Advances in Experimental Medicine and Biology 680

Hamid R. Arabnia
Editor

Advances in Computational Biology

Springer

Advances in Experimental Medicine and Biology

For further volumes:
http://www.springer.com/series/5584

Advances in Experimental Medicine and Biology

Volumes published in the series

Volume 672
BIOSURFACTANTS
Edited by Ramkrishna Sen

Volume 673
MODELLING PARASITE TRANSMISSION AND CONTROL
Edited by Edwin Michael and Robert C. Spear

Volume 674
INTEGRINS AND ION CHANNELS: MOLECULAR COMPLEXES AND SIGNALING
Edited by Andrea Becchetti and Annarosa Arcangeli

Volume 675
RECENT ADVANCES IN PHOTOTROPHIC PROKARYOTES
Edited by Patrick C. Hallenbeck

Volume 676
POLYPLOIDIZATION AND CANCER
Edited by Randy Y.C. Poon

Volume 677
PROTEINS: MEMBRANE BINDING AND PORE FORMATION
Edited by Gregor Anderluh and Jeremy Lakey

Volume 678
CHEMO FOG: CANCER CHEMOTHERAPY RELATED COGNITIVE IMPAIRMENT
Edited by Robert B. Raffa and Ronald J. Tallarida

Volume 679
MIPS AND THEIR ROLE IN THE EXCHANGE OF METALLOIDS
Edited by Thomas P. Jahn and Gerd P. Bienert

Volume 680
ADVANCES IN COMPUTATIONAL BIOLOGY
Edited by Hamid R. Arabnia

Hamid R. Arabnia
Editor

Advances in Computational Biology

Springer

Editor
Hamid R. Arabnia
University of Georgia
Dept. Computer Science
30602 7404 Athens Georgia
415 Boyd Graduate Studies
Research Ctr.
USA

ISSN 0065 2598
ISBN 978 1 4419 5912 6 e ISBN 978 1 4419 5913 3
DOI 10.1007/978 1 4419 5913 3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2010934907

Printed on acid free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

It gives us great pleasure to introduce the book entitled “Advances in Computational Biology." The book is composed of a collection of papers received in response to an announcement that was widely distributed to academicians and practitioners in the broad area of computational biology. Also, selected authors of accepted papers of BIOCOMP’09 proceedings (International Conference on Bioinformatics and Computational Biology: July 13 16, 2009; Las Vegas, NV, USA) were invited to submit the extended versions of their papers for evaluation.

The collection of papers in Part I, present advances in Bioinformatics Databases, Data Mining, and Pattern Discovery. Microarray, Gene Expression Analysis, and Gene Regulatory Networks are presented in Part II. Followed by Part III that is composed of a collection of papers in the areas of Protein Classification and Structure Prediction, and Computational Structural Biology. Part IV is composed of a number of articles that discusses significant issues in the areas of Comparative Sequence, Genome Analysis, Genome Assembly, and Genome Scale Computational Methods. Drug Design, Drug Screening, and related topics are discussed in Part V. Part VI presents various Computational Methods and Diagnostic Tools in Biomedical applications. Lastly, general topics in the area of Bioinformatics are presented in Part VII.

We are very grateful to the many colleagues who helped in making this project possible. In particular, we would like to thank the members of the BIOCOMP’09 (http://www.world-academy-of-science.org/) Program Committee who provided their editorial services for this book project. The BIOCOMP’09 Program Committee members were:

- *Dr. Niloofar Arshadi, University of Toronto, ON, Canada*
- *Prof. Ruzena Bajcsy, Member, National Academy of Engineering; IEEE Fellow; ACM Fellow; University of California, Berkeley, USA*
- *Prof. Alessandro Brawerman, CEO, Globaltalk, Brazil and Positivo University, Brazil*

- *Ashu M. G. Solo (Publicity Chair), Principal/R&D Eng., Maverick Technologies America Inc.*
- *Dr. Quoc-Nam Tran, Lamar University, Texas, USA*
- *Prof. Alicia Troncoso, Head of CS Dept., Pablo de Olavide University, Seville, Spain*
- *Dr. Mary Qu Yang (Vice Chair), National Institutes of Health, USA*
- *Prof. Lotfi A. Zadeh, Member, National Academy of Engineering; IEEE Fellow, ACM Fellow, AAAS Fellow, AAAI Fellow, IFSA Fellow; Director, BISC; University of California, Berkeley, California, USA*
- *Prof. Yanqing Zhang, Georgia State University, Atlanta, Georgia, USA*
- *Dr. Leming Zhou, University of Pittsburgh, Pittsburgh, PA, USA*
- *Members of WORLDCOMP Task Force for Artificial Intelligence*
- *Members of WORLDCOMP Task Force for Computational Biology*
- *Members of WORLDCOMP Task Force for High-Performance Computing*
- *Members of WORLDCOMP Task Force for Pattern Recognition*

Last but not least, we would like to thank the editorial staff at Springer; in particular, we are indebted to Ms. Melanie Wilichinsky (the editorial manager) and her assistant, Ms. Meredith Clinton, who provided excellent professional service to us to make this project feasible.

Hamid R. Arabnia, PhD

Professor, Computer Science; Fellow, Int'l Society of Intelligent Biological Medicine (ISIBM); Editor-in-Chief, The Journal of Supercomputing (Springer); Co-Editor/Board, Journal of Computational Science (Elsevier); Member, Advisory Board, IEEE Technical Committee on Scalable Computing (TCSC); Department of Computer Science, The University of Georgia, Athens, GA 30602-7404, USA

Contents

Part I
Bioinformatics Databases, Data Mining, and Pattern Discovery

Chapter 1
A Classification Method Based on Similarity Measures of Generalized Fuzzy Numbers in Building Expert System for Postoperative Patients

Pasi Luukka

Abstract In this research, we concentrate to build an expert system for a problem where task is to determine where patients in a postoperative recovery area should be sent to next. Data set created from postoperative patients is used to build proposed expert system to determine, based on hypothermia condition, whether patients in a postoperative recovery area should be sent to Intensive Care Unit (ICU), general hospital floor, or go home. What makes this task particularly difficult is that most of the measured attributes have linguistic values (e.g., stable, moderately stable, unstable, etc.). We are using generalized fuzzy numbers to model the data and introduce new fuzzy similarity based classification procedure which can deal with these linguistic attributes and classify them accordingly. Results are compared to existing result in literature, and this system provides mean classification accuracy of 66.2% which compared well to earlier results reported in literature.

Keywords Classification methods · Expert systems · Generalized fuzzy numbers · ICU · Postoperative patients · Similarity measures

1.1 Introduction

In this chapter, we have tackled the situation where the problem is to build an expert system which can make a decision of where patients in a postoperative recovery area should be sent to next. Expert system will learn to make the decision with the help of existing data. What makes this problem especially challenging and also interesting is that, data is represented with linguistic statements such as that patients internal temperature can be, e.g., low, mid, or high. In this case, the data, which we

P. Luukka
Laboratory of Applied Mathematics, Lappeenranta University of Technology, PO Box 20, FIN 53851 Lappeenranta, Finland
e mail: pasi.luukka@lut.fi

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 1,

use, is mostly in the form of experts linguistic assessments instead of crisp measured numbers. Most of the uncertainty in the data, in this case, stems for linguistic attributes and also from the fact that data is collected from patients and humans are not exactly alike. So in equal or comparable situation, the data collected does not show identical states if we try to measure the data too precisely. Because of this, the linguistic representation seems to be a better choice. This kind of uncertainty is typically found with data taken from living beings and reflects the rich variability of nature.

In expert systems dealing with the analysis of the data and learning from it to make decisions, we are faced with situation where data is almost always burdened with uncertainty of different kinds. Possibly the uncertainty is even increased by the method of analysis applied [2]. Fuzzy logic is a logic that allows vagueness, imprecision, and uncertainty. Zadeh [10] introduced the concept of fuzzy sets and the respective theory that can be regarded as the extension of the classical set theory. One of the fundamental mathematical constructs is the similarity measure. In the same way, as the notion of the fuzzy subset generalizes that of the classical subset, the concept of similarity can be considered as being a multivalued generalization of the classical notion of equivalence [11].

In this chapter, we introduce new fuzzy similarity-based classification procedure for the problem at hand. Earlier methods to tackle this task at hand used Linguistic Hard C-Means (LHCM) [1], regular Hard C-Mean (HCM) with two variants, learning system LERS (Learning from Examples based on Rough Sets) [9], and combination where this data is first preprocessed using PCA for fuzzy data [4], and then it is defuzzified to get it in crisp form and after this similarity classifier [5] is used to make the classification decision. While the last of these methods [5] manage to get the highest mean classification accuracy, it suffers from one drawback which can affect the decision in this type of case where we are dealing with linguistic attributes. In the method described in [5], PCA for fuzzy data [4] can deal with the linguistic data, but for similarity classifier [6] we need to defuzzify the data first in order to use it in proposed form. In defuzzification, we are basically transforming the data into a single number. The problem with this approach is that it loses information, and this information loss can effect the classification decision. In this chapter, we propose a new similarity-based classification system, which can now deal with similarities of generalized fuzzy numbers, and hence does not need the defuzzification which was the drawback of earlier method. This way we can avoid the possible or even likely information loss occurring with expert system introduced in [5] and should be able to have a more efficient expert system for problem at hand.

1.2 Postoperative Patient Data Set

Postoperative patient data set is freely available in [7]. Creators of this data set are Sharon Summers, School of Nursing, University of Kansas Medical Center and Linda Woolery, School of Nursing, University of Missouri. In this data, problem

Table 1.1 Postoperative patient data sets linguistic attributes

Attribute	L CORE	L SURF	L 02	L BP	SURF STBL	CORE STBL	BP STBL
Linguistic label	Low Mid High	Low Mid High	Poor Fair Good Excellent	Low Mid High	Unstable Mod stable Stable	Unstable Mod stable Stable	Unstable Mod stable Stable

is to determine where patients in a postoperative recovery area should be sent to next. Because hypothermia is a significant concern after surgery, the attributes correspond roughly to body temperature measurements. There are 90 samples in this data and seven linguistic attributes:

1. L-CORE (patient's internal temperature): (high (>37), mid (≥ 36 and ≤ 37), low (<36))
2. L-SURF (patient's surface temperature): (high (>36.5), mid (≥ 36.5 and ≤ 35), low (<35))
3. L-O2 (oxygen saturation in %):(excellent (≥ 98), good (≥ 90 and <98), fair (≥ 80 and <90), poor (<80))
4. L-BP (last measurement of blood pressure): (high ($>130/90$), mid ($\leq 130/90$ and $\geq 90/70$), low ($<90/70$))
5. SURF-STBL (stability of patient's surface temperature): (stable, mod-stable, unstable)
6. CORE-STBL (stability of patient's core temperature) (stable, mod-stable, unstable)
7. BP-STBL (stability of patient's blood pressure) (stable, mod-stable, unstable)

In addition, there is also one numerical attribute named COMFORT, which is patient's perceived comfort at discharge, measured as an integer between 0 and 20. Moreover, there is also label information decision such as whether patient is sent to intensive care unit (ICU), prepared to go home, or be sent to general hospital floor. In Table 1.1, linguistic values of each attribute are given. Linguistic attributes are then presented as generalized fuzzy numbers which are used by the classification system (see [5] for more about this issue).

1.3 Similarity Measures of Generalized Fuzzy Numbers

Similarity measure, which is used in the similarity classifier, now needs to be in such form, that it can deal with generalized fuzzy numbers. Wei and Chen introduced in [8] such a similarity measure. First, we briefly review basic concepts of generalized fuzzy numbers from Chen [3], and then we introduce the similarity measure they proposed. Chen [3] represented a generalized trapezoidal fuzzy number $\tilde{A}$ as $\tilde{A} = (a, b, c, d; w)$, where a, b, c, and d are real values and $0 < w \leq 1$ as shown in Fig. 1.1. The membership function $\mu_{\tilde{A}}$satisfies the usual needed

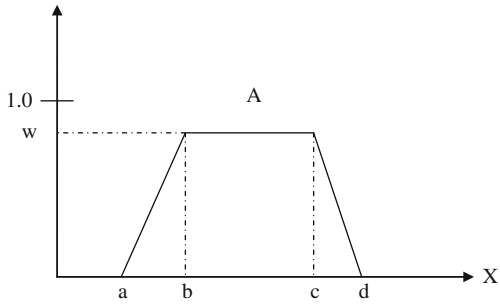


Fig. 1.1 A generalized trapezoidal fuzzy number $\tilde{A} \quad (a, b, c, d; w)$

conditions (see [8]). Assume we have two generalized trapezoidal fuzzy numbers $\tilde{A}$ and $\tilde{B}$, where $\tilde{A} = (a_1, a_2, a_3, a_4; w_a)$ and $\tilde{B} = (b_1, b_2, b_3, b_4; w_b)$.

Now, for example, addition for the generalized trapezoidal fuzzy numbers is

$$\begin{aligned} \tilde{A} + \tilde{B} &= (a_1, a_2, a_3, a_4; w_a) + (b_1, b_2, b_3, b_4; w_b) \\ &= (a_1 + b_1, a_2 + b_2, a_3 + b_3, a_4 + b_4, \min(w_a, w_b)) \end{aligned} \tag{1.1}$$

For scalar multiplication, subtraction, multiplication, and division of fuzzy numbers see, for example, [3] or [8]. Degree of similarity between two generalized fuzzy numbers can be computed by following the similarity introduced in [8] as

$$S(\tilde{A}, \tilde{B}) = \left(1 - \frac{\sum_{i=1}^{4} |a_i - b_i|}{4}\right) \times \frac{\min(P(\tilde{A}), P(\tilde{B})) + \min(w_a, w_b)}{\max(P(\tilde{A}), P(\tilde{B})) + \max(w_a, w_b)} \tag{1.2}$$

where $S(\tilde{A}, \tilde{B}) \in [0, 1]$; $P(\tilde{A})$ and $P(\tilde{B})$are defined as follows:

$$P(\tilde{A}) = \sqrt{(a_1 - a_2)^2 + w_a^2} + \sqrt{(a_3 - a_4)^2 + w_a^2} + (a_3 - a_2) + (a_4 - a_1) \tag{1.3}$$

$$P(\tilde{B}) = \sqrt{(b_1 - b_2)^2 + w_b^2} + \sqrt{(b_3 - b_4)^2 + w_b^2} + (b_3 - b_2) + (b_4 - b_1) \tag{1.4}$$

The larger the value of $S(\tilde{A}, \tilde{B})$ is, the more the similarity is between the generalized fuzzy numbers $\tilde{A}$ and $\tilde{B}$.

1.4 Similarity Classifier of Generalized Fuzzy Numbers

We would like to classify a set X of objects into N different classes $C_1, \ldots, C_N$ by their attributes. We suppose that t is the number of different kinds of attributes $f_1, \ldots, f_t$ that we can measure from objects. We suppose that the value for the

magnitude of each attribute is normalized. The objects we want to classify are basically vectors consisting of generalized fuzzy numbers. First one must determine for each class the ideal vector $v_i = (v_i(f_1), \ldots, v_i(f_t))$ that represents class i as well as possible. This vector can be user-defined or calculated from some sample set X_i of vectors $x = (x(f_1), \ldots, x(f_t))$ which are known to belong to class C_i. In this expert system, we used mean values to calculatev_i, which is

$$v_i(r) = \left(\frac{1}{\#X_i} \sum_{x \in X_i} x(f_r) \right) \quad \forall r = 1, \ldots, t \tag{1.5}$$

This can be done in practice simply by applying formula (1.1) and scalar multiplication to calculate mean values. Once the ideal vectors, which now in practice the mean vectors of generalized fuzzy numbers have been determined, then the decision to which class an arbitrarily chosen $x \in X$ belongs to is made by comparing it to each ideal vector. The comparison can be done, e.g., by using similarity measure (1.2) introduced in [8]

$$S\langle x, v\rangle = \left(\frac{1}{t} \sum_{r=1}^{t} S_r\langle x, v\rangle \right) \tag{1.6}$$

where

$$S_r\langle x, v\rangle = \left(1 - \frac{\sum_{i=1}^{4} |x_{ri} - v_{ri}|}{4} \right) \times \frac{\min(P(x_r), P(v_r)) + \min(w_{x_r}, w_{v_r})}{\max(P(x_r), P(v_r)) + \max(w_{x_r}, w_{v_r})} \tag{1.7}$$

Now, when the similarities between the fuzzy sample vector, which we want to classify, and the ideal vectors have been computed, then we decide that $x \in C_m$ if

$$S\langle x, v_m\rangle = \max_{i=1,\ldots,N} S\langle x, v_i\rangle \tag{1.8}$$

So in other words, sample belongs to the class in which it has the highest similarity value.

1.5 Experimental Results and Comparison

The postoperative patient's data set was first represented in trapezoidal fuzzy numbers. After this, we can use it directly with our newly derived classifier or we can first preprocess this data by using principal component analysis for fuzzy data

[4], and then use similarity classifier for generalized fuzzy numbers. Both results are calculated. PCA is considered to be the best linear dimension reduction technique in the mean-square error sense so to use it in preprocessing step seems reasonable. Since there was no prior knowledge of what would be the best reduced dimension, the dimension reduction was done for all dimensions from two to seven. In the classifier, data was splitted half. One half for the training and the other half for the testing. This procedure for randomly splitting data in half was done 100 times and mean classification accuracies were computed.

Classification results can be seen in Table 1.2. As can be seen from Table 1.2, the highest mean accuracy of 66.22% was achieved with dimension two. Variance with this dimension was 0.0270. If we present mean classification accuracy in 99% confidence interval using Student's t distribution $\mu \pm t_{1\ \alpha/2} S_\mu / \sqrt{n}$, we get accuracy of 66.22 ± 4.04. If we compare the results with the results achieved without preprocessing, we notice that preprocessing with PCA for fuzzy data managed to enhance the results by 6% units.

Classification results are compared to results presented in literature in Table 1.3. There results are compared with the that of the earlier methods such as Linguistic Hard C-Means (LHCM) [1], regular Hard C-Mean (HCM) with two variants, and learning system LERS (Learning from Examples based on Rough Sets) [9]. In regular HCM, numeric data set was used (HCM1). There linguistic values were mapped as a numeric value, e.g., L-COREs value low, was simply 34, etc. (see more about that in [1]). Centroids from fuzzy numbers were used as numerical values in the second case (HCM2).

As can be seen from the results comparison Table 1.3, the method based on PCA for fuzzy data and similarity classifier for generalized fuzzy numbers are

Table 1.2 Classification results for postoperative patient data set where data is first preprocessed using PCA for fuzzy data and after this classified using new similarity classifier for generalized fuzzy numbers. In first column, there is number of dimensions used, in second column maximum classification accuracy (in %), in third column mean accuracies are reported and for fourth column variances. In the last row also, results from similarity classifier using generalized fuzzy numbers without preprocessing is reported

Method	Dimension	Max (%)	Mean (%)	Variance
PCA + SIM	2	82.22	66.22	0.0270
	3	73.33	61.63	0.0331
	4	82.22	63.11	0.0339
	5	73.33	63.63	0.0224
	6	84.44	61.26	0.0252
	7	73.33	63.93	0.0247
SIM	7	71.11	59.70	0.0187

Table 1.3 Classification results comparison for postoperative patient data set. In first row, method is reported and in second row, classification accuracy

Method	LERS	HCM1	HCM2	LHCM	PCA + SIM	PCA + new SIM
Accuracy	48%	50%	51.1%	53.3%	62.7%	66.2%

managed with the highest accuracy of 66.2%, whereas the second highest mean accuracy was achieved with PCA for fuzzy data and similarity classifier using Łukasiewics similarity for crisp data where defuzzification was needed. The third highest accuracy was achieved with Linguistic Hard C-Mean with accuracy of 53.3%.

1.6 Discussion

In this chapter, we are dealing with a decision making system which can cope with highly heterogeneous information. Here, information does not come only in crisp measured numbers, but it can also deal with interval type number, and moreover linguistic data that come from experts, e.g., doctors. Such a heterogeneous data is first represented by generalized fuzzy numbers. After this, PCA for fuzzy numbers is used to make a linear combination of it to a lower dimensional subspace. In third step, earlier a defuzzification was needed for the older similarity classifier but now new similarity classifier for generalized fuzzy numbers is given, and it can be used directly to get the classification decision. In doing so, we can avoid the possible and even likely information loss coming from defuzzification and hence obtain more accurate results. This new procedure was applied to a problem where we needed to make a decision of where patients in a postoperative recovery area should be sent to next. Using this newly derived similarity classifier for generalized fuzzy numbers, we can avoid information loss coming from defuzzification, and we can manage to enhance previous results. Proposed method managed to classify with 66.2% mean classification accuracy where previous methods managed with 62.7% and third highest mean accuracy was 53.3%. Results show that the method clearly has potential in this area.

References

1. Auephanwiriyakul, S., Theera Umpon, N. (2004). Comparison of linguistic and regular hard C means in postoperative patient data. Journal of Advanced Computational Intelligence and Intelligent Informatics 8(6), pp. 599 605.
2. Bandemer, H., Näther, W. (1992). Fuzzy data analysis, Kluwer Academic Publisher, Dordrecht.
3. Chen, S.H. (1999). Ranking generalized fuzzy number with graded mean integration. In Proceedings of the eighth international fuzzy systems association world congress, Vol. 2, pp. 899 902. Taipei, Taiwan, Republic of China.
4. Denoeux, T., Masson, MH. (2004). Principal component analysis of fuzzy data using auto associative neural networks. IEEE Transactions on Fuzzy Systems 12(3), pp. 336 349.
5. Luukka, P. (2009), PCA for fuzzy data and similarity classifier in building recognition system for post operative patient data. Expert Systems with Applications 36, pp. 1222 1228.
6. Luukka, P., Leppälampi, T. (2006). Similarity classifier with generalized mean applied to medical data. Computers in Biology and Medicine 36, pp. 1026 1040.

7. Newman, D.J., Hettich, S., Blake, C.L., Merz, C.J. (1998). UCI Repository of machine learning databases [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, Department of Information and Computer Science.
8. Wei, S.H., Chen, S.M. (2009). A new approach for fuzzy risk analysis based on similarity measures of generalized fuzzy numbers. Expert Systems with Applications 36, pp. 589 598.
9. Woolery, L., Grzymala Busse, J., Summers, S., Budihardjo, A. (1991). The use of machine learning program $LERS_{LB}$ 2.5 in knowledge acquisition for expert system development in nursing. Computers in Nursing 9, pp. 227 234.
10. Zadeh, L. (1965). Fuzzy sets. Information and Control 8, pp. 338 353.
11. Zadeh, L. (1971). Similarity relations and fuzzy orderings. Information Science 3, pp. 177 200.

Chapter 2
Efficient Support Vector Machine Method for Survival Prediction with SEER Data

Zhenqiu Liu, Dechang Chen, Guoliang Tian, Man-Lai Tang, Ming Tan, and Li Sheng

Abstract Support vector machine (SVM) is a popular method for classification, but there are few methods that utilize SVM for survival analysis in the literature because of the computational complexity. In this paper, we develop a novel L_1 penalized SVM method for mining right-censored survival data (L_1 SVMSURV). Our proposed method can simultaneously identify survival-associated prognostic factors and predict survival outcomes. It is easy to understand and efficient to use especially when applied to large datasets. Our method has been examined through both simulation and real data, and its performance is very good with limited experiments.

Keywords Support vector machine · SVM · Survival analysis · Prognostic factors · SEER

2.1 Introduction

Survival prediction and prognostic factor identification play an important role in medical research. Survival data normally include a variable indicating whether some outcome under consideration (such as death or recurrence of a disease) has occurred within a specific follow-up time. The modeling technique has to consider that for some patients the follow-up may end before the event occurs. In other words, we must take into account patients for whom the event has not occurred during the follow-up period but might have occurred just after it. This makes it more difficult to apply a standard machine learning method for survival prediction.

Z. Liu
University of Maryland at Baltimore, Baltimore, MD, USA
e mail: zliu@umm.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 2,

Many models for survival prediction have been proposed in the statistical literature. However, most of them are designed for small datasets and not suitable for large data mining. The most popular one is Cox proportional hazards model [1, 2, 7, 14], in which model parameters are estimated with partial log likelihood maximization. Another one is the accelerate failure time (AFT) model [5, 15, 16]. AFT is a linear regression model in which the response variable is the logarithm or a known monotone transformation of a failure (death) time. Even though the semi-parametric estimation of an AFT model with an unspecific error distribution has been studied extensively in the literature, the model has not been widely used in practice, mainly due to the difficulties in computing model parameters [4]. Recently, AFT has been applied to gene expression data with a small size but a large dimension [9, 11]. Survival prediction with the area under the ROC curve (AUC) maximization has also been proposed [8] for high-dimensional gene expression data with a small size. However, it is much more difficult to apply these methods for survival data with a large size and a high dimension.

The size in a retrospective surveillance, epidemiology and end results (SEER) dataset is usually very large. For example, the lung cancer data set from SEER that contains the records of lung cancer patients who were diagnosed from 1973 to 2002 has more than 500,000 patients; and the breast cancer data set also has more than 500,000 patients at the same time span. Cancer data also contain a high percentage of censored observations. With the above mentioned lung cancer data set, for instance, 34% of total records involve censored times. Ignoring records that contain censored observations or treating censored data as the actual life-times will produce biases in survival modeling. Traditional statistical methods fail to deal with large survival data sets. New methods are required for mining the SEER data efficiently.

Though there are many publications for L_1SVM in the classification framework [10], there are few SVM methods for survival analysis because of the computational complexity. In this paper, we propose a novel L_1 [12, 13] SVM approach for survival outcome predictions (L_1 SVMSURV). At the same time, L_1 SVMSURV is utilized to automatically identify survival-associated prognostic factors. The proposed models are evaluated with simulation and real data using the global AUC summary (GAUCS) measure [3]. The paper is organized as follows. In Sect. 2.2, we formulate L_1 SVMSURV under the accelerate failure time framework and introduce the GAUCS measure. Experiments with simulation and real survival data for model performance evaluation are given in Sect. 2.3. Finally, conclusions and remarks are provided in Sect. 2.4.

2.2 L_1 SVMSURV for Censored Survival Outcomes

Consider a set of n independent observations $\{T_i, \delta_i, \mathbf{x}_i\}_{i=1}^n$, where δ_i is the censoring indicator and T_i is the survival time (event time) if $\delta_i = 1$ or censored time if $\delta_i = 0$, and $x_i = (x_{i1}, x_{i2}, \ldots, x_{im})^t$ is the m-dimensional input vector of the ith sample. Let

$\mathbf{w} = (w_1, w_2, \ldots, w_m)^t$ be a vector of regression coefficients and $\phi(\mathbf{x}_i)$ be the nonlinear transformation of $\mathbf{x}_i$ in the feature space. The AFT model is defined as

$$M(\mathbf{x}_i) = \mathbf{w}^t \phi(\mathbf{x}_i), \quad i = 1, \ldots, n, \tag{2.1}$$

where $M(\mathbf{x}_i) > logT_i$ if $\delta_i = 0$ and $M(\mathbf{x}_i) = logT_i$ if $\delta_i = 1$. Because there are both equality and inequality constraint in the model, simple least square solution will fail to work. We use $L_1 = \sum_{i=1}^{n} |w_i|$ as the penalty for the sparse solution and have the quadratic L_1 SVMSURV:

$$\begin{gathered} min \frac{1}{2n} \sum_{i=1}^{n} \xi_i^2 + \lambda \sum_{i=1}^{n} |w_i| \\ \text{s.t. } |\mathbf{w}^t \phi(\mathbf{x}_i) - logT_i| < \xi_i, \quad \text{if} \quad \delta_i = 1, \\ \mathbf{w}^t \phi(\mathbf{x}_i) > logT_i - \xi_i, \quad \text{if} \quad \delta_i = 0 \\ \xi_i \geqslant 0, \quad \forall\, 1 \leqslant i \leqslant n. \end{gathered} \tag{2.2}$$

When ties in the event times are presented, variables associated with each tied time appear in the constraints independently. We can define an index function $I(\delta_i) = 1$ if $\delta_i = 1$, and I is defined as $I(\delta_i) = 1$ if $logT_i \geqslant \mathbf{w}^t \phi(\mathbf{x}_i)$ and 0, otherwise, when $\delta_i = 0$. Then, we have

$$J(\mathbf{w}; \lambda) = \frac{1}{2n} \sum_{i=1}^{n} I(\delta_i) \{\mathbf{w}^t \phi(\mathbf{x}_i) - logT_i\}^2 + \lambda \sum_{i=1}^{n} |w_i|. \tag{2.3}$$

Since $|w_i|$ does not have the first order derivative at 0, we will use the similar procedure proposed by Liu et al. [8] for parameter estimation. Rewrite $J(\mathbf{w}, 0)$ as a function of the kth parameter w_k, and let the remaining parameters $\mathbf{w}_{\ k}$ be fixed. We have

$$\begin{aligned} J(\mathbf{w}; 0) &= \frac{1}{2n} \sum_{i} I(\delta_i) (\mathbf{w}^t \phi(\mathbf{x}_i) - logT_i)^2, \\ &= \frac{1}{2} b_k w_k^2 + c_k w_k + d_k, \end{aligned} \tag{2.4}$$

where

$$\begin{aligned} b_k &= \tfrac{1}{n} \sum_{i} I(\delta_i) \phi_k^2(\mathbf{x}_i), \\ c_k &= \tfrac{1}{n} \sum_{i} I(\delta_i) (logT_i - \mathbf{w}^t{}_k \Phi_{\ k}(\mathbf{x}_i(\mathbf{x}_i)), \\ d_k &= \tfrac{1}{2n} \sum_{i} I(\delta_i) (logT_i - \mathbf{w}^t{}_k \Phi_{\ k}(\mathbf{x}_i))^2. \end{aligned}$$

Equation (2.4) is a quadratic function of w_k, and we have the following first order derivative w.r.t. w_k:

$$\frac{\partial J(\mathbf{w};0)}{\partial w_k} = b_k w_k + c_k.$$

Since $\mathbf{w}$ is not differentiable at 0, the first derivative of $J(\mathbf{w},\lambda)$ is a step function:

$$\partial_{w_k} J(\mathbf{w};\lambda) = \begin{cases} \{(b_k w_k - c_k) - \lambda\}, & w_k < 0 \\ [-c_k - \lambda, -c_k + \lambda], & w_k = 0 \\ \{(b_k w_k - c_k) + \lambda\}, & w_k > 0 \end{cases} \tag{2.5}$$

We therefore can update each coefficient w_i with the following equation:

$$w_k(c_k) = \begin{cases} (\lambda + c_k)/b_k, & c_k < -\lambda \\ 0, & c_k \in [-\lambda, \lambda] \\ (-\lambda + c_k)/b_k, & c_k > \lambda \end{cases} \tag{2.6}$$

The model performance is evaluated by the GAUCS measure. In survival analysis, the AUC is a time dependent measure. We will utilize the GAUCS measure for the comparison of model performance. The GAUCS is defined by averaging over t: GAUCS $= 2\int AUC(t)g(t)S(t)\,\mathrm{d}t = Pr(M_j > M_k | t_j < t_k)$, which indicates the probability that the subject who died (case) at an earlier time has a larger value of the risk score. In the above equation, $S(t)$ and $g(t)$ are the survival and corresponding density functions, respectively.

2.3 Computational Results

Simulation data: Simulation studies were conducted to evaluate the performance of the proposed method under different assumptions. The following describes the method to generate input data with censored survival outcomes that emulate the mechanisms presented by the actual data. We first sample 12-dimensional input data $\mathbf{x}$ with 10,000 training and test samples, respectively, from a multivariate normal distribution with zero means and variance covariance matrix Σ. Σ is set to have the same correlation coefficient ρ for all input variables and different $\rho = 0, 0.2, 0.4, 0.6,$ and, 0.8 will be chosen to assess the performance of the proposed method. We then choose model parameters $\mathbf{w} = [-1.9, 2.8, 1.7, 0, 0, 0, 0, 0, 0, 0, 0, 0]^T$ and calculate $Z = \mathbf{w}^t\mathbf{x} + \varepsilon$. Hence, this model is only associated with the first three input variables plus random noise. Finally, we compute $H = Aexp(-0.5Z)$ with $A = 100$, sample the survival time T_i from Weibull random number generating function, and build $d_i = (\mathrm{rand}(1,1) + C) * T_i$, such that the censoring status is $\delta_i = T_i < d_i$. Different Cs give different portions of censored data.

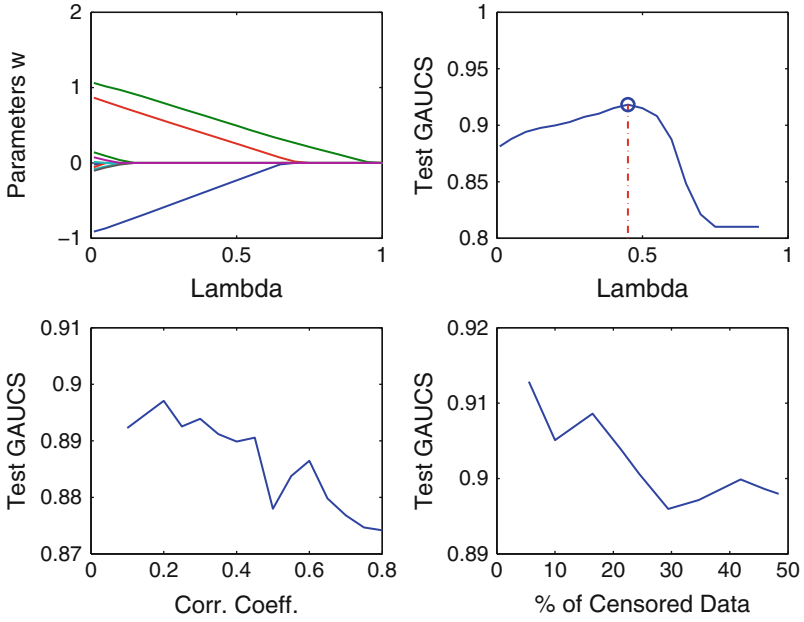


Fig. 2.1 Regularization path and test GAUCS

We analyze the simulation data with L_1 SVMSURV and build the model with the training data and evaluate the performance of the model with the test data. The regularization parameter λ is determined with tenfold cross-validation with training data only. To prevent bias arising from the specific data, we simulate the data 100 times. The average computational results are reported in Fig. 2.1.

The upper left subfigure is the regularization path with different λs. It shows that only three nonzero w_is are left when $\lambda \geqslant 0.1$ and all w_is go to zero when $\lambda = 1$. The upper right subplot shows that the value of the optimal test GAUCS is 0.9183 with $\lambda = 0.45$ for simulation data with no censoring. Combining both upper subplots, we conclude that our proposed method correctly selects the three variables. The lower left subplot shows the average value of the test GAUCS with different correlation coefficients ρ among input variables. As correlations among input variables increase from 0.1 to 0.8, the average value of the test GAUCS over 100 test data sets decreases from 0.898 to 0.873. Finally, the lower right subplot shows that the value of the test GAUCS also decreases with the increase of the percent of censored data, but the decrease is not statistical significant, which indicates that the proposed model is robust.

SEER prostate data: Prostate cancer is the most common cancer, other than skin cancers, and the second leading cause of cancer death in American men, behind only lung cancer. In this paper, we study the SEER prostate cancer registry data from 2000 to 2003 in the states of Greater California, Kentucky, Louisiana, and New Jersey. There are in total 104,363 patients in the database. For the comparison

purpose, we only study 13,975 samples which do not contain missing tumor size information. Fifteen potential prognostic factors are included in the model, including age, race, married status, grade, size of tumor, clinical extension of tumor, lymph node involved, number of positive nodes examined, number of nodes examined, surgery performed, radiation, radiation sequence with surgery, stage (TNM), PSA marker, and number of primaries. We divide the data into training and test data with a roughly equal size. The regularization parameter is again determined by tenfold cross-validation. To prevent bias and overfitting from a specific grouping, we partition the data 100 times and the average λ and test GAUCS values are shown in Fig. 2.2.

The left subplot shows that the optimal test GAUCS $= 0.7707$ is reached at $\lambda = 0.05$. There are five prognostic factors associated with survival, i.e., age (0.179 ± 0.002), tumor size (-0.002 ± 0.0003), the clinical extension of tumor (-0.001 ± 0.0002), radiation (0.1 ± 0.007), and stage (-0.06 ± 0.0008), where the values in the parenthesis are the mean and standard deviation of w_i for each prognostic factor. They indicate that patients who have cancer in their later age and those who have been exposed to radiations may survive longer and patients with a large tumor size, clinical extension of tumor, and late stage may die earlier. Our conclusion that younger men with prostate cancer have shorter survival times is consistent with the finding in [6]. While the reasons for this unexpected but excited finding are not clear, one explanation may be that young men with prostate cancer may have biologically more aggressive forms of the disease than the forms

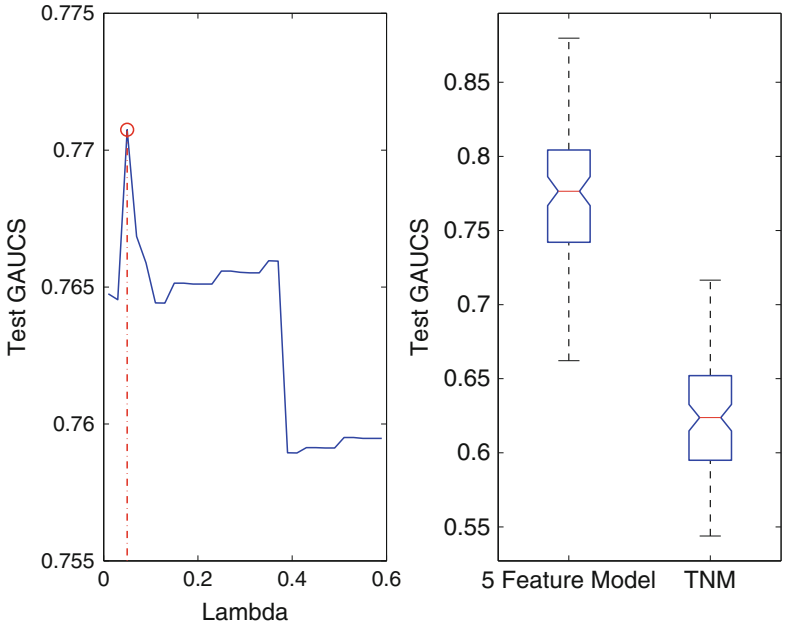


Fig. 2.2 Average test GAUCS with SEER data

diagnosed in older men. Additional studies are needed to determine what, if any, underlying differences exist between prostate cancer found in young men and that found in older men. These studies may help clinicians improve screening in young men and could ultimately lead to the development of better treatment strategies for young patients. Finally, the right subplot shows the performance comparison of our proposed model with the TNM stage system. It is seen that the average value of the test GAUCS increases roughly 22% from 0.62 to 0.77.

2.4 Conclusions and Remarks

Analysis of censored failure time data containing a huge number of samples is important in practice, especially for data containing millions of patients' records. How to mine these databases and identify important prognostic factors presents a class of interesting and challenging questions. In this paper, we propose a L_1 penalized SVM method for simultaneous variable selection and estimation. The simulation studies and the real example on prostate cancer illustrate that the proposed method can effectively reduce the dimension of the input variables and select important survival-associated prognostic factors while providing satisfactory estimation and prediction. More work can be done regarding improvement and validation of the proposed method. This constitutes our future work. For example, the asymptotic properties of the model will be studied in our future work, and we will apply the proposed method to other cancer data sets.

Acknowledgment This work was partially supported by NIH Grant 1R03CA133899 01A210 and NSF CCF 0729080.

References

1. Cox DR (1972). Regression models and life tables (with discussion). Journal of Royal Statistical Society, Series B 34:187 220.
2. Gui J, Li H (2005). Penalized Cox regression analysis in the high dimensional and low sample size settings, with applications to microarray gene expression data. Bioinformatics 21:3001 3008.
3. Heagerty PJ, Zheng Y (2005). Survival model predictive accuracy and ROC curves. Biometrics 61(1):92 105.
4. Jin Z, Lin DY, Wei LJ, Ying ZL (2003). Rank based inference for the accelerated failure time model. Biometrika 90:341 353.
5. Kalbfleisch JD, Prentice RL (1980). The Statistical Analysis of Failure Time Data. New York: John Wiley.
6. Lin DW, Porter M, Montgomery B (2009). Treatment and survival outcomes in young men diagnosed with prostate cancer: a Population based Cohort Study. Cancer 115 (13):2863 2871.
7. Liu Z, Jiang F (2009). Gene identification and survival prediction with Lp penalty and novel similarity measure. International Journal of Data Mining and Bioinformatics 3(4):398 408.

8. Liu Z, Gartenhaus RB, Chen X, Howell C, Tan M (2009). Survival prediction and gene identification with penalized global AUC maximization. Journal of Computational Biology 16(12):1661 1670.
9. Ma S, Huang J (2007). Additive risk survival model with microarray data. BMC Bioinformat ics 8:192.
10. Mangasarian OL (2006). Exact 1 norm support vector machines via unconstrained convex differentiable minimization. Journal of Machine Learning Research 7:1517 1530.
11. Sha N, Tadesse MG, Vannucci M (2006). Bayesian variable selection for the analysis of microarray data with censored outcomes. Bioinformatics 22(18):2262 2268.
12. Tibshirani R (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B 58(1):267 288.
13. Tibshirani R (1997). The lasso method for variable selection in the Cox model. Statistics in Medicine 16(4):385 395.
14. Van Houwelingen HC, et al. (2006). Cross validated Cox regression on microarray gene expression data. Statistics in Medicine 25:3201 3216.
15. Wei LJ (1992). The accelerated failure time model: a useful alternative to the Cox regression model in survival analysis. Statistics in Medicine 11:1871 1879.
16. Ying ZL (1993). A large sample study of rank estimation for censored regression data. Annals of Statistics 21:76 99.

Chapter 3
Searching Maximal Degenerate Motifs Guided by a Compact Suffix Tree

Hongshan Jiang, Ying Zhao, Wenguang Chen, and Weimin Zheng

Abstract Compared to a mismatched consensus motif, a degenerate consensus motif is more suitable for modeling position-specific variations within motifs. In the literature, the state-of-art methods using degenerate consensus motifs for de novo motif finding use a naïve enumeration algorithm, which is far from efficient. In this paper, we propose an efficient algorithm to extract maximal degenerate consensus motifs from a set of sequences based on a compact suffix tree. Our algorithm achieved a time complexity about n times lower than that of a naïve enumeration, where n is the average length of source sequences. To demonstrate the efficiency and effectiveness of our proposed algorithm, we applied it to finding transcription factor binding sites. It is validated on a popular benchmark proposed by Tompa. The executable files of our algorithm can be accessed through http://hpc.cs.tsinghua.edu.cn/bioinfo.

Keywords Motif discovery · Degenerate motif · Suffix tree

3.1 Introduction

De novo motif discovery with the aim of biomolecule interaction prediction and sequence annotation is an important process in the postgenome era. In this process, three simple motif models, i.e., *mismatched consensus* (consensus sequence allowing some mismatches), *degenerate consensus* (consensus sequence with degenerate symbols), and *position weight matrix* (PWM) [1], are widely used. Accordingly, the mainstream approaches [2 11] for motif finding can be classified into consensus-based approaches [2 6] which are based on word counting and profile-based approaches [7 11] which are based on multiple sequences alignment.

H. Jiang (✉)
Department of Computer Science and Technology, Institute of High Performance Computing, Tsinghua University, Beijing 100084, China
e mail: hongshan.jiang@gmail.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 3,

Among these simple models, the degenerate motif combines the merits of the other two. In an improved benchmark [12], degenerate consensus shows a similar performance as PWM does, both of which overwhelm the mismatched consensus.Apparently, the advantage of degenerate consensus over the mismatched consensus lies in its ability to model the widely existing position specific variability [1].

Recently, various effective studies [3, 10, 11] employ degenerate consensus motifs (or *degenerate motifs* in brief) as motif representation for de novo motif finding. However, to the best of our knowledge, all existing algorithms use a naïve enumeration method to extract degenerate motifs, which is far from efficient.

In this paper, we propose a novel consensus-based approach to efficiently extract degenerate motifs and apply it to the task of finding transcription factor binding sites (TFBSs). This work is different from previous related works [4 6], all of which focus on extracting mismatched consensus motifs. Our contributions are: (1) The problem of finding common maximal degenerate motifs was formalized; (2) An algorithm was depicted succinctly to solve the problem based on a compact suffix tree [13] instead of a suffix trie; (3) Properties of suffix trees were used to reduce redundant operations in extracting motifs with unknown lengths.

The rest of this paper is organized as follows. In Sect. 3.2, for the *common motifs* problem, an algorithm based on a compact suffix tree is presented, accompanied with computational complexity analysis and an extension to deal with unknown length motifs. In Sect. 3.3, experiment results on real biological data [14] that validate the efficiency of the approach. The last section concludes this paper.

3.2 Methods

3.2.1 Common Motifs Problem

Terminologies. Given an alphabet Σ, the set of its extended symbols, each denoting a set of symbols in Σ, is noted as Σ_E. For clarity, the set of extended symbols each denotes a single symbol in Σ is noted as Σ'. A *degenerate motif* (or *pattern* in brief) m is a sequence of extended symbols, i.e., $m \in {\Sigma_E}^+$. The *ambiguous degree* of a pattern m, noted as $G(m)$, is the number of its containing degenerate symbols each denoting multiple symbols in Σ. A *word* w is a sequence which only contains exact symbols, i.e., $w \in {\Sigma'}^+$. The set of *occurrences* of a word w in sequence S is the set of suffixes in S each has w as a prefix. The set of *instances* of a pattern m in sequence S, noted as $\mathrm{Inst}(m, S)$, is defined as the words in m which has at least one occurrence in S. The set of *occurrences* of a pattern m in sequence S, noted as $\mathrm{Occ}(m, S)$, is defined as the occurrences of all words of m in S. The set of *concrete patterns* of a pattern m, noted as $\mathrm{Con}(m)$, is defined as the set of patterns satisfying: (1) each pattern has the same length as m; (2) every symbol of the pattern denotes a subset of the corresponding symbol of m; (3) at least one of the symbols is a true subset. Given a set of sequences S, the patterns which cannot be more concrete without

losing any occurrences in S are called *breadth-maximal patterns* (or *maximal patterns* in brief) w.r.t. S, noted as $B_{\max}(S)$. More formally, $B_{\max}(S) = \{m | m \in {\Sigma_E}^+, \mathrm{Occ}(m', S) \subsetneq \mathrm{Occ}(m, S), \forall m' \in \mathrm{Con}(m)\}$.

Common motifs problem. Given a set of sequences $S = \{S_i | S_i \in \Sigma^+, i = 1, 2, \ldots, N\}$, a vector of frequency thresholds $(p_1, p_2, \ldots, p_N)$, minimum quorum q, length k, and maximum ambiguous degree g, search for all degenerate motifs m, such that $m \in B_{\max}(S), |m| = k, |G(m)| \leq g, \sum_{i=1}^{N} (|\mathrm{Occ}(m, S_i)| \geq p_i) \geq q$.

Note that the reason why we do not consider the length-maximal patterns is that patterns in real biosequences are usually in a narrow length range which can be implied from prior knowledge.

3.2.2 Searching Degenerate Motifs Guided by a Suffix Tree

Suffix trees and preprocessing. A suffix tree is a kind of tree data structure that stores all suffixes of a sequence [13]. In practical applications, all paths in a suffix tree are compressed to achieve linear space complexity. A more general form of a suffix tree, called general suffix tree (GST), stores all suffixes of multiple sequences with the sequence IDs tagged. For the *common motifs* problem, a preprocessing phase constructs the GST of the source sequences using Ukkonen's algorithm [15]. Similar to that of the preprocessing introduced in [4], the time and space complexity of the preprocessing here are both $O(nN^2)$, where n is the average length of the source sequences, and N is the number of sequences.

Extracting maximal degenerate motifs. Given a GST T, the process to extract all required maximal degenerate motifs can be done by traversing T to guide the enumeration of patterns on a virtual pattern trie.[1] The compressed information in GST helps pruning the enumeration. Before evolving the algorithm, we must ensure the bread-optimal property of the found motifs, which is guaranteed by Lemma 3.1.

Lemma 3.1. *A degenerate motif $m \in {\Sigma_E}^+$ is a breadth-maximal motif w.r.t. a set of sequences S, i.e., $m \in B_{\max}(S)$, if and only if for each ambiguous position j of m, the union of the jth symbols of $\mathrm{Inst}(m, S)$ equals the jth symbol of m, i.e., $\forall j \in G(m), \bigcup_{w \in \mathrm{Inst}(m,S)} w_j = m_j$, $\therefore$ where w_j and m_j are the jth symbol of corresponding word or pattern.*

Proof. Necessity: $\because \forall w \in \mathrm{Inst}(m, S)$ and j, $w_j \subseteq m_j \therefore \forall j \in G(m), \bigcup_{w \in Inst(m,S)} w_j \subseteq m_j$. If $m \in B_{\max}(S)$, then $\forall j \in G(m), \bigcup_{w \in \mathrm{Inst}(m,S)} w_j = m_j$. Otherwise, if $\exists i \in G(w)$, such that $\bigcup_{w \in \mathrm{Inst}(m,S)} w_i \subsetneq m_i$, then $\exists m' \in Con(m)$ and $\mathrm{Occ}(m', S) = \mathrm{Occ}(m, S)$ (here $m'_i = \bigcup_{w \in \mathrm{Inst}(m,S)} w_i, m'_j = m_j, \forall j \neq i$), which contradicts with the definition of $B_{\max}(S)$.

[1] A tree where each edge is labeled by a single symbol.

Sufficiency: If $\forall j \in G(m)$, $\bigcup_{w \in \mathrm{Inst}(m,S)} w_j = m_j$, then $m \in B_{\max}(S)$. The reason is that $\forall m' \in \mathrm{Con}(m)$, $\exists i$ such that $m'_i \subsetneq m_i$, $\therefore m'_i \subsetneq \bigcup_{w \in \mathrm{Inst}(m,S)} w_i$, which means $\exists w' \in \mathrm{Inst}(m,S)$, such that $w'_i \not\subseteq m'_i$, hence $\mathrm{Occ}(m',S) \subsetneq \mathrm{Occ}(m,S)$.

Different from those approaches that search for mismatched consensus motifs based on a suffix tree, our approach checks ambiguous degree of a degenerate motif. In addition, when prolonging a pattern, we use alphabet Σ_{E} instead of Σ. Lemma 3.2 gives the main recurrence of the algorithm, where pair (v,p) denotes the path leading from root and ending at a position p-symbols far before reaching v; $\mathrm{parent}(v)$ denotes the parent node of v on the (compact) suffix tree.

Lemma 3.2. *A pair (v,p) is the path corresponding to an instance of pattern $m\alpha$ with an ambiguous degree of g', if and only if the ambiguous degree of m is $g'-1$ and $\alpha \in \Sigma_{\mathrm{E}} - \Sigma'$, or the ambiguous degree of m is g' and $\alpha \in \Sigma'$, in addition, one of the following two conditions is verified:*

(On the node) A pair $(\mathrm{parent}(v), 0)$ is a path corresponding to an instance of m and the first symbol of the label on the edge between $\mathrm{parent}(v)$ and v belongs to the subset of Σ denoted by α, furthermore, the length of the edge label is $p+1$.

(On the edge) A pair $(v, p+1)$ is a path corresponding to an instance of pattern m and the reciprocal pth symbol of the label on the edge to node v belongs to the subset of Σ denoted by α.

The pseudo code of the algorithm is listed in Fig. 3.1. Here, m is the pattern being checked; Inst_m is its instances; Occ_m is its occurrences; "len" is the length of m; g_m is the degenerate degree of m. Note that we do not use a set of "extended symbols" as Sagot did [4]. Though seems trickier, the effect of Sagot's optimization actually depends on the data.

3.2.3 Computational Complexity

With respect to the computational complexity of the extraction phase, we have the following two lemmas, whose proofs are omitted due to length limitation.

Lemma 3.3. *Using the algorithm presented in Fig. 3.1 to solve the common motifs problem requires $O(nN(N+ck)A(g,k))$ time, where n is the average length of the source sequences, N is the sequence number, k is the specified motif length, c is a constant factor, $A(g,k) = \sum_{i=0}^{g} [\binom{k}{i} (2^{|\Sigma|} - 1)^i]$ is the possible maximum number of degenerate patterns with maximum ambiguous degree of g which verify valid patterns of a single path of length k in the suffix tree.*

Compared to that of a naïve enumeration, i.e., $O(n^2N(N-k+1)A(g,k))$[2] [11], it is approximately n times lower, where n is the average length of source sequences.

[2]The original statement is: $O(n^2N(N \quad k+1)\binom{k}{g})$, since they use a specific Σ_E.

```
1  bool IsMaximal(m, Inst_m, len)
2    for i=1,len // len - length of m
3      if(m_i ∈ Σ_E - Σ'){ // ambiguous site
4        set = ∅;
5        for each w ∈ Inst_m
6          set = set ∪ w_i; // ith symbol of w
7        if(set ≠ m_i) return false;
8      }
9    return true;
```

```
10 void ExtractMotif(m, Inst_m, len, g_m)
11 // initial: m=ε, Inst_m={ε}, len=0, g_m=0
12   if(len ≥ k){
13     KeepModel(m); return;
14   }
15   for each symbol α in Σ_E do
16     if( g_m ≥ g && α ∈ Σ_E − Σ') continue;
17     Inst_mα = ∅;
18     for each path P in Inst_m do{
19       if(P ends on an internal node v){
20         for each child u of v
21           x = u.label[0];
22           if(x ⊆ α) Inst_mα = Inst_mα ∪ Px;
23       }else if(P ends on the edge reaching v){
24         x = v.label[idx];// P ends at idx
25         if(x ⊆ α) Inst_mα = Inst_mα ∪ Px;
26       }
27     }
28     if(Inst_mα = ∅) continue;
29     if(IsMaximal(mα, Inst_mα, len+1)){
30       count = 0;
31       for i=1,N do { // calculate quorum
32         bits(i) = Logical_OR(|Occ_mα(i)| > p_i);
33         count = count + bits(i);
34       }
35       if(count ≥ q)
36         ExtractMotif(mα, Inst_mα, len+1, g+δ(α ∈ Σ_E − Σ'));
37     }
38     release(Inst_mα);
```

Fig. 3.1 Sketch of degenerate motifs extraction procedure

Lemma 3.4. *Using the algorithm presented in Fig. 3.1 to solve the common motifs problem requires $O(nN(N+ck))$ space, where n is the average length of the source sequences, N is the number of sequences, k is the specified motif length, c is a constant factor.*

3.2.4 Extension to Extract Motifs with Unknown Lengths

Utilizing the properties of a suffix tree, we can modify the algorithm to find the motifs in a length range with negligible additional computational cost. More specifically, two parameters, namely, $k_{\min}$ and g_{ratio} are used to specify the

```
1   if(len ≥ k_min) {
2     if(g_m ≤ len · g_ratio) KeepModel(m);
3     if(len >= k) return;
4   }
```

Fig. 3.2 Extension to extract motifs with unknown lengths

Table 3.1 Running time of normal version and extended version

Run time (s)	Fly (8)	Human (26)	Mouse (12)	Yeast (10)
Normal	326.84	744.66	110.93	58.71
Extended	118.88	206.36	28.45	11.48

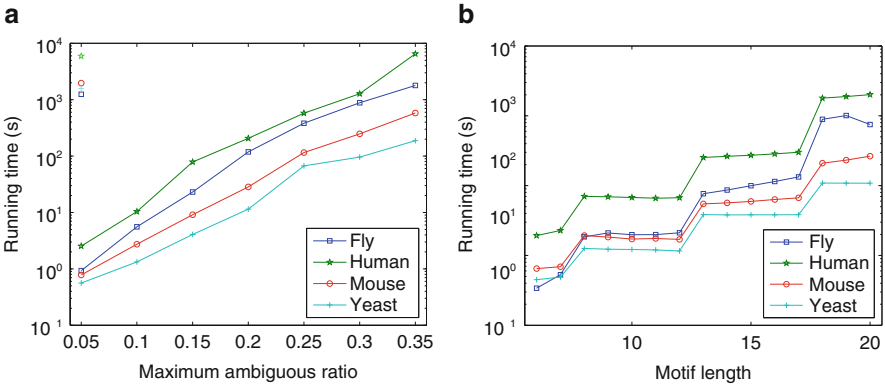


Fig. 3.3 Running time scaling over maximum ambiguous ratios and motif lengths.
Note: In (**a**), the length range is set to [6, 20]; the dots on the *upper left* area correspond to the running time of conducting naïve enumeration with the maximum ambiguous ratio set to 0.05. In (**b**), the maximum ambiguous degree increases as the motif length increases, given the fixed ratio of 0.2

minimum length allowed and the maximum ambiguous ratio allowed, respectively. In addition, the code on lines 12 14 of Fig. 3.1 should be replaced with the codes in Fig. 3.2.

3.3 Experimental Results

We implemented the algorithm as a console application using Visual C++ 2005. The program inputs a set of related sequences and outputs the top candidate patterns after extraction and postprocessing. We conducted experiments on the data sets from Tompa's benchmark [14]. The test platform is a laptop with a Core Duo U2500 (1.2 GHz) CPU and 2G main memory. The parameters are set according

to the common experiences in this field. In this process, the Log-Likelihood-Ratio postprocessing strategy is integrated.

To evaluate the merit of using the extension in Fig. 3.2, we compared the running time of using this extension with the normal version,[3] as summarized in Table 3.1. For each species, we conduct experiments on all its datasets and summarize the running time. The numbers of datasets are marked in the parenthesis after the species names.

To validate the main contribution of our work, two sets of experiments were conducted to test the running time scaling over maximum ambiguous ratios and motif lengths respectively, as illustrated by Fig. 3.3.

3.4 Conclusion

In this paper, we proposed a novel algorithm based on a suffix tree to solve the common-motifs problem w.r.t. degenerate consensus representation. Given a set of sequences, our algorithm can efficiently find all maximal degenerate motifs of unknown lengths in a large length range. Integrated with appropriate postprocessing strategies, our algorithm can be used for effectively finding real biological motifs. In the future, we plan to extend this algorithm to cope with more complicated situations, such as motifs with flexible gaps and combinatory regulatory modules.

References

1. Stormo, G.D.: 'DNA binding sites: representation and discovery', Bioinformatics, 2000, 16, (1), pp. 16 23
2. Bussemaker, H.J., Li, H., and Siggia, E.D.: 'From the cover: building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis', Proc Natl Acad Sci USA, 2000, 97, (18), pp. 10096 10100
3. Sinha, S., and Tompa, M.: 'Discovery of novel transcription factor binding sites by statistical overrepresentation', Nucleic Acids Res, 2002, 30, (24), pp. 5549 5560
4. Sagot, M F.: 'Spelling approximate repeated or common motifs using a suffix tree'. In 'Proceedings of the 1998 3rd Latin American Symposium, Apr 20 24 1998' (1998), p. 374
5. Marsan, L., and Sagot, M F.: 'Extracting structured motifs using a suffix tree algorithms and application to promoter consensus identification'. In 'Extracting structured motifs using a suffix tree algorithms and application to promoter consensus identification' (ACM, 2000), pp. 210 219
6. Pavesi, G., Mauri, G., and Pesole, G.: 'An algorithm for finding signals of unknown length in DNA sequences', Bioinformatics, 2001, 17, (suppl 1), pp. S207 S214
7. Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., and Wootton, J.C.: 'Detecting subtle sequence signals. A Gibbs sampling strategy for multiple alignment', Science, 1993, 262, (5131), p. 208
8. Bailey, T.L., and Elkan, C.: 'Fitting a mixture model by expectation maximization to discover motifs in biopolymers', Proc Int Conf Intell Syst Mol Biol, 1994, 2, pp. 28 36

[3]Where multiple runs were conducted, each with a fixed length within the length range.

9. Hertz, G., and Stormo, G.: ‘Identifying DNA and protein patterns with statistically significant alignments of multiple sequences’, Bioinformatics, 1999, 15, (7), pp. 563 577
10. Vishnevsky, O.V., and Kolchanov, N.A.: ‘ARGO: a web system for the detection of degenerate motifs and large scale recognition of eukaryotic promoters’, Nucleic Acids Res, 2005, 33, (suppl 2), pp. W417 W422
11. Peng, C.H., Hsu, J.T., Chung, Y.S., Lin, Y.J., Chow, W.Y., Hsu, D.F., and Tang, C.Y.: ‘Identification of degenerate motifs using position restricted selection and hybrid ranking combination’, Nucleic Acids Res, 2006, 34, (22), pp. 6379 6391
12. Sandve, G.K., Abul, O., Walseng, V., et al.: ‘Improved benchmarks for computational motif discovery,’ BMC Bioinformatics, 8, p. 193, 2007
13. Weiner, P.: ‘Linear pattern matching algorithms’. In ‘Proceedings of the 14th Annual Sym posium on Switching and Automata Theory (swat 1973), Volume 001973’, pp. 1 11
14. Tompa, M., Li, N., Bailey, T.L., et al.: ‘Assessing computational tools for the discovery of transcription factor binding sites’, Nat Biotechnol, 2005, 23, (1), pp. 137 144
15. Ukkonen, E.: ‘On line construction of suffix trees’, Algorithmica (New York), 1995, 14, (3), p. 249

Chapter 4
Exploring Motif Composition of Eukaryotic Promoter Regions

Nikola Stojanovic and Abanish Singh

Abstract Gene expression in eukaryotic organisms is regulated by complex interactions of enzymes, some of which directly bind to DNA. Although this binding is mostly nonspecific, one can still characterize and then search for motifs appearing with significant frequency and consistency, which are generally presumed to be target sites for transcription factors. We here report an algorithm for the identification of such motifs and subsequent genome-wide scans for their conglomerations at loci other than these provided as input sequences, for which we usually select promoter regions of coexpressed genes. The initial testing of the software implementing this method has been performed on human *Hox* gene clusters.

Keywords DNA motifs · Gene regulation · Promoters · Gene expression · Coregulation

4.1 Motivation

Eukaryotic gene transcription is regulated by the networks of protein DNA and protein protein interactions, directing chromatin remodeling, histone modifications, formation of initiation complexes, and RNA polymerase elongation. The prevailing opinion, corroborated by some studies but being increasingly disputed, is that most of these interactions take place within a few hundred bases upstream of the transcription start positions. However, even as sites important for the regulation

N. Stojanovic (✉)
Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA
e mail: nick@cse.uta.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 4,

of genes have been found at distal loci, around alternative first exons, in introns and sequences located downstream of the transcription start sites or even genes themselves, core promoter regions are still considered the most important for the gene expression. It is also generally accepted that the binding of transcriptional enzymes is in large part directed by target motifs in DNA sequence. However, while there are proteins which bind specifically to exact sequences of bases, most transcription factors are rather promiscuous in their choice of a preferred binding motif. These variations may affect the strength of binding, but they also suggest that sequence letters generally constitute only a part of the regulatory signal.

Focusing on the motif composition only, the search for transcription factor binding sites has become one of the most popular subfields of bioinformatics, and many algorithms have been developed over about two decades of intensive research. The early approaches generally suffered from large false positive ratios, so more recent methods have concentrated on the incorporation of additional information to the raw sequence data, even as they have so far neglected epigenetic effects. They often relied on phylogenetic conservation, or on the search for clusters whose elements matched experimentally confirmed consensus motifs retrieved from databases such as TRANSFAC [9] or Jaspar [2]. The latter methods exploited the fact that proteins involved in the initiation of transcription rarely, if ever, act in isolation.

With the advances in microarray technology, large sets of putatively coexpressed genes became available, stimulating the development of methods to detect conserved motifs in their upstream sequences. It is intuitive that if a group of genes is coordinately regulated, it should be controlled by similar sets of transcription factors. From the hypothesis that protein binding is largely directed by DNA sequence motifs, it follows that same motifs should be present in regulatory sequences of coexpressed genes, moreover as a cluster or clusters. This has led to the exploitation of motif overrepresentation in related target sequences, and in particular, these upstream of genes whose expression has been observed to be correlated, in searching for statistically significant conglomerations.

The efforts to characterize noncoding functional DNA have intensified along with the recent sequencing of a large number of eukaryotic genomes. Whereas there has been some success in the computational recognition of binding sites, the existing methods are not satisfactory. Apart from the inherent lack of precision due to the usage of sequence information alone, many problems stem from the fact that genomes of higher eukaryotes are not a random assembly of letters. Protein-binding targets in DNA are short, and even when assuming a pure Poisson model the count of short repeats expected by chance can be large in any meaningful sequence segment, dictating an excessive redundancy necessary to guarantee significance. Furthermore, our earlier study [7] has shown that the number of repeated motifs in repeat-masked random intergenic DNA was far greater than expected, most likely due to similar evolutionary origins of many genomic fragments. In consequence, any search for overrepresented sequences is bound to return many results, and, depending on what is being searched for, most would likely be false positives. Taking into account the nonspecific protein binding and often

noncontiguous arrangement of the bases important for the interaction, many true signals would remain undetected, resulting in relatively low sensitivity in addition to low specificity. However, methods are constantly being improved, with new generations of bioinformatics tools expected to rely on a significant volume of information about the physical and chemical aspects of the binding process. The efforts in that direction have already started [3, 4].

We have previously developed software [6] to rapidly identify significantly repeated short (5 25 bases) motifs in groups of DNA sequences, presumably promoter regions of coexpressed genes. Our core method strives to detect all significant elements, and then filter the results according to the number of regions sharing the motif, score, composition (simple sequence, tandem repeat structure or perfectly conserved bases at loci suggesting a helix-loop-helix, zinc fingers or leucine zipper structure, for instance), database hits, clustering patterns, or positional conservation. The motifs we discover are approximate, since the transcription factor binding is generally determined by a very small number of bases (sometimes as small as three) which need not even be adjacent. In such cases, popular methods like PWM matching or alignment analysis may not be appropriate, the former suffering from the restriction that motifs need to have been previously characterized, and the latter requiring their positional conservation. In contrast, our approach can identify previously unknown motifs, and it permits their rearrangements within the regulatory regions.

We have extended the finding of the modules with the genome-wide discovery of the layouts similar to these we have identified in the original collections of promoters of putatively coexpressed or otherwise related genes. Although the biological relevance of the discovered sets still needs to be established, as they need not necessarily be binding sites for transcriptional proteins even when their frequency and conglomeration implies *some kind* of a function; our runs on the human genome have indeed found many other genes whose promoter regions exhibited substantial similarities with the original sets of promoters. These genes may be coregulated with the original set, or be otherwise related in some aspects of their expression.

4.2 Methods

Our core tool, originally reported in [6], receives a set of sequences in which significant motifs should be located, and outputs a comprehensive list which is nevertheless filtered by starting criteria: minimal motif length, minimal number of input sequences in which the motif should appear, likelihood of its occurrence by chance, and the permitted degeneracy. We start by identifying all exactly repeated strings of length two or more in the input sequences, both strands, using a suffix tree [8] data structure. After the original list has been built, we identify the repeats which co-occur at least twice, at a constant distance, and name them potential mates. This identification can be efficiently done using indexing of the seed element

positions. We use all sites of co-occurrence to build the putative consensus of a variable motif, noting the conservation of each individual position. We determine whether the percentage of conserved characters exceeds a given threshold (95% in our tests), and then apply a heuristic measure of consensus *stability*, the fraction of degenerate bases, and statistical measure of its *viability*, indicating whether the motif with so many substitutions can still be considered unlikely by chance. If a consensus has been successfully formed, we try to find additional locations in the sequence where this motif could have occurred with a limited number of substitutions, and then eliminate the constitutive occurrences of the seeds and other elements which have been factored in. We combine the consensus motifs with the remaining repeats, reindex the list, and attempt to collapse substantially overlapping elements. This is necessary because components of the motifs may have separate occurrences elsewhere and thus could have been identified as separate, distinct repeats. If their separate occurrences permit the inclusion in the broader consensus, under the same conditions of stability and viability, that is done now. We iteratively repeat the process until no further extensions are possible, after which we report all significant elements.

We then use the assembled list of the shared motifs to explore whether there are any additional loci in other segments, representing upstream sequences of a catalog of genes, entire chromosomes or genomes, which exhibit motif groupings similar to these identified in the promoter sequences of the initially considered genes. These may be just sites sharing the same evolutionary origins with the input sequences; however, their conservation would imply a function, whether protein binding or not. In cases when they would indeed present a regulatory module, which is likely when they occur within promoters, that would imply coregulation with the initial set. This may prove useful if these additional genes have not been previously characterized.

Looking at all motifs from the initially generated list may or may not be productive, depending on the list size (which in our experiments varied widely), but certain types of consensuses were generally of lesser interest. First, they would be those whose p-values, reflecting the probability that their occurrence might be due to chance, is relatively high. Our default settings require the motif p-value to be less than 0.01 to make it reportable; however, it is a user-definable parameter, and we do not recommend genome-wide searches for motifs with p-values larger than 10^{-5}. Second, in our experiments, we have filtered out motifs which could be considered to represent tandem repeats or simple sequence. One would generally not expect to see many Poly-A tail remnants in promoters, but most lists we have looked at featured elements, some scoring very high, with suspiciously high concentration of A's or T's. We have thus used the Shannon's entropy of the motifs, taking into account only non-N characters, and assigning the probability to any of the four letters according to the relative contribution of that letter to the motif composition (so, for instance, motif AAtCgNNCCT would have $p(\mathrm{A}) = 0.25$, $p(\mathrm{C}) = 0.375$, $p(\mathrm{G}) = 0.125$, and $p(\mathrm{T}) = 0.25$ and thus $H(\mathrm{AAtCgNNCCT}) = 1.91$). The decision about the acceptable motif entropy is

left to the user and the nature of the application, and in most of our genome-wide search tests, we have discarded motifs featuring less than 1.0 bit of entropy (out of possible 2.0 bits, reflecting four equal contributors).

As our motifs are approximate, and in order to enable fast genome-wide search, we expand them to a set of exact variants. We merge all expansions into a comprehensive set, and maintain an indexing scheme so that for each variant we can easily reconstruct the original consensus it instantiates. We then construct a finite state automaton [1] to simultaneously match all variants of all motifs, in time linear with the size of the examined sequence(s) and the combined lengths of all variants.

In order to locate the conglomerations of motifs similar to these in our set, we look at windows of size W equal to the size of the promoter regions we have initially considered. We report the window if the number of motifs which have matched within it exceeds a threshold value T, calculated as follows. Let m be the total number of motifs M_i, $i = 1, m$ selected for the genome-wide search, and let n_i be the number of exact instances of the approximate motif M_i. Let l_i be the length of motif M_i. We then calculate the probability p_i of motif M_i matching in an instance of a window of size W as $p_i = n_i \frac{W}{4^{l_i}}$. Since we need a unified probability measure for all motifs in our set, and the motifs were not that much different from one another, we derive it as the average of all probabilities p_i, i.e., $p = \sum_{i=1}^{m} \frac{p_i}{m}$. The expected number of motifs matched in any window is then $\lambda = mp = m \sum_{i=1}^{m} \frac{p_i}{m} = \sum_{i=1}^{m} n_i \frac{W}{4^{l_i}} = W \sum_{i=1}^{m} \frac{n_i}{4^{l_i}}$. We use this λ to calculate the probability of k or more motifs matching within the window. If random variable X models the number of matches, using the Poisson distribution, we get $P\{X \geq k\} = 1 - \sum_{i=0}^{k} \frac{\lambda^k e^{\lambda}}{k!}$. If the total number of windows we consider is N, the expected number of these in which k or more separate motif instances that would match would be $E(k) = N \times P\{X \geq k\}$. This corresponds to the classical measure of *e*-value, or the *expect*-value. In our experiments, we have set the threshold T as the value of k for which $\mathrm{E}(k)<1$, although the user of our software is free to select any other T, depending on the needs of his/her particular application.

These methods have been implemented as a part of our Web server located at http://bioinformatics.uta.edu/toolkit/motifs. However, the Web interface provides only a part of the functionality of the full FTP distribution, at http://bioinformatics.uta.edu/toolkit/download.

4.3 Performance

The running time of our detection of the shared motifs in the input sequences has been analyzed in [6], and this process can be done efficiently. Once the list of motifs has been formed, their expansion into exact variants may take time exponential in their size. However, the consensuses are generally short (usually less than 15 characters) and do not feature many variable positions (since every such position

hurts the motif score and *p*-value), so a highly variable item is not likely to be reported. Consequently, the expansion of the motifs and their incorporation into a finite state automaton generally takes just a second or two. The subsequent genome-wide search is performed in linear time, and even when this analysis scans millions of bases, the result is returned in time suitable for online work. Although the space requirements of our approach are not modest, they do not exceed the memory necessary to store the sequences themselves and all candidate motifs and their original positions, plus the exact variants of the motifs which need to be matched. We have also validated the core functionality of our software and the calculations of thresholds, *p*-values and *e*-values on simulated random sequences, with and without planted motifs (this was necessary because of the compositional bias of eukaryotic DNA and its incomplete characterization at the present state of genomics). We have also used several well-studied coregulated gene sets for the verification on real data.

For a preliminary biological validation of the genome-wide search, we have used the Build 36.1 finished human genome assembly (hg18), 1,000 bases upstream of annotated transcription starts for RefSeq genes with annotated 5′ UTRs, containing 5′ sequences of 23,570 genes (our program can also scan complete chromosomal sequences; however, in that case, the interpretation of the results would be much more complex). We have further concentrated on four *Hox* gene clusters. *Hox* genes code for transcription factors which regulate the formation of the anterior posterior axis of an animal during early embryonic development, acting on a large number of downstream genes. Since this axis is common throughout the evolution, *Hox* clusters are well conserved, often over regions much longer than expected under a simple model of coding sequence and transcriptional regulation.

Whereas *Hox* genes are not straightforwardly coregulated, some are controlled by similar transcriptional machinery, and we have postulated that this should be particularly the case for these with redundant function in different clusters. Studies have found that *cis*-acting elements were more likely to be found in the close proximity of the anterior genes, with posterior ones being regulated in increasingly complex and spatially distant ways [5], therefore, we have paid special attention to the anterior genes conserved over all four clusters. We have also tried to find modules which may be shared among most *Hox* genes in a single cluster, and check the results of the genome-wide scan for the presence of *Hox* genes from other clusters.

Looking at paralogs conserved over all four clusters, there were only three such groups, corresponding to *Hox13*, *Hox9*, and *Hox4*. Although we have found genes which were sharing significant groups of motifs with upstream sequences of *Hox13* and *Hox9*, none of these genes were from *Hox*, which was anticipated from the findings in [5]. However, the situation was different at the anterior end. The search for motifs (minimal length 7, minimal significance score 0.99, minimal *p*-value 0.01, and minimal motif complexity 1.0 bits) in 1,000 bases upstream of the transcription start site of all four *Hox4* genes (*HoxA4*, *HoxB4*, *HoxC4*, and *HoxD4*) has returned 32 significant shared motifs. Genome-wide search for conglomerations of these motifs in upstream sequences of other genes performed with maximal *e*-value of 0.5 (requiring a match of at least 20 out of 32 motifs) has

returned only five hits: four original *Hox4* genes and *HoxC6*. Actually, *HoxC6* would have been returned even with a far stricter *e*-value, since it shared 29 motifs with all *Hox4* genes, practically eliminating any possibility of a random match.

We next considered the motifs shared among upstream sequences of all *Hox* genes in one cluster. Whereas the results have not been conclusive, related genes from other clusters were indeed present among our results more than warranted by chance, occasionally but consistently.

4.4 Discussion

Although there is much work yet to be done on the verification of the biological significance of our findings, including negative controls, we are enthusiastic about the promise of our approach. Our software was finding motifs and further matching them genome-wide efficiently, and it has shown good sensitivity in cases when it could have been measured. Depending on the number of sequences among which the motifs needed to be shared, the number of initially identified elements tended to be higher than one would expect; however, this is a common problem shared by all motif finders, and stems from the structure of genomes, or at least eukaryotic genomes. In another study [7], we have described and quantified the remarkable microrepetitive structure of the human genome, and shown that it is almost impossible to reliably identify functional motifs solely based on the occurrence counts, even after all known repeats (such as satellites and transposon copies) have been masked and thus removed from consideration. This also applies to the genome-wide matches of the identified motifs, as the number of hits was consistently far larger than the targeted *e*-values. These were not random hits, at least not random by the simple-minded equal letter probability model, but it is difficult to believe that they all represented conglomerations of functional elements, and binding signals for transcriptional complexes in particular. It is more likely that most matches were due to still poorly understood compositional features of immediate upstream sequences of the genes, and elucidating them will be a challenging but necessary task.

Acknowledgements This work has been supported by NIH grant 5R03LM009033 02 to NS.

References

1. A.V. Aho and M.J. Corasick. Efficient string matching: an aid to bibliographic search. *Commun. ACM*, 18:333 340, 1975.
2. J.C. Bryne, E. Valen, M.H. Tang, T. Marstrand, O. Winther, I. da Piedade, A. Krogh, B. Lenhard, and A. Sandelin. JASPAR, the open access database of transcription factor binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res*., 36:D102 D106, 2008.
3. Y. Cai, J. He, X. Li, L. Lu, X. Yang, K. Feng, W. Lu, and X. Kong. A novel computational approach to predict transcription factor DNA binding preference. *J. Proteome Res*., 8(2):999 1003, 2009.

4. C. Huttenhower, K.T. Mutungu, N. Indik, W. Yang, M. Schroeder, J.J. Forman, O.G. Troyanskaya, and H.A. Coller. Detailing regulatory networks through large scale data integration. *Bioinformatics*, 25(24):3267 3274, 2009.
5. J. Sharpe, S. Nonchev, A. Gould, J. Whiting, and R. Krumlauf. Selectivity, sharing and competitive interactions in the regulation of Hoxb genes. *EMBO J*., 17:1788 1798, 1998.
6. A. Singh and N. Stojanovic. An efficient algorithm for the identification of repetitive variable motifs in the regulatory sequences of co expressed genes. In *Proceedings of the 21st Inter national Symposium on Computer and Information Sciences*, volume 4263 of *LNCS*, pages 182 191. Springer Verlag, 2006.
7. A. Singh, C. Feschotte, and N. Stojanovic. A study of the repetitive structure and distribution of short motifs in human genomic sequences. *Int. J. Bioinform. Res. Appl*., 3:523 535, 2007.
8. P. Weiner. Linear pattern matching algorithms. In *Proceedings of the 14th IEEE Symposium on Switching and Automata Theory*, pages 1 11, 1973.
9. E. Wingender. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinformatics*, 9:326 332, 2008.

Chapter 5
Large-Scale Analysis of Phylogenetic Search Behavior

Hyun Jung Park, Seung-Jin Sul, and Tiffani L. Williams

Abstract Phylogenetic analysis is used in all branches of biology with applications ranging from studies on the origin of human populations to investigations of the transmission patterns of HIV. Most phylogenetic analyses rely on effective heuristics for obtaining accurate trees. However, relatively little work has been done to analyze quantitatively the behavior of phylogenetic heuristics in tree space. A better understanding of local search behavior can facilitate the design of better heuristics, which ultimately lead to more accurate depictions of the true evolutionary relationships. In this paper, we present new and novel insights into local search behavior for maximum parsimony on three biological datasets consisting of 44, 60, and 174 taxa. By analyzing all trees from search, we find that, as the search algorithm climbs the hill to local optima, the trees in the neighborhood surrounding the current solution improve as well. Furthermore, the search is quite robust to a small number of randomly selected neighbors. Thus, our work shows how to gain insights into the behavior of local search algorithm by exploring a large diverse collection of trees.

Keywords Biological datasets · Maximum parsimony · Phylogenetic analysis · Phylogenetic search · Search algorithm · Search behavior

5.1 Introduction

Phylogenetic trees represent the genealogical relationships between a group of organisms (or taxa), where leaves represent the organisms of interest and edges represent the evolutionary relationships. Inferring evolutionary trees is not a trivial task and is often reformulated as an NP-hard optimization problem. Here, trees are given a score, where trees with better scores are believed to be better

H.J. Park (✉)
Department of Computer Science, Rice University, Houston, TX, USA
e mail: hp6@cs.rice.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 5,

approximations of the truth. Given the exponential number of potential hypotheses (or trees) for a set of taxa, an exhaustive exploration of the tree space is not possible. Instead, phylogenetic inference relies on effective heuristics for obtaining good-scoring trees. However, relatively little work has been done to analyze quantitatively the *behavior* of phylogenetic heuristics in tree space. A better understanding of search behavior can drive the design of better heuristics that ultimately lead to more accurate reconstructions of phylogenetic trees.

In this paper, we exploit the data mining and information visualization opportunities that exist among the large collection of trees found by a phylogenetic heuristic. In particular, unlike traditional studies, we do not solely focus on analyzing the most parsimonious trees found during a search. By analyzing thousands of trees which have a variety of scores our study helps provide insights regarding a phylogenetic search behavior in tree space.

Our study is based on local search heuristics for maximum parsimony on three biological datasets of 44, 60, and 174 taxa. Several interesting and striking facts are uncovered from our study. Our first observation is that as a search climbs the hill to the local optima, the trees in the neighborhood surrounding the current solution improve as well. In fact, there is minimal overlap between the trees that are explored during a search. Hence, one can think of a neighborhood as a population of solutions surrounding the current solution and the population improves as the search moves forward. Similar observations are found concerning the topological distances between trees. In other words, trees that are further apart (closer together) in score are further (closer) topologically as measured by their Robinson-Foulds distance [7, 10]. Since the neighborhoods of trees are improving along with the current solution, a small number of random neighbor selections do not impact significantly the performance of a search. In fact, with this strategy, our SLS heuristic established a new best score on our 60 taxa biological dataset.

The remaining sections provide an overview of our study examining phylogenetic search behavior based on analyzing thousands of trees. For the interested reader, complete details of all experiments can be found at [6].

5.2 Simple Local Search

Our Simple Local Search (SLS) maximum parsimony heuristic operates by successively exploring the neighborhood of a current solution and moving to one of its neighbors. First, SLS creates a random sequence addition (RSA) to create the initial starting tree. Starting trees can also be based on neighbor-joining (NJ) [9] or by generating a starting tree randomly. Once we have a tree T, we improve it by rearranging its edges in a way that improves its maximum parsimony score. In SLS, we use the standard rearrangement operators: Nearest Neighbor Interchange (NNI), Subtree Pruning and Regrafting (SPR), and Tree Bisection and Reconnection (TBR).

With a mechanism for generating a neighborhood, we must decide the neighboring tree T' which should be selected. SLS uses a *first improvement algorithm* to select a neighbor. If $score(T') < score(T)$, then T' is accepted to replace the current tree, T. The search continues until there is no neighboring tree, T', with a better score than the current tree, T. If no better neighbor can be found, a local optimum has been reached and SLS terminates. SLS could also operate extremely greedily by selecting the best tree from a neighborhood. Our experiments use SLS with a first improvement strategy since it performs better than a best improvement strategy.

The search history of a phylogenetic heuristic is the set of neighbors selected along the search path to a local optimum. Given that we are interested in understanding the behavior of phylogenetic search, a phylogenetic heuristic's search history is of interest. Popular maximum parsimony heuristics such as PAUP* [11] do not provide such behavior.

Formally, the sequence of trees encountered along the search path is defined as $P = (t_1, \ldots, t_m)$. For a path P, the search examines tree t_i before tree t_j, where $0 \leq i < j \leq m$. There are m trees on the search path, where t_1 represents the initial (or starting) tree and t_m the final tree (e.g. local optimum). Consider a neighborhood relation $N_\beta(t)$ which generates the neighboring trees of t using the rearrangement scheme β, where $\beta \in \{NNI, SPR, TBR\}$. For example, $N_{TBR}(t)$ produces all of the TBR neighbors of tree t. To capture all of the β neighbors along a search path P, $N_\beta(t) = \{t' | t' \text{ is a } \beta \text{ neighboroft}\}$. $\hat{N}_\beta(P)$ is a mapping between a search path, P, and a list of neighboring tree sets, such that $N_\beta(t)$ is the i^{th} tree set in $\hat{N}_\beta(P)$ and t_i is the i^{th} tree in P.

Each run i of the heuristic results in a search path, P_i. The complete set of trees are $T = \bigcup_{i=1}^{k} \hat{N}_\beta(P_i)$, where k is the total number of runs. In our experiments, the number of runs, k, is 5. For the search path P_i of run i, we are interested in the following trees, $\phi(P_i) = (t_{0\%}, t_{20\%}, t_{40\%}, \ldots, t_{100\%})$. The ϕ operator selects the trees of interest along the search path, P_i. That is, we are interested in the initial tree ($t_{0\%}$), the tree that represents the 20% completion point of the search ($t_{20\%}$), etc. If $|(P_i)| = 100$ trees, then $t_{20\%}$ would represent the 20^{th} tree (t_{20}) of P_i. The final tree on the path represents the end of the search (100% search completion).

5.3 Experimental Results

We used the following biological datasets as input to study the behavior of our SLS heuristic: (i) 44 taxa dataset (17,028 sites) of placental mammals [4], (ii) 60 taxa dataset (2,000 sites) of ensign wasps [1], and (iii) 174 taxa dataset (1,867 sites) of insects [2]. In our experiments, both SLS established best scores of 43,085, 8698, and 7440 for Datasets #1, #2, and #3, respectively. SLS performs similarly to PAUP* [11] in terms of finding similar scoring trees, and SLS finds the best-scoring trees for Datasets #2 and #3. Our SLS algorithm is implemented in C++. We used the HashRF algorithm to compute the Robinson-Foulds (RF) topological $t \times t$

distance matrix between the collection of t trees [10]. The SLS heuristic was run five times on each of the biological datasets. All experiments were run on an Intel Pentium D platform with 3.0 GHz dual-core processors with 2 GB of total memory.

Given that TBR is the most popular neighborhood used in a phylogenetic search, we exclusively consider the *behavior* of the SLS (TBR) heuristic in the remainder of the paper. However, the trends found from analyzing SLS (TBR) apply to SLS (NNI) and SLS (SPR) as well. Figure 5.1 shows all of the MP scores in $N_{TBR}(t_{p\%})$, which is the set of trees in the TBR neighborhood of tree $t_{p\%}$. Here, $t_{p\%}$ represents the tree that represents the $p\%$ completion of the search. For example, $t_{0\%}$ is the initial tree and $t_{20\%}$ is the tree that represents the 20% completion point of the search. For Dataset #1, the average number of MP scores depicted in each boxplot is 92,281.3. There are 233,120.8 and 250,000 MP scores represented by each box plot in Datasets #2 and #3, respectively. The actual average neighbors for Dataset #3 are 3,716,369.7, but each box plot represents a sampling of 250,000 of those trees.

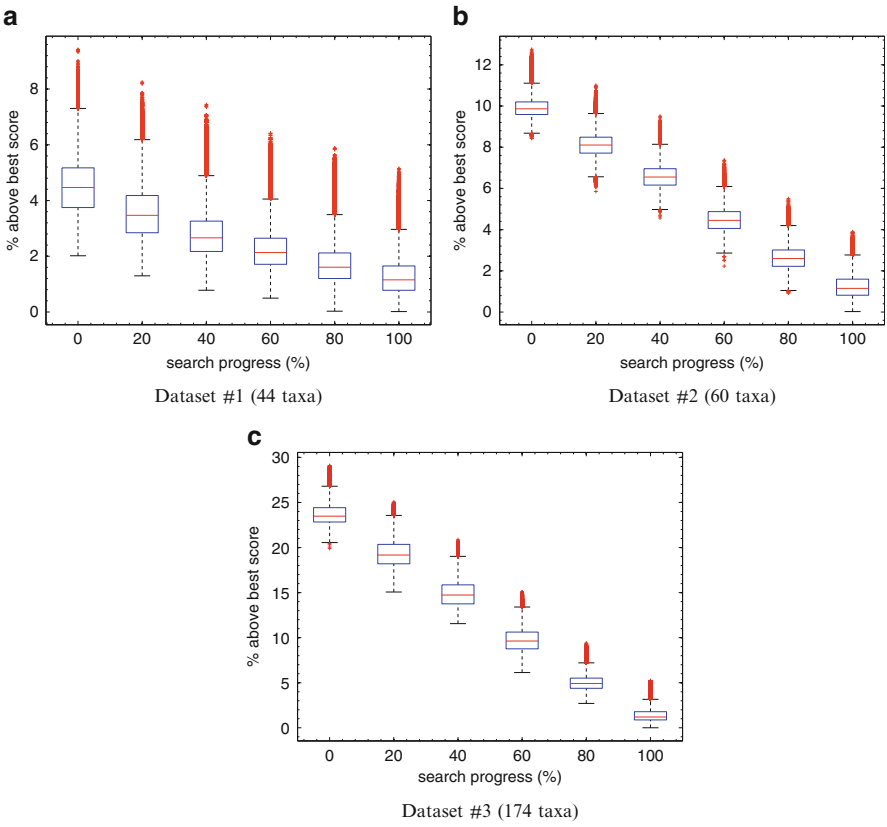


Fig. 5.1 The distribution of the MP scores in a TBR neighborhood as the search progresses toward local optima. For each interval $(0\%, 20\%, \ldots, 100\%)$, all tree scores from the neighborhood of the current tree is shown for all five runs

The plots clearly show that the entire distribution of MP scores in a TBR neighborhood improves as the search progresses toward the local optimum. Once the search reaches 100% search progress (i.e. a local optimum is reached), the plots show that the TBR neighborhood in terms of MP scores has improved dramatically over the TBR neighborhood of the starting trees. Thus, although the SLS heuristic is designed to improve each tree t_i selected on the search path P, Fig. 5.1 clearly shows that the entire neighborhood is improving as well. So, not only is the best getting better, but the worse scoring trees in a neighborhood improve as well.

Figure 5.2 provides the topological distances between the neighboring trees along the search path, P, and the best-known tree for a dataset. Here, we use a heat map representation, where each value in the two-dimensional matrix is represented as a color. Darker (lighter) colors represented smaller (higher) RF rates, which is the RF distance normalized. Given that RF is a dissimilarity measure, higher RF rates means that the trees are more dissimilar than lower RF rates. An RF rate of 0% indicates that the trees are identical. The heat map clearly demonstrates that as the search progress by improving upon the current scores found, the neighboring trees also get topologically closer to the best-known trees. Trees that are in the neighborhood of the local optima have the smallest RF distances to the best-known trees. Hence, our set of heat maps clearly shows that there is a strong correlation between MP scores and their RF distance from the best trees.

The previous figures demonstrated that as SLS improves upon the tree t_i, its neighbors improve as well. Significant time during a phylogenetic search is spent generating and scoring neighbors in order to make a good decision regarding selecting the "best" neighbor. We were curious as to how sensitive the search is regarding selecting a neighbor. If the search is not that sensitive in terms of its

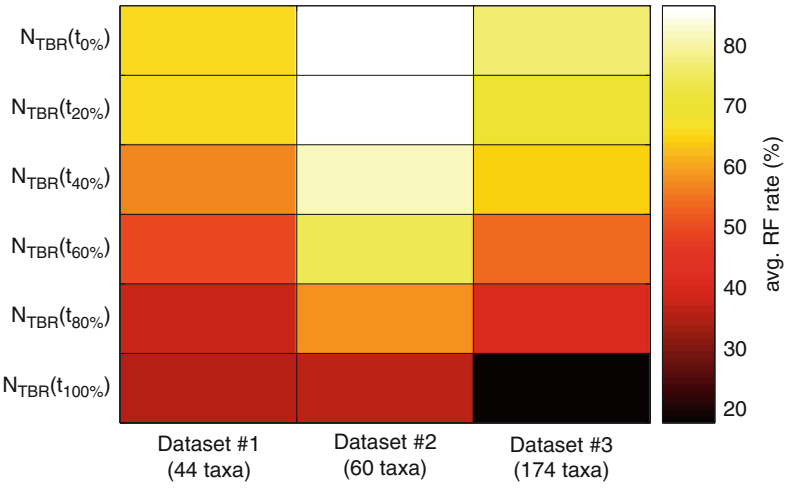


Fig. 5.2 RF distances between the neighborhood trees and the best tree found for each dataset. Each heat map entry is the average of an all to all comparison of 10,000 trees sampled from the neighborhoods of interest

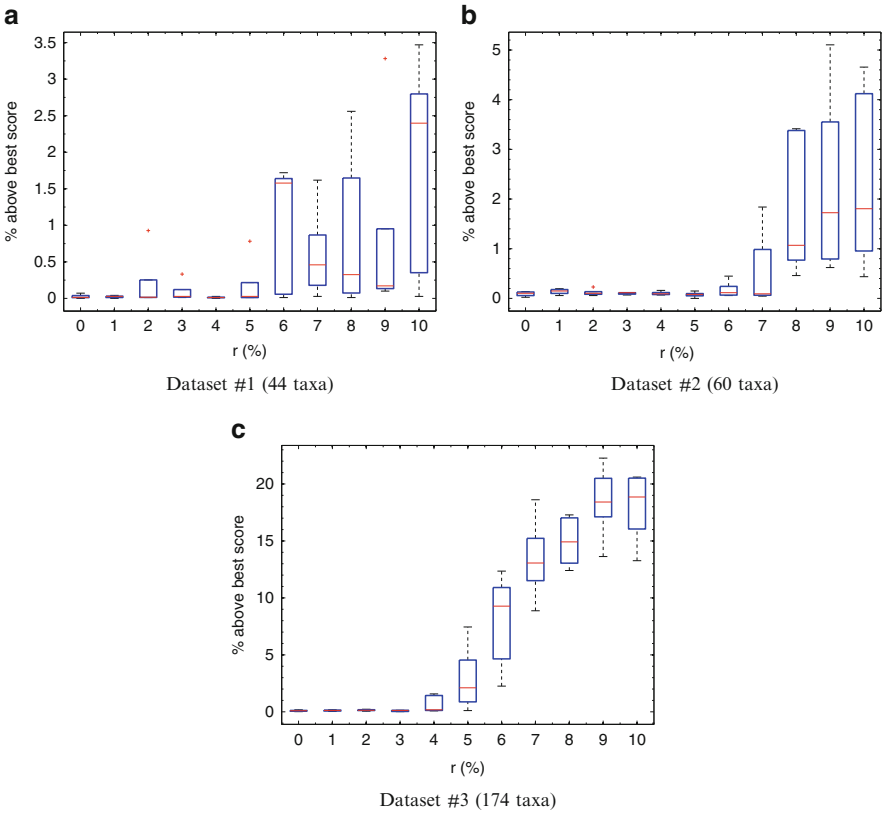


Fig. 5.3 The performance of the SLS algorithm when random neighbors are selected. Our original SLS algorithm (r 0%) always chooses the first improving neighbor for its next move on the search path. However, for $r \geqslant 1$, there is an $r\%$ chance that the next tree (t_{i+1}) on the search path is selected randomly from the TBR neighborhood of t_i (i.e. $N_\beta(t_i)$)

overall performance, then time can be saved, which translates into being able to perform larger phylogenetic analyses. Figure 5.3 shows the performance of the SLS algorithm when a random neighbor is selected $r\%$ of the time. Here, $r = 0\%$ represents our standard first improvement algorithm, each tree on the search path is based on the first neighbor that improves upon the current score. For $r \geq 1$, there is an r% chance that the next tree (t_{i+1}) on the search path is selected randomly from $N_\beta(t_i)$. (In case a local optimum is not reached, our r experiments used a search path limit of 1,000 trees so that the search would terminate. However, all of our experiments terminated on a local optimum.

In Fig. 5.3(a), the SLS runs with $1 \leq r \leq 5\%$ result in median values that are similar to SLS runs with no randomly selected neighbors ($r = 0\%$). As r approaches 10%, the search cannot recover as the scores it finds are much further away from the best score. Similar trends occur in Datasets #2 and #3.

5.4 Conclusions

To understand the behavior of local search heuristics, we implemented SLS for solving the maximum parsimony problem on moderately-sized datasets. We show that our SLS algorithm performs comparably to PAUP*. Thus, we lose minimal (if any) accuracy in using our algorithm for analyzing the behavior of local search heuristics. Furthermore, by analyzing thousands of trees with a diverse range of MP scores, our experiments with SLS show that there is a strong correlation between MP scores and topological distance. Of course, since our local search heuristic is greedy, parsimony scores improve as the search progresses toward local optima. More enlightening, however, is that neighborhood trees surrounding the current best tree improve as well. In fact, the search is quite robust to a small percentage of random neighbor selections, which provides evidence why search strategies such as parsimony ratchet [5], which takes backward moves by reweighting the characters in the dataset.

By analyzing the behavior of local searches, better phylogenetic heuristics can be designed. For example, by knowing that there are several good, but competing solutions within a neighborhood, a variety of different neighbor selection strategies (such as simulated annealing) are worthy of further investigation especially in the context of investigating their behavior based on the analysis techniques presented here. In the future, we plan to apply our approach to more powerful metaheuristics such as parsimony ratchet [5], TNT [3], and Rec-I-DCM3 [8], which will allow us to analyze much larger datasets.

Acknowledgments Funding for this project was supported by the National Science Foundation under grants DEB 0629849 and IIS 0713618. The authors wish to thank Bill Murphy and Matt Yoder for providing us with the biological datasets used in this study.

References

1. A. R. Deans, J. J. Gillespie, and M. J. Yoder (2006) An evaluation of ensign wasp classifi cation (Hymenoptera: Evanildae) based on molecular data and insights from ribosomal rna secondary structure. *Syst. Ento.*, 31:517 528.
2. J. Gillespie, C. McKenna, M. Yoder, R. Gutell, J. Johnston, J. Kathirithamby, and A. Cognato (2005) Assessing the odd secondary structural properties of nuclear small subunit ribosomal rna sequences (18s) of the twisted wing parasites (Insecta: Strepsiptera). *Insect Mol. Biol.*, 15:625 643.
3. P. A. Goloboff, J. S. Farris, and K. C. Nixon (2008) Tnt, a free program for phylogenetic analysis. *Cladistics*, 24(5):774 786.
4. W. J. Murphy, E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally, C. J. Douady, E. Teeling, O. A. Ryder, M. J. Stanhope, W. W. de Jong, and M. S. Springer (2001) Resolution of the early placental mammal radiation using bayesian phylogenetics. *Science*, 294:2348 2351.
5. K. C. Nixon (1999) The parsimony ratchet, a new method for rapid parsimony analysis. *Cladistics*, 15:407 414.
6. H. J. Park, S. J. Sul, and T. L. Williams (2009) Large scale analysis of phylogenetic search behavior. Technical Report TR CS 2009 12 1, Department of Computer Science and Engineering, Texas A& M University.

7. D. F. Robinson and L. R. Foulds (1981) Comparison of phylogenetic trees. *Math Biosci*, 53:131 147.
8. U. Roshan, B. M. E. Moret, T. L. Williams, and T. Warnow (2004) Rec I DCM3: a fast algorithmic techniques for reconstructing large phylogenetic trees. In *Proc. IEEE Computer Society Bioinformatics Conference (CSB 2004)*, pp 98 109. IEEE Press.
9. N. Saitou and M. Nei (1987) The neighbor joining method: A new method for reconstructiong phylogenetic trees. *Mol. Biol. Evol.*, 4:406 425.
10. S. J. Sul and T. L. Williams (2008) An experimental analysis of robinson foulds distance matrix algorithms. In *European Symposium of Algorithms (ESA'08)*, volume 5193 of *Lecture Notes in Computer Science*, pp 793 804. Springer, New York.
11. D. L. Swofford (2002) PAUP*: Phylogenetic analysis using parsimony (and other methods), Sinauer Associates, Underland, Massachusetts, Version 4.0.

Chapter 6
Silicosection and Elucidation of the Plant Circadian Clock Using Bayesian Classifiers and New Genemining Algorithm

Sandra Smieszek, Rainer Richter, Bartlomiej Przychodzen, and Jaroslaw Maciejewski

Abstract Datasets with a high dimensional feature space, advancing statistical methods, and computational efficiency were analyzed to uncover the rules of the circadian rhythms. The aim of the study was to uncover the identity, the dynamic behavior, and the interactions among the components of the circadian clock. Transcriptional profiling has exposed the regulon conferring benefits for circadian biology and bioinformatics. Circadian plant time course gene expression data was examined, this was the prerequisite for Naive Bayes classifiers which were trained and led to expression model with a success rate of up to 87%. The model showed new combinatorial rules, including presence of elements and their frequencies in driving particular phases. Implementation of Genemining V2.3 multipotent algorithm showed the specific combinations of elements responsible for expression patterns, highlighting the role of GATA motifs. State-of-the-art technologies allowed for a model in silico, the first such model was made using time course circadian data.

Keywords Circadian clock · Naive Bayes classifiers · Genemining V2.3 · Time course data · Supervised soft clustering · CUDA

6.1 Introduction

Circadian rhythms are ubiquitous and exhibit exquisite precision. They are characterized by being free running, entrained by Zeitgebers, under constant conditions displaying 24-h periodicity and exhibiting temperature compensation [1]. What

S. Smieszek (✉)
Department of Molecular Biology, Royal Holloway, University of London, Surrey, UK
e mail: s.smieszek@rhul.ac.uk

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 6,

ignites core promoters is of paramount importance for understanding spatiotemporal circadian expression profiles. In the era of systems biology, the impact factor of high throughput technologies in data mining and in silico approaches is indescribable, but so are the challenges. A "Good-Turing estimator formula" of this day and age is required to break the code, which is the elusive and enigmatic architecture of networks underlying temporal sensitivity of processes giving rise to circadian rhythms. Having everyone active all at once is a waste of energy, thus nature designed efficient expression providing fitness, synchrony, and constant wonder as "Flower Clock" described by Karl von Linne portrays [2]. Unambiguous identification of input signals, *cis*-regulatory elements and their corresponding transcription factors driving the expression of the circadian clock, therefore establishing the molecular architecture of the circadian clock is a great challenge, one that tormented minds for years especially since Bussemaker made the first network attempts [3]. Genomewide expression profiling allows for snapshots at a genomic level. Various algorithms, idiosyncratic software, have been created, decreasing type I and type II errors. The need for statistical assessment led to the development of various algorithms like CORRCOS, COSOPT, MC-FFTSD32, CIRCCORR, L-S, Haystack, and Feature Selection Templates, and it became a norm to create optimal algorithms for one's data [4]. Eukaryotic expression can be approached at three fundamental levels as [5] proposed: sequences of DNA where transcription factors bind *cis*-regulatory elements, level of chromatin remodeling (DNA methylation), and at the level of nuclear architecture, positions on chromosomes. Further studies add dimension to the regulatory control of gene expression, the notion that noncoding RNAs control transcription [6]. It can be stipulated that it would be interesting to decipher such a network in other eukaryotes, particularly *Arabidopsis thaliana*. The study of yeast led to success in understanding gene regulatory networks [7]. As Beer and Tavazoie described in 2004, new frameworks are required for connecting expression of genes, signaling molecules, transcription factors and DNA targets, and revealing forces driving the sequential compartmentalization of the transcriptome of *Arabidopsis thaliana*. Although it is seen across taxonomic groups, there is little conservation of the molecular components [1]. Models were proposed by McClung et al. [2]. The trend evolved from one loop model involving transcription factors *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) and *LATE ELONGATED HYPOCOTYL* (*LHY*) acting as repressors and their activator *TIMING OF CAB EXPRESSION 1* (*TOC1*), the fundamental figure which does not function in a canonical way [8]. The complexity can be explained by robustness and flexibility. The three loop model involves the morning expressed CCA1/LHY-PRR9/PRR7 loop, the evening oscillator Y(GI)-TOC1 loop tied together via the core CCA1/LHYTOC1/X loop [9]. Between 6 and 12 genes are involved. However, the *gi toc1* and *prr9 prr7 toc1* mutants are worrisome as results showed that GI is not sufficient, no epistatic interaction between PRR9/PRR7 and TOC1 was seen and mutants could still generate robust rhythms [9]. The proposed model requires another loop explaining described mutants and is not yet complete. Analytically deciphering the transcriptional feedback loops that drive the clock is imperative. The approach is essentially different from phylogenetic analysis and examination of orthologs [10].

cis-Regulatory elements exert powers over gene expression patterns and from evolutionary perspective seem to be responsible for adjustment. Known elements investigated include the Evening Element (EE), Morning Element (ME), GATA, G box, Starch box (SBX), Telo box (TBX), Protein box (PBX), I Box, E box, TATA box and many others particularly belonging to GATA family. The establishment of a circadian group of genes can lead to regulatory elements that serve as input for circadian predictive model employing Naive Bayes classifiers. Soft clustering of genes done in parallel was used to confirm known and potential elements. A Bayesian network similar to the one in the yeast example was created considering improvements proposed by Yuan et al. in [11]. The aim was to produce the first such network on plant time course expression data. Meticulous in silico testing using a repertoire of algorithms to eliminate false positives and to ensure that a lack of susceptibility to false negatives was employed, setting stringent, optimal criteria.

6.2 Results

6.2.1 Transcriptome Profiling

DNA microarray technology was used to decipher network architecture of the molecular clock that controls "temporal compartmentalization of processes." The GeneChip Arabidopsis ATH1 Genome Array representing 24,000 genes was used resulting in high-density microarray expression data. Collection happened in 4-h intervals over two circadian cycles resulting in 13 samples per plant, assayed on the ATH1 array. The initial stage of investigation posed two fundamental questions: whether periodicity exists and in which phases of a day do specific genes peak. Stringent criteria were set to select genes that complied with the predicted model of expression step validated using multiple criteria. A double filter was used as actively expressed genes below a set threshold were eliminated (average expression below 10). Genes that passed underwent further analysis of phase, amplitude of oscillation. Apart from selecting genes at this stage, certain universal principles that went accordingly with the hypothesis were adhered to. Selection was for fold change (excluding peak and trough variation <1.50), not absolute values. According to Lomb Scargle, 398 genes were passed to the next step ($p_{\text{corrected}} < 0.05$). According to these results, at least 1.8% of transcriptome is strictly controlled by circadian clock. Phase expression is volatile nonetheless stringent statistical criteria regarding oscillatory character of every gene profile, should have eliminated false positives entirely. Output generated by Lomb Scargle, that is 398 genes selected, was finally divided into six phase bins. Each bin was enriched in genes with similar peak of expression in the array. Correctness of the whole dataset is shown by the heat map which beautifully portrays the "time course-ness" of the data Fig. 6.1 [12].

The heat map was done on the entire circadian set of 398 genes and compared to random set. Principle component analysis (Fig. 6.2) shows the stringent clustering

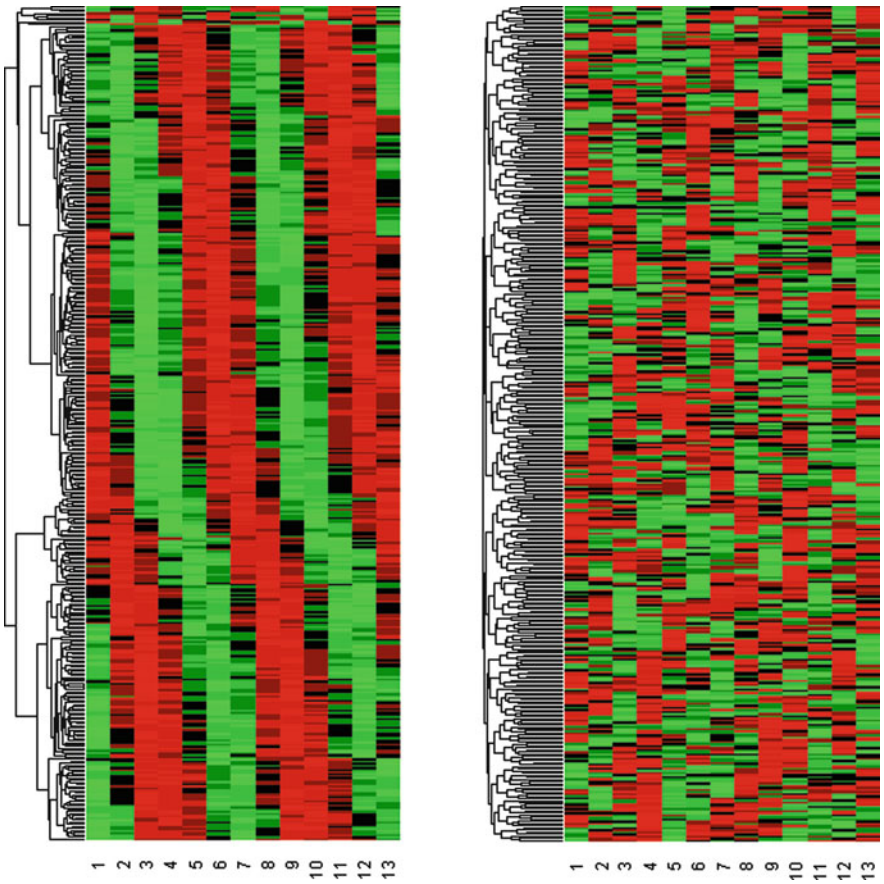


Fig. 6.1 Heat map to the *far left* represents carefully selected genes in phases. When compared to the heat map on the *right* which is made on random genes, the rigid selection is visible. *Green color* indicates maximum and *red* represents minimum. Heat map on the *left side* shows rhythmicity of clustered genes compared to the arrhythmic pattern on the *right side*. Corresponding phases of peaking are represented

in the two consecutive periods. It demonstrates the orthogonal components of the input vectors retaining components with highest variance [13]. Soft supervised clustering resulted in several strong clusters, similar to phase bins created with LS, with minor differences. Troublesome was the fact that nonoscillatory bins are selected where the particular interest is in oscillating ones. This was the main reason to use it only for validation purposes. Particular interest was in words that were overrepresented among genes in phase bins, by using criteria of at least one occurrence in at least 50% genes per bin, stringent cut-off criteria. These motifs will finally serve as an input for Bayesian classifier to find best scoring networks correctly predicting expression pattern. AlignACE proved to be the most effective method relying on Gibbs Sampling and was selected as a tool scanning for

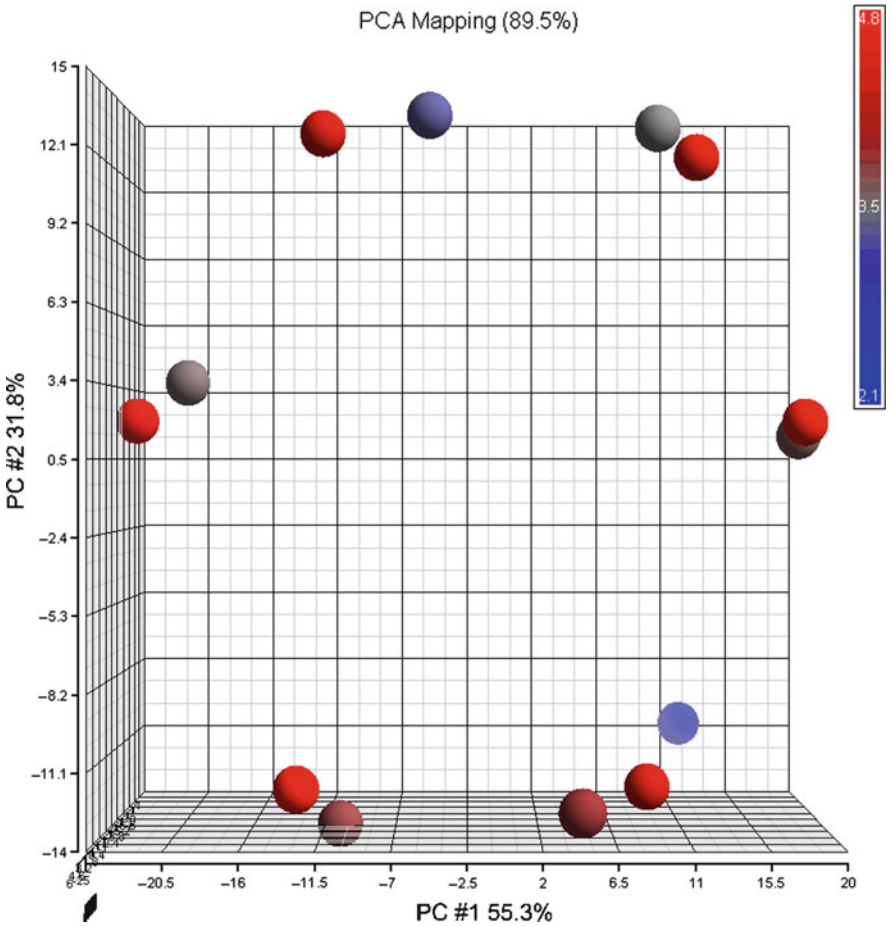


Fig. 6.2 Principal component analysis allowed for the simplification of multidimensional datasets showing the global patterns in the expression of strictly chosen set of genes by condensing the multivariate data, capturing the maximal covariation. The graph was obtained by plotting principal components of 398 genes in circadian set. There are apparent clusters that resemble the phases and are repeated due to consecutive cycles. No outliers are visible. The coherence is high 89.3%. Phase 1 is represented by *three dots* and proceeds to the right and around the cycle

cis-regulatory elements. Apart from confirming Evening Element, TATA box, G box, GATA family motifs, including AGATA, I box, GATA motif 6 passed calibrated threshold significantly in correct phase.

6.2.2 Genemining Output

Apart from the initial transcriptome analysis, a more advanced procedure which was investigating pairs and triplets using a Genemining V2.3 algorithm specifically

Phase	Pait Motif 1	Pait Motif 2
1	AGATA450	TATAAAA500
2	CAAAAAA250	TATAAAA500
3	AATATC500	TATAAAA500
4	ACGTC350	AATATC450
5	CAAAAAA450	TATAAAA500
6	AGATA450	ACCAC500

Fig. 6.3 Strongest pair in phase 3 which includes TATA box and evening element. Sequence constraints (*X*1, *X*2) selected by each network and the combinations which are predictive of a particular expression pattern are presented. The outcome of genes participating in the given pattern for each state are shown as filled for presence and empty for absence. The probabilities of being in specific expression patterns are represented by *color* from *weak green* to *strong red*. The values are not weighted

designed for such a purpose was carried out. It enables comparison of every single combination on a given dataset, using logical operator AND. Output is given as a pair of features with the highest predictive power. Prediction is weighted mean of correctly predicted elements in both tested groups from which one is composed of genes from one phase solely versus remaining genes. Results in Fig. 6.3 demonstrate the strongest pairs showing constraints and the changes in expression that happen at every combination. The output of TATA box and EE was highly significant as was EE with E box core, consecutively in phases 3 and 4 fundamentally changing the outlook upon circadian perspective being never described before.

6.2.3 Bayes Output

Bayesian analysis resulted in networks containing nodes which in this study represent the positional constraint of particular motif chosen by AlignACE. Length analyzed was 1,000 bp upstream. The aim of probability network, like Bayesian a network, is to predict certain class of an object giving the available data describing each element. The intention was to show how unique gene expression pattern can be predicted by inputing information about upstream DNA sequence. A well known, efficient method of fivefold cross-validation was used to estimate more precisely our predictions. Genes within each of six clusters were randomly divided into five sets of equal sizes. For each of five sets, 20% was taken out to act as a template for test set, the remaining 80% was used to create training set on which Bayesian Network was done with results in Table 6.1. Our Bayes classifier was trained on input from each of five training sets and preselected motifs taken from AlignACE.

Table 6.1 Output of Bayesian network considering the input variables and parent nodes. The results are high on prediction surpassing the barrier of 80%

Expression pattern	Number of genes		Number of genes	
	In training sets	Correctly predicted	In test sets	Correctly predicted
1	46	0.83	46	0.79
2	38	0.82	38	0.77
3	85	0.74	85	0.69
4	31	0.87	31	0.81
5	76	0.69	76	0.66
6	122	0.63	122	0.59
	398	0.76	398	0.71

Input was remodeled in such a fashion that not solely present in 1,000 bp upstream, as either true or false was used, but rather position from ATG starting sequence. Position from starting sequence was implemented in 50 bp increments. Orientation of a motif was not used in accordance with [11]. Subsequently, models trained were used to predict the belongingness of each element in test set. Doubts remain whether simple dichotomization is appropriate. However, results showed that 0 1 approach led to true predictions, meaning the model can be used in the future with added players. Bayesian network learned combinatorial codes for gene regulation of circadian code. Dense networks, which had more than four nodes were penalized. Overfitting is the usual outcome of many networks with worsening overall true prediction.

Models tend to progress toward more complex as being highly significant, therefore model with optimal number of parent nodes was chosen. Results of prediction obtained in test sets were in line with those in the training set, reducing the probability that they are due to chance. For the six classes representing six phases, six models were fitted and the genes in the test set were assigned to the class with the optimal model. For classes 1 6 the fraction of correctly predicted genes for a particular network was calculated and genes were added to a class with the highest fraction. As an example, cluster 1 is used and its preselected e motifs to describe naïve Bayes model fitting procedure. The class label variable was noted as Y and the preselected top e covariates as $X1, \ldots, Xe$. Figure 6.4 demonstrates the words, including the known ones such as AGATA and EE that will be further analyzed and that cover the day comprehensively.

6.3 Discussion

6.3.1 Silicosection and Bayes

The clock perhaps involves the multitude of low and high penetrance alleles which in performance with exogenous entraining *Zeitgebers* result in variable outcomes not easily detectable by initial transcriptome platforms, that is why Bayesian

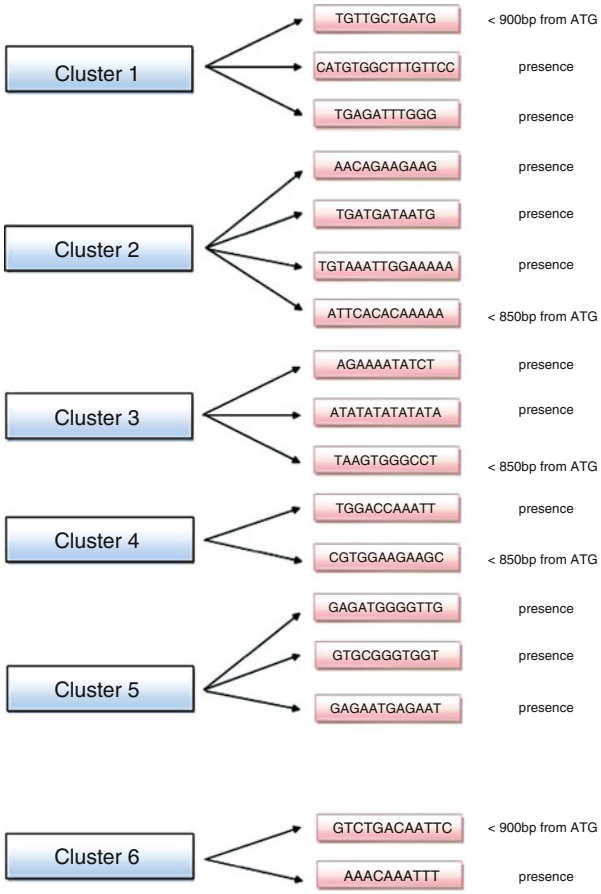


Fig. 6.4 Output of Bayes classifiers resulted in couple of nodes per cluster highlighting known elements like the evening element and at the same time proposing new circadian model and combinatorial rules with even 87% confidence. The model is comprehensive and certainly more than single element is responsible per phase. The model was unsupervised. The dyads will require experimental validation

network was employed exemplifying the potential of fertile transcriptome data. In the era of systems biology where "uncertainty is ubiquitous" studies utilizing probability networks provide a way to elucidate at the components of the regulon. A major paradigm shift took place as the combinations are responsible and require validation. The presence of Evening Element together with TATA box in phase 3 demonstrated the power of the approach and showed the importance of experimental validation of the rest. The models had a prediction rate up to 87%. Bayesian classifiers outperformed other tools proving to be effective tools for combinatorial rules of gene regulation. Results showed the importance of combinations in large-scale studies. Implementation of Genemining algorithm is a step forward in that direction. Again the pairs and triplets effectively covered all the phases. Apart from

phases, the study involved examination of ratios that is not only pairs of different motifs but frequencies of particular motif. The Bayes model proposed several new elements involving Operator OR among which EE is present. They require experimental validation. It would be interesting to create supervised model creating the input, rather than using one of AlignACE. Noisy, dimensional, significantly undersampled, pretentiously rich, with artifacts as Hemant Ishwaran called them, are some of the characteristics of microarray data with its unprecedented abilities [14]. Maybe now methylation arrays on local sites should be investigated. From previous studies, it is clear that site directed mutagenesis can change peak transcription due to the conversion of EE into CBS [15]. Current research confirms it revealing partner sequences equally responsible.

6.3.2 *Overrepresented Previously Known Motifs*

The percent of genes that are strictly circadian-regulated appears to be lower than expected, that is 1.8% resembling crème de la crème. It is the lowest reported value, however, that does not mean other genes are not circadian but these certainly are. The 1.8% circadian genes provide high signal to noise within the data making the snapshots rather clear. Multiple methods prevent in any way the results from being false, simply overly stringent. Top scoring genes included LHY, CCA1, PRR7, Sigma Factor 5 involved in the initiation of transcription and interestingly GATA-4 among others that are important. GI unfortunately was not on the array. Type-IV zinc-finger proteins have been described extensively in animals and fungi as related to light signal, never with these direct roles though [16].

6.3.3 *Novel Role of GATA Family Motifs*

Many promoters contain a conserved GATA motif in their 5-upstream region especially involved in light dependent transcriptional regulation [17]. GATA transcription factor family has orthologs across species, including other plants like rice, peas, however, in both fungi and vertebrates. Already long ago, investigating the *Dra* site I of plastocyanin (Pc) using wet lab techniques revealed this interesting motif with orthologs in pea promoter [18]. This 9 bp motif known as the AGATA box (AAAAGATAT) with flanking box II and box III being sites where GT-1 protein binds became of central importance as this protein is involved in light dependent transcription. Proof for the binding comes out of methylation in rbcS-3A gene in pea. Pc gene having AGATA box and neighboring GT-1 binding sites was the beginning. The AGATA box was the possible candidate for 3AF5 binding site. Due to light regulation involvement, it appears that this binding in Pc and ribulose-1,5-bisphosphate carboxylase-3A genes is the way that light signal initiates gene expression. As many genes in particular phase resemble the GATA motifs and

GT-1 functioning, maybe this machinery mediates signal input to the clock. Another suspicious thing was the expression observed at night as if the clock not light would interplay. CAB GATA Factor 1(CGFB1), which binds circadian regulatory element in the *Arabidopsis CAB2* promoter, requires matching sequence binding as GT-1 [19]. Similarities between CGF-1 and GT-1 suggest that GT-1-like factor can be involved in circadian regulation of other circadian genes apart from the *CAB2* gene [19]. If different factors can bind the same sequence, that is equally good and bad. GATA factors and their domains like CONSTANS, CO-like, and TOC1 (CCT) are found in other proteins involved in signaling like TOC1 positive regulator of famous transcription factors LHY CCA1 [20]. The presence of these and similar I-boxes in famous clock-regulated *RBCS*, *CAB,* and GAP genes confirms the hypothesis. Their deletion decreases the activity of promoters. G-boxes often signal the presence of the other motif which would suggest the clue to circadian expression [20]. GATA factors, initially believed to interact with the a/tGATAg/a motif, are of interest as it appears that they might show the degree of difference in binding due to circadian control as the results would suggest with some fluctuation in phases up to 85% of genes per phase. Plant type-IV zinc-finger proteins like the ones in top-scoring genes would essentially be playing important roles. Parallel evidence coming from other species made the *Z*-scores and frequencies of these groups of motifs, interesting and the results suggest experimental validation of specific once is worthwhile too. It appears not all are on the ATH1 array unfortunately 8 are missing from the 29. GATA 1, 3, 7, 8, 25 peak around CCA1 around hour 24 in silico confirmation gave further evidence [16].

6.4 Conclusion

Microarrays represent, in essence the best high throughput, density technology to decipher networks like the transcriptional circadian network investigated. Algorithms used certainly show the involvement of the GATA family motifs in circadian biology. The idea of acetylation and methylation nearby occurrence made it possible to identify significant numbers of human elements within a short 30 Mb section, new signatures and combinatorial codes are becoming clear in circadian field [21]. Pairs, ratios, and frequency analysis coupled with Bayes network highlighted the importance of groups per phase rather than single motifs creating a new *cis* model. Amid this outpouring of results, certain stand out for uncovering the intricacies of how central oscillator gives rise to rhythms. From bioinformatics perspective, Genemining V2.3 and the network are powerful in terms of throughput but foremost applications as supervised networks can be tested. Understanding the system is the Hamilton's curve of today, and so is prediction of gene expression from sequence on time course data. The understanding of higher order structure of *cis*-regulatory modules should allow modifications that could result in increased hybrid vigor. Idiosyncratic analysis of transcriptome datasets is slowly progressing to the interactome research of the future [14].

6.5 Methods

6.5.1 Mining for Patterns

Hybridization data was obtained from a public database (NASCArrays), courtesy Dr. Kieron Edwards. Sets were validated using the data provided by Todd Michael with accession number EMEXP1304 [15]. Data mining included several methods for extraction and validation.

Filter Feature Selection algorithm was applied. A modified method of pattern fitting was used to select genes with oscillatory pattern which was the adapted version of the feature selection template matching, SINEMINE. Genes with p value of <0.05 after Bonferroni correction, mean expression value of 10 or more, and a minimum of 1.5-fold between the minimum and maximum expression level were selected. Ultimately, genes were grouped with respect to their specific peak time.

The Haystack algorithm successfully reduces type I and type II errors. It is amplitude independent relying on linear least squares regression. Information regarding amplitude comes from "variance of the linear regression detrended original time series" [4]. Low sampling rate of two consecutive periods and stochastic noise-posed obstacles that make Lomb Scargle suitable.

Lomb Scargle algorithm, written by Earl Glynn was the mainstream method applied. The idea behind using the Lomb Scargle stems from Fast Fourier being appropriate for spaced data, with more than two periods, not necessarily the situation with high throughput analysis and low sampling [22]. It is the method of choice for nonuniformly spaced data. It involves the calculation of the normal Fourier spectrum. It can be called "slow" and better adaptations of FFT will shortly be available. It behaves well up to Nyquist limit, however, beyond spurious events can be observed. As a correction for multiple testing, FDR was applied to output data.

6.5.2 Motif Search

LS output served as input for ELEMENT, MEME, and AlignACE. Methods include enumerative and alignment approaches [10]. As the Beer and Tavazoie stated "heroic experimental effort to elucidate," the regulon code are required involving other data sets [7].

Genes per bin were inputed to ELEMENT with settings set to 500 bp upstream. "Z-score profiles" were then created for words of interest. FDR was applied to the output p-values. Such p-value inspection allowed for setting the threshold. It would be best if the Z-score profiling would lead to representation of different words among separate phase bins. Words occurring at least once in more than 50% of genes within a bin were considered for further analysis. Nonetheless, Z-scores did not reflect expected complexity among found motifs, selecting inadequate ones.

MEME software gave estimates of *E*-values focusing on outputs with lowest 0.001 instead of randomness artifact. Background model is 0-order Markov. MEME computes the significance of their product through computation of each column individually. The *E*-value of motif therefore depends on several factors and those are width, log likelihood ratio, number of hits, size of training set, 0-order portion of the background model, and way of site distribution. It is stringent and prone to false negatives which were the reasons to pick AlignACE as a main motif scanner.

While motif search programs concentrate on statistical overrepresentation within, some function exceptionally considering statistical underrepresentation and position. To the best of knowledge, of particular efficacy, was AlignACE which was the method of choice and considered 1,000 bp upstream of the genes of interest.

6.5.3 Clustering

Similarity in behavior can be assessed by Euclidean distance, dot products, and correlation coefficient looking at shape [12]. Soft clustering was used to confirm the validity of the gene content of phase bins selected by LS. Having certain a priori hypothesis allows for supervised, soft clustering which is noise robust, assigns genes to several clusters creating information rich groups and differentiates strength, shows the ability of clusters to represent genes and global clustering structure [23]. Hard does not fit time course data. For example, *K* means clustering which suffers from too stringent criteria such as one gene per cluster which prevents differentiation, and how well gene is portrayed by centroid causes the loss of information. Far from trivial selection of *k* parameter is worrisome. It suffers from sensitivity to noise and detection of random clusters. Soft, partitional algorithm was implemented called the fuzzy *c* means showing shared genes. Mfuzz R package clustering function was used. The focus was on clear signal interactions. Methods like random forest were considered, however, seemed inappropriate [24]. Clustering whether exclusive, inclusive, intrinsic, extrinsic, hierarchical, and partitional with regards to Kleinberg's impossibility theorem, none is perfectly consistent [13].

6.5.4 Genemining V2.3

The program focuses on a paradigm shift from single motifs to groups of motifs and the interplay between them. The software itself might be used for other tasks such as interplay between SNPs, but this time its input data was tailored for motif interaction findings. It executes an exhaustive evaluation of all possible combinations of motifs/SNP's AND'd together for a given model size (e.g., pairs, triples, etc.). For each model, a measure of differentiation between a test group and a control group is determined. To achieve a rating of 100%, every sample in the test group would have a particular pattern of base pairs which *none* of the control group

has. The top M models are output. The calculations are computationally intensive due to their combinatorial nature. For a given set of n motifs with a model size of r, there are nCr (n items taken r at a time) combinations which result in order n^r complexity. A one million point SNP array with 1,000 samples would require order 10E15 comparisons for even the simplest pairwise analysis. Genemining version 2.x implemented a densified data model and calculation algorithm which resulted in a 7X (700%) speed improvement. Version 3.x has been ported to NVIDIA's CUDA architecture to leverage the price/performance of inexpensive massively parallel processing general purpose graphics processing unit (GPGPU) hardware. Performance gains are expected to scale significantly higher on newer, faster, and more parallel cards. While Bayesian Networks are prone to choose models with logical operator OR, it is of great importance to test the same dataset with models considering exclusively the logical operator AND.

6.5.5 Learning Structure of Bayesian Network

Bayesian networks as described by Pearl make models of joint multivariate probability distributions reality. None was done up to date on plant time course data. The aim was to relate similar expression patterns (output) with sequence elements (input), where sequence elements and corresponding transcription factors should lead to correct expression patterns. A BN can be used to compute the conditional probability of one node, given values assigned to the other nodes, therefore it can be used as a classifier that gives the posterior probability distribution of the class node given the values of other attributes. In our classification process, we used BN PowerPredictor. The classification process in each case of test set was done by choosing a class label, the value of class variable with the highest scoring posterior probability, based on instantiations of the feature nodes. Finally, the top scoring network was defined by percentage of correctly predicted genes in each class.

Acknowledgments We thank Paul Glynn Earl from Stowers Institute, Todd Michael from Rutgers University, Yuan Yuan from Harvard University, Matthias Futchik from Humboldt University for many immense comments and advice.

References

1. Eckardt, N. "A wheel within a wheel: temperature compensation of the circadian clock". *The Plant Cell* 18: 1105 1108, (2006).
2. Ueda, H. "Systems biology flowering in the plant clock field". *Molecular Systems Biology* 2: 60, (2006).
3. Bussemaker, H., Li, H., Siggia, E. "Regulatory element detection using correlation with expression". *Nature Genetics* 27: 167 171, (2001).

4. Straume, M. “DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning”. *Methods in Enzymology* 383: 149 166, (2004).
5. Janga, S., Collado Vides, J., Babu, M. “Transcriptional regulation constrains the organization of genes on eukaryotic chromosomes”. *Proceedings of National Academy of Sciences of the United States of America* 105: 15761 15766, (2008).
6. Tan, K., Tegner, J., Ravasi, T. “Integrated approaches to uncovering transcription regulatory networks in mammalian cells”. *Genomics* 91: 219 231, (2008).
7. Beer, M., Tavazoie, S. “Predicting gene expression from sequence”. *Cell* 117: 185 198, (2004)
8. McClung, CR. “Plant circadian rhythms”. *The Plant Cell* 18: 792 803, (2006)
9. Ito, S., Kawamura, H., Niwa, Y., Nakamichi, N., Yamashino, T., Mizuno, T. “A genetic study of the Arabidopsis circadian clock with reference to the TIMING OF CAB EXPRESSION 1 (TOC1) gene”. *Plant Cell Physiology* 50: 290 303, (2009)
10. Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P., Van de Peer, Y. “Computa tional approaches to identify promoters and *cis* regulatory elements in plant genomes”. *Plant Physiology* 132: 1162 1176, (2003).
11. Yuan, Y., Guo, L., Shen, L., Liu, J. “Predicting gene expression from sequence: a reexamina tion”. *PLOS Computational Biology* 3: e243, (2007).
12. Eisen, M. “Cluster analysis and display of genome wide expression patterns”. *Proceedings of National Academy of Sciences of the United States of America* 95: 14863 14868, (1998).
13. Nayak, A., Stojmenovic, I. “Handbook of Applied Algorithms”. Hoboken: Wiley, 2008.
14. Zhu, M., Wu, Q. “Transcription network construction for large scale microarray datasets using a high performance computing approach”. *BMC Genomics* 9: 841, (2008).
15. Michael, T., McClung, C. “Phase specific circadian clock regulatory elements in Arabidop sis”. *Plant Physiology* 130(2): 627 638, (2002).
16. Manfield, W., Devlin, P., Jen, C., Westhead, D., Gilmartin, P. “Conservation, convergence, and divergence of light responsive, circadian regulated, and tissue specific expression pat terns during evolution of the Arabidopsis GATA gene family”. *Plant Physiology* 143: 941 958, (2007).
17. Teakle, G., Manfield, I., Graham, J., Gilmartin, P. “*Arabidopsis thaliana* GATA factors: organisation, expression and DNA binding characteristics”. *Plant Molecular Biology* 50: 43 57, (2002).
18. Fisscher, U., Weisbeek, P., Smeekens, S. “Identification of potential regulatory elements in the far upstream region of the *Arabidopsis thaliana* plastocyanin promoter”. *Plant Molecular Biology* 26: 873 886, (1994)
19. Anderson, S., Kay, S. “Functional dissection of circadian clock and phytochromes regulated transcription of the Arabidopsis CAB2 gene”. *Proceedings of National Academy of Sciences of the United States of America* 92: 1500 1504, (1995).
20. Reyes, J.C., Muro Pastor, M.I., Florencio, F.J. “The GATA family of transcription factors in Arabidopsis and rice”. *Plant Physiology* 134: 1718 1732, (2004).
21. Heintzman, N.D., Stuart, R.K., Hon, G., Fu, Y., Ching, C.W., Hawkins, R.D., Barrera, L.O., Van Calcar, S., Qu, C., Ching, K.A., Wang, W., Weng, Z., Green, R.D., Crawford, G.E., Ren, B. “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome”. *Nature Genetics* 39: 284 285, (2007).
22. Glynn, E., Chen, J., Mushegian, A. “Detecting periodic patterns in unevenly spaced gene expression time series using Lomb Scargle periodograms”. *Bioinformatics* 22: 310 316, (2006).
23. Futschik, M., Charlisle, B. “Noise robust clustering of gene expression time course”. *Journal of Bioinformatics and Computational Biology* 4: 965 988, (2007).
24. Ishwaran, H., Rao, S., Kogalur U. “BAMarray™: Java software for Bayesian analysis of variance for microarray data”. *BMC Bioinformatics* 7: 59, (2006).

Chapter 7
ChemBrowser: A Flexible Framework for Mining Chemical Documents

Xian Wu, Li Zhang, Ying Chen, James Rhodes, Thomas D. Griffin, Stephen K. Boyer, Alfredo Alba, and Keke Cai

Abstract The ability to extract chemical and biological entities and relations from text documents automatically has great value to biochemical research and development activities. The growing maturity of text mining and artificial intelligence technologies shows promise in enabling such automatic chemical entity extraction capabilities (called "Chemical Annotation" in this paper). Many techniques have been reported in the literature, ranging from dictionary and rule-based techniques to machine learning approaches. In practice, we found that no single technique works well in all cases. A combinatorial approach that allows one to quickly compose different annotation techniques together for a given situation is most effective. In this paper, we describe the key challenges we face in real-world chemical annotation scenarios. We then present a solution called ChemBrowser which has a flexible framework for chemical annotation. ChemBrowser includes a suite of customizable processing units that might be utilized in a chemical annotator, a high-level language that describes the composition of various processing units that would form a chemical annotator, and an execution engine that translates the composition language to an actual annotator that can generate annotation results for a given set of documents. We demonstrate the impact of this approach by tailoring an annotator for extracting chemical names from patent documents and show how this annotator can be easily modified with simple configuration alone.

Keywords Chemical annotator · ChemBrowser · Name entity recognition · Flexible framework

X. Wu (✉)
IBM China Research Lab, Beijing 100193, China
e mail: wuxian@cn.ibm.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 7,

7.1 Introduction

Leveraging widely available content from diverse information sources, such as web, patents, scientific articles, and news for biochemical research, is becoming increasingly critical. However, the volume and the diversity of the information post significant challenges for scientists to effectively digest such information and derive insights. Text analytics is poised as an emerging area to tackle such problems in the biochemical space.

One of the initial and fundamental steps of biochemical research deals with extracting and understanding chemical and biological entities that are being discussed in documents. Such entities, once extracted, can be used for subsequent analysis, such as understanding their relationships, similarities and differences, and trends. We call such entity extraction process an annotation process in this paper.

Chemical name extraction is a Named Entity Recognition (NER) task which is a fundamental task in natural language processing with lots of previous approaches, such as Markovian Models (HMM [1], MeMM [2]) and conditional random field (CRF) [3]. In the past few years of our work on chemical name annotation, we have experimented with the above approaches on diverse chemical documents, including articles, patents, web data, etc. We found that

- The effectiveness of supervised approaches highly depends on the quality and quantity of the training data set. However, in practice, it is often impractical to collect such training data.
- Each individual technique alone may work on one data set and in one situation but fail on others.
- In practice, scientists often need to be able to quickly customize annotators for their own purposes and then validate immediately. However, many techniques are expensive to run, making it even more difficult to adapt to different situations.

These observations suggest that real world data and usage scenarios require flexible annotation techniques that have the following properties:

- Adaptive to different situations with little development efforts.
- Fast runtime to allow rapid development and testing cycles and easy tuning.
- Low cost on computation and processing power to allow for annotation over large text corpus.
- High accuracy in annotation results.

In this paper, we present such a new approach, which is embedded in a working system called ChemBrowser. Overall, ChemBrowser captures a flexible annotation framework that contains three key components:

- A suite of basic and unique annotation processing units, such as the dictionary processing unit, OCR rule engine, filters, etc.
- A high-level language that describes how the basic annotation processing units can be composed and linked to form an annotator in a workflow.

- A runtime execution engine which executes the language description as a runtime annotator.

The rest of the paper is organized as follows: Section 7.2 describes the key real-world challenges in chemical annotation, especially on patents, an important corpus for biochemical research. Section 7.3 presents the flexible annotation framework we developed. Section 7.3 shows our experimental results with ChemBrowser on different patent data sets. Section 7.4 concludes and outlines future work.

7.2 Chemical Annotation Framework

Given the challenges described above, we concluded that it is going to be extremely difficult to find a single magic annotation algorithm that would work for all situations. Instead, what is needed is a flexible framework that allows one to quickly configure and construct a chemical annotator for different situations. In addition, such composite annotators must be easy to use, easy to test, fast to validate against data sets, high quality, and high performance for large corpuses. This section describes such a chemical annotation framework.

7.2.1 Chemical Annotation Processing Units

ChemBrowser includes a suite of simple but efficient annotation units. Such basic units can be modified or evolved overtime. We describe several key units that we have developed over time for annotating patents:

- *Non-Chemical Term Filter* (*also called English dictionary filter*)*:* Due to the diversity of chemical nomenclature, a straightforward dictionary-based approach is impossible. Instead, we created a dictionary of nonchemical terms which contains all words in a given language that are not chemically related. Such dictionaries can be easily obtained for many languages.
- *Pattern Filter:* Although chemical nomenclature is diverse, there are common string patterns in chemical names. We developed such a pattern filter which contains common chemical string patterns.
- *NGram Filter:* Stores the N-Grams which are usually a part of a chemical name. Vasserman [5] has studied the N-Gram methods and indicated good performance overall.
- *Length Filter and Number Filter:* Normally, chemical names are at least four characters long, and the names typically cannot be all numbers.

An intelligent combination of the above processing units can be highly effective. For example, the Non-Chemical Term filter ensures that all chemical names are retained (100% recall), and the pattern filter ensures high precision. With the

composition of these simple processing units and filters in some meaningful manner, one can potentially construct highly efficient and highly effective annotators. These processing units can be customized and developed over time as well.

7.2.2 High-Level Modeling and Workflow Composition Language

To effectively enable the users to compose high-quality annotators from basic units, we develop a high-level modeling language to describe the logics and workflow that link the units. The language contains six types of language elements as listed in Table 7.1. *Tokenizer* is used to segment text into tokens. *Decision* unit functions as various kinds of annotation filters to process tokens and distribute tokens to different subsequent units, such as the pattern filter listed in Sect. 7.2.1. Other forms of string processors are classified as *Processing* unit, such as OCR correction, which just modifies the token and passes them to next unit. *Input*, *Output*, and *Repository* describe the linkages among various components and the storage for intermediate or final results.

In the ChemBrowser implementation, we adopted XML as the modeling language framework, since XML is well known, easy to edit and process. All language components are described using XML. Given the processing units and the language files, ChemBrowser constructs a runtime annotator using a Runtime Annotation Engine. Such an engine parses the intermediate modeling language, and uses the mapping file to map the modeling language to real processing unit executables. Then, it executes the entire workflow as a composite annotator.

7.2.3 Chemical Annotator for Patents

Figure 7.1 illustrates the workflow of a real-world chemical annotator for patents. The annotating process is divided into two steps. In the first step, the content of the patents are tokenized to individual terms by blank space, line break, punctuations, and other splitting characters. Each term is sent to a series of filters to determine

Table 7.1 The definition of unit types in annotator modeling language

Type	Function
Tokenizer	Segment text into tokens with a split parameter
Decision	Filter the tokens by certain conditions, such as patterns or length etc, and send to different successors
Processing	Modify the token, such as correcting potential OCR errors, removing punctuations in the head or the tail of the token
Output	Save the tokens to Repository unit
Input	Read the tokens from Repository unit
Repository	Store the intermediate or final annotating results

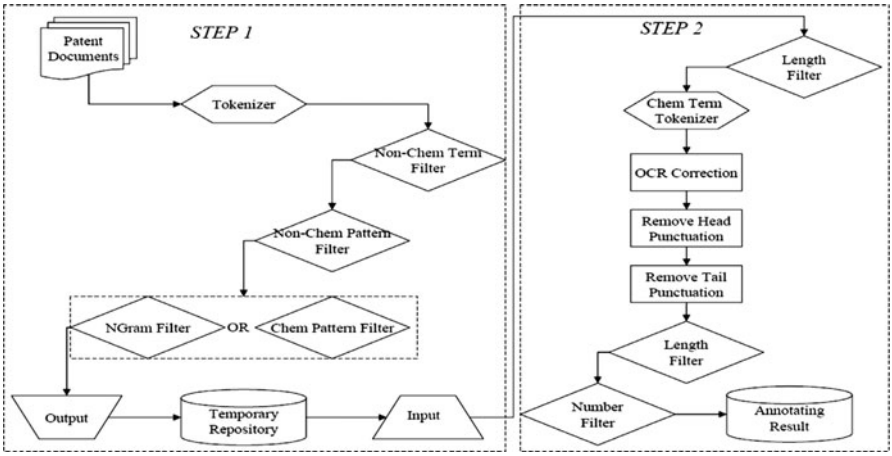


Fig. 7.1 Chemical annotator for patents built in our framework

whether this token is a part of a chemical term or not. All the identified terms are saved to the temporary repository and the consecutive ones are connected and grouped as one single unit. In the second step, since a unit of multiple terms may be composed by several chemical names, the units are picked sequentially and split again with rules summarized from the nomenclature of chemical names, such as common suffix and prefix strings, etc. The split segments are treated as candidate chemical names and are sent to several processing units to modify the OCR errors and remove the punctuation attached in the head or the tail. Since chemical names are usually longer than four characters containing no numbers, the candidate names are further verified by "Number Filter" and "Length Filter" and finally saved in the "Annotating Result Repository."

Note the recall of both the "Non-Chem Term Filter" and the "Non-Chem Pattern Filter" is very high (i.e., candidate set can be reduced without losing the real chemical names). On the other hand, "Chem Pattern Filter" and "NGram Filter" are precision-enhancing filters. The combination of those filters can enhance the quality of the annotator. The tokens that passed all three decision units are saved into the temporary cache for subsequent processing. This annotator can be integrated into a chemical search and analysis system to enable chemical compound search and affinity analysis in patent data, such as in [4].

7.3 Experiments

We ran the annotator shown in Fig. 7.1 on a patent corpus, since patents are extremely rich in chemicals. The corpus contains close to ten million patents from 1976 to 2008. We identified 3,036,803 patents that had at least one chemical

Table 7.2 Change of annotation results when one decision unit is removed

		EP	US	WO
Chemical annotator		810	10,046	3,326
Annotator removing	Non Chem Term Filter	+13	+505	+353
	Non Chem Pattern Filter	+23	+1,867	+895
	Pattern Filter	10	38	30
	NGram Filter	4	+30	+55
	Length Filter	0	0	0
	Number Filter	0	0	0

structure. In total, our chemical annotator extracted 75,110,891 chemical occurrences. A chemical occurrence is defined as a unique compound per patent. These chemical occurrences resulted in 4,218,237 unique structures. The unique composition of such processing units and workflow is extremely effective in identifying chemicals in patents.

To evaluate the effectiveness of our overall approach, we devise two sets of test cases, both intend to extract chemical names out of patents. The first case evaluates the effectiveness of the overall framework by showing how results might change when we modify the elements of annotators in the framework. The second case shows the effectiveness of processing units by calculating the number of tokens being modified.

In our experiments, we used a randomly selected subset of patents from the patent corpus that contained at least one chemical structure. Our subset contained 100 US patents, 100 WOPCT patents, and 100 European (EP) patents. As described earlier, WO and EP patents are more prone to OCR errors. We performed the experiments on a ThinkCentre with a two-way processor of 2.16 GHz and 2 GB memory.

As shown in Table 7.2, our annotator ran on the three patent data sets and identified 810, 10,046, and 3,326 chemical names, respectively. This is our baseline result. We then modified the annotator by editing the high-level modeling language progressively. Specifically, we removed some decision units from the annotator and performed experiments on the modified annotator. From the results, we see some processing units have dramatic effects on annotation results on this data set. For example, the "Non-Chem Term Filter" and "Non Chem-Pattern Filter" are key processing units. Removing them causes the annotator to pick up a significant number of false positives. However, removing "Length Filter" and "Number Filter" had no impact on this set of patents at all. In reality, when we expand the data set, those filters are still useful.

In addition to changes made to the decision units, we also tested out the effect of processing units. We examined the impact of the OCR correction and punctuation processing units. Table 7.3 lists the percentage of the tokens that are modified by each of the three processing units. Note that significant tokens require processing on punctuation. For OCR error correction, EP and WO patents required more OCR corrections than the US patents. To evaluate the speed of our annotator run, we also

Table 7.3 Ratio of tokens modified by the processing unit

Process unit	EP (%)	US (%)	WO (%)
OCR correction	12.3	7.0	15.5
Remove head punctuation	3.8	3.5	2.2
Remove tail punctuation	77.4	63.0	99.8

compared the processing times of our annotator against machine learning-based annotators. We found that our annotator is more than 20 times faster than the ones we experimented with. One machine learning-based annotator could take on average up to 6 s to annotate each patent. Annotating ten million patents would take close to 700 days. Our annotator can run over all patent in the corpus in less than 1 day.

7.4 Conclusion

This paper presents a flexible framework for chemical annotation. The framework is designed to allow users to quickly compose effective annotators for different situations using simple and easy-to-understand processing units and workflows. The overall framework consists of three key elements: a suite of basic processing units, a high-level modeling language that describes the composition of the processing units in some form of workflow and a runtime execution engine that executes the description of runtime annotation purposes. The overall approach has been embedded in the ChemBrowser solution.

We have presented experimental results of using such composite annotators on patent data and show that with such a framework, one can quickly modify annotation processing units or workflows to adjust to different situations, such as OCR errors in patents. In real world situations, we have found that these approaches have proven to be extremely effective when compared to pure machine learning-based approaches. In addition, using simple and efficient processing units to form annotators is fast, which is important for annotating a large data corpus.

References

1. T. R. Leek. Information extraction using hidden Markov models. Master's thesis, University of California, San Diego, CA, 1997.
2. A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *Proc. 17th International Conf. on Machine Learning*, pages 591 598. Morgan Kaufmann, San Francisco, CA, 2000.
3. J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282 289. Morgan Kaufmann, San Francisco, CA, 2001.

4. J. Rhodes, S. Boyer, J. Kreulen, Y. Chen, and P. Ordonez. Mining patents using molecular similarity search. In *Proc. Pacific Symp. Biocomputing*, pages 304 315, 2007.
5. A. Vasserman. Identifying chemical names in biomedical text: An investigation of substring co occurrence based approaches. In D. M. Susan Dumais and S. Roukos, editors, HLT NAACL 2004: Student Research Workshop, pages 7 12, Boston, MA, May 2 May 7 2004. Association for Computational Linguistics.

Chapter 8
Experimental Study of Modified Voting Algorithm for Planted (*l*,*d*)-Motif Problem

Hazem M. Bahig, Mostafa M. Abbas, and Ashraf Bhery

Abstract We consider the planted (*l*,*d*)-motif search problem, which consists of finding a substring of length *l* that occurs in each s_i in a set of input sequences $\{s_1,\ldots,s_t\}$ with at most *d* substitutions. In this paper, we study the effect of using Balla, Davila, and Rajasekaran strategy on voting algorithm practically. We call this technique, modified voting algorithm. We present an experimental study between original and modified voting algorithms on simulated data from (9,*d*) to (15,*d*). The comparison shows that the voting algorithm is faster than its modification in all instances except the instance (15,3). We also study the effect of increasing *h*, which is proposed by Balla, Davila, and Rajasekaran on the modified voting algorithm. From this study, we obtained the values of the number of sequences that make the running time of modified voting algorithm less than the voting algorithm and minimum. Finally, we analyze the experimental results and give some observations according to the relations: (1) *l* is fixed and *d* is variable. (2) *l* is variable and *d* is fixed. (3) *l* and *d* are variables. (4) (*l*,*d*) is challenging.

Keywords DNA motif · Planted (*l*,*d*)-motif · Voting algorithm · Binding site · Exact algorithm

8.1 Introduction

Given a number of DNA sequences, the motif searching problem is the task of discovering a particular base sequence that appears (perhaps in a slightly mutated form) in every given sequence. Finding similar patterns (motifs) in a set of sequences has many applications in molecular biology such as locating binding

H.M. Bahig (✉)
Computer Science Division, Department of Mathematics, Faculty of Science, Ain Shams University, Cairo 11566, Egypt
e mail: hbahig@asunet.shams.edu.eg

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 8,

sites and finding conserved regions in unaligned sequences. Pevzner and Sze [8] gave a precise definition of motif searching problem as follows.

Planted (l,d)-Motif Problem (PMP): Suppose there is a fixed but unknown string M (motif) of length l. Given t length-n sequences, each of which contains a planted d-variant of M, we want to determine M without a priori knowledge of the positions of the planted d-variants. A length-l sequence s' to be a d-variant of another length-l sequence s if the Hamming distance between s' and s is at most d.

Numerous exact algorithms have been proposed to solve PMP [2, 4 7, 9, 10]. The exact algorithms always output the planted motif.

Buhler and Tompa [3] studied the limitation of PMP and found that when l, t, and n are fixed, d must not be larger than some threshold. The threshold is determined by considering the expected number $E(l,d)$ of random patterns P with at least one variant in each sequence. The value of $E(l,d)$ is equal to $4^l(1-(1-p_d)^{n-l+1})^t$, where p_d is the probability that the Hamming distance between a length-l pattern P and a randomly generated length-l string is less than or equal to d. The general formula for p_d is given by $p_d = \sum_{k=0\ldots d} {}^{l}C_k\,(1/4)^{l-k}\,(3/4)^k$.

Based on the value of $E(l,d)$ some instances have been probabilistically proved as "difficult to be solved" due to the existence of several motifs by random chance for such instances. In this case, we call it challenging instances [2]. Examples of these instances are [2]: (9,2), (11,3), (12,3), (13,4), (14,4), (15,4), and (15,5). If the value of $E(l,d)$ is very small ($E(l,d) \ll 1$), then the instance is solvable.

The effect of Bella's suggestion [2] on voting algorithm (one of the exact algorithms [5]) called modified voting algorithm. In [1], the authors studied the modified voting algorithm on challenging instance only.

In this paper, we study the effect of modified voting algorithm on simulated data from (9,d) to (15,d). We also studied the effective part in the running time for modified voting algorithm. The experimental results show that the modified voting algorithm is slower than voting algorithm in most instances. We also study the effect of increasing the value of h that is proposed by Balla et al. [2] on modified voting algorithm. From this study, we obtained the values of the number of sequences that make the running time of modified voting algorithm less than the voting algorithm and minimum. Finally, we analyze the experimental results and give some observations according to the relations (1) l is fixed and d is variable. (2) l is variable and d is fixed. (3) l and d are variables. (4) (l,d) is challenging.

The structure of paper is as follows. In Sect. 8.2, we present an overview for voting algorithm and its modification. The experimental comparison between voting algorithm and its modification is given in Sect. 8.3. In Sect. 8.4, we study the effect of increasing h on the modified voting algorithm. We also discuss the results and give some observations according to different relations. Finally, the conclusion in Sect. 8.5.

8.2 Modified Voting Algorithm

In this section, we give an overview of voting algorithm (VA) and modified voting algorithm (MVA). The main idea behind VA [5] is that each length-l substring v in the input sequence gives one vote to all length-l sequences s in $N(v,d)$, where $N(v,d)$

is the set of neighborhoods that contains all d-variants of length-l sequence s. The pseudocode for VA is given in [5].

Balla et al. [2] suggested the number of sequences, h, that are required by an algorithm to distinguish strong motif candidates from spurious signals. The value of h is given by

$$h = 1 + \log_{1/p_{2d}} n \tag{8.1}$$

Based on the value of h, Balla et al. proposed a strategy to improve practical performance of several exact algorithms for PMP such as voting, PMSP, and PMSi [5, 6, 10] by a factor of the time elapsed in generating and processing the neighborhood of the reminder sequences. The strategy consists of two phases. Phase 1: constructs a set C of common patterns that contains all neighborhoods of l-mers from h input sequences, where $h < t$. Phase 2: for each element M' in the set C, we vote M' in the reminder sequences, t h, by using string matching. The result of applying this strategy on voting algorithm is called MVA. It consists of five steps. Steps 1 3 represent the first phase while the reminder steps represent the second phase.

Algorithm: MVA.

Begin

1. *Compute* h *from (8.1).*
2. *Construct a set* C *of all neighborhoods for each* l*-mer from* h *input sequences as lines 1 8 in VA [5].*
3. *Construct a set* C′ *by selecting all neighborhoods for voting* h *as lines 9 12 in VA [5].*
4. *For each pattern in* C′*, scans the remainder input sequences,* t h*, and vote it up.*
5. *Select the set of possible candidate motifs by sorting* C′ *using integer sorting algorithm in descending order.*

End

8.3 Experimental Results

In this section, we present an experimental comparison between original and modified voting algorithms for PMP on simulated data from (9,*d*) to (15,*d*), including the solvable and challenging instances.

We implement the two algorithms on PC Intel Pentium 4 processor. The processor has a clock speed of 3.2 MHz and memory size 512 MB. The PC works under Windows XP operating system. Both algorithms are implemented by using Borland C++ compiler.

We compute the running time of each algorithm by testing the algorithm 50 iterations for each (l,d), and then we take the average of them. In each iteration, we generate the simulated data by using the same method that is used in many articles [5, 6, 10]. Each input instance of the simulated data consists of $t = 20$ random sequences and $n = 600$ nucleotide. The probability of occurrence for each nucleotide is 1/4. In our experiment, we use :

(1) s and m to represent the time in seconds and minutes, respectively.
(2) T_{VA} and T_{MVA} to represent the running time for VA and MVA, respectively.
(3) T_{MVA1} and T_{MVA2} to represent the running time for the first and the second phases, respectively, in case of MVA.
(4) $N_{\text{ex}}(l,d,h)$ to represent the average number, experimentally, of motifs in h input sequences for 50 iterations in case of (l,d).

We considered that: (1) If the running time of both algorithms are less than 1 s, then we can neglect the comparison. (2) The running time of two algorithms is equal if they are equal in seconds and minutes, if exists, even the millisecond are different.

Table 8.1 presents the running time estimation as a function of l and d for both VA and MVA keeping n and t constant. Note that in case of (11,1) and (12,1), the running time for VA and MVA is less than 1 s, so we omit these results from the table. From Table 8.1, we observe the following remarks.

1. The running time of MVA is slower than VA in all instances except the case (15,3). The last column in Table 8.1 shows how MVA takes many times as much time as VA, where $T_{\text{MVA}} \simeq \text{times} \times T_{\text{VA}}$.

Table 8.1 Comparison between T_{VA} and T_{MVA}

l	d	$E(l,d)$	T_{VA}	h	$N_{\text{ex}}(l,d,h)$	T_{MVA1}	T_{MVA2}	T_{MVA}	$T_{\text{MVA2}}/T_{\text{MVA}}$	Times
9	1	1.46E 19	0.023s	2	1,291	0.005s	3.136s	3.141s	0.998	139
9	2	1.59973	0.133s	4	24,973	0.044s	37.265s	37.309s	0.999	285
10	1	6.19E 30	0.06s	2	546	0.017s	1.373s	1.39s	0.988	23
10	2	6.11E 08	0.49s	3	13,143	0.084s	27.323s	27.407s	0.997	60
11	2	5.43E 17	0.71s	3	2,875	0.148s	6.732s	6.88s	0.978	9.7
11	3	4.72084	4.425s	4	291,491	1.1s	7.901m	7.919m	0.998	107
12	2	1.09E 26	1.22s	3	631	0.347s	1.637s	1.984s	0.825	1.6
12	3	3.19E 07	8.22s	4	45,480	1.661s	1.662m	1.689m	0.984	12
13	1	5.23E 62	0.95s	2	73	0.867s	0.205s	1.072s	0.191	1.1
13	2	1.14E 36	2.13s	2	6,194	0.991s	17.056s	18.047s	0.945	8.5
13	3	8.14E 16	11.71s	3	42,456	2.537s	1.763m	1.805m	0.977	9.25
13	4	5.23252	1.027m	6	627,381	18.35s	16.355m	16.66m	0.982	16
14	1	7.83E 73	2.7s	2	36	2.609s	0.101s	2.71s	0.037	1
14	2	8.45E 47	4.23s	2	3,060	2.834s	8.746s	11.58s	0.755	2.74
14	3	5.05E 25	17.71s	3	9,277	4.881s	24.173s	29.054s	0.832	1.6
14	4	4.20E 07	1.7m	4	421,753	21.938s	15.492m	15.858m	0.977	9
15	1	1.06E 83	10.25s	2	17	10.128s	0.051s	10.179s	0.005	1
15	2	4.90E 57	11.77s	2	1,579	10.034s	4.554s	14.588s	0.312	1.2
15	3	1.59E 34	31.68s	3	2,157	13.337s	5.841s	19.178s	0.305	0.61
15	4	2.17E 15	2.66m	4	53,887	38.153s	2.172m	2.808m	0.774	1
15	5	2.84202	11.76m	7	1,542,909	4.201m	41.263m	45.464m	0.908	3.8

2. In the case of challenging instances, (9,2), (11,3), (12,3), (13,4), (14,4), (15,4), and (15,5) the following observations are true.
   (a) The effective part of MVA is the second phase of the algorithm since $T_{\mathrm{MVA2}} \gg T_{\mathrm{MVA1}}$.
   (b) If we rearrange the challenging instances, (*l*,*d*), according to $N_{\mathrm{ex}}(l,d,h)$, then T_{MVA} for (*l*,*d*) increases with increase $N_{\mathrm{ex}}(l,d,h)$.
   (c) The ratio between T_{MVA} and T_{VA} decreases when *l* increases in the following cases: (i) $E(l,d) < 1$, i.e., (12,3), (14,4), and (15,4). (ii) $E(l,d) > 1$, i.e., (9,2), (11,3), (13,4), and (15,5).

Therefore, MVA does not improve the practical performance of VA in most instances from (9,1) to (15,5).

8.4 Effect of Increasing *h* on MVA

From Sect. 8.3, we conclude that the value of *h* does not improve the performance of VA in most instances up to (15,5). Therefore, in this section, we search experimentally for another value of the number of sequences that is applied by voting to make the performance of MVA better than VA according to the running time. In order to find this value, we study experimentally the effect of changing *h* from $h + 1$ to $t - 1$ on MVA for each instance (*l*,*d*). We use *nsv* to represent the number of sequences that is applied by voting (first phase), where $h < nsv < t$. Figures 8.1 8.8 represent the effect of increasing *h* on MVA, for the instances (15, $d \leq 3$), (14, $d \leq 3$), and (13,2$\leq d \leq 3$).

Note that due to the maximum number of pages allowed by the publisher, we omit the figures for the cases (9,*d*), (10,*d*), (12,*d*), and (13,1). The experimental data are available upon request.

We use the following marks in the figures.

(a) Two black triangles down to indicate the region (values of *nsv*) that makes $T_{\mathrm{MVA}} < T_{\mathrm{VA}}$. This region is called the decreasing region, *dr*. The first point in *dr* is denoted by *fdr*.

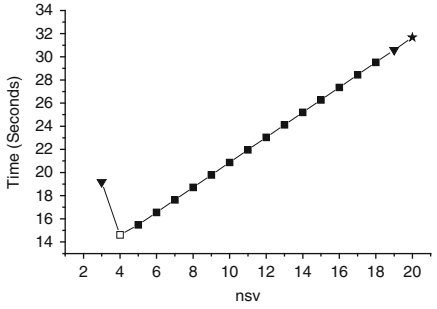


Fig. 8.1 T_{MVA} for (15,3), *nsv* *h*. . .*t*

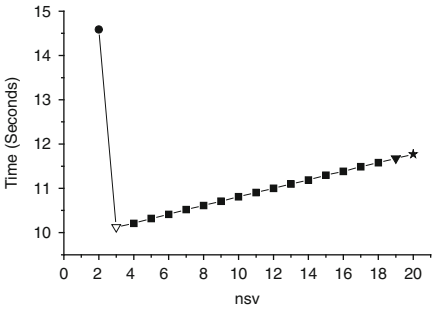


Fig. 8.2 T_{MVA} for (15,2), *nsv h. . .t*

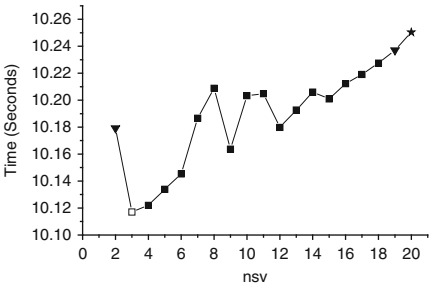


Fig. 8.3 T_{MVA} for (15,1), *nsv h. . .t*

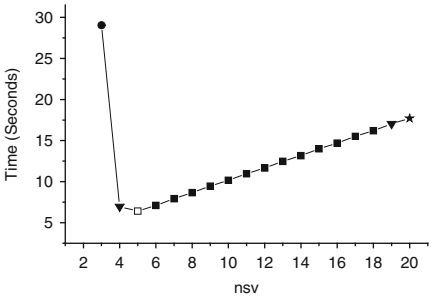


Fig. 8.4 T_{MVA} for (14,3), *nsv h. . .t*

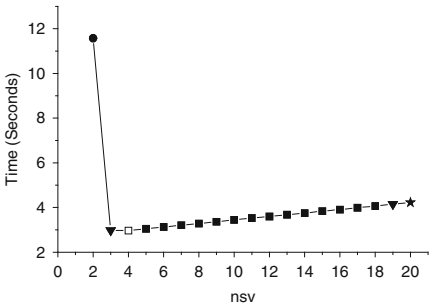


Fig. 8.5 T_{MVA} for (14,2), *nsv h. . .t*

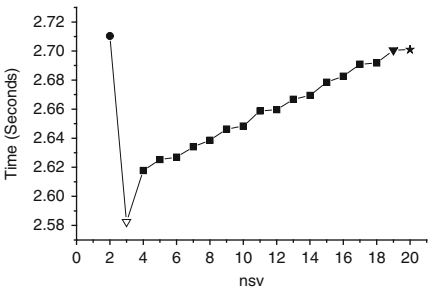


Fig. 8.6 T_{MVA} for (14,1), *nsv* *h*. . .*t*

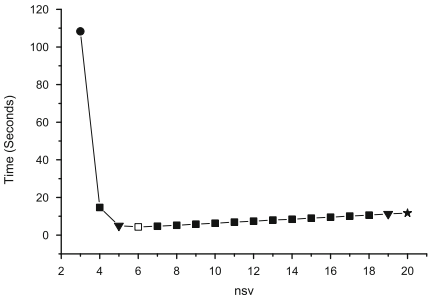


Fig. 8.7 T_{MVA} for (13,3), *nsv* *h*. . .*t*

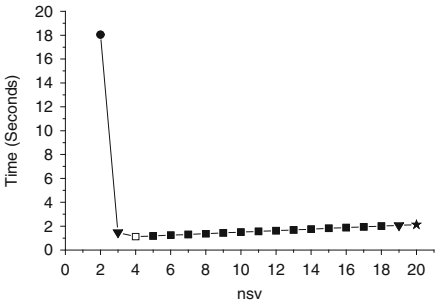


Fig. 8.8 T_{MVA} for (13,2), *nsv* *h*. . .*t*

(b) White box to indicate the value of *nsv* that makes T_{MVA} minimum. This value is called the best number of sequences that is applied by voting, *bnsv*.
(c) Black star to indicate $nsv = t$ (i.e., T_{VA}).
(d) Black circle to indicate $nsv = h$.
(e) White triangles down to indicate that $bnsv = fdr$.

Table 8.2 Values of *dr* and *bnsv* for each (*l*,*d*).

l	*d*	*E*(*l*,*d*)	*h*	*dr*	*bnsv*	*l*	*d*	*E*(*l*,*d*)	*h*	*dr*	*bnsv*
9	1	1.46E 19	2	[5,19]	6	13	3	8.14E 16	3	[5,19]	6
9	2	1.59973	4	[16,19]	17	13	4	5.23252	6	[11,19]	13
10	1	6.19E 30	2	[4,19]	4	14	1	7.83E 73	2	[3,19]	3
10	2	6.11E 08	3	[7,19]	9	14	2	8.45E 47	2	[3,19]	4
11	1	1.57E 40	2	[3,19]	4	14	3	5.05E 25	3	[4,19]	5
11	2	5.43E 17	3	[5,19]	6	14	4	4.20E 07	4	[6,19]	8
11	3	4.72084	4	[13,19]	15	15	1	1.06E 83	2	[2,19]	3
12	1	3.10E 51	2	[3,19]	3	15	2	4.90E 57	2	[3,19]	3
12	2	1.09E 26	3	[4,19]	5	15	3	1.59E 34	3	[3,19]	4
12	3	3.19E 07	4	[7,19]	9	15	4	2.17E 15	4	[5,19]	6
13	1	5.23E 62	2	[3,19]	3	15	5	2.84202	7	[9,19]	12
13	2	1.14E 36	2	[3,19]	4						

From our experimental results (for all instances from (9,*d*) to (15,*d*)), we observe the following:

1. There are many consecutive values of *nsv* that make $T_{\mathrm{MVA}} < T_{\mathrm{VA}}$, see Table 8.2.
2. For each (*l*,*d*), the value of *h* lies outside the *dr* except the instances (15,1) and (15,3) the value of *h* is equal to *fdr*. Also, the last value of *dr* is always $t - 1$.
3. For each (*l*,*d*) and (*l*,*d*′) such that $d < d'$, the following observations are true:
   (a) The value of *bnsv* and *fdr* for (*l*,*d*) is smaller than the value of *bnsv* and *fdr* for (*l*,*d*′), respectively.
   (b) The *dr* for (*l*,*d*) is larger than or equal to the *dr* for (*l*,*d*′).
4. For each (*l*,*d*) and (*l*′,*d*) such that $l < l'$, the following observations are true:
   (a) The value of *bnsv* and *fdr* for (*l*,*d*) is larger than or equal to the value of *bnsv* and *fdr* for (*l*′,*d*), respectively.
   (b) The *dr* for (*l*,*d*) is smaller than or equal to *dr* for (*l*′,*d*).
5. For each (*l*,*d*) and (*l*′,*d*′) such that $E(l',d') < E(l,d)$, the following observations are true for solvable instances:
   (a) The value of *bnsv* and *fdr* for (*l*′,*d*′) is smaller than or equal to the value of *bnsv* and *fdr* for (*l*,*d*), respectively.
   (b) The *dr* for (*l*′,*d*′) is larger than or equal to the *dr* for (*l*,*d*).
6. In the case of the challenging instances (9,2), (11,3), (12,3), (13,4), (14,4), (15,4), and (15,5), the following observations are true:
   (a) For all (*l*,*d*) such that $E(l,d) < 1$, the *dr* are almost nearly equal and the values of *bnsv* are nearly equal.
   (b) For any (*l*,*d*) and (*l*′,*d*′) such that (i) $E(l,d) > 1$, and (ii) $l' > l$. The *dr* for (*l*′,*d*′) is larger than the *dr* for (*l*,*d*).
   (c) For any (*l*,*d*) and (*l*′,*d*′) such that (i) $E(l,d) > 1$, and (ii) $l' > l$. The value of *bnsv* for (*l*′,*d*′) is smaller than the value of *bnsv* for (*l*,*d*).

8.5 Conclusion

In this paper, we study the effect of using modified voting algorithm on DNA PMP. We show that VA is performed better in practice than MVA on most motif instances up to (15,5). We also determined, experimentally, the values of the number of sequences that are needed to make the running time of MVA faster than VA. Finally, we discussed the experimental results and gave some observations according to the relations: (1) l is fixed and d is variable. (2) l is variable and d is fixed. (3) l and d are variables. (4) (l,d) is challenging.

Acknowledgments We are grateful to Henry Leung and Francis Chin for providing us with a complete program for voting algorithm. We also thank M.M. Mohie Eldin for useful discussion.

References

1. M M Abbas, H M Bahig (2009) Performance and analysis of modified voting algorithm for planted motif search. In Proc. of the 7th ACS/IEEE International Conference on Computer Systems and Applications, 725 731.
2. S Balla, J Davila, S Rajasekaran (2006) On the challenging instances of the planted motif problem. Technical Report, Department of Computer Science and Engineering, University of Connecticut, Storrs, CT.
3. J Buhler, M Tompa (2002) Finding motifs using random projections. Journal of Computa tional Biology, 9(2):225 242.
4. A M Carvalho, A T Freitas, A L Oliveira, M F Sagot (2005) A highly scalable algorithm for the extraction of CIS regulatory regions. In Proc. of the 3rd Asia Pacific Bioinformatics Conference, 273 282.
5. F Y Chin, H C Leung (2005) Voting algorithms for discovering long motifs. In Proc. 3rd Asia Pacific Bioinformatics Conference, 261 271.
6. J Davila, S Balla, S Rajasekaran (2006) Space and time efficient algorithms for planted motif search. LNCS, Vol. 3992, 822 829.
7. L Marsan, M F Sagot (2000) Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. Journal of Computational Biology, 7(3 4):345 362.
8. P Pevzner, S H Sze (2000) Combinatorial approaches to finding subtle signals in DNA sequences. In Proc. of the 8th International Conference on Intelligent Systems for Molecular Biology, 269 278.
9. N Pisanti, A M Carvalho, L Marsan, M F Sagot (2006) RISOTTO: Fast extraction of motifs with mismatches. LNCS, Vol. 3887, 757 768.
10. S Rajasekaran, S Balla, C H Huang (2005) Exact algorithms for planted motif problems. Journal of Computational Biology, 12(8), 1117 1128.

Chapter 9
Prediction of Severe Sepsis Using SVM Model

Shu-Li Wang, Fan Wu, and Bo-Hang Wang

Abstract Sepsis is an infectious condition that results in damage to organs. This paper proposes a severe sepsis model based on Support Vector Machine (SVM) for predicting whether a septic patient will become severe sepsis. We chose several clinical physiology of sepsis for identifying the features used for SVM. Based on the model, a medical decision support system is proposed for clinical diagnosis. The results show that the prognosis of a septic patient can be more precisely predicted than ever. We conduct several experiments, whose results demonstrate that the proposed model provides high accuracy and high sensitivity and can be used as a reminding system to provide in-time treatment in ICU.

Keywords Sepsis · Severe sepsis · Support vector machines · SVM

9.1 Introduction

Sepsis, also called "blood stream infection", is a life threatening disease calling for urgent and comprehensive care. Generally, sepsis is caused by the presence of bacteria or other infectious organisms, or by their toxins in the blood or in other tissues of the body. Sepsis may accompany some clinical symptoms of systemic illness, such as fever, chills, malaise (generally feeling "rotten"), low blood pressure, and mental status changes. Severe sepsis, liable to cause deaths, is a more critical situation of sepsis associated with organ dysfunction, hypoperfusion, or hypotension. Sepsis and severe sepsis not only have high mortality, but also become a major burden to healthcare costs. According to the report in [6], there are more than 750,000 cases of severe sepsis occurring annually in the U.S., and the cost for treating patients with severe sepsis is approximately 17 billion dollars each year in the U.S.

F. Wu (✉)
National Chung Cheng University, Chia Yi, Taiwan
e mail: kfwu@mis.ccu.edu.tw

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 9,

Since severe sepsis causes high mortality, the prevention of the septic patients from becoming severe-septic ones has become an important issue. Periodically monitoring the at-risk patients is highly recommended for medical staffs so as that they can diagnose and treat the septic patients as early as possible. In the past, a common method to evaluate the mortality of a septic and severe-septic patient in intensive care unit (ICU) is Acute Physiological and Chronic Health Evaluation (APACHE) II scoring system, in which the higher score always associates with a higher risk of death. Through APACHE II scoring system, medical practitioners can determine the mortality of sepsis patients. Since the severe-septic patients have higher probability of death than the septic patients, a method capable of identifying the severity of septic patients in time is necessary. Though APACHE II scoring system can determine the mortality of sepsis patients, how to identify a patient who is at the risk of becoming severe sepsis in the next few minutes is still not addressed before.

Support Vector Machine (SVM) is known as an innovative algorithm in machine learning with higher accuracy than other methods for data classification [1]. In the past, SVM has been successfully applied to clinical decision problems, such as survival time classification of breast cancer patients [2] and fault diagnostics [3]. Due to the characteristic of SVM, the proposed approach is expected to precisely predict whether septic patients are at the risk of becoming the severe-septic ones if these patients have early diagnosis of severe sepsis.

9.2 Literature Review

Sepsis caused by overwhelming reaction of the patient is a critical disease in the recent years. In 1991, the Society of Critical Care Medicine (SCCM) and the American College of Chest Physicians (ACCP) convened a conference in order to provide a conceptual and practical framework to define the systemic inflammatory response to infection, which will cause sepsis-associated organ dysfunction [4]. The response to infection can be classified into three types in different stages, namely, system inflammatory response syndrome (SIRS), sepsis, and severe sepsis [5]. Septic patient is defined as a patient who has systemic inflammatory response and has evidences to be infected (i.e., SIRS and evidence of infection). For those patients with sepsis, the systemic response to infections may spin out of control. The body's state of balance will become unsettled, damaging one or more vital organs such as the heart, kidneys, or livers. Severe sepsis, which may result from any types of infection, even influenza, is a more critical one that is subset of sepsis with acute organ dysfunction [6].

Since systemic inflammation may lead to severe sepsis, septic shock, or even multiple organ dysfunction [6], early treatment of infection and systemic inflammation are recommended. In the past years, many therapies for septic patients have been proposed: Pittet et al. [7] proposed a method called dynamic analysis of ICU patients to perform a bedside prediction of mortality for bacteremic sepsis. They also believed that a statistic procedure can be applied to the diagnosis of septic

patients by adjusting the above factors [7]. Several measures of the markers have been evaluated, but only a few of them are suitable for clinical use, such as body temperature, protein procalcitonin (PCT) as well as C-reactive protein (CRP), and some cytokines like TNF-*, IL-6, and IL-8 [8]. Among them, IL-6 and IL-8 are very closely related to the severity of the physiological response to infection and systemic inflammation [8].

Recently, Fabiàn Jaimes et al. [9] are the first ones who adopted decision support tools on septic patients. They used the logistic regression and artificial neural networks (ANNs) to predict the mortality in patients with suspected sepsis. The two approaches use clinical variables such as age, immunosuppressive systemic disease, general systemic disease, shock index, temperature, and so on, to predict the death of patients after the admission to the emergency room (ER). However, in the cross-validation procedure, both ANNs and logistic regression takes longer training time than SVM [1].

SVM is a novel method for data classification developed by Vapnik 1998 [10]. SVM has better performance and more precision in classification compared to other approaches such as K-nearest neighbor classifier and neural network [1]. Thus, SVM has been widely used in various areas, such as recognition, reliability evaluation, bioinformatics, and the medical, such as the survival time classification, fault diagnostics [3], and assessing the severity of idiopathic scoliosis from surface topography.

Based on statistical learning theory [10], SVM follows structural risk minimization principle to create a classifier, using a linear hyperplane to maximize the distance between classes. The classification procedure involves training data and testing data, each of which consists of data instances. For each instance of training data set, it can be described with a target value (i.e., class identification) and several attributes (i.e., features). For example, each of patient data might have a class identification called “mild” or “severe” to indicate its severity of outcome and also have several attributes, such as age, gender, and APACHE II scores. SVM aims to find an optimal hyperplane with the metric of a maximal margin called a *separating* hyperplane, to divide all training data belonging to the two classes into two halfspaces. The points lying on the boundaries of the halfspaces are called *support vectors*. Each of the two halfspaces is our target category (class). For example, in this case, they could be considered as mild and severe classes, respectively.

9.3 Materials and Methods

In the data preparation for training and testing of the model, 1,000 historical septic patient records between October 2002 and March 2006 from a medicine center located in the south Taiwan were collected. The records of patients are collected from admission to the ER if their admission diagnoses are confirmed as septicemia. All patient records are screened during the observation period in the ICU to determine whether they fulfilled the inclusion criteria for sepsis. The patients

with diagnoses of sepsis are subsequently screened to evaluate whether the sepsis criteria were indeed fulfilled for the patients and determine whether they reach the exclusion criteria.

The historical patient records used in the model for predicting the severity of a septic patient are composed of feature attributes and the class identification. The feature attributes include age, gender, cause of sepsis (which is classified by ICD-9 code), laboratory variables (i.e., IL-6, CRP), and historical outcomes (i.e., mild or severe, accounting for whether the patients can be moved out of the ICU).

Pathophysiology of sepsis is characterized by the whole body inflammatory reaction and concurrent activation of the host's anti-inflammatory mechanisms. Among these mechanisms, there are several chemical mediators, such as TNF-* and IL-*, are released by the patient's body. These mediators are involved in the processing of damaged tissue. The balance between proinflammatory and anti-inflammatory reactions is directly related to the outcome of the septic patient. Strongly activated phagocytes and high levels of pro-inflammatory cytokines occur in patients who are at risk of developing circulatory shock and multiple organ dysfunction. Extensive anti-inflammatory reaction, which is characterized by the presence of high levels of circulating anti-inflammatory cytokines and impaired innate and adaptive immune functions, renders critically ill patients prone to secondary infections.

Since the identification and control of the infection source are equally important for the prevention of sepsis progression, the cause of sepsis becomes one of the important attributes. We believed that patients could possibly benefit from a therapy aiming at preventing the cause of inflammatory response. Another important attribute is APACHE II scores. According to Pearce et al. [11], APACHE II scores is a quantitative measure that can help machine learning technology to predict the severity of acute disease; Åke Andrén Sandberg et al. also proved that it is a better choice to predict the severe outcome of a patient using a biochemical marker or a scoring system than only based on the description in textbooks [8]. In summary, there are six attributes, which are expected to help the prediction of the outcome of patients with sepsis, under consideration.

Several types of kernel functions exist for different kinds of problems, and each kernel function has its suitableness for specific problems. For example, some well-known problems with large amount of features, such as text classification and DNA problems, are reported to be classified more correctly if the kernel function is a linear function. In the reports of the literature, the RBF kernel function is a decent choice for most problems, since the kernel function usually performs better than other kernels in terms of the generalization ability. Furthermore, the RBF kernel function only uses a pair of parameters (C, γ), where parameter C denotes a penalty parameter in RBF kernel function, and parameter γ denotes the basic kernel parameter of that function. Compared to other kernel functions, the values of parameters C and γ are easier to adjust during the training phase.

For predicting the outcome of a septic patient, a severe sepsis model through the software, Weka [12], is constructed at first. Weka, developed by the University of Waikato in New Zealand, is a widely used tool of machine learning algorithms for data mining and classification tasks. We adopted Weka classifier function, called

SMO, as our classifier. SMO function implements John C. Platt's sequential minimal optimization algorithm for training a support vector classifier with polynomial or RBF kernels. This implementation will replace all missing values and transforms nominal attributes into binary ones. For each attributes as well as the class identification are transformed into numeric value. In addition, the data is first transformed into the standard format of Weka software. That is, all data sets are formatted as ARFF format. This format starts with a declaration of its name, followed by a list of all the attributes in the data sets, and the actual data is attached at last. The following is an example of ARFF data format for the input of Weka:

```
@relation "septic patient outcome prediction"
@attribute Age INTEGER [20 100]
@attribute "Body Temperature (C)" real
@attribute "Heart Rate beats/min" real
@attribute "Respiratory Rate (/min)" real
@attribute "White Blood Cells (K)" real
@attribute "APACHE II scores" real
@attribute Severity {0 1}
@data
72, 36.2, 35, 0, 33.6, 50, 1
66, 34.0, 29, 0, 26.6, 31, 0
64, 37.2, 0, 0, 23.3, 32, 1
66, 35.6, 23, 94, 28.1, 21, 0
40, 34.5, 35, 68, 43.1, 33, 1
```

Before the data is sent to the training process, the input data should be performed a simple scaling on the whole dataset. The primary reason of this procedure is due to the kernel values in a kernel function that often associate with the inner product of attribute values.

As discussed before, the RBF kernel function utilizes two parameters, C and γ, during the training procedure. The subsequent task is to find the optimal values of the two parameters to increase the accuracy of results. For the reason, we performed the cross-validation. In detail, the input data set is divided into n parts of equal size. Then, for every training procedure, $(n - 1)$ parts are chosen as a training set and the remaining one part is as test set. The advantage of cross-validation is that we may get the local optimal parameters in some particular situation, but these parameters may not be the global optimal parameters for all collection.

Traditionally, the performance of a classifier is evaluated by its accuracy. The results of a prediction can be classified into four types, namely, True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). In this paper, the above two values mean that the proposed SVM can correctly predict the results in those cases. On the contrary, the value of FP is the number of FP classified cases in a prediction, and the value of FN is the number of FN classified cases in a prediction.

According to [13], Veropoulos et al. proposed a brand new technique for evaluating the performance of SVM by using receiver-operating characteristic (ROC) curve. ROC curves were initially developed to assess the quality of radar [14]. In

recent years, ROC curves have been widely used as a method to analyze the accuracy of diagnostic tests in medical area. A good prediction method will yield a graph with points in the upper left corner of ROC space, which means 100% sensitivity (i.e., all true positives are found) and 100% specificity (i.e., no false positives are found). A completely random predictor may give a straight line (called *nondiscrimination line*) from bottom left to top right at an angle of 45° from the horizontal. The reason is that when the value of the threshold rises, there will be equal numbers of true positives and false positives. Results below this nondiscrimination line indicate the classifier giving wrong results. Typically, the larger amounts of the Area Under the Curve (AUC) means the higher capability the classifier has. According to the literature, a fair AUC is around 0.5 0.75, a good AUC is around 0.75 0.92, a very good AUC is around 0.92 0.97, and an excellent AUC is around the ratio of 0.97 1.

In the training procedure, 1,000 patient records of septic patients are used to construct our model. Since the abundant amount of data of body temperature, heart rate, respiratory rate, and white blood cells, we used the average values of those data for each patient in first 24 h admission. From the past patient records of the 1,000 septic patients, the statistical data shows that 419 records (about 42%) yield the worse outcomes that turn into severe sepsis, and 581 records (about 58%) remain mild status or yield better outcomes that can be moved out of the ICU.

We also evaluate the performance of our septic model. It is easy to compute that the sensitivity is 0.8744, the specificity is 0.8837, and the accuracy is 0.8841. Figure 9.1 shows the operating characteristic curves of the proposed severe sepsis model as a diagnostic test in elderly septic patients. The AUC in this figure is 0.9421, which shows that the parameters used in the proposed model have very good predictive capability.

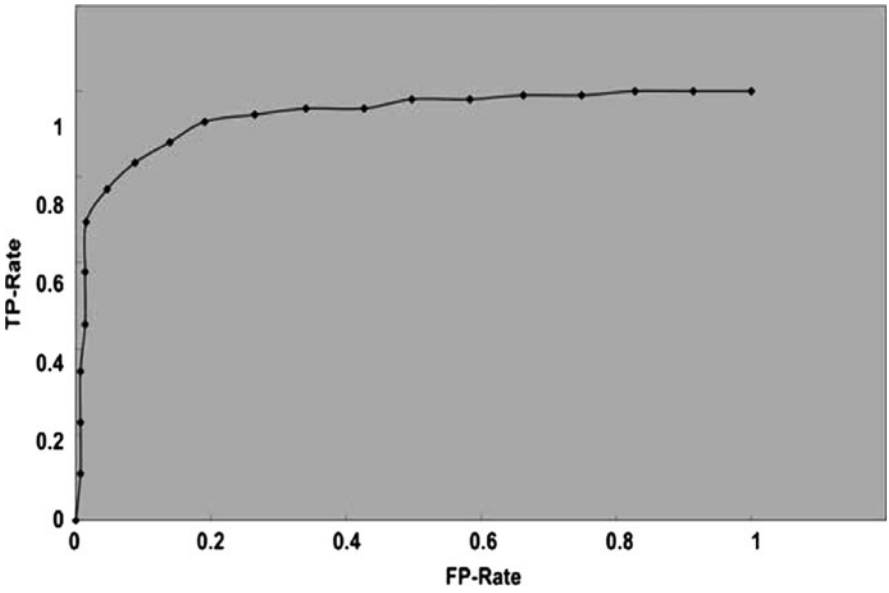


Fig. 9.1 ROC curves of severe sepsis model

9.4 Conclusion

Predicting the outcome of patients with sepsis is always a challenge. Even cytokine concentrations in plasma as well as the APACHE II score cannot be expected to predict patient outcomes accurately. In this paper, we proposed an innovative method based on SVM to monitor as well as predict the outcome of septic patients. This model chooses several practical measurements as the indicators by getting the opinions of medical staffs in ER as well as much literature related to the outcome prediction of septic patients. We use the ROC curve to visualize the performance of the proposed sever sepsis model and evaluate the results. From the results, we show that the chosen clinical predictors can provide significant supports in the severe sepsis model.

References

1. Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, V. Vapnik. International Conference on Artificial Neural Networks, pp. 53 60, 1995, Comparison of learning algorithms for handwritten digit recognition.
2. Y J. Lee, O.L. Mangasarian, W.H. Wolberg. Comput. Optim. Appl. 2003; 25:1 3 ABI/INFORM global, survival time classification of breast cancer patients.
3. S. Pöyhönen, M. Negrea, A. Arkkio, H. Hyötyniemi, H. Koivo. Support vector classification for fault diagnostics of an electrical machine. Proceedings of the 6th International Conference on signal processing, Vol. 2, pp. 1719 1722, 2002, Beijing, China
4. M.M. Levy, M.P. Fink, J.C. Marshall, E. Abraham, D. Angus, D. Cook, J. Cohen, S.M. Opal, J L Vincent, G. Ramsay. For the International Sepsis Definitions Conference Lippincott Williams & Wilkins, 2001, SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference.
5. N.C. Riedemann, R F. Guo, P.A. Ward. SCIENCE IN MEDICINE, J. Clin. Invest. 2003; 112:460 467 The enigma of sepsis.
6. http://www.sepsis.com/.
7. D. Pittet, B. Thievent, R.P. Wenzel, N. Li, R. Auckenthaler. Crit. Care Med. 1996; 153(2): 684 693 Beside prediction of mortality from bacteremic sepsis. A dynamic analysis of ICU patients.
8. Å.A. Sandberg, A. Borgström. JOP. J. Pancreas (Online) 2002; 3(5):116 125 Early prediction of severity in acute pancreatitis. Is this possible?
9. F. Jaimes, J. Farbiaz, D. Alvarez, C. Martínez. Crit. Care 2005; 9:R150 R156 Comparison between logistic regression and neural networks to predict death in patients with suspected sepsis in the emergency room.
10. V.N. Vapnik. 1998; John Wiley, New York, Statistical Learning Theory.
11. C.B. Pearce, S.R. Gunn, A. Ahmed, C.D. Johnson. Pancreatology 2006; 6:123 131 Machine learning can improve prediction of severity in acute pancreatitis using admission values of APACHE II score and C reactive protein.
12. I.H. Witten, E. Frank. 2005; Morgan Kaufmann, San Francisco, Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition.
13. K. Veropoulos, C. Campbell, N. Cristianini. 1999; International Joint Conference on AI, pp. 55 60, Controlling the sensitivity of support vector machines.
14. J.A. Swets, R.M. Pickett. 1982; Academic Press, New York, Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.

Chapter 10
Online Multi-divisive Hierarchical Clustering for On-Body Sensor Data

Ibrahim Musa Ishag Musa, Anour F.A. Dafa-Alla, Gyeong Min Yi, Dong Gyu Lee, Myeong-Chan Cho, Jang-Whan Bae, and Keun Ho Ryu

Abstract Data mining applications over on-body sensor data have earned great attention in recent years. We propose a novel Online Multi-divisive Hierarchical Clustering Method on on-body sensor data. Our method evolves tree-like top down hierarchy cluster, which splits and agglomerates clusters as needed. Experimental results prove a competing quality for our method over existing ones.

Keywords Agglomeration · Cluster accuracy · Hierarchical clustering · On-body sensor

10.1 Introduction

On-body sensors become essential part of online health care systems. Utilizing such technology produces streaming time series data. Mining structural and temporal relationships or hidden dependencies patterns in multiple time series framework is significant domain [1].

The problem of streaming time series clustering has been addressed in many papers [2 4]. In sample clustering, samples are grouped together. Whereas in variable clustering, variables (attributes) are grouped into similar groups so as to reduce intracluster dissimilarities while increasing dissimilarities between different clusters. Hierarchical clustering is superior to other methods because it does not involve the user in specifying the cluster's number such as partitional clustering, and it does not require the whole data to be available at once as BIRCH [5]. And it has a linear time complexity with respect to the size of the input [6]. Variable clustering is useful for applications with high dimensionality like on-body sensor data.

A.F.A. Dafa Alla (✉) and K.H. Ryu (✉)
Database and Bioinformatics Laboratory, Chungbuk National University, Chungbuk, South Korea
e mail: anwarking@dblab.chungbuk.ac.kr; khryu@dblab.chungbuk.ac.kr

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 10,

This paper is organized as follows. The related works, our Online Multi-divisive Hierarchical Clustering Method, the experimental works, and the conclusion are discussed sequentially.

10.2 Proposed Framework

Our proposed framework for online mining streaming time series data consists of three parts. First, a time series data from on-body sensors is continuously preprocessed to allow similarity measure by removing the missing values and transforming categorical data into numerical one. Next, the framework applies the preprocessed data to our clustering method online. It stores the cluster's structure in a database to allow the incoming queries. Finally, the structure is visualized to the end user. Figure 10.1 illustrates our framework design.

Our unsupervised clustering method, which is an enhanced version of the Online Divisive Agglomerative Clustering (ODAC) [3], improves splitting of the cluster since ODAC sometimes ends up giving inaccurate clusters. And even in the semi-fuzzy version of ODAC [7], there is still duplication of clusters due to assigning some variables to two clusters. ODAC was based on monitoring the diameter of the cluster which is the maximum dissimilarity in one cluster and the variable variance in case of clusters with single variable. The Pearson's correlation coefficient between time series is used to measure similarity and dissimilarity using rooted normalized one-minus correlation [3]. It gives 0, 1 values allowing Hoeffding [2], to take statistically significant decisions.

10.2.1 Multiple Splits

For splitting the cluster in addition to the first maximum, we also consider the next and third maximum dissimilarities. In case of dissimilarities sharing at least one point with the pivots (the center points) and being larger than the cluster's radius (a half of the maximum dissimilarity), and when the second pivot's dissimilarity is larger than the radius; we choose its other end point to be the third or fourth pivot. Thus, allowing maximum of four splits at the same time. $d_2 = d(x_3, x_4)$ and $d_3 = d(x_5, x_6)$ are second and third maximum distances, if there is any point shared with the points the maximum dissimilarity, the distance $d^{\sim}$from this variable and the other end of the maximum

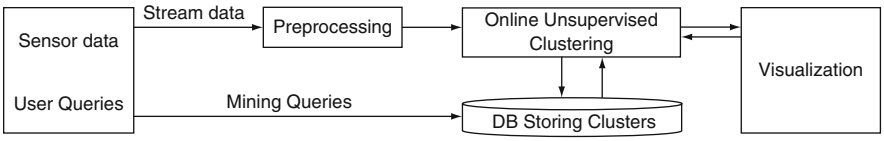


Fig. 10.1 Our proposed framework

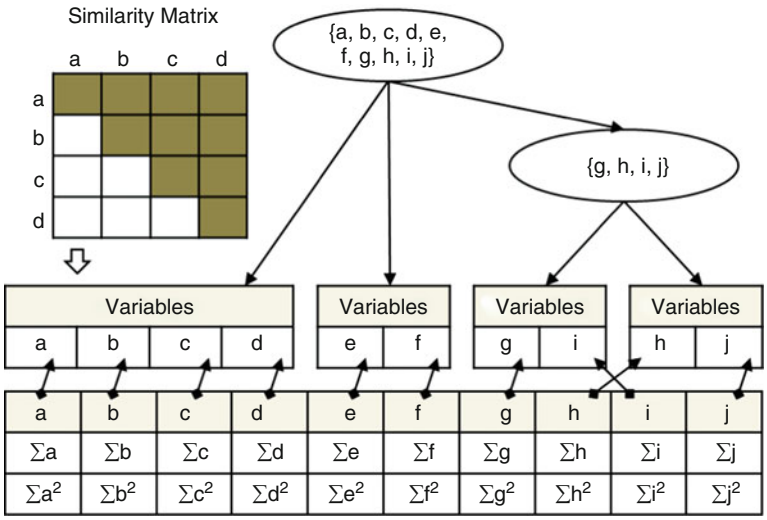


Fig. 10.2 ODAC and our proposed splitting enhancement (E ODAC)

dissimilarity is measured, if it is greater than the half of d_1, then its variables are considered as pivots, and hence end up having three or four actual clusters. For instance, let $d_1 = d(x_1, x_2)$ and $d_2 = d(x_2, x_3)$, since variable x_2 is shared with the first and second maximum dissimilarities, distance $d^{\sim} = d(x_1, x_3)$ is found and test the condition $d^{\sim} \geq d_1 \Rightarrow x_3$ the x_3 is a third cluster's center it split the cluster into three clusters with x_1, x_2, x_3 as cluster centers.

10.2.2 Agglomeration

Our algorithm splits streaming data, further detects the wrong splitting so as to be re-aggregating in its parent as shown in Fig. 10.2. Using time frame to read, the time series stream is calculated to summarize statistics at leaf. It checks for splits or aggregation when detecting to change new structure is announced feeding to the leaf nodes. The maximum number that can be generated after splitting one cluster is four clusters which enhance memory space and processing time since samples are independent.

10.3 Experimental Evaluations

In our experimental studies, we choose real training datasets from Physiological Data Modelling Contest (2004) [8]. It consists of 10,000 h; with several variables. We selected nine attributes with accelerometer, heat flux, Galvanic skin response, skin temperature, and near-body temperature. The accelerometer is 2-axis MEMS

device that measures forces exerted on body and the gravity information. The proprietary heat flux sensor is a robust and reliable device that measures the amount of heat being dissipated by the body. Galvanic Skin Response represents electrical conductivity between two points on the user's arm. Skin temperature is the body's core temperature. The near-body temperature sensor measures the air temperature immediately around the user's armband. Further constructed datasets as done in [3], by dividing a dataset based on userID. In the following two sections, we explain our studies on the cluster accuracy specification and the execution time measurement, where E-ODAC proves better qualities than ODAC.

10.3.1 Cluster Accuracy Specification

Internal criteria cophenetic correlation coefficient (CPCC) [9] is used with relative criteria Dunn's validity index [9, 10] for cluster accuracy specification. CPCC is given by the following (10.1), where Q is the cophenetic proximity matrix, q_{ij} is proximity level, P is proximity matrix, and $M = n(n-1)/2$ as shown by (10.2).

$$\text{CPCC} = \frac{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} p_{ij}\, q_{ij} - \mu P \mu Q}{\sqrt{\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} p_{ij}^2 - \mu P^2\right)\left(\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} q_{ij}^2 - \mu Q^2\right)}} \tag{10.1}$$

$$\mu P = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} p_{ij} => \mu Q = \frac{1}{M}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} q_{ij} \tag{10.2}$$

Whereas Dunn's validity index (DVI) is defined as

$$\text{DVI} = min\left\{\frac{d(c_i, c_j)}{max_k\,\{diam(c_k)\}}\right\} \tag{10.3}$$

The elements connectivity as shown in Table 10.1 proves the accuracy of E-ODAC, where Usr 1, the final number of clusters gained by applying E-ODAC are four, compared to three clusters gained by ODAC. However, it shows higher CPCC value over ODAC indicating more compact clusters [9]. With Usr 6 both methods show the

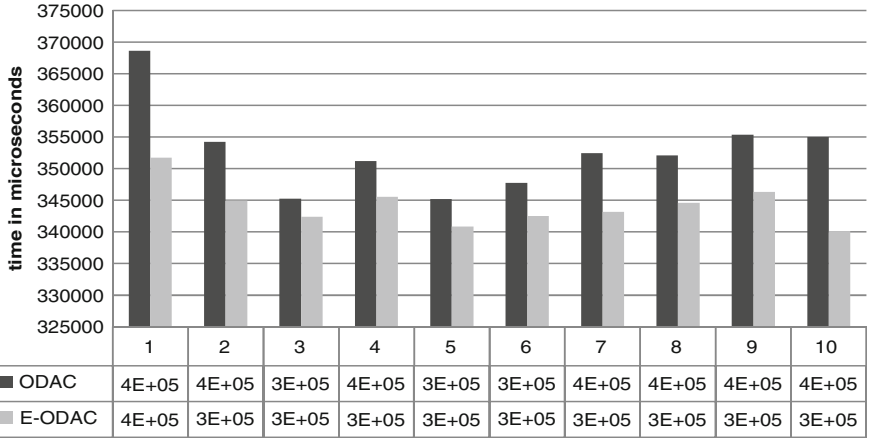
Table 10.1 CPCC and DVI index for ODAC and E ODAC

Dataset	CPCC index		DVI index	
	ODAC	E ODAC	ODAC	E ODAC
Usr 1	0.417481	**0.518265**	1.00858	**1.02321**
Usr 6	0.604084	0.604084	0.91373	0.91373
Usr 25	0.48397	**0.594837**	0.95215	**0.95341**

same value, whereas with Usr 25 E-ODAC gives rather more compact value. To insure the benefits of our method, the values of DVI are listed in Table 10.1. Again E-ODAC appears to outperform ODAC in case of Usr 1 and Usr 25.

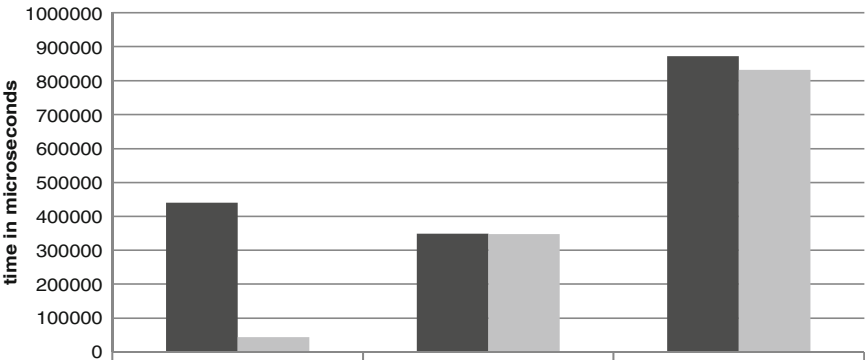
10.3.2 Execution Time Measurement

Figure 10.3 illustrates the sequence of ten execution times for usrID6 in both algorithms while Fig. 10.4 shows the average execution time for datasets of usrID1, usrID6, and usrID25.



	1	2	3	4	5	6	7	8	9	10
ODAC	4E+05	4E+05	3E+05	4E+05	3E+05	3E+05	4E+05	4E+05	4E+05	4E+05
E-ODAC	4E+05	3E+05	3E+05	3E+05	3E+05	3E+05	3E+05	3E+05	3E+05	3E+05

Fig. 10.3 Ten execution times of usrID6



	UsrID_1	UsrID_6	UsrID_25
ODAC	439665.7	349396.5	871765.3
E-ODAC	43810.6	347547.2	832418.9

Fig. 10.4 Average execution time

10.4 Conclusion

Our novel framework for hierarchically clustering on-body sensor data consists of three parts; preprocessing to remove outliers and resolve missing data (E-ODAC), a database back end for storing the resulting cluster structure, and finally data visualization. Experimental results showed that E-ODAC outperforms ODAC in terms of cluster evaluation indexes and time complexity.

Acknowledgment This research was supported by the grant of the Korean Ministry of Education, Science and Technology (The Regional Core Research Program/Chungbuk BIT Research Ori ented University Consortium): the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF No. 2010 0001732): the research grant of the Chungbuk National University in 2008.

References

1. Gaurav N. P., and Balakrishnan P., "Storage, Retrieval, and Communication of Body Sensor Network Data", *Proceeding of the 16th ACM international conference on Multimedia*, pp. 1161 1162, 2008.
2. Jiawei H., and Micheline K., "*Data Mining Concepts and Techniques. 2nd Edition*", Morgan Kaufmann: San Francisco, CA, pp. 482 496, 2006.
3. Pedro P. R., Joao G., and Joao P. P., "ODAC: Hierarchical Clustering of Time Series Data Streams", *IEEE Transactions on Knowledge and Data Engineering*, 20, pp. 615 627, 2008.
4. Denton A., "Kernel Density Based Clustering of Time Series Subsequences Using a Contin uous Random Walk Noise Model", In *Proceedings of the Fifth IEEE International Confer ence on Data Mining*, IEEE Computer Society: Washington, DC, pp. 122 129, 2005.
5. Tian Z., Raghu R., and Miron L., "BIRCH: An Efficient Data Clustering Method for Very Large Databases", *ACM SIGMOD Record*, pp. 103 114, 1996.
6. Abdun N. M., Christopher L., and Parampalli U., "An Efficient Clustering Scheme to Exploit Hierarchical Data in Network Traffic Analysis", *IEEE Transactions on Knowledge and Data Engineering*, 20(6), pp. 752 767, 2008.
7. Pedro P. R., and Joao G., "Semi fuzzy Splitting in Online Divisive Agglomerative Cluster ing", *Lecture Notes in Computer Science*, 4874, pp. 133 144, 2007.
8. http://www.cs.utexas.edu/users/sherstov/pdmc/.
9. Rui X. and Donald C. W., "Clustering", *IEEE Press Series on Computational Intelligence*, pp. 263 278, 2009.
10. Dunn J. C., "Well Separated Clusters and Optimal Fuzzy Partitions," *An International Journal of Cybernetics and Systems*, 4(1), pp. 95 104, 1974.

Chapter 11
On Quality Assurance and Assessment of Biological Datasets and Related Statistics

Maria Vardaki and Haralambos Papageorgiou

Abstract The complexity of modern biological database management systems indicates the need of integrated metadata repositories for harmonized and high-quality assured data processing. Such systems should allow for the derivation of specific producer-oriented indicators monitoring the quality of the final datasets and statistics provided to the end-users. In this paper, we offer a quality assurance and assessment framework for biological dataset management from both the producers' and users' perspective. In order to assist the producers in high-quality end-results, we consider the integration of a process-oriented data/metadata model enriched with quality declaration metadata, like quality indicators, for the entire process of dataset management. With the automatic manipulation of both data and "quality" metadata, we assure standardization of processes and error detection and reduction. Regarding the user assessment of final results, we discuss trade-offs among certain quality components (such as accuracy, timeliness, relevance, comparability, etc.) and offer indicative user-oriented quality indicators.

Keywords Biological datasets · Data processing · Database management · Quality assurance · Statistics

11.1 Introduction

Biological databases are an important tool in assisting scientists to understand and explain biological phenomena and contain, among others, huge amounts of datasets collected by diversified sources and processed using various metadata-based information systems. Assessing the quality of the data provided and mainly the corresponding statistics produced from datasets comparisons has become

H. Papageorgiou (✉)
Department of Mathematics, University of Athens, Panepistemiopolis, Athens 17584, Greece
e mail: hpapageo@math.uoa.gr

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 11,

increasingly important to clinical investigators, policy-makers, medical research centers, data analysts and to patients themselves. Since data services in most cases lack access to supporting data, data quality should remain the main responsibility of the data provider.

In order to achieve this, data providers should (1) embody harmonization and transformation procedures in their workflow processing to minimize human errors and, therefore, increase the quality of biological information and statistics provided and (2) allow for the derivation of specific producer-oriented indicators which will monitor the quality of the final datasets and statistics they provide to the end-users.

In this paper, we concentrate on the quality of biological datasets analysis and statistics produced by (1) presenting a producer-oriented quality assurance approach for the management of biological datasets and (2) discussing user-oriented quality assessment criteria and related monitoring and performance indicators emerging from the guidelines and perspectives of International Organizations and National approaches.

For the producer-oriented quality assurance, we extend the analysis and dissemination part of a previously introduced data/metadata model [9] developed for the automation of the series of processes of an experiment with the integration of specific indicators and other quality declaration metadata. We stress the importance of automatically manipulating both data and metadata with the use of a highly structured statistical model and the incorporation of transformations, measuring output quality.

For the user-oriented quality assessment, we not only consider a number of generally accepted criteria, like relevance, timeliness and punctuality, accessibility, comparability, etc., but we also indicate the trade-offs between them. We selected to provide a comparative description of both producers' and users' quality requirements in Sect. 11.2 of the paper, thus clarifying our two approaches introduced in Sects. 11.3 and 11.4, respectively.

11.2 Quality Assessment Criteria and Trade-Offs

Quality in Statistics is defined by National and International Organizations with reference to a number of usually overlapping dimensions (criteria) and indicators for the assessment of any survey.

Initially, we should clarify that "quality dimensions (or criteria)" and "quality indicators" are considered by National and International Organizations from two different perceptions: They are used, like for example in OECD's "Health Care Quality Indicators" project [1], to measure Health care performance and their aim is to maintain, restore, or improve health. Examples of defining clinical indicators for quality improvement are given by Mainz [6]. In the other case, quality indicators may be applied a posteriori, at the end of a process (or a stage of it) or on statistics to monitor the quality of the already produced results.

Regarding the former point of view, Legido-Quigley et al. [4] have provided detailed information on health-care quality strategies in EU Member States.

However, in this paper, we focus on the latter case, considering also related work already published by International Organizations, like for instance, the OECD quality framework for statistics [7], the IMF Data Quality Assessment Framework [3], etc. Eurostat has also defined a list of "Standard Quality Indicators" [5] as well as a list of Quality Performance Indicators [2]. Since producers are also users of the biological datasets, we will focus on the following commonly required quality assessment criteria:

- *Relevance*: It reflects the degree to which the output datasets/statistics/product meet current and potential user needs. It highly depends on the requirements and views of different user categories concerning the other quality components.
- *Accuracy*: In the general statistical sense, it denotes the closeness of computations or estimates to the exact or true values. The difference between the two values is the error.
- *Timeliness*: It reflects the length of time between its availability and the event or phenomenon it describes.
- *Punctuality*: It refers to the time lag between the release date of data and the target date when it should have been delivered.
- *Accessibility*: It refers to the physical conditions in which users can obtain data, i.e., how to order, delivery time, and available formats (paper, CDROM, etc.).
- *Comparability*: It is the ability of allowing for comparisons of data quality over time and from different sources.
- Apart from these criteria, usually a *Cost/Benefit* analysis should also be considered.
- Although there are common requirements between producers and users of biological information, when developing quality indicators, we have to take into account specific trade-offs, some of which are briefly described as follows:
- The data producers are interested not only in the quality of final output results but also in the quality of a number of intermediate steps of the process (data input, editing and imputation, aggregation, weighting, etc.).
- Users of biological datasets or statistics like researchers, practitioners, government, etc., are in general not in a position to assure for themselves the overall quality of output results. For example, while the relevance, timeliness, and accessibility of data may be immediately apparent to a user, other dimensions of quality, especially accuracy, and comparability (mainly in case of merged datasets) cannot be deduced from inspection of the product alone.
- Time spent developing an experimental study can contribute substantially to lack of timeliness. On the other hand, time spent to define patients' response to a therapeutic model is vital.
- At study design, the desired timeliness of information is related to relevance (for what period does the information remain useful for its main purposes?).
- For retrospective surveys, when combining biological datasets from various time periods, timeliness becomes less important relative to accuracy and comparability.

- From users' perspective, we can state that (1) without relevance, the other dimensions are unimportant; (2) given relevance, without timeliness and accessibility, the data are not available when they are needed; and hence, (3) only when relevance, timeliness, and accessibility are satisfied, accuracy becomes important. Of course, a similar remark holds for comparability.

11.3 Quality Assurance: Producer-Oriented Approach

There are various information sources devoted to the diffusion of experimental and other biological data, and currently medical organizations and research projects are making efforts to integrate metadata in their systems and develop common schemas to both increase information exchange and make this information publicly available. In the form of classifications and codes, metadata define demographics, diagnoses, and care of patients for the purposes of documentation, communication, transaction, and monitoring.

In general, metadata are defined as "data about data". This term includes, for example, information about the conducted survey (sampling units and method used, the population studied, and the eligibility criteria), gene or protein characteristics, etc. The semantic power of metadata can greatly increase the usage of data by providing end-users with additional semantics necessary to reconstruct the context of data stored [8]. The simplest benefit of a metadata-enabled statistical software system for a user who analyzes tabulated data is that it can warn him of a possible error when, for example, he attempts to merge patient data with different eligibility criteria or conflicting risk factors.

Furthermore, any user of the biological information would benefit if he could trace back quickly the processing history of specific datasets during the management of various experiments. A way of achieving this goal is by using highly structured statistical models that consist of logical structures and operators. Also, we should define and integrate several transformations for the manipulation of data. The structures specify which data and metadata items we will capture; the operators permit the execution of processing of metadata items included in the model; and the transformations ensure the validity and automation of data/metadata manipulations.

Vardaki et al. [9] have introduced such a statistical, process-oriented metadata model to describe the entire medical research process, enabling a more active role for any associated metadata. The proposed model, although process-oriented and broad enough, can also be expanded to embrace metadata items of forthcoming challenges due to the Object-Oriented technique followed, which allows further extensions of the already integrated model without database redesign. The interesting point is that, mainly due to this extension possibility, we can propose and incorporate, as metadata, specific producer-oriented quality indicators that would automatically indicate the quality of the datasets provided (from one experiment or combined datasets).

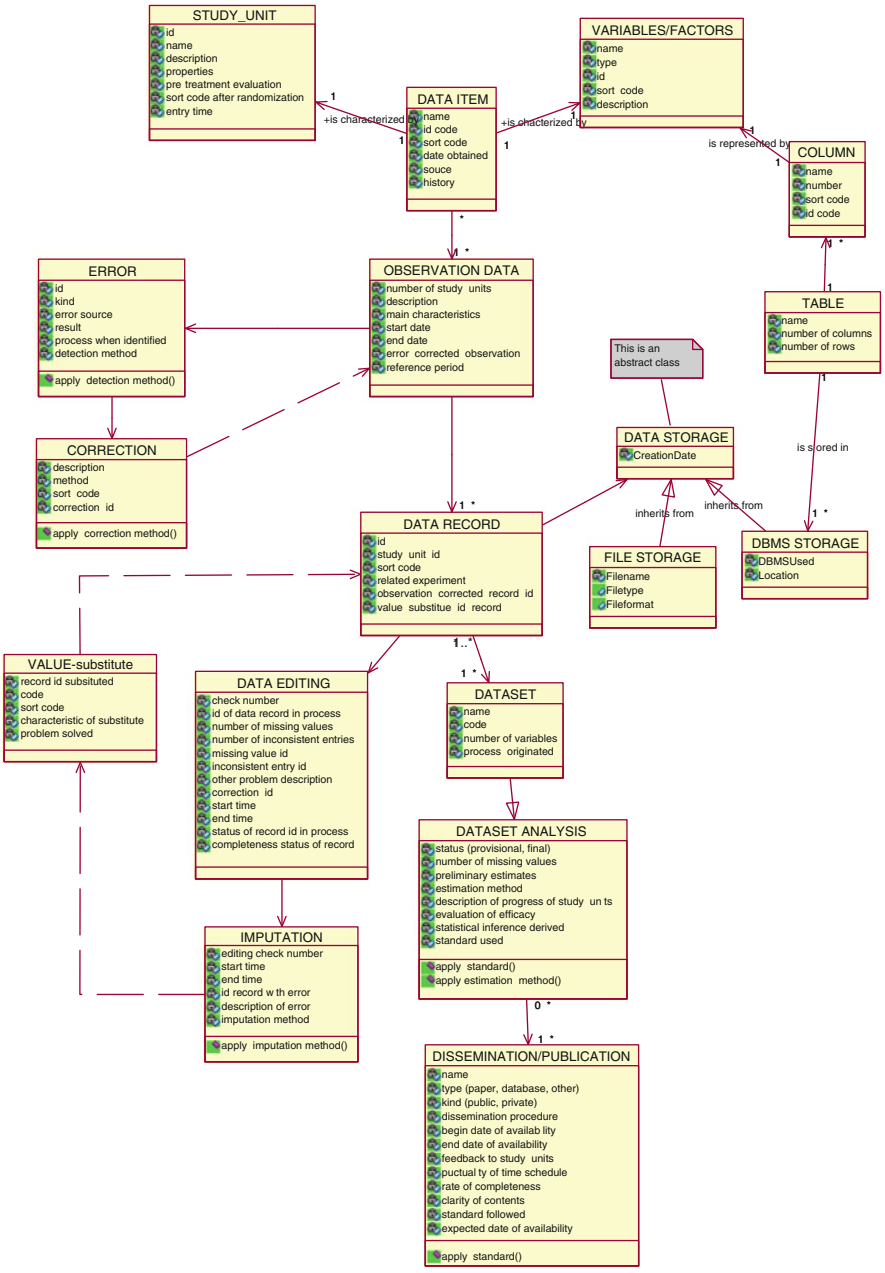


Fig. 11.1 Quality assurance metadata model

We use the main considerations of the previously designed UML metadata model and extend it with quality assurance metadata in the analysis and publication part, as illustrated in Fig. 11.1. During the data collection stage, a number of errors like coverage, sampling, and non-response errors may occur and can, where possible, be automatically corrected as illustrated in [9]. Regarding observation errors (measurement errors detected during editing phase processing errors, imputation errors, etc.), quality assurance of the sequence of processes discussed below may minimize them.

Suppose that a set of Variables/Factors documents the characteristics of one or more units of a study (Study units) and each Data Item is a single entry characterizing a pair (Study unit id, Variable id). Sets of such items represent the Observation Data, where observation errors (Error) may be detected (detection method). According to a list of errors and correction methods a priori integrated into the system, the required Correction method is applied, where possible. Then, the corrected observation is ready to replace the erroneous one. The observation data are translated into Data Records in the database. Data Editing is the application of checks that identify missing, invalid, or inconsistent entries which deteriorate quality and does not allow for meaningful automatic comparisons. Records that fail the data entry checks will, therefore, be rejected in whole or in part depending on the conventions being used. In case an Imputation stage is permitted, the corrected data record (correction id) replaces during this phase the erroneous one by the substitution of a Record id. This can be realized by executing the operation "Apply imputation method", and the Value-Substitute produced replaces the erroneous Data Record. Sets of data records form a Dataset, which is practically represented in tabular form in the database, with columns corresponding to the variables and rows describing the characteristics measured for each unit of study. After Data Storage in a DataBase, the Analysis of the datasets takes place. Information such as missing data, preliminary estimates, the applied estimation methods, the progress of the experiment, etc., is essential for the final results, evaluation.

The Dissemination/Publication can be either in paper format or in the form of a database like the GenBank® (http://www.ncbi.nlm.nih.gov/Genbank/). Therefore, the model not only considers the requirements for the Standards followed during each data management process, but also offers derivation of quality indicators automatically calculated by the system, like for instance, the following measuring accuracy, timeliness, and completeness:

- *Editing rate* = (number of observations edited)/(total number of observations).
- *Imputation rate* = (number of observations imputed)/(total number of observations), both of which refer to a single variable of interest.
- *Editing ratio* = (total number of edited values)/(total number of final values).
- *Unweighted imputation ratio* = (total number of imputed values)/(total number of all final values), which refer to all variables checked during the editing and imputation processes.
- *Punctuality of time-schedule of publication* which is calculated as the difference (in time units) between the actual and the scheduled date of publication.

Furthermore, data producers can create an overall accuracy quality indicator like the *Total Survey Error* (for a key variable).

11.4 Quality Assessment Monitoring: User-Oriented Requirements

In this section, we will propose specific user-oriented indicators for the assessment of data analysis results according to the main quality criteria already discussed in Sect. 11.2, which play an important role in user satisfaction.

11.5 Relevance

If we consider each user category as a special case, then the aggregation performed to produce relevance indicators should be made according to either the variable "user category" or the variable "user need" but not both at the same time. We may consider the indicator:

Rate of Available Products (Datasets, Statistics, etc.)

which is generally calculated by *dividing the number of products provided by the* "*producer*" *by the number of products requested by the* "*user*".

This indicator may also be considered as a measure for completeness.

11.5.1 Accuracy

Since a number of errors can be corrected during the analysis and dissemination phase, the main quality indicator that can be broadly understandable by every user category is the

Rate of difference between initial and corrected (final) results.

This percentage can be calculated for each key variable of an experiment, and it monitors the corrective actions performed in order to publish the final results.

11.5.2 Timeliness and Punctuality

We can use the indicator:

Rate of difference between the date of the announced release of datasets/statistics and the date of their actual release.

11.5.3 Accessibility

An indicator monitoring the available databases where a user may retrieve the datasets or key statistics of interest can be of value, like the following:

Number of databases used for disseminating specific product.

11.6 Suggestions for Further Research

The presented framework on quality assurance and assessment is not limited to health statistics or biological databases but can be applied in nearly all socio-economic areas.

It appears that further research on the topic can be threefold: (1) integrate a consistent quality framework defining specific quality acceptance levels for each process stage; (2) develop for biological statistics the so-called quality reports [2] to promote harmonized quality reporting across statistical processes and their outputs, and hence to facilitate comparisons; and (3) develop an overall "User satisfaction index" defined as the degree of satisfaction with services and products for different user categories. This indicator will be based on satisfaction surveys among user categories, taking also into account sections of the previously mentioned quality reports.

Acknowledgment This research was partially funded by the University of Athens, Special Account for Research Grants, Grant no. 70/4/8758.

References

1. Arah O.A., Westert G.P., and Hurst J. et al. (2006). Conceptual framework for the OECD Health Care Quality Indicators project. International Journal for Quality in Health Care, Vol. 18, pp. 5 13.
2. Eurostat (2009). Quality in Statistics ESS quality and performance indicators 2009 progress report, accessed in CIRCA database, members only library.
3. IMF (2003). Data Quality Assessment Framework. Internet: http://dsbb.imf.org/vgn/images/pdfs/dqrs factsheet.pdf , accessed October 12th, 2009.
4. Legido Quigley H., McKee M., Nolte E., and Glinos I.A. (2008). Assuring the Quality of Health Care in the European Union: A Case for Action, World Health Organization: Copenha gen, ISBN 9289071931.
5. Linden H., and Papageorgiou H. (2004). Standard Quality Indicators. European Conference on Quality and Methodology in Official Statistics: Mainz, Germany. Also published in Statistical Research Reference Material, Japan Statistical Research Institute, Hosei University, No 93, pp. 61 86. Tokyo, Japan. ISSN 0288 8734, 2006.
6. Mainz J. (2003). Defining and classifying clinical indicators for quality improvement. Interna tional Journal for Quality in Health Care, Vol. 15, pp. 523 530.

7. OECD (2003). Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities. Internet: http://www.oecd.org/dataoecd/26/38/21687665.pdf, accessed October 12th, 2009.
8. Vardaki M., and Papageorgiou H. (2007). Statistical data and metadata quality assessment, in Handbook of Research on Public Information Technology, eds. G.D. Garson and M. Khosrow Pour, Vol. 2, IGI Global, USA, pp. 604 614.
9. Vardaki M., Papageorgiou H., and Pentaris F. (2009). A statistical metadata model for clini cal trials' data management. Computer Methods and Programs in Biomedicine, Vol. 95, pp. 129 145.

Chapter 12
Pattern Recognition-Informed Feedback for Nanopore Detector Cheminformatics

A. Murat Eren, Iftekhar Amin, Amanda Alba, Eric Morales, Alexander Stoyanov, and Stephen Winters-Hilt

Abstract Pattern recognition-informed (PRI) feedback using channel current cheminformatics (CCC) software is shown to be possible in "real-time" experimental efforts. The accuracy of the PRI classification is shown to inherit the high accuracy of our offline classifier: 99.9% accuracy in distinguishing between terminal base pairs of two DNA hairpins. The pattern recognition software consists of hidden Markov model (HMM) feature extraction software, and support vector machine (SVM) classification/clustering software that is optimized for data acquired on a nanopore channel detection system. For general nanopore detection, the distributed HMM and SVM processing used here provides a processing speedup that allows pattern recognition to complete within the time frame of the signal acquisition where the sampling is halted if the blockade signal is identified as not in the desired subset of events (or once recognized as nondiagnostic in general). We demonstrate that Nanopore Detection with PRI offers significant advantage when applied to data acquisition on antibody-antigen system, or other complex biomolecular mixtures, due to the reduction in wasted observation time on eventually rejected "junk" (nondiagnostic) signals.

Keywords Channel current · Cheminformatics · Pattern recognition · Real-time · Base-level DNA classification · Support Vector Machine (SVM) · Hidden Markov Model (HMM)

12.1 Introduction

Pattern recognition-informed (PRI)sampling and experimental feedback opens the door to a whole new realm of signal stabilization and device capabilities. In this chapter, we describe results from recent PRI sampling experiments on the nanopore

S. Winters Hilt (✉)
Department of Computer Science, University of New Orleans, New Orleans, LA 70148, USA
Research Institute for Children, Children's Hospital, New Orleans, LA 70118, USA
e mail: winters@cs.uno.edu; swinters@chnola esearch.org

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 12,

detector platform [1 3]. The signal analysis involves the use of Machine Learning (ML) algorithms such as Hidden Markov Models (HMMs) [2, 4 6] and support vector machines (SVMs) [7 10]. The machine learning ML algorithms are amenable to distributed computational solutions (for the HMMs, in particular [11]), which permits a computational speed-up that ensures real-time operational feedback on the nanopore detector applications studied. Although a specific application is examined here in detail, a similar approach can be used in many experimental efforts. The results and application of the PRI method, with distributed HMM and SVM algorithms are generally applicable in knowledge discovery contexts involving stochastic sequential analysis and/or classification/clustering. Some background material on the nanopore detector and the channel current cheminformatics (CCC) architecture are given as follows.

12.1.1 Nanopore Detector

A nanometer-scale channel can be used to associate ionic current measurements with single-molecule channel blockades. The α-hemolysin channel, used in the experiments described here, self-assembles such that a single channel of α-hemolysin can be isolated in lipid bilayer. This provides an inexpensive and highly reproducible method to construct a nanopore-based detector that is informed by single molecule interactions with a nanometer scale channel. See [1, 2] for further details. An example of a direct interaction/modulation of the channel ion current is shown in [3]. Indirect channel modulation or transduction is described in [12] and is being used for generalized binding analysis at the single interaction-complex level. In [1, 2], PRI sampling is done with mixtures involving five DNA hairpin molecules (see Sect. 12.2 for sequence details).

12.1.2 Channel Current Cheminformatics Architecture

Preliminary work to establish real-time control of a nanopore detector has been described in [6]. The work is based on live, streaming, measurements, and fast pattern recognition identification of blockading (“captured”) analytes. Real-time *sampling* control of a nanopore detector, alone, has been proposed to boost nanopore detector sampling productivity by orders of magnitude, depending on the mix of desirable signal classes vs. undesirable in the data being analyzed. In a real-time setting the challenge is to perform the HMM-based feature extraction sufficiently quickly (whereas the SVM is trained offline, and so operates very quickly online).

The real-time experimental linkage between the DAQ and the computational facilities is implemented using LabWindows development environment that connects via TCP/IP to our ML nodes that run our “in-house” CCC methods (see Sect. 12.2). Data acquired with LabWindows is passed to the CCC software server

on a streaming real-time basis. The classification results are then quickly returned to the LabWindows automation software for experimental feedback control. As suggested in [6], the real-time classification inherits the 99.9% accuracy of the non real-time implementation (established in prior work [2]) as nothing has changed in regards to the features extracted and the classifier used. Thus, the full power of HMM and SVM methodologies can be leveraged into numerous "real-time" experimental protocols that would employ PRI methods.

12.2 Methods

12.3 Nanopore Detector

The experimental setup is described in detail in [1, 2]. Each experiment is conducted using one α-hemolysin channel inserted into a diphytanoyl phosphatidylcholine/hexadecane bilayer across a 25-μ-diameter horizontal Teflon aperture, as described previously [1, 2]. Seventy microliter chambers on either side of the bilayer contain 1.0 M KCl buffered at pH 8.0 (10 mM HEPES/KOH) except in the case of buffer experiments where the salt concentration, pH, or identity may be varied. Voltage is applied across the bilayer between Ag and AgCl electrodes. DNA control probes are added to the *cis* chamber at 10 or 20 μM final concentration. All experiments are maintained at room temperature (23 ± 0.1°C), using a Peltier device.

12.4 Test Molecules

The eight base-pair and nine base-pair hairpin molecules used in this study were previously studied in [1, 2, 13]. The sequences of these hairpins are: 9GC [5′-GTTCGAACG TTTT CGTTCGAAC-3′], 9TA [5′-TTTCGAACGTTTT CGTTCG AAA-3′], 8GC [5′-GTCGAACG TTTT CGTTCGAC-3′], 7CG [5′-CTGAACG TTTT CGTTCAG-3′] and 6GC [5′-CGAACG TTTT CGTTCG-3′]. The eight base-pair DNA hairpin is identical to the core nine base-pair subsequence, except the terminal base-pair, which is 5′-G•C-3′.

12.4.1 Channel Current Cheminformatics

A capture signal generated with the nanopore apparatus is filtered and amplified before it is sent through the DAQ. The data acquisition device converts the analog signal to digital format for use in the display and recording of data in binary Axon (Molecular Devices) format. In the pattern recognition feedback loop, the first 200 ms detected after drop from baseline are sent via TCP-IP protocol to the

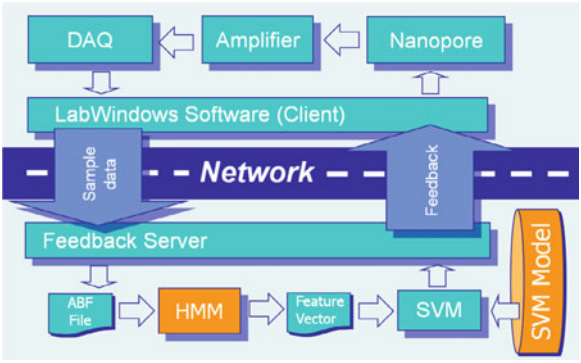


Fig. 12.1 Labwindows/feedback server architecture with distributed CCC processing. The HMM learning (on line) and SVM learning (off line) are network distributed processes for N fold speed up, where N is the number of computational threads in your cluster network

HMM software, which generates a profile for each signal sent. The HMM-generated profile is processed with the SVM classifier to compare the real-time signal with previous training data in order to determine whether the signal is acceptable according to the experimenter's choice of model (see Fig. 12.1). If the signal is acceptable, the message to continue recording is sent to the LabWindows software. If not, a message is sent to LabWindows to eject the molecule, and the amplifier briefly reverses the polarity to eject the molecule from the channel.

For the successful real-time feedback experiments described in Sect. 12.5, only two computers, a client, and a server were needed. In general, the server consists of a cluster of computers to distribute the HMM, and possibly SVM, processes. The Client runs Microsoft Windows XP to visualize and record the entire experiment by using LabWindows. Our in-house implementation of LabWindows acquisition software is able to detect blockades using a tFSA while also recording and visualizing the experiment. Our implementation for channel current analysis also has the critical functionality to change the polarity of current so as to eject any molecules from pore when necessary. The Server computer runs Pardus Linux 2007.3. The hardware for both the Client and Server consists of PCs with 2.4 GHz AMD CPUs, with 2 GB memory.

12.4.2 HMM Feature Extraction and SVM Classification

The HMM implements 50 states as determined by making 50 bins of the blockade current data. The quantized data goes through one round of Expectation-Maximization to obtain transition probabilities after running the Viterbi algorithm to obtain the most probable path of states that created the signal (see [2] for details).

12.4.3 SVM Classification

The online discriminatory speed of a trained SVM is simply that of evaluating an inner product, and so its operational constraint on the PRI feedback endeavor is negligible compared to that of the HMM feature extraction stage. For this reason, there is little discussion of SVMs in this paper, even though SVMs comprise much of the complexity of the HMM/SVM PRI feedback system.

12.5 Results

The nanopore experiments with PRI sampling are first done for PRI binary sampling with a 1:70 mixture of 9GC:9TA. Figures 12.2 and 12.3 show how molecular signals appear in terms of their blockade attributes in the online setting (with event-observation time on the vertical axis in Fig. 12.3). In Fig. 12.4, the PRI sampling acquisition results are shown, with the rarer 9GC molecules properly identified, and sampled for a full 5-s duration, while other molecules are rejected typically in a fraction of a second (with the prototype network setup used here). A major speedup can be achieved with extra nodes for server side computation and various optimizations.

The robustness of the results is then explored when there are numerous other classes present for multiclass PRI sampling (see Fig. 12.5). In Figs. 12.6 and 12.7,

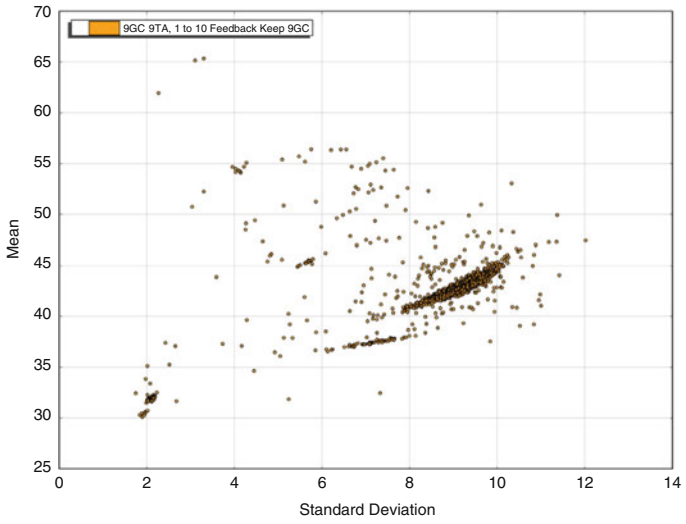


Fig. 12.2 Standard deviation vs. Mean on a {9GC,9TA} mixture in a 1:70 ratio. Clusters of 9GC and 9TA signal groups are identifiable, nearly full blockade signal group also present

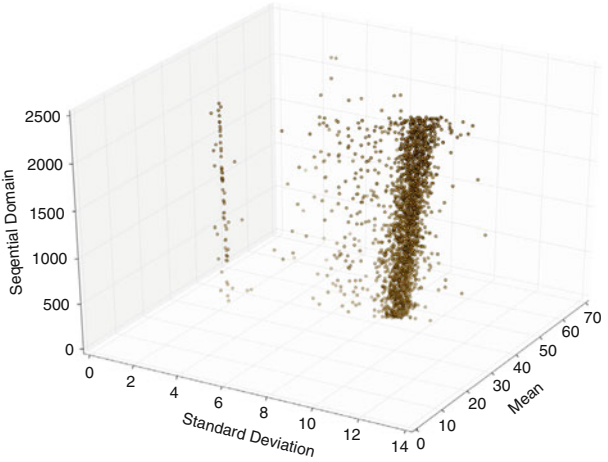


Fig. 12.3 Standard deviation vs. Mean vs. event observation time (vertical axis). Drift in the {9GC,9TA} signal is seen as the experiment proceeds due to evaporative concentration of the background salt. This results in altered environment for the DNA hairpins, one where the increasing magnitude of the blockade standard deviation is thought to be due to stronger (and noisier) DNA hairpin channel blockades

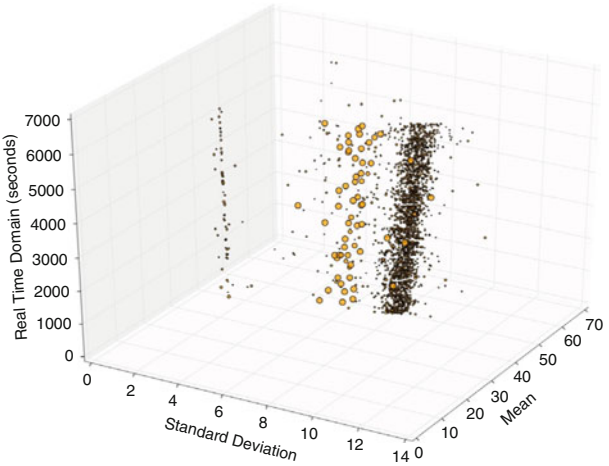


Fig. 12.4 Standard deviation vs. Mean vs. event observation time vs. PRI informed sample observation time (fourth dimension represented as the radius of the data point). This figure shows a successful real time operation on the PRI sampling method on the ND platform. 9GC signal is selected for observation and it is at a 1:70 lower concentration than the decoy 9TA DNA hairpins. As can be seen, only 9GC signals are held for the lengthier observation time, all other molecules being rejected promptly upon identification (the smaller diameter events points corre spond to short lived events), where the brief duration of the event is dictated by the active, PRI control, of the device voltage

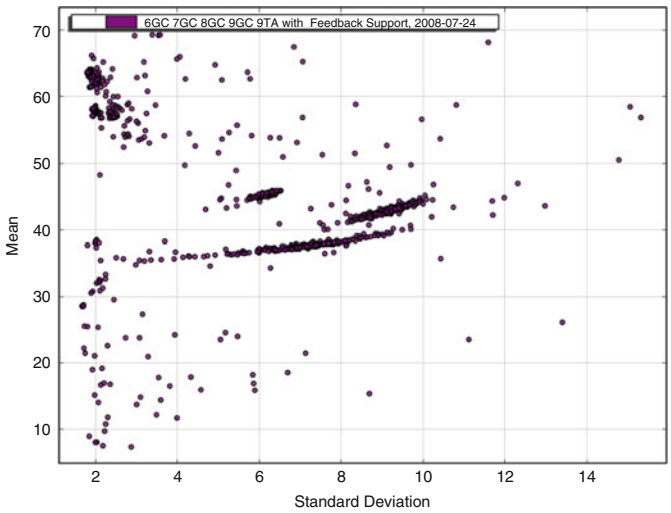


Fig. 12.5 Standard deviation vs. Mean on a {6GC,7GC,8GC,9GC,9TA} mixture. Clusters of the different species of blockade signal are clearly identifiable (and the nearly full blockade signal class is also present)

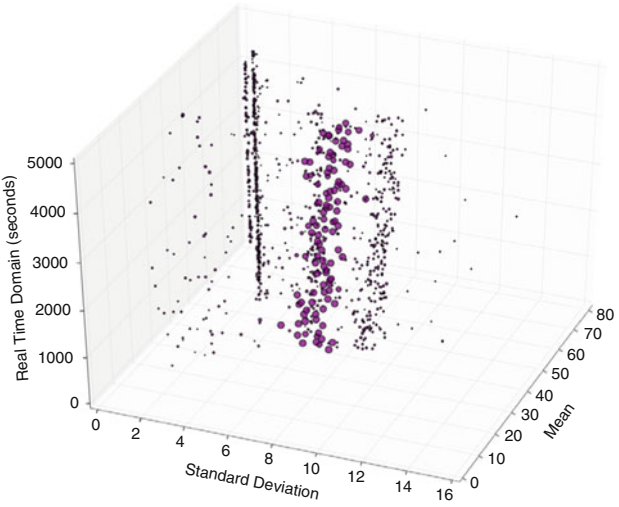


Fig. 12.6 Standard deviation vs. Mean vs. event observation time (vertical axis) vs. PRI informed sample observation time (fourth dimension represented as the radius of the data point). Drift in the signal is seen as the experiment proceeds, as before. Similar strong classification performance is demonstrated for this five class test as with the prior two class test

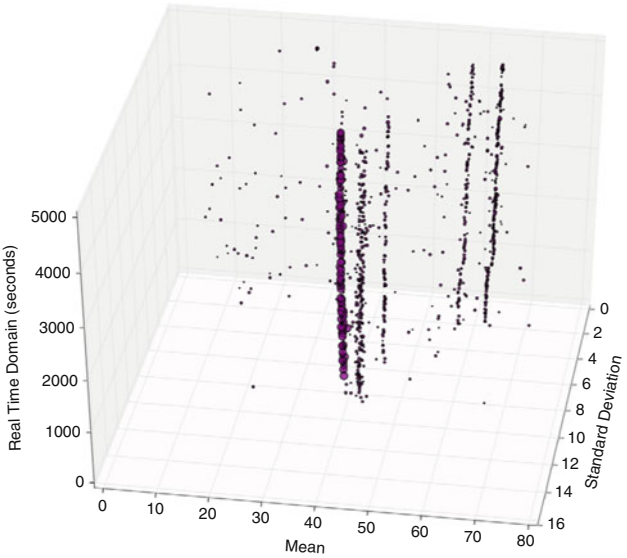


Fig. 12.7 Shows a rotated view of the results shown in Fig. 12.6. The successful 99.9% accurate separation of the 9GC from the {6GC,7GC,8GC,9TA} signals can be seen more clearly from this perspective. *Note*: the actual discriminating features used by the SVM classifier are not based on the mean and standard deviation statistical features plotted, but on a 150 component feature extraction based on HMM emission and transition probabilities (see [2] for details)

an approximately 1:70 mixture of 9GC:{6GC,7GC,8GC,9TA} is examined, with 9GC sample time again boosted correctly as indicated.

12.6 Discussion

Sample Boosting and Nanomanipulation via PRI Selection: PRI sampling is done with mixtures involving the five DNA hairpin control molecules examined in [1, 2, 13]. In the nanopore transduction detection context [3, 14], it is hypothesized that auxiliary molecules consisting of these same control molecules can be covalently linked to sensing moieties of interest to provide the beginnings of a generalized detection platform. Once covalently linked, however, further optimization/selection on the control molecule portion, as regards stem length tuning or base-pair alterations, is usually needed to reacquire a highly structured modulatory signal (with stationary statistics) [12]. Once a bifunctional molecule has been engineered to desired channel modulation and target-analyte interaction, the nanopore detector can be operated with the transduction molecule and signal analysis software to classify the different blockade signals. As far as the signal processing software is concerned, however, pattern recognition that resolves different hairpin blockades, or the same hairpin blockader with/without

complexation at its binding moiety, is practically the same. Thus, the five DNA hairpin PRI-sampling study examined here demonstrates a capability for nano-manipulation when observing reactants, via the mechanism of recognition and appropriate selection.

12.7 Conclusions

The primary purpose of this experiment was to develop an implementation of PRI experimental protocol for more specific and efficient collection of signals in nanopore cheminformatics experiments. In the Results, PRI-sampling is shown to boost the acquisition rates on molecules of interest by orders of magnitude, greatly extending the applicability of the Nanopore's inherent serial-event detection capability.

A secondary purpose was to explore the resolving/tracking power of the PRI system when applied to binding experiments. The clear binding behavior shown ("tracked") in the Results indicates that population-based binding studies using the nanopore detector can be done, and suggests that sufficient sensitivity to state might be possible for tracking an individual binding history in future efforts along these lines.

Acknowledgments Federal funding was provided by NIH K 22 (SWH PI, 5K22LM008794). State funding was provided from a LaBOR Enhancement (SWH PI).

References

1. Vercoutere W, Winters Hilt S, Akeson M et al (2001) Rapid discrimination among individual DNA hairpin molecules at single nucleotide resolution using an ion channel. Nat. Biotechnol. 19, 3, 248 252.
2. Winters Hilt S, Vercoutere W, DeGuzman VS, Deamer DW, Akeson M, Haussler D (2003) Highly accurate classification of Watson Crick basepairs on termini of single DNA mole cules. Biophys. J. 84, 967 976.
3. Winters Hilt S (2007) The alpha hemolysin nanopore transduction detector: single molecule binding studies and immunological screening of antibodies and aptamers. BMC Bioinformat ics 8, Suppl 7, S9.
4. Durbin R (1998) Biological sequence analysis: probalistic models of proteins and nucleic acids. Cambridge, UK/New York: Cambridge University Press, xi, 356 p.
5. Winters Hilt S (2006) Hidden Markov model variants and their application. BMC Bioinfor matics 7, Suppl 2, S14.
6. Winters Hilt S, Baribault C (2007) A novel, fast, HMM with duration implementation for application with a new, pattern recognition informed, nanopore detector. BMC Bioinformatics 8, Suppl 7, S19.
7. Vapik V (1995) The nature of statistical learning theory. New York: Springer Verlag.
8. Platt JC (1998) Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines, Microsoft Research, Technical Report MSR TR 98 14.

9. Winters Hilt S, Yelundur A, McChesney C, Landry M(2006) SVM clustering. BMC Bio informatics 7, S2 S4.
10. Winters Hilt S and S Merat (2007) SVM clustering. BMC Bioinformatics 8, Suppl 7, S18.
11. Churbanov A, Winters Hilt S (2008) Implementing EM and Viterbi algorithms for hidden Markov model in linear memory. BMC Bioinformatics 9, 228.
12. Winters Hilt, S. Nanopore transduction analysis of biotin streptavidin binding. Submitted to BMC Biotechnology.
13. Vercoutere W, Winters Hilt D, Akeson M et al (2003) Discrimination among individual Watson Crick base pairs at the termini of single DNA hairpin molecules. Nucleic Acids Res. 31, 1311 1318.
14. Winters Hilt S (2006) Nanopore detector based analysis of single molecule conformational kinetics and binding interactions. BMC Bioinformatics 7, Suppl 2, S21.

Chapter 13
An MLP Neural Network for ECG Noise Removal Based on Kalman Filter

Sara Moein

Abstract In this paper, application of Artificial Neural Network (ANN) for electrocardiogram (ECG) signal noise removal has been investigated. First, 100 number of ECG signals are selected from Physikalisch-Technische Bundesanstalt (PTB) database and Kalman filter is applied to remove their low pass noise. Then a suitable dataset based on denoised ECG signal is configured and used to a Multilayer Perceptron (MLP) neural network to be trained. Finally, results and experiences are discussed and the effect of changing different parameters for MLP training is shown.

Keywords Kalman filter · Noise removal · MLP training · Dataset · Performance

13.1 Introduction

An electrocardiogram (ECG) signal is the electrical activity of the heart that is caused by the impulses that travel through the heart. It provides information about the heart rate, rhythm, and morphology. A typical ECG wave of a normal heartbeat consists of three parts: *P* wave, *QRS* complex, and *T* wave. Figure 13.1 depicts the *PQRST* shape of ECG signal [1]. The *P* wave reflects the activation of the right and left atria. The *QRS* complex shows depolarization of the right and left ventricles. The *T* wave that is after *QRS* complex reflects ventricular activation [2, 3].

However, the presence of artifacts and noise within the signal may influence the diagnosis. Artifacts and noise are generated by biological and environmental recourses [4]. Mechanical movement of electrodes and power line interferences causes ECG artifacts [5].

S. Moein
Multimedia University, Cyberjaya, Malaysia
e mail: sara.moein08@mmu.edu.my

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 13,

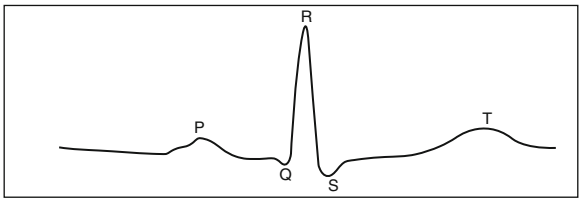


Fig. 13.1 Morphology of *PQRST* of recorded ECG signal from a normal human

13.2 Literature Review

Noise removal from ECG signal has been extensively studied [4 8, 17]. One particular study conducted by SadAbadi et al. [9] defined a formula to determine the window length for the noise removal process. Their proposed method was able to preserve *QRS* complex characteristic points, especially *Q* and *S* waves.

Lian and Hoo [10] proposed a cost-effective way to remove the noise by the use of Finite Impulse Response (FIR) filters. Their proposed filter uses the multiplication-free Recursive Running Sum (RRS) filters to achieve computational efficiency. The filter is realized using multiplication-free RRS and is able to remove either 50 or 60 Hz power line noise by toggling a switch. Simulated results from the filter have shown satisfactory performance in the filtering of the ECG signal, producing a clean output at a relatively computational cost. On the other hand, previous studies [11, 12] show that artificial neural network has been used for the classification of ECG signal. Sordo [12] presented a Multilayer Perceptron (MLP) neural network for ECG classification and showed that MLP is fast in learning for ECG signal classification. Here, the objective is to design a MLP to denoise the ECG signal based on Kalman filter.

13.3 Review of Kalman Filter

The Kalman filter is a powerful tool that plays an increasing role in solving the problems of the real world. It is an efficient recursive filter that estimates the state of a linear dynamic system from a series of noisy measurements. A previous study has shown that the Kalman filter is a good estimator for a large class of problems with increase in effectiveness for even larger classes [13]. This article aims to prove the application of this filter for noise removal. The Kalman filter consists of two steps: (1) prediction and (2) correction. In the first step, the state is predicted with a dynamic model. In the second step, it is corrected with the observation model, so that the error covariance of the estimator is minimized. In this sense, it is an optimal estimator. Assume that the system variables, represented by the vector$X$, are governed by the equation $X_{k+1} = AX_k + W_k$, where W_k is random process noise,

and the subscripts on the vectors represent the time step. A relates the state at the previous time step $k-1$ to the state at the current step k. There are many alternative ways to formulate the Kalman filter equations. One of the formulations is given in (13.1) (13.4) as follows:

$$S_k = P_k + R \tag{13.1}$$

$$K_k = AP_k S_k{}^{1} \tag{13.2}$$

$$\hat{X}_{k+1} = A\hat{X}_k + K_k(Z_{k+1} - A\hat{X}_k) \tag{13.3}$$

$$P_{k+1} = AP_k A^T + Q - AP_k S_k \;{}^{1}P_k A^T \tag{13.4}$$

Here, the superscripts -1 and T indicate matrix inversion and matrix transposition, respectively, S is the covariance of the innovation, K is the gain matrix, and P is the covariance of the prediction error. In (13.4), the first term used to derive the state estimate at time $k+1$ is just A times the state estimate at time k. $(AP_k A^T)$would be the state estimate if we did not have a measurement. In other words, the state estimate propagates in time just like the state vector. The term $(AP_k S_k \;{}^{1}P_k A^T)$ in (13.4) is called the correction term, and it represents how much time is required to correct the propagated estimate due to our measurement.

13.4 Multilayer Perceptron

A multilayer perceptron (MLP) is a network of simple neurons called perceptrons. The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. More explanation is available in artificial neural network books [14].

13.5 Proposed Method

13.5.1 Data Collection

13.5.1.1 Database

The database that is used for this study is Physikalisch-Technische Bundesanstalt (PTB) database, an ECG database of healthy volunteers and patients with different heart diseases [15]. The ECG signals are affected with noise and accidental

disturbance in some of the signals. A total of 100 signals are collected for noise removal. The frequency of noise is different in each ECG signal. Therefore, Kalman filter parameters values are variable. The objective is to remove the noise using Kalman filter from ECG signals to prepare a dataset for multilayer perceptron training.

13.5.1.2 Kalman Filter for Noise Removal

All 100 noisy signals must be denoised using Kalman filter. The tuning of Kalman filter plays a critical role in the Kalman filter design. Mentioned below are some parameters of the Kalman filter that were used to denoise the signal of a healthy human. f_0 and f_s indicate the band of frequency for filtering the low pass noise in ECG signal. See Fig. 13.2. There are two other signals that are denoised using the Kalman filter (Figs. 13.3 and 13.4). The Kalman filter is implemented with MATLAB.

$Q = 10^{\ 1} - 2 \times \max|X|$, $R = \text{Variance}(X)$, $A = 1$, $f_0 = 100$, $f_s = 80$, $S_0 = \{0, 0, \ldots, 0\}$, $\hat{X}_0 = \{x_0, x_1, \ldots, x_{n\ 1}\}$, $k_0 = \{0, 0, \ldots, 0\}$, $w = 2\pi(f_0/f_s)$, $Z_0 = \{0, 0, \ldots, 0\}$, $p_0 = \{0, 0, \ldots, 0\}$

Here, X is the input signal and n is the number of samples.

Based on the Kalman filter initialization, it is true to classify signals to six classes, considering that Parameters of Kalman filter for each class are unique.

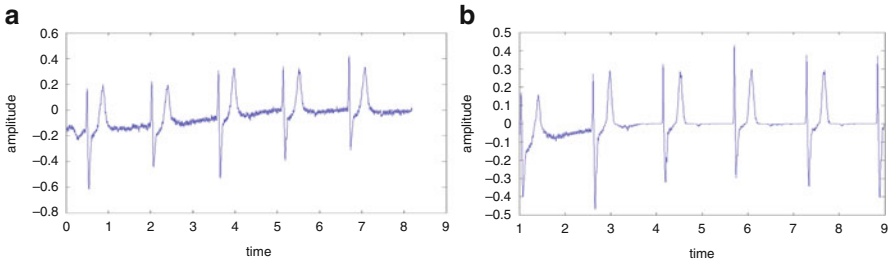


Fig. 13.2 ECG signal of a healthy human. Noisy ECG (**a**); ECG after using Kalman filter (**b**)

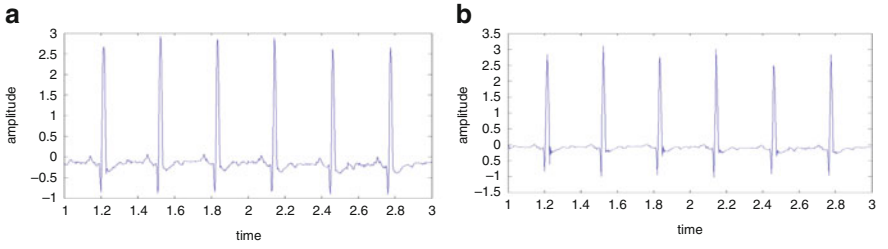


Fig. 13.3 ECG signal of myocardial infarction. Noisy ECG (**a**); ECG after using Kalman filter (**b**)

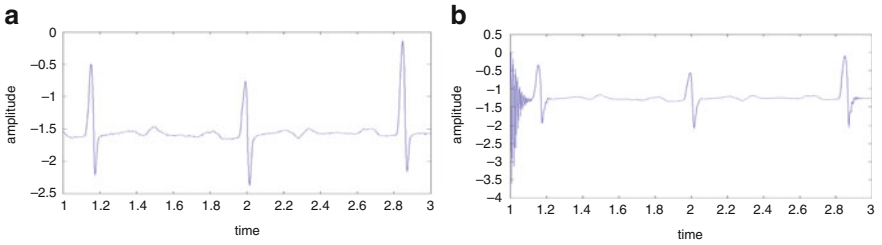


Fig. 13.4 ECG signal with accidental disturbance. Noisy ECG (**a**); ECG after using Kalman filter (**b**)

Table 13.1 Assigning label to denoised ECG signals based on Kalman filter parameters

Denoised signal name	f_s	f_0	w		Q			Denoised signal label
ECG of normal human	100	80	w	$2 \times p_i \times f_0/f_s$	Q	$1e$	$2 \times \max\|X\|$	0
Heart failure ECG	100	40	w	$2 \times p_i \times f_0/f_s$	Q	$1e$	$2 \times \max\|X\|$	1
Myocardial infarction ECG	80	100	w	$2 \times p_i \times f_0/f_s$	Q	$2e$	$2 \times \max\|X\|$	2
Miscellaneous ECG	100	60	w	$p_i \times f_0/f_s$	Q	$1e$	$2 \times \max\|X\|$	3
Dysrhythmia ECG	80	60	w	$2 \times p_i \times f_0/f_s$	Q	$2e$	$2 \times \max\|X\|$	4
Bundle branch block ECG	80	40	w	$2 \times p_i \times f_0/f_s$	Q	$2e$	$2 \times \max\|X\|$	5

Table 13.1 shows each class of signals based on Kalman filter parameters and the label that is assigned to each class. This classification has been done based on f_s, f_0, w, and Q.

13.5.2 Dataset for MLP Training

For training a MLP neural network, it is important to have a valid dataset from denoised ECG signals [16]. Since the objective is to train the MLP for noise removing, the target value will be the denoised ECG signal. The denoised ECG signals are obtained using the Kalman filter, as mentioned in the previous section.

The provided dataset is a collection of 100 ECG signals whose statistical features are extracted and used as the attributes of the dataset. Each row of the dataset presents one noisy ECG signal, and each column is an attribute of ECG. Considering that, the last column is the value of the label that is assigned to each denoised ECG signal. Since all signals are denoised with Kalman filter in the previous section, the label of each denoised signal is available. Table 13.2 shows a part of the dataset.

Table 13.2 Part of the dataset for training MLP

No. of noisy ECG	Variance	Standard deviation	Mean	Denoised signal label
1	0.2139	0.4625	0.7980	4
2	0.0675	0.2597	0.5737	4
3	0.0712	0.2669	7.4824	5
4	0.0163	0.1277	3.1738	2
5	0.0133	0.1151	0.8195	2
6	0.0541	0.2327	2.1734	3
7	0.0239	0.1545	0.2710	2
8	0.0609	0.2468	2.3842	3
9	0.0328	0.1812	1.5237	1
10	0.0675	0.2597	0.5737	1
11	0.0293	0.1711	0.0399	0
12	0.0679	0.2605	0.5616	3
13	0.0115	0.1072	0.9313	2
14	0.0018	0.0429	7.6588	5

13.5.3 MLP Training

The three-layer perceptron is the network to be trained for the proposed method. A back propagation Generalized Delta Rule (GDR) is used to train the MLP. Three is the number of nodes in input layer, since that is the number of features in the dataset. In addition, one node is assumed as the number of nodes in output layer. Number of nodes in hidden layer is the point that can affect the performance of a trained neural network. We train the MLP with different number of nodes in hidden layer and after analyzing, the number of nodes in this layer that give better results can be indicated. On the other hand, number of training iteration is the other effective parameter that must be considered.

13.6 Experiment and Results

As mentioned in the previous section, there are 100 records in the dataset for training the MLP. Due to the lack of cases in our dataset and to increase the validity and generality of the results, a k-folding scheme with $k = 5$ was applied. In this method, the training procedure is repeated k times, each time with 80% of the samples in the dataset for training and the remaining 20% for testing. The 20% testing section is non-overlapping. Figure 13.5a c presents the training error of MLP neural network for 20 of the test samples with various numbers of nodes in the hidden layer and training iteration. Results show that the best performance for MLP training is achieved when there were 2,000 training iteration and 15 nodes in the hidden layer. The performance is more than 90% since error for all test samples is less than 0.5.

Finally, the trained MLP indicates to which class each noisy signal is related. All parameters of Kalman filter are presented in Table 13.1.

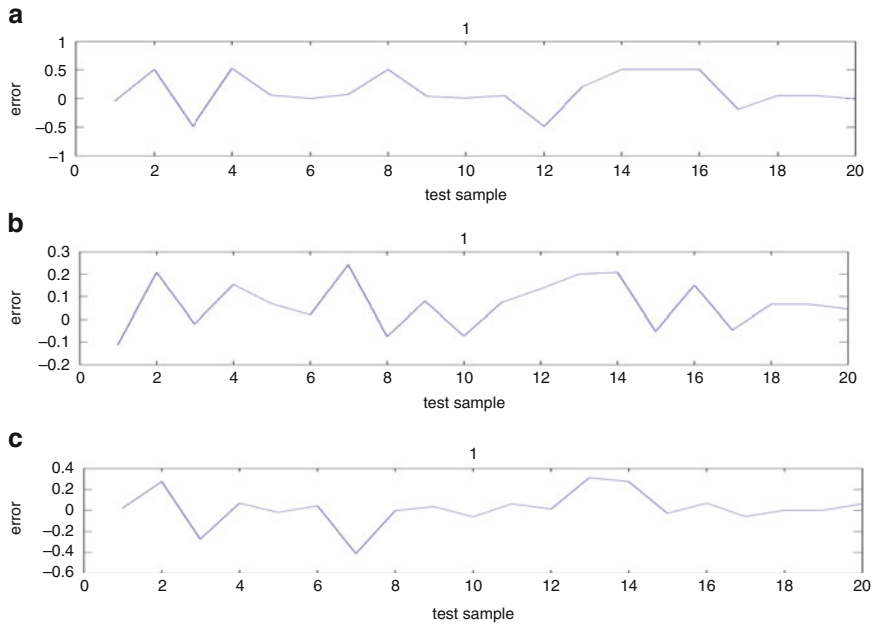


Fig. 13.5 Output error for test samples. (**a**) MLP training, five nodes in hidden layer, 500 training iterations; (**b**) MLP training, ten nodes in hidden layer, 1,000 training iterations; (**c**) MLP training, 15 nodes in hidden layer, 2,000 training iterations

13.7 Conclusion

In this paper, application of MLP for noise removal from ECG signal using Kalman filter is investigated. Results show that the prepared dataset is acceptable for training a MLP for noise removal. Also, the number of nodes in the hidden layer and the number of training iteration are very important for increasing the performance. As future works, one may concentrate on the problem of advanced training algorithm for neural networks. Using fuzzy toolbox for increasing the accuracy of data and results might be of interest too.

References

1. Sornmo L, Laguna P (2005) Bioelectrical Signal Processing in Cardiac and Neurological Applications. Elsevier: Amsterdam.
2. Information about heart, http://www.mdsr.ecri.org/summary/detail.aspx? doc id 8246.
3. Behbahani S (2007) Investigation of Adaptive Filtering for Noise Cancellation in ECG signals, Second International Multi Symposiums on Computer and Computational Sciences.

4. Ziarani A K, Konrad A (2004) A nonlinear adaptive method of elimination of power line interferences in ECG signals. *IEEE Transactions on Biomedical Engineering*, 49(6), pp. 540 547.
5. Bozic S M (1983) Digital and Kalman Filtering: An Introduction to Discrete Time Filtering and Optimal Linear Estimation. Wiley: New York.
6. Willems J L, Arnaud P et al (1987) A reference data base for multi lead electrocardiographic computer measurement programs. *Journal of American College of Cardiology*, 10, pp. 1313 1321.
7. Losada R A (2004) Design finite impulse response digital filters. Part II. *Microwaves & RF*, 43, 70 84.
8. Moein S, Monadjemi S A, Moallem P (2008) A Novel Fuzzy Neural Based Medical Diagnosis System. WASET, Egypt, vol 26.
9. SadAbadi H, Ghasemi M et al (2007) A Mathematical Algorithm for ECG Signal Denoising Using Window Analysis. Biomedical Paper Medical Faculty University Palacky Olomouc Czech Republication, 151(1), pp. 73 78.
10. Lian Y, Hoo P C (2006) Digital elliptic filter application for noise reduction in ECG signal. *WSEAS Transactions on Electronics*, 3(1), pp. 65 70.
11. Orfanidis S J (1996) Introduction to Signal Processing. Prentice Hall: Upper Saddle River, NJ.
12. Sordo M (2002) Introduction to Neural Networks in Healthcare. OpenClinical:. Knowledge Management for Medical Care, Harvard, http://www.openclinical.org.
13. Moein S, Khasimatol S et al (2009) ECG Noise and Artifact Removal Using Kalman Filter, ISSAP Malaysia.
14. Hassoun M H (1995) Fundamentals of Artificial Neural Network. Massachusetts Institute of Technology, ISBN 0 262 08239 x.
15. A database for heart signals, http/www.physionet.org/physiobank/database/PTB.
16. Moein S (2008) Hepatitis Diagnosis by Training a MLP Artificial Neural Network. world comp08 conference, vol 14, Las Vegas, USA.
17. Saramaki T (1993) Finite impulse response filter design. In: Mitra S K, Kaiser J F (Eds) Handbook for Digital Signal Processing. Wiley Interscience: New York.

Chapter 14
Discovery of Structural Motifs Using Protein Structural Alphabets and 1D Motif-Finding Methods

Shih-Yen Ku and Yuh-Jyh Hu

Abstract Although the increasing number of available 3D proteins structures has made a wide variety of computational protein structure research possible, yet the success is still hindered by the high 3D computational complexity. Based on 3D information, several 1D protein structural alphabets have been developed, which can not only describe the global folding structure of a protein as a 1D sequence, but can also characterize local structures in proteins. Instead of applying computationally intensive 3D structure alignment tools, we introduce an approach that combines standard 1D motif detection methods with structural alphabets to discover locally conserved protein motifs. These 1D structural motifs can characterize protein groups at different levels, e.g., families, super families, and folds in SCOP, as group features.

Keywords Protein structure · Structural alphabet · Motif

14.1 Introduction

As the rapid growth of protein structural information, biologists require accurate classification to understand and rationalize the variety in proteins [1]. To ensure a more easily constructed and better comprehensible classification, it is desired that we provide only essential characteristic structural descriptions of protein functional parts. With such a classification, we can assign a novel protein to known categories, and thus predict its structures and functions. The task of extracting characteristic structural features for classification is difficult and becomes more challenging for small proteins, where the characteristic statistics are marginal owing to short protein chains, or for proteins that only share low sequence similarity.

Y.J. Hu (✉)
Department of Computer Science, Institute of Biological Engineering, National Chiao Tung University, Hsinchu, Taiwan
e mail: yhu@cs.nctu.edu.tw

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 14,

In functionally related protein families, there usually exist conserved local structural characteristics, e.g., the binding sites for metal-binding proteins. Given that the conservation in local active sites is likely to reflect similar biological functions, 3D patterns of local active sites can be used to predict the functions of previously unknown proteins [2]. These conserved structural features themselves represent significant motifs, which can be identified and described in various ways. Unlike most previous works on protein local structures, we describe a combinatorial approach to structural motif discovery. It first converts protein 3D structures into 1D structural alphabet letters, and then identifies conserved local features as 1D structural alphabet sequence motifs. There are several advantages of 1D structural alphabet over the conventional 3D co-ordinates. First, 1D representation of protein structures is more efficient in comparison and more economical in storage. Second, many commonly used 1D sequence tools can be directly applied to protein structure and sequence analysis. Third, 1D-based approaches can serve as pre-processors to filter out irrelevant proteins prior to the application of more computationally intensive 3D structure analysis tools.

14.2 Discovery of Structural Alphabet Motifs

The discovery of structural motifs can be divided into two stages. Given a structural alphabet, we can first transform a set of functionally or structurally related proteins, e.g., SCOP family, into a 1D representation. Different alphabets were derived based on different design philosophies [3 6]. Their size can vary from a dozen to nearly a hundred. They reflect different structural characteristics and have various applications. In different domains, we can adopt an appropriate structural alphabet to transform amino acid sequences or protein 3D structures into 1D structural alphabet sequences as required. In the second stage, we can apply a sequence motif detection algorithm to discover significant motifs from the 1D structural alphabet sequences. Like structural alphabets, a significant number of motif detection tools have been developed based on different objective functions, motif representations, and search strategies [7 10].

For this study, we designed a structural alphabet [6] that contains 18 letters, five of which represent the helix structure, eight for the sheet, and the rest for the coil. To discover structural motifs, we used MEME [7], which adopts an expectation maximization approach to find motifs represented as weight matrices. Unlike IUPAC-IUB codes, motifs described in weight matrices are more flexible because a weight matrix can show each alphabet letter preference in every motif position. Besides, a weight matrix can be easily transformed to IUPAC-IUB codes or regular expressions when necessary, but not vice versa. We call the motifs found by MEME from the structural alphabet sequences simple motifs. When the local properties in protein structures are too complicated, e.g., multiple binding sites or sub-domains, to capture in a simple motif, we can further combine several simple motifs into a compound motif. To avoid the computational complexity of combining matrices,

we transform simple motifs to regular expressions first, and then combine them to a compound motif. A compound motif example looks like the following.

$M_1(20,\ 50)M_2(0,\ 6)M_3$, where M_1, M_2, and M_3 are simple motifs, and the numbers in the parentheses denote the range of residue separation between motifs.

$$
\begin{aligned}
M_1 &= \text{SP[PS][SN]N[NE]EE},\\
M_2 &= \text{[WE][NE]EEACWGQS},\\
M_3 &= \text{TTTTTTTTTLK[TG][SH]WNMR[DQ]},
\end{aligned}
$$

where letters in brackets denote the possible structural alphabet letters in that particular motif position.

We show in Fig. 14.1 a general framework for structural motif discovery in which the structural alphabet and the motif finding algorithm can be replaced when necessary in different applications.

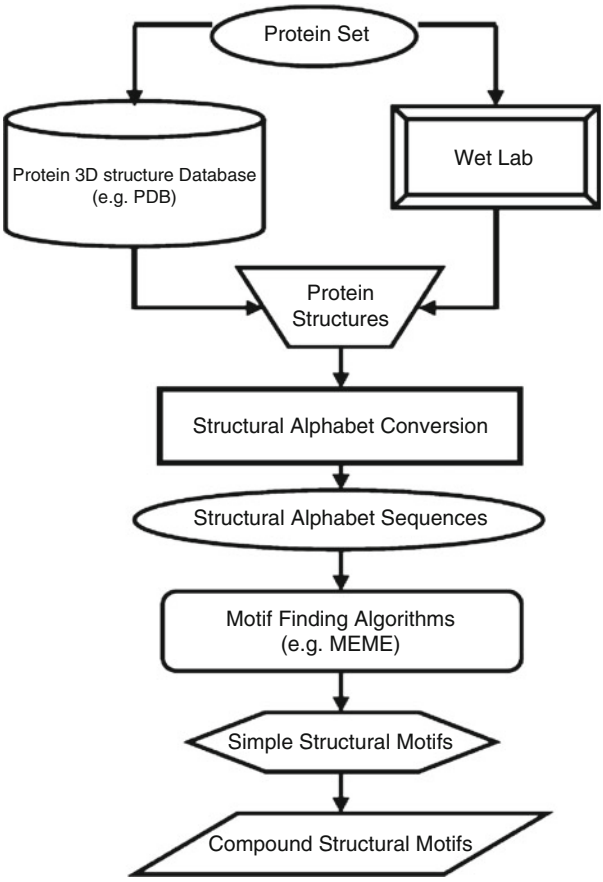


Fig. 14.1 System flow of structural motif discovery

The structural alphabet motifs can characterize the local structure features conserved in functionally related proteins. Based on the motif analysis in alphabet letter preference, alphabet letter occurrence distribution, and its significance, we may get a deeper insight of protein structures. To show the difference in alphabet conservation between structural alphabet motifs and the corresponding amino acid motifs, we showed one example motif for a SCOP family in Fig. 14.2a, which indicates that the structural alphabet motif is more conserved than the amino acid motif.

Besides alphabet conservation, we can also study the distribution of alphabet letter occurrences in each position of the structural and the amino acid motif, respectively. An example is shown in Fig. 14.2b and c. From the histograms, we can analyze the number of occurrences of each alphabet letter in a particular position of a motif. From the comparison of occurrences between structural alphabet and amino acids, we can derive the relationships between protein sequences and structures, e.g., the preference of structural alphabet for specific amino acids. Relationships of this kind about motifs can be further refined as building blocks to predict the structures of novel protein sequences.

14.3 Application Example of Structural Motifs

The C2H2 zinc finger is one of the best-studied metal-binding domains. It was first observed as a repeated zinc-binding motif with DNA-binding properties in the Xenopus transcription factor IIIA, and the term "zinc finger" is now largely used to denote any compact domain stabilized by a zinc ion [11, 12]. The domains from C2H2-like fingers consist of a β-hairpin followed by an α-helix that forms a left-handed ββα-unit, where two zinc ligands are contributed by a zinc knuckle at the end of the β-hairpin and other two ligands come from the C-terminal end of the α-helix [13, 14]. To demonstrate that our approach is capable of characterizing the structural ββα-unit, we analyzed the structural motifs discovered from 156 C2H2 zinc finger proteins in SCOP. A motif found was considered to match a (sub-)domain correctly if more than half of the residues in the (sub-)domain were included in the motif. If any simple or compound motif correctly corresponded to a (sub-)domain, we claimed that this (sub-)domain was recovered successfully (i.e., a hit). In Table 14.1, we present the compound motif found to characterize the (sub-)domains, and its coverage. The results suggested that using protein structural alphabet combined with 1D motif-finding algorithm was able to recover the structural (sub-)domains in proteins. We show some C2H2 zinc finger proteins with structural motifs marked in Fig. 14.3.

14.4 Conclusion

In this chapter, we introduced a general framework for structural motif discovery and proposed the applications of the motifs. Two major components in our framework are (1) the structural alphabet used to describe protein structures and (2) the

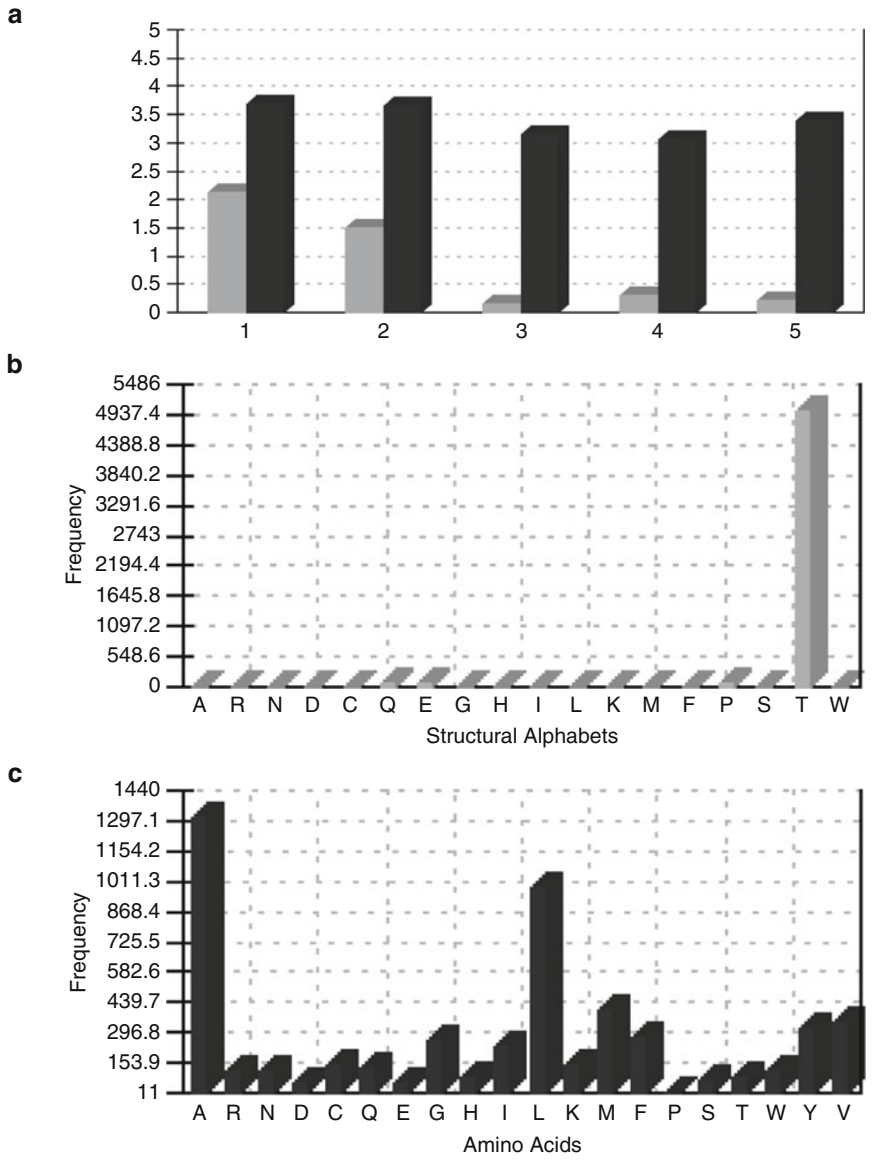


Fig. 14.2 Examples of motif analysis. (**a**) Histogram of entropy in each position of the structural alphabet motif and its corresponding amino acid motif. The *x* axis indicates the positions, and the *y* axis shows the entropy. Entropy of structural alphabet motif is colored in *gray*, and the amino acid motif in *black*. The lower the entropy, the more conserved the alphabet. (**b**, **c**) Histograms of alphabet letter occurrence distribution in the first position of the structural alphabet motif and its corresponding amino acid motif. The *x* axis indicates all the alphabet letters, 18 in structural alphabet and 20 in amino acids. The *y* axis shows the number of occurrences for a particular alphabet letter. (**b**) The distribution of structural alphabet letter occurrences in the first position. (**c**) The distribution of amino acid occurrences in the first position within the corresponding amino acid motif

Table 14.1 Summary of motifs mapping to C2H2 zinc finger ββα unit that consists of β hairpin and α helix

Structural (sub)domains	Motifs	Hit[a]	Coverage (%)[b]
β Hairpin	[FH]CWNA[RC]QK(0 2) [GN][HE][NE]AC [AW]RQ	131	83.9
α Helix	[GN][HE][NE]AC[AW]RQ(0 5)TTTTTT[PL] [KPL]	142	91.0
ββα Unit	[FH]CWNA[RC]QK(0 2) [GN][HE][NE]AC [AW]RQ(0 5)TTTTTT[PL][KPL]	124	79.5
Total		156	100

[a]We called it a hit for a structural (sub)domain when more than half of the (sub)domain residues were contained in a motif. We presented the count of hits of different (sub)domains
[b]Coverage was defined as the ratio of the count of hits to the number of zinc finger proteins, e.g., if total 156 and Hit 131, then coverage 131/156 83.9%

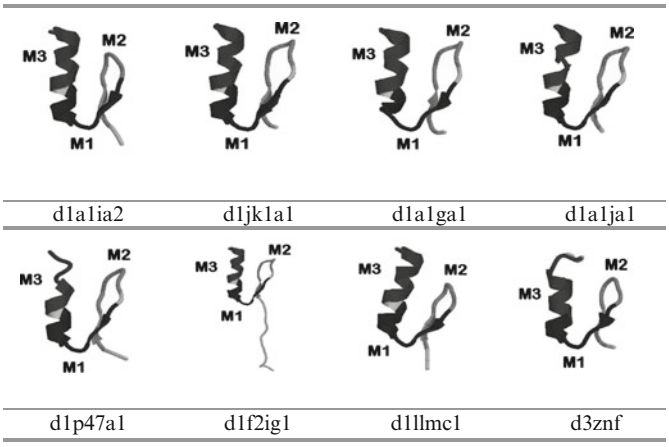


Fig. 14.3 Examples of C2H2 zinc finger protein structures. The simple motifs that map to the β hairpin and the α helix are highlighted in different *gray* levels, where M_1 [GN][HE][NE]AC [AW]RQ, M_2 [FH]CWNA[RC]QK, and M_3 TTTTTT[PL][KPL]. The compound motif mapping to the ββα unit is [FH]CWNA[RC]QK(0 2) [GN][HE][NE]AC[AW]RQ(0 5)TTTTTT [PL][KPL]

motif-finding algorithm used to discover significant local structure features. In our evaluation experiments, we used the structural alphabet designed in [6] and a widely used motif detection algorithm, MEME [7]. These components can be flexibly replaced with others when necessary to increase the applicability in different domains. The current results showed that using structural alphabets combined with 1D motif-finding algorithms could successfully identify biologically meaningful sub-domains in proteins.

We plan to continue the work in the following directions. First, many structural alphabets and quite a few motif detection algorithms have been developed based on

different design philosophies and application domains. We intend to incorporate other structural alphabets and motif-finding algorithm into our system. We expect to discover more kinds of motifs in a wider variety of protein structures. Second, the analysis of the distribution of alphabet occurrences and conservations within the motifs provides a different point of view from which to investigate the conserved evolutionary relationships in proteins as well as an alternative way in which to assist in protein structure prediction. We intend to design a protein function predictor using structural motifs as important features. Based on the motifs, the functions of novel proteins can be predicted by classifying them in to protein groups with known functions. Third, as many other protein structure or protein function prediction systems are available, we also plan to verify the possibility of using structural alphabet-based methods as a pre-processor. Compared with most prediction strategies typically based on 3D information, alphabet-based methods have much lower computational complexity. Thus they can help other predictors constrain the search space efficiently by filtering out irrelevant predictions in advance.

References

1. Berman HM, Battistuz T, Bhat TN, et al (2002) The protein data bank. Acta Crystallogr D Biol Crystallogr 58:899 907.
2. Unger R, Harel D, Wherland S, Sussman JL (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 5:355 373.
3. de Brevern AG (2005) New assessment of a structural alphabet. In Sillico Biol 5:26.
4. Camproux AC, Gautier R, Tuffery P (2004) A hidden Markov model derived structural alphabet for proteins. J Mol Biol 339:591 605.
5. Offmann B, Tyagi M, de Brevern AG (2007) Local protein structures. Curr Bioinformatics 2:165 202.
6. Ku S, Hu Y (2008) Protein structure search and local structure characterization. BMC Bioinformatics 9:349.
7. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res 34:W369 W373.
8. Roth FP, Hughes JD, Estep PW, Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole genome mRNA quantitation. Nat Bio technol 16:939 945.
9. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically signifi cant alignments of multiple sequences. Bioinformatics 15:563 577.
10. van Helden J, Andre B, Collado Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol 281:827 842.
11. Laity JH, Lee BM, Wright PE (2001) Zinc finger proteins: new insights into structural and functional diversity. Curr Opin Struct Biol 11:39 46.
12. Iuchi S (2001) Three classes of C2H2 zinc finger proteins. Cell Mol Life Sci 58:625 635.
13. Grishin NV (2001) Treble clef finger a functionally diverse zinc binding structural motif. Nucleic Acids Res 29:1703 1714.
14. Wang B, Jones DN, Kaine BP, Weiss MA (1998) High resolution structure of an archaeal zinc ribbon defines a general architectural motif in eukaryotic RNA polymerases. Structure 6:555 569.

Chapter 15
Biological Databases at DNA Data Bank of Japan in the Era of Next-Generation Sequencing Technologies

Yuichi Kodama, Eli Kaminuma, Satoshi Saruhashi, Kazuho Ikeo, Hideaki Sugawara, Yoshio Tateno, and Yasukazu Nakamura

Abstract The Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ) has operated biological databases since 1987 in collaboration with NCBI and EBI. As one of the three major public databases, CIB-DDBJ has run four primary databases DDBJ, CIBEX, DDBJ Trace Archive (DTA), and DDBJ Read Archive (DRA) to collect, archive, and provide various kinds of biological data. As the massively parallel new sequencing platforms are increasingly in use, huge amounts of the raw data have been produced. To archive these raw data, we at CIB-DDBJ began operating a new repository, the DDBJ Read Archive (DRA). To accommodate efficiently the processed data as well, we have developed a new pipeline, the DDBJ Read Annotation Pipeline that deals with both data submission and analysis. For data produced by the next generation platforms, the three archives DRA, DDBJ, and CIBEX, which are interconnected by the pipeline, collect the raw, processed sequence, and quantitative data, respectively. The public biological databases at CIB-DDBJ, EBI, and NCBI will together construct world-wide archives for biological data by data sharing to accelerate research in life sciences in the era of next generation sequencing technologies.

Keywords Biological database · CIBEX · DDBJ · DDBJ omics archive · DDBJ read annotation pipeline · DDBJ read archive · DDBJ trace archive · Next-generation sequencing platform

Y. Tateno, (✉)
Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, Yata, Mishima, Shizuoka 411 8540, Japan
e mail: ytateno@genes.nig.ac.jp

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 15,

Abbreviations

CIB-DDBJ	Center for Information Biology and DDBJ
CIBEX	Center for Information Biology gene EXpression database
DDBJ	DNA Data Bank of Japan
DOR	DDBJ Omics aRchive
DRA	DDBJ Read Archive
DTA	DDBJ Trace Archive
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
ERA	European Read Archive
GEO	Gene Expression Omnibus
INSDC	International Nucleotide Sequence Database Collaboration
MGED	Microarray Gene Expression Data
MIAME	Minimum Information About a Microarray Experiment
MINSEQE	Minimum Information about a high-throughput Nucleotide SEQuencing Experiment
NCBI	National Center for Biotechnology Information
SRA	Short Read Archive
UHTS	Ultra High-Throughput Sequencing

15.1 Introduction

Since 1987 we have operated databases of biological molecules at the Center for Information Biology and DNA Data Bank of Japan (CIB-DDBJ, http://www.ddbj.nig.ac.jp) of the National Institute of Genetics. They now include a nucleotide sequence database (DDBJ) [1 3], a microarray database (CIBEX, http://cibex.nig.ac.jp) [4], and two new ones. The first one has been run in collaboration with the EMBL-Bank at the European Bioinformatics Institute (EBI, http://www.ebi.ac.uk) in UK and GenBank at the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov) in USA for more than 20 years in the framework of the International Nucleotide Sequence Database Collaboration (INSDC, http://insdc.org). The three databanks exchange the collected data on a daily basis so that users can retrieve essentially the same amount and quality of data from each database. According to the latest DDBJ Release 79 as of September 2009, DDBJ contributes about 15% of the total INSD data. Almost all INSD data from Japanese researchers are submitted to DDBJ.

The second one, CIBEX (Center for Information Biology gene EXpression database), is in compliance with the Minimum Information About a Microarray Experiment (MIAME) [5], which was prepared and proposed by the Microarray Gene Expression Data society (MGED, http://mged.org). It is noted that CIBEX will soon be replaced with a new database, DDBJ Omics aRchive (DOR,

http://trace.ddbj.nig.ac.jp/dor/index e.shtml). MIAME also has been adopted by ArrayExpress (http://www.ebi.ac.uk/microarray-as) at EBI and Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo) at NCBI. Following the international scientific recommendation that when one is to submit a manuscript about a newly sequenced gene or genome to a journal, the person is required to register beforehand the sequence data at DDBJ, EMBL-Bank, or GenBank [6]. MGED also asks one to register the relevant microarray data at GEO, ArrayExpress, or CIBEX before the submission of the manuscript to a journal [7, 8]. The submitter of either data in return receives an internationally recognized accession number that is to be listed in the paper published. While this mechanism assures the submitter the property and priority of the submitted data, it makes the reader of the paper easily accessible to the data by simple data retrieval on the corresponding accession number at those public databases.

As mentioned above, the three databanks of INSDC have collaboratively collected and released "processed" nucleotide sequence data. In late 1990s, the automated capillary platforms were introduced to many sequencing centers, which began to produce the raw data at the genomic scale. The raw data mean the intact data produced without artificial treatments and thus containing vectors and linkers. Researchers soon recognized the importance of the raw data for the quality assessment of sequences and high-level re-analyses such as re-base-calling and re-assembly. In 2001, NCBI thus launched the Trace Archive (http://www.ncbi.nlm.nih.gov/Traces) to archive the DNA sequence chromatograms (traces), base-calls, and quality estimates for single-pass reads from various large-scale sequencing projects. In collaboration with the NCBI Trace Archive and the EBI Ensembl Trace Server (http://trace.ensembl.org), the DDBJ Trace Archive (DTA, http://trace.ddbj.nig.ac.jp/dta/dta index e.shtml) has collected the raw data since 2008.

As the next generation sequencing platforms such as Roche 454, Illumina Genome Analyzer, and Applied Biosystems SOLiD are increasingly in use, huge amounts of sequence data have been produced at a number of laboratories. The amounts of sequence data produced at once, for example, can extend to 100 million sequences per single run [9]. This created a problem: how to deal with huge raw data at the cost of the computer storage and labor. However, since the importance of these raw data was obvious, NCBI first began to collect and release them in 2007, followed by EBI in 2008. We at CIB-DDBJ also began to accept the raw data as the DDBJ Read Archive (DRA, http://trace.ddbj.nig.ac.jp/dra/index e.shtml) in 2008. Then, in 2009 the representatives of DRA, ERA, and SRA met together and discussed for the first time the possible collaboration in dealing with the raw data that would come out one after another worldwide. They have agreed to accept, release, and exchange the data in a common format and content, and issue their own accession numbers to the submitted data, as INSDC has done for more than 20 years. They have also agreed to meet once a year on the occasion of the INSDC annual meeting and discuss the collaborative matters on data collection, release, and exchange.

In this paper, we report our activities of DRA, CIBEX, and an analytical pipeline, the DDBJ Read Annotation Pipeline, which are intended to promote research in life sciences in the era of the next-generation sequencing methodology.

15.2 DDBJ Read Archive

Next-generation sequencing platforms are revolutionizing life sciences. These instruments are producing vastly more sequence data than was ever possible with the capillary technology. In 2007, NCBI set out the Short Read Archive (SRA) to accommodate the data from next-generation sequencing platforms. Then, early in 2008, EBI began to operate the European Read Archive (ERA) and late in the same year CIB-DDBJ started to accept data produced by the next-generation sequencing platforms. To accept them, we first prepared temporary submission files at CIB-DDBJ and uploaded them to SRA, and then began operating a new repository, the DDBJ Read Archive (DRA), to archive the raw data from the new platforms just after the collaborative meeting mentioned above. DRA/ERA/SRA are intended to be public repositories of data output by "primary analysis" phase of the sequencing platform, which is the set of the sequences, the instrument data indicating the degree of reliability for each base-call (quality) and signal intensity measurements (intensity). DRA/ERA/SRA neither accept raw image data for its unacceptably large size, nor approve sequences with no quality measures. Following the agreement in the collaborative meeting, DRA started to issue its own internationally recognized accession numbers with prefix "DR." The submissions to DRA have so far mostly been from Japan and a few from other Asian countries. As the number of next-generation platforms in Japan and other Asian countries increases steadily, we expect that the amount of data submitted to DRA will be tremendously large. We will certainly need much larger computer storages and more staff for software engineering than those at present.

15.2.1 DRA Data Model

The sequence and other data are represented in a single flat file in the INSD data model. However, this data model is impractical to accommodate millions of sequences (reads) that were produced by the next-generation platform, simply because the data size is too huge to fit in the model. DRA instead uses the same data model developed mainly by SRA at NCBI. In this model, the metadata and sequence data, including the quality and intensity, are completely separated to allow us to design the format for the former independently of the data size of the latter. The metadata are composed of six objects, Submission (DRA), Study (DRP),

Experiment (DRX), Sample (DRS), Run (DRR), and Analysis. All objects but Analysis are accessioned with the prefix, as indicated in parentheses, followed by a six-digit number (e.g., DRA000001). Study, Experiment, Sample and Run objects are the core of the metadata information to describe how the data were obtained. A unique number is automatically assigned to each read with the Run accession number as the prefix. The Submission object is a package of metadata and/or data objects and a directive for what to do (e.g., to release or modify) with those objects. The Analysis object is a package associated with assembly, alignment, and quality control that are intended for downstream usage. The metadata objects are represented in XML documents in which each structure is defined by the respective XML schema. The objects are related to one other and compose hierarchy as shown in Fig. 15.1, which enables us to annotate efficiently the submitted data at each level with minimum redundancy of the metadata. This design also allows users to navigate from the individual reads to metadata and retrieve the data in a unit of Run or Experiment.

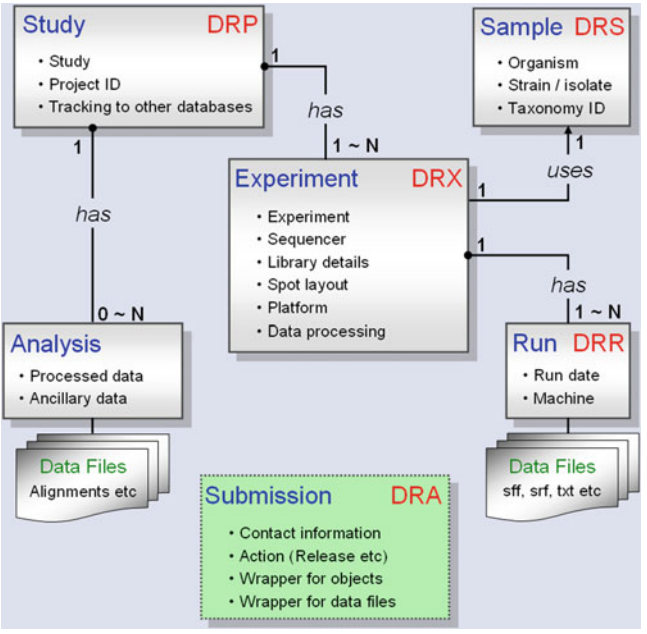


Fig. 15.1 DRA metadata model. Submission, Study, Experiment, Sample, Run, and Analysis objects are represented with their typical contents. DRA, DRP, DRX, DRS and DRR indicate prefix of accessions assigned to each object. Relationships between objects are represented as *has* and *uses* with the allowable number of each object. Data files contain sequences, quality, and other data

15.2.2 DRA Submission Systems

Now we are developing an online submission system, D-way, to facilitate large-scale data submissions. As a part of the system, we have released Microsoft Excel spreadsheets, the DRA sheets, for the submission of the metadata to DRA. By using the sheets, the submitters are able to create the metadata files by simply filling the necessary pieces of information in the fields of familiar Excel files. They can choose one of the DRA sheets prepared for three major platforms, 454, Genome Analyzer, and SOLiD. For the submitter's convenience, (1) every field in the DRA sheets is explained in a pop-up form, (2) the compulsory and optional fields are distinguished from each other by their color, (3) the fixed fields have pre-entered values, and (4) the DRA sheets contain an Excel macro to generate the metadata XML files by a single click, as shown in Fig. 15.2. The submitter can deposit the metadata either in the Excel file or in the XML file that can be validated by the DRA Meta Checker (http://trace.ddbj.nig.ac.jp/DRAMeta-Checker/contents/check.jsp). This checker first examines the uploaded XML files against the corresponding XML schema, and then factors that cannot be examined by the schema, such as the reference integrity in the XML documents and the consistency between the entered values, by referring to the relevant external databases such as the taxonomy and project databases in NCBI.

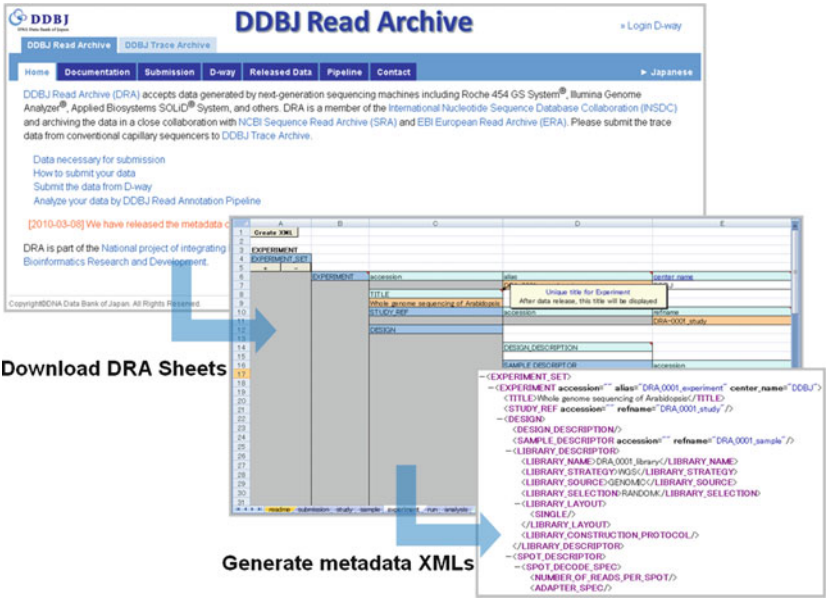


Fig. 15.2 DRA sheets. DRA sheets can be downloaded from the DRA website. Submitters can create their metadata files by filling the relevant pieces of information in the fields of the Excel file. They can easily generate metadata XML files by using the Excel macro, and can deposit their metadata either in the Excel file or in the XML file

The checker displays detailed error, warning, and usage messages after the validation process to help the submitter to correct their metadata accordingly by themselves.

For data transfer, the submitter can use the FTP service of DDBJ or record the data on a hard disk, and send it to DRA by a return-paid courier service. Once we receive the data, our DRA team validates the contents, issues accession numbers to the submitter and uploads the validated data to SRA by using Aspera (http://www.asperasoft.com). The DRA team works closely with large-scale sequencing centers to establish an automatic high-throughput submission pipeline between the centers and DRA.

By using D-way, the submitters will be able to send their metadata, confirm the status of the data processing and the accession numbers, specify the release date, and view the submitted data by logging with their accounts. We will improve our web submission system by integrating the validation and management of submission into a single system to make the submitters less demanding in the creation and submission of the metadata files.

15.2.3 DDBJ Read Annotation Pipeline: An Analytical Tool for DRA

We are also developing a pipeline, the DDBJ Read Annotation Pipeline, by which to process raw sequencing reads submitted to DRA and simultaneously generate submission files for DDBJ and CIBEX, because the pipeline helps submitters to analyze easily huge number of reads and submit processed data to DDBJ and CIBEX in addition to submitting the raw data to DRA. The pipeline consists of two parts, a basic and a high-level analysis part, as shown in Fig. 15.3. The basic part processes mapping the reads on a given reference genome and de novo assembly of the reads by using FASTQ file (sequence and quality) as an input, and the high-level part combines the automatic and manual annotations of the reads for SNP detection, expression tag counting, and others (Fig. 15.3). This pipeline has the following three characteristics. First, it can generate submission files so that the submitters can readily register the processed sequence data at DDBJ or processed quantitative data at CIBEX, and add the mapping/assembly results as an Analysis object to the DRA Submission object. Second, this pipeline can process millions of reads efficiently and effectively on a CIB-DDBJ cluster computing system. Third, users can select appropriate analytical tools from multiple candidates by manipulating in graphical user interface only.

As a preliminary step of the development, analytical tools for SNP detection have been implemented to the high-level analysis part. Currently, this pipeline has the automatic annotation system only, but a manual annotation step is indispensable to improve automatically annotated results [10]. To support this manual step, a user support function will be added to the pipeline.

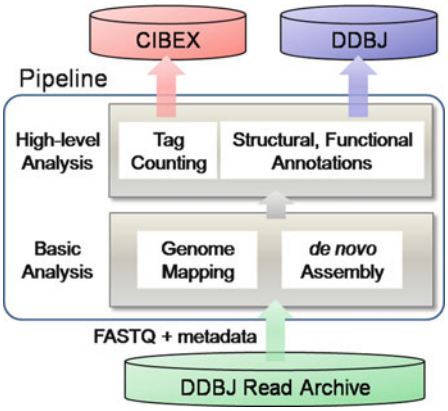


Fig. 15.3 DDBJ Read Annotation Pipeline interconnects DRA, DDBJ, and CIBEX. The pipeline processes the FASTQ and metadata files deposited in DRA as an input. In the basic analysis part, the pipeline maps the reads on the reference genome and assembles the reads de novo. In the high level analysis part, the pipeline performs sequence tag counting, and structural and functional annotations by using results of the basic analysis. The resulting files can readily be submitted to CIBEX (quantitative data) and DDBJ (annotated sequence data)

15.3 CIBEX

15.3.1 CIBEX: Array-Based Data

The DNA microarray technology was developed in the late 1990s [11] and is widely used to measure gene expression levels, as the number of sequenced genome increases. Microarray data were first stored in several databases in their original formats, making it difficult to compare the results from different experiments and platforms. To solve this problem and to improve data sharing and usage by establishing a standard data presentation, MGED was organized at EBI in 1999. In 2001, the MGED society published the first version of guideline called MIAME [5]. At present, many scientific journals require the deposition of MIAME-compliant microarray data to GEO, ArrayExpress, or CIBEX before the publication of the relevant paper [7, 8]. As microarray technology is improved and used in epigenetics, SNP and genomic variation studies, and others, the three archives have extended their ranges to accommodate these functional genomics data.

CIBEX has accepted and distributed microarray data compliant with MIAME since 2004 [4]. The contents of the data are composed of the metadata and intensity data. The metadata are given in a line-based simple text file format containing the information about the contact, publication, experiment, microarray type, protocol, sample, sample data relationship, data processing, and others, to enable users to interpret the overall experiment and design. The data of microarray design, and raw

and processed intensity data of hybridization results are provided in a tab-delimited text file with a header describing each field. This simple, flexible text file format allows the submitter to add any pieces of information by adding columns. CIBEX has so far released over 1,800 hybridizations. Most data were submitted from Japan, and the number of submissions to CIBEX steadily increases.

CIBEX provides a stand-alone submission tool so that the submitter can create a MIAME-compliant submission file by simply entering pertinent information in each field in the file and specifying the intensity data files, and send the submission file to CIBEX online. CIBEX annotators then review the contents and if they find no problems, a CIBEX accession number is issued with prefix "CBX." On the contrary, users can carry out keyword search against the metadata, view the list of experiments and arrays, and download the search result in a tab-delimited text file format.

15.3.2 CIBEX: Sequencing-Based Data

The next-generation sequencing platforms are gradually replacing the DNA microarray in measuring molecular abundance at the genomic level, because the new platforms can count the molecules with much higher scale and accuracy without cross-hybridization and background. To accommodate the data produced by the new platforms or the ultra high-throughput sequencing (UHTS) data, the MGED society has proposed the Minimum Information about a high-throughput Nucleotide SEQuencing Experiment (MINSEQE) guideline for standardizing UHTS data (http://www.mged.org/minseqe). GEO and ArrayExpress have started to accept UHTS data in compliance with MINSEQE. CIBEX will soon follow them. As next-generation sequencers are increasingly used to quantify molecules, researchers submit their raw data to DRA and processed data to CIBEX. Both GEO and ArrayExpress developed the submission brokering systems to SRA and ERA, respectively. For example, if researchers submit two sets of submission files (GEO metadata with processed data and SRA metadata with raw data) to GEO, GEO automatically registers the raw data to SRA, and links both data by the GEO and SRA accession numbers. CIB-DDBJ will develop a brokering system similar to that of GEO's for CIBEX and DRA. In addition, the DDBJ Read Annotation Pipeline mentioned above will process DRA-registered raw data to sequence reads with counts and transfer the processed data to CIBEX.

GEO, ArrayExpress, and CIBEX have not fully exchanged their data among them. So far, ArrayExpress has imported Affymetrix and Agilent microarray data from GEO [12]. However, GEO and ArrayExpress started to exchange their UHTS data in 2009. The two databases also prepare for exchanging the microarray data by mapping the GEO and ArrayExpress metadata. CIB-DDBJ has decided to join this international collaboration to realize the world-wide data sharing of the quantitative functional genomics data.

15.4 Future Direction

We are now in the era of the next-generation ultra high-throughput technologies and enjoy huge amounts of genomic scale data produced in many fields of life sciences. As one of the three major public biological databases, we at CIB-DDBJ collect, archive, and provide various kinds of the primary data in cooperation with the data submitters and database users. In particular, we continuously run four primary public databases, DDBJ, DDBJ Trace Archive (DTA), DDBJ Read Archive (DRA), and CIBEX, in collaboration with NCBI and EBI. For data produced by the traditional capillary-type sequencers, DTA, DDBJ and CIBEX have collected the raw, processed sequence, and quantitative data, respectively. On the contrary, for data produced by the next-generation platforms, DRA, DDBJ, and CIBEX collect the raw, processed sequence, and quantitative data, respectively. Because the UHTS data are huge in size, it is critical that the latter three databases are interconnected by the DDBJ Read Annotation Pipeline that deals with both data submission and analysis. Once the submitters input their raw instrument data with DRA metadata into the pipeline, they will be able to obtain the processed data in file formats that can be readily submitted to DDBJ and CIBEX.

One problem with the pipeline is due to the data formats of the three databases. For historical reasons, the metadata formats of DDBJ (flat file), CIBEX (line-based text file), and DRA (XML file) are different from one another. While some part among the three databases is common, the other is database specific. Therefore, it is desirable to have a system by which the metadata submitted in one of the three databases are automatically transferred to the other two metadata. We will develop such a conversion tool of the metadata formats among DDBJ, CIBEX, and DRA.

CIB-DDBJ has not been in international collaboration with GEO and ArrayExpress, but will soon participate in it. CIB-DDBJ, EBI, and NCBI will together construct public world-wide archives for raw and processed sequences and quantitative UHTS data by data sharing to accelerate research in life sciences.

Acknowledgments We gratefully acknowledge the support of all members of CIB DDBJ. In particular, we thank Takako Mochizuki for constructing the pipeline. We also especially thank Daisuke Fukuda, Keisuke Yamamoto, Wataru Kodachi, and Masahiro Fujimoto for DRA development; Toshinori Yagi for the development of CIBEX; and Drs. Takashi Gojobori, Toshi hisa Takagi, and Kousaku Okubo for their supports and understandings.

DDBJ is funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan with a management expenses grant for national university cooperation. The DDBJ Read Archive and DDBJ Trace Archive are supported partially by the Integrated Database Project (http://lifesciencedb.mext.go.jp/en) by MEXT and by the Institute for Bioinformatics Research and Development, Japan Science and Technology Agency.

References

1. Tateno Y, Gojobori T (1997) DNA Data Bank of Japan in the age of information biology. Nucleic Acids Res 25:14 17.

2. Sugawara H, Ikeo K, Fukuchi S, Gojobori T, Tateno Y (2009) DDBJ dealing with mass data produced by the second generation sequencer. Nucleic Acids Res 37:D16 D18.
3. Kaminuma E, Mashima J, Kodama Y, Gojobori T, Ogasawara O, Okubo K, Takagi T, Nakamura Y (2010) DDBJ launches a new archive database with analytical tools for next generation sequence data. Nucleic Acids Res 38:D33 D38.
4. Ikeo K, Ishi i J, Tamura T, Gojobori T, Tateno Y (2003) CIBEX: center for information biology gene expression database. C R Biol 326:1079 1082.
5. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze Kremer S, Stewart J, Taylor R, Vilo J, Vingron M (2001) Minimum information about a microarray experiment (MIAME) toward standards for microarray data. Nat Genet 29:365 371.
6. Brunak S, Danchin A, Hattori M, Nakamura H, Shinozaki K, Matise T, Preuss D (2002) Nucleotide sequence database policies. Science 298:1333.
7. Editorial (2002) Microarray standards at last. Nature 419:323.
8. Ball CA, Brazma A, Causon H, Chervitz S, Edgar R, Hingamp P, Matese JC, Parkinson H, Quackenbush J, Ringwald M, Sansone, SA, Sherlock G, Spellman P, Stoeckert C, Tateno Y, Taylor R, White J, Winegarden N (2004) Submission of microarray data to public repositories. PLoS Biol 2:1276 1277.
9. Ansorge WJ (2009) Next generation DNA sequencing techniques. Nat Biotechnol 25:195 203.
10. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, Fujishiro T, Katoh M, Kohara M, Kishida Y, Minami C, Nakayama S, Nakazaki N, Shimizu Y, Shinpo S, Takahashi C, Wada T, Yamada M, Ohmido N, Hayashi M, Fukui K, Baba T, Nakamichi T, Mori H, Tabata S (2008) Genome structure of the legume, *Lotus japonicas*. DNA Res 15:227 239.
11. Brown PO, Botstein D (1999) Exploring the new world of the genome with DNA microarrays. Nat Genet 21:33 37.
12. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A, Holloway E, Lukk M, Malone J, Mani R, Pilicheva E, Rayner TF, Rezwan F, Sharma A, Williams E, Bradley XZ, Adamusiak T, Brandizi M, Burdett T, Coulson R, Krestyaninova M, Kurnosov P, Maguire E, Neogi SG, Rocca Serra P, Sansone SA, Sklyar N, Zhao M, Sarkans U, Brazma A (2009) ArrayExpress update from an archive of functional genomics experiments to the atlas of gene expression. Nucleic Acids Res 37: D868 D872.

Part II
Microarray, Gene Expression Analysis, and Gene Regulatory Networks

Chapter 16
Comparison of Microarray Preprocessing Methods

K. Shakya, H.J. Ruskin, G. Kerr, M. Crane, and J. Becker

Abstract Data preprocessing in microarray technology is a crucial initial step before data analysis is performed. Many preprocessing methods have been proposed but none has proved to be ideal to date. Frequently, datasets are limited by laboratory constraints so that the need is for guidelines on quality and robustness, to inform further experimentation while data are yet restricted. In this paper, we compared the performance of four popular methods, namely *MAS5*, *Li & Wong pmonly* (*LWPM*), *Li & Wong subtractMM* (*LWMM*), and Robust Multichip Average (*RMA*). The comparison is based on the analysis carried out on sets of laboratory-generated data from the Bioinformatics Lab, National Institute of Cellular Biotechnology (NICB), Dublin City University, Ireland. These experiments were designed to examine the effect of Bromodeoxyuridine (5-bromo-2-deoxyuridine, *BrdU*) treatment in deep lamellar keratoplasty (*DLKP*) cells. The methodology employed is to assess dispersion across the replicates and analyze the false discovery rate. From the dispersion analysis, we found that variability is reduced more effectively by *LWPM* and *RMA* methods. From the false positive analysis, and for both *parametric* and *nonparametric* approaches, *LWMM* is found to perform best. Based on a complementary *q*-value analysis, *LWMM* approach again is the strongest candidate. The indications are that, while *LWMM* is marginally less effective than *LWPM* and *RMA* in terms of variance reduction, it has considerably improved discrimination overall.

Keywords Dispersion analysis · False positives · Microarray · Nonparametric · Parametric · Preprocessing

K. Shakya (✉)
Dublin City University, Dublin 9, Ireland
e mail: kabita.shakya@gmail.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 16,

16.1 Introduction

Microarray technology allows the monitoring of expression levels of thousands of genes, simultaneously, which in turn helps to explore gene sequence information and ultimately gene function(s). Since microarray gene expression data are characterized by high dimensionality and noisiness, the initial steps of microarray experiments are very crucial in terms of feeding "clean" data to the downstream analysis, namely, identification of gene expression patterns.

An important initial step in microarray technology is data preprocessing. Preprocessing removes systematic errors between arrays, introduced by labeling, hybridization and scanning. Several such methods have been developed so far [8], but this paper focuses on four popular methods which comprise basic tools for much experimental work. These are: *MAS5* (Microarray Suite version 5, core to the Affymetrix system [4] providing instrument control, data acquisition and analysis for the entire genechip platform) [5], two alternatives of the *Li & Wong* method, *Li & Wong pmonly* (*LWPM*) and *Li & Wong subtractMM* (*LWMM*), and Robust Multichip Average (RMA). *RMA* consists of three steps: a background adjustment, quantile normalization, and finally summarization [2]. *LWPM* method ignores the Mismatch probe intensities (MM) while *LWMM* uses the Perfect Match (PM) *MM* value to adjust the nonspecific binding (NSB) during background adjustment [15]. These methods are compared on the basis of dispersion across replicates [2, 10] distinguished as explained in Sect. 16.2.1 and also in terms of false discovery rates.

16.2 Materials and Methods

The dataset used for this comparison is laboratory-generated (experiments performed: Bioinformatics Lab, National Institute of Cellular Biotechnology (NICB), Ireland). The experiments were performed on Affymetrix GenechipTM, Human Genome U133 set (HG-U133A) and were designed to investigate patterns of gene expression changes in deep lamellar keratoplasty (*DLKP*) cells treated with thymidine analog (5-bromo-2-deoxyuridine, *BrdU*), during three different periods, 0 (control), 3, and 7 days. For each time point, three microarrays were used [9]. The dataset is thus modest, but typical of experiments targeted to exploratory analysis.

16.2.1 Dispersion Analysis

Dispersion analysis is used to assess the ability of each preprocessing method to reduce systematic error introduced during the treatment stage. A method giving large dispersion implies that, at the analysis stage, some genes are falsely declared

to be differentially expressed, and vice versa. The approach is commonly based on two precision criteria:

(i) The ability to minimize differences in pairwise comparisons between the arrays of replicates: Theoretically, genes are not differentially expressed across replicates, but should produce similar values. *MA*-plots [12] are used here for pairwise comparisons, as these conveniently illustrate the distribution of intensity values and log ratios and can give a quick overview of the data. In *MA*-plots, methods which minimize the distance between the loess curve [3] and the $M = 0$ line are considered optimal, as a gene is less likely to be falsely declared as differentially expressed in such cases. The four preprocessing methods were applied to the datasets for comparison using this criterion. For each of the four methods, nine *MA*-plots were produced, (for three replicates at three time points). The absolute distance between the line $M = 0$ and the loess curve was measured for every intensity. Finally, the mean was calculated for each intensity across the nine comparisons. The loess function becomes prohibitively time consuming for datasets corresponding to more than 20,000 probes, and it was not possible to include all 506,944 probes on the array used in experiment. Hence, we chose to apply the process for ten random samples of 5,000 probes each, and then to calculate the mean of the ten vectors. The final mean vector was found to be almost the same for the ten random samples, which argues for good consistency. Averaging over a number of samples also improved reliability even though excessive smoothing can obscure finer details.

(ii) The precision of the expression measures, estimated by the standard deviation across the replicates: After preprocessing, most noise would be removed from the data, and the biological replicates should have similar values. Thus, for a given method, high residual standard deviation across the replicates (for each gene, and/or time point) implies poor reliability. The following process was applied: for each method and gene, at each time point, the standard deviation and the mean were calculated across the replicates. To investigate the behavior of the standard deviation of the mean, a loess curve based on these calculated values was fitted in order to visualize the trend in the data.

16.2.2 False Positive Analysis

Microarray experiments measure expression values of thousands of genes simultaneously. Many of these are not in fact expressed, but repeated statistical testing at levels of significance ($\alpha \sim 0.05$) can lead to a large number of false positives. The number of false positives generated after preprocessing is used as a second criterion for comparison and measures the specificity of the preprocessing technique.

Procedures of testing for similar levels of expression between any two genes may be either *parametric* or *nonparametric* with use typically dependent on sample size. With a smaller number of replicates, the assumption of normality is less

robust. A *nonparametric* approach such as *Wilcoxon* [11] or *Mann Whitney*, while not reliant on distributional assumptions of sample measurements, has less ability to distinguish between methods and a larger number of replicates are desirable. In a typical laboratory investigation, practical restrictions can apply so that frequently the approach is to perform complementary analysis as an internal checking procedure. Consequently, some reliance is placed on the ability of preprocessing techniques to detect poor quality results.

In the work presented here, false positive analysis is based on *FWER* and *FDR* measures. The family-wise error rate (*FWER*) [10] is defined as the probability of occurrence of at least one false positive (V) over all true null hypotheses (corresponds to no relationship between gene expression measurement and response variable). Thus,

$$\text{FWER} = pr(V \geq 0) \tag{16.1}$$

The one-step *Bonferroni* method [5], together with the sequential *Westfall & Young* (*maxT*) adjusted p-value method [14], were used to estimate the *FWER* statistics. The advantage of the latter is that it takes the dependence structure between genes into account, giving improved power in many cases (e.g., where there is positive dependence between genes).

The false discovery rate (*FDR*), is an alternative and less stringent concept of error control, which is defined to be the expected proportion of erroneously rejected null hypotheses (Type I errors) among all those rejected.

$$\text{FDR} = E[V/R|R > 0]\ Pr\,(R > 0) \tag{16.2}$$

Controlling the *FDR* is desirable, since the quantity controlled is more directly relevant than that for the *FWER* (i.e., statistical power is improved for the former).

False-positive analysis was carried out, using libraries available with the *R* "*Bioconductor*" software: *multtest* and q-value. The *multtest* package implements multiple testing procedures for controlling different Type I error rates, for *FWER* and *FDR*. For our analysis, the procedures used to control *FWER* and *FDR* were (a) *Bonferroni* and *maxT* and *proc-2* correction [1].

The *pFDR* [13] is the expected proportion of Type I error among the rejected hypotheses when the number of rejected hypotheses is strictly positive.

$$p\text{FDR} = E(V/R|R > 0) \tag{16.3}$$

The *pFDR* can be interpreted in Bayesian terms. If we wish to perform m identical tests of a null hypothesis vs an alternative hypothesis based on the statistics $T_1, T_2, \ldots, T_m$, for a given significance region Γ,

$$p\text{FDR}(\Gamma) = E[V(\Gamma)/R(\Gamma)|R(\Gamma) > 0] \tag{16.4}$$

Associated with the *pFDR*, an adjusted p-value known as the q-value is defined. The q-value gives each hypothesis test a significance in terms of a given error rate. It measures the minimum false discovery rate that is incurred when the test result is deemed significant. For an observed statistic $T = t$, the q-value of t is:

$$q\text{-value}(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} p\text{FDR}(\Gamma_\alpha) \tag{16.5}$$

where $\{\Gamma_\alpha\}$ is the set of nested rejection regions.

Equation (16.5) indicates that the q-value is a measure of the strength of an observed statistic, with respect to *pFDR*; it is the minimum *pFDR* that can occur when rejecting a statistic with value t for the set of nested significance regions. Thus, it relates to false discovery rate.

The q-value module of *Bioconductor* has been used here for *pFDR* analysis, for both *parametric* and *nonparametric* estimation. The q-value package functions permit computation and presentation of q-values for multiple comparison features.

16.3 Results

16.3.1 Dispersion Analysis

Figure 16.1a illustrates the results of analysis for the minimization of differences across replicates (criterion (i), Sect. 16.2.1). Both the *RMA* and *LWPM* methods perform better than *MAS*5 and *LWMM* in which they minimize the variability around the $M = 0$ line and show potential for relatively few genes to be falsely declared as differentially expressed. The same hierarchy of performance across

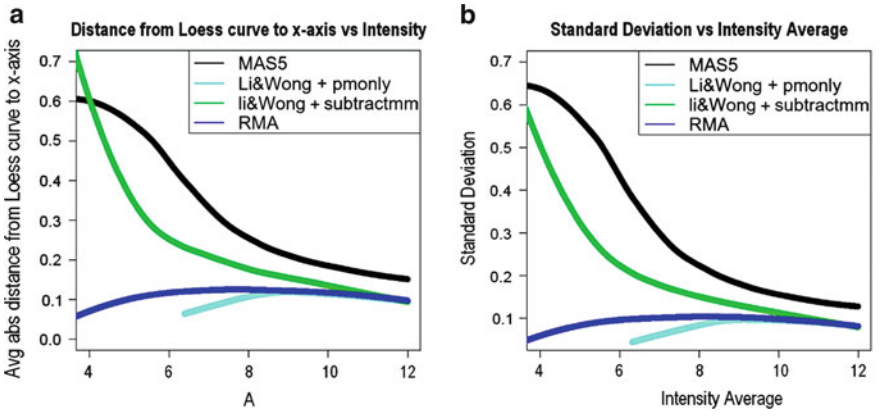


Fig. 16.1 (**a**, **b**) Comparison of the four preprocessing methods based on the two dispersion criteria

these methods is found in Fig. 16.1b, corresponding to the analysis under criterion (ii), Sect 16.2.1, and also in Fig. 16.2. Better performing methods both use the *PM* Correction, which may explain the effect observed.

To obtain a numerical estimate of the relative precision, linear models were fitted between replicates, (three pairwise comparisons between the three replicates), for each time point and for each method. From the 36 R^2 values obtained, an average was taken of the three values for each method and each time point. In Table 16.1, these values are summarized by averaging over time points and methods (mean row). Again, *RMA* and *LWPM* have the highest R^2 values, indicating that the amount of variation explained is higher compared to other methods and implying better reproducibility and precision. This is a crude measure of goodness of fit, but acts as our indicator of relative reliability.

16.3.2 False Positive Analysis

16.3.2.1 *Multtest* Results

Using the *multtest* library, analysis was performed on the number of rejected hypotheses as a function of the adjusted *p*-value, (calculated according to equations, Sect. 16.2.2). The objective is to highlight those preprocessing methods that have a

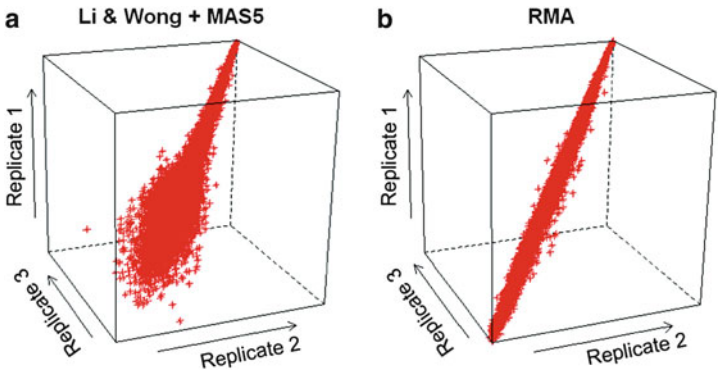


Fig. 16.2 Plot of the three replicates on the *x*, *y*, *z* axis for the four methods of Table 16.1. Note the high variability of differentially expressed genes for (**a**) and to a lesser extent for (**b**)

Table 16.1 Average R^2 associated with each time point and each method

Sample	*MAS*	*LWPM*	*LWMM*	*RMA*
Time 1	0.9265326	0.9904766	0.9724816	0.9952248
Time 3	0.9165260	0.9857380	0.9603595	0.9930388
Time 7	0.9080216	0.9649691	0.9434947	0.9850423
Mean	0.9170267	0.9803946	0.9587786	0.9911020

minimum number of false positives for a given number of differentially expressed genes, that is, to find those methods, which have a high number of differentially expressed genes for a large adjusted *p*-value range.

The number of rejected hypotheses is larger for the *FDR* method as compared to the *FWER* (Fig. 16.3), in agreement with the view that the *FWER* criterion is more conservative than the *FDR*. Nevertheless, better discrimination between preprocessing methods is achieved for the *FWER* criterion, with *RMA* and *LWMM* able to identify considerably more differentially expressed genes than *MAS*5 and *LWPM* for a given adjusted *p*-value. For *FDR* procedures, *LWMM* is slightly improved, whereas *LWPM* performs slightly worse. Curves for *MAS*5 and *RMA* overlap over partial ranges so that real differences between methods are small.

Comparing the results of *parametric* and *nonparametric* approaches showed that *Wilcoxon* + *MaxT* correction and *Wilcoxon* + *Bonferroni* correction (*FWER* criterion) produced the same shape curves. Similarly, *Wilcoxon* + *BH* and *Wilcoxon* + *Proc_2* (*FDR* criterion). In general, the *nonparametric* approach is unsatisfactory in terms of discrimination, especially for the *FWER* criterion, as all preprocessing methods produce similar curves. For the two *FDR* procedures, *LWMM* is slightly improved, whereas *RMA*, which performed well in *parametric* analyses, is slightly worse. Given the closeness of the curves, however, the difference is not marked.

Summary Results of False Positive Analyses

Figure 16.4 summarizes the results of the false positive correction procedures, Sect. 16.3.2.1. For each of the four false positive correction procedures, (two for each of *FWER* and *FDR*) and for each *p*-value, the four values (number of rejected hypotheses, corresponding to the four preprocessing methods), are divided by the largest of these so that hierarchy and relative variation between the four are preserved (normalization). Then for each *p*-value, and for each of the four preprocessing methods, the normalized values (for four positive-correction approaches) are averaged. This mean value is used to plot mean percentage of the best value for the four false positive correction procedures vs adjusted *p*-value.

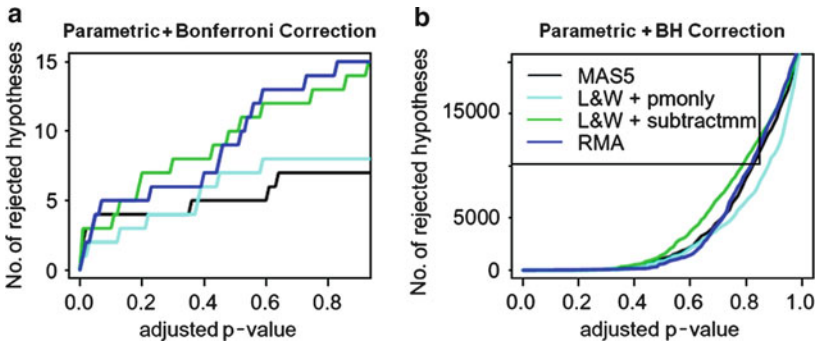


Fig. 16.3 *Parametric* approach

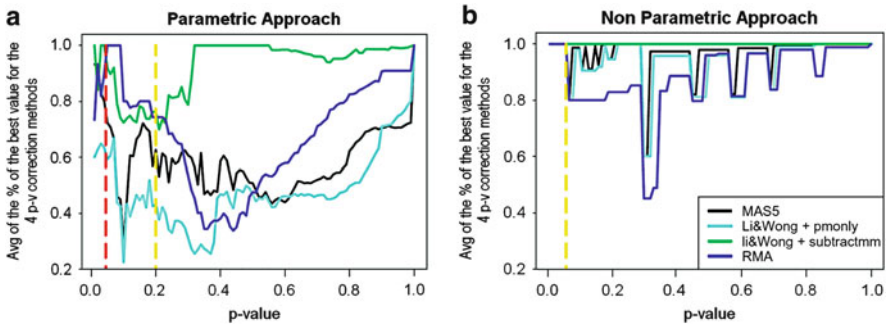


Fig. 16.4 Summarization of false positive correction procedures with (**a**) the *parametric* approach and (**b**) the *nonparametric* approach

Summary results are shown in Fig. 16.4a for *parametric* and Fig. 16.4b for *nonparametric* approaches. Based on these, *LWMM* outperforms the other preprocessing methods for $0.2 \leq p < 1$ and, somewhat surprisingly, for very low p-value $0 < p \leq 0.045$ for the *parametric* approach, and for $0.06 \leq p < 1$ for *non-parametric* approach: (*LWMM* performance is distinctly improved here). For midrange $0.045 \leq p \leq 0.2$, *RMA* outperformed *LWMM* in the *parametric* case. The best preprocessing method is indeterminate, however, for $0 \leq p \leq 0.06$ (*nonparametric*). The 0.06 threshold here is possibly due to the low number of replicates for the *Wilcoxon* test.

16.3.2.2 Results Based on *q*-Value Library

The q-value library of *Bioconductor* was used to estimate *pFDR* (Storey 2008). Results are similar to those for the *multtest* library in both *parametric* and *nonparametric* cases. All four preprocessing methods give similar shaped curves; however, *LWMM* and *RMA* performed best for $0 < q \leq 0.6$ and $q \geq 0.6$, respectively (*parametric* case). For the *nonparametric*, *LWMM* was best across all q-values, implying that a higher number of differentially expressed genes are captured.

16.4 Conclusion

An analysis of four different microarray preprocessing methods, namely, *MAS5*, *LWPM*, *LWMM,* and *RMA*, was performed with respect to their dispersion and false positive analysis. Dispersion comparisons indicate that technical variability is addressed more effectively by *LWPM* and *RMA* methods: (supported by the results of the fitted linear model and R^2, Coefficient of determination measures).

Comparison of false positive rates in ranges $0.045 \leq p \leq 0.2$ and $p > 0.2$ indicates *RMA* and *LWMM*, respectively, performed best (*parametric* case). *LWMM* also outperformed other methods, $0 < p \leq 0.045$ (possibly due to small

sample size). For $0 < p \leq 0.06$ (*non-parametric*), all methods performed equivalently well with *LWMM* for $p \geq 0.06$. In q-value tests for *pFDR* analysis, *LWMM* outperformed *RMA* (for *nonparametric* and $0 < q \leq 0.06$, *parametric*), supporting previous analyses.

Given that sample size is relatively small for these data, that is methods are less robust for small p, results of *false positive* analysis indicate *LWMM* to be the *best preprocessing method*. Based on *dispersion* analysis results, *LWPM outperformed* other methods. While the choice of method may depend on analysis purpose, these control the two very relevant experimental criteria (not necessarily of equal importance in an investigation) so that one is typically preferred. Consequently, results presented here, even for modest-sized dataset, do provide some distinct guidelines on preprocessing method choice for microarray data.

References

1. Benjamini, Y., Krieger, A., Yekutieli, D. Two staged Linear Step Up FDR Controlling Procedure, Technical Report, Tel Aviv University and Department of Statistics, Wharton School, University of Pennsylvania (2001)
2. Bolstad, B. M., Irizarry, R. A., Astrand, M., Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformat ics, **19**(**2**), 185 193 (2003)
3. Cleveland, W.S. Visualizing Data, Summit, New Jersey: Hobart Press (1993)
4. Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., Speed, T. P. A benchmark for affymetrix genechip expression measures. Bioinformatics, **20**(**3**), 323 331 (2004)
5. Gordon, A., Glazko, G., Qiu, X., Yakovlev, A. Control of the mean number of false discoveries, Bonferroni and stability of multiple testing. Ann. Appl. Statist., **1**(**1**), 179 190 (2007)
6. Hubbell, E., Liu, W. M., Mei, R. Robust estimators for expression analysis. Bioinformatics, **18** (**12**), 1585 1592 (2002)
7. Irizarry, R. A., Hobbs, B., Collin F., Beazer Barclay Y.D., Antonellis K. J., Scherf U., Speed, T. P. Exploration, normalization and summaries of high density oligonucleotide array probe level data. Biostatistics, **4**(**2**), 249 264(2003)
8. Mastrogianni A., Dermatas E., Bezerianos A. Robust pre processing and noise reduction in microarray images. Proceeding (555), Biomedical Engineering (2007)
9. McMorrow, J. Ph.D. thesis. Dublin City University, Ireland (2006)
10. Novak, J. P., Kim, S. Y., Xu, J., Modlich, O. et al. : Generalization of DNA microarray dispersion properties: microarray equivalent of t distribution. Biol Direct, **1**(**27**), doi:10.1186/ 1745 6150 1 27, (2006)
11. Speed, T. P. Statistical Analysis of Gene Expression Microarray Data. CRC Press, ISBN 1584883278, 9781584883272
12. Stafford, P. (Ed.) Methods in Microarray Normalization (Drug Discovery Series), USA, CRC Press, ISBN 1420052780, 9781420052787
13. Storey, J. D. The positive false discovery rate: a Bayesian interpretation and the q value. Ann. Statist., **31**(**6**), 2013 2035 (2001)
14. Westfall, P. H., Young, S. S. Resampling Based Multiple Testing: Examples and Methods for p Value Adjustment. Wiley, England (1993)
15. Wu, Z., Irizarry, R. A., Gentleman, R., Martinez Murillo, F., Spencer, F. A model based background adjustment for oligonucleotide expression arrays. J. Am. Stat. Assoc., **99**(**468**), 909 917 (2004)

Chapter 17
A Robust Ensemble Classification Method Analysis

Zhongwei Zhang, Jiuyong Li, Hong Hu, and Hong Zhou

Abstract Apart from the dimensionality problem, the uncertainty of Microarray data quality is another major challenge of Microarray classification. Microarray data contain various levels of noise and quite often high levels of noise, and these data lead to unreliable and low accuracy analysis as well as high dimensionality problem. In this paper, we propose a new Microarray data classification method, based on diversified multiple trees. The new method contains features that (1) make most use of the information from the abundant genes in the Microarray data and (2) use a unique diversity measurement in the ensemble decision committee. The experimental results show that the proposed classification method (DMDT) and the well-known method (CS4), which diversifies trees by using distinct tree roots, are more accurate on average than other well-known ensemble methods, including Bagging, Boosting, and Random Forests. The experiments also indicate that using diversity measurement of DMDT improves the classification accuracy of ensemble classification on Microarray data.

Keywords Classification methods · CS4 · Diversified multiple trees · Diversity measurement · Microarray classification · Microarray data

17.1 Introduction

The primary purpose of Microarray data classification is to build a classifier from the classified historical Microarray data, and then use the classifier to classify future incoming data or predict the future trend of data. Due to the advent of DNA Microarray technology, vast amount of DNA microarray datasets has been widely

Z. Zhang (✉)
Department of Mathematics and Computing, University of Southern Queensland, Toowoomba, QLD, Australia
e mail: zhongwei@usq.edu.au

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 17,

available for Microarray data classification. However, as a new technology, Microarrays present the following new statistical problems to the Microarray data classification:

1. *Curse of dimensionality problem*. Microarray data contain a huge number of genes with small number of samples and this problem has prevented many existing classification systems from direct dealing with this type of databases.
2. *Robustness problem*. In addition, DNA Microarray database contains high level of noise, irrelevant and redundant data, which will lead to unreliable and low accuracy of analysis. Most of the current systems are not robust enough to handle these types of data properly.

Ensemble decision tree classification methods [4, 6] have shown promise for achieving higher classification accuracy than single classifier classification method, such as C4.5 [8]. The essence of ensemble methods is to create diversified classifiers in the decision committee. Aggregating decisions from diversified classifiers is an effective way to reduce bias existing in individual trees. However, if the classifiers in the committee are not unique, the committee has to be very large to create certain diversity in the committee.

Up to date, all ensemble decision tree methods have been designed with diversity in mind [1, 9]. However, among those methods, most of them, such as Boosting and Bagging, do not guarantee that each ensemble decision tree in the committee is different from outputs, namely identical trees and overlapping genes are not prohibited from an ensemble committee. Identical trees decrease the diversity of an ensemble committee, and noise in one gene may affect a number of ensemble decision trees; the noise will ultimately affect the reliability of Microarray classification. Therefore, committees built on these methods may not be as effective as a committee that contains no identical trees and overlapping genes.

A quick fix to improve diversity in the ensemble decision tree committee is to include a set of diversified decision trees with no overlapping genes. If classifiers in the ensemble decision tree committee are not guaranteed to be different from each other, the committee must be very large, in order to create certain diversity in the committee. This behooves us to pay special attention while designing our algorithm. One concern for such a split is that it might break down some attribute combinations that are good for classification. However, an apparent benefit of such trees is that a noise attribute cannot affect more than one tree in the committee. Considering that Microarray data normally contain much noise and many missing values, the idea of using diversified trees with no overlapping genes may provide a better solution.

The rest of this chapter is organized as follows. In Sect. 2, we discuss the measurement of diversity. In Sect. 3, we introduce our diversified multiple decision tree algorithm (DMDT). In Sect. 4, we show our experimental results. In Sect. 5, we conclude the paper.

17.2 Measurement of Diversity

All ensemble decision tree classification methods generate a set of decision trees to form a committee. Due to the different approaches applied to generate the committee, the decision trees in the final ensemble committee could be diverse from each other in certain ways. In the past decades, measuring diversity has become a very important issue in the research of Microarray ensemble classification methods [1, 5, 9].

Measuring outputs is a most natural way to measure the diversity of ensemble classifiers [1]. The output from measuring the classifiers in a committee may give a result of total different, partially different, or identical classifiers. If the classifiers in a committee are all identical, we can say that these classifiers are not diversified; if the classifiers are partially different, we can say that they are diversified. When the classifiers are totally different or unique to each other, we say that the classifiers are maximally diversified.

Many statistical diversity measures are also available, such as diversity of errors [1, 2, 7], and pairwise and non-pairwise diversity measures [1, 5, 9]. It is desirable if every classifier in an ensemble committee can agree on most samples that are predicted correctly. At the same time, we also expect that they do not make the same incorrect predictions on testing samples. Those methods are also very important measurements of diversity, because if their errors were correlated, classification prediction would not lead to any performance gain by combining them.

The approach of measuring diversity based on statistical methods has drawbacks. There is a lack of robustness consideration in Microarray classification in terms of incorrect and missing data values. Identical trees are excluded from the ensemble committee since they are not helpful in improving the prediction accuracy of classification. However, this measurement allows overlapping genes among diversified trees. Overlapping genes are a problem for reliable Microarray data classification.

In our proposed method, diversity is measured by the difference of outputs for Microarray data classification problems. The degree of diversity is dependent on how many overlapping genes are included between the decision trees of an ensemble committee.

Definition 1 (Degree of diversity) *Given a data set D with n attributes, $A = \{att_1, \cdots, att_n\}$; C is an ensemble decision tree committee with $k (k > 1)$ individual decision trees generated from D, $C = \{c_1, \cdots, c_k\}$; $c_i \in C$ and $c_j \in C$ are any single decision trees; c_i contains a set of attributes $A_{c_i} = \{att \mid att \in A\}$ and $A_{c_j} = \{att \mid att \in A\}$;*

Let $|A_{c_i} \cap A_{c_j}|$ = the number of elements contained in $A_{c_i} \cap A_{c_j}$, and $|A_{c_i} \cup A_{c_j}|$ = the number of elements contained in $A_{c_i} \cup A_{c_j}$. Then the degree of diversity between c_i and c_j is

$$DD = 1 - \frac{|A_{c_i} \cap A_{c_j}|}{|A_{c_i} \cup A_{c_j}|} \quad (0 \leq DD \leq 1).$$

When an ensemble committee contains only decision trees which have totally different outputs, or unique trees with no overlapping genes, we say that the ensemble committee is maximally diversified. According to Definition 1, $DD = 1$ for the unique decision trees.

Definition 2 (Unique decision trees) *c_i and c_j are called unique decision trees if $A_{c_i} \cap A_{c_j} = \phi$.*

We say an ensemble decision tree classification method has greater diversity when its decision trees have a higher degree of different outputs with less overlapping genes. It is clear that diversified decision trees have a DD value between 0 and 1.

Definition 3 (Diversified decision trees) *If $A_{c_i} \neq A_{c_j}$ and $A_{c_i} \cap A_{c_j} \neq \phi$, then c_i and c_j are called diversified decision trees.*

Similarly, if all decision trees in an ensemble decision tree committee are identical, the degree of its diversity would be 0.

Definition 4 (Identical decision trees) *We call c_i and c_j are identical decision trees if $A_{c_i} = A_{c_j}$.*

17.3 Diversified Multiple Decision Trees Algorithm

We design a diversified multiple decision tree (DMDT) algorithm to deal with the problems of Microarray classification, namely, small samples versus high dimensions and noisy data. DMDT aims to improve the accuracy and robustness of ensemble decision tree methods. In our proposed algorithm, we avoid the overlapping genes among alternative trees during the tree construction stage. DMDT guarantees that the constructed trees are truly unique and maximizes the diversity of the final classifiers. Our DMDT algorithm is presented in Algorithm 1. The DMDT algorithm consists of the following two steps.

17.3.1 Tree Construction

The main idea is to construct multiple decision trees by re-sampling genes. All trees are built on all of the samples but with different sets of genes. We conduct re-sampling data in a systematic way. First, all samples with all genes are used to build the first decision tree. The decision tree is built using the C4.5 algorithm. After the decision tree is built, the used genes appearing in the decision tree are removed from the data. All samples with the remaining genes are used to build the second decision tree. Then the used genes are removed and so on. This process

repeats until the number of trees reaches a preset number. As a result, all trees are unique and do not share common genes.

Algorithm 1: *Diversified multiple decision trees algorithm (DMDT).*

1. TREECONSTRUCTION (D, $\mathcal{T}$, DD, n)

INPUT: A Microarray dataset D, the degree of diversity DD and the number of trees n.

OUTPUT: A set of disjointed trees $\mathcal{T}$

let $\mathcal{F} = \phi$

let $DD = 1$

for $i = 0$ *to* $n - 1$ **do**

call C4.5 to build tree T_i on D;

remove genes used in T_i from D;

$\mathcal{T} = \mathcal{T} \cup T_i$.

endfor

Output $\mathcal{T}$;

2. *CLASSIFICATION*($\mathcal{T}$, x, n)

INPUT: A set of trained trees $\mathcal{T}$, a test sample x, and the number of trees n.

OUTPUT: A class label of x

let $vote(i) = 0$, where $i = 1$ *to* c = the number of classes.

for $j = 1$ *to* n **do**

let c be the class outputted by T_j;

$vote(c) = vote(c) \times accuracy(T_j)$;

endfor

Output c that maximizes $vote(c)$;

17.3.2 Classification

Since the k th tree has only used the genes that have not been selected by the previously created $k - 1$trees, the quality of the k th tree might be decreased. To fix this problem, we take a vote approach; that is to say, the final predicted class of an unseen sample is determined by the weighted votes from all constructed trees. Each tree is given the weight of its training classification accuracy rate. When the vote is a tie, the class predicted by the first tree is preferred. Since all trees are built on the original data set, all trees are accountable on all samples. This avoids the unreliability of voting caused by sampling a small data set. Since all trees make use of different sets of genes, the trees are independent. This adds another merit to this diversified committee. One gene containing noise or missing values affects only one tree, and not multiple trees. Therefore, it is expected to be more reliable in Microarray data classification where noise and missing values prevail.

17.4 Experimental Results and Discussion

In this section, we first present the accuracy of individual methods and the average prediction accuracy of the six methods [3, 4], which are all based on the tenfold validation technique [6, 10, 11]. Table 17.1 shows the individual and average accuracy results of the six methods based on the tenfold cross-validation method.

Our DMDT outperforms other ensemble methods. For instance, compared to the single decision tree, DMDT is a more favorable ensemble method and outperforms C4.5 by 10.0% on average.

From Table 17.1, we notice that CS4 also performs very well and improves the accuracy by 8.4% on average. Random Forests, AdaBoost C4.5, and Bagging C4.5 improve the accuracy on average by up to 4.3%. More specifically,

1. among the five ensemble methods used in our experiments, DMDT turns to be the most favorable classification algorithm with the highest accuracy, which improves the accuracy of classification on all cancer data sets by up to 26.7%.
2. CS4 is comparable to DMDT in the test which improves the accuracy of classification on all data sets by up to 17.4%.
3. Bagging C4.5 also outperforms C4.5 on all data sets by up to 9.6%.
4. Random Forests improves the accuracy on lung cancer, lymphoma, leukemia, and prostate data sets by up to 19.1%, but fails to improve the accuracy on breast cancer, colon, and ovarian data sets. AdaBoost C4.5 can only improve the accuracy on Lung cancer, lymphoma, and leukemia, and decreases the accuracy performance on the breast cancer and colon data sets.

It is interesting to see that traditional ensemble decision tree algorithms do not always outperform a single tree algorithm. This is because the traditional ensemble methods assume that a training data set has a large number of samples with small numbers of attributes. As a result, the re-sampled data set is only slightly different from the original data set. The trees constructed on these re-sampled data are still reliable. However, in Microarray data analysis, the problem that we are facing is completely the opposite: a small number of samples with large numbers of attributes (genes). As most Microarray data contain less than 200 samples, a slight change of samples may cause a dramatic structural change in the training data set. The trees constructed on such unreliable data sets are more likely to lead to higher

Table 17.1 Average accuracy of five data sets with six classification algorithms

Data set	C4.5	Random Forests	AdaBoost C4.5	Bagging C4.5	CS4	DMDT
Breast cancer	62.9	61.9	61.9	66.0	68.0	64.3
Lung cancer	95.0	98.3	96.1	97.2	98.9	98.9
Lymphoma	78.7	80.9	85.1	85.1	91.5	94.1
Leukemia	79.2	86.1	87.5	86.1	98.6	97.5
Colon	82.3	75.8	77.4	82.3	82.3	85.8
Average	79.62	80.6	81.6	83.3	87.9	88

risk of the problem of unreliability. This risk affects the performance of classification. In contrast, DMDT, and CS4 are designed specially for Microarray data analysis. DMDT keeps the alternative trees using all available samples in order to minimize the impact of the unreliability problem.

17.5 Conclusions

In this chapter, we studied the concept of diversity measurement in ensemble classifiers. We then proposed an algorithm that diversifies trees in the ensemble decision tree committee. We conducted experiments on six Microarray cancer data sets. We conclude that the proposed DMDT performs the best among all algorithms used in the experiments. DMDT is more resistant to the noise data while it has the highest classification accuracy rate. From the robustness point of view, Random Forests is comparable to DMDT and outperforms other compared algorithms. Without increase in the noise data level, CS4 is comparable to DMDT. However, its performance decreases while comparing with DMDT when the noise level increases in the training and test data.

References

1. Matti Aksela and Jorma Laaksonen. Using diversity of errors for selecting members of a committee classifier. Pattern Recognition, 39(4):608 623, 2006.
2. M. Brown, W. Grundy, D. Lin, N. Cristianini, C. Sugnet, T. Furey, M. Jr, D. Haussler. Knowledge based analysis of Microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences of the United States of America, 97:262 267, 2000.
3. T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. Machine Learning, 40(2):139 158, 1998.
4. Mordechai Gal Or, Jerrold H. May, William E. Spangler. Using decision tree models and diversity measures in the selection of ensemble classification models. In: Nikunj C. Oza, Robi Polikar, Josef Kittler, Fabio Roli, editors, Multiple Classifier Systems, Lecture Notes in Computer Science, 3541:186 195. Springer, 2005.
5. Ludmila I. Kuncheva, J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 51(2):181 207, 2003.
6. Jinyan Li, Huiqing Liu. Ensembles of cascading trees. In ICDM, 585 588, 2003.
7. Derek Partridge, Wojtek Krzanowski. Distinct failure diversity in multiversion software. Technical report, Dept. Computer Science, University of Exeter, sec@dcs.exeter.ac.uk, 1999.
8. J. R. Quinlan. C4.5: Programs for machine learning. Morgan Kaufmann, San Mateo, California, 1993.
9. Geoffrey I. Webb, Zijian Zheng. Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques. IEEE Transactions on Knowledge and Data Engineering, 16(8):980 991, 2004
10. C. Yeang, S. Ramaswamy, P Tamayo, et al. Molecular classification of multiple tumor types. Bioinformatics, 17(Suppl 1):316 322, 2001.
11. Heping Zhang, Chang Yung Yu, Burton Singer. Cell and tumor classification using gene expression data: Construction of forests. Proceedings of the National Academy of Sciences of the United States of America, 100(7):4168 4172, 2003

Chapter 18
k-NN for the Classification of Human Cancer Samples Using the Gene Expression Profiles

Manuel Martín-Merino

Abstract The *k*-Nearest Neighbor (*k*-NN) classifier has been applied to the identification of cancer samples using the gene expression profiles with encouraging results. However, the performance of *k*-NN depends strongly on the distance considered to evaluate the sample proximities. Besides, the choice of a good dissimilarity is a difficult task and depends on the problem at hand. In this chapter, we introduce a method to learn the metric from the data to improve the *k*-NN classifier. To this aim, we consider a regularized version of the kernel alignment algorithm that incorporates a term that penalizes the complexity of the family of distances avoiding overfitting. The error function is optimized using a semidefinite programming approach (SDP). The method proposed has been applied to the challenging problem of cancer identification using the gene expression profiles. Kernel alignment *k*-NN outperforms other metric learning strategies and improves the classical *k*-NN algorithm.

Keywords Metric learning · Combination of dissimilarities · Multiple kernel learning · Microarrays

18.1 Introduction

The *k*-Nearest Neighbor (*k*-NN) classifier has been widely applied to the identification of cancer samples using the gene expression profiles. However, *k*-NN relies strongly on the distance considered to evaluate the object proximities. The choice of a dissimilarity that reflects accurately the proximities among the sample profiles is a difficult task and depends on the problem at hand. Moreover, there is no optimal

M. Martín Merino
Computer Science Department, Universidad Pontificia de Salamanca, C/Compañía 5, 37002 Salamanca, Spain
e mail: mmartinmac@upsa.es

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 18,

dissimilarity in the sense that each dissimilarity reflects different features of the data and misclassifies frequently a different subset of patterns [2]. Therefore, different dissimilarities should be integrated in order to reduce the misclassification errors.

In this paper, we propose a method to combine a set of heterogeneous dissimilarities that reflect different features of the data. To this aim, a linear combination of dissimilarities is learnt considering the relation between kernels and distances. Each dissimilarity is embedded in a feature space using the Empirical Kernel Map [5]. Next, learning the dissimilarity is equivalent to optimize the weights of the linear combination of kernels. The combination of kernels is learnt in the literature [1, 3] maximizing the alignment between the input kernel and an idealized kernel. However, this error function does not take into account the generalization ability of the classifier and is prone to overfitting.

Our approach considers a regularized version of the kernel alignment proposed by [1]. The linear combination of kernels is learnt in a Hyper Reproducing Kernel Hilbert Space (HRKHS) following the approach of hyperkernels proposed in [8]. This formalism that exhibits a strong theoretical foundation is less sensitive to overfitting, and allows us to work with infinite families of distances.

The algorithm has been applied to the identification of human cancer samples using the gene expression profiles with remarkable results.

18.2 Metric Learning Using Regularized Kernel Alignment

To incorporate a linear combination of dissimilarities into k-NN, we follow the approach of hyperkernels developed by [8]. To this aim, each distance is embedded in a RKHS via the Empirical Kernel Map (see ref. [5] for details). Next, a regularized version of the alignment that incorporates a L_2-penalty over the complexity of the family of distances considered is introduced. The solution to this regularized quality functional is searched in an HRKHS. This allows to minimize the quality functional using a semidefinite programming approach (SDP).

Let $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_m\}$ be a finite sample of training patterns, where $y_i \in \{-1, +1\}$. Let K be a family of semidefinite positive kernels. Our goal is to learn a kernel of dissimilarities [5] $k \in K$ that represents the combination of dissimilarities and that minimizes the empirical quality functional defined by:

$$Q_{emp}^{align}(K, X, Y) = 1 - A(K, X, Y) = 1 - \frac{y^T K y}{m \parallel K \parallel_F}, \tag{18.1}$$

where K is the input kernel matrix and $A(K, X, Y)$ is the empirical alignment between the input kernel matrix K and an ideal kernel matrix $y^T y$. The empirical alignment has been introduced in [1] and evaluates the similarity between kernels.

However, if the family of kernels K is complex enough, it is possible to find a kernel that achieves training error equal to zero overfitting the data. To avoid this problem, we introduce a term that penalizes the kernel complexity in an HRKHS [8].

The quality functional optimized by the regularized kernel alignment can be written as:

$$Q_{reg}(k,X,Y) = Q_{emp}^{align}(K,X,Y) + \frac{\lambda_Q}{2} \parallel k \parallel_{\underline{H}}^2, \tag{18.2}$$

where $\parallel \parallel_{\underline{H}}$ is the L_2 norm defined in the HRKHS generated by the hyperkernel $\underline{k}$. λ_Q is a regularization parameter that controls the complexity of the resulting kernel. By the Representer Theorem [8], the kernel k that minimizes (18.2) can be written as a linear combination of hyperkernels.

$$k(x,x') = \sum_{i,j=1}^{m} \beta_{ij} \underline{k}((x_i,x_j),(x,x')) \tag{18.3}$$

However, we are only interested in solutions that give rise to positive semidefinite kernels. The following condition over the hyperkernels [8] allow us to guarantee that the solution is a positive semidefinite kernel.

Given a hyperkernel $\underline{k}$ with elements such that for any fixed $\underline{x} \in \underline{X}$, the function $k(x_p,x_q) = \underline{k}(\underline{x},(x_p,x_q))$ with $x_p,\ x_q \in X$ is a positive semidefinite kernel and $\beta_{ij} \geq 0$ for all $i,j = 1,\ldots,m$, then the kernel

$$k(x_p,x_q) = \sum_{i,j=1}^{m} \beta_{ij} \underline{k}(x_i,x_j,x_p,x_q) \tag{18.4}$$

is positive semidefinite.

Now, we address the problem of combining a finite set of dissimilarities. As we mentioned earlier, each dissimilarity can be represented by a kernel using the Empirical Kernel Map. Next, the hyperkernel is defined as:

$$\underline{k}(\underline{x},\underline{x}') = \sum_{i=1}^{n} c_i k_i(\underline{x}) k_i(\underline{x}') \tag{18.5}$$

where each k_i is a positive semidefinite kernel of dissimilarities, c_i is a constant ≥ 0, and n is the number of dissimilarities.

Now, we show that $\underline{k}$ is a valid hyperkernel: First, $\underline{k}$ is a kernel because it can be written as a dot product $\langle \underline{\Phi}(\underline{x}), \underline{\Phi}(\underline{x}') \rangle$, where

$$\underline{\Phi}(\underline{x}) = (\sqrt{c_1} k_1(\underline{x}), \sqrt{c_2} k_2(\underline{x}), \ldots, \sqrt{c_n} k_n(\underline{x})) \tag{18.6}$$

Next, the resulting kernel (18.4) is positive semidefinite because for all $\underline{x}, \underline{k}(\underline{x}, (x_p, x_q))$ is a positive semidefinite kernel and β_{ij} can be constrained to be ≥ 0. Besides, the linear combination of kernels is a kernel and therefore is positive semidefinite. Notice that $\underline{k}(\underline{x}, (x_p, x_q))$ is positive semidefinite if $c_i \geq 0$ and k_i is pointwise positive for training data. Both Laplacian and multiquadratic kernels verify this condition.

Finally, we show that the resulting kernel is a linear combination of the original k_i. Substituting the expression of the hyperkernel (18.5) in (18.4), the kernel is written as:

$$k(x_p, x_q) = \sum_{l=1}^{n} \left[c_l \sum_{i,j=1}^{m} \beta_{ij} k_l(x_i, x_j) \right] k_l(x_p, x_q) \tag{18.7}$$

The previous approach can be extended to an infinite family of distances. In this case, the space that generates the kernel is infinite dimensional. Therefore, in order to work in this space, it is necessary to define a hyperkernel and to optimize it using a HRKHS. Let k be a kernel of dissimilarities. The hyperkernel is defined as follows [8]:

$$\underline{k}(\underline{x}, \underline{x}') = \sum_{i=0}^{\infty} c_i (k(\underline{x}) k(\underline{x}'))^i. \tag{18.8}$$

where $c_i \geq 0$ and $i = 0, \ldots, \infty$. In this case, the nonlinear transformation to feature space is infinite dimensional. Particularly, we consider all powers of the original kernels that are equivalent to transform nonlinearly the original dissimilarities.

$$\underline{\Phi}(\underline{x}) = (\sqrt{c_1} k(\underline{x}), \sqrt{c_2} k^2(\underline{x}), \ldots, \sqrt{c_n} k^n(\underline{x})) \tag{18.9}$$

where n is the dimensionality of the space which is infinite in this case.

As for the finite family, it can be easily shown that $\underline{k}$ is a valid hyperkernel provided that the kernels considered are pointwise positive. The inverse multiquadratic and Laplacian kernels satisfy this condition. Evaluating the infinite sum, we can get an expression for the harmonic hyperkernel $\underline{k}(\underline{x}, \underline{x}') = (1 - \lambda_h)/(1 - \lambda_h k(\underline{x}) k(\underline{x}'))$ [8], where λ_h is a regularization parameter in the interval $(0, 1)$.

18.2.1 *Kernel Alignment* k-NN

We start with some notation that is used in the kernel alignment algorithm. For $p, q, r \in R^n$, $n \in N$, let $r = p \circ q$ be defined as element by element multiplication, $r_i = p_i \times q_i$. Define the hyperkernel Gram matrix $\underline{K}$ by $\underline{K}_{ijpq} = \underline{k}((x_i, x_j), (x_p, x_q))$, the kernel matrix $K = reshape(\underline{K}\beta)$(reshaping an m^2 by 1 vector, $\underline{K}\beta$, to an $m \times m$ matrix), where β are the linear coefficients in (18.3) that allow us to compute the kernel as a linear combination of hyperkernels. Finally, 1 is a vector of ones.

The optimization of the regularized quality functional (18.2) for the kernel alignment in a HRKHS can be written as:

$$\max_{k \in \underline{H}} \min_{\beta} \quad y^T K y + \frac{\lambda_Q}{2} \beta^T \underline{K} \beta \tag{18.10}$$

$$\text{subject to} \quad \beta \geq 0 \tag{18.11}$$

where we have considered that $Q_{emp}^{align} = y^T K y$ and according to the representer theorem $\| k \|_{\underline{H}}^2 = \beta^T \underline{K} \beta$. $\underline{K}$ denotes the hyperkernel matrix of dimension $m^2 \times m^2$ and λ_Q is a regularization parameter that controls the complexity of the family of kernels considered. The constraint $\beta \geq 0$ guarantees that the resulting kernel is positive semidefinite.

The optimization problem (18.10) can be easily solved using a SDP approach:

$$\min_{\beta} \quad \frac{1}{2} t_1 + \frac{\lambda_Q}{2} t_2 \tag{18.12}$$

$$\text{subject to} \quad \beta \geq 0 \tag{18.13}$$

$$\| \underline{K}^{\frac{1}{2}} \beta \| \leq t_2, 1^T \beta = 1 \tag{18.14}$$

$$\begin{bmatrix} K & y \\ y^T & t_1 \end{bmatrix} \geq 0 \tag{18.15}$$

Once the kernel is learnt, the first k-NN are identified considering that the Euclidean distance in feature space can be written exclusively in terms of kernel evaluations:

$$d_e^2(x_i, x_j) = k(x_i, x_i) + k(x_j, x_j) - 2k(x_i, x_j) \tag{18.16}$$

where k is the kernel of dissimilarities learnt by the regularized kernel alignment algorithm introduced previously.

18.3 Experimental Results

The algorithms proposed have been applied to the identification of several human cancer samples using microarray gene expression data.

The first dataset [7] consists of frozen tumor specimens from newly diagnosed, previously untreated Mediastinal large B-cell Lymphoma (MLBCL) patients

(34 samples) and Diffuse large B-cell lymphoma (DLBCL) patients (176 samples). The second problem compares primary cancers that have not spread beyond the breast to those who have metastasized to axillary lymph nodes at the time of diagnosis. We identified tumors as "reported negative" (24) when no positive lymph nodes were discovered and "reported positive" (25) for tumors with at least three identifiably positive nodes [10]. The third dataset [6] consists of 60 samples containing 39 medulloblastoma survivors and 21 treatment failures. All the datasets have been standardized subtracting the median and dividing the interquantile range. The rescaling were performed based only on the training set to avoid bias.

To assure a honest evaluation of all the classifiers, we have performed a double loop of cross-validation [2]. The performance of the classifiers has been evaluated considering the proportion of samples misclassified to estimate the optimal parameters avoiding overfitting. The stratified variant of cross-validation keeps the same proportion of patterns for each class in training and test sets. This is necessary in our problem because the class proportions are not equal.

Regarding the value of the parameters, $c_i = 1/n$ for the finite family of distances, where n is the number of dissimilarities that is fixed to 6 in this paper. The regularization parameter $\lambda_Q = 1$ gives good experimental results for all the problems considered. Finally, for the infinite family of dissimilarities, the regularization parameter λ_h in the harmonic hyperkernel has been set up to 0.6 which gives an adequate coverage of various kernel widths. Smaller values emphasize only wide kernels. All the base kernel of dissimilarities have been normalized so that they have the same scale. Three different kernels have been considered, linear, inverse multiquadratic, and Laplacian.

The optimal values for the kernel parameters, the number of genes and the nearest neighbors considered have been set up by cross-validation using a grid search strategy.

We have compared with the Lanckriet formalism [4] that allows us to incorporate a linear combination of dissimilarities into the SVM considering the connection between kernels and dissimilarities, the Large Margin Nearest Neighbor algorithm [9] that learns a Mahalanobis metric maximizing the k-NN margin in input space and the classical k-NN with the best dissimilarity for a subset of six measures widely used in the microarray literature.

From the analysis of Table 18.1, the following conclusions can be drawn.

Kernel alignment k-NN outperforms two widely used strategies to learn the metric, such as Large Margin NN and Lanckriet SVM. The first one is prone to overfitting and does not help to reduce the error of k-NN based on the best dissimilarity. Similarly, our method improves the Lanckriet formalism particularly for Breast LN problem in which the sample size is smaller.

Kernel alignment k-NN is quite insensitive to the kind of nonlinear kernel employed.

Kernel alignment k-NN considering an infinite family of distances outperforms k-NN with the best distance and the v-SVM, particularly for breast cancer and lymphoma DLBCL-MLBCL. The infinite family of dissimilarities helps to reduce

Table 18.1 Empirical results for the kernel alignment k NN based on a combination of dissimilarities

Technique	DLBCL MLBCL (%)	Breast LN (%)	Medulloblastoma (%)
k NN (Best distance)	10	6	10
v SVM (Best Distance)	11	8.16	13.3
Kernel align. k NN (Finite family, linear kernel)	10	6	11.66
Kernel align. k NN (Infinite family, linear kernel)	10	4	10
Kernel align. k NN (Finite family, inverse kernel)	10	8	10
Kernel align. k NN (Infinite family, inverse kernel)	9	4	10
Kernel align. k NN (Finite family, laplacian kernel)	9	6	8.33
Kernel align. k NN (Infinite family, laplacian kernel)	9	4	10
Lanckriet SVM	11	8.16	11.66
Large Margin NN	17	8.50	13.3

We have included two widely used learning metric strategies

the errors of the finite counterpart particularly for breast cancer. This suggests that for certain complex nonlinear problems, the nonlinear transformation of the original dissimilarities helps to improve the classifier accuracy. We report that only for the Medulloblastoma and with Laplacian base kernel the error is slightly larger for the infinite family. This suggests that the regularization term controls appropriately the complexity of the resulting dissimilarity.

18.4 Conclusions

In this paper, we propose two methods to incorporate in the k-NN algorithm a linear combination of non-Euclidean dissimilarities. A penalty term has been added to avoid the overfitting of the data. The algorithm has been applied to the classification of complex cancer human samples.

The experimental results suggest that the combination of dissimilarities in an HRKHS improves the accuracy of classifiers based on a single distance particularly for nonlinear problems. Besides, this approach outperforms other learning metric strategies widely used in the literature and is robust to overfitting.

References

1. N. Cristianini, J. Kandola, J. Elisseeff, and A. Shawe Taylor, "On the kernel target alignment", *Journal of Machine Learning Research,* vol. 1, pp. 1 31, 2002.
2. R. Gentleman, V. Carey, W. Huber, R. Irizarry, and S. Dudoit, "*Bioinformatics and Compu tational Biology Solutions Using R and Bioconductor*", Berlin: Springer Verlag, 2006

3. J. Kandola, J. Shawe Taylor, and N. Cristianini, “Optimizing kernel alignment over combina tions of kernels”, NeuroCOLT, Tech. Rep, 2002.
4. G. Lanckriet, N. Cristianini, P. Barlett, L. El Ghaoui, and M. Jordan, “Learning the kernel matrix with semidefinite programming”. *Journal of Machine Learning Research* vol. 3, pp. 27 72, 2004.
5. E. Pekalska, P. Paclick, and R. Duin, “A generalized kernel approach to dissimilarity based classification”. *Journal of Machine Learning Research,* vol. 2, pp. 175 211, 2001.
6. S.E.A. Pomeroy, “Prediction of central nervous system embryonal tumour outcome based on gene expression”. *Nature*, vol. 415, pp. 436 442, 2002
7. K. Savage et al, “The molecular signature of mediastinal large B cell lymphoma differs from that of other diffuse large B cell lymphomas and shares features with classical hodgkin lymphoma”, *Blood,* vol. 102(12), pp. 3871 3879, 2003.
8. C. Soon Ong, A. Smola, and R. Williamson, “Learning the kernel with hyperkernels”, *Journal of Machine Learning Research,* vol. 6, pp. 1043 1071, 2005.
9. K.Q. Weinberger, L.K. Saul, “Distance metric learning for large margin nearest neighbor classification”, *Journal Of Machine Learning Research,* vol. 10, pp. 207 244, 2009.
10. M. West et al, “Predicting the clinical status of human breast cancer by using gene expression profiles”, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98(20), pp. 11462 11467, 2001.

Chapter 19
The Application of Regular Expression-Based Pattern Matching to Profiling the Developmental Factors that Contribute to the Development of the Inner Ear

Christopher M. Frenz and Dorothy A. Frenz

Abstract The biomedical literature has always played a critical role in the development of hypotheses to test, experimental design, and the analysis of study results. Yet, the ever-expanding body of biomedical literature is starting to present new challenges, in which locating pertinent literature from among the millions of published research articles is often a challenging task. A regular expression-based pattern matching method has been developed to profile the various gene and protein factors that may play a role in various tissues contained within an organism. This methodology has been demonstrated through the profiling of the various factors that are involved in the development of the inner ear, and is shown to be both effective and accurate.

Keywords Bioinformatics · Gene expression · Inner ear · Text-mining

19.1 Introduction

As the biological sciences have matured, the problems that researches have sought to tackle have increasingly grown in complexity so that many studies no longer look at a single aspect in isolation but try to ascertain the biological make-up and significance of biological systems and processes that comprise multiple components. This trend can largely be observed in the development of the field of systems biology, which seeks to define and model the complex interaction networks that

C.M. Frenz (✉)
Department of Computer Engineering Technology, New York City College of Technology (CUNY), 300 Jay Street, Brooklyn, NY 11201, USA
e mail: cfrenz@citytech.cuny.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 19,

comprise biological systems, such as gene regulation and metabolism [1]. While the study of discrete components of these systems is still a valuable endeavor, it is becoming increasingly common to see profiling techniques, such as DNA microarrays and antibody arrays, applied to such problem domains because they allow for the expression levels of hundreds to thousands of components simultaneously. Such technologies can be used to illustrate the patterns of gene and protein expression at key developmental stages as well as in response to treatments that are applied to cells and tissues of interest. For many scientists, such profiling experiments have begun to play a key role in the identification of genes and proteins that might provide useful candidates for further study [2, 3].

The advent of such profiling techniques, however, does not invalidate the utility of one of the most time-tested means of candidate selection and hypothesis generation, which entails using published studies as a source of information to aid in the conception and development of new research hypotheses. The proven nature of this approach is readily observed by reading through any published scientific study and noting all of the cited literature that was accredited with providing information used to conceptualize the study or information that proved critical to the interpretation of the study's findings. Yet, despite the utility of such information, a growing problem is the ability to actually find information pertinent to your needs. PubMed, one of the most widely used sources for locating biological literature contains abstracts for over 17 million published articles [4]. The vastness of this repository often makes it prohibitively large for researchers to locate relevant information and thus researchers are increasingly turning to computational methods of data extraction and information processing as study aids. To date, such computational approaches have been used to extract information pertaining to gene and protein names [5], intermolecular relationships [6], molecular biological descriptors [7], and mutations [8].

A common method for extracting information from biological text repositories is to make use of regular expression-based pattern matching, since regular expressions allow for textual patterns to be defined. For example, a mutation in a protein is typically identified according to the convention of having a single letter that represents the amino acid in the wild type sequence, followed by the number that represents the position of the amino acid within the sequence, followed by the letter that represents the amino acid that is present in the mutated form of the sequence. Regular expressions provide a readily available means of expressing such nomenclature conventions, and the regular expression

$$[ARNDCEQGHILKMFPSTWYV]/d + [ARNDCEQGHILKMFPSTWYV]$$

can be used to match any mutation listed in the literature, which follows that mutation convention. As such, regular expression-based pattern matching provides an ideal information search and extraction tool, as it can be used to express many nomenclatures or data types that would be hard to express using standard keyword-based search techniques [8, 9].

This study seeks to apply the principles of gene and protein expression profiling typically encapsulated in microarrays and antibody arrays to biological

information, by demonstrating a methodology that is capable of extracting growth and transcription factor expression patterns from published biological literature, using regular expression-based pattern matching. The inner ear is used as an organ system for purposes of testing this system, since it provides for a great diversity of cell and tissue types and thus allows for a greater range of expression possibilities than other organ systems would likely offer.

19.2 Methods

The PREP.pl Perl script is a freely available open source software that has been developed to allow for the application of regular expression-based pattern matching techniques to abstracts contained within the PubMed database. The script is available from http://bioinformatics.org/project/?group id=494, and makes an ideal choice for this study, since it has been demonstrated to successfully process a data set relevant to the human ear. The script allows the user to provide a regular expression as well as a list of one or more keywords. It then uses the keywords to search PubMed, downloads all of the abstracts returned from the PubMed search as being pertinent to those keywords, and then searches through those abstracts to identify the ones that contain the user-defined regular expression. As such, the script provides a useful basis for information extraction and search refinement-based techniques [10].

The PREP.pl software formed the basis of the software used in this study but was modified to allow multiple regular expressions to be defined. One set of regular expressions contained text patterns that matched key proteins and genes of interest to the development of the inner ear, and a second set of regular expressions contained text patterns that matched the different cell and tissue types found in the inner ear. The script only returned matches in cases where a regular expression from each set was found to match elements contained within an abstract. The regular expressions used to define the text patterns associated with the genes and proteins of interest are defined in Table 19.1 and the parts of the ear are listed in Table 19.2. In all cases, matching against these text patterns was performed in a

Table 19.1 Gene and protein classes that play a role in the development of the inner ear and the regular expressions used to match them in article abstracts

Developmental factor	Regular expression
GATAs	gata(\s+|)?\d+
FGFs	fgf(\s+|)?\d+
BMPs	bmp(\s+|)?\d+
Connexins	Cx(\s+|)?\d+|Connexin(\s+|)?\d+
shh	shh
EGF	EGF
EPO	EPO\s+?
Myosin	MY[O,H,L]\d+[A H]?|MYLIP|MYLK2?| [Mm]yosin

Table 19.2 Parts of the inner ear which will be sought to be matched to developmental factors discussed in the literature

Inner ear parts		
Cochlea	Pillar cell	Macula
Organ of Corti	Tectorial membrane	Spiral limbus
Inner hair cell	Endolymphatic duct	Stria Vascularis
Outer hair cell	Endolymphatic sac	Reissner's membrane
Hair cell	Semicircular ducts	Basilar membrane
Hensen's cell	Cristae ampullaris	Spiral ganglion
Dieter's cell	Vestibule	Vestibular ganglion
Cells of Claudius	Utricle	Capsule
Cells of Boettcher	Saccule	

case insensitive manner to minimize the occurrence of false negatives with the data set. The expressions used were also designed with the goal of allowing matching to all potential nomenclature variants, where possible measures were taken to reduce the occurrence of false positives when testing revealed it to be necessary. For example, "epo" was found to occur in words found in several abstracts and thus the expression was modified to require a space after epo to eliminate such words from registering a positive result for erythropoetin.

Once the code modifications and regular expressions were deemed successful, the modified script was used to search PubMed with the keywords "inner ear".

19.3 Results and Discussion

At the time of execution, over 42,000 abstracts were returned as being relevant to the search term inner ear and were thus processed by the modified PREP.pl code to determine instances, where an inner ear part and a developmental factor were to be found within the same abstract. The various developmental factors and inner ear part combinations discovered by the program are listed in Table 19.3. The number of instances of each combination is not considered because it is hypothesized to be more a function of the popularity of a given research area rather than a result of the biological significance of a given combination.

However, the reliability of the pairing of each developmental factor to a specified part of the inner ear is noteworthy. Gata2 and Gata3, transcription factors that demonstrate partially overlapping expression domains in the early developing inner ear become distinct later on. These later expression patterns, which are highly conserved between species, are well-demarcated to the appropriate region or cell-type of the inner ear using the approach of expression-based pattern matching. However, a false positive was noted, i.e., Gata3 is not expressed in the endolymphatic duct, but was matched to it in an abstract that described the expression of Gata3 and Pax2, of which only Pax2 is present in this inner ear structure [11]. Likewise, myosins play an important role in the development and functionality of the inner ear, particularly in the inner and outer hair cells of the cochlea, where the data mining technique correctly assigned the expression of various myosins. It is noteworthy that two false positives

Table 19.3 Genes and proteins expressed in the inner ear matched with their expression locations within the biomedical literature

Developmental Factors	Cochlea	Organ of corti	Inner hair cell	Outer hair cell	Hair cell	Pillar cell	Tectorial membrane	Endolymphatic Duct	Endolymphatic Sac	Semicircular Duct	Cristae ampularis	Vestibule	Utricle	Saccule	Macula	Spiral ganglion	Stria vascularis	Basilar membrane	Spiral limbus	Vestibular ganglion	Capsule
GATA2	X																				
GATA3	X			X				X							X	X					
FGF1	X				X											X					
FGF2	X			X	X	X										X				X	
FGF3	X				X			X		X					X						X
FGF8		X			X	X														X	
FGF9																					X
FGF10	X				X																
FGF16	X																				
FGF19	X																			X	
FGF20	X																				
BMP2	X				X														X		X
BMP4	X				X			X	X	X	X		X		X	X				X	X
Cx26	X	X		X	X							X		X		X	X		X		
Cx29	X															X					
Cx30	X																				
Cx31	X				X																
Cx32	X																				
Cx36	X																				
Cx43	X																				
shh	X				X																X
EGF	X	X		X	X	X		X	X			X	X		X	X	X				
EPO	X								X							X					
Myosin	X	X	X	X	X		X											X			
Myo1A	X																				
Myo1C					X																
Myo6					X																
Myo7								X													
Myo7A	X	X			X																
Myo15	X																				
Myo15A	X				X																
MyH9	X																				

occurred when myosin was linked to the tectorial membrane [12] and the basilar membrane [13]. Examination of the content of these abstracts revealed that the mentions of myosin were not in relation to the inner ear parts being discussed.

Numerous members comprise both the fibroblast growth factor (FGF) and connexin (Cx) gene families, with each family having multiple members that are expressed and functional during inner ear development. Identification of the particular FGF or Cx that is expressed in the diverse regions of the inner ear can thus be a rather cumbersome task, but has overall been effectively accomplished using the data mining methodology. It is, however, important to note that care must be executed in the selection of appropriate and current terminology, both for the developmental factor and region of the inner ear that is of interest. For example, FGF4 is endogenously expressed in tissues relevant to the developing inner ear, i.e., in tissues from which the inner ear forms [14], but was not included in the match (Table 19.3) because the ectodermal tissue in which it is localized was not in the "target" list. In addition, FGF2 plays an important role in the development of the otic capsule [15]; however at the time of publication of the pertinent literature, FGF2 was more commonly referred to as basic FGF and thus FGF2 and capsule did not match.

Bone morphogenetic protein (BMP), Sonic hedgehog (Shh), and epidermal growth factor (EGF) are well established in their particular roles in inner ear development and auditory (hearing) or vestibular (balance) function, being expressed in various tissues in the developing and mature inner ear, as well as in tissues in the process of remodeling or repair (e.g., BMP, EGF). The data mining technique has effectively extracted the pertinent expression data relating to these factors and linked it to the appropriate cells and tissues of the inner ear. Of interest, even a factor that was not anticipated to be expressed in the developing inner ear, i.e., erythropoietin (EPO) was identified to the cochlea and spiral ganglion neurons [16] and to the endolymphatic sac, where it is involved in tumorigenesis [17].

In all, the simultaneous regular expression-based pattern matching for developmental factors and tissue types has been demonstrated to provide an effective means of profiling the occurrence of these factors within the various tissues of interest. The occurrences of false positives were not prevalent within the data set and the false negatives discovered could be easily rectified by improving upon the regular expressions used for the various developmental factors and tissue types. Thus, despite these minor caveats, the methodology presented has the potential to provide biomedical scientists a rapid way to search the biomedical literature for the information pertinent to both experimental design and data analysis.

References

1. T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins, "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science,* vol. 301, pp. 102 105, 2003.
2. G. S. Chaga, "Antibody arrays for determination of relative protein abundances," *Meth Mol Biol,* vol. 441, pp. 129 151, 2008.

3. M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary DNA microarray," *Science,* vol. 270, pp. 467 470, 1995.
4. D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko, "Database resources of the National Center for Biotechnology Information," *Nucleic Acids Res,* vol. 36, pp. D13 D21, 2008.
5. J. E. Leonard, J. B. Colombe, and J. L. Levy, "Finding relevant references to genes and proteins in Medline using a Bayesian approach," *Bioinformatics,* vol. 18, pp. 1515 1522, 2002.
6. M. Yoshida, K. Fukuda, and T. Takagi, "PNAD CSS: A workbench for constructing a protein name abbreviation dictionary," *Bioinformatics,* vol. 16, pp. 169 175, 2000.
7. M. A. Andrade, and P. Bork, "Automated extraction of information in molecular biology," *FEBS Lett,* vol. 476, pp. 12 17, 2000.
8. F. Horn, A. L. Lau, and F. E. Cohen, "Automated extraction of mutation data from the literature: application of MuteXt to G protein coupled receptors and nuclear hormone recep tors," *Bioinformatics,* vol. 20, pp. 557 568, 2004.
9. C.M. Frenz, *Pro Perl Parsing*. Berkeley, CA: Apress, 2005.
10. C. M. Frenz, "Deafness mutation mining using regular expression based pattern matching," *BMC Med Inform Decis Mak,* vol. 7, p. 32, 2007.
11. G. Lawoko Kerali, M. N. Rivolta, and M. Holley, "Expression of the transcription factors GATA3 and Pax2 during development of the mammalian inner ear," *J Comp Neurol,* vol. 442, pp. 378 391, 2002.
12. L. M. Friedman, A. A. Dror, and K. B. Avraham, "Mouse models to study inner ear development and hereditary hearing loss," *Int J Dev Biol,* vol. 51, pp. 609 631, 2007.
13. S. J. Harvey, R. Mount, Y. Sado, I. Naito, Y. Ninomiya, R. Harrison, B. Jefferson, R. Jacobs, and P. S. Thorner, "The inner ear of dogs with X linked nephritis provides clues to the pathogenesis of hearing loss in X linked Alport syndrome," *Am J Pathol,* vol. 159, pp. 1097 1104, 2001.
14. T. J. Wright, E. P. Hatch, H. Karabagli, P. Karabagli, G. C. Schoenwolf, and S. L. Mansour, "Expression of mouse fibroblast growth factor and fibroblast growth factor receptor genes during early inner ear development," *Dev Dyn,* vol. 228, pp. 267 272, 2003.
15. D. A. Frenz, W. Liu, J. D. Williams, V. Hatcher, V. Galinovic Schwartz, K. C. Flanders, and T. R. Van de Water, "Induction of chondrogenesis: requirement for synergistic interaction of basic fibroblast growth factor and transforming growth factor beta," *Development,* vol. 120, pp. 415 424, 1994.
16. P. Caye Thomasen, N. Wagner, B. Lidegaard Frederiksen, K. Asal, and J. Thomsen, "Erythropoietin and erythropoietin receptor expression in the guinea pig inner ear," *Hear Res,* vol. 203, pp. 21 27, 2005.
17. T. W. Vogel, A. O. Vortmeyer, I. A. Lubensky, Y. S. Lee, M. Furuta, B. Ikejiri, H. J. Kim, R. R. Lonser, E. H. Oldfield, and Z. Zhuang, "Coexpression of erythropoietin and its receptor in endolymphatic sac tumors," *J Neurosurg,* vol. 103, pp. 284 288, 2005.

Chapter 20
Functionally Informative Tag SNP Selection Using a Pareto-Optimal Approach

Phil Hyoun Lee, Jae-Yoon Jung, and Hagit Shatkay

Abstract Selecting a representative set of single nucleotide polymorphism (SNP) markers for facilitating association studies is an important step to uncover the genetic basis of human disease. Tag SNP selection and functional SNP selection are the two main approaches for addressing the SNP selection problem. However, little was done so far to effectively combine these distinct and possibly competing approaches. Here, we present a new multiobjective optimization framework for identifying SNPs that are both informative tagging and have functional significance (FS). Our selection algorithm is based on the notion of Pareto optimality, which has been extensively used for addressing multiobjective optimization problems in game theory, economics, and engineering. We applied our method to 34 disease-susceptibility genes for lung cancer and compared the performance with that of other systems which support both tag SNP selection and functional SNP selection methods. The comparison shows that our algorithm always finds a subset of SNPs that improves upon the subset selected by other state-of-the-art systems with respect to both selection objectives.

Keywords Functional SNP · Lung cancer · Multiobjectve optimization · SNP · Tag SNP

20.1 Introduction

Identifying single nucleotide polymorphisms (SNPs) that underlie the etiology of common and complex diseases is of primary interest in current molecular epidemiology, medicine, and pharmarco-genomics. Knowledge of such SNPs is expected to enable timely diagnosis, effective treatment, and ultimately prevention of human

P.H. Lee (✉)
Center for Human Genetics Research, Department of Medicine, Harvard Medical School and Massachusetts General Hospital, Boston, MA 02114, USA
e mail: phlee@pngu.mgh.harvard.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 20,

disease. However, the vast number of SNPs on the human genome, which is estimated at more than 11 million [14], poses challenges to obtain and analyze the information of all the SNPs, especially for large-scale population-based association studies. As a result, the strategy of SNP marker selection has been a crucial element of study design to increase the statistical power and the coverage of association studies, even at the genome-wide level [12]. In this chapter, we address the problem of selecting representative SNP markers for supporting effective disease-gene association studies.

The two main approaches for addressing the SNP selection problem are *tag SNP selection* and *functional SNP selection*. The tag SNP selection approach was motivated by the nonrandom association among SNPs, called *linkage disequilibrium* (LD) [7]. When high LD exists between SNPs, the allele information of one can usually be inferred from the others. Thus, we can select a subset of SNPs that still retains most of the allele information of the original set. Under this SNP selection approach, possible association between a disease phenotype and unselected SNPs is assumed to be *indirectly* captured through the selected tag SNPs.

On the other hand, the functional SNP selection approach aims to *directly* select a subset of SNPs that are likely to be disease-causing [13]. For example, nonsynonymous SNPs that radically change the amino acid composition of a protein are highly likely to distort the protein's function, and are therefore more likely to underlie disease. SNPs that disrupt regulatory sites can affect gene expression, and can therefore alter tissue-specificity and cellular activity of relevant proteins. Directly genotyping and analyzing these SNPs that are likely to be responsible for disease etiology is expected to reduce false positive errors and to increase reproducibility of association studies [13].

In recent years, there have been a few efforts to combine these two SNP selection approaches into one selection framework [6, 16]. Most of the work, however, considers tag SNP selection and functional SNP selection as two separate optimization problems, and solves them each separately, while combining the resulting SNP sets as a final step. The number of selected SNPs in the resulting set can thus be much larger than necessary.

To address this issue, we previously proposed a greedy selection algorithm in which both tagging informativeness and functional significance (FS) of SNPs are incorporated into a single objective function expressed as a weighted sum [10]. However, this formulation is limited by the fact that the selected set of SNPs depends on the predefined weighting factors, whose optimal value is unknown a priori in most cases. Here, we introduce a new multiobjective SNP selection system that, as one of its main contributions, overcomes this limitation by using the well-established, game-theoretic notion of *Pareto optimality* [9]. To the best of our knowledge, this idea was not applied before in any genetic variation study.

20.2 Problem Formulation

Given a set of p SNPs on a target genomic locus, we aim to find a subset of at most k SNPs that are most functionally significant as well as most informative in representing the allele information of the whole set of SNPs. We call this problem

functionally informative tag SNP selection. As a basis, we use our own previous formulation of the problem [10], but extend it here in the context of Pareto optimality. We also redefine objective functions to incorporate the widely used concept of pairwise linkage disequilibrium (LD).

Suppose that our target locus contains p consecutive SNPs. We represent each SNP as a discrete random variable, X_j ($j = 1,...,p$), whose possible values are the four nucleotides, $\{a,c,g,t\}$. Let $V = \{X_1,...,X_p\}$ denote a set of random variables corresponding to the p SNPs. A haplotype dataset, D, that contains the allele information of n haplotypes, each of which consists of the p SNPs in V, is provided as input. We are also given the set of FS scores for the p SNPs, which we denote by $E = \{e_1,...,e_p\}$. Last, for a subset of SNPs, S, where S is a subset of V, we define an objective function, $f(S|D,E)$, to reflect both the allele information carried by the SNPs in S about the remaining SNPs in (V S) and the FS scores represented by the SNPs in S.

The problem of *functionally informative tag SNP selection* is then to find among all possible subsets of the original SNPs in the set V, an optimal subset of SNPs, S, of maximum size k, based on the objective function, $f(S|D,E)$. We define the objective function, $f(S|D,E)$, as an ordered pair of two simpler objective functions, $f_1(S|D)$ and $f_2(S|D,E)$, where the former measures the allelic information of a SNP set S, while the latter measures the FS as follows.

Definition 1. *Information-Based Objective $f_1(S/D)$.*

$$f_1(S|D) = \frac{1}{p}\sum_{j=1}^{p} I_1(X_j, S),$$

where α is a parameter value ($0 < \alpha < 1$) and I_1 is a modified indicator function formally defined as:

$$I_1(X_j, S) = \begin{cases} 1, & \text{if } \exists X_s \in S \text{ such that } \mathrm{LD}(X_j, X_s|D) \geqslant \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

Definition 2. *Function-Based Objective $f_2(S/D,E)$.*

$$f_2(S|D,E) = \frac{\sum_{j=1}^{p} (e_j \cdot I_2(X_j, S))}{\sum_{j=1}^{p} e_j},$$

where I_2 is a modified indicator function formally defined as:

$$I_2(X_j, S) = \begin{cases} 1, & \text{if } X_j \in S, \\ ld_j, & \text{if } X_j \notin S \text{ and } \exists X_s \in S \text{ such that } \mathrm{LD}(X_j, X_s|D) \geqslant \alpha, \\ 0, & \text{otherwise.} \end{cases}$$

Definition 3. *Functionally Informative Objective Function* f(S|D,E).

$$f(S|D,E) = \langle f_1(S|D), f_2(S|D,E) \rangle$$

Note that we aim to optimize these two distinct and possibly competing objectives, $f_1(S|D)$ and $f_2(S|D,E)$ *simultaneously*. To achieve this goal, we adopt the notion of Pareto optimality defined as follows.

Definition 4. *Pareto Optimality. Let S_i and S_j be two distinct subsets of V, of the same size, k. S_i is said to dominiate S_j if and only if ($f_1(S_i) \geq f_1(S_j)$ and $f_2(S_i) > f_2(S_j)$) or ($f_1(S_i) > f_1(S_j)$ and $f_2(S_i) \geq f_2(S_j)$). S_i is called Pareto optimal if and only if no other subset of V dominates S_i.*

In summary, among all possible SNP subsets of maximum size k, we aim to select all Pareto optimal subsets of functionally informative tag SNPs. In the problem of finding k, most informative tag SNPs has been proven to be NP-complete by Bafna et al. [2]. We compute our information-based objective independently of the function-based objective, which means that considering the two selection objectives simultaneously does not reduce the complexity of the problem. In the next section, we thus propose a heuristic framework (which, like all heuristics, looks for a locally optimal solution), to address the problem of functionally informative tag SNP selection in the context of Pareto optimality.

20.3 Pareto-Based Functionally Informative Tag SNP Selection

Our selection algorithm is based on the multiobjective simulated annealing (SA) algorithm [8], which has been successfully used for addressing many combinatorial optimization problems [4].

First, we choose an initial subset of k SNPs using our weighted sum-based greedy selection algorithm [10] as a current solution, S_c, and compute the score-pair, $f(S_c|D,E)$. Second, while a temperature parameter, T, is greater than a minimum threshold, $T_{\min}$, the following three steps are repeated.

1. A neighbor set of the current solution, S_c, referred to as S_n, is generated (as explained later in this section).
2. If S_n is Pareto optimal among the sets we examined, S_n is added to the Pareto optimal solutions with respect to the examined sets, PO, and replaces S_c for the next iteration; otherwise, it replaces S_c with a probability P_{accept}. The probability P_{accept} is updated as a function of $f(S_c|D,E)$, $f(S_n|D,E)$, and T.
3. The temperature T is reduced by a rate of r_c.

This whole procedure is repeated M times. In the experiments described here, we empirically set the simulated annealing parameters as follows: $T_0 = 1.3$,

$r_c = 0.9999$, $T_{\min} = 0.001$, and $M = 10^4$. The temperature T is reduced by a rate of r_c.

To guide an efficient SA search, we introduce two heuristics for generating a new neighbor solution. First, in order to find a neighbor SNP set that is likely to dominate the current set, S_c, we utilize the score of each SNP with respect to the two selection objectives, f_1 and f_2. That is, for each SNP X_i, we compute the objective scores, $f_1(\{X_i\}|D)$ and $f_2(\{X_i\}|D,E)$, before starting the search. When generating a new neighbor for the current set of functionally informative SNPs, S_c, we first determine whether to focus on the information-based objective, f_1, or on the function-based objective, f_2, by flipping an unbiased coin.

Suppose that the information-based objective f_1 is selected. We now select a SNP X_r from S_c to be replaced by a SNP X_a from $(V - S_c)$. X_r is chosen with probability, P_{remove}, which is *inversely* proportional to its f_1 score, while X_a is chosen with probability, P_{add}, which is *directly* proportional to its f_1 score. That is,

$$P_{\text{add}} = \frac{f_1(\{X_i\}|D)}{\sum_{X_i \in (V \quad S_c)} f_1(\{X_i\}|D)} \quad \text{and} \quad P_{\text{remove}} = \frac{(f_1(\{X_i\}|D))^{\ 1}}{\sum_{X_i \in S_c} (f_1(\{X_i\}|D))^{\ 1}}.$$

A second heuristic is used to expedite and diversify the coverage of the search space. Instead of generating a new neighbor by replacing one SNP at a time, we simultaneously replace several SNPs in the initial search period, and gradually decrease the number of replaced SNPs as the search progresses. As a result, farther neighbors are examined in the initial stages of the search, diversifying the search area, while the later stages, which are expected to search closer to the optimum, focus on neighbors that are closer to the current solution. This strategy helps avoid local optima. In the next section, we show the utility of these two heuristics by comparing the performance of our selection algorithm with and without them.

The time complexity of each iteration is $O(p)$, where p is the number of candidate SNPs. As this iteration is repeated for maximum $(M \cdot \log(T_{\min}/T_0)/\log r_c)$ times, the overall complexity of our selection algorithm is $O(p \cdot M \cdot \log(T_{\min}/T_0)/\log r_c)$. The computation procedure of the pairwise LD between p SNPs is $O(n \cdot p^2)$, where n is the number of haplotypes, and p is the number of SNPs in dataset D.

20.4 Results and Conclusion

We applied our method to 34 disease-susceptibility genes for lung cancer, as summarized by [17]. The list of SNPs linked to the genes, including 10k upstream and downstream regions, was retrieved from the dbSNP database [14]. The haplotype datasets for the genes were downloaded from the HapMap consortium for the CEU population (public release #20/phase II) [15]. We obtained the FS scores for SNPs from the F-SNP database [11].

We compare the performance of our system with that of two state-of-the-art SNP selection systems that support both tag SNP selection and functional SNP selection: SNPselector [16] and TAMAL [6]. The compared systems share the same goal as ours, namely, selecting an informative set of tag SNPs with significant functional effects. However, they address tag SNP selection and functional SNP selection as two separate optimization problems, while we address it as a single multiobjective optimization problem.

TAMAL and SNPselector do not allow users to specify the maximum number of selected SNPs. We thus first apply TAMAL and SNPselector to the dataset of 34 genes, and apply our system on the same dataset to select the same number of SNPs as selected by each compared system. We denote our full-fledged multiobjective simulated annealing algorithm that employs the two heuristics by SA_1. In addition, we demonstrate the utility of our heuristics by examining the performance of two baseline search algorithms for identifying (locally) Pareto optimal solutions: (1) The proposed simulated annealing algorithm without the proposed two heuristics, which we denote by SA_0; and (2) A naïve selection algorithm that randomly generates M solutions and identifies (locally) Pareto optimal subsets within the M solutions. We refer to this naïve selection algorithm as *RS*.

Table 20.1 summarizes the evaluation results of the three Pareto optimal search algorithms, SA_1, SA_0, and *RS* against the two compared systems, SNPselector and TAMAL (due to the space issue, only the results for the first 14 genes based on their alphabetical gene symbols are shown). The first, second, and third columns show

Table 20.1 Evaluation results of three Pareto optimal search algorithms, SA_1, SA_0, and *RS* against the two compared systems, SNPselector and TAMAL

Gene symbol	Locus	Total SNP#	SNPselector				TAMAL			
			k	SA_1	SA_0	*RS*	k	SA_1	SA_0	*RS*
ADRB2	5q31	153	41	100	100	33.3	17	66.6	100	50†
APEX1	14q11.2	83	27	100	100	100	19	100	100	100
ATR	3q22	181	36	100	100	100	20	100	100	50
CDKN1A	6p21	116	34	100	100	100	20	100	100	100
CYP1A1	15q22	49	34	100	100	100	10	100	100	75
CYP1B1	2p21	172	51	100	100	100	28	100	100	100
EPHX1	1q42.1	148	27	80	25	14.2†	23	25	.	.
ERCC2	19q13.3	210	27	100	100	100	30	50	50	.
ERCC4	16p13.3	289	41	100	100	44.4	49	50	50	20†
ERCC5	13q22	261	43	88	25	.	43	11.1	.	.
GSTP1	11q13	70	27	100	100	100	14	100	100	50
LIG4	3q33	107	27	100	100	100	27	20	100	25†
MBD1	18q21	65	24	100	100	100	19	100	100	50
MGMT	10q26	550	36	100	100	71.4	81	20	25†	.

Under the name of each compared system, the leftmost column shows the number of SNPs, k, selected by the compared system, for the corresponding gene. The remaining three columns, SA_1, SA_0, and *RS*, typically show the e_1 score (i.e., the percentage of the identified Pareto optimal solutions that *dominate* the compared systems solution), computed for each of the respective Pareto optimal search algorithms. In the few cases where the solutions are *dominated* by the compared systems solution, e_2 is shown (denoted by †). Cases where there is no dominating or dominated solution are indicated by a dot

gene symbols, genomic locus, and the total number of SNPs linked to each gene, respectively. The remaining columns are divided into two parts, corresponding to the two compared systems. In each part, the leftmost column shows the number of SNPs, k, which is chosen by each compared system for the corresponding gene. The remaining three columns, SA_1, SA_0, and RS show the evaluation measure, e_1 (the majority of the cases) or e_2 (designated by †) computed for the corresponding search algorithm, respectively. When both of e_1 and e_2 are 0, which means that the compared solution is neither dominant nor dominated by our Pareto optimal solutions, we display it with a dot.

Overall, the SA_1 algorithm that uses the two proposed heuristics always finds Pareto optimal subsets that *dominate* the compared solutions. The difference between our dominating solution and the compared system solution is statistically significant with respect to both selection objectives. Using the paired t-test with 5% significance level, p-values are 8.29e-179 for $f_1(S|D)$ and 8.15e-157 for $f_2(S|D,E)$ in the case of SNPselector, and 7.02e-073 and 5.76e-005 for TAMAL. In contrast, the naïve SA_0 algorithm, which does not employ any heuristics, fails to find dominating solutions in eight cases (shown as † or $\cdot$ in Table 20.1). In three cases, SA_0's solutions are dominated by the compared system's (shown as † in Table 20.1). The random search algorithm RS fails to find dominating solutions in 23 cases, while producing dominated solutions in 11 cases.

In this chapter, we presented a new multiobjective optimization framework for selecting functionally informative tag SNPs, based on the notion of pareto optimality. The proposed work shows a new application of established multiobjective optimization frameworks in human genetics and medicine. Moreover, the work clearly demonstrates that combining distinct problem solving criteria into one unified process is possible, and indeed improves upon separate optimization approaches. We expect that the proposed multiobjective optimization approach can be applied to solve other types of optimization problems in human genetics and medicine [5], for example, designing new drug combinations [3] or modeling biological networks [1].

Acknowledgments This work was supported by HS's NSERC Discovery grant 298292 04 and CFI New Opportunities Award 10437.

References

1. Adiwijaya BS, Barton PI, Tidor B (2006) Biological network design strategies: Discovery through dynamic optimization. Mol Biosyst. 2(12): 650 659.
2. Bafna V, Halldórsson BV, Schwartz R et al (2003) Haplotypes and informative SNP selection algorithms: Don't block out information. In Proceedings of the 7th International Conference on Computational Molecular Biology (RECOMB). 19 27.
3. Calzolari D, Bruschi S, Coquin L, et al (2008) Search algorithms as a framework for the optimization of drug combinations. PLoS Comput Biol. 4(12): e1000249.
4. Czyzak P, Jaszkiewicz A (1998) Pareto simulated annealing A metaheuristic technique for multiple objective combinatorial optimization. J Multi Criteria Decis Anal. 7: 34 47.

5. Handl J, Kell DB, Knowles J (2007) Multiobjective optimization in bioinformatics and computational biology. IEEE/ACM Trans Comput Biol Bioinform. 4(2): 279 292.
6. Hemminger BM, Saelim B, Sullivan PF (2006) TAMAL: An integrated approach to choosing SNPs for genetic studies of human complex traits. Bioinformatics. 22(5): 626 627.
7. Johnson GCL, Esposito L, Barratt BJ, et al (2001) Haplotype tagging for the identification of common disease genes. Nat Genet. 29(2): 233 237.
8. Kirkpatrick S, Gelatt C, Vecchi M (1983) Optimization by simulated annealing. Science. 22: 671 680.
9. Kirman AP (1987) Pareto as an economist, 5, 804 808. In Durlauf S N and Blume L E (ed), The New Palgrave: A Dictionary of Economics. Palgrave Macmillan, Hampshire, England.
10. Lee PH, Shatkay H (2007) Two birds, one stone: Selecting functionally informative tag SNPs for disease association studies. In the Proceedings of the 7th Workshop of Algorithms in Bioinformatics (WABI). 61 72.
11. Lee PH, Shatkay H (2009) An integrative scoring system for ranking SNPs by their potential deleterious effects. Bioinformatics. 25(8): 1048 1055.
12. Pettersson FH, Anderson CA, Clarke GM, et al (2009) Marker selection for genetic case control association studies. Nat Protoc. 4(5): 743 752.
13. Rebbeck TR, Spitz M, Wu X (2004). Assessing the function of genetic variants in candidate gene association studies. Nat Rev Genet. 5: 589 597.
14. Sherry ST, Ward MH, Kholodov M, et al (2001) dbSNP: The NCBI database of genetic variation. Nucl Acids Res. 29(1): 308 311.
15. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature. 437: 1299 1320.
16. Xu H, Gregory SG, Hauser ER et al (2005) SNPselector: A web tool for selecting SNPs for genetic association studies. Bioinformatics. 21(22): 4181 4186.
17. Zhu Y, Hoffman A, Wu X et al (2008) Correlating observed odds ratios from lung cancer case control studies to SNP functional scores predicted by bioinformatics tools. Mutat Res. 639: 80 88.

Chapter 21
KMeans Greedy Search Hybrid Algorithm for Biclustering Gene Expression Data

Shyama Das and Sumam Mary Idicula

Abstract Microarray technology demands the development of algorithms capable of extracting novel and useful patterns like biclusters. A bicluster is a submatrix of the gene expression datamatrix such that the genes show highly correlated activities across all conditions in the submatrix. A measure called Mean Squared Residue (MSR) is used to evaluate the coherence of rows and columns within the submatrix. In this paper, the KMeans greedy search hybrid algorithm is developed for finding biclusters from the gene expression data. This algorithm has two steps. In the first step, high quality bicluster seeds are generated using KMeans clustering algorithm. In the second step, these seeds are enlarged by adding more genes and conditions using the greedy strategy. Here, the objective is to find the biclusters with maximum size and the MSR value lower than a given threshold. The biclusters obtained from this algorithm on both the bench mark datasets are of high quality. The statistical significance and biological relevance of the biclusters are verified using gene ontology database.

Keywords Biclusters · Gene expression · Greedy search · KMeans · Microarray · MSR

21.1 Introduction

The relative abundance of the mRNA of a gene under a specific condition or sample is called the expression level of a gene which can be measured using microarray technology. Gene expression data are arranged in the form of a matrix. Rows represent genes and columns represent experimental conditions. Clustering is one

S. Das (✉)
Department of Computer Science, Cochin University of Science and Technology, Kochin, Kerala, India
e mail: shyamadas777@gmail.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 21,

of the most popular data mining techniques for the analysis of gene expression data. Clustering partition conditions or genes into similar groups. However, clustering has some limitations which can be overcome by using biclustering. Biclustering is clustering applied along the row and column dimensions simultaneously. This approach identifies group of genes that show similar expression level under a specific subset of experimental conditions. Cheng and Church were the first to apply biclustering to gene expression data [1]. Biclustering is a local model, whereas clustering represents a global model. In this work, high quality bicluster seeds are generated initially using the KMeans clustering algorithm. They are then enlarged using a greedy method. A greedy strategy makes a choice that maximizes a local gain with the hope that this choice will result in a globally good solution.

21.2 Identification of Biclusters with Coherent Values

Biclusters with coherent values are identified in this work. They are biologically more relevant than biclusters with constant values. The degree of coherence is measured by Mean Squared Residue (MSR) or Hscore. It is the sum of the squared residue score. The residue score of an element *bij* in a submatrix *B* is defined as $RS(bij) = bij - bIj - biJ + bIJ$. Hence, Hscore or MSR of bicluster *B* is

$$\text{MSR(B)} = \frac{1}{|I \parallel J|} \sum i \in I, J \in J(RS(bij))^2,$$

where *I* denotes the row set, *J* denotes the column set, *bij* denotes the element in a submatrix, *biJ* denotes the *i*th row mean, *bIj* denotes the *j*th column mean, and *bIJ* denotes the mean of the whole bicluster. If the MSR of a matrix is less than certain threshold δ, then it is a bicluster and called δ bicluster where δ is the MSR threshold. The value of δ depends on the dataset. For Yeast dataset, the value of δ is 300 and for Lymphoma dataset the value of δ is 1,200. There is correlation in the matrix if the MSR value is low. The volume of a bicluster or the bicluster size is the product of the number of rows and the number of columns in the bicluster. The larger the volume and the smaller the MSR of the bicluster, the greater the quality of the bicluster.

21.3 Description of the Algorithm

In this work, seeds are generated using KMeans clustering algorithm, and they are enlarged using greedy strategy. In the seed generation phase, the gene expression dataset is partitioned into p gene clusters and q sample clusters by using KMeans algorithm. Each gene cluster is further divided into sets of ten genes, and the sample clusters are divided into sets of five samples based on cosine angle distance from the cluster centre. The number of gene clusters having maximum ten genes is *m* and the number of sample clusters having maximum five conditions is *n*. They are

combined to form $n \times m$ submatrices. Submatrices with low MSR value are selected as seeds. Seed is a tightly coregulated submatrix of a gene expression data matrix that can accommodate more genes and conditions within the MSR threshold. More conditions and genes are added to the bicluster in the seed growing phase. In this phase, the genes and conditions that are not included in the seed are maintained as separate lists. Then, using the greedy search algorithm, the best element is selected from the condition list or gene list and added to the bicluster. The quality of the element is determined by the MSR value of the bicluster after including the element in the bicluster. The element which results in minimum MSR value when added to the bicluster is considered as the best element. The seed growing phase continues till the MSR of the bicluster reaches the given threshold [2].

21.3.1 Greedy Search Algorithm

```
Algorithm  greedyseedgrowing(seed, δ)
bicluster := seed
//Column_or_Row_List is the list of genes or conditions not included in the bicluster. These
//two lists are maintained separately. Search starts from condition list followed by gene list.
Calculate Column_or_Row_List
While (MSR(bicluster) <= δ)
  Num_elem=size(Column_or_Row_List)
          for  i:=1: Num_elem
                  bicluster=bicluster+ Column_or_Row_List [i]
                  Column_or_Row_List_msr[i]= MSR(bicluster)
                  Remove Column_or_Row_List[i] from bicluster
          end(for)
find minimum value in Column_or_Row_List_msr  and corresponding index K
biclust=biclust+ Column_or_Row_List [K]
delete Column_or_Row_List [K] from Column_or_Row_List
end(while)
end(greedyseedgrowing)
```

21.4 Experimental Results

21.4.1 Datasets Used

Experiments are conducted on the Yeast and Lymphoma datasets. The algorithm is implemented in Matlab. The Yeast dataset is based on Tavazoie et al. [3]. The dataset contains 2,884 genes and 17 conditions. The values in the dataset are

integers in the range 0 600. Missing values are represented by −1. Human B-cell Lymphoma dataset contains 4,026 genes and 96 conditions. The dataset is downloaded from the website for supplementary information for the article by Alizadeh et al. [4]. The values in the Lymphoma dataset are integers in the range −750 to 650. There are 47,639 (12.3%) missing values in the Lymphoma dataset. Missing values are represented by 999. The preprocessed dataset is obtained from http://arep.med.harvard.edu/biclustering. Missing data in the Lymphoma dataset are replaced with uniformly distributed random numbers between −800 and 800 [1].

21.4.2 Bicluster Plots for Yeast and Lymphoma Datasets

In Fig. 21.1, biclusters obtained by the KMeans greedy search algorithm on the Yeast and Lymphoma datasets are shown. From the bicluster plots, it is clear that genes show a similar behavior under a set of conditions. All the biclusters obtained from the Yeast dataset are having MSR less than 300. And all the biclusters obtained from the Lymphoma dataset are having MSR less than 1,200.

The first row shows eight biclusters of the Yeast dataset, and the second row shows eight biclusters of the Lymphoma dataset. The details about the biclusters are reported in the following format (label of bicluster, number of genes, number of conditions, volume of the bicluster, MSR, row variance). (a, 17, 17, 289, 99.3497, 407.47), (b, 108, 17, 1,836, 194.5204, 472.34), (c, 14, 17, 238, 97.8389, 507.63), (d, 33, 17, 561, 99.9639, 506.14), (e, 147, 17, 2,499, 200.2474, 396.04), (f, 31, 17, 527, 97.9121, 613.89), (g, 1,405, 9, 12,645, 299.8968, 348.07), (h, 10, 17, 170, 66.4403, 522.23), (p, 11, 94, 1,034, 1,194.40, 5,317.5), (q, 40, 66, 2,640, 918.25, 1,156.4), (r, 30, 80, 2,400, 1,175.9, 3,466.3), (s, 21, 9, 189, 476.12, 6,183.5), (t, 26, 81, 2,106, 1,196.8, 3,906), (u, 10, 83, 830, 1,182.1, 5,070.1), (v, 53, 35, 1,855, 723.41, 788.7), (w, 292, 9, 2,628, 1,196.9, 3,359.1).

21.5 Comparison

The performance of KMeans greedy search algorithm in comparison with that of SEBI [5], Cheng and Church's algorithm (CC) [1], the algorithm FLOC by Yang et al. [6], and DBF [7] for the Yeast and Lymphoma datasets are listed in Table 21.1 given below. In the case of KMeans greedy search algorithm presented here for the Yeast dataset, the average number of conditions is better than that of CC, FLOC, and DBF. Average gene number, average volume, and largest bicluster size are greater than those of all other algorithms. Average MSR is better than that of all other algorithms listed in the table except DBF. In multiobjective evolutionary computation [8] maximum number of conditions obtained is only 11. In this method, biclusters are obtained with all 17 conditions. For the Yeast dataset, the maximum number of genes obtained by KMeans greedy search algorithm in all the 17 conditions is 147 with MSR value 200.2474 (label of bicluster is (e) in Fig. 21.1).

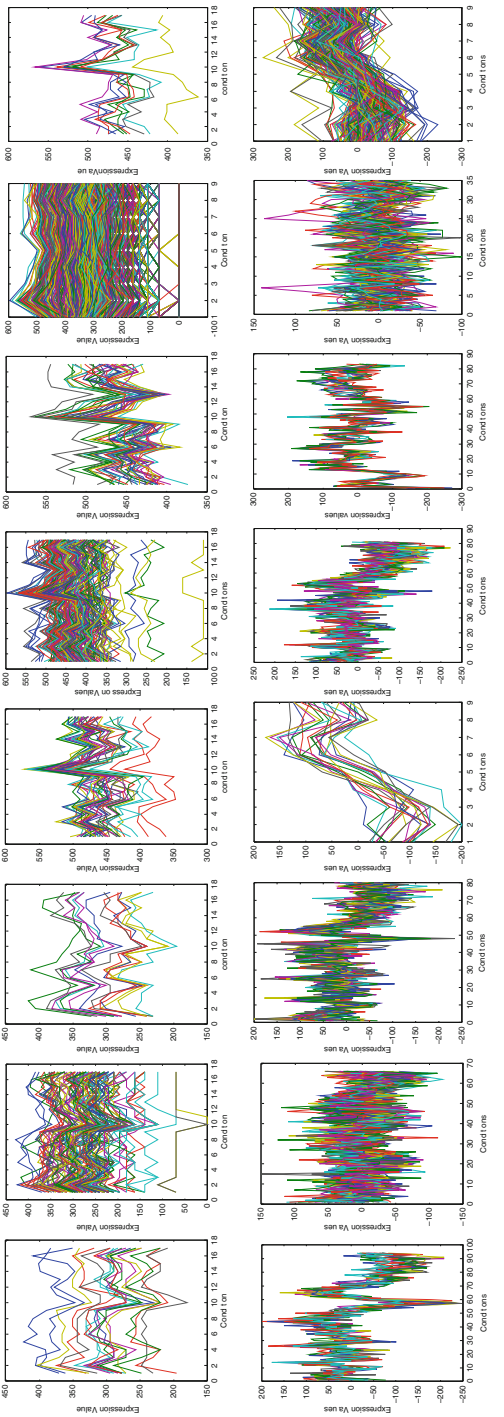


Fig. 21.1 Biclusters obtained by using KMeans greedy search hybrid algorithm

Table 21.1 Performance comparison between KMGS and other Algorithms

Algorithm	Datasets	Avg. MSR	Avg. volume	Avg. gene num.	Avg. cond. num	Largest Bicl. size
KMGS	Yeast	185.88	4,684.29	515.21	13.36	12,645
	Lymphoma	1,007.99	1,710.25	60.38	57.13	
SEBI	Yeast	205.18	209.92	13.61	15.25	1,394
	Lymphoma	1,028.84	615.84	14.07	43.57	
CC	Yeast	204.29	1,576.98	166.71	12.09	4,485
	Lymphoma	850.04	4,595.98	269.22	24.50	
FLOC	Yeast	187.54	1,825.78	195.00	12.80	2,000
DBF	Yeast	114.70	1,627.00	188.00	11.00	4,000

KMGS KMeans Greedy Search

For Lymphoma dataset, the average gene number is greater than SEBI. The average value of condition is better than that of all other algorithms. The average volume is better than SEBI. The average MSR is lower than SEBI. Normally, multiobjective algorithms produce biclusters of larger size compared to greedy algorithms. But in the case of multiobjective evolutionary computation [8], the maximum number of conditions obtained is only 40 for the Lymphoma dataset. On the other hand, in this study, one bicluster with 94 conditions is obtained. The row variance of this bicluster is also above 5,000 (label of bicluster is (p) in Fig. 21.1).

21.6 Statistical Significance and Biological Relevance

Biclusters can be evaluated by using prior biological knowledge [9]. Existence of biologically similar genes is a proof that a specific biclustering technique can produce biologically relevant results. Here, biological relevance is verified by using a small bicluster of size 22×17. For this purpose, GO annotation database is used. In this database, gene products are described in terms of associated biological process, components, and molecular functions in a species-independent manner. For evaluating the statistical significance for the genes in each bicluster, p-values are used. The extent to which the genes in the bicluster match with the different GO categories is indicated by p-values. Smaller p-values indicate better match. Yeast genome gene ontology term finder [10] is a database available in the Internet which can be used to evaluate the biological significance of biclusters. In the bicluster selected for testing the biological significance, there are 22 genes, namely, YAL003W, YBL072C, YBL092W, YBR048W, YBR084C-A, YDL061C, YDL075W, YDL081C, YDL083C, YDL130W, YDL136W, YDL191W, YDL208W, YDL228C, YDR012W, YDR025W, YDR050C, YDR064W, YDR382W, YDR447C, YDR450W, and YDR471W.

The Table 21.2 given below shows the significant GO terms used to describe the set of 22 genes of the bicluster for the process, function, and component ontologies. The common terms are described with increasing order of p-values or decreasing

Table 21.2 GO terms of 22 genes in a small bicluster obtained using KMGS Algorithm

Process	Function	Component
Translation (19, 2.13e 15), Cellular biopolymer biosynthetic process (19, 1.50e 09), Biopolymer biosynthetic process (19, 1.52e 09), cellular protein metabolic process (19, 1.82e 09), gene expression (20, 4.46e 09), cellular macromolecule metabolic process (21, 8.92e 07)	Structural constituent of ribosome (18, 1.88e 23) , structural molecule activity (18, 4.77e 20)	Cytosolic ribosome (18, 9.24e 25) cytosolic part (18, 2.08e 23), ribosomal subunit (18, 2.72e 22) ribosome (19, 3.39e 21), ribonucleoprotein complex (20, 2.94e 17), cytosol (18, 2.44e 14), nonmembrane bounded organelle (19, 1.26e 11)

order of significance. In Table 21.2 the first entry of the column with the title process contains Translation (19, 2.13e-15) which means that 19 out of the 22 genes of the bicluster are involved in the process of translation and their *p*-value is 2.13e-15. Second entry indicates that 19 out of 22 genes are involved in cellular biopolymer biosynthetic process and *p*-value is 1.50e-09. First entry of column function contains structural constituent of ribosome (18, 1.88e-23) which means that 18 out of the 22 genes are involved in this function. This proves that the bicluster contains biologically similar genes. Hence, the method used here is capable of identifying biologically relevant biclusters. Very low *p*-values indicate high statistical significance.

21.7 Conclusion

In this paper, a new algorithm is developed for identifying biclusters from the microarray gene expression data. This KMeans greedy search hybrid algorithm is implemented on both benchmark datasets. In the first step, KMeans clustering algorithm is used to produce bicluster seeds. Then, these seeds are enlarged by greedy method in which the node with minimum incremental increase in MSR score is selected and added to the bicluster in each iteration. Hence, it is possible to get biclusters having more genes and conditions with high coherence. Some of the biclusters have very high row variance also. The statistical significance and biological relevance of biclusters obtained in this method are verified using gene ontology database. In this study, the maximum number of genes (147) is obtained in all the 17 conditions with the minimum MSR value (200.2474) for the Yeast dataset. A bicluster with the maximum number of conditions (94) is obtained for the Lymphoma dataset. The biclusters obtained here show similar upregulation and downregulation under a set of conditions. In terms of size and MSR value, the biclusters obtained in this method are far better than the biclusters obtained in many of the metaheuristic algorithms.

References

1. Yizong Cheng and George M. Church, “Biclustering of expression data”, Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology, 2000. pp. 93 103.
2. Anupam Chakraborty and Hitashyam Maka, “Biclustering of Gene Expression Data Using GeneticAlgorithm”, Proceedings of Computation Intelligence in Bioinformatics and Compu tational Biology CIBCB, 2005. pp. 1 8. Shyama Das and Sumam Mary Idicula, “Modified Greedy Search Algorithm for Biclustering Gene Expression Data”, Proceedings of ADCOM 2009.
3. Tavazoie S., Hughes J. D., Campbell M. J., Cho R. J. and Church G. M., “Systematic determination of genetic network architecture”, Nature Genetics, 1999. Vol. 22, no. 3, pp. 281 285.
4. Alizadeh, A. A. et al., “Distinct types of diffuse large B cell lymphoma identified by gene expression profiling”, Nature, 2000. Vol. 43, no. 6769, pp. 503 511.
5. Federico Divina and Jesus S. Aguilar Ruize, “Biclustering of Expression Data with Evolu tionary computation”, IEEE Transactions on Knowledge and Data Engineering, Vol. 18, 2006. pp. 590 602.
6. Yang J., Wang H., Wang W. and Yu P., “Enhanced Biclustering on Expression Data”, Proceedings of the Third IEEE Symposium on BioInformatics and BioEngineering (BIBE’03, 2003). pp. 321 327.
7. Zhang Z., Teo A., Ooi B. C. and Tan K. L., “Mining deterministic biclusters in gene expression data”, Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioen gineering (BIBE’04), 2004. pp. 283 292.
8. Banka H. and Mitra S., “Multi objective evolutionary biclustering of gene expression data”, Journal of Pattern Recognition, Vol. 39, 2006. pp. 2464 2477.
9. Amos Tanay, Roded Sharan and Ron Shamir, “Discovering Statistically significant Biclusters in Gene Expression Data,” Bioinformatics, Vol. 18 Suppl 1, 2000. pp. S136 S144.
10. SGD GO Termfinder [http://db.yeastgenome.org/cgi bin/GO/goTermFinder].

Chapter 22
Robust Stability Analysis and Design Under Consideration of Multiple Feedback Loops of the Tryptophan Regulatory Network of *Escherichia coli*

A. Meyer-Baese, F. Theis, and M. R. Emmett

Abstract The tryptophan system present in *Escherichia coli* represents an important regulatory unit described by multiple feedback loops. The role of these feedback loops is crucial for the analysis of the dynamical behavior of the tryptophan synthesis. We analyze the robust stability of this system which models the dynamics of both fast state, such as transcription and synthesis of free operator, and slow state, such as translation and tryptophan synthesis under consideration of nonlinear uncertainties. In addition, we analyze the role of these feedback loops as key design components of this regulatory unit responsible for its physiological performance. The range of allowed parameter perturbations and the conditions that ensure the existence of asymptotically stable equilibria of the perturbed system are determined. We also analyze two important alternate regulatory designs for the tryptophan synthesis pathway and derive the stability conditions.

Keywords *E. coli* · Multiple feedback loops · Stability analysis · Tryptophan regulatory network

22.1 Introduction

Biological systems exhibit very complex regulatory mechanisms regardless of whether the situation would require a lower complexity for ensuring nominal functionality. Like the engineering counterpart, the biological system has to robustly function under operational systems. The tryptophan (trp) operon represents one of the best-known molecular systems based on the available experimental data and is therefore like the lactose operon and the phage lambda switch an important candidate for dynamical modeling and analysis. The trp operon of *Escherichia coli*

A. Meyer Baese (✉)
Department of Scientific Computing, Florida State University, Tallahassee, FL 32310 4120, USA
e mail: ameyerbaese@fsu.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 22,

produces the enzymes that are needed to synthesize tryptophan (an essential amino acid) from chorismate. If tryptophan is available, the *E. coli* consumes it and the trp operon is switched off. This process is accomplished based on three different regulatory mechanisms: repression, feedback enzyme inhibition, and transcription attenuation. These three distinct negative feedback loops are quantified using the Hill function.

Several mathematical models are formulated for the tryptophan system [1, 4, 5] which were verified based on the available experimental data. In [5], a detailed mathematical modeling was determined based on these three feedback loops, and their role was assessed in both parameter and structure perturbations. The results based on simulations of the tryptophan system showed that this system is highly robust to parameter perturbations while vulnerable to structural perturbations due to its multiple feedback loops [1].

In this paper, we present a general robust stability analysis method for the trp operon. The system under study models the dynamics of both fast state, such as transcription and synthesis of free operator, and slow state, such as translation and tryptophan synthesis under consideration of nonlinear uncertainties. We analyze the mathematical system as an uncertain two-time scale system and show that we can obtain maximal bounds for fluctuating operating parameters [2]. As shown in [5], the regulation of tryptophan is achieved based on a feedback system of three loops with different tasks, namely, genetic regulation, mRNA attenuation, and enzyme inhibition. The tryptophan is fed back into the system such that it regulates its own synthesis. We mathematically analyze the role of these multiple feedback loops found in the tryptophan synthesis in the overall dynamics of this system. Thus, we enhance the results obtained in [1] by giving a quantitative evaluation of the robustness domain.

22.2 Robust Stability Analysis

The objective of this study is to discuss the robustness properties of the gene regulatory unit of the tryptophan metabolic system based on the mathematical model from a rigorous analytic standpoint and to apply the obtained theoretical results to nonlinear uncertain singularly perturbed systems. In the following, we present the model equations describing the dynamic transcription, translation, and tryptophan synthesis. The mathematical description of the tryptophan regulatory network is given as [1]

$$\begin{aligned}
\dot{O}_{\mathrm{R}} &= k_1 O_{\mathrm{t}} C_1(T) - k_{\mathrm{d}1} O_{\mathrm{R}} - \mu O_{\mathrm{R}} \\
\dot{mRNA} &= k_2 O_{\mathrm{R}} C_2(T) - k_{\mathrm{d}2} mRNA - \mu mRNA \\
\dot{E} &= k_3 mRNA - \mu E \\
\dot{T} &= k_4 C_3(T) E - g \frac{T}{T + K_{\mathrm{g}}} - \mu T
\end{aligned}$$

Table 22.1 Parameter values of the tryphtophan model [1]

Name	Symbol	Value
Kinetic rate constant for free operator	k_1	50 min^{-1}
Kinetic rate constant for mRNA	k_2	15 min^{-1}
Kinetic rate constant for transcription	k_3	90 min^{-1}
Kinetic rate constant for tryptophan synthesis	k_4	59 min^{-1}
Total operator site concentration	O_t	3.32 nM
Specific growth rate of *Escherichia coli*	μ	0.01 min^{-1}
Degradation rate constant for OR	k_{d1}	0.5 min^{-1}
Degradation rate constant for mRNA	k_{d2}	15 min^{-1}
Half saturation constant	K_g	0.2 μM
Kinetic constant for uptake of tryptophan	g	25 μM min^{-1}
Half saturation constant repression	$K_{\text{i},1}$	3.53 μM
Half saturation constant attenuation	$K_{\text{i},2}$	0.04 μM
Half saturation constant inhibition	$K_{\text{i},3}$	810 μM
Sensitivity of genetic regulation	η_H	1.92

where $\dot{O}_\text{R}$ represents the free operator, $m\dot{R}NA$ the concentration of mRNA, $\dot{E}$, and $\dot{T}$ the concentrations of enzyme anthranilate synthase and tryptophan, respectively, in the cell. Parameters k_1, k_2, k_3, and k_4 represent the kinetic rate constants for the synthesis of free operator, mRNA transcription, translation, and tryptophan synthesis, respectively, while O_t, μ, k_{d1}, and k_{d2} represent the total operator site concentration, specific growth rate of *E. coli*, and degradation rate constants of free operator $\dot{O}_\text{R}$ and $m\dot{R}NA$. The specific parameter values for the tryptophan model are given in Table 22.1.

The three controllers $C_1(T)$, $C_2(T)$, and $C_3(T)$ are included in the feedback mechanisms and describe repression, attenuation, and inhibition, respectively, and are modeled by Hill functions:

$$C_1(T) = \frac{K_{\text{i},1}^{\eta_\text{H}}}{K_{\text{i},1}^{\eta_\text{H}} + T^{\eta_\text{H}}}, C_2(T) = \frac{K_{\text{i},2}^{1.72}}{K_{\text{i},2}^{1.72} + T^{1.72}}, C_3(T) = \frac{K_{\text{i},3}^{1.2}}{K_{\text{i},3}^{1.2} + T^{1.2}}$$

where $K_{\text{i},1}$, $K_{\text{i},2}$ and $K_{\text{i},3}$ represent half-saturation constants and η_H the sensitivity of genetic regulation to tryptophan concentration. A better visualization of the feedback mechanisms is illustrated in the block diagram derived from [5] shown in Fig. 22.1.

We proceed by a shifting of the equilibrium of the system to the origin and obtain thus a new system with the variables $\tilde{E} = E - E^*$, $\tilde{T} = T - T^*$, $\tilde{O}_\text{R} = O_\text{R} - O_\text{R}{}^*$, and $m\tilde{R}NA = mRNA - mRNA^*$

$$\begin{aligned}
\varepsilon\dot{\tilde{O}}_\text{R} &= \tfrac{\mu}{B}\tfrac{k_1 O_\text{t}}{k_{\text{d1}}+\mu}(C_1(\tilde{T}+T^*)) - C_1(T^*)) - \tfrac{\mu}{B}\tilde{O}_\text{R} \\
\in m\dot{\tilde{R}}NA &= \tfrac{\mu k_2 \tilde{O}_\text{R} C_2(\tilde{T}+T^*)}{k_{\text{d1}}+\mu} + \tfrac{k_2 O_\text{R}^* \mu}{k_{\text{d1}}+\mu}(C_2(\tilde{T}+T^*) - C_2(T^*)) - \mu m\tilde{R}NA, \\
\dot{\tilde{E}} &= k_3 m\tilde{R}NA - \mu\tilde{E}, \\
\dot{\tilde{T}} &= k_4 C_3(\tilde{T}+T^*)(\tilde{E}+E^*) - k_4 C_3(T^*)E^* - \left[g\tfrac{\tilde{T}+T^*}{T+T^*+K_\text{g}} - g\tfrac{T^*}{T^*+K_\text{g}}\right] - \mu\tilde{T}.
\end{aligned}$$

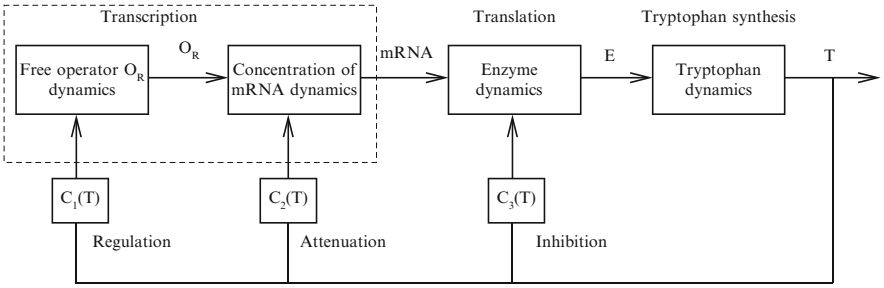


Fig. 22.1 Block diagram of the dynamical behavior of the tryptophan synthesis

The above system can be transformed into a new system that is represented by a slow and fast subsystem. This system is a superposition of a linear nominal system and perturbation terms.

In more general terms, we can rewrite our system in the following form by using the notation $x_{\text{slow}} = [\tilde{E},\tilde{T}]$ and $x_{\text{fast}} = [\tilde{O}, m\tilde{R}NA]$

$$\dot{x}_{\text{slow}} = A_{11}x_{\text{slow}} + A_{12}x_{\text{slow}} + f_{\text{slow}}(x_{\text{slow}}, x_{\text{fast}})$$
$$\varepsilon\dot{x}_{\text{fast}} = A_{21}x_{\text{slow}} + A_{22}x_{\text{fast}} + f_{\text{fast}}(x_{\text{slow}}, x_{\text{fast}}),$$

where

$$\varepsilon = \frac{\mu}{k_{\text{d2}} + \mu},$$

$$B = \frac{\mu + k_{\text{d2}}}{\mu + k_{\text{d1}}},$$

$$A_{12} = \begin{pmatrix} 0 & k_3 \\ 0 & 0 \end{pmatrix},$$

$$A_{22} = \begin{pmatrix} \frac{\mu}{B} & 0 \\ 0 & \frac{\mu}{0} \end{pmatrix},$$

$$A_{21} = 0$$

and

$$A_{11} = \begin{pmatrix} -\mu & 0 \\ 0 & -\mu \end{pmatrix}.$$

Also, we have

$$f_{\text{slow}} = \left(k_4 C_3(\tilde{T} + T^*)(\tilde{E} + E^*) - k_4 C_3(T^*)E^* - \left[g \frac{\tilde{T} + T^*}{\tilde{T} + T^* + K_{\text{g}}} - g \frac{T^*}{T^* + K_{\text{g}}} \right] \right)$$

and

$$f_{\text{fast}} = \begin{pmatrix} \frac{\mu}{B} \frac{k_1 O_{\text{t}}}{k_{\text{d1}} + \mu} (C_1(\tilde{T} + T^*)) - C_1(T^*)) \\ \frac{\mu k_2 \tilde{O}_{\text{R}} C_2(\tilde{T} + T^*)}{k_{\text{d1}} + \mu} + \frac{k_2 O_{\text{R}}^* \mu}{k_{\text{d1}} + \mu} (C_2(\tilde{T} + T^*) - C_2(T^*)) \end{pmatrix}.$$

The nonlinear uncertainties are bounded by

$$||f_{slow}(x_{slow}, x_{fast})|| \leq \alpha_1 ||x_{slow}|| + \beta_1 ||x_{fast}||$$
$$||f_{fast}(x_{slow}, x_{fast})|| \leq \alpha_2 ||x_{slow}|| + \beta_2 ||x_{fast}||$$

where α_1, α_2, β_1, and β_2 are positive constants and $||f||$ is the norm of the perturbed terms.

The linear nominal system form of the above system is given as

$$\dot{x}_{slow}(t) = A_{11} x_{slow}(t) + A_{12} x_{fast}(t)$$
$$\varepsilon \dot{x}_{fast}(t) = A_{21} x_{slow}(t) + A_{22} x_{fast}(t),$$

and its stability is given by the following lemma.

Lemma 22.1 [?, 3]. If A_{22} is nonsingular and both A_{22} and $A_0 = A_{11} - A_{12}A^{-1}_{22} A_{21}$ are Hurwitz matrices, then there exists an $\varepsilon^* > 0$ such that the system (22.7) is asymptotically stable for all $\varepsilon \in [0, \varepsilon^*)$.

Proof. The goal is to determine the upper bounds α_{i}, β_{i}, and ε^* such that the system (22.1) is asymptotically stable assuming that A_{22} and A_0 are Hurwitz. It implies immediately for our system that the Hurwitz property is guaranteed. We will define the following state transformation

$$\begin{bmatrix} x_{slow}(t) \\ x_{fast}(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \frac{k_3 \varepsilon}{\mu} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \varsigma(t) \\ \eta(t) \end{bmatrix}$$

with

$$H = \begin{pmatrix} 0 & -\frac{k_3}{\mu} \\ 0 & 0 \end{pmatrix}.$$

We thus obtain

$$\begin{bmatrix} \varsigma(t) \\ \eta(t) \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \frac{\varepsilon k_3}{\mu} \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_{slow}(t) \\ x_{fast}(t) \end{bmatrix}.$$

Using the above transformation, we can derive a new transformed system

$$\begin{aligned} \dot{\varsigma}(t) &= A_{11}\varsigma(t) - \varepsilon A_{11} \cdot A_{12} \cdot A_{22}{}^{1}\eta(t) + f_1(\varsigma(t), \eta(t)), \\ \varepsilon\dot{\eta}(t) &= -A_{22}\eta(t) + f_2(\varsigma(t), \eta(t)), \end{aligned}$$

where the transformed nonlinear uncertainties are given by

$$\begin{aligned} f_{1\_}(\varsigma(t), \eta(t)) &= f_1[\varsigma(t) + \varepsilon H\eta, \eta(t)] - Hf_2[\varsigma(t) + \varepsilon H\eta, \eta(t)] \\ f_{2\_}(\varsigma(t), \eta(t)) &= f_2[\varsigma(t) + \varepsilon H\eta, \eta(t)]. \end{aligned}$$

The above nonlinear uncertainty can be estimated such as, $||f_i(\varsigma(t), \eta(t))|| \leq \alpha_i||\varsigma(t)|| + \beta_i||\eta(t)||$ with $\alpha_1 = \alpha_1 + \alpha_2||H||$, $\beta_1 = \varepsilon[\alpha_1 + \alpha_2||H||]|\ |H||+\beta_2||H||$, $\alpha_2 = \alpha_2$, and $\beta_2 = \varepsilon\alpha_2||H|| + \beta_2$. The study of the stability of the transformed system (22.10) is equivalent to the stability of the original system (22.1).

Theorem. The system (22.1) is asymptotically stable for $0 \leq \varepsilon < \max \varepsilon^*$, where $\varepsilon^* = \min(\varepsilon_1{}^*, \varepsilon_2{}^*)$ is given by (22.20) and for any $\gamma > 0$, if the inequalities

$$a = 2 - \frac{2\alpha_1}{\mu} - \gamma\left[\frac{||\alpha_{12}||}{\mu} + \frac{\beta_1}{\mu} + \frac{B}{\mu}\alpha_2\right] \geqslant 0$$

$$b = 2 - 2\beta_2\frac{B}{\mu} - \frac{1}{\gamma}\left[\frac{||\alpha_{12}||}{\mu} + \frac{\beta_1}{\mu} + \alpha_2\frac{B}{\mu}\right] \geqslant 0$$

with $\|\alpha_{12}\| = \varepsilon k_3$, $\alpha_1 = \alpha_1 + \alpha_2||H||$, $\beta_1 = \varepsilon[\alpha_1 + \alpha_2||H||]||H||+\beta_2||H||$, $\alpha_2 = \alpha_2$, and $\beta_2 = \varepsilon\alpha_2||H|| + \beta_2$, are satisfied. Then

$$\mu = \beta_2 B > 0$$

and

$$\mu - \alpha_1 - \alpha_2||H|| > 0$$

and

$$4(\mu - \alpha_1 - \alpha_2||H||)(\mu - B\beta_2) > (B\alpha_2 + \beta_2||H||)^2$$

Proof. As a Lyapunov function for the new transformed system, we choose $V(\zeta(t), \eta(t)) = \zeta^{\mathrm{T}}P\zeta + \varepsilon\eta^{\mathrm{T}}R\eta$, where P, and R are the solutions of the Lyapunov equations

$$A_0^{\mathrm{T}}P + PA_0 = -2Q_{\mathrm{P}},$$
$$A_{22}^{\mathrm{T}}R + RA_{22} = -2Q_{\mathrm{R}}.$$

Choosing $Q_{\mathrm{P}} = Q_{\mathrm{R}} = I$, we obtain $P = -A_{11}^{-1}$ and $R = -A_{22}^{-1}$. For the derivative of the Lyapunov function, we can show that using (22.10) and (22.17)

$$\dot{V}(\varsigma(t), \eta(t)) \leqslant \dot{\varsigma}^{\mathrm{T}}P\varsigma + \varsigma^{\mathrm{T}}P\dot{\varsigma} + \varepsilon\dot{\eta}^{\mathrm{T}}R\eta + \varepsilon\eta^{\mathrm{T}}R\dot{\eta}$$
$$= -2\varsigma^{\mathrm{T}}\varsigma - 2\eta^{\mathrm{T}}\eta + 2\varsigma^{\mathrm{T}}Pf_1 + 2\varsigma^{\mathrm{T}}\varepsilon||H||\eta + 2\eta^{\mathrm{T}}f_2.$$

We will make use of the inequality $2||\zeta||||\eta|| \leqslant \gamma||\zeta||^2 + 1/\gamma||\eta||^2$ valid for any $\gamma > 0$ in our Lyapunov function and thus obtain

$$\dot{V}(\varsigma, \eta) \geqslant -\left(2 - \frac{2\alpha_1}{\mu} - \gamma\left[\frac{||\alpha_{12}||}{\mu} + \frac{\beta_1}{\mu} + \alpha_2\frac{B}{\mu}\right]\right)||\varsigma||^2$$
$$-\left(2 - 2\beta_2\frac{B}{\mu} - \frac{1}{\gamma}\left[\frac{||\alpha_{12}||}{\mu} + \frac{\beta_1}{\mu} + \alpha_2\frac{B}{\mu}\right]\right)||\eta||^2,$$

from which we see that $\dot{V}(\zeta, \eta) \leq 0$ means that $a, b \geq 0$ with

$$a = 2 - \frac{2\alpha_1}{\mu} - \gamma\left[\frac{||\alpha_{12}||}{\mu} + \frac{\beta_1}{\mu} + \frac{B}{\mu}\alpha_2\right] \geqslant 0$$

and

$$b = 2 - 2\beta_2\frac{B}{\mu} - \frac{1}{\gamma}\left[\frac{||\alpha_{12}||}{\mu} + \frac{\beta_1}{\mu} + \alpha_2\frac{B}{\mu}\right] \geqslant 0$$

is a sufficient condition for $V(\zeta, \eta) \leq 0$. We thus obtain two equations for ε

$$0 \leqslant \varepsilon \leqslant \varepsilon_1^* = \frac{E - \gamma N}{O\gamma}, 0 \leqslant \varepsilon \leqslant \varepsilon_2^* = \frac{M\gamma - N}{D\gamma + O}$$

with $M = 2\mu - 2\beta_2 B$, $N = B\alpha_2 + \beta_2//H//$, $O = k_3 + \alpha_1//H// + \alpha_2//H//^2$, $D = 2B\alpha_2//H//$, and $E = 2(\mu - \alpha_1 - \alpha_2//H//)$. Since ε_1^* is strictly decreasing and continuous on $\gamma > 0$ and ε_2^* is strictly increasing and continuous on $\gamma > 0$, it follows that $\varepsilon_2^* - \varepsilon_1^*$ is strictly increasing and continuous on $\gamma > 0$. There is a unique $\gamma^* > 0$ such that $\varepsilon_1^*(\gamma^*) = \varepsilon_2^*(\gamma^*)$. This represents the value of $\gamma > 0$ for which $\varepsilon^* = max_{\gamma>0} min\,(\varepsilon_1^*, \varepsilon_2^*)$. By equating ε_1^* and ε_2^*, we obtain from the quadratic polynomial $(\mathrm{MO} + \mathrm{ND})\gamma^2 - \mathrm{DE}\gamma - \mathrm{OE} = 0$ the maximum value of ε^*. This quadratic polynomial has one positive real solution $\gamma^* > 0$ given by

Table 22.2 Stability conditions for two alternate regulatory designs for the tryptophan synthesis pathway

Removal of tryptophan effect	Absence of attenuation and inhibition
$M \quad \beta_2 B > 0$	
$M \quad \alpha_1 \quad \alpha_2 \lVert H \rVert > 0$	$\mu \quad \alpha_1 \quad \alpha_2 \lVert H \rVert > 0$
$4(\mu \quad \alpha_1 \quad \alpha_2 \lVert H \rVert)(\mu \quad \beta_2 B) > (B\alpha_2 + \beta_2 \lVert H \rVert)^2$	$4\mu(\mu \quad \alpha_1 \quad \alpha_2 \lVert H \rVert) > (B\alpha_2)^2$

$$\gamma^* = \frac{\mathrm{DE} + \sqrt{(\mathrm{DE})^2 + 4O(\mathrm{MO} + \mathrm{ND})}}{2(\mathrm{MO} + \mathrm{ND})}.$$

By substituting the above expression into ε^*, we obtain max.

$$\max_{\gamma > 0} \varepsilon^* = \frac{M\gamma^* - N}{O + D\gamma^*} = \frac{E - \gamma^* N}{O\gamma^*}.$$

As shown in [1], we consider two distinct alternate regulatory designs for the tryptophan synthesis pathway: (1) mutated regulatory structure with the effect of tryptophan removed and (2) absence of attenuation and inhibition to regulate the tryptophan synthesis. As a consequence of the fact that $\varepsilon^* > 0$, we get for the two distinct alternate regulatory designs the following stability conditions to be fulfilled, as depicted in Table 22.2.

22.3 Conclusion

We analyzed and identified the relationship between the design of the tryptophan regulatory unit and its physiological function. This robust design of the regulatory mechanism is needed in order to deal with uncertainties while ensuring a stable operating point. We used concepts from the theory of uncertain singularly perturbed systems and applied these results to study robustness properties of the tryptophan system in *E. coli*. This system represents an important control system, where three processes transcription, translation, and tryptophan synthesis, are in series and describe negative feedback loops of the genetic repression, mRNA attenuation, and enzyme inhibition. In addition, it is a multiple time-scale system combining a coupled nonlinear fast and slow dynamics. In this sense, we established robustness stability results for the reduced-order model and determined the conditions that ensure the existence of asymptotically stable equilibria of this model. A sufficient condition for the nonnegative singular perturbation parameter, representing in our context the sum of the specific growth rate of *E. coli* and the degradation rate constant for mRNA, is derived for the uncertain reduced-order model under the condition that the nonlinear uncertainties are bounded.

Acknowledgment The authors thank Felix Buggenthin for his efficient help on typesetting the manuscript. FT acknowledges support by the Helmholtz Alliance on Systems Biology (Project CoReNe).

References

1. S. Bhartiya, N. Chaudhary, K. Venkatesh and F. Doyle (2006), Multiple feedback loop design in the tryptophan regulatory network of *E. coli* suggests a paradigm for robust regulation of processes in series, Journal of the Royal Society Interface, vol. 3, pp. 383 391.
2. Ali Saberi and Hassan Khalil (1984), Quadratic type lyapunov functions for singularly perturbed systems, IEEE Transactions on Automatic Control, vol. AC 29, no. 6, pp. 542 550.
3. M. Santillan and E. Zeron (2004), Dynamic influence of feedback enzyme inhibition and transcription attenuation on the tryptophan operon response to nutritional shifts, Journal of Theoretical Biology, vol. 231, pp. 287 298.
4. Z. Shao (2004), Robust stability of two time scale systems with nonlinear uncertainties, IEEE Transactions on Automatic Control, vol. 49, no. 2, pp. 258 261.
5. K. Venkatesh, S. Bhartiya and A. Ruhela (2004), Multiple feedback loops are key to a robust dynamic performance of tryptophan regulation in *E. coli*, FEBS Letters, vol. 563, pp. 234 240.

Chapter 23
FM-GA and CM-GA for Gene Microarray Analysis

Lily R. Liang, Rommel A. Benites Palomino, Zhao Lu, Vinay Mandal, and Deepak Kumar

Abstract In this paper, we propose two new approaches, FM-GA and CM-GA, to identify significant genes from microarray datasets. FM-GA and CM-GA combine our innovative FM-test and CM-test with genetic algorithm (GA), respectively, and leverage the strengths of GA. The performance of FM-GA and CM-GA was evaluated by the classification accuracy of decision trees constructed with the selected genes. Experiments were conducted to demonstrate the superiority of the proposed method over other approaches.

Keywords Genetic algorithms · Gene microarray · Fuzzy set theory · CM-test · FM-test

23.1 Introduction

Genetic Algorithms (GAs) are powerful tools for solving global optimization problems [1]. They have been proved to have the capability of finding optimal solutions in a large search space within a limited amount of time. However, they may not be applied directly on microarray datasets to identify significant genes. The large number of genes involved makes a GA chromosome extremely long, often several thousands of bits. Thus it is unpractical to apply GA directly to solve this kind of problem. In the work of Yeh et al.[2], the method of *t*-test is proposed to pre-screen the genes before GA is applied. The work produced interesting and convincing results. However, it is known that *t*-test has the following limitations: it cannot

L.R. Liang (✉)
Department of Computer Science and Information Technology, University of the District of Columbia, Washington, DC 20008, USA
e mail: LLiang@udc.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 23,

distinguish two divergent sets with close means and is very sensitive to extreme values [3, 4].

In this paper, two new approaches, FM-GA and CM-GA, are proposed to identify significant genes from microarray datasets. FM-GA and CM-GA integrate our fuzzy-set-theory-based divergence measurements, fuzzy membership test (FM-test), and cluster misclassification test (CM-test) [3, 4] with standard genetic algorithm, respectively, and leverage the strengths of GA [5]. It is proved that FM-test and *t*-test overcome the limitation of *t*-test and out perform *t*-test in identifying significant genes. We conducted experiments on several sets of microarray data and compared FM-GA and CM-GA with other approaches. FM-GA and CM-GA are discussed in detail in Sect. 23.2. Experimental results are demonstrated in Sect. 23.3. We give conclusions in Sect. 23.4.

23.2 Methodology

Both FM-GA and CM-GA are composed of the following two phases: (1) selecting initial list of genes with FM-test or CM-test, respectively, and (2) shortening the list with genetic algorithm.

23.2.1 Selecting Initial List of Significant Genes by FM-Test or CM-Test

In this step, we use either FM-test or CM-test to measure the divergence of each gene between two different groups of samples. CM-test and FM-test are proposed in [3, 4] as innovative approaches that overcome the limitations of *t*-test and rank sum test. The fundamental idea of these two tests is to consider the two groups of values as samples from two different fuzzy sets. We then examine the membership value of each sample with respect to each of these two fuzzy sets. A *d*-value is calculated for each gene using FM-test or CM-test. The top 500 genes with the largest *d*-values are selected and used in the input for the next step.

FM-test [3]. Given two sets S_1 and S_2, the convergence degree between S_1 and S_2 in FM-test is defined as

$$c_{\mathrm{FM}}(S_1, S_2) = \frac{\sum_{e \in s_1} f_{FS_2}(e) + \sum_{f \in S_2} f_{FS_1}(f)}{|S_1| + |S_2|}, \tag{23.1}$$

where FS_1 and FS_2 are the corresponding fuzzy sets. The divergence, FM *d*-value, between S_1 and S_2 is defined as

$$d_{\mathrm{FM}}(S_1, S_2) = 1 - c_{\mathrm{FM}}(S_1, S_2). \tag{23.2}$$

CM-test [4]. If an element of one fuzzy set belongs more to the other fuzzy set, then we say that the element is *misclassified*. The divergence of the original two sets is measured by aggregating the number of misclassified elements and the degree of misclassification as follows:

$$c_{\mathrm{CM}}(S_1, S_2) = \alpha^* + (1 - \alpha)^* T_2. \tag{23.3}$$

Here, α is a weight whose value ranges from 0 to 1, and T_1, and T_2 are normalized number of misclassified samples and degree of misclassification, respectively. T_1 and T_2 are defined in (23.4) and (23.5).

$$T_1 = \frac{|M_{\mathrm{FS2}}(S_1) + M_{\mathrm{FS1}}(S_2)|}{|S_1| + |S_2|} \tag{23.4}$$

$$T_2 = \frac{\sum\limits_{e \in s_1} m(e, S_2) \sum\limits_{f \in S_2} m(f, S_1)}{|S_1| + |S_2|} \tag{23.5}$$

CM d-value, the cluster misclassification divergence degree, between S_1 and S_2 is defined as

$$d_{\mathrm{CM}}(S_1, S_2) = 1 - c_{\mathrm{CM}}(S_1, S_2). \tag{23.6}$$

23.2.2 Shortening the List of Genes with GA

In this step, the genes identified in the previous step are subjected to GA to select the most significant ones among them. A standard genetic algorithm was adopted from [1].

- *Encoding*. Chromosome length is the number of genes in the list generated by the previous step. The genes are binary encoded. A "1" means that the gene is present and a "0" that it is not.
- *Fitness evaluation*. We define the fitness value of a chromosome as the classification accuracy obtained by a classifier using the genes presented in that chromosome. We used decision tree C4.5 as the classifier. Library functions of Weka were used for the implementation [6].
- *Selection*. The individual chromosomes are selected using a roulette wheel, where each individual is weighted according to its fitness value.
- *Crossover*. Two selected individual chromosomes are crossovered under a given probability. Their offspring will be the new individuals for the next population.

- *Mutation*. We mutate the chromosomes by changing one randomly selected bit from 0 to 1 or vice versa with a given probability.

The evaluation, selection, crossover, and mutation procedures are repeated for each generation until the maximum number of generations is reached. Then, the best individual of the last GA population is used to determine the genes that should be present in the final significant gene list. Only the genes represented as "1" in that individual are selected to be in the list.

23.3 Experimental Results

Our experiments were performed on four sets of microarray data below.

Lung Cancer Data [7]. Ten individuals, out of which five were diagnosed with lung cancer and five were not. 22,283 genes.

CNS Cancer Data [8]. Sixty individuals, out of which thirty-nine were survivors and twenty-first were treatment failures. 7,129 genes.

Diabetes Dataset No. 1 [9]. Thirty-four individuals, out of which seventeen were diagnosed with diabetes and seventeen were not. 15,056 different genes.

Diabetes Dataset No. 2 [10]. Ten individuals, out of which five were insulin-sensitive (IS) and five were insulin-resistant (IR). 10,831 genes. Only the genes with no null expression values were analyzed.

We compared FM-GA and CM-GA with *t*-GA to identify the benefit brought by FM-test and CM-test. We also compared FM-GA and CM-GA with FM-test, CM-test, and *t*-test to identify the benefits brought by GA. As shown in Table 23.1, of the four datasets, (1) CM-GA gives better results than CM-tests alone on all sets of data, (2) FM-GA gives better results than FM-test alone on three sets of data and equal accuracy on one set of data, and (3) FM-GA and CM-GA have accuracies that are higher than or equal to that of *t*-GA and *t*-test.

To evaluate the results of CM-GA and FM-GA, we selected top ten differentially regulated genes in the lung cancer data and searched published literature to validate their relevance in tumor growth and development. The most significant gene identified by CM and CM-GA method is the PI 3-kinase or PI3K, which is one of the extensively studied and commonly activated signaling proteins in human

Table 23.1 Classification accuracy (%) obtained by decision trees constructed with genes identified

	Lung cancer	CNS Tumor	Diabetes dataset No. 1	Diabetes dataset No. 2
CM GA	100	76.67	91.18	90
FM GA	90	80	91.18	90
t GA	90	70.49	70.59	90
CM test	90	59.02	79.41	80
FM test	80	58.33	73.53	90
t test	90	46.67	32.35	80

cancers [11]. It plays essential roles in many cellular processes, including cell survival, proliferation, and differentiation. PI3 kinases are major effector molecules downstream of receptor tyrosine kinases (RTKs) and G protein-coupled receptors (GPCRs). The PI3Ks transduce signals from various growth factors and cytokines into intracellular messages by generating phospholipids, which activate the serine threonine protein kinase AKT (also known as protein kinase B (PKB)) and other downstream effector pathways. PI3 kinases are important molecular targets of several experimental drugs and chemotherapeutic and chemopreventive compounds [11]. Identification of PI3 kinases is highly significant and validates the methods to identify relevant genes. Adenosine monophosphate deaminase 1 (AMPD1) was also identified as a significant gene. AMPD1 catalyzes the deamination of AMP to IMP and plays an important role in the purine nucleotide cycle. Although differential regulation of AMPD1 is not known, adenine nucleotide catabolism is altered in tumor tissues, suggesting functional investigation of AMPD1 in cancer cells [12]. The cytochrome P450 proteins are monooxygenases involved in drug metabolism and synthesis of cholesterol, steroids, and other lipids [13]. Expression, allelic imbalances, and mutations such as SNPs of both CYP3A4 and CYP2A13 that were identified are associated with cancer risk, prognosis, and drug resistance [14]. Another important oncogene family that was identified as significant using this method was the ETS family of transcription factors that has been characterized in various cancers [15]. The members of this gene have been shown to be the target of various oncogenic signaling pathways, such as ras-raf-MAPK signaling cascade. Caspase 10 or the Fas-associated death domain protein interleukin-1b-converting enzyme 2 mRNA is processed during apoptosis and caspase 8, and in turn cleaves caspase 3 and 7 executioner caspases to induce apoptotic cell death. Inhibition of apoptosis is the hallmark of cancer cells and modulation/mutations in caspase 10 have been shown in various cancers [16]. FBN1 or Fibrillin 1 is a member of the fibrillin family of glycoproteins in calcium-binding microfibrils that supports elastic and nonelastic connective tissue throughout the body. Isolated uncharacterized reports of FBN1 alterations in human cancers suggest further investigation of this molecule. Another important target that was identified using our method was GNAI1, the Guanine nucleotide-binding protein (G protein), alpha-inhibiting activity polypeptide 1. G proteins are known molecular switches in oncogenic signaling pathways [17]. Overall, the validation of top ranking molecules based on the published literature suggests that CM-GA and FM-GA are efficient and superior methods for identifying significantly modulated genes in target gene expression datasets.

23.4 Conclusions and Future Work

In this paper, we proposed CM-GA and FM-GA as new approaches to identify significant genes of diseases. Experiments were conducted on real-world datasets. Results were compared with CM-test, FM-test, *t*-test, and *t*-GA. CM-GA and

FM-GA have the highest overall accuracy among all these methods. Our future research will include applying this approach to other feature selection problems and investigating ways of improving the performance of GA.

Acknowledgment This work was supported by the Agriculture Experiment Station at the University of the District of Columbia.

References

1. Goldberg DE (1989) Genetic Algorithms in Search, Optimization, and Machine Learning. Addison Wesley, Reading.
2. Yeh JY (2008) Applying Data Mining Techniques for Cancer Classification from Gene Expression Data. Cybernetics and Systems 39: 583 602.
3. Liang LR, Lu SY, Wang XN et al (2006) FM test: A Fuzzy Set Theory Based Approach to Differential Gene Expression Data Analysis. BMC Bioinformatics 7 (Suppl 4): S7.
4. Liang LR, Lu SY, Lu Y et al (2006) CM test: An Innovative Divergence Measurement and Its Application in Diabetes Gene Expression Data Analysis. In proceedings of 2006 IEEE International Conference on Granular Computing. Atlanta, GA.
5. Benites Palomino RA, Liang RL, Lu Z et al (2009) Identifying Significant Genes with FM/CM GA. In Proceedings of 11th WSEAS International Conference on Mathematical and Computational Methods in Science and Engineering (MACMESE'09). Baltimore, Maryland, USA.
6. Hall M, Frank E, Holmes G et al (2009) The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1).
7. Wachi S, Yoneda K, Wu R (2005) Interactome transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics 21(23): 4205 4208.
8. Pomeroy SL, Tamayo P, Gaasenbeek M et al (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. Nature 415: 436 442.
9. Mootha VK, Lindgren CM, Eriksson KF et al (2003) PGC 1alpha responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nature Genetics 34: 267 273.
10. Yang X, Pratley RE, Tokraks S et al (2002) Microarray profiling of skeletal muscle tissues from equally obese, nondiabetic insulin sensitive and insulin resistant Pima Indians. Diabe tologia 45(11): 1584 1593.
11. Engelman JA (2009) Targeting PI3K signaling in cancer: opportunities, challenges and limitations. Nature Reviews Cancer 8: 550 562.
12. Tong X, Zhao F, Thompson CB (2009) The molecular determinants of de novo nucleotide biosynthesis in cancer cells. Current Opinion in Genetics and Development 1: 32 37.
13. Androutsopoulos VP, Tsatsakis AM, Spandidos DA (2009) Cytochrome P450 CYP1A1: Wider roles in cancer progression and prevention. BMC Cancer 9: 187.
14. Timofeeva MN, Kropp S, Sauter W et al (2009) CYP450 polymorphisms as risk factors for early onset lung cancer: Gender specific differences. Carcinogenesis 30: 1161 1169.
15. Hahne JC, Okuducu AF, Sahin A et al (2008) The transcription factor ETS 1: Its role in tumour development and strategies for its inhibition. Mini Reviews in Medicinal Chemistry 11: 1095 1105.
16. Harada K, Toyooka S, Shivapurkar N et al (2002) Deregulation of caspase 8 and 10 expression in pediatric tumors and cell lines. Cancer Research 62: 5897 5901.
17. Hurst JH, Hooks SB (2009) Regulator of G protein signaling (RGS) proteins in cancer biology. Biochemical Pharmacology 78(10): 1289 1297.

Part III
Protein Classification & Structure Prediction, and Computational Structural Biology

Chapter 24
Novel Features for Automated Cell Phenotype Image Classification

Loris Nanni, Sheryl Brahnam, and Alessandra Lumini

Abstract The most common method of handling automated cell phenotype image classification is to determine a common set of optimal features and then apply standard machine-learning algorithms to classify them. In this chapter, we use advanced methods for determining a set of optimized features for training an ensemble using random subspace with a set of Levenberg Marquardt neural networks. The process requires that we first run several experiments to determine the individual features that offer the most information. The best performing features are then concatenated and used in the ensemble classification. Applying this approach, we have obtained an average accuracy of 97.4% using the three best benchmarks for this problem: the 2D HeLa dataset and both the endogenous and the transfected LOCATE mouse protein subcellular localization databases.

Keywords Pattern classification and recognition · Image processing in medicine and biological sciences

24.1 Introduction

A major goal in the biological sciences is to understand the function of proteins at the cellular level [2]. Protein localization is also critical in many practical applications, such as in the design of drug screening systems and for drug discovery and early diagnosis of disease. In [8], for example, an automated system was able to learn to identify cancer by comparing the subcellular protein location patterns in normal tissues and those in cancerous tissues.

S. Brahnam, (✉)
Computer Information Systems, Missouri State University, 901 S. National, Springfield, MO 65804, USA
e mail: sbrahnam@missouristate.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 24,

Automated protein subcellular localization was first proposed by Eisenhaber and Bork [6] and Nakai and Horton [16]. Their methods, however, resulted in marginal performance, ranging in accuracy from 60 to 80%. Over the last decade, new methods based on the visual inspection of fluorescence microscopy images have been studied extensively.

Most work to date train common image descriptors using general purpose classifiers. Of note are the Haralick texture measures [13], Zernike moments [10], Gabor filters, and many other ad hoc measures [5]. More recently, new machine learning algorithms that are based on fusion methods at the feature level or score level have been developed. At the feature level, vectors are created by concatenating several descriptors (see, for example, [3, 10, 14]). Studies in this problem domain that are based on feature level fusion primarily include those methods proposed by Murphy's group [1, 3, 4]. In these studies, several types of descriptors (named Subcellular Location Features, or SLF) are concatenated using a method proposed in [10]. A multiresolution approach, proposed in [2], trained classifiers using descriptors extracted from different resolution spaces. Some examples of fusion at the score level are Refs. [15, 17, 18].

In this chapter, we propose an optimized set of features for training a random subspace of Levenberg Marquardt neural networks that is based on previous work [17, 18]. We run experiments to optimize the performance of variants of the Local Binary Pattern texture descriptors, the wavelet descriptor, and the Haralick descriptor. We then concatenate the best performing features and use them to train an ensemble. We test our approach using data from three different benchmark datasets.

24.2 Descriptors for Cell Phenotype Images

Two types of descriptors are proposed and examined for this problem [17]: global descriptors, which are extracted from the whole image and local descriptors, which are extracted by focusing on various local regions within the image (see [17] for details). In order to deal with the nonuniform illumination problem, images are preprocessed by applying the contrast limited adaptive histogram equalization (using adapthisteq.m in MATLAB 7). Below, we describe the descriptors used in our proposed approach.

Invariant Local Binary Patterns: It is an extensively studied local texture operator [19]. The Local Binary Pattern is a histogram that is based on the $\mathrm{LBP}_{P,R}$ statistical operator. It is calculated by examining the joint distribution of gray scale values of a circularly symmetric neighbor set of P pixels around a pixel $\mathbf{x}$ on a circle of radius R. In this study, we use the long descriptor that is obtained by concatenating three histograms calculated with: $(P = 8; R = 1)$, $(P = 16; R = 2)$, and $(P = 24; R = 3)$.

Local Ternary Patterns: It is a generalization of the Local Binary Pattern [21]. The difference between the gray value of a pixel $\mathbf{x}$ from the gray values in one of its

neighborhood $\mathbf{u}$ assumes the three values by application of a threshold τ (here $\tau = 3$): 1 if $\mathbf{u} \geq \mathbf{x} + \tau$; -1 if $\mathbf{u} \leq \mathbf{x} - \tau$; else 0. The ternary pattern is split into two binary patterns by considering the positive and negative components. The histograms computed from these two patterns are then concatenated. In this study, we use the descriptor obtained by concatenating three histograms calculated with: ($P = 8$; $R = 1$) and ($P = 18$; $R = 2$).

Threshold adjacency statistics: The threshold adjacency statistics (*TAS*) descriptor was first proposed by Hamilton et al. [10] to classify phenotype images. TAS is based on multiple binarizations of the image using different thresholds and the calculation of a histogram for each of these binary images.

Haralick texture features: The Haralick texture features descriptor was proposed 30 years ago by Haralick [11]. It is based on the Spatial Gray Level Dependence Matrices (SGLD), each element in the matrix is a count of the total number of pairs of gray levels i and j at a distance d along the direction θ.

Thirteen features are calculated from a SGLD matrix at a fixed angle θ: energy, correlation, inertia, entropy, inverse difference moment, sum average, sum variance, sum entropy, difference average, difference variance, difference entropy, and two information measures of correlation. There are several methods for extracting features using SGLD:

- *Haar1*, which concatenates the features extracted by considering the two angles of 0° and 90°, with $d = 1$, as proposed in [17]
- *Haar2*, which adds the mean and the range of the 13 measures, thus increasing the number of features to 26, as proposed in [2]
- *Haar3*, which concatenates the features extracted by considering the four angles 0°, 45°, 135°, and 90°, with $d = 1$
- *Haar4*, which concatenates the features extracted by considering the four angles 0°, 45°, 135°, and 90°, with $d = 1$ and $d = 3$

Wavelet features: The method we used is that proposed by Huang and Murphy [14], where the average energy of the three high-frequency components is calculated up to the tenth level decomposition using both the scaling and the wavelet functions of the Coiflets wavelet.

24.3 The Proposed Approach

The approach we propose in this chapter for classifying protein subcellular localization images extends state-of-the-art methods for calculating a set of local and global descriptors [15, 18]. These descriptors are then used to train a random subspace of Levenberg Marquardt neural networks [9] with five hidden nodes. The random subspace method [12] reduces the feature set by randomly drawing a subset of all available features to train classifiers within an ensemble. The random subspace ensemble is made up of 100 neural networks combined by sum rule training each using a random subset of 2/3 of all the features.

24.4 Experimental Results

Our experimental results were obtained using a fivefold cross validation procedure, where we consider the accuracy as the performance indicator. The benchmark datasets used in our experiments are the following:

- *LOCATE mouse protein subcellular localization database* [7]. This dataset is actually two: the endogenous and the transfected datasets. They contains approximately 50 images, with each image containing between 1 and 13 cells, per class.
- *2D HeLa dataset* [2]. This dataset contains 862 images.

In Table 24.1, we compare global descriptors. LTP clearly outperforms LBP, and *HAAR4* is the best Haralick-based method. In Table 24.2, we compare results using the different wavelets descriptors. In Table 24.3, we report the performance obtained by the Coiflet wavelet as we vary the number of levels of decomposition.

In Table 24.4, we report the performance of several global descriptors:

- LPQ, specifically, the descriptor as proposed in [20]
- Gabor, where a bank of Gabor filters with five scales and six orientations is used and the mean and the standard deviation of the resulting image are taken as the texture features [14]
- Edge, using the five edge features (ranging from *SLF1.9* to *SLF1.13* of *SLF* descriptors) as proposed in [1]

Table 24.1 Accuracy obtained using different feature sets

Descriptor	LBP	LTP	HAAR1	HAAR2	HAAR3	HAAR4
LOCATE endogenous	0.930	0.957	0.819	0.836	0.831	0.895
LOCATE transfected	0.841	0.890	0.828	0.845	0.878	0.918
2D HeLa	0.875	0.908	0.817	0.859	0.828	0.879
Average	0.882	**0.918**	0.821	0.846	0.845	0.897

Table 24.2 Comparison among different wavelets

	Haar	Db4	Sym2	Coif2	Bior2.2	Rbio2.2
LOCATE endogenous	0.872	0.884	0.882	0.901	0.878	0.874
LOCATE transfected	0.868	0.867	0.880	0.894	0.852	0.890
2D HeLa	0.822	0.850	0.830	0.852	0.826	0.822
Average	0.854	0.867	0.864	**0.882**	0.852	0.862

Table 24.3 Comparison among different number of levels of decomposition

	5	10	15
LOCATE endogenous	0.842	0.901	0.910
LOCATE transfected	0.826	0.894	0.900
2D HeLa	0.764	0.852	0.832
Average	0.8107	**0.882**	0.880

Table 24.4 Comparison among different texture descriptors

Descriptor	Length of the feature vector	LOCATE endogenous	LOCATE transfected	2D HeLa
LBP	54	0.930	0.841	0.875
TAS	27	0.936	0.848	0.805
Haar1	26	0.803	0.841	0.846
Daub	30	0.911	0.867	0.852
Edge	5	0.515	0.663	0.582
Hull	3	0.137	0.115	0.191
Gabor	60	0.889	0.830	0.830
Zer	49	0.776	0.662	0.589
LPQ	256	0.868	0.876	0.785
LTP	64	0.957	0.890	0.908
HAAR4	104	0.895	0.918	0.879
WAVE	30	0.901	0.894	0.852

Table 24.5 Accuracy obtained by different approaches in the three tested datasets

Method	Length of the feature vector	LOCATE endogenous	LOCATE transfected	2D HeLa	Average
Chebira	78			0.954	
Hamilton	53	0.947	0.933	0.889	0.923
Lin	680			0.936	
Nanni	107	0.984	0.933	0.942	0.953
L + G	145	0.984	0.963	0.958	0.968
GLO1	121	0.989	0.959	0.949	0.965
GLO2	255	0.991	0.953	0.955	0.966
Conc	305	0.995	0.970	0.958	0.974

- Hull, using the features ranging from *SLF1.14* to *SLF1.16* of the *SLF* descriptor as proposed in [1] to measure the convex hull of a cell image
- Daub, which are the descriptors extracted from the Daubechies wavelet [14]
- Zer, the Zernike features

All the texture descriptors are used to train the ensemble of neural networks. In this way, the results are directly comparable. In Table 24.5, we report the performance obtained by concatenating different feature sets. In Table 24.5, the results of these previous studies are named as follows:

- *Chebira*, the multiresolution approach proposed by Chebira et al. [2]
- *Hamilton*, the ensemble of neural networks trained using the concatenation between the Haralick texture features and the TAS features, as proposed in [10]
- *Lin*, a novel AdaBoost method proposed in [15]
- *Nanni*, the method proposed in [17]
- *L + G*, the method proposed in [18]
- GLO1, the concatenation of the global descriptors (*LTP*, *WAVE*, and *TAS*)
- GLO2, the concatenation of the global descriptors (*LTP*, *WAVE*, *HAAR4* and *TAS*)

- *Conc*, the concatenation of the global descriptors (*LTP*, *WAVE*, *HAAR4* and *TAS*) and two local descriptors (*LBP* and *HAAR1*) extracted as in [17]

Notice that the performance obtained by *Conc* is higher than the performance obtained by the other state-of-the-art methods.

24.5 Conclusion and Discussion

This chapter is focused on the study of texture descriptors for training an ensemble of machine learning algorithms for cell phenotype image classification. Based on an analysis of prior research, we propose a new method for automating subcellular protein localization based on a fusion of Levenberg Marquardt neural networks using random subspace combined by Borda count.

The ensemble proposed in this work has been tested on three datasets which are the most widely used benchmarks for comparing automated cell phenotype image classification approaches: the 2D HeLa dataset and both the endogenous and transfected LOCATE mouse protein sub-cellular localization databases.

In our classification experiments, we obtain an average accuracy of 97.4%, using the above cited datasets.

References

1. Boland MV & Murphy RF (2001). A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of HeLa cells. *Bioinformatics*, 17:1213 1223.
2. Chebira A, Barbotin Y, Jackson C, Merryman T, Srinivasa G, Murphy RF & Kovačević J (2007). A multiresolution approach to automated classification of protein subcellular location images. *BMC Bioinformatics*, 8:210.
3. Chen X & Murphy RF (2005). Objective clustering of proteins based on subcellular location patterns. *Journal of Biomedicine Biotechnology*, 2:87 95.
4. Chen S C, Zhao T, Gordon GJ & Murphy RF (2007). Automated image analysis of protein localization in budding yeast. *Bioinformatics*, 23:66 71.
5. Conrad C, Erfle H, Warnat P, Daigle N, Lorch T, Ellenberg J, Pepperkok R & Eils R (2004). Automatic identification of subcellular phenotypes on human cell arrays. *Genome Research*, 14(6):1130 1136.
6. Eisenhaber F & Bork P (1998). Wanted: subcellular localization of proteins based on sequence. *Trends in Cell Biology*, 8:169 170.
7. Fink JL, Aturaliya RN, Davis MJ, Zhang F, Hanson K, Teasdale MS & Teasdale RD (2006). LOCATE: a protein subcellular localization database. *Nucleic Acids Research*, 34(database issue):D213 D217.
8. Glory E, Newberg J & Murphy RF (2008). Automated comparison of protein subcellular location patterns between images of normal and cancerous tissues. In *ISBI*, 304 307.
9. Hagan MT & Menhaj M (1999). Training feed forward networks with the Marquardt algo rithm. *IEEE Transactions on Neural Networks*, 5(6):989 993.

10. Hamilton N, Pantelic R, Hanson K & Teasdale RD (2007). Fast automated cell phenotype classification. *BMC Bioinformatics*, 8:110.
11. Haralick RM (1979). Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):768 804.
12. Ho TK (1998). The random subspace method for constructing decision forests. *IEEE Trans actions on Pattern Analysis and Machine Intelligence*, 20(8):832 844.
13. Huang K & Murphy RF (2004). Automated classification of subcellular patterns in multicell images without segmentation into single cells. In *IEEE International Symposium on Biomedi cal Imaging: Nano to Macro*, Arlington, VA, USA, 1139 1142.
14. Huang K & Murphy R (2004). Boosting accuracy of automated classification of fluorescence microscope images for location proteomics. *BMC Bioinformatics*, 5:78, doi:10.1186/1471 2105 5 78.
15. Lin CC, Tsai Y S, Lin Y S, Chiu T Y, Hsiung C C, Lee M I, Simpson JC & Hsu C N (2007). Boosting multiclass learning with repeating codes and weak detectors for protein subcellular localization. *Bioinformatics*, 23(24):3374 3381.
16. Nakai K & Horton P (1999). Psort: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Science*, 24:34 35.
17. Nanni L & Lumini A (2008). A reliable method for cell phenotype image classification. *Artificial Intelligence in Medicine*, 43(2):87 97.
18. Nanni L & Lumini A (2009). Ensemble of neural networks for automated cell phenotype image classification. *Biomedical Image Analysis and Machine Learning Technologies: Appli cations and Techniques*.
19. Ojala T, Pietikainen M & Maeenpaa T (2002). Multiresolution gray scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971 987.
20. Ojansivu V & Heikkilä J (2008). Blur insensitive texture classification using local phase quantization. In *Proc. 3rd International Conference on Image and Signal Processing (ICISP 2008)*, volume 5099 of LNCS, Springer, Berlin, 236 243.
21. Tan X, Triggs B (2007). Enhanced local texture feature sets for face recognition under difficult lighting conditions. In *Analysis and Modelling of Faces and Gestures*, volume 4778 of LNCS, Springer, Heidelberg, 168 182.

Chapter 25
A Relational Fuzzy *C*-Means Algorithm for Detecting Protein Spots in Two-Dimensional Gel Images

Shaheera Rashwan, Talaat Faheem, Amany Sarhan, and Bayumy A.B. Youssef

Abstract Two-dimensional polyacrylamide gel electrophoresis of proteins is a robust and reproducible technique. It is the most widely used separation tool in proteomics. Current efforts in the field are directed at the development of tools for expanding the range of proteins accessible with two-dimensional gels. Proteomics was built around the two-dimensional gel. The idea that multiple proteins can be analyzed in parallel grew from two-dimensional gel maps. Proteomics researchers needed to identify interested protein spots by examining the gel. This is time consuming, labor extensive and error prone. It is desired that the computer can analyze the proteins automatically by first detecting, then quantifying the protein spots in the 2D gel images. This paper focuses on the protein spot detection and segmentation of 2D gel electrophoresis images. We present a new technique for segmentation of 2D gel images using the Fuzzy *C*-Means (FCM) algorithm and matching spots using the notion of fuzzy relations. Through the experimental results, the new algorithm was found out to detect protein spots more accurately, then the current known algorithms.

Keywords 2D Gel images · Protein spot detection · Fuzzy *C*-Means algorithm · Fuzzy relations

25.1 Introduction

The last decade in life sciences was deeply influenced by the development of the "Omics" technologies (genomics, transcriptomics, proteomics, and metabolomics), which aim for a global view on biological systems. With these tools at hand, the

S. Rashwan (✉)
Informatics Research Institute, Mubarak City for Science and Technology, Borg ElArab, Alexandria, Egypt
e mail: rashwan.shaheera@gmail.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 25 ,

scientific community is striving to build functional models to develop a global understanding of the living cell [1 3].

The analysis of the proteome as the final level of gene expression started out with techniques based on 2D gel electrophoresis [4, 5] and extended its reach with semigel-free and shot gun gel-free liquid chromatography mass spectrometry (LC MS)-based techniques in recent years.

Quantitative analysis based on LC MS techniques is still in an early stage when considering available software and algorithms. Here, we focus on the computerized analysis of 2D gels which are widely used in the scientific community. Two-dimensional gels may separate up to 10,000 protein spots on one gel [6]. In a suitably equipped and experienced lab environment, 2D gels are easy to handle, and they can be produced in a highly parallelized way.

On a proteome map, one can detect all spots of a whole experiment in a single gel image, whereas the average images proposed earlier suffer from dilution effects for weak and rare spots. The spots detected there can serve as a spot consensus pattern that is valid for the whole gel set of the experiment. The consensus spot pattern is then transferred according to the warping transform and used on all gels. This allows for 100% matching spots and, in turn, complete expression profiles for reliable statistical analysis [7, 8].

The goals of this step Protein Spot detection are to find the spot positions, find their surrounding boundary, and determine their quantities. There are two basic approaches that are used in current software: image segmentation and model-based quantitation. The segmentation approach partitions the image into nonoverlapping segments, essentially classifying each pixel as belonging to a certain spot, or as being part of the background between spots. The advantage of this approach is that the image is clearly separated into spots and "nonspot" areas which are easy to assess by a user. If the software allows editing of spot boundaries, then any desired spot shape can, in principle, be obtained. Model-based approaches try to model a spot's intensity as a Gaussian normal distribution or some variant thereof. A spot's quantity and boundaries are then derived from the model (Fig. 25.1).

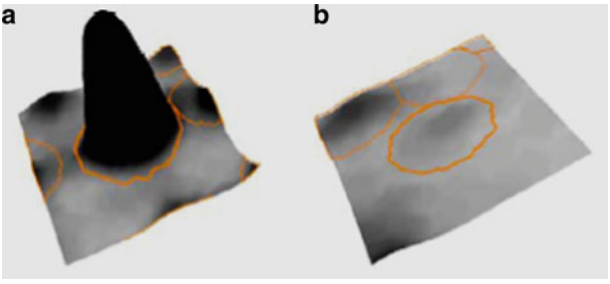


Fig. 25.1 Spot boundaries for high (**a**) and low (**b**) abundance spots

25.2 Previous Developments

Previous developments in this area have employed a wide range of techniques, and a majority of applications address the matching of 2D gel electrophoresis images for proteins as opposed to DNA. This is because 2D gel electrophoresis images for proteins have a relatively more uniform background and are somewhat easier to work with than the 2D gel electrophoresis images for DNA.

Kim et al. [9] proposed a hierarchical segmentation based on thresholding and the detection of watersheds. They first preprocess the images to remove noise and enhance contrast, then thresholding is applied which produces large regions. A watershed detection algorithm is then applied recursively on these regions until only a single blob is detected which is considered to be a spot.

Sugahara et al. [10] smoothed image regions by averaging pixel intensities using an mxm window and performed a thresholding operation which ultimately subtracted the background, and then created a binary image for spot detection.

Takahashi et al. [11] performed image enhancement and smoothing before defining local maxima in order to label the spots.

Morris et al. [12] developed a very accurate and robust method of detecting spots in 2D gel electrophoresis images. Their process involves an "average gel" which is created by first using registration software to create an alignment of all gels being used. The pixel intensities are then averaged across the aligned gels.

In their paper [13], Umer Z. ljaz et al. presented a technique that uses the clustering techniques like *K*-means and Fuzzy *C*-Means (FCM) to distinguish between different types of protein spots and unwanted artifacts.

Christopher S. Hoeflich et al. [14] presented a new technique using the labeling of each image pixel as either a spot or nonspot and use a Markov Random Field (MRF) model and simulated annealing for inference. Neighboring spot labels are then connected to form spot regions.

Dimitris K. Iakovidis et al. [15] presented a novel approach to unsupervised protein spot detection in 2D-PAGE images based on a genetic algorithm. This algorithm searches within a multidimensional parameter space to determine, in parallel, the parameters of multiple diffusion models that optimally fit the characteristics of possible spots.

In their paper [16], Tsakanikas et al. introduce the use of Active Contours without Edges coupled with Contour Transform-based image enhancement for extracting accurately the gel image foreground (regions with spots) from the background.

In [17], Yoon et al. to find protein spots more accurately and reliably from gel images, propose Reversible Jump Markov Chain Monte Carlo method (RJMCMC) to search for underlying spots which are assumed to have Gaussian-distribution shape.

In this work, we intend to present a novel algorithm that uses the FCM algorithm as a primary step in the segmentation process.

The paper is organized as follows. Section 25.3 presents the FCM segmentation algorithm. Protein Spot detection using FCM and introducing the notion of fuzzy relations and a parameter estimation of the Relational Fuzzy *C*-Means (RFCM) new algorithm are described in Sects. 25.4 and 25.5, respectively. Algorithm validation on 2D gel images of human leukemia is shown in Sect. 25.6. Discussion and conclusions are presented in Sect. 25.7.

25.3 The Fuzzy *C*-Means Segmentation Algorithm

FCM method, also known as Fuzzy ISODATA, which was originally introduced by Bezdek in 1981 as an extension to Dunn's algorithm [18] is the most widely used fuzzy clustering algorithm in practice.

FCM is a data clustering technique based on optimizing the objective function:

$$J(U,V) = \sum_{j=1}^{C} \sum_{i=1}^{N} (\mu_{ij})^m \left\| x_i - v_j \right\|^2 \tag{25.1}$$

It requires every data point in the data set to belong to a cluster to some membership degree. The purpose of the FCM is to group data points into different specific clusters. Let $X = \{x_1, x_2, \ldots, x_N\}$ be a collection of data.

By minimizing the objective function (25.1), X is classified into C homogeneous clusters. Where μ_{ij} is the membership degree of data x_i to a fuzzy cluster set v_j, $V = \{v_1, v_2, \ldots, v_c\}$ are the cluster centers. $U = (\mu_{ij})_{N*c}$ is a fuzzy partition matrix, in which each μ_{ij}indicates the membership degree for each data point in the data set to the cluster j. The value of U should satisfy the following conditions:

$$\mu_{ij} \in [0,1], \quad \forall i = 1, \ldots, N, \quad \forall j = 1, \ldots, C \tag{25.2}$$

$$\sum_{j=1}^{C} \mu_{ij} = 1, \quad \forall i = 1, \ldots, N \tag{25.3}$$

The $\left\| x_i - v_j \right\|$is the Euclidean distance betweenx_i and v_j. The parameter m is called fuzziness index that control the fuzziness of membership of each datum. The goal is to iteratively minimize the aggregate distance between each data point in the data set and cluster centers until no further minimization is possible. The whole FCM process can be described in the following steps.

Step 1: Initialize the membership matrix U with random values, subject to satisfying conditions (25.2) and (25.3).

Step 2: Calculate the cluster center V by using the following equation

$$v_j = \frac{\sum_{i=1}^{N} (\mu_{ij})^m x_i}{\sum_{i=1}^{N} (\mu_{ij})^m}, \quad \forall j = 1, \ldots, C \tag{25.4}$$

Step 3: Get the new distance:

$$d_{ij} = \left\| x_i - v_j \right\|, \quad \forall i = 1, \ldots, N, \quad \forall j = 1, \ldots, C \tag{25.5}$$

Step 4: Update the Fuzzy partition matrix U:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}}, \quad \text{if} \quad d_{ij} \neq 0 \tag{25.6}$$

Else, $\mu_{ij} = 1$

Step 5: If the termination criteria have been reached, then stop. Else go back to step 2.

The suitable termination criteria can be set by checking whether the objective function is below a certain tolerance value or if its improvement compared to the previous iteration is below a certain threshold. Moreover, the maximum number of iteration cycles can be used as a termination criterion as well.

25.4 Protein Spot Detection Utilizing the Relational Fuzzy *C*-Means Algorithm

In this section, we present our new algorithm for matching protein spots in the 2D gel images. We called the new algorithm the RFCM as it builds on the traditional FCM algorithm but is modified by introducing the notion of fuzzy relations in order to differentiate spot pixels from the varying background.

The algorithm is composed of four steps, where the first step of it is to apply the FCM to the image to produce preliminary clusters. Then, these clusters are then internally refined to identify the inner spots by separating the background pixel from the contained pixel in the cluster by applying steps 2 4 on the clusters. A summary of the steps of the proposed algorithm is given below.

Step 1: Apply the FCM algorithm presented in Sect. 25.3 with C more than 2. The output is the partitioning of pixels in the image to different clusters each having a center value v.

Step 2: For each two pixels x, y belonging to two different clusters, create a fuzzy relation between x and y named $I(x, y)$, where $I(x, y)$ define the degree of closeness between intensities of pixels x and y.

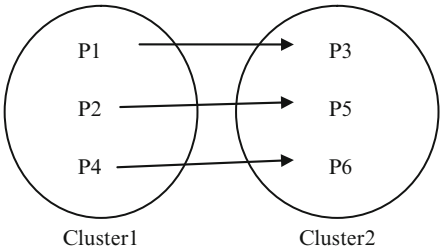


Fig. 25.2 Representation of the fuzzy relation between pixels in two clusters

Step 3: Compare pixels x, y
if pixel x is much more darker than pixel y
then pixel x is a spot pixel
else if pixel x is much more lighter than pixel y
then pixel y is a spot pixel
else if difference between intensities is low and one of the pixels is a spot pixel
then the other is a spot pixel also
end if

Step 4: Mark spot pixels and differentiate them from the background. This is done by assigning the spot pixels to the maximum value of centers of clusters and the background nonspot pixels to the minimum value of centers of clusters.

In Fig. 25.2, a representation of a fuzzy relation between two points in two different clusters: clusters 1 and 2, $I(x, y)$, is shown, where x and y are any two points each in one different cluster. Here, the arrow represents the degree of closeness between two pixels from different clusters.

25.5 Parameter Estimation

In step 3 of the proposed RFCM algorithm, we compare pixels x, y of different clusters. This comparison represents the membership function of the fuzzy relation established between both pixels. It can be represented either by the ratio or the difference between the level intensities of both of them.

In both cases, we need a parameter which identifies the lower bound of degree of darkness between both pixels. Whenever the difference between the level intensities of both of them exceeds a certain parameter we referred to this parameter by β then the pixel representing the darker pixel value is a spot pixel. We have shown to represent the fuzzy relation between the two pixels by the difference between the level intensities of both of them. Thus, we can rewrite step 3 in the algorithm as follows:

Step 3: Compare x, y
if absolute(gv(x) $-$ gv(y)) $> \beta$

then pixel representing
max(gv(x) − gv(y)) is a spot pixel
end if

Where gv refers to gray value of the corresponding pixel x or y.

Of course, the value of β which gives the best results in 2D gel electrophoresis images is a key issue in the algorithm and has to be estimated experimentally.

25.6 Experimental Results

The LECB 2D PAGE gel images database is available for public use. It contains data sets from four types of experiments with over 300 gif images with annotation and landmark data in html, tab-delimited, and xml formats. It could be used for samples of several types of biological materials and for test data for 2D gel analysis software development and comparison with other similar samples. PAGE is polyacrylamide gel electrophoresis. The LECB was the U.S. National Cancer Institute's Laboratory of Experimental and Computational Biology. The database is available at two Web sites [20, 21]. We choose the *Human leukemias* (Eric Lester, Peter Lemkin) data sets to show our results.

In our work, we used these data and applied our algorithm to 2D gel images of autologous human lymphoblastoid cell lines. The results are shown in the following figures. Figures 25.3, 25.7, and 25.11 show three test cases 2D gel electrophoresis images of a patient human leukemias. Figures 25.4, 25.8, and 25.12 show the gradient images of the 2D gel images presented in Figs. 25.3, 25.7, and 25.11. The results of applying the current FCM Segmentation algorithm on these images at $C = 2$, are shown in Figs. 25.5, 25.9, and 25.13. The results of applying the Proposed RFCM Segmentation algorithm on these images at $C = 6$ and $\beta = 20$ are shown in Figs. 25.6, 25.10, and 25.14.

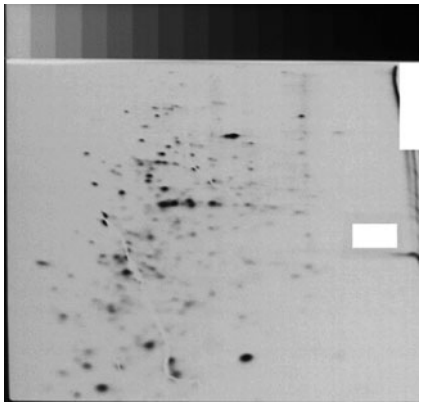


Fig. 25.3 2D Gel electrophoresis image of a patient human leukemias

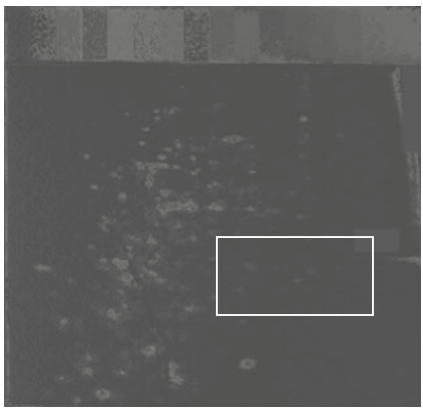


Fig. 25.4 The gradient image of 2D gel electrophoresis image in Fig. 25.3

Fig. 25.5 The gradient image of 2D gel electrophoresis image in Fig. 25.4 after applying the Fuzzy *C* Means segmentation algorithm *C* 2

Fig. 25.6 The gradient image of 2D gel electrophoresis image in Fig. 25.4 after applying the Proposed Relational Fuzzy *C* Means segmentation algorithm *C* 6, β 20

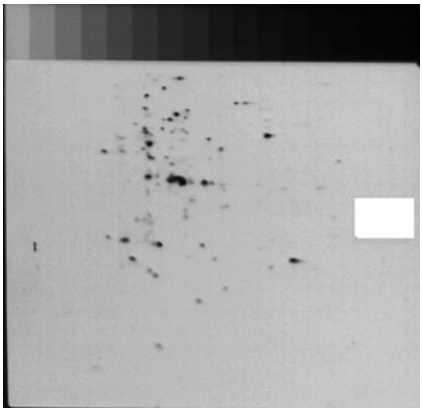


Fig. 25.7 2D Gel electrophoresis image of second patient human leukemias

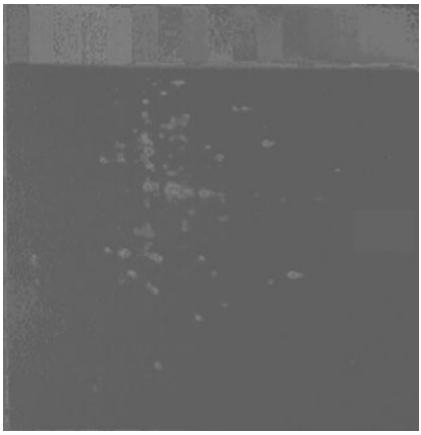


Fig. 25.8 The gradient image of 2D gel electrophoresis image in Fig. 25.7

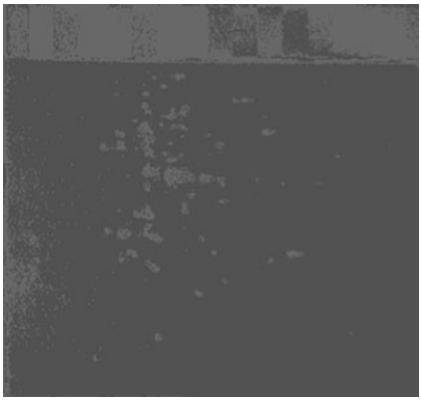


Fig. 25.9 The gradient image of 2D gel electrophoresis image in Fig. 25.8 after applying the Fuzzy *C* Means Segmentation algorithm *C* 2

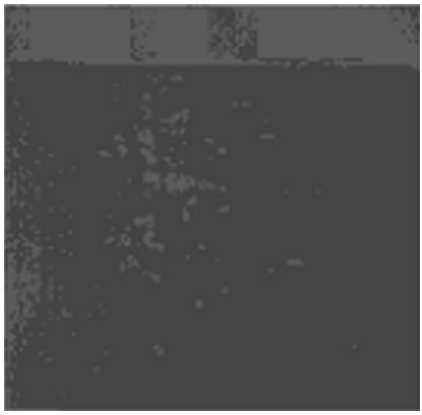


Fig. 25.10 The gradient image of 2D gel electrophoresis image in Fig. 25.8 after applying the Relational Fuzzy *C* Means Segmentation algorithm *C* 6, β 20

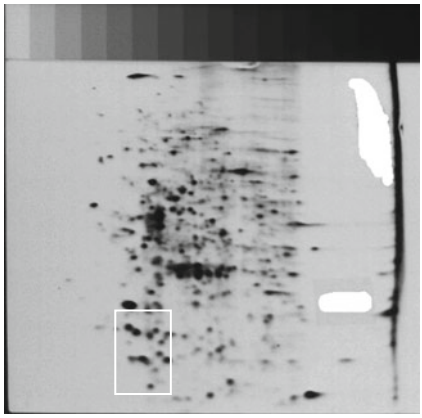


Fig. 25.11 2D Gel electrophoresis image of third patient human leukemias

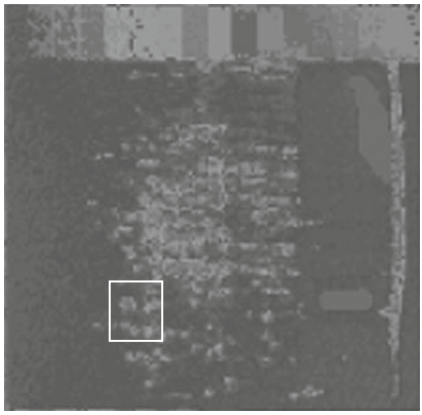


Fig. 25.12 The gradient image of 2D gel electrophoresis image in Fig. 25.11

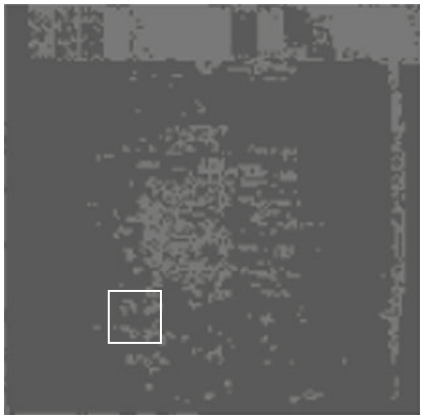


Fig. 25.13 The gradient image of 2D gel electrophoresis image in Fig. 25.12 after applying the Fuzzy C Means Segmentation algorithm C 2

Fig. 25.14 The gradient image of 2D gel electrophoresis image in Fig. 25.12 after applying the Relational Fuzzy C Means Segmentation algorithm C 6, β 20

25.7 Discussion

In this work, we presented a new algorithm based on the notion of fuzzy relations to segment and detect protein spots in 2D gel electrophoresis images. This technique shows high performance and detects the protein spots precisely, as shown in Figs. 25.6 and 25.14, even the less dark spots in the image appears (shown by squares) while in Figs. 25.5 and 25.13 when applying the FCM algorithm those protein spots which affects the spot quantization step in the whole process of 2D gel image analysis disappear totally.

When the image has clear nonoverlapping spots, both algorithms have similar results as shown in Figs. 25.7 25.10. However, for the 2D gel images with overlapping spots as for the third patient, the FCM algorithm divided the same spot that

can yield to ambiguity in the spot quantization while the new algorithm did not have the same problem, please refer to Figs. 25.11 25.13.

For future work, we suggest the development of fuzzy relations to obtain better results. The parameters representing the degree of closeness must be defined for the enhancement and improvement of the RFCM algorithm. The use of intuitionistic fuzzy relations to identify the degree of noncloseness between pixels and the hesitation margin can be investigated also.

References

1. N.G. Anderson, A. Matheson, N.L. Anderson, "Back to the future: the human protein index (HPI) and the agenda for post proteomic biology", Proteomics, 1:3, 2001.
2. V.C. Wasinger, S.J. Cordwell, A. Cerpa Poljak, J.X. Yan, A.A. Gooley, M.R. Wilkins, M.W. Duncan, R. Harris, K.L. Williams, I. Humphrey Smith, "Large scale amino acid analysis for proteome studies", Electrophoresis, 16, 1995.
3. W.P. Blackstock, M.P. Weir, "Proteomics: quantitative and physical mapping of cellular proteins", Trends in Biotechnology, 17:122 127, 1999.
4. P.H. O'Farrell, "High resolution 2 D electrophoresis of proteins", Journal of Biological Chemistry, 250:4007 4021, 1975.
5. J. Klose, "Protein mapping by combined isoelectric focusing and electrophoresis in mouse tissues: a novel approach to testing for induced point mutations in mammals", Human Genetik, 26:231 243, 1975
6. J. Klose, U. Kobalz, "2 D electrophoresis of proteins: an updated protocol and implications for a functional analysis of the genome", Electrophoresis, 16:1034 1059, 1995.
7. B. Voigt, T. Schweder, M.J. Sibbald, D. Albrecht, A. Ehrenreich, J. Bernhardt, J. Feesche, K.H. Maurer, G. Gottschalk, J.M. van Dijl, M. Hecker, "The extracellular proteome of *Bacillus licheniformis* grown in different media and under different nutrient starvation conditions", Proteomics, 6(1):268 281, 2006.
8. D. Höper, J. Bernhardt, M. Hecker, "Salt stress adaptation of Bacillus subtilis: a physiological proteomics approach", Proteomics, 6(5):1550 1562, 2006.
9. Y. Kim, J. Kim, Y. Won, Y. In, "Segmentation of protein spots in 2D gel electrophoresis images with watersheds using hierarchical threshold" In LNCS Computer and Information Sciences ISCIS 2003, 2869:389 396, 2003.
10. Y. Sugahara, Y. Hayashizaki, I. Tanihata, "An automatic image analysis system for RLGS LMS", Mammalian Genome, 9:643 651, 1998.
11. K. Takahashi, M. Nakazawa, Y. Watanabe, "DNAinsight: an image processing system for 2 D gel electrophoresis of genomic DNA", Genome Informatics, 8:135 146, 1997.
12. J.S. Morris, B.N. Clark, H.B. Gutstein, "Pinnacle: a fast, automatic and accurate method for detecting and quantifying protein spots in 2 dimensional gel electrophoresis data", Bioinformatics, 24(4):529 536, 2008.
13. U. Zeeshan Ijaz, S. Ullah Chaudhary, M. Sang Don, K. Youn Kim, "Computational Strategies for Protein Quantitation in 2D Electrophoresis Gel Image Processor for Matlab", In Proceedings of the 2007 Frontiers in the Convergence of Bioscience and Information Technologies, 129 134, 2007.
14. C.S. Hoeflich and J.J. Corso, "Segmentation of 2D Gel Electrophoresis Spots Using a Markov Random Field", In Proceedings of SPIE Conference on Medical Imaging, 2009.
15. D.K. Iakovidis, D. Maroulis, E. Zacharia, S. Kossida, "A Genetic Approach to Spot Detection in 2 D Gel Electrophoresis images", In Proceedings of International Conference on Information Technology in Biomedicine (ITAB), Ioannina, Greece, Oct 2006.

16. P. Tsakanikas, E.S. Manolakos, "Active Contours Based Segmentation of 2DGE Proteomics Images", In Proceedings of Eusipco 2008, 2008.
17. J.W. Yoon, S.J. Godsill, C. Kang, T S. Kim "Bayesian inference for 2D gel electrophoresis image analysis", In Bioinformatics Research and Development, 343 356, 2007.
18. J.C. Dunn, "A fuzzy relative of the ISODATA process and its use in detecting compact, well separated clusters", Journal of Cybernetics, 3(3):32 57, 1974.
19. M. Berth, F. Michael Moser, M. Kolbe, J. Bernhardt, "The state of the art in the analysis of 2 D gel electrophoresis images", Applied Microbiology Biotechnology, 76:1223 1243, 2007.
20. http://www.ccrnb.ncifcrf.gov/2DgelDataSets.
21. http://www.bioinformatics.org/lecb2dgeldb.

Chapter 26
Assigning Probabilities to Mascot Peptide Identification Using Logistic Regression

Jinhong Shi and Fang-Xiang Wu

Abstract We propose a method to assign probabilities to Mascot peptide identification by using logistic regression. Three key scores, Mascot ions score (MIS), identity threshold, and homology threshold, are integrated into the logistic regression model. Two features in the model are constructed as the differences between MIS and the two thresholds, respectively. Newton Raphson method is then adopted to solve the model and the weight vector is estimated by maximizing the likelihood of training data. By applying the method to two datasets with known validity, the results demonstrate that the proposed method can assign accurate probabilities to Mascot peptide identifications and have a high discrimination power to separate correct and incorrect peptide identifications.

Keywords Logistic regression · Mascot ions score (MIS) · Mascot peptide identification · Newton-raphson method

26.1 Introduction

Mascot [1] is one of the most popular peptide identification search engines. Generally, it provides three scores, Mascot ions score (MIS), identity threshold, and homology threshold, to distinguish peptide identifications. MIS is given by $10\log_{10}(P)$, where P is the probability that a peptide spectrum match is a random event. MIS measures the similarity between experimental spectra and theoretical spectra, but the algorithm to compute the probability P is unknown to users. Mascot Identity Threshold (MIT) is defined as $10\log_{10}(20p/qmatch)$, where $qmatch$ is the number of candidate peptides with masses close to the precursor ion mass in a

J. Shi (✉)
Division of Biomedical Engineering, University of Saskatchewan, Saskatoon, SK, Canada S7N 5A9
e mail: jinhong.shi@usask.ca

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 26,

database, and p is the significance threshold. It is reported that MIT is conservative in discriminating peptide identifications. In some cases, all peptides for one spectrum have scores lower than MIT, but there exist peptides with scores which are significant outliers from the distribution of all scores. To cope with this situation, Mascot provides a lower empirical threshold called Mascot homology threshold (MHT). The algorithm to compute MHT is not released either.

Mascot uses a threshold method to discriminate correct and incorrect peptide identifications. This method is useful when the goal stays on the level of peptide identification since it can identify many correct peptides. However, the final goal of proteomics is not peptide identification but protein inference. To facilitate the subsequent protein inference, it is necessary to perform statistical analysis on peptide identifications given by search engines. For one peptide identification, it is desired to associate it with a probability which can represent the likelihood that it is correct. With the associated probabilities to peptide identifications, it would be convenient to use statistical models to address protein inference problem [2 4].

There have been methods proposed to assign probabilities to peptide identifications given by search engines. Keller et al. [4] used bimodal and EM algorithm to assign probabilities to Sequest [5] search results. This algorithm is then adopted by the commercial software Scaffold [6] and has been extended to the analysis of search results from X!Tandem and Mascot. The advantage of this algorithm is that it models the distribution of all discriminant scores, and uses Bayesian model to assign probabilities. However, it did not provide any theoretical explanations why the distributions of scores are subject to the assumed standard statistical distributions. In addition, Feng et al. [7] used a generalized Mascot scoring function to improve the accuracy of peptide identification probabilities. The problem with this function is that it cannot guarantee the results always to be probabilities.

In this paper, the logistic regression is used to assign probabilities to peptide identifications based on Mascot scores. MIS, MIT, and MHT are used to construct two features in the model, which is then solved by Newton Raphson method. Two datasets, ISB and TOV, with known validity are used to evaluate the performance of the method. The results show that the probabilities assigned are accurate and have a high power to discriminate the correct and incorrect peptide identifications.

26.2 Methods

Logistic regression is a good choice of classification technique for approaching problems in bioinformatics. One obvious advantage is that it can integrate all possible risk factors into one model and assure that the results are probabilities, which is exactly what we expect when performing the statistical analysis of proteomic data. Moreover, this characteristic can help to avoid the problem in Feng's model [7].

We use logistic regression to assign probabilities to Mascot peptide identifications and classify the results. A two class classification is considered here, that is, the correct and the incorrect peptide identifications. If we take the identification of one peptide as a Bernoulli trial, denoted by the random variable X, then we have

$$X = \begin{cases} 1, & \text{correctidentification,} \\ 0, & \text{incorrect identification.} \end{cases} \tag{26.1}$$

Under general assumptions [8], the posterior probability of X can be written as a logistic function acting on a linear combination of a feature vector $\mathbf{\Phi}$ so that

$$p(X = 1|\Phi, \mathbf{w}) = \frac{\exp(\mathbf{w}^{\mathrm{T}}\Phi)}{1 + \exp(\mathbf{w}^{\mathrm{T}}\Phi)} \tag{26.2}$$

$$p(X = 0|\Phi, \mathbf{w}) = \frac{1}{1 + \exp(\mathbf{w}^{\mathrm{T}}\Phi)}, \tag{26.3}$$

where $\mathbf{w}$ is the weight vector. Equations (26.2) and (26.3) represent the posterior probabilities of a correct and incorrect peptide identification given its feature vector $\mathbf{\Phi}$, respectively. Notice that (26.3) follows directly from (26.2) because the sum of the two probabilities must be 1. Two features ϕ_1 and ϕ_2 are incorporated into the logistic regression model, i.e.,

$$\phi_1 = \mathrm{MIS} - \mathrm{MIT} \quad \text{and} \quad \phi_2 = \mathrm{MIS} - \mathrm{MHT} \tag{26.4}$$

with the feature vector $\Phi = [1, \phi_1, \phi_2]^{\mathrm{T}}$ and the weight vector $\mathbf{w} = [w_0, w_1, w_2]^T$. Instead of using three scores separately, we choose the differences between scores and thresholds as features because it is easy to understand how Mascot distinguishes peptide identifications. The larger the difference, the higher the probability of a peptide identification being correct. Additionally, we expect that the weights of ϕ_1 and ϕ_2 are positive to be consistent with Mascot peptide identification criteria.

At this point, assigning probabilities to peptide identifications is reduced to estimate the model parameters. If feature values are subject to Gaussian distributions with the same covariance matrix, then naive Bayes classifiers or the Fisher's LDA can be used to estimate the weight vector $\mathbf{w}$. However, this assumption usually cannot be satisfied by experimental data. Hence, we estimate the weights directly from the training data by maximizing the conditional data likelihood, i.e.,

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \left(\prod_{l=1}^{L} p(X^l|\Phi^l, \mathbf{w}) \right), \tag{26.5}$$

where L is the number of the training data, and X^l and Φ^l are the random variable and the feature vector of the l^{th} peptide identification, respectively. Equation (26.5) is equivalent to maximizing the conditional log likelihood as follows

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \mathcal{L}(\mathbf{w}), \tag{26.6}$$

where the conditional log likelihood $\mathcal{L}(\mathbf{w})$ is given by

$$\mathcal{L}(\mathbf{w}) = \sum_{l=1}^{L} \left[X^l \left(w_0 + \sum_{i=1}^{2} w_i \phi_i^l \right) - \ln\left(1 + \exp\left(w_0 + \sum_{i=1}^{2} w_i \phi_i^l \right) \right) \right], \tag{26.7}$$

where $\phi_i{}^l$ denotes the value of the i^{th} feature of the l^{th} peptide identification. There is no analytic solution to (26.6), and so we choose the Newton Raphson method to find a numerical solution. By taking the first and second derivatives of $\mathcal{L}(\mathbf{w})$ w.r.t $\mathbf{w}$ to obtain the gradient vector and the Hessian matrix, respectively, the iterative equation to optimize the weight vector $\mathbf{w}$, starting at some point $\hat{\mathbf{w}}_0$, is given by

$$\hat{\mathbf{w}}_{n+1} = \hat{\mathbf{w}}_n - \mu \left[\nabla^2 \mathcal{L}(\hat{\mathbf{w}}_n) \right]^{-1} \nabla \mathcal{L}(\hat{\mathbf{w}}_n), \quad n \geq 0, \tag{26.8}$$

where μ is a constant controlling the convergence speed. Here, we assume that $\det(\nabla^2 \mathcal{L}(\mathbf{w}^*)) \neq 0$ for all critical points $\mathbf{w}^*$. Since the conditional log likelihood function $\mathcal{L}(\mathbf{w})$ is concave w.r.t. $\mathbf{w}$, the result will converge to a global maximum.

26.3 Results

26.3.1 Data and ROC Curve

In this study, we use receiver operating characteristic (ROC) curves to evaluate the performance of the proposed method based on two reference datasets.

ISB dataset consists of 37,044 tandem mass spectra from 18 control proteins [9]. The spectra are separated into 124 files with 300 spectra or less in each file to meet the requirement of the online Mascot MS/MS Ions Search (Version 2.2, http://www.matrixscience.com). Then, they are searched with most Mascot parameters being default except *Database* using *NCBInr*, *Instrument* using *ESITRAP*, and *Report top* 50 hits. When saving the results, we select the format *csv* and check MIS, MIT, and MHT to be saved. Note that we use MIT as MHT when MHTs are not reported. To label the data, we search the reported peptides in the 18 control proteins plus 15 contaminants [10]. If the peptide is found in the 33 proteins, then it is labeled as "correct", otherwise "incorrect".

TOV dataset, a part of the whole TOV dataset in [11], includes 24,000 tandem mass spectra and is also searched by Mascot with the same parameters for ISB dataset. It is searched with Sequest as well, and the outputs are validated by PeptideProphet [4] with 95% as the threshold. Then, Mascot search results are

labeled by referring to these correctly identified peptides. In all, 4,992 reports at rank 1 are obtained, with 1,580 true and 3,412 false reports, respectively.

ROC curves are used to measure the discrimination power of our method. An ROC is a plot of "sensitivity (TPR)" vs "(1-specificity) (FPR)" for a binary classifier as the threshold is varied. If an ROC curve is a diagonal line and the area under the curve (AUC) is 0.5, the results are just random guesses. The ROC curve for a perfect classification is the point (0, 1). Actual classifications correspond to curves locating between the ideal and random plots with AUC $\in$ (0.5, 1). The greater the AUC, the higher the average discrimination power of classifiers.

26.3.2 Weight Vector Estimation from ISB Training Data

We first feed this model a training set, *ISBTrainingDataI,* with a good separation of correct and incorrect peptide identifications, to primarily test our method. The ROC curve is not shown due to the space limit, and AUC = 0.8845, which shows that the logistic regression works very well on the dataset. The estimated weights are $w_0 = 1.7785$, $w_1 = 0.1327$, and $w_2 = 0.0455$. However, the performance is not as good as expected when the weights are applied to ISB test data. One possible reason is the happening of overfitting. As discussed before, both w_1 and w_2 are supposed to be >0 since MIS above both thresholds means peptides have higher confidences to be correctly identified than cases when MIS exceeds one or none of them. However, the estimated $w_2 < 0$, which is probably the result of overfitting.

To avoid overfitting the model, we use another dataset, called *ISBTrainingDataII,* comprising of 567 true and 1,909 false reports, to estimate the weight vector. The weights estimated are $w_0 = 2.4667$, $w_1 = 0.1447$, and $w_2 = 0.0180$, with $w_1 > 0$ and $w_2 > 0$ as expected. The resultant ROC curve is shown in Fig. 26.1 with AUC = 0.8613. Although the average classification power for this dataset is not as strong as *ISBTraningDataI*, the performance is still very promising.

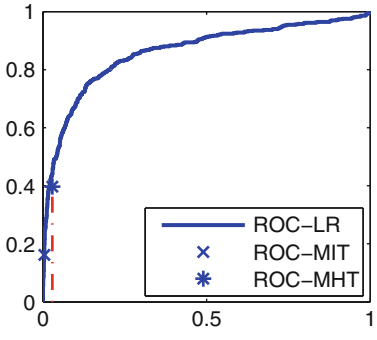


Fig. 26.1 ROC curve of *ISBTraningDataII*

26.3.3 Evaluation on Testing Datasets

We use positive predictive value (PPV) as the effectiveness measure to choose a threshold on the estimated probabilities by logistic regression so as to maximize the PPV on the training dataset. A plot of PPV vs. threshold is given in Fig. 26.2. It can be seen that the threshold 0.55 can achieve a similar PPV as MHT and the threshold 0.85 can achieve a similar PPV as MIT. We choose 0.55 as the threshold to keep a

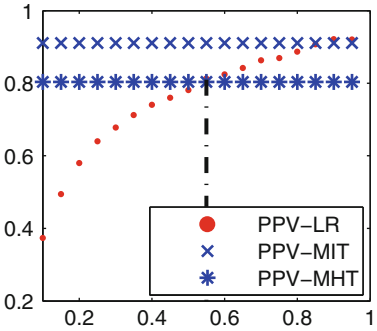


Fig. 26.2 Plot of PPV vs. threshold of ISB training data

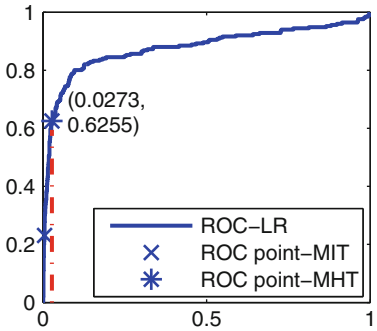


Fig. 26.3 ROC curve of ISB test data

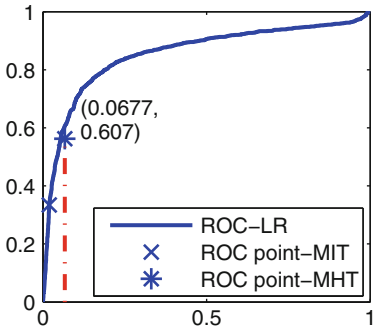


Fig. 26.4 ROC curve of TOV test data

Table 26.1 Comparisons of the number of peptides identified as TP, FP, TN, and FN between the proposed logistic regression (LR) model and Mascot threshold methods (PPV is also compared)

	ISB			TOV		
	LR	MIT	MHT	LR	MIT	MHT
True Positive (TP)	145	58	157	1,036	528	889
False Positive (FP)	42	9	51	286	67	228
True Negative (TN)	1,824	1,857	1,815	3,126	3,345	3,184
False Negative (FN)	106	193	94	544	1,052	691
PPV	0.7754	0.8657	0.7548	0.7837	0.8874	0.7959

similar PPV as MHT because MIT is too conservative, and apply the threshold 0.55 on all the testing datasets in the following analysis.

The estimated weights are applied on an ISB and a TOV test dataset to measure the performance of the proposed method. ROC curves are shown in Fig. 26.3 with AUC = 0.8768 and Fig. 26.4 with AUC = 0.8547, respectively. Both ROC and AUC indicate that the proposed method has a high discrimination power for the two test datasets. The point * and the point × are (FPR, TPR) in the ROC space when MHT and MIT are used as thresholds, respectively. The datatips are the points on the ROC curves of the proposed method which have the same FPR as MHT. It can be seen that the proposed method achieves a same sensitivity (0.6255) for ISB test data and a higher sensitivity (0.607) for TOV test data than MHT. Although the sensitivity is not increased for ISB test data, the probabilities assigned by the proposed method are more meaningful for users to judge their peptide identifications and can facilitate subsequent protein inference. Moreover, the results for TOV test data show that the proposed method has good cross experiment usability.

We also compare the number of identified peptides in the test datasets between the proposed method and Mascot threshold methods, which is shown in Table. 26.1. It can be seen that the proposed method (LR) does not report as many correct peptide identifications as MHT does for ISB test data, but it achieves a higher PPV than MHT. Sometimes, a higher PPV is more useful in reporting peptide identification results, especially for the direct use of the results without further reconfirmation. For the TOV dataset, the proposed method (LR) identifies more TPs than MHT while also introduces more FPs at the same time. Although there is always a trade-off between TP and FP, hopefully this conflict can be mitigated by combining more information from the deeper understanding of mass spectrometry experiments in the future.

26.4 Conclusions

A method has been proposed to assign probabilities to Mascot peptide identifications by the use of logistic regression, and it can be easily extended to other search engines. The results of the two reference datasets are promising, which show that the method can generate accurate probabilities to Mascot peptide identifications

and can generally increase the sensitivity at the same specificity as Mascot threshold methods. In other words, the proposed method can provide effective statistical analysis for Mascot peptide identifications. For the future work, we are going to find more information to alleviate the conflict between TP and FP, and integrate this statistical analysis for Mascot peptide identification into protein inference.

Acknowledgments This study is supported by Natural Sciences and Engineering Research Council of Canada (NSERC). We would like to thank Dr. Keller for generously providing the ISB dataset and Dr. Poirier for providing the TOV dataset used in this paper.

References

1. Perkins D N, Pappin D J C, Creasy D M et al (1999) Probability based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis, 20: 3551 3567.
2. Nesvizhskii A I and Aebersold R (2005) Interpretation of shotgun proteomic data: The protein inference problem. Mol Cell Proteomics, 4(10): 1419 1440.
3. Nesvizhskii A I, Keller A, Kolker E et al (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem, 75: 4646 4658.
4. Keller A, Nesvizhskii A I, Kolker E et al (2002) Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem, 74: 5383 5392.
5. Eng J K, McCormack A L and Yates J R III (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom, 5: 976 989.
6. Searle B C, Turner M and Nesvizhskii A I (2008) Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies. J Proteome Res, 7: 245 253.
7. Feng J, Naiman D Q and Cooper B (2007) Probability model for assessing proteins assembled from peptide sequences inferred from tandem mass spectrometry data. Anal Chem, 79: 3901 3911.
8. Bishop C M (2006) Pattern Recognition and Machine Learning. Springer, Singapore.
9. Keller A, Purvine S, Nesvizhskii A I et al (2002) Experimental protein mixture for validating tandem mass spectral analysis. OMICS, 6(2): 207 212.
10. Klimek J, Eddes J S, Hohmann L et al (2008) The standard protein mix database: A diverse data set to assist in the production of improved peptide and protein identification software tools. J Proteome Res, 7: 96 103.
11. Gagne J P, Gagne P, Hunter J M et al (2005) Proteome profiling of human epithelial ovarian cancer cell line TOV 112D. Mol Cell Biochem, 275: 25 55.

Chapter 27
Genome-Wide EST Data Mining Approaches to Resolving Incongruence of Molecular Phylogenies

Yunfeng Shan and Robin Gras

Abstract Thirty-six single genes of 6 plants inferred 18 unique trees using maximum parsimony. Such incongruence is an important challenge. How to reconstruct the congruent tree is still one of the most challenges in molecular phylogenetics. For resolving this problem, a genome-wide EST data mining approach was systematically investigated by retrieving a large size of EST data of 144 shared genes of 6 green plants from GenBank. The results show that the concatenated alignments approach overcame incongruence among single-gene phylogenies and successfully reconstructed the congruent tree of 6 species with 100% jackknife support across each branch when 144 genes was used. Jackknife supports of correct branches increased with number of genes linearly, but the number of wrong branches also increased linearly. For inferring the congruent tree, a minimum of 30 genes were required. This approach may provide potential power in resolving conflictions of phylogenies.

Keywords Data mining · Genome-wide · Phylogeny

27.1 Introduction

It is well known that different single genes often reconstructed different phylogenetic trees. Recent research shows that this problem still exists [1 3]. Such incongruence is caused by many reasons such as insufficient number of informative sites,

Y. Shan (✉)
School of Computer Science, University of Windsor, 401 Sunset Avenue, Windsor, ON, Canada N9B 3P4
e mail: yunfeng@uwindsor.ca

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 27,

lateral gene transfer, unrecognized paralogy, and variable evolutionary rates of different genes [3 5].

Multiple-gene approaches have been studied [6, 7]. Kluge believed that phylogenetic analysis should always be performed using all the evidences [8], but Miyamoto and Fitch argued that partitions (including genes) should not be combined [9]. Concatenated alignments of a couple of genes improved supports [3 10]. Concatenating alignments into one from genome-scale 106 genes based on 7 yeast genomes [11] for phylogenetic analysis proclaimed ending incongruence [12]. However, there is doubt whether it is a "true tree" [13]. Generally, determining the phylogeny of microbes was difficult due to the lack of discernible morphological characters of microbes [14]. Minimum evolution (ME) showed the different trees from maximum likelihood (ML) and maximum parsimony (MP) for the same dataset when base biases were not adjusted [15]. Therefore, this critical problem must be clarified further for this approach. Actually, there are organisms for which the phylogeny is firmly established by fossil records and morphological characteristics [3]. If this set of organisms is used, it becomes possible to determine how reliable an approach is for overcoming the incongruence. In this study, such well-known green plants were selected [16 19]. The six species included two gymnosperms: *Picea glauca* and *Pinus taeda,* two monocots: *Oryza sativa* and *Triticum aestivum,* and two eudicots: *Populus tremula* and *Arabidopsis thaliana*. *Ginkgo biloba* was specified as the outgroup. Five commonly used methods: ML, ME, neighbour-joining with unweighted least squares (NJUW), neighbour-joining with absolute difference (NJAD), and MP were used to compare consistence.

There are some evident limitations for the genome-scale approach [11]. Especially, all of concatenated alignments must include the same set of taxa. This requirement limits many species representations because only dozens of species have completed genome sequence data up to date.

Data mining is a powerful tool for retrieving data and is widely used in a variety of areas. An alternate solution is to retrieve enough EST sequence data of shared genes from publicly available sequence databases such as GenBank using data mining approach since GenBank includes many more species that have EST sequence data than the number of species that have completed sequencing genomes. The GenBank size is huge and also growing fast. GenBank and its collaborating databases, EMBL and DDBJ, have reached a milestone of 100 billion bases from over 165,000 organisms by August 22, 2005.

27.2 Results and Discussion

Single-gene trees showed wide incongruence. Thirty-six genes reconstructed 21 unique rooted trees using ME, 20 using NJUW, and 18 using MP.

Figures 27.1 and 27.2 show the rooted tree and the unrooted tree of the concatenated alignments of 36 genes, and Fig. 27.3 shows the unrooted tree of the concatenated alignments of 144 genes inferred by ML, ME, NJUW, NJAD, and MP

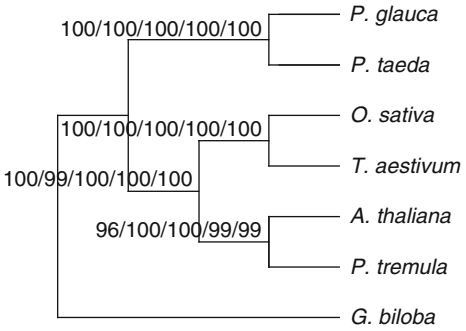


Fig. 27.1 The rooted tree of the concatenated alignments of 36 genes. *Ginkgo biloba* was specified as the outgroup. Numbers above branches indicate supports (L/ME/NJUW/NJAD/MP, bootstrap for NJAD and jackknife for others)

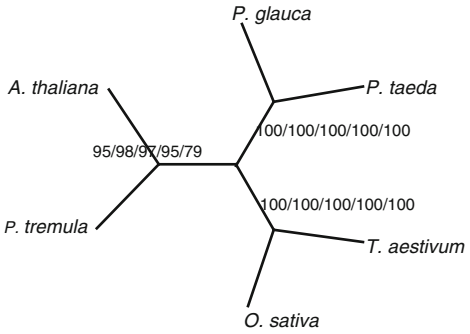


Fig. 27.2 The unrooted tree inferred from analysis of the concatenated alignments of 36 genes. See legends in Fig. 27.1

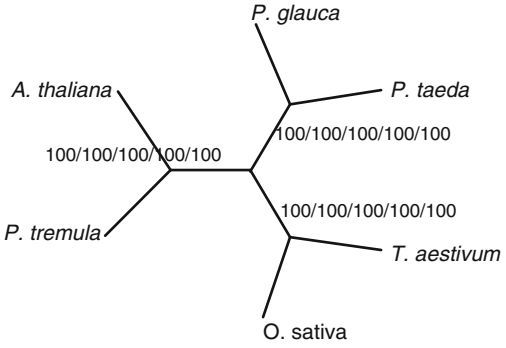


Fig. 27.3 The unrooted tree inferred from analysis of the concatenated alignments of 144 genes. See legends in Fig. 27.1

methods. The topologies of these rooted or unrooted trees all show a single tree and are consistent with those based on traditional fossil, morphological and anatomical [3], and phylogenetic analysis [16 19]. Therefore, these phylogenetic trees (Figs. 27.1 27.3) are considered as congruent trees. The concatenation of 106 genes recovered a single species tree of 7 yeasts with 100% bootstrap supports of all branches [11]. However, ME tree [15] was different from those of MP and ML

when base biases were not adjusted [11, 15]. Rokas et al. also observed topological differences between the tree obtained from a study on 75 yeast species but only 8 commonly sequenced genes and their tree from 106 genes of 7 species [11, 20].

Although Soltis et al. doubted whether the tree is a "true tree", they advocated the use of multiple-gene or genome-scale approaches [13]. The results of this study strongly show that the genome-wide data mining approach is effective to overcome incongruence regardless of the phylogenetic methods.

Jackknife support values for all branches of the rooted tree based on the concatenated alignments of 36 single genes were 100% except for the branch of *P. tremula A. thaliana* . The support for the branch of *P. tremula A. thaliana* was 99% for NJAD or MP, or 96% for ML (Fig. 27.1). Similarly, support values for all branches of the unrooted tree based on 36 single genes were 100% except that for the branch of *P. tremula A. thaliana* was 95%, 98%, 97%, 95%, or 79% for ML, ME, NJUW, NJAD, or MP, respectively (Fig. 27.2). Supports for all branches of the unrooted tree based on 144 genes were 100% (Fig. 27.3). These results suggested that more genes get greater support values. 100% support for the congruent tree from the concatenated alignments of 144 single genes regardless of the phylogenetic methods, even the simplest method such as NJAD, demonstrated the power of large datasets by means of the genome-wide data mining approach in resolving the incongruence. Multiple-gene approaches such as 3 combined genes of angiosperms [13], 4 combined proteins of eukaryotes [10], 6 and 9 genes of flowering plants [21], 23 proteins of bacteria, archea, and eucarya [22], 63 genes of baculovirus [7], and 106 concatenated genes of yeasts [11] strengthened supports and improved the consistence of phylogenies. Those previous results were consistent with the results of the study.

100% of ten random replicates of 30-gene concatenated alignments got the congruent tree with an average jackknife support value of at least 85% across all branches. These results show that the number of genes was about 30. Rokas et al. [11]

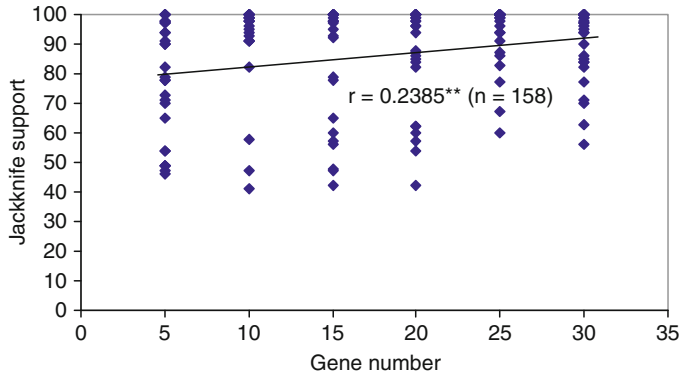


Fig. 27.4 The relationship between jackknife support values of the correct branch (branch of *P. tremula A. thaliana*) of phylogenetic trees and the number of genes. ** indicates the statistical significant level at 0.01

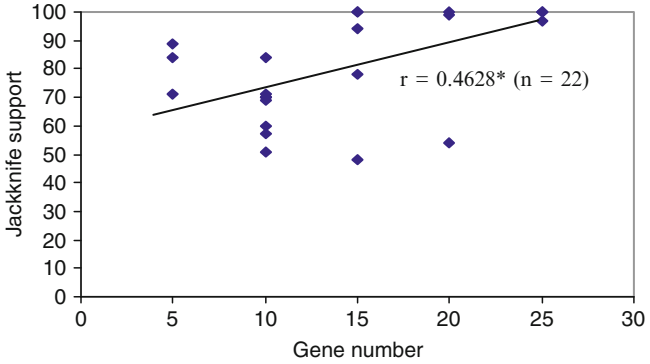


Fig. 27.5 The relationship of jackknife supports of wrong branches (alternative branches of branch of *P. tremula A. thaliana*) and the number of genes. * indicates the statistical significant level at 0.05

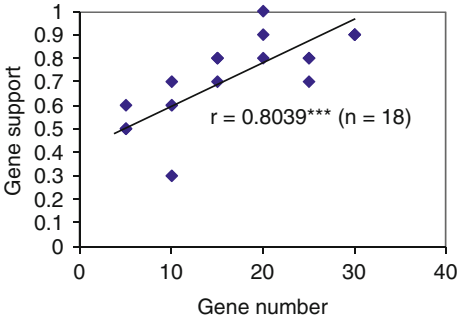


Fig. 27.6 The relationship between gene support percentages of phylogenetic trees and the numbers of genes. *** indicates the statistical significant level at 0.001

proposed the number of genes sufficient to support all branches of the species tree ranged from a minimum of 8 20 based on the 106 concatenated genes of yeasts [11].

The jackknife support values of the branch of *P. tremula A. thaliana* (correct branch) increased with the numbers of genes linearly ($r = 0.2385$**, $n = 158$) (Fig. 27.4). On the other hand, the supports of wrong branches did also increase with the number of genes linearly (Fig. 27.5) ($r = 0.4628$*, $n = 22$). 100% support did appear in wrong branches. This means that the concatenation of multiple genes increases the support regardless of the correctness of branches. Systematic errors might accumulate with concatenating multiple genes [15].

The bootstrap or jackknife method is used as a representation of confidence in phylogenetic topologies. However, 100% support does not mean that the branch is 100% correct. 100% support may occur in an alternative branch [15]. High bootstrap support does not necessarily signify "the truth" [13].

Gene-support percentage is the percentage calculated by the number of correct gene trees divided by the total number of gene trees. The Gene-support percentages of phylogenetic trees significantly increased with the number of genes linearly

when the number of genes increased from 5 to 30 ($r = 0.8039$***, $n = 18$) (Fig. 27.6). The Gene-support percentages to reconstruct the "true tree" were only 30 60% when the number of genes was 5, 10, or 15. When 30 genes were used, 100% Gene-support percentage was observed. High Gene-support percentage is the appropriate criterion for evidence of getting the "true tree".

When phylogenetic trees were inferred using NJ including slices with gaps [18], they were identical with the rooted tree of 36 genes (Fig. 27.1) and the unrooted tree of 144 genes (Fig. 27.3). However, the branch of *P. tremula A. thaliana* was not consistent between the rooted tree and the unrooted of 36 genes. These results show that gaps in alignments resulted in incongruence of phylogenies. They also suggest that concatenated alignments from more genes resulted in more consistent trees and the sufficient number of genes can overcome incongruence caused by gaps.

ML usually performs well, but is a very computationally intensive method. For general case, it is thought that the problem of ML phylogeny is NP-hard [23]. A parallel version for ML could help to solve some problems of the genome-scale approach.

27.3 Methods

Nucleotide sequences were retrieved from the public EST database of GenBank. Homologous genes were identified by BLASTN v2.2.6 with the highest available BLASTN score hit (e-value <0.0009). A program SeqMiner.pl was used (Available upon request from authors). 144 shared genes of the six species, except for *G. biloba*, and 36 shared genes of all 6 species were retrieved.

Single genes were separately aligned using Clustalx with default settings [18]. All gene alignments were edited to simply exclude positions with gaps for further analysis, except others specially stated, and then concatenated into one large alignment for further phylogenetic analysis.

PAUP*4.0b10 [19] was used for tree inference. Each nucleotide dataset was analyzed under the optimality criteria of ML, distance, which included NJUW and ME, and maximum parsimony. The ML analyses were conducted assuming that the *ti/tv* ratio was unequal and estimated for a nucleotide substitution model. The NJUW and ME analyses were performed assuming the HKY85 model of nucleotide substitution. The MP analyses were performed with unweighted parsimony. The jackknife consensus tree was searched using the branch-and-bound algorithm for MP, and the full heuristic search for ML, NJUW, and ME. Support for each branch was tested with the bootstrap/jackknife analysis. 100 replicates were used for ML and 1,000 for NJUW, ME, and MP. NJAD trees were calculated using Clustalx by default settings [18], whose bootstrap replicates were 1,000. The correlation analyses were performed using the SAS system for Windows V8. Trees and sequence datasets are available from the authors upon request.

Acknowledgments Part of this research was conducted by YS at Atlantic Bioinformatics Centre. This work was partially supported by the NSERC grant ORGPIN 341854, the CRC grant 950 2 3617, and the CFI grant 203617.

References

1. Goodman M, Romero Herrera AE, Dene H, Czelusniak J, Tashian RE (1982) Amino acid sequence evidence on the phylogeny of primates, other eutherians. In: Goodman M (Ed) Macromolecular Sequences in Systematic and Evolutionary Biology. Plenum, New York, pp. 115 191.
2. Hedges SB (1994) Molecular evidence for the origin of birds. Proc Natl Acad Sci USA 91:2621 2624.
3. Russo CAM, Takezaki N, Nei M (1996) Efficiencies of different genes and different tree building methods in recovering a known vertebrate phytogeny. Mol Biol Evol 13:525 536.
4. Doolittle WF (1999) Phylogenetic classification and the universal tree. Science 284: 2124 2128.
5. Wolf YI, Rogozin IB, Grishin NV, Koonin EV (2002) Genome trees, the tree of life. Trends Genet 18:472 479.
6. Yang Z (1996) Maximum likelihood models for combined analysis of multiple sequence data. J Mol Evol 42:587 596.
7. Huelsenbeck JP, Bull JJ, Cunningham CW (1996) Combining data in phylogenetic analysis. TREE 11:152 158.
8. Kluge AG (1989) A concern for evidence and a phylogenetic hypothesis of relationships among Epicrates (Boidae, Serpentes). Syst Zool 38:7 25.
9. Miyamoto MM, Fitch WM (1995) Testing species phylogenies and phylogenetic methods with congruence. Syst Biol 44:64 76.
10. Baldauf SL, Roger AJ, Wenk Siefert I, Doolittle WF (2000) A kingdom level phylogeny of eukaryotes based on combined protein data. Science 290:972 977
11. Rokas A, Williams BL, King N, Carroll SB (2003) Genome scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798 804.
12. Gee H (2003) Evolution, ending incongruence. Nature 425:782.
13. Soltis DE et al (2004) Genome scale data, angiosperm relationships, and ending incongru ence: A cautionary tale in phylogenetics. Trends Plant Sci 9:477 483.
14. Fitz Gibbon ST, House CH (1999) Whole genome based phylogenetic analysis of free living microorganisms. Nucl Acids Res 27:4218 4222.
15. Phillips MJ, Delsuc F, Penny D (2004) Genome scale phylogeny and the detection of systematic biases. Mol Biol Evol 21:1455 1458.
16. Panchen AL (1992) Classification, Evolution and the Nature of Biology. Cambridge Univer sity Press, Cambridge.
17. Cronquist A (1981) An Integrated System of Classification of Flowering Plants. Columbia University Press, New York.
18. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The ClustalX windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucl Acids Res 25:4876 4882.
19. Swofford D (2002) PAUP*: Phylogenetic Analysis Using Parsimony (*and Other Methods) (Version 4). Sinauer Associates, Sunderland, MA.
20. Kurzman CP, Robnett CJ (2003) Phylogenetic relationships among yeasts of the 'Saccharo myces complex' determined from multigene sequence analysis. FEM Yeast Res 3:417 432.
21. Barkman TJ et al (2000) Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. Proc Natl Acad Sci USA 97:13166 13177.
22. Brown JR, Douady JD, Italia MJ, Marchall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. Nat Genet 28:281 285.
23. Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368 376.

Chapter 28
Building a Parallel Between Structural and Topological Properties

Omar Gaci

Abstract In this chapter, we study the amino acid interaction networks. An amino acid interaction network is a graph whose vertices are the protein's amino acids and whose edges are the interactions between them. Using a graph theory approach, we identify a number of properties of these networks. Some of them are common to all proteins, while others depend on the structure arrangement. We rely on the latter group of properties to illustrate the correlation between structural and topological properties. Then, we propose a topological space where proteins from a same family tend to be grouped.

Keywords Protein structure · Interaction network · Topological space

28.1 Introduction

In their natural environment, proteins adopt a native compact three-dimensional form. The process is called folding and is not fully understood. The process is a result of interactions between the protein's amino acids which form chemical bonds.

In this study, we treat proteins as networks of interacting amino acid pairs [2]. In particular, we consider the subgraph induced by the set of amino acids participating in the secondary structure also called Secondary Structure Elements (SSE). We call this graph SSE interaction network (SSE-IN). We carry out a study to describe the structural families of proteins when they are represented as interaction networks. We show how the properties of these networks are related to the structure of the corresponding protein. Thus, we propose a topological space, where proteins from the same family tend to be grouped. By this way, we draw a parallel between structural and topological properties.

O. Gaci
LITIS Laboratory, 25 rue Philippe Lebon, Le Havre, France
e mail: omar.gaci@gmail.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 28,

28.2 A Topological Study

The purpose of our work is to offer a graph theory interpretation of the hierarchical protein classifications. Indeed, when a protein belongs to a hierarchical level according to its structural properties, then one can say that the corresponding protein SSE-IN also belongs to the same level. Thus, the topological properties of a SSE-IN are a consequence of the protein structural family. It implies that a SSE-IN is described by specific topological properties relative to the protein structural classification.

The first step before studying the protein SSE-IN is to select them according to their SSE arrangements. We have computed topological measures for three families of each hierarchical classification, namely, CATH and SCOP (see Table 28.1).

We have chosen these three families by classification, in particular, because of their huge protein number. Thus, each family provides a broad sample guarantying more general results and avoiding fluctuations. Moreover, these six families contain proteins of very different sizes, varying from several dozens to several thousand amino acids in SSE.

28.2.1 Diameter and Mean Distance

Table 28.2 (column *D*) shows the average diameter for each one of the studied families. We observed very close diameters between *TIM Barrel* and *TIM beta/alpha-barrel* and also between *Lysozyme* and *Lysozyme-like* families. This is

Table 28.1 Families studied, mainly due to their protein number in CATH v3.1.0 and SCOP v1.73

Name	Type	Class	Proteins
Rossmann fold	CATH	αβ	2576
TIM Barrel	CATH	αβ	1051
Lysozyme	CATH	Mainly α	871
Globin like	SCOP	All α	733
TIM β/α barrel	SCOP	α/β	896
Lysozyme like	SCOP	α+β	819

Table 28.2 Average of metrics values for each family [1]

Name	*l*	*D*	*δ*	*z*
Rossmann fold	7.26	18.84	0.033	7.20
TIM Barrel	7.79	19.83	0.030	7.17
Lysozyme	4.99	12.81	0.038	6.82
Globin like	6.64	15.65	0.034	7.69
TIM β/α barrel	7.86	20.09	0.029	7.15
Lysozyme like	5.03	12.85	0.042	6.81

The column *l* regroups the average mean distances of SSE IN. The column *D* represents the average diameter, δ is the average density, and *z* the average mean degree for each studied family

explained by the fact that each pair of families contains almost the same proteins, in other words, *Lysozyme* topology in CATH is the equivalent of *Lysozyme-like* fold level in SCOP.

The diameter being an upper bound of distances in interaction networks, we expect that the mean distance l will be lower than D. Table 28.2 (column l) confirms this. Again, we observed very close values between the equivalent SCOP and CATH families for the reasons discussed above. But we can also see that different families have values which allow discrimination between them based on this parameter. It is interesting to note that the ratio D/l is about 2.5 for all the families. This last property is a characterization of all proteins' SSE-IN.

28.2.2 Density and Mean Degree

The density measures the ratio between the number of available edges and the number of all possible edges. Results presented in Table 28.2 (column δ) show that the two families, *TIM Barrel* and TIM beta/alpha-barrel, have the minimum density. It has a consequence on their SSE-IN topology. When the density is low, the network is less connected and consequently, the diameter and the average distance are higher. Comparing these results to Table 28.2 (columns l, D and δ), one can see the inversely proportional relation between density on one hand and diameter and average distance on the other.

The mean degree is presented in Table 28.2 (column z). The observed values are close enough between one family and another. That is why the mean degree is not a discriminating property, but rather a property characterizing all proteins' SSE-IN.

28.2.3 Degree Distribution

We compute the cumulative degree distribution for all proteins' SSE-IN of studied families. A sample of our results is presented in Fig. 28.1. We can remark that the curves follow a power law distribution, which can be approximated by the following power-law function:

$$p(k) = 141.29\ k^{-\alpha}, \quad \alpha = 2.99 \pm 0.6. \tag{28.1}$$

We observe the same results for all the studied proteins. To explain this behavior, we have to rely on two facts. First, the mean degree of all proteins' SSE-IN evolves weakly (see Table 28.2, column z). Second, the degree distribution, see Fig. 28.2, follows a Poisson distribution whose peak is reached for a degree near z. These two facts imply that for degrees lower than the peak, the cumulative degree distribution decreases slowly, and after the peak, its decrease is fast compared to an exponential one. Consequently, all proteins' SSE-IN studied have a similar cumulative degree distribution which can be approximated by a unique power-law function.

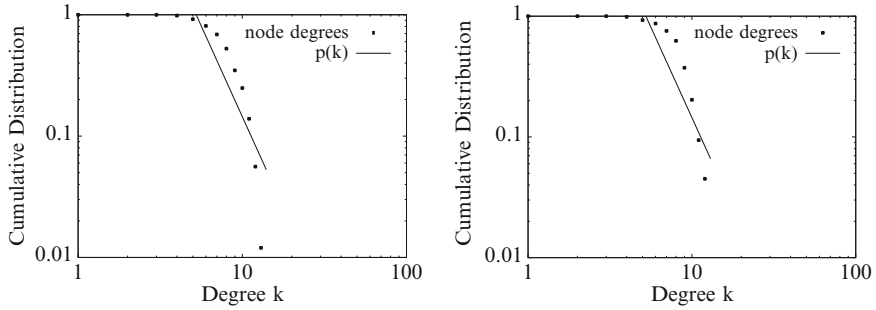


Fig. 28.1 Cumulative degree distribution for 1RXC from Rossman fold (*left*), and 1HV4 from TIM beta/alpha barrel (*right*)

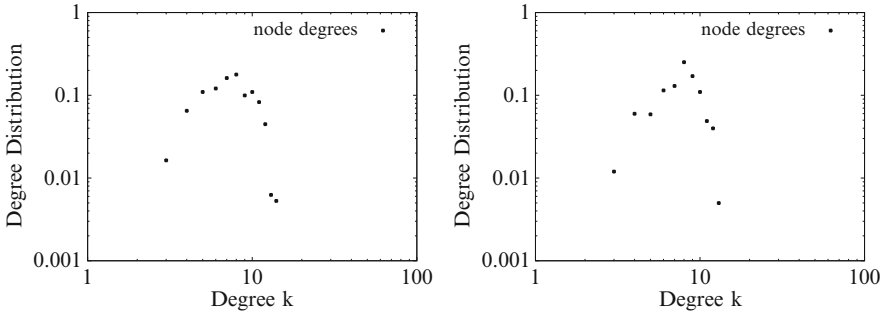


Fig. 28.2 Degree distribution for 1RXC from Rossman fold (*left*), and 1HV4 from TIM beta/alpha barrel (*right*)

28.3 A Topological Space

In the previous section, we give different means to describe a protein structural family characterizing its SSE-IN. Some of the properties, like diameter and density, allow to discriminate two distinct families, while others, like mean degree and degree distribution, are general properties of all SSE-IN. Thus, proteins having similar structural properties and biological functions will also have similar SSE-IN properties. In this way, our model allows us to draw a parallel between biology and graph theory.

Here, we exploit this hypothesis by proposing a topological space where a protein is described by its SSE-IN topology. Then, we want to project the structural families into this topological space to put in evidence that the proteins from a same family have SSE-IN, which are grouped in this topological space. Consequently, we have to determine the dimensions of this topological space, that is, we want to identify the topological criteria that are able to discriminate the SSE-IN according to their families.

Table 28.3 SCOP fold families from ALL alpha class used to build our topological space

SCOP ID	Family name	Protein number
46457	Globin like	817
46688	DNA/RNA binding 3 helical bundle	370
47472	EF Hand like	313
48507	Nuclear receptor ligand binding domain	223
48112	Heme dependent peroxidases	207
48618	Phospholipase A2, PLA2	186
47112	Histone fold	156
46625	Cytochrome c	148
48263	Cytochrome P450	146

To build our topological space, we rely on the study done in the previous section and we apply it on another dataset, see Table 28.3. This new dataset is composed only of structural families from the *All Alpha* class in SCOP v1.73 classification.

First, we know that the mean distances and also the density are the discriminant metrics among the SSE-IN from different structural families. We plot a 3D topological space, see Fig. 28.3, where the x axis represents the SSE-IN size, denoted as N, the y axis represents the densities, denoted as G, and the z axis represents the mean distances, denoted as L. The plots confirm that the study done in the previous section is reliable since the dimensions we use provide a topological space where the proteins' SSE-IN from the same structural family are grouped. Consequently, the parallel between structural and topological properties can be illustrated through the topological space we propose.

Second, we remark that among the protein SSE-IN belonging to a same structural family, there are some which have a very close size. Then, the proteins are grouped around a particular value n to form clusters. Thus, we can also describe the structural families' topological space describing the cluster properties that they form.

To characterize the clusters observed in the topological space, we have to define them. A cluster, denoted $c_{n\ i}$, defined in the neighborhood of a specific SSE-IN size equals to i satisfy:

$$p_{\text{cluster}} \in i \pm \text{radius} \geq r\, p_{\text{family}}, \tag{28.2}$$

where, p_{cluster} designates the number of proteins in the cluster and p_{family} is the total number of proteins considered in the structural family. The parameter r is a threshold, and we use the value of 25%. With this definition, we suppose that some clusters overlap each others. In this case, we merge them and consider that the cluster center equals to i and the cluster radius is the average length between the minimum and the maximum SSE-IN size involved in the cluster.

Table 28.4 shows how the clusters appear in our topological space. We remark that each family has a specific cluster distribution, meaning that the topological space we built is reliable to regroup the SSE-IN according to their structural families. The radius is, in the most case, around 50, meaning that the clusters

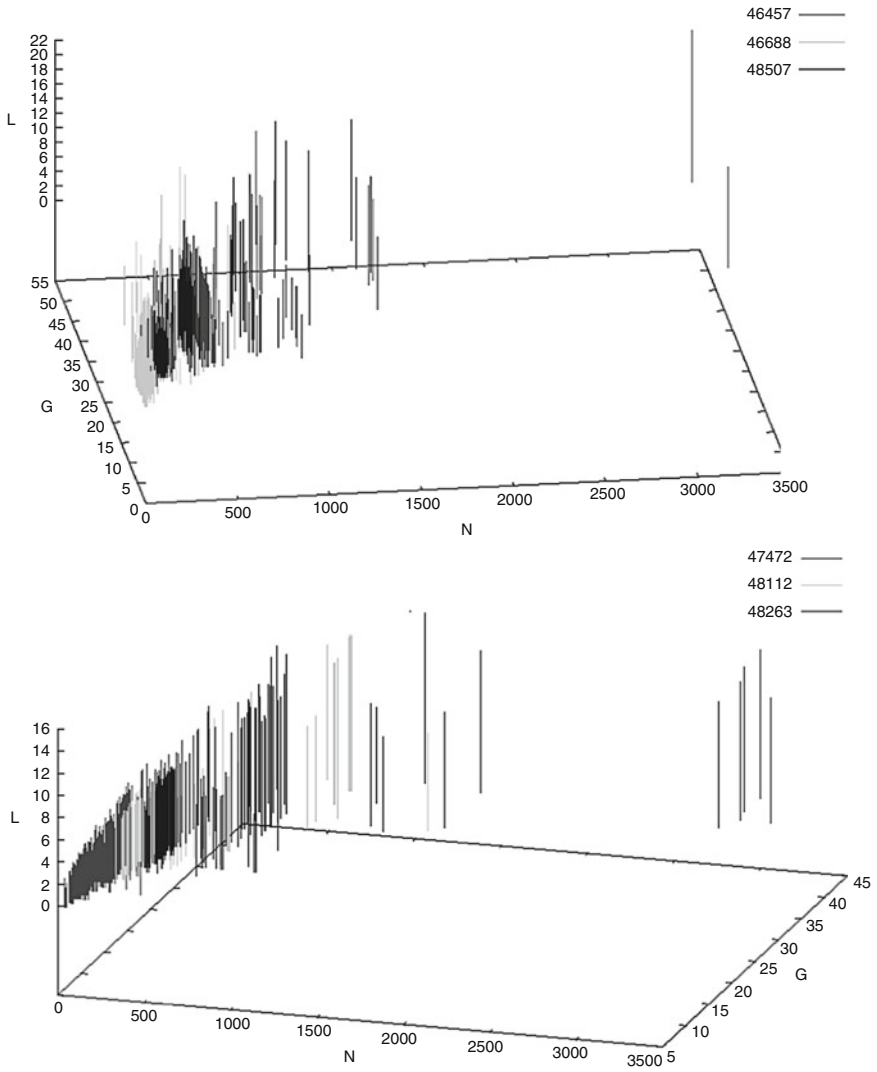


Fig. 28.3 A topological space. The *x* axis represents the SSE IN size, *y* the density, and *z* the average distances. Proteins from a same family tend to be grouped

regroup proteins whose sizes are comparable. The cluster sizes show how the proteins from a family are grouped around a particular neighborhood.

This cluster description is actually a consequence of the family composition. Indeed, the families regroup proteins having a notably close-enough size because their secondary structures are similar.

Table 28.4 Cluster description for each family

SCOP ID	Cluster center	Cluster radius	Cluster size
46457	125	45	33.5
	485	45	39.4
46688	60	60	67.1
47472	90	70	78.6
48507	195	45	51.6
48112	190	50	77.3
48618	75	45	67.2
47112	595	45	76.9
46625	60	60	76.4
48263	275	45	50.7

The cluster size is expressed as the percentage of the total protein number in families

28.4 Conclusion

In this chapter, we consider a protein as an interaction network of amino acids (SSE-IN) and study some of the properties of these networks. It appears that specific properties, like diameter and density, allow to discriminate two distinct families, whereas others are common to all SSE-IN. Thus, proteins whose structural properties are similar will also have similar SSE-IN properties. In this way, our model allows us to draw a parallel between biology and graph theory.

To illustrate the parallel between structural and topological properties, we propose a topological space whose dimensions are strong enough to discriminate among SSE-IN from different structural families. Then, the topological space shows some clusters where the proteins from a same family are grouped. The description of these clusters contributes to distinguish, by a new means, the structural families relying on the topological criterion. Through our topological space, we propose a means to describe a structural family by topological measures.

References

1. Diestel R (2000) Graph Theory. Princeton: Springer.
2. Dokholyan NV, Li L, Ding F, Shakhnovich EI (2002) Topological determinants of protein folding. Proc Natl Acad Sci USA. 99(13):8637 8641

Chapter 29
GNCPro: Navigate Human Genes and Relationships Through Net-Walking

Guozhen Gordon Liu, Elvena Fong, and Xiao Zeng

Abstract The use of computational applications in biological research is significantly lagging behind other scientific research areas such as physics, mathematics, and geology; more in silico tools are needed. The increasing complexity of biological data makes it more and more difficult for scientists to verify their hypotheses and results against existing discoveries. GNCPro is a free data integration and visualization tool for gaining comprehensive overviews of such complicated biological knowledge. In particular, GNCPro warehouses and encodes biological information as binary relationships. When represented graphically, these binary relationships take on the form of edges that connect the genes and proteins, which are represented by nodes. By using distinguishing features such as colors, shape, and opacity, GNCPro provides a stimulating visual experience in which the user can quickly identify groups of genes by annotations and the types of relationships involved. GNCPro integrates human gene expressions, regulations, gene product modifications, and interactions into one platform while delivering a simple and powerful user interface for systems biology study. Availability: http://GNCPro.sabiosciences.com.

Keywords Molecular interactions · Computational systems biology · Biological data mining and knowledge discovery · Biological databases and information retrieval · Biological data integration and visualization

29.1 Introduction

The rapid accumulation of scientific publications makes it difficult for the scientist to keep up with new discoveries related to his/her area of research. The recent developments in high-throughput experimental systems have enabled

G.G. Liu (✉)
SABiosciences Corporation, 6951 Executive Way, Frederick, MD 21703, USA
e mail: gzliu@sabiosciences.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 29,

the generation of biological information at an explosive speed. For example, using yeast two-hybrid screening [1, 2] or mass spectrometry [3, 4] approaches, researchers can generate information about protein protein interactions or protein complexes for a complete proteome in a short term. Using DNA microarray technologies, researchers can survey the expression levels of thousands of genes on a tissue-by-tissue basis at a time [5 7]. Furthermore, NextGen genomic sequencing technologies [8, 9] allow for the generation of data for gene expressions, gene mutations/deletions, and SNPs at a speed never seen before. With such high-throughput technologies and discoveries, new algorithms and computational tools are needed to store, process, and analyze the resulting amount of data and articles.

While many scientists are aware of the existence of genomic information and have used available sequence information in the public domain for their chip and primer designs, few scientists are aware of the existence of proteomic and functional data gathered through high-throughput experiments, let alone use them efficiently. For many biological scientists, the finding and gathering of published high-throughput proteomic data is not an easy job. To effectively use all available information about a biological system whether genomic, genetic, epigenetic, or proteomic we need new methods for data encoding, integration, visualization, analysis, and reporting. GNCPro (Gene Network Central Pro) is the answer to these challenges. GNCPro will aid the biologists in working effectively with the large amount of data available to generate new hypotheses and design experiments.

29.2 Methods

The GNCPro software package is implemented utilizing a three-tier architecture with a Graphic User Interface (GUI) at the frontend, a web server in the middle, and a MySQL database (http://en.wikipedia.org/wiki/MySQL) at the backend. The GUI is embedded in the web browser.

The initial user interface is a PHP-based (http://php.net/index.php) html form that takes a user query and sends the request to the web server. The web server then queries the database server, retrieves the answer, and formats and forwards the contents back to the client browser. An applet program then interprets and displays the results in a graphical format.

In the MySQL database (http://en.wikipedia.org/wiki/MySQL), biological information is encoded as binary relationships between genes. The relationships can be physical interactions, genetic regulations, coexpressions [10, 11], chemical modifications (also referred to as posttranslational modification), etc. These relationships were derived from three different types of resources: textmining, datamining, and data acquisition.

To accommodate user querying and navigation, two online graphic navigation systems have been developed: a text-based SVG platform (powered by Adobe: http://www.adobe.com/svg/) and a Java Applet-based MEDUSA platform (http://sourceforge.net/projects/graph-medusa/). In the GUI, the biological entities such as

genes and proteins are presented as nodes, and the relationships between them are presented as edges. The edges can be directional, such as in chemical modifications and up/down regulations, or nondirectional, such as in physical interactions and coexpressions. Different colored edges encode for different relationships. In addition to color, edges can be further distinguished by opacity and solidity. A solid edge indicates a relationship that is backed by experimental evidence; dashed edges are used for the relationships that have been established by computational predictions.

Text mining: Abstracts from PubMed are downloaded from NCBI (http://www.ncbi.nlm.nih.gov/pmc/about/ftp.html) and split into sentences. Each sentence is broken up into single words. Starting from the end of the sentence, and moving "backwards" in the sentence, the biological entities are extracted and identified with the help of a precompiled dictionary of human genes. Gene relationships between the extracted entities (such as "gene A" positively regulates "gene B") are then identified with the help of another dictionary for relationships.

The text-mining results are then pooled and validated automatically and manually to remove contradicting statements for any particular relationships between two genes. Our text-mining result has an accuracy of 80%.

Data acquisition: High-throughput protein protein interaction data [12 15] and microarray profiling data for human gene expression in a variety of tissues [5 7] were downloaded as supplementary tables or from the links provided by the authors. Additional microarray profiles were downloaded from NCBI/GEO (http://www.ncbi.nlm.nih.gov/sites/entrez?db=geo). We also licensed protein protein interaction data from the Human Protein Reference Database [16]. Protein protein interaction predictions were downloaded from ProLinks [17]. A subset of the coexpression data used by GNCPro was downloaded from the GEMMA database [10].

Data mining: Gene coexpression relationships were calculated according to Lee et al. [10]; the tissue expression profiles for human genes were calculated using Shannon's Entropy formula [18].

29.3 Results

29.3.1 Graphic User Interface

The graphic user interface is presented in a web browser as an embedded form with a navigation menu (http://GNCPro.sabiosciences.com). Users can start using GNCPro from any one of the following five entry methods: (a) Input a gene or a custom list of genes (or upload a custom list of genes); (b) Browse existing pathways and load the genes that belong to the selected pathway; (c) Navigate the gene ontology and select the genes that are annotated by a specific ontology term; (d) Search human disease terms and select all the genes that are associated with a disease; (e) Start from a previously saved job ID. In order to save a job for later access, the user will need to register an account with GNCPro. The registration is completely free.

The users can find additional information about GNCPro in the "About," "FAQ," and "Tutorial" sections.

29.3.2 Network Navigation Pad

Once the query results are returned, a graph (Fig. 29.1) is presented to the user. In this graph, genes and proteins are encoded by nodes, which come in different shapes and colors to represent different annotations. The edges indicate the relationships between genes and proteins and are characterized by different attributes: color to distinguish between relationship types, such as physical interactions, functional and transcriptional regulations, modifications, coexpressions, etc.; solid versus dashed states to indicate experimentally-discovered versus computationally-predicted relationships; arrowheads to illustrate directionality of relationship (if applicable); and opacity to suggest relationship "strength" or confidence.

29.3.3 Data Integration

Through textmining, we acquired 93,155 binary pairs that describe regulatory relationship between two genes. Regulatory relationships can be further subclassified as up, down, activating, or inhibiting regulations. In cases where the text mining program was not able to determine whether a regulation was positive or negative, we simply assigned the connection as a "regulation" to describe the directional relationship between the two genes. From collected existing pathways, we extracted 48,977 pairs of gene gene relations and added them to the database. 2,353 pairs of chemical modifications (posttranslational modifications) and 32,804 pairs of physical protein protein interactions were acquired from HPRD and also added to the database. From the predicted protein protein interactions, we added 43,364 pairs that had a 90% confidence level to the GNCPro database. With 216,267 pairs of coexpression relationships and 122,700 pairs of regulation relationships involving transcription factors and target genes, GNCPro has harnessed a total of 559,620 binary pairs of gene gene relations. These genes cover 20,307 unique genes out of a predicted approximate total of 25,000 genes in the human genome.

29.4 Discussions

MEDUSA is used by STRING [19], a popular data service provider for known and predicted protein protein interactions. In order to provide users with an interface "looked familiar", we have chosen MEDUSA as the basic tool for graph

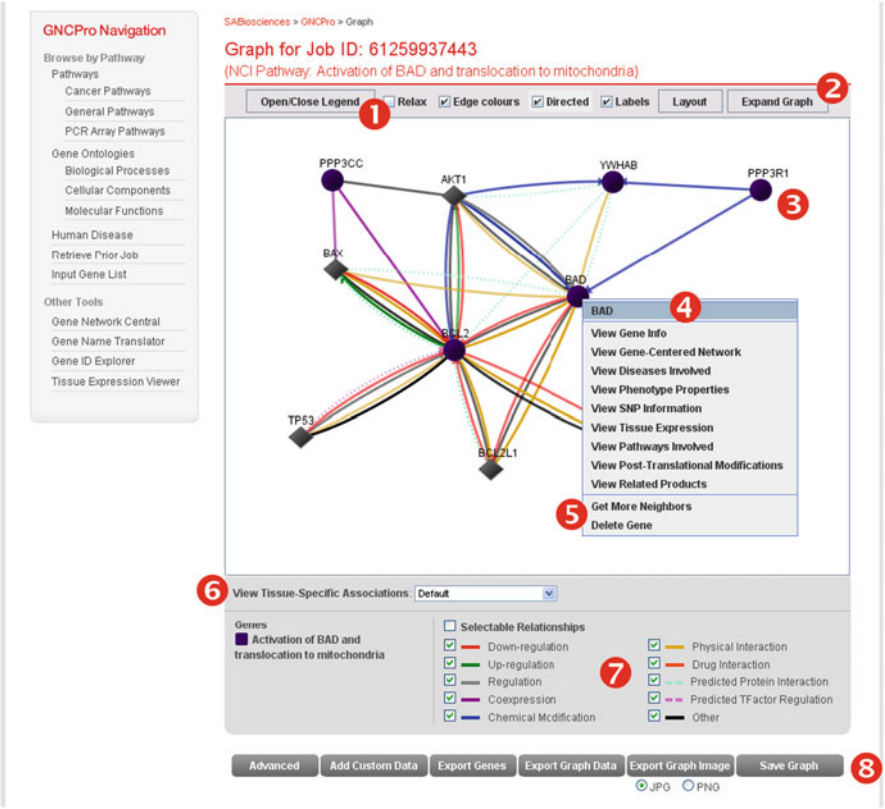


Fig. 29.1 A brief overview of the network navigation pad and major features. (1) Manipulate the appearance of the graph with simple controls to toggle certain options on/off and relax the graph. (2) Expand graph by adding neighbors with one click button access. (3) Highly interactive graph features labels that popup when mousing over nodes and edges, clickable edges and nodes that point to supporting evidence and additional information, and nodes that can be dragged and dropped around the graph. (4) Right click on a node to open a menu that allows quick and convenient access to a variety of useful resources. (5) Expand and delete individual genes to conduct a more focused research. (6) Use tissue overlays to view networks in context of tissues and explore tissue specific relationships. (7) Toggle relationships on/off in the legend to view relation ships of interest. (8) Click on simple buttons to add custom data, export data in either graphical or textual formats for use with other programs, or save data to continue research later. The custom data option allows users to add and view their own data against existing discoveries

manipulation. Compared to Cytoscape [20] and the IM browser [21], we have reduced those functions that are aimed at the whole network while focusing, at the same time, on those functions that might help users who are conducting pathway- and biomarker-related research. Furthermore, when comparing GNCPro to other similar software, GNCPro offers more comprehensive data integration and a more user-friendly navigation interface. Our aim is to allow a regular user to master the program in less than 15 min. All the functions are immediately available

on the Network Navigation Pad. You do not need to search through pull-down menus to find the right function. The legend is simple and comes with every graph and can be easily turned off if desired.

Even with such a simple interface, GNCPro is powerful enough to help scientists to identify top-interacting partners for a group of genes of interest, find candidates to bridge the gap between genes in a fragmented pathway, identify potential substrates for a kinase (and vice versa), predict potential regulatory targets for a transcription factor (and vice versa), view genes and networks within the context of tissues, access supporting data for a given relationship with one click of the mouse and integrate and view their own discoveries against the background of well-known and existing data. And, last but not the least, the user can export the graph and data for sharing and publication purposes, or even save the graph to continue their research later.

Advanced users can use GNCPro to aid them in generating hypotheses, modeling the signal transduction pathways, gaining an insight into a pathway's dynamic behavior, avoiding adverse drug effects, and increasing productivity in biomarker identifications.

GNCPro will change the way how biomedical data are searched and presented. The semantic representation of biological knowledge by physically displaying the relationships between individual genes is more impressive than the plain text description. New biological discoveries can be easily integrated into this platform. Scientists and even bioinformaticians will appreciate the fact that GNCPro has so much data integrated in it. GNCPro is the software of choice in the era of systems biology.

Conflict of interest: none declared.

References

1. Uetz P, Finley RL Jr (2005) From protein networks to biological systems. FEBS Lett 579 (8):1821 1827.
2. Rajagopala SV, Uetz P (2009) Analysis of protein protein interactions using array based yeast two hybrid screens. Meth Mol Biol 548:223 245.
3. Gstaiger M, Aebersold R (2009) Applying mass spectrometry based proteomics to genetics, genomics and network biology. Nat Rev Genet 10(9):617 627.
4. Yates JR, Ruse CI, Nakorchevsky A (2009) Proteomics by mass spectrometry: approaches, advances, and applications. Annu Rev Biomed Eng 11:49 79.
5. Su AI et al. (2004) A gene atlas of the mouse and human protein encoding transcriptomes. Proc Natl Acad Sci U S A 101(16):6062 6067.
6. Ge X et al. (2005) Interpreting expression profiles of cancers by genome wide survey of breadth of expression in normal tissues. Genomics 86(2):127 141.
7. Hsiao L L et al. (2001) A compendium of gene expression in normal human tissues reveals tissue selective genes and distinct expression patterns of housekeeping genes. Physiol Genom 7(2):97 104.
8. Voelkerding KV et al. (2009) Next generation sequencing: from basic research to diagnostics. Clin Chem 55(4):641 658.

9. Mardis ER (2008) The impact of next generation sequencing technology on genetics. Trends Genet 24(3):133 141.
10. Lee HK et al. (2004) Coexpression analysis of human genes across many microarray data sets. Genome Res 14(6):1085 1094.
11. Yan X et al. (2007) A graph based approach to systematically reconstruct human transcrip tional regulatory modules. Bioinformatics 23(13):577 586.
12. Lim J et al. (2006) A protein protein interaction network for human inherited ataxias and disorders of Purkinje cell degeneration. Cell 125(4):801 814.
13. Rual JF et al. (2005) Towards a proteome scale map of the human protein protein interaction network. Nature 437(7062):1173 1178.
14. Stelzl U et al. (2005) A human protein protein interaction network: a resource for annotating the proteome. Cell 122(6):957 968.
15. Friedman A, Perrimon N (2007) Genetic screening for signal transduction in the era of network biology. Cell 128(2):225 231.
16. Prasad TSK et al. (2009) Human protein reference database 2009 update. Nucleic Acids Res 37:D767 D72.
17. Bowers PM et al. (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol 5(5):R35.
18. Schug J et al. (2005) Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol 6(4):R33.
19. Jensen LJ et al. (2009) STRING 8 a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37(Database issue):D412 D416.
20. Shannon P et al. (2003) Cytoscape: a software environment for integrated models of biomo lecular interaction networks. Genome Res 13(11):2498 2504.
21. Pacifico S et al. (2006) A database and tool, IM browser, for exploring and integrating emerging gene and protein interaction data for *Drosophila*. BMC Bioinformatics 7:195.

Chapter 30
Small-Scale Modeling Approach and Circuit Wiring of the Unfolded Protein Response in Mammalian Cells

Rodica Curtu and Danilo Diedrichs

Abstract The accumulation of unfolded proteins in the endoplasmic reticulum (ER) activates a mechanism whose primary functions are to sense any perturbation in the protein-folding capacity of the cell, and correct the situation to restore homeostasis. This cellular mechanism is called the unfolded protein response (UPR). We propose a biologically plausible computational model for the UPR under ER stress in mammalian cells. The model accounts for the signaling pathways of PERK, ATF6, and IRE1 and has the advantage of simulating the dynamical (timecourse) changes in the relative concentrations of proteins without any a priori steady-state assumption. Several types of ER stress can be assumed as input, including long-term (eventually periodic) stress. Moreover, the model allows for outcomes ranging from cell survival to cell apoptosis.

Keywords Endoplasmic reticulum stress · Unfolding protein response · Signaling pathways · PERK · ATF6 · IRE1

30.1 Introduction

The endoplasmic reticulum (ER) is a large intracellular organelle that plays an essential role to the functionality and survival of the cell; it is a major calcium storage site as well as a site where secretory and membrane proteins are modified, folded, and assembled. Perturbations to the ER that affect its protein-folding capacity can be induced in several ways, for example, through pathogenic infections, chemical insult, genetic mutation, or nutrient deprivation. Consequently, the ER experiences an accumulation of unfolded or misfolded proteins, a situation which is generally termed *ER stress* [10]. The cellular mechanism that addresses

R. Curtu (✉)
Department of Mathematics, University of Iowa, Iowa City, IA 52242, USA
e mail: rodica curtu@uiowa.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 30,

ER stress is called the *unfolded protein response* (UPR) and it is initiated in mammalian cells by three ER-resident transmembrane proteins: PERK, ATF6, and IRE1 [7, 10].

The UPR is a complex signaling network consisting of both transcriptional and translational steps that yield an overall improvement in the ER protein-folding function. While most of the network components and interactions are known (and some are hypothesized) [5], how the UPR acts as a whole and why under certain conditions it allows for differential outcomes such as adaptation or cell death (apoptosis) are still open questions [8, 9]. Moreover the UPR secretory pathways involve overlapping and feedback loops, making experimental manipulations based on intuitive reasoning alone costly and very difficult. In this context, mathematical modeling becomes imperative and, when coupled to real data, a very efficient tool.

In this paper, we propose a biologically plausible computational model for the UPR in the ER under stress. The model accounts for the signaling pathways of PERK, ATF6, and IRE1 proteins, and it is constructed using the biochemical (kinetic-like equations) framework. The calibration of parameters of the model as well as the model validation necessitate comparison against experimental data and they are beyond the scope of the paper. We focus instead on the description of the model's conceptual principles and the definition of its variables and equations.

Some modeling attempts of the UPR were recently initiated by research teams investigating either: (1) the paradox of simultaneous activation by the UPR of both adaptive and pro-apoptotic pathways [9], or (2) the importance of a translation attenuation mechanism for the proper functioning of mammalian secretory cells under stress such as insulin-producing pancreatic beta-cells [11]. The former is a very simplistic, linear ordinary differential equation system based on crude assumptions such as uniform rates of production and degradation for all proteins along the signaling pathway. Mainly, it was designed to fit the values of chaperone BiP, and of CHOP and GADD34 obtained from the experimental data. The latter model does account for some nonlinearities by assuming Michaelis Menten kinetics but works under the assumption that the upregulation of chaperones is much slower than the translation attenuation mechanism; that corresponds to an ad hoc reduction of the concentrations of most of the proteins involved in the UPR to their steady-state condition. Moreover the model is valid only if restoration of homeostasis is possible and this excludes the alternative case of apoptosis.

In contrast to these examples, the model we propose for the UPR is much more general. It incorporates what is known to date about this signaling network while refraining from any a priori assumption on the timescales involved in the system. We believe the correct timescales would rise naturally from the model once its parameters (for example, the rate constants of production and degradation) are estimated based on experimental data. Knowing the timescales will then allow us to simplify the model in its relevant functional regime. The model has also the advantage of simulating the dynamical changes in the relative concentrations of proteins and it is not restricted to instantaneous snapshots resulting from steady-state conditions. We include the proteins that play a key role in the response to ER stress and take into account how each of them regulates the activity of the others.

Table 30.1 Main components of the intracellular signaling pathway activated by the accumulation of unfolded proteins in the endoplasmic reticulum

Acronyms	Full name
PERK	Pancreatic endoplasmic reticulum kinase
eIF2α	Eukaryotic initiation factor 2 alpha
ATF4	Activating transcription factor 4
ATF6	Activating transcription factor 6
IRE1	Inositol requiring kinase 1
XBP1	X box binding protein 1
BiP	Binding immunoglobulin protein (or Grp78: glucose regulated protein 78)
CHOP	(Controlled amino acid therapy)/enhancer binding protein Homologous Protein
GADD34	Growth arrest and DNA damage protein 34

Therefore, in the model, the former are the variables and the latter correspond to the model equations. Several types of ER stress can be assumed as input, including long-term (eventually periodic) stress. In addition, the model allows for both survival and apoptosis as possible outcomes; the switch between survival and apoptotic biological states corresponds in the model to a dynamical switch between two distinct attractors. Experimentally, this can be reflected, for example, by the levels reached by CHOP and/or GADD34.

As Kim et al. point out in a recent review [3], both small-scale and large-scale computational techniques are used in cell biology modeling. Small-scale models are particularly useful when the key components and wiring of regulatory circuits are known and the focus is on the understanding of the functionality of the circuit. They typically rely on thermodynamics or kinetic-like approaches and they were quite successful in prokaryotes (for example, in *Escherichia coli* [2]) due to their relatively simple regulatory networks, but also in lower eukaryotic systems such as *yeast* [1]. On the other hand, large-scale models are used to infer circuit components and wiring when there is significant lack of knowledge about them. They include linear, Bayesian, and logical (Boolean) models. We choose to formulate the UPR problem using a small-scale approach because the scientific literature (e.g., [4, 5, 7, 10]) provides sufficient details about the components and biochemical interactions in the UPR circuit in mammalian secretory cells. We therefore take advantage of the available information and incorporate it in a new quantitative model for the UPR under ER stress. Note that we use in the paper acronyms such as PERK, IRE1, CHOP, and several others; they are well-established in the cell biology community but for completeness we have also included their full names in Table 30.1.

30.2 Modeling of the UPR and Construction of the Wiring Diagram

The ability for a cell to fold proteins is critical for its survival and for maintaining the well-functioning of the (multicellular) organism. Protein folding occurs in lumen of the endoplasmic reticulum (ER) where the ribosomes translate the information

from the mRNA into a chain of amino-acids. This chain must then be folded into a precise three-dimensional structure that allows the newly created protein to function. After a protein is successfully folded, it exits the ER and either moves to another part of the cell to perform a function internal to the cell or it is secreted out of the cell to be used in another part of the organism. Different types of environmental insult can produce stress in the ER and affect the cell's ability to fold proteins. This creates an increase in the amount of unfolded or misfolded proteins in the ER which remain in the ER until they are either correctly folded, or until they are degraded.

The accumulation of unfolded proteins activates a mechanism called the UPR whose primary functions are to sense any perturbation in the protein-folding capacity of the cell, and correct the situation in order to restore homeostasis. In mammalian cells the UPR signaling pathways are initiated by three ER stress sensors: PERK, ATF6, and IRE1. In the absence of stress, these three ER-transmembrane proteins exist in complex form, associated with the ER-resident protein BiP. (The latter is an abundant chaperone that assists in the process of protein folding.) ER stress leads to the dissociation of BiP from PERK, ATF6, and IRE1 thus effectively activating the last three as well as freeing BiP to assist with the protein folding. Additional key players in the signaling cascade of the UPR are the eukaryotic translation initiation factor-2 (eIF2), the transcription factor ATF4 and the proteins XBP1, CHOP, and GADD34.

Our modeling approach to UPR accounts for different biochemical processes that a network component may undergo (synthesis, degradation, association, dissociation, phosphorylation, dephosphorylation, splicing, transcription, or translation) as well as for the effect each component may have on the others. All these are summarized in the so-called *wiring diagram* (see Fig. 30.1) in which the nodes (in bold letters) are the UPR components while the arrows represent their biochemical transformations and/or interactions; small letters denote mRNAs and capital letters denote proteins. The wiring diagram is then converted to a mathematical model according to the following rule: each node in the diagram becomes a variable in the model, and each arrow pointing to or from a node corresponds to a term in the variable's differential equation. The arrows in the wiring diagram may represent several types of biochemical reactions:

- *Synthesis/transcription and degradation*. The synthesis of some mRNAs and proteins is indicated by an arrow pointing to their corresponding node. The degradation of mRNA or a protein is drawn as an arrow pointing away from the node. If, for example, the degradation arrow does not have a destination that means the model does not track the products of the degradation.
- *Translation*. An arrow from the mRNA to the protein indicates its translation.
- *Association/Dissociation*. In the absence of stress, the sensor proteins PERK, ATF6, and IRE1 exist in complex form, associated with BiP. When the UPR is initiated these complexes dissociate. When the response is no longer needed, PERK and IRE1 re-associate with BiP.
- *Phosphorylation/Dephosphorylation*. PERK, IRE1, and eIF2α are phosphorylated when a phosphate group (PO_4) binds to the molecule. Phosphorylation

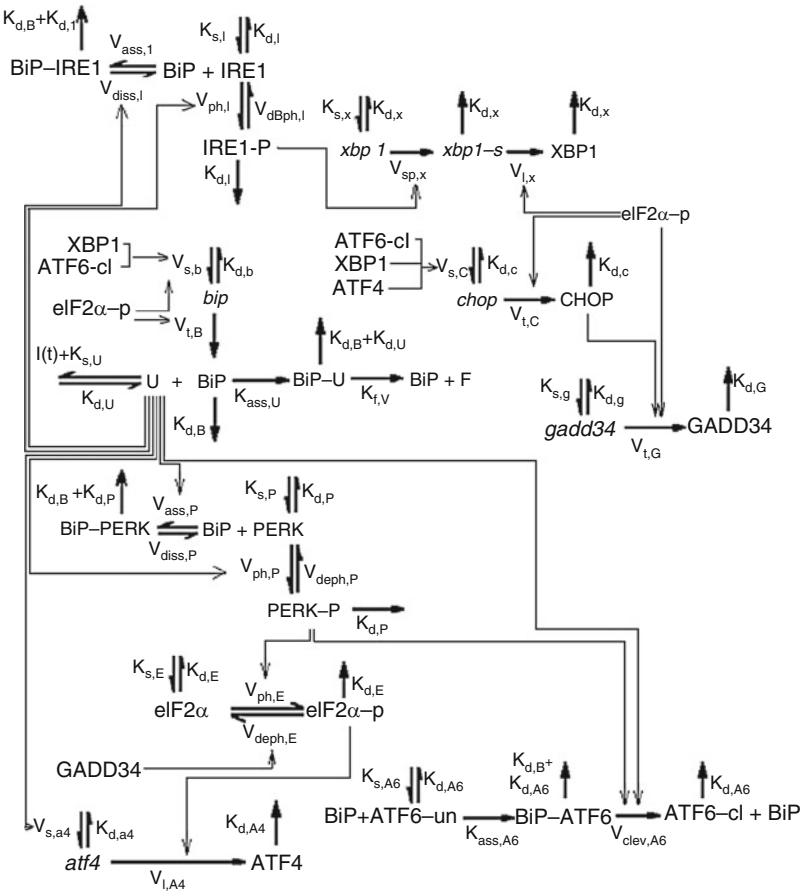


Fig. 30.1 Wiring diagram for the unfolded protein response in the endoplasmic reticulum of mammalian cells under stress

effectively activates certain properties of these proteins. The reaction is reversible through a dephosphorylation process.

- *Cleavage*. The arrow from [BiP ATF6] to [ATF6-cl + BiP] indicates the dissociation of BiP from the ATF6 followed by immediate cleavage of the free ATF6 molecule. This reaction is irreversible.
- *Splicing*. The arrow from xbp1 to xbp1-s indicates the splicing of the xbp1 mRNA. This reaction is irreversible.
- *Folding*. When a BiP chaperone comes together with an unfolded protein, the unfolded protein is converted to a folded protein. This reaction is irreversible.

To maintain a consistent notation in the definition of the UPR model, we also adopt the following convention: The proportionality rate coefficients in all equations will be denoted by either the letter k or the letter V followed by two indices:

the first index corresponds to the process involved, while the second index is associated to the protein's type. Whenever the proportionality rate coefficient is constant, the letter k will be assigned; whenever the rate coefficient depends on other network components (it is a function of other variables), letter V will be used instead. For example, $k_{s,P}$ means the synthesis rate (s) of PERK (P) and is assumed to be constant; on the other hand $V_{t,A4}$ is the rate of translation (t) of ATF4 ($A4$) and it is considered a function of the concentration [eIF2α-p] of phosphorylated-eIF2α.

Let us now explain the principles behind the construction of the UPR model.

30.2.1 Signaling Cascade Initiated by PERK Activation

In ER-stressed cells, PERK undergoes trans-autophosphorylation by oligomerization which leads to phosphorylation of eIF2α (the α-unit of eukaryotic translation initiation factor-2); then the eIF2α-phosphorylation stimulates the synthesis of ATF4-protein from its mRNA [5]. PERK activation by ER stress is, however, rapidly reversible (activated PERK is dephosphorylated within minutes of restoring ER homeostasis and re-enters its complex form with BiP). Moreover, phosphorylated eIF2α is also subject to negative regulation through baseline dephosphorylation and through the negative feedback loop induced by GADD34 (the growth arrest and DNA-damage protein-34).

Due to the dimerization condition PERK + PERK $\rightarrow$ 2 PERK we choose in our model quadratic kinetics for PERK phosphorylation; in addition, since association/dissociation of PERK with/from BiP is directly related to ER stress and since ER stress is reflected by an increase in unfolded proteins, we take these reaction rates to be functions of [U], the concentration of unfolded proteins. That is, we define $V_{ass,P} = V_{ass,P}([U])$ and $V_{diss,P} = V_{diss,P}([U])$. Similarly, since phosphorylation and dephosphorylation rates of PERK depend on the level of ER stress we define them as $V_{ph,P} = V_{ph,P}([U])$ and $V_{deph,P} = V_{deph,P}([U])$, respectively. In addition, we consider constant rates of synthesis $k_{s,P}$ and degradation $k_{d,P}$ for PERK and a constant rate $k_{d,B}$ of degradation for BiP. The concentrations of the BiP PERK complex, free PERK and phosphorylated PERK are denoted by [BiP PERK], [PERK], and [PERK-p] and they satisfy the following differential equations

$$\frac{d[\text{BiP}-\text{PERK}]}{dt} = -(k_{d,B}+k_{d,P})[\text{BiP}-\text{PERK}] + V_{ass,P}[\text{BiP}][\text{PERK}]$$

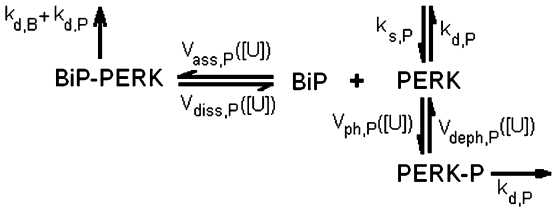
$$- V_{diss,P}[\text{BiP}-\text{PERK}],$$

$$\frac{d[\text{PERK}]}{dt} = k_{s,P} - k_{d,P}[\text{PERK}] - V_{ass,P}[\text{BiP}][\text{PERK}] + V_{diss,P}\,[\text{BiP}-\text{PERK}]$$

$$- V_{ph,P}[\text{PERK}]^2 + V_{deph,P}[\text{PERK}-\text{p}]^2,$$

$$\frac{\mathrm{d}[\mathrm{PERK}-\mathrm{p}]}{\mathrm{d}t} = -k_{\mathrm{d,P}}[\mathrm{PERK}-\mathrm{p}] + V_{\mathrm{ph,P}}[\mathrm{PERK}]^2 - V_{\mathrm{deph,P}}[\mathrm{PERK}-\mathrm{p}]^2. \quad (30.1)$$

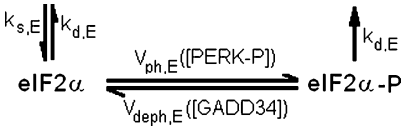
Note that a simple way to describe the modeling equations for PERK given by (30.1) is to draw the corresponding wiring diagram. We include below the wiring diagram for PERK and we will do this again for all the other UPR variables; the wiring diagram will always follow the system of differential equations modeling the UPR variables and it is intended to summarize the biochemical processes and interactions specific to that component.



Since the phosphorylation of PERK catalyzes the phosphorylation of eIF2α and the protein GADD34 initiates the dephosphorylation of eIF2α we define $V_{\mathrm{ph,E}} = V_{\mathrm{ph,E}}([\mathrm{PERK}-\mathrm{p}])$ and $V_{\mathrm{deph,E}} = V_{\mathrm{deph,E}}([\mathrm{GADD34}])$. For example, in the linear case they are $V_{\mathrm{ph,E}} = k'_{\mathrm{ph,E}} + k''_{\mathrm{ph,E}}[\mathrm{PERK}-\mathrm{p}]$ and $V_{\mathrm{deph,E}} = k'_{\mathrm{deph,E}} + k''_{\mathrm{deph,E}}[\mathrm{GADD34}]$. Assuming constant rates of synthesis $k_{\mathrm{s,E}}$ and degradation $k_{\mathrm{d,E}}$, the concentrations [eIF2α] and [eIF2α-p] of free and phosphorylated eIF2α satisfy the equations

$$\frac{\mathrm{d}[\mathrm{eIF2}\alpha]}{\mathrm{d}t} = k_{\mathrm{s,E}} - k_{\mathrm{d,E}}[\mathrm{eIF2}\alpha] - V_{\mathrm{ph,E}}[\mathrm{eIF2}\alpha] + V_{\mathrm{deph,E}}[\mathrm{eIF2}\alpha - \mathrm{p}],$$

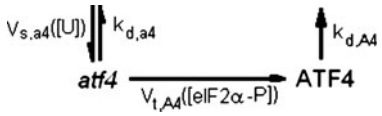
$$\frac{\mathrm{d}[\mathrm{eIF2}\alpha - \mathrm{p}]}{\mathrm{d}t} = -k_{\mathrm{d,E}}[\mathrm{eIF2}\alpha - \mathrm{p}] + V_{\mathrm{ph,E}}[\mathrm{eIF2}\alpha] - V_{\mathrm{deph,E}}[\mathrm{eIF2}\alpha - \mathrm{p}]. \quad (30.2)$$



In the modeling of ATF4 during the ER stress we focus on the transcription step of its mRNA (*atf4*) as well as on the protein synthesis. ER stress upregulates the expression of *atf4* mRNA (though the specific mechanism is unclear [6]), while the phosphorylation of eIF2α stimulates the translation of ATF4 protein. We include these steps into the model by assuming constant rates of degradation for *atf4* mRNA and ATF4 protein ($k_{\mathrm{d},a4}$, $k_{\mathrm{d},A4}$), translation rate for ATF4 depending on phosphorylated eIF2α ($V_{\mathrm{t},A4} = V_{\mathrm{t},A4}([\mathrm{eIF2}\alpha\text{-p}])$) and rate of transcription of *atf4* depending on the amount of unfolded proteins ($V_{\mathrm{s},a4} = V_{\mathrm{s},a4}([\mathrm{U}])$). Then the concentrations [*atf4*] and [ATF4] will be computed according to the differential equations

$$\frac{\mathrm{d}[\mathrm{atf4}]}{\mathrm{d}t} = -(k_{\mathrm{d},a4} + V_{\mathrm{t},A4})[\mathrm{atf4}] + V_{\mathrm{s},a4},$$

$$\frac{\mathrm{d}[\mathrm{ATF4}]}{\mathrm{d}t} = -k_{\mathrm{d},A4}[\mathrm{ATF4}] + V_{\mathrm{t},A4}[\mathrm{atf4}]. \tag{30.3}$$



30.2.2 Signaling Cascade Initiated by IRE1 Activation

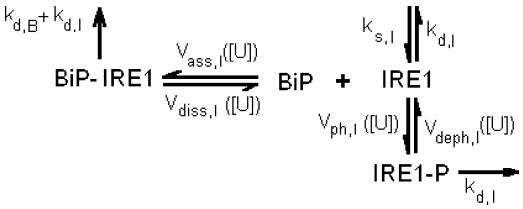
The ER-resident transmembrane protein IRE1 oligomerizes in response to unfolded proteins allowing for trans-autophosphorylation; unlike PERK, however, the only known substrate of the IRE1 kinase is IRE1 itself [5]. An increase in unfolded proteins causes the BiP IRE1 complex to dissociate allowing for dimerization (IRE1 + IRE1 $\rightarrow$ 2 IRE1) and phosphorylation of IRE1; on the other hand, when the folding environment in the ER is restored to normal, IRE1 is rapidly inactivated by dephosphorylation and reformation of the BiP IRE1 complex. Using the notation [IRE1], [BiP IRE1] and [IRE1-p] for concentration of the kinase in its free, complex with BiP, or phosphorylated state, and defining constant rates for synthesis and degradation of IRE1 ($k_{\mathrm{s,I}}$ and $k_{\mathrm{d,I}}$) and variable rates for association with and dissociation from BiP ($V_{\mathrm{ass,I}} = V_{\mathrm{ass,I}}([\mathrm{U}])$ and $V_{\mathrm{diss,I}} = V_{\mathrm{diss,I}}([\mathrm{U}])$) and for phosphorylation and dephosphorylation of IRE1 ($V_{\mathrm{ph,I}} = V_{\mathrm{ph,I}}([\mathrm{U}])$, $V_{\mathrm{deph,I}} = V_{\mathrm{deph,I}}([\mathrm{U}])$) we can write the model equation for this UPR component. As a reminder the constant $k_{\mathrm{d,B}}$ is the rate of degradation for BiP.

$$\frac{\mathrm{d}[\mathrm{BiP}-\mathrm{IRE1}]}{\mathrm{d}t} = -(k_{\mathrm{d,B}} + k_{\mathrm{d,I}})[\mathrm{BiP}-\mathrm{IRE1}] + V_{\mathrm{ass,I}}[\mathrm{BiP}][\mathrm{IRE1}] - V_{\mathrm{diss,I}}[\mathrm{BiP}-\mathrm{IRE1}],$$

$$\frac{\mathrm{d}[\mathrm{IRE1}]}{\mathrm{d}t} = k_{\mathrm{s,I}} - k_{\mathrm{d,I}}[\mathrm{IRE1}] - V_{\mathrm{ass,I}}[\mathrm{BiP}][\mathrm{IRE1}] + V_{\mathrm{diss,I}}[\mathrm{BiP} - \mathrm{IRE1}],$$

$$- V_{\mathrm{ph,I}}[\mathrm{IRE1}]^2 + V_{\mathrm{deph,I}}[\mathrm{IRE1} - \mathrm{p}]^2,$$

$$\frac{\mathrm{d}[\mathrm{IRE1} - \mathrm{p}]}{\mathrm{d}t} = -k_{\mathrm{d,I}}[\mathrm{IRE1} - \mathrm{p}] + V_{\mathrm{ph,I}}[\mathrm{IRE1}]^2 - V_{\mathrm{deph,I}}[\mathrm{IRE1} - \mathrm{p}]^2. \tag{30.4}$$



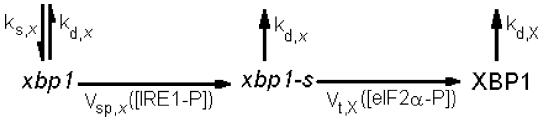
The mechanism through which trans-autophosphorylation of the kinase domain of IRE1 causes the splicing of the mRNA *xbp1* that encodes the transcription factor XBP1 (X-box binding protein-1) is not well understood. What is known is that IRE1 cuts the precursor *xbp1* mRNA twice, excising an intervening fragment or intron; then certain mRNA fragments are ligated, generating a spliced mRNA (*xbp1-s*) which encodes an activator of UPR target genes [5]. The translation of XBP1 protein from *xbp1*-spliced mRNA is expected to be subject to the general translational repression caused by the presence of phosphorylated eIF2α.

Let us then consider [*xbp1*], [*xbp1-s*], and [XBP1] the concentrations of *xbp1* mRNA, *xbp1*-spliced mRNA, and XBP1 protein, and define constant rate of synthesis and degradation for *xbp1* mRNA ($k_{s,x}$, $k_{d,x}$), constant rate of degradation for XBP1 protein ($k_{d,X}$), and variable rates for mRNA splicing and translation ($V_{sp,x} = V_{sp,x}([\text{IRE1-p}])$, $V_{t,X} = V_{t,X}([\text{eIF2}\alpha\text{-p}])$). Then the equations for XBP1 are

$$\frac{\mathrm{d}[\text{xbp1}]}{\mathrm{d}t} = k_{s,x} - k_{d,x}[\text{xbp1}] - V_{sp,x}[\text{xbp1}],$$

$$\frac{\mathrm{d}[\text{xbp1} - \text{s}]}{\mathrm{d}t} = -(k_{d,x} + V_{t,X})[\text{xbp1} - \text{s}] + V_{sp,x}[\text{xbp1}],$$

$$\frac{\mathrm{d}[\text{XBP1}]}{\mathrm{d}t} = -k_{d,X}[\text{XBP1}] + V_{t,X}[\text{xbp1} - \text{s}]. \tag{30.5}$$



30.2.3 Activation of ATF6 Through Cleavage

ATF6 is another transmembrane protein that is activated under ER stress. An increase of unfolded proteins in the ER initiates a trafficking event of ATF6 from the ER to the Golgi [5]; in the Golgi apparatus ATF6 is cleaved by Golgi-resident proteases to release the cytosolic fragment DNA-binding portion ATF6f. The cytosolic fragment of ATF6 moves then to the nucleus to activate gene expression.

In our model [ATF6-un] and [ATF6-cl] denote the concentrations of free uncleaved (inactive precursor) ATF6 and cleaved (cytosolic fragment/transcription factor) ATF6, respectively, while [BiP ATF6] is the concentration of the complex of ATF6 with BiP. We assume constant rates of synthesis ($k_{s,A6}$) and degradation ($k_{d,A6}$) for ATF6, constant association rate of the BiP ATF6 complex ($k_{ass,A6}$) but the rate of cleavage for ATF6 is taken to depend on phosphorylated PERK and on the amount of unfolded proteins ($V_{clev,A6} = V_{clev,A6}([\text{PERK-p}], [\text{U}])$); again $k_{d,B}$ is the rate of degradation for BiP.

$$\frac{\mathrm{d}[\mathrm{BiP}-\mathrm{ATF6}]}{\mathrm{d}t} = -(k_{\mathrm{d,B}} + k_{\mathrm{d},A6})[\mathrm{BiP}-\mathrm{ATF6}] - V_{\mathrm{clev},A6}[\mathrm{BiP}-\mathrm{ATF6}],$$

$$+ k_{\mathrm{ass},A6}[\mathrm{BiP}][\mathrm{ATF6}-\mathrm{un}],$$

$$\frac{\mathrm{d}[\mathrm{ATF6}-\mathrm{un}]}{\mathrm{d}t} = k_{\mathrm{s},A6} - k_{\mathrm{d},A6}[\mathrm{ATF6}-\mathrm{un}] - k_{\mathrm{ass},A6}[\mathrm{BiP}][\mathrm{ATF6}-\mathrm{un}],$$

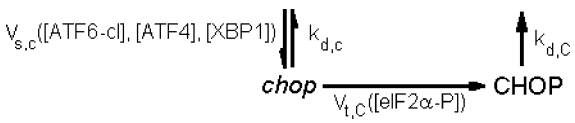
$$\frac{\mathrm{d}[\mathrm{ATF6}-\mathrm{cl}]}{\mathrm{d}t} = -k_{\mathrm{d},A6}[\mathrm{ATF6}-\mathrm{cl}] + V_{\mathrm{clev},A6}[\mathrm{BiP}-\mathrm{ATF6}]. \tag{30.6}$$

30.2.4 Overlapping of the UPR Signaling Pathways and Feedback Loops

The phosphorylation of eIF2α inhibits the translation of most mRNAs in the cell ribosome and so reduces the load of newly synthesized proteins. In particular, it represses the translation of *chop* mRNA into CHOP protein. On the other hand, the transcription factor ATF4 stimulates the transcription of *chop* mRNA. Likewise ATF6 in its cleaved form and XBP1 are thought to contribute to *chop* mRNA expression leading to a cross-talk between all three UPR signaling pathways. We incorporate this information in our model for the UPR by defining the concentration of *chop* mRNA and of CHOP protein ([*chop*] and [CHOP]), their rates of degradation as constants $k_{\mathrm{d,c}}$ and $k_{\mathrm{d,C}}$, and the rate of transcription of *chop* mRNA and the rate of translation of CHOP as $V_{\mathrm{s,c}} = V_{\mathrm{s,c}}$ ([ATF6-cl], [ATF4], [XBP1]) and $V_{\mathrm{t,C}} = V_{\mathrm{t,C}}$ ([eIF2α-p]). For example, we may assume $V_{\mathrm{s,c}}$ to be linear with respect to each of its variables $V_{\mathrm{s,c}} = k_{\mathrm{s,c}}(\varepsilon_{A6}[\mathrm{ATF6}-\mathrm{cl}] + \varepsilon_{A4}[\mathrm{ATF4}] + \varepsilon_X[\mathrm{XBP1}])$; in order to account for the inhibitory effect of its argument, we may assume that $V_{\mathrm{t,C}}$ decreases with [eIF2α-p]. Then the equations for CHOP are the following

$$\frac{\mathrm{d}[\mathrm{chop}]}{\mathrm{d}t} = -(k_{\mathrm{d,c}} + V_{\mathrm{t,C}})[\mathrm{chop}] + V_{\mathrm{s,c}},$$

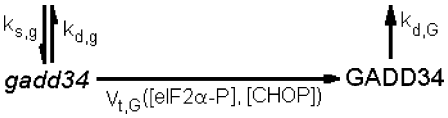
$$\frac{\mathrm{d}[\mathrm{CHOP}]}{\mathrm{d}t} = -k_{\mathrm{d,C}}[\mathrm{CHOP}] + V_{\mathrm{t,C}}[\mathrm{chop}]. \tag{30.7}$$



The protein GADD34 is another key component of the UPR under ER stress. The translation of GADD34 is stimulated by CHOP but it is also subject to the general translational repression caused by the phosphorylation of eIF2α. Based on experimental observations that show sigmoid-like changes in GADD34 associated to linear changes in CHOP [6] we hypothesize a nonlinear relationship between the amount of CHOP and of GADD34 in the ER under stress. GADD34 also influences the dephosphorylation of eIF2α making their reciprocal interaction considerably complex and leading to an important (negative) feedback loop in the UPR signaling network. The protein GADD34 and its mRNA *gadd34* are modeled by their concentrations [GADD34] and [*gadd34*] that satisfy the differential (8). We consider that the rate of transcription and degradation of *gadd34* mRNA are constants $k_{s,g}$ and $k_{d,g}$ and the rate of degradation of protein GADD34 is $k_{d,G}$; the rate of translation is, however, defined as a function of both [eIF2α-p] and [CHOP], $V_{t,G} = V_{t,G}$ ([eIF2 α-p], [CHOP]).

$$\frac{\mathrm{d}[\mathrm{gadd34}]}{\mathrm{d}t} = k_{s,g} - (k_{d,g} + V_{t,G})[\mathrm{gadd34}] \ ,$$

$$\frac{\mathrm{d}[\mathrm{GADD34}]}{\mathrm{d}t} = -k_{d,G}[\mathrm{GADD34}] + V_{t,G}[\mathrm{gadd34}] \ . \tag{30.8}$$



The ER chaperone BiP is nevertheless the most important player of the UPR. It is involved in the activation of all three UPR subnetworks due to the complexes it forms with PERK, ATF6, and IRE1, and it is itself affected by the UPR signaling cascades. In the absence of stress, BiP exists in complex form in association with PERK, ATF6, and IRE1. The accumulation of unfolded proteins causes BiP to dissociate from these three ER-stress sensors leading not only to their activation, but also freeing BiP to convert the unfolded proteins into folded proteins. The cleavage of ATF6 regulates the expression of *bip* mRNA; it is also hypothesized that XBP1 contributes to *bip* mRNA expression, and that PERK and eIF2α are necessary for upregulation of *bip* mRNA as well (although the mechanism is not clear). Moreover, the translation of protein BiP from *bip* mRNA depends on the phosphorylation of eIF2α.

We summarize all these observations in a set of differential equations for *bip* mRNA and BiP protein of, say, concentrations [*bip*] and [BiP]. The rates of degradation for BiP mRNA and protein are $k_{d,b}$ and $k_{d,B}$, and the rate of translation depends on the phosphorylation of eIF2α, $V_{t,B} = V_{t,B}$ ([eIF2α-p]). The rate of transcription of *bip* mRNA is defined as $V_{s,b} = V_{s,b}$([ATF6-cl],[XBP1],[eIF2α-p]), for example $V_{s,b} = k_{s,b}$ ($\tilde{\varepsilon}_{A6}$ [ATF6-cl]+ $\tilde{\varepsilon}_{X}$ [XBP1]+ $\tilde{\varepsilon}_{E}$ [eIF2α-p]).

Let us note that as a chaperone BiP assists in the protein-folding process in the ER. Assuming that unfolded proteins go through an association with a chaperone

BiP in order to be folded, we define $k_{\text{ass,U}}$ the rate of association of [BiP] and unfolded proteins [U] into a complex [BiP-U], and $k_{\text{f,U}}$ the rate of protein folding by the chaperone BiP. Note that the model does not track the proteins in their folded form (F). Then the equations for BiP become

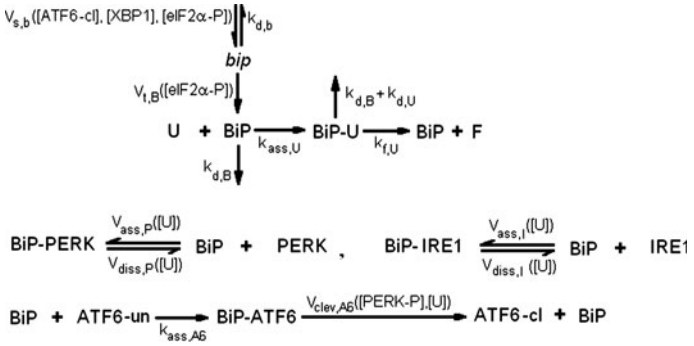
$$\frac{\mathrm{d}[\text{bip}]}{\mathrm{d}t} = -(k_{\text{d,b}} + V_{\text{t,B}})[\text{bip}] + V_{\text{s,b}} \ ,$$

$$\frac{\mathrm{d}[\text{BiP}]}{\mathrm{d}t} = -k_{\text{d,B}}[\text{BiP}] + V_{\text{t,B}}[\text{bip}] - k_{\text{ass,U}}[\text{BiP}][\text{U}] + k_{\text{f,U}}[\text{BiP}-\text{U}],$$

$$- V_{\text{ass,P}}[\text{BiP}][\text{PERK}] + V_{\text{diss,P}}[\text{BiP}-\text{PERK}] - V_{\text{ass,I}}[\text{BiP}][\text{IRE1}],$$

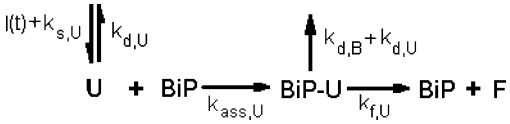
$$+ V_{\text{diss,I}}[\text{BiP}-\text{IRE1}] - k_{\text{ass,A6}}[\text{BiP}][\text{ATF6}-\text{un}] + V_{\text{clev,A6}}[\text{BiP}-\text{ATF6}],$$

$$\frac{\mathrm{d}[\text{BiP}-\text{U}]}{\mathrm{d}t} = -(k_{\text{d,B}} + k_{\text{d,U}})[\text{BiP}-\text{U}] + k_{\text{ass}}[\text{BiP}][\text{U}] - k_{\text{f,U}}[\text{BiP}-\text{U}] \ . \quad (30.9)$$



The model is complete once we add the equation for the amount [U] of unfolded proteins. We define constant rates for the synthesis and degradation of the unfolded proteins in the absence of stress ($k_{\text{s,U}}$ and $k_{\text{d,U}}$) but note that under ER stress the synthesis rate of [U] is certainly significantly modified. We associate a time-dependent function $I(t)$ to the generically termed *ER stress*, and consider $I(t)$ to be the external input to the UPR network. Obviously, to account for both ER stress and no ER stress situations the synthesis rate of the unfolded proteins must be adjusted to $k_{\text{s,U}} + I(t)$. Then [U] satisfies

$$\frac{\mathrm{d}[\text{U}]}{\mathrm{d}t} = I(t) + k_{\text{s,U}} - k_{\text{d,U}}[\text{U}] - k_{\text{ass,U}}[\text{BiP}][\text{U}] \ . \quad (30.10)$$



30.3 Discussion and Further Directions

We constructed a biologically plausible computational model ((30.1) (30.10)) for the UPR in the mammalian secretory cells under ER stress. The model incorporates up-to-date knowledge regarding the signaling cascades of PERK, ATF6, and IRE1 and it accounts for the cross-talk and feedback loops in the network due to other UPR key components such as eIF2α, XBP1, ATF4, CHOP, GADD34, and BiP. We have only focused on the description of the model and did not discuss in this paper issues such as model calibration or model validation. We agree that the complexity of the UPR signaling pathways poses particular challenges to the validation of the model but we also note that recent advances in experimental techniques indicate the possibility of generating an accurate *quantitative* experimental description of the UPR network. Therefore, it seems reasonable to address the problem of the UPR through computational modeling. A model sufficiently general to allow for simulation of the UPR internal components, for flexibility in the choice of strength and type of the stressor, and that does not impose constraints on possible outcomes becomes a necessity; the system of differential (30.1) to (30.10) is built having precisely these goals in mind.

A first attempt to compare experimental data for the UPR in mammalian cells to a quantitative model was made in [9]. The model is very simplistic; however, the idea to combine computational modeling with quantitative cell biology data for the UPR is extremely valuable. This is a very good example that quantitative information can be extracted even for complex eukaryotic systems such as the UPR in mammalian cells. For example, degradation rate constants were derived from measurements of protein and mRNA half-lives, while production rate constants were derived from measurements of steady-state levels of proteins and mRNAs for CHOP, GADD34, and BiP (see Supplemental material in [9]). To conclude, computational modeling as we propose in this paper can be used in connection with quantitative experimental data for the UPR under ER stress, and may soon prove its utility in validating and/or predicting hypotheses about the functionality of the UPR in complex cells.

Acknowledgments We thank Tom Rutkowski for helpful discussions on the problem of unfolded protein response, and for his feedback and critiques during the development of the model.

References

1. Chen K, Calzone L, Csikasz Nagy A, et al (2004) Integrative analysis of cell cycle control in budding yeast. Molecular Biology of the Cell 15:3841 3862
2. Guido N, Wang X, Adalsteinsson D, McMillen D, et al (2006) A bottom up approach to gene regulation. Nature 439:856 860
3. Kim H, Shay T, O'Shea E, Regev A (2009) Transcriptional regulatory circuits: Predicting numbers from alphabets. Science 325:429 432

4. Marciniak S, Yun C, Oyadomari S, et al (2004) CHOP induces death by promoting protein synthesis and oxidation in the stressed endoplasmic reticulum. Genes and Development 18:3066 3077
5. Ron D, Walter P (2007) Signal integration in the endoplasmic reticulum unfolded protein response. Nature Reviews. Molecular Cell Biology 8:519 529
6. Rutkowski T (2009) Personal communication
7. Rutkowski T, Kaufman R (2004) A trip to the ER: coping with stress. Trends in Cell Biology 14(1):20 28
8. Rutkowski T, Kaufman R (2007) That which does not kill me makes me stronger: Adapting to chronic ER stress. Trends in Biochemical Sciences 32(10):469 476
9. Rutkowski T, Arnold S, Miller C, et al (2006) Adaptation to ER stress is mediated by differential stabilities of pro survival and pro apoptotic mRNAs and proteins. PLOS Biology 4(11):2024 2041
10. Shen X, Zhang K, Kaufman R (2004) The unfolded protein response A stress signaling pathway of the endoplasmic reticulum. Journal of Chemical Neuroanatomy 28:79 92
11. Trusina A, Papa F, Tang C (2008) Rationalizing translation attenuation in the network architecture of the unfolded protein response. Proceedings of the National Academy of Sciences of the United States of America 105(51):20280 20285

Chapter 31
Random-Walk Mechanism in the Genetic Recombination

Youhei Fujitani, Junji Kawai, and Ichizo Kobayashi

Abstract We have explained some experimental data of the homologous recombination and the genetic interference in terms of one-dimensional random walk over discrete sites. We first review our previous results. Next, we modify our random-walk model for the homologous recombination into a continuous-site model, and discuss a possible explanation for the previous experimental data obtained by means of the plasmid having one-side homology. Finally, we show that a reaction between an intermediate and a product is indispensable in explaining the genetic interference in terms of our reaction-diffusion model.

Keywords Genetic interference · Homologous recombination · Homology · Plasmid · Random walk model · Reaction-diffusion model

31.1 Introduction and a Brief Review

Many enzymes have been revealed to be involved in the genetic recombination mainly by experiments for the homologous recombination of *Escherichia coli* [15]. Enzymes similar to some of them are identified in eukaryotes, suggesting that the similar mechanism should work in meiotic crossing-over. Though the interaction between reactants is complicated [13, 14], it appears crucial to properties of some phenomena that a key intermediate structure moves with thermal fluctuation along the one-dimensional (1D) space given by the DNA molecule. As reviewed below, we have succeeded in explaining some data in terms of 1D random walk over discrete sites.

Y. Fujitani (✉)
Keio University, 223 8522 Yokohama, Japan
e mail: youhei@appi.keio.ac.jp

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 31,

Although the frequency of homologous recombination was thought to be linear with respect to the homology length, the logarithmic replot of data shows the cubic dependence (Fig. 31.1a). Fujitani et al. explained it by regarding the branch migration as random walk over a 1D lattice [11]. Let N denote the homology length measured by the base-pair. We assume the random walk to be symmetric, i.e. forward and backward transitions to share the same rate, g, and its step size to be the base base interval. A branch point is assumed to be produced with a small enough probability α per site only at the initial time, and then to walk randomly until its annihilation. After the annihilation, its resolution to a recombinant follows, or otherwise the reaction is aborted. We define h and k so that the product gh gives the transition rate of the annihilation and that ghk gives that of the resolution. The reaction is also aborted when the random-walker reaches either end of the homology, i.e. the boundary is totally absorbing [20]. The probability that the branch point is located at a site j at time t, denoted by $p_j(t)$, follows the master equation,

$$\frac{\mathrm{d}p_j}{\mathrm{d}t} = gp_{j+1}(t) + gp_{j\ \ 1}(t) - g(2+h)p_j(t) \quad \text{for} \quad 1 \le j \le N, \tag{31.1}$$

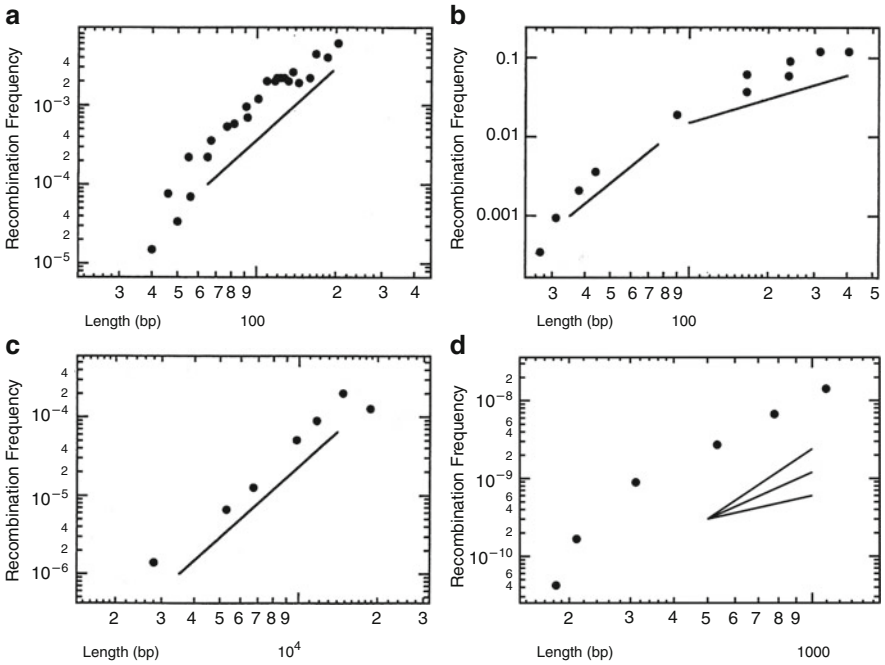


Fig. 31.1 Logarithmic plots of experimental data. (**a**) Bacteriophage $T4 \times T4$ in wild type *Escherichia coli*. [19]. (**b**) $\lambda \times$ plasmid in wild type *E. coli*. (AB1157) [18]. (**c**) Gene targeting with isogenic DNA in mouse cells [2]. (**d**) Replot of Fig. 4 of Ref. [3], discussed in Sect. 31.2. *Lines*, provided for reference, have a slope of three in **a** and **c**, slopes of three and unity in **b**, and slopes of three, two, and unity in **d**

together with $p_j(0) = \alpha$ and $p_0(t) = p_{N+1}(t) = 0$. The recombination frequency after long enough time would be measured experimentally, and can be obtained as

$$ghk \int_0^\infty \mathrm{d}t \sum_{j=1}^{N} p_j(t) = k\alpha\{N + 1 - \tanh\phi(N + 1)\coth\phi\} \tag{31.2}$$

$$\approx \begin{cases} hk\alpha N^3/12 & \text{for} \quad N \ll 2/\sqrt{h} \\ k\alpha\left(N - 2/\sqrt{h}\right) & \text{for} \quad N \gg 2/\sqrt{h} \end{cases}, \tag{31.3}$$

where ϕ is defined so that $h = 4\sinh^2\phi$. We usually have $h \ll 1$ to obtain (31.3). Our random-walk (RW) model thus predicts that the dependence shifts from the cubic one to the linear one as the homology length increases to pass across $\sim 2/\sqrt{h}$.

We can explain the cubic dependence in Fig. 31.1a by assuming that the range of smaller length was examined in this experiment. The cubic dependence itself, yielded by a size effect, does not depend on any fitting parameter. It comes out if the homology length is small enough to be felt by a branch point during its random walk. The map expansion phenomenon and the very rapid drop-off phenomenon [1, 21] can be explained by this dependence [7, 8]. If the homology length is larger, the linear dependence comes out from the probability of the initial production of a branch point in the homologous region. Another dataset in Fig. 31.1b may show the shift. We also found the cubic dependence rather robust to changes in the boundary condition [6] and to asymmetry of the random walk [9]. Thus, the RW model can explain the cubic dependence observed in vivo, where the branch migration would be asymmetric [15] and the boundary would not be totally absorbing.

In the gene-targeting, a vector is constructed so that it has a DNA region homologous to a part of the recipient genome [17]. The region is separated into two subregions by an intervening marker-gene region. A dataset for the mammalian gene-targeting system [2] can be also read as the cubic dependence (Fig. 31.1c). However, to understand this dataset totally, we should modify the RW model, where the random walk is assumed to be over a stretch of homology. We will return to this point later.

We can find another kind of random-walker in a much larger length-scale phenomenon the meiotic crossing-over between a pair of homologous chromosomes. One crossover point appears to interfere with occurrence of another in the neighborhood. This genetic interference was clearly observed in Drosophila and Neurospora, between which the plots of the coincidence (normalized density correlation of crossover points) against the genetic distance (distance measured in terms of expected number of crossover points) are very similar. Recombination between pairs of homologous regions scattered along the chromosomes is thought to lead to a crossing-over. It was suggested that some premeiotic contact points between intact duplexes prime the homologous recombination [22]. We proposed a reaction-diffusion (RD) model by assuming that contact points randomly walk to search for the homologous region and that they can be immobilized to mature into

crossover points [10]. Random-walkers are annihilated pairwise whenever they collide with each other, and any of them is annihilated whenever it encounters a crossover point. These reactions are symbolically written as $A \to B$, $A + A \to \emptyset$, and $A + B \to B$, respectively, where A and B denote a contact point and a crossover point, respectively. We impose the periodic boundary condition for simplicity. The coincidence is given by the final $B - B$ density correlation divided by the final squared B-particle density.

Our RD model is based not on the genetic distance but on the physical distance measured by the step of the contact point, specifying what happens physically, unlike the genetic model [5]. Our model has two parameters the initial density of the random-walkers and the rate of its processing into a crossover point. It was numerically shown that, as the former increases and/or the latter decreases, plotted curves of the coincidence against the genetic distance converge on a unique curve, which can explain the similarity without fine adjustment of parameter values [10].

31.2 Continuous-Site Model for the Homologous Recombination

We rewrite the RW model so that the branch point moves over continuous sites to describe a random-walker on which some force is exerted [20]. Let δx be the site site interval in the discrete-site model, and we introduce $x = j\delta x$, $P(x,t) \equiv p_j/\delta x$, $D \equiv g(\delta x)^2$, $H \equiv h/(\delta x)^2$, $A = \alpha/\delta x$, and $L = N\delta x$. A potential V is defined so that $-V'$ gives the force, where the prime $'$ indicates the derivative. The time-evolution equation of $P(x,t)$ turns out to be the Fokker Planck equation with a damping term,

$$\frac{\partial}{\partial t}P(x,t) = \frac{\partial}{\partial x}(V'(x)P(x,t)) + D\left(\frac{\partial^2}{\partial x^2} - H\right)P(x,t) \quad \text{for} \quad 0 \le x \le L. \tag{31.4}$$

If we put $V \equiv 0$ in the above, the resultant equation corresponds with (31.1). The initial condition and the totally absorbing boundary condition are, respectively, given by $P(x,0) = A$ and $P(0,t) = P(L,t) = 0$. The recombination frequency is a function of L here, and is denoted by $R(L)$. As in (31.2), we have $R(L) = DHk \int_0^\infty \mathrm{d}t \int_0^L \mathrm{d}x\, P(x,t)$.

As described below, we can calculate the recombination frequency if we put

$$V(x) = -2D\,\mathrm{lncos}\{\pi\gamma(x - L/2)/L\} \quad \text{with} \quad 0 \le \gamma < 1. \tag{31.5}$$

This function is convex, and symmetric with respect to $x = L/2$. See the inset of Fig. 31.2. Thus, the branch point moves as if it were connected to an end of a spring with small enough original length, of which the other end is fixed at the center of the

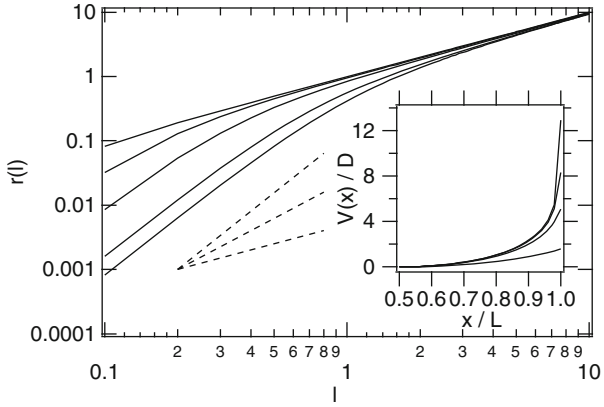


Fig. 31.2 *Curves* represent $r(l)$ with γ $0.999, 0.99, 0.95, 0.7$ and 0 from the top, respectively. *Dashed lines* with slopes of three, two, and unity are provided for reference. We plot $V(x)/D$ against x/L in the *inset*, where *curves* are obtained with γ $0.999, 0.99, 0.95$, and 0.7 from the top, respectively

region. As the parameter γ is smaller, the function is more flat, i.e., the spring is weaker. When $\gamma = 0$, the random-walker is free as in the original RW model.

Let us introduce $\tau = Dt$ and $Q(x,\tau) \equiv P(x,t)\exp\{DHt + (V(x)/2D)\}$. Defining the Fourier transform $Q_n(\tau)$ as the integral of $Q(x,\tau)(2/L)\sin(n\pi x/L)$ over $0 \le x \le L$, we can use (31.4) to relate $Q_n(\tau)$ with $Q_n(0)$, which can be calculated from the initial condition. Introducing $l \equiv L\sqrt{H}/\pi$ and $r(l) \equiv R(L)\sqrt{H}/(\pi kA)$, we can find

$$r(l) = l - \frac{2l\cos(\gamma\pi/2)}{\pi\cosh(\Delta_l\pi/2)}\int_0^{\pi/2} \mathrm{d}y \frac{\cosh\Delta_l y}{\cos\gamma y}, \tag{31.6}$$

where $\Delta_l \equiv \sqrt{l^2 - \gamma^2}$ with its imaginary part being negative for $l<\gamma$. We used (1.445.5) of Ref. [12] with $m = 0$ in deriving (31.6). Putting $\gamma = 0$ in (31.6) yields $r(l) = l - (2/\pi)\tanh(\pi l/2)$, which recovers (31.3), and is represented by the curve in the bottom of Fig. 31.2. Introducing $z \equiv e^{\pm 2iy}$, and the line segments $C_\pm$ running from $z = 0$ to $e^{\pm i\gamma\pi}$ in the z-plane, we find the integral in (31.6) given by $I_+ - I$, where

$$\begin{aligned} I_\pm &\equiv \int_{C_\pm} \frac{\mathrm{d}z}{2i\gamma} \frac{z^{\beta}}{1+z} \\ &= \frac{e^{\mp i\gamma\beta\pi}F(1,1,2-\beta;1/(1+e^{\mp i\gamma\pi}))}{2i\gamma(1-\beta)(1+e^{\mp i\gamma\pi})}. \end{aligned} \tag{31.7}$$

Here, $\beta \equiv (1 - i\Delta/\gamma)/2$ and F is Gauss' hypergeometric function. As shown in Fig. 31.2, dependence of $r(l)$ on l in the smaller-l range changes from the cubic one to the linear one as γ is larger. Then, the random-walker more hardly reaches an end of the homology, i.e. more hardly feels the homology length.

de Vries and Wackernagel studied transformation of *Acinetobacter* sp. by means of plasmid vectors [3]. In Fig. 4 of [3], where a plasmid has a homologous region on one side of the marker-gene region and a nonhomologous region on the other side, the recombination frequency appears to have the square dependence on the homology length above ~ 300 bp, as shown in our Fig. 31.1d. The branch point in the homologous region would be subject to an external force because the plasmid should have some elasticity and a region of the plasmid is stuck to the nonhomologous region. We can model this recombination by means of (31.5) if the spring is weaker as the homologous region, and thus the whole plasmid, are longer. In Fig. 31.2, $r(l)$ with $\gamma = 0.99$ shows l^2-dependence in the smaller-l range, which could explain the square dependence observed in the one-side homology system.

In the usual gene-targeting system, the two homologous regions are on both sides of the marker-gene region, and a randomly walking branch point is produced in each of the regions. We could formulate this problem of two random-walkers by using the two-variable Fokker Planck equation [20].

31.3 A Model without $A + B \to B$ for the Interference

Eliminating the reaction $A + B \to B$ from our RD model, we can proceed with calculations in another way, of which results are described below. Let N be the number of lattice-sites, and the particle distribution can be labeled by $\{m\} \equiv \{m_1, m_2, \ldots, m_N\}$ and $\{n\} \equiv \{n_1, n_2, \ldots, n_N\}$, where m_j and n_j are the numbers of A- and B-particles at a site j, respectively. Defining the time τ so that the transition rate of the random walk is unity, we write $P_{\{m\};\{n\}}(\tau)$ for the probability of the particle distribution at time τ, and write $\langle \cdots \rangle_\tau$ for the average over the particle distribution. The coincidence is given by $s_l \equiv \langle n_j n_{j+l} \rangle_\infty / \left(\langle n_j \rangle_\infty \right)^2$ for $l>0$, where l is a physical distance. The genetic distance is given by $l \langle n_j \rangle_\infty / 2$. Initially, no B-particle exists, while we assume an independent Poisson distribution for A-particles. We thus have $P_{\{m\};\{n\}}(0) = \prod_{j=1}^{N} a^{m_j} e^{\ a}\delta_{0n_j}/(m_j!)$, where a is the average of the initial A-particle number at each site. Also assuming the reaction $A \to \emptyset$, we define H as the total rate of $A \to \emptyset$ and $A \to B$, and define κ ($0<\kappa\leqslant 1$) as the ratio of the rate of $A \to B$ to the total rate. We introduce the transition rate of the reaction $A + A \to \emptyset$, which is denoted by λ. In the large-λ limit, the pair annihilation always occurs at collision of A-particles, as in the RD model. The master equation is given by

$$\frac{\mathrm{d}P}{\mathrm{d}\tau} = \sum_{j=1}^{N} \Big[\sum_{e(j)} \{(m_e + 1) P_{m_j\ \ 1, m_e+1} - m_j P\} + H\{\kappa(m_j + 1) P_{m_j+1; n_j\ \ 1} + (1 - \kappa) \times (m_j + 1) P_{m_j+1} - m_j P\} + \lambda\{(m_j + 1)(m_j + 2) P_{m_j+2} - m_j(m_j - 1) P\}\Big], \tag{31.8}$$

where, for brevity, the variable τ is dropped, and a variable in the subscript is written only if different from $(\{m\},\{n\})$, e.g. P means $P_{\{m\};\{n\}}(\tau)$ while P_{m_j+1} means $P_{\ldots,m_j\ 1,m_j+1,m_{j+1},\ldots;\{n\}}(\tau)$. In the first term, $\sum_{e(j)}$ means the sum over the nearest-neighbor sites e of the site j. We stipulate that $P_{\{m\};\{n\}}(\tau)=0$ if any of m_js and n_js is negative. The three terms in the braces describe diffusion of A-particles, $A\to B$ and $A\to\emptyset$, and $A+A\to\emptyset$, respectively.

Applying the path-integral method [4, 16], we can calculate the B-particle density $\langle n_j\rangle_\tau$ and its connected correlation $\langle n_j n_{j+l}\rangle_\tau - \langle n_j\rangle_\tau^2$ by means of the diagrammatic perturbation expansion with respect to λ. After some algebra, we obtain

$$\langle n_j\rangle_\infty = \frac{\kappa H}{2\lambda}\ln\left(1+\frac{2a\lambda}{H}\right)+\lambda^2\frac{\kappa a^2}{\pi H\sqrt{H}}\arctan\frac{\pi}{\sqrt{H}} \tag{31.9}$$

up to the order of λ^2, and

$$s_l = 1-\frac{\lambda}{4\sqrt{H}}(1+l\sqrt{H})e^{\ l\sqrt{H}} \tag{31.10}$$

up to the order of λ. These results agree with the simulation results as far as $\lambda\ll 1$ (Fig. 31.3a). Simulation results in Fig. 31.3b suggest that, even in the large-λ limit, s_1 cannot be smaller than ~ 0.5 in this model, unlike in the experimental data and the RD model. The reaction $A+B\to B$ would be indispensable to the genetic interference.

Part of the work by YF was supported by Keio Gakuji Shinko Shikin, while part of the work by IK by "Grants-in-Aid for Scientific Research" from JSPS (2137001,

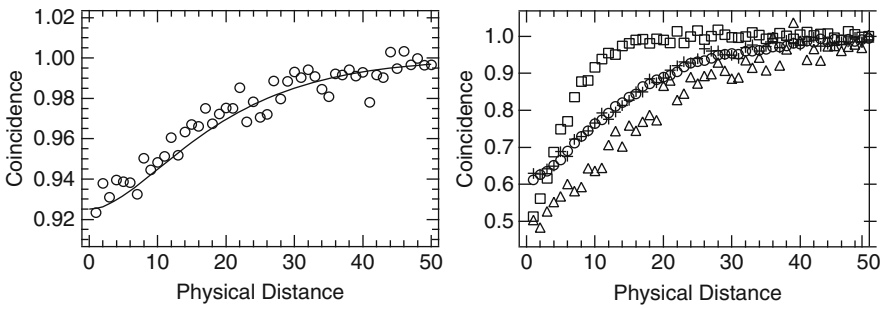


Fig. 31.3 (**a**) Plots of s_l against $l(>0)$ for (a,H,λ,κ) $(0.005,0.01,0.03,1.0)$. The *curve* is calculated with (31.10). Numerical study of (31.8) with N 10^5 gives *circles*, each of which is calculated over 10^4 samples, and yields $\langle n_j\rangle_\infty$ 4.936×10^{-3}, which agrees with the value of (31.9), 4.937×10^{-3}. (**b**) We calculate s_l numerically with N 10^4, and plot it against l. We prohibit simultaneous presence of more than one particle at a site, which corresponds with taking the large λ limit. The parameter values are (a,H,κ) $(0.005,0.1,1.0)$ for *squares*, $(0.05,0.01,1.0)$ for *circles*, $(0.05,0.01,0.3)$ for *crosses*, and $(0.005,0.01,1.0)$ for *triangles*. A data point is calculated over 10^4 samples except 5×10^4 samples for a *square*. From Junji Kawai, Master thesis, Keio University (1999)

19657002) and the Global COE program “Deciphering Biosphere from Genome Big Bang” to IK.

References

1. Datta A, Hendrix M, Lipsitch M et al (1997) Dual roles for DNA sequence identity and the mismatch repair system in the regulation of mitotic crossing over in yeast. Proc Natl Acad Sci USA 94: 9757 9762.
2. Deng C, Capecchi MR (1992) Reexamination of the gene targeting frequency as a function of the extent of homology between the targeting vector and the target locus. Mol Cell Biol 12: 3365 3371.
3. de Vries J, Wackernagel W (2002) Integration of foreign DNA during natural transformation of *Acinetobacter* sp. by homology facilitated illegitimate recombination. Proc Natl Acad Sci USA 99: 2094 2099.
4. Doi M (1976) Stochastic theory of diffusion controlled reaction. J Phys A: Math Gen 9: 1479 1495.
5. Foss E, Lande R, Stahl FW et al (1993) Chiasma interference as a function of genetic distance. Genetics 133: 681 691.
6. Fujitani Y, Kobayashi I (1995) Random walk model of homologous recombination. Phys Rev E 52: 6607 6622.
7. Fujitani Y, Kobayashi I (1997) Mismatch stimulated destruction of intermediates as an explanation for map expansion in genetic recombination. J Theoret Biol 189: 443 447.
8. Fujitani Y, Kobayashi I (1999) Effect of DNA sequence divergence on homologous recombi nation as analyzed by a random walk model. Genetics 153: 1973 1988.
9. Fujitani Y, Kobayashi I (2003) Asymmetric random walk in a reaction intermediate of homologous recombination. J Theoret Biol 220: 359 370.
10. Fujitani Y, Mori S, Kobayashi I (2002) A reaction diffusion model for interference in meiotic crossing over. Genetics 161: 365 372.
11. Fujitani Y, Yamamoto K, Kobayashi I (1995) Dependence of frequency of homologous recombination on the homology length. Genetics 140: 797 809.
12. Gradshteyn IS, Ryzhik IM (1994) Tables of integrals, series, and products. Academic Press, San Diego.
13. Kornyshev AA, Lee DJ, Leikin S et al (2007) Structure and interactions of biological helices. Rev Mod Phys 79: 944 991.
14. Leach DRF (1996) Genetic recombination. Blackwell Science, Oxford.
15. Lloyd RG, Low KB (1996) Homologous recombination. In: Neidhardt FC (ed) Escherichia coli and Salmonella. ASM Press, Washington, DC.
16. Peliti L (1985) Path integral approach to birth death process on a lattice. J Phys (Paris) 46: 1469 1483.
17. Sedivy JM, Joyner AL (1993) Gene targeting. Oxford University Press, Oxford.
18. Shen P, Huang HV (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics 112: 441 457.
19. Singer BS, Gold L, Gauss P et al. (1982) Determination of the amount of homology required for recombination in bacteriophage T4. Cell 31: 25 33.
20. van Kampen NG (1992) Stochastic processes in physics and chemistry, North Holland, Amsterdam.
21. Vulić M, Dionisio F, Taddei F et al (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc Natl Acad Sci USA 94: 9763 9767.
22. Weiner B, Kleckner N (1994) Chromosome pairing via multiple interstitial interactions before and during meiosis. Cell 77: 977 991.

Chapter 32
Critical Assessment of Side Chain Conformation Prediction in Modelling of Single Point Amino Acid Mutation

Anna Marabotti and Angelo Facchiano

Abstract We assessed the ability of three widely used and freely available programs for side chain repacking to simulate the structural effects of single point mutations. The programs tested seem to produce sufficiently reliable predictions only when mutations involve residues characterized by limited flexibility. The change in size and/or polarity of the mutant side chain can affect the performances of the different programs. The correlation between the quality of predictions, exposure to solvent, and B-factors of the corresponding residues in crystal is also investigated. This analysis may provide non-experts with insights into what types of modelling protocols work best for such a problem as well as providing information to developers for improving their algorithms.

Keywords Amino acid · Conformation prediction · Mutation · Protein structure · Quality of predictions

32.1 Introduction

A main task for protein science is to understand and predict how mutations can affect protein structure and functions. This is of great interest for protein engineering [8], or to simulate the effects of mutations related to genetic diseases [1]. Historically, various approaches have been proposed to simulate point mutations in proteins [15, 19 21]. In the last years, high-performance computer facilities have prompted people to use also molecular dynamics simulations [4, 5, 9]. However, these simulations are time consuming and difficult to apply to large systems. Recently, a fully automated protocol named "Mutate model" especially focused

A. Marabotti (✉)
Institute of Biomedical Technologies, National Research Council, Via F.lli Cervi 92, 20090 Segrate (MI)
e mail: anna.marabotti@itb.cnr.it

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 32,

at modelling point mutations [6] was developed in the frame of the popular and widely used program MODELLER [18]. Results seem to be encouraging, although further improvement is needed especially in better defining the scoring function and in increasing the conformations sampled [6].

Often, also programs for side chain repacking (for a review, see [10]) are used to simulate point mutations. However, they are especially focused on the whole protein modelling, to refine the results of homology modelling or to add side chains to the backbone obtained e.g. with ab initio methods. Their published performances in global reconstruction of side chains during comparative modelling are quite encouraging [22], but side chains exposed to solvent still represent a major problem, due to their high conformational freedom. Instead, the real suitability of these programs for facing this specific problem is still an unexplored issue.

To address this question, we have assessed the performances of three freely available software for side chain repacking in predicting the correct conformation of side chains after introducing a point mutation in a protein structure. The three programs (SCWRL [3], SCAP [23], and NCN [16]) are among the most popular and the best performing in this field [10] and they are representative of the different approaches that can be applied to solve this problem.

32.2 Methods

We report here a very short description of methods used. A full description can be found at: http://bioinformatica.isa.cnr.it/side-chain-replacement/.

The structures of the wild-type phage T4 lysozyme [11] and of the "pseudo-wt" lysozyme carrying the mutations C54T and C97A [13], obtained by X-ray crystallography and deposited in the PDB database [2] (PDB codes: 3LZM and 1L63, respectively) were chosen as starting points to model the mutants included in the benchmark. The final benchmark includes 107 structures of mutants (resolution between 1.65 and 2.60 Å) carrying only a single mutation (for "pseudo-wt," in addition to the two cited mutations), without ligands, and excluding mutation to proline.

SCWRL3.0 [2] and SCAP [23] were freely downloaded from the related Web servers, whereas NCN [16] was obtained by direct request to the authors. Single point mutants were modelled on the crystallographic structure of wild-type or "pseudo-wt" lysozymes by using facilities of the repacking program (for SCWRL3.0 and SCAP), or by editing the PDB file (for NCN), as indicated in the software manual. Then, we allowed the programs to rebuild all the side chains of each mutant lysozyme. By comparison to the mutant structures obtained by X-ray crystallography, the accuracy of the side chain conformer predictions was assessed in terms of dihedral angle deviation, calculated with the program CHI, implemented in the JACKAL package [23], and RMSD value, calculated using the McLachlan algorithm [12] as implemented in the program ProFit v. 2.3.5.1, developed by Dr. A.C.R. Martin (see http://www.bioinf.org.uk/software/profit).

The residues with at least one atom included in a distance of 5 Å from the mutant side chain were identified as "neighbours" of a mutation. The solvent exposure of the residues was calculated with the program NACCESS [7]. B-factors were extracted from the PDB files of the reference structures. We defined as "conserved in size" all the residues with no more than 10% of difference in their volume, calculated according to Zamyatin [25]. We classified Gly, Ala, Val, Leu, Ile, Met, Phe, Trp as "non-polar amino acids" Asn, Cys, Gln, Ser, Thr, Tyr as "polar amino acids" and Arg, Lys, Glu, Asp, His as "charged amino acids".

32.3 Results

Table 32.1 shows the results for the single side chain mutated, in terms of dihedral angle correctness and residue RMSD. SCWRL3.0 gives the best results. Table 32.2 reports the results concerning the "neighbours" of the mutations. In this case, NCN performs better. Table 32.3 shows that all predictors perform better when a large residue is replaced by a smaller one, rather than when the opposite case occurs. NCN and SCWRL3.0 perform similarly for angle accuracy, whereas SCAP performances evaluated on RMSD are better than those of NCN in the case of the "large to small" mutation. Looking at data ordered for changes in polarity between the wild type and mutant residue shown in Table 32.3, we noted that NCN is generally the best performing in terms of χ_1 and χ_{1+2} accuracy for the most abundant group (unchanged polarity). On the contrary, mutations involving changes in polarity are better predicted by SCWRL3.0 in terms of dihedral accuracy, and by NCN in terms of RMSD. The best performances for SCAP and NCN are obtained when the original residue is replaced by one with opposite charge, whereas in this case SCWRL3.0 performs worse than in the other cases.

Fig. 32.1a shows the correlation of the residue RMSD with the percentage of relative solvent accessible surface area (SASA) calculated on the side chains

Table 32.1 Prediction accuracy for mutated side chain only

Parameter	NCN	SCAP	SCWRL
Dihedral accuracy (%)	57.0[a]/53.5[b]	49.1[a]/43.5[b]	64.9[a]/60.0[b]
Residue RMSD (Å)	1.33[c]/1.74[d]	1.34[c]/1.80[d]	1.20[c]/1.57[d]

[a]Result for χ_1 dihedral angle
[b]Result for χ_{1+2} dihedral angle
[c]Result when Cβ atom was included in calculations
[d]Result when Cβ atom was not included in calculations

Table 32.2 Prediction accuracy for residues surrounding the mutation

Parameter	NCN	SCAP	SCWRL
Dihedral accuracy (%)	78.2[a]/73.6[b]	73.4[a]/69.4[b]	76.7[a]/72.7[b]
Residue RMSD (Å)	1.01[c]/1.21[d]	1.00[c]/1.23[d]	1.08[c]/1.30[d]

Notes as for Table 32.1

Table 32.3 Prediction accuracy for physico chemical features of the mutation in the residues surrounding the mutation

Type of mutation	No. of mutants	Dihedral accuracy (%)			Residue RMSD (Å)		
		NCN	SCAP	SCWRL	NCN	SCAP	SCWRL
Size unchanged	69	80.7[a]/ 761[b]	76.4[a]/ 72.6[b]	78.4[a]/ 74.4[b]	0.91[c]/ 1.11[d]	0.95[c]/ 1.16[d]	1.04[c]/ 1.25[d]
Large to small	39	77.0[a]/ 72.9[b]	72.8[a]/ 68.1[b]	75.3[a]/ 72.7[b]	1.04[c]/ 1.26[d]	0.98[c]/ 1.21[d]	1.13[c]/ 1.37[d]
Small to large	42	75.2[a]/ 70.0[b]	69.1[a]/ 65.5[b]	75.4[a]/ 69.8[b]	1.13[c]/ 1.35[d]	1.12[c]/ 1.35[d]	1.09[c]/ 1.31[d]
Unchanged polarity	67	77.7[a]/ 73.7[b]	75.1[a]/ 69.9[b]	76.7[a]/ 72.2[b]	1.02[c]/ 1.22[d]	1.04[c]/ 1.26[d]	1.12[c]/ 1.35[d]
Decreased polarity	39	79.0[a]/ 73.4[b]	71.9[a]/ 68.8[b]	75.5[a]/ 72.6[b]	0.98[c]/ 1.21[d]	0.99[c]/ 1.23[d]	1.04[c]/ 1.28[d]
Increased polarity	38	77.9[a]/ 73.4[b]	72.0[a]/ 69.6[b]	78.5[a]/ 74.1[b]	0.97[c]/ 1.18[d]	0.95[c]/ 1.17[d]	0.99[c]/ 1.19[d]
Charge inversion	6	80.1[a]/ 73.6[b]	73.7[a]/ 67.1[b]	73.7[a]/ 68.9[b]	1.21[c]/ 1.38[d]	1.04[c]/ 1.20[d]	1.35[c]/ 1.57[d]

Notes as for Table 32.1

included in the neighbours of mutation. For buried residues (SASA less than 10% [17]), all programs greatly improve their accuracy. Nevertheless, we did not find a continuous loss of prediction correctness when SASA values increase. Since high B-factor values can be used as indicators of flexibility and dynamics of protein structure [24], we evaluated the correlation between B-factors and performances of the side chain repacking programs (Fig. 32.1b). For all programs there is a more direct relationship between B-factor amplitude and correctness of the prediction.

32.4 Discussion

Based on our results, none of the programs tested can be considered fully reliable to predict and simulate the effect of a single point mutation. By considering the features of the three programs, the differences in their performances could be ascribed to the definition of the energy function and the conformational search, but also to the use of a larger rotamer library. In particular, NCN seems to be more sensitive than the other two programs to the effects of electrostatic fields. Its potential energy function includes not only steric terms such as van der Waals parameters, but also electrostatics and H-bonding terms [16]. Anyway, despite a very simple scoring function, SCWRL3.0 results are comparable to those of the other two repacking programs.

The reliability of repacking programs to predict the effects of single point mutations on its environment is higher when a residue is replaced by a smaller one with similar polarity. Instead, the prediction of the effects of the replacement of a charged and relatively small residue with a bigger one with different polarity features is expected to be less reliable. This could also be due to the need of a relevant backbone rearrangement, not allowed in these programs.

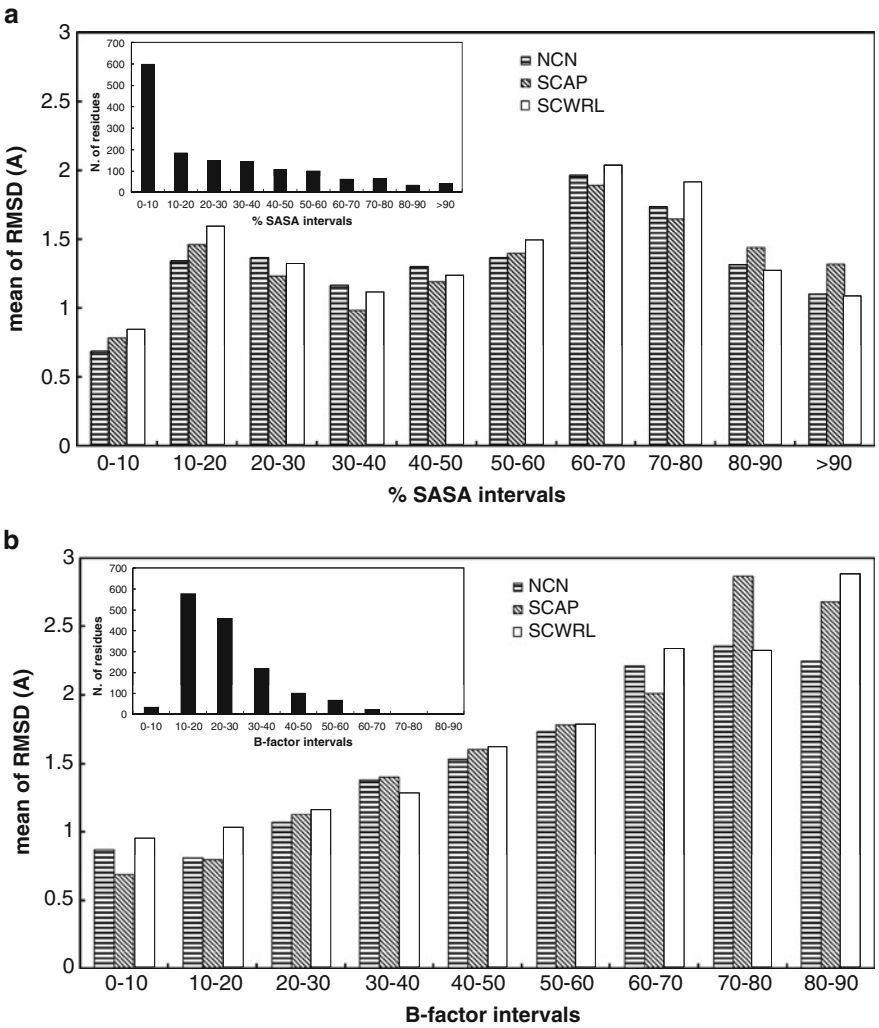


Fig. 32.1 Relationships between reliability of predictions, SASA, and B factor for residues surrounding the mutation. (**a**) The bars report, for each software, the mean of residue RMSD of the side chains belonging to this subset, partitioned into SASA intervals. The *inset* shows the number of side chains for each SASA interval. (**b**) The bars report, for each software, the mean of residue RMSD of the side chains belonging to this subset, partitioned into B factor intervals. The *inset* shows the number of side chains for each B factor interval

The performances of the programs for predicting side chain conformation are strictly related to the conformational freedom of the side chains in a protein [10]. In general, authors developing side chain repacking programs evaluate performances towards SASA. In our analysis the reliability of all predictors continuously degrades as B-factors increase, whereas this is not true for SASA, so that B-factors

seem a more reliable measure of the side chain conformational freedom. The evaluation of software performances with respect to SASA has also two other disadvantages. Firstly, SASA calculated with different software using different criteria are very difficult to compare [14]. Secondly, the concept of "buried side chains" is differently used to test the programs, since the threshold of accessible surfaces for this definition ranges between 10 and 30% [10]. On the contrary, B-factors are experimental data univocally defined for each crystallographic structure as a measure of imprecision in the protein coordinates. Our results suggest to use B-factors instead of SASA to evaluate the reliability of the programs with respect to the flexibility of side chains. In any case, people aiming at predicting the conformation of a side chain, especially when it is expected to be highly flexible, should treat with care their results. If a residue is characterized by high B-factor, also its conformation fixed in the crystallographic coordinates may be highly imprecise, and therefore it is not possible to deduce reliable structural information of the effects of the mutation of that residue.

In conclusion, side chain repacking programs should be used carefully to predict the effects of single point mutations. Results should be always treated with care and interpretation of the effects of mutations should take into account also the quality of the starting structure.

Acknowledgements A preliminary version of this work was presented at the Annual Meeting of the Bioinformatics Italian Society (BITS), Naples, 26 28/4/2007. This work was partially supported by "Italia USA Farmacogenomica Oncologica" Conv. no. 527/A/3A/5 and "Progetto CNR Bioinformatics."

References

1. Antonarakis SE, Krawczak M, Cooper DN (2008) The nature and mechanisms of human gene mutation. In Valle D et al (eds) The online metabolic and molecular bases of inherited disease, McGraw Hill, Columbus, USA
2. Berman H, Henrick K, Nakamura H et al (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35:D301 D303
3. Canutescu A, Shelenkov AA, Dunbrack RL (2003) A graph theory algorithm for rapid protein side chain prediction. Protein Sci 12:2001 2014
4. el Bastawissy E, Knaggs MH, Gilbert IH (2001) Molecular dynamics simulations of wild type and point mutation human prion protein at normal and elevated temperature. J Mol Graph Model 20:145 154
5. Falconi M, Biocca S, Novelli G et al (2007) Molecular dynamics simulation of human LOX 1 provides an explanation for the lack of OxLDL binding to the Trp150Ala mutant. BMC Struct Biol 7:73
6. Feyfant E, Sali A, Fiser A (2007) Modeling mutations in protein structures. Protein Sci 16:2030 2041
7. Hubbard SJ, Campbell SF, Thornton JM (1991) Molecular recognition. Conformational analysis of limited proteolytic sites and serine proteinase protein inhibitors. J Mol Biol 220:507 530
8. Li IT, Pham E, Truong K (2007) Current approaches for engineering proteins with diverse biological properties. Adv Exp Med Biol 620:18 33

9. Liu B, Bernard B, Wu JH (2006) Impact of EGFR point mutations on the sensitivity to gefitinib: insights from comparative structural analyses and molecular dynamics simulations. Proteins 65:331 346
10. Marabotti A (2008) Modeling the conformation of side chains in proteins: approaches, problems and possible developments. Curr Chem Biol 2:200 214
11. Matsumura M, Wozniak JA, Sun DP et al (1989) Structural studies of mutants of T4 lysozyme that alter hydrophobic stabilization. J Biol Chem 264:16059 16066
12. McLachlan AD (1982) Rapid comparison of protein structures. Acta Crystallogr A 38:871 873
13. Nicholson H, Anderson DE, Dao pin S et al (1991) Analysis of the interaction between charged side chains and the alpha helix dipole using designed thermostable mutants of phage T4 lysozyme. Biochemistry 30:9816 9828
14. Novotny J, Bruccoleri R, Karplus M (1984) An analysis of incorrectly folded protein models. Implications for structure predictions. J Mol Biol 177:787 818
15. Novotny M, Seibert M, Kleywegt GJ (2007) On the precision of calculated solvent accessible surface areas. Acta Crystallogr D Biol Crystallogr 63:270 274
16. Peterson RW, Dutton PL, Wand AJ (2004) Improved side chain prediction accuracy using an ab initio potential energy function and a very large rotamer library. Protein Sci 13:735 751
17. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232:584 599
18. Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234:779 815
19. Shih HHL, Brady J, Karplus M (1985) Structure of proteins with single site mutations: a minimum perturbation approach. Proc Natl Acad Sci USA 82:1697 1700
20. Snow ME, Amzel LM (1986) Calculating three dimensional changes in protein structure due to amino acid substitutions: the variable region of immunoglobulins. Proteins 1:267 279
21. Summers NL, Carlson WD, Karplus M (1987) Analysis of side chain orientations in homolo gous proteins. J Mol Biol 196:175 198
22. Xiang Z (2006) Advances in homology protein structure modeling. Curr Protein Pept Sci 7:217 227
23. Xiang Z, Honig B (2001) Extending the accuracy limits of prediction for side chain con formations. J Mol Biol 311:421 430
24. Yuan Z, Bailey TL, Teasdale RD (2005) Prediction of protein B factor profiles. Proteins 58:905 912
25. Zamyatin AA (1972) Protein volume in solution. Prog Biophys Mol Biol 24:107 123

Chapter 33
Temporal Anomaly Detection: An Artificial Immune Approach Based on T Cell Activation, Clonal Size Regulation and Homeostasis

Mário J. Antunes and Manuel E. Correia

Abstract This paper presents an artificial immune system (AIS) based on Grossman's tunable activation threshold (TAT) for temporal anomaly detection. We describe the generic AIS framework and the TAT model adopted for simulating T Cells behaviour, emphasizing two novel important features: the temporal dynamic adjustment of T Cells clonal size and its associated homeostasis mechanism. We also present some promising results obtained with artificially generated data sets, aiming to test the appropriateness of using TAT in dynamic changing environments, to distinguish new unseen patterns as part of what should be detected as normal or as anomalous. We conclude by discussing results obtained thus far with artificially generated data sets.

Keywords Artificial immune systems · Pattern recognition · Anomaly detection · Homeostasis

33.1 Introduction

The vertebrate immune system (IS) [4] evolved to become a highly complex defence mechanism, with the ability to recognize foreign substances (pathogens) and to distinguish between those that correspond to the harmless (self) from those that are related to some form of intrusion (non-self). The IS is composed by the innate and adaptive layers, being supported by a complex set of cellular structures. Antigen Presenting Cell (APC) digests and converts pathogens into small *peptides* which are then presented to T cells, a lymphocyte, through a molecular structure denominated "MCH/Peptide Complex." T cells have a specific set of *receptors* that *binds* with a certain degree of affinity with the peptides that are being presented by APCs.

M.J. Antunes (✉)
School of Technology and Management, Polytechnic Institute of Leiria, Morro do Lena, Alto do Vieiro, 2411 901 Leiria, Portugal
e mail: mario.antunes@ipleiria.pt

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 33,

Artificial immune systems (AIS) [7] comprises a full body of models and algorithms devised by theoretical immunologists that describe and successfully predict certain aspects of the IS behaviour. These algorithms and models constitute the basis and source of inspiration behind the developments of AIS for anomaly detection, being usually divided into two major groups [10]. The first group comprises all AIS based on the classical Burnet's negative selection theory [5], like Kim's research [9]. The other group includes those that take inspiration on Matzinger's danger theory [11], comprising Aickelin's research the most well known [10]. However, these systems suffer from some well-documented serious limitations [13, 14], which motivated us to investigate the appropriateness of developing an AIS based on a rather different biological theoretical perspective about the IS [1 3]: the Grossman's Tunable activation thresholds (TAT) hypothesis [8]. TAT posits that each individual immune cell has its own TAT whose value reflects the recent history of interactions with the surrounding environment. The potentially autoimmune (self) lymphocytes, which are continuously exposed to body antigens, end up raising their activation thresholds and thus become unresponsive. In contrast, lymphocytes that are not auto reactive and recognize external microorganisms end up with low activation thresholds becoming thus fully responsive towards non-self antigens. TAT behaviour is thus completely dynamic. Its current state and emergent collective behaviour depends on the rate of change and intensity of the signalling for each cell that resulted from past interactions with the antigens each cell happens to get in contact with throughout time [15]. In summary, for the AIS designer TAT assumes no prior "classification" of antigens as either "self" or "non-self," and it is expected to automatically adjust each one of the individual cell activation dynamics with the current environment, throughout time.

In our framework [1, 3], we have also included two immunological concepts: cells clonal size regulation mechanisms and a dynamic equilibrium based on the sharing of finite resources (homeostasis). The former is related to the accepted idea from Biology that immune cells proliferate after being activated, cloning themselves and enabling a faster reaction for a second future encounter with that same pathogen [8]. We mimic this concept by associating to each artificial cell a *weight* expressed by the number of clones of each T cell. The later derives from the idea that the size of the immune cell repertoire is constantly changing by adapting to the current environmental circumstances. In this paper we aim to introduce the minimal TAT model adopted (Sect. 33.2) and to describe the deployed TAT-based AIS for anomaly detection (Sect. 33.3). We also present results obtained with artificially generated data sets (Sect. 33.4). Finally, we discuss the results obtained, draw some conclusions and delineate guidelines for future research (Sect. 33.5).

33.2 A Minimal TAT Model

The TAT theory hypothesizes that immune cell activation depends on a threshold that is adjusted dynamically and at each point in time it corresponds to the integrated received signalling past history. Every interaction between the cell receptor and the

peptide ligands presented by the APC, results in an intracellular competition between "excitation" and "de-excitation" signalling pathways, causing the cell to adapt to the stimulus by increasing or decreasing its activation threshold. Therefore, cells with different antigen specificity will have different activation thresholds as they are exposed to different stimuli. During its lifetime, each cell changes and adapts its responsiveness according to its interaction with the environment [8].

We have adopted a minimal mathematical model of TAT for T cells [6]. Briefly, the model states that T-cell activation is controlled by two enzymes that respond to antigenic signals delivered by APCs: Kinase (K) and Phosphatase (P). Antigenic signals (S) lead to a linear increase of both K and P activities until they reach a plateau that is proportional to the intensity of the stimulus. For the same signal S, K increases faster than P, but if the signal persists P will eventually reach a higher plateau. Similarly, on signalling absence, K returns to the basal level at a faster rate than P. A complete explanation of the adopted model can be found in [1 3, 6].

33.3 The TAT-Based AIS Framework

In the TAT-AIS, as with other AIS [7], there is a direct mapping of system components with the relevant biological IS immune counterparts. In [1 3] there is a comprehensive description of both the immunological metaphor adopted in our TAT model and the TAT-AIS core framework adopted.

The data sets used in the experiments have been artificially generated through a stochastic procedure. Each data set is comprised by APCs that represent a timely ordered set of events and the corresponding observed patterns. The core of each APC is composed by a list of strings separated by white space, corresponding to artificial peptides (PEPTIDEs) that are being presented by the artificial APC. For system performance evaluation purpose all APCs have been pre-tagged as "NORMAL" or "ALERT," as illustrated in the example that follows:

apc:1232:NORMAL: cbac accb fggf (...) eehf efge abcd bcad cadc bdbd daca
apc:1233:ALERT: abcc aBCB PSQP (...) ABAB abdd 3421 PPPR DADA bQPR

The data sets basic alphabet generators we have used for our experiments are shown in Table 33.1. All the PEPTIDEs of a given artificial data set belong to the set of all possible fixed length strings (we have used length four) that can be obtained, allowing for character repetition, from the given alphabets.

Table 33.1 Alphabet used to create the PEPTIDEs

Alphabet	Data sets	APC tag
{a,b,c,d}	Training, testing	Normal
{a,A,B,C,D}	Vaccination, testing	Abnormal
{b,P,Q,R,S}	Testing	Abnormal
{e,f,g,h, 1,2,3,4, *,+,@, }	Testing	Normal

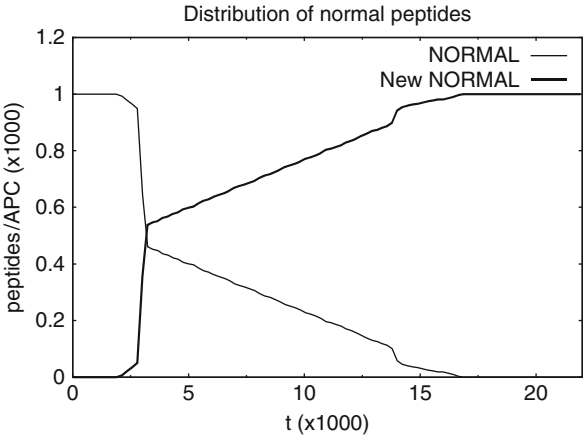


Fig. 33.1 Distribution of normal occurrences during training and testing phases

We then use these PEPTIDEs to randomly assemble an APC. If the APC has at least one abnormal PEPTIDE then it is tagged as "ALERT," otherwise, the APC is tagged as "NORMAL." The PEPTIDEs are randomly chosen from the list of admissible strings with a certain time distribution, as illustrated in Fig. 33.1.

The normal behaviour presented in the training dataset changes progressively in the testing phase, where normal APC content is gradually replaced by new unseen normal PEPTIDEs. The APCs corresponding to unseen abnormal behaviour appears sporadically in the testing dataset in a randomly temporal order. We have also included two different sets of normal PEPTIDEs ({1,2,3,4} and {*,+,@,=}) that only appear in the testing phase and only occur in a certain period of time.

The TAT-AIS simulator requires some parameters to run. These are related to K and P dynamics, and are comprised by their initial values (K_0 and P_0), the slopes that define the way K (ϕP) and P (ϕP) change, the plateau values ($K_{\max}$ and $P_{\max}$), the affinity threshold and a value for the detection threshold.

In order to reduce the number of simulation parameters and therefore to simplify their run time optimization, we have chosen to derive K_0 and P_0 and opted to fix all the parameters with the exception of $K_{\max}$ and ϕP, as described below:

K_0	$S_0 \times K_{\max}$	$P_{\max}$	$\left(\frac{K_{\max}}{P_{\max}}\right) \times K_{\max}$
P_0	$S_0 \times P_{\max}$	ΦP	$\left(\frac{\Phi P}{\Phi K}\right) \times \Phi K$

The ratios $\frac{K_{\max}}{P_{\max}}$ and $\frac{\Phi P}{\Phi K}$ varies between 0 and 1 and are optimized in the vaccination phase by an implementation of a meta-heuristic approach using a simplex algorithm for non-linear optimization [12].

The artificial T-cell (TCELL) clonal size is a number that represents the size of the sub-population of clones of that TCELL. TCELLs are created with an initial clonal size value (we adopted C_0=2). Each time an APC is processed, the clonal

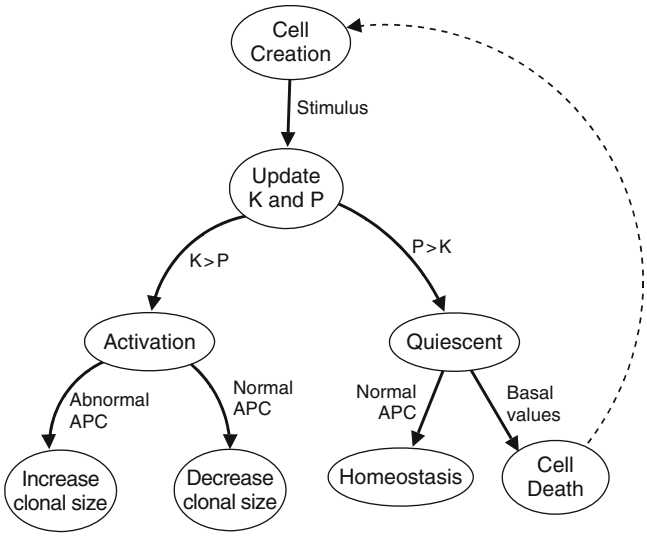


Fig. 33.2 TCELL life cycle

size of affected TCELLs is updated in order to reflect the accuracy level of detection each TCELL obtain through time. We defined that C increases and decreases by units of two. For TCELLs with $P > K$ we fixed the increase in the order of 0.5 units. These were the constant values that gave us good results in practice [3].

Figure 33.2 depicts the TCELL life cycle. At first, a new TCELL is created when the pattern represented by the PEPTIDE does not match any one in the current repertoire.

When there is a match with an appropriate affinity, each matching TCELL updates its K and P values according to the signal sent by the APC [3]. If a TCELL becomes active, there are two possible outcomes: the APC is abnormal, then the TCELL proliferates and increases its clonal size; the APC is normal, then the TCELL decreases its clonal size. If the TCELL remains quiescent (with $P > K$) it will contribute to the homeostasis process. Finally, a TCELL will die if it is not stimulated for a certain predefined period of time, thus decreasing gradually its K and P values to its initial values (K_0 and P_0) and its clonal size is below a predefined minimum.

33.4 Evaluation and Results

Our working hypothesis is that TAT-AIS is able to recognize new unseen patterns and is able to further distinguish between those patterns that are considered self from others, included in APCs related to abnormal activities. We have tested the

system with two different artificial data sets specially constructed to include a gradual variation of normal behaviour throughout time. The affinity measure corresponds to the number of equal characters at the same position in the PEPTIDE and TCELL string identifiers. We have tested the data sets with two different affinity measures for each run: one (25%) and two (50%) equal characters. The detection threshold, calculated by the ratio between bound and activated TCELLs clonal size, was fixed at 0.2 (20%) for all the experiments. The training comprises 2,000 APCs (500 for the vaccination phase) and the testing phase has 20,000 and 30,000 APCs for each dataset, respectively. The TAT parameters have been fixed with the following values: $H_{\max} = 256$ correspond to the repertoire size of normal TCELLs for homeostasis; $C_{\max}$ is the clonal size reached by a TCELL that tends to recognize abnormal patterns and is calculated by $C_{\max} = H_{\max} \times 0.2$; S_0, $K_{\max}$ and ΦK are equal to 10. The optimized parameters achieved in the "vaccination" phase and the results obtained for each experiment are described in Table 33.2. The repertoire size corresponds to the number of TCELLs used in each execution phase.

In spite of the great number of patterns that the system is exposed to, TAT-AIS possesses an efficient cell death mechanism (apoptosis) that is capable of maintaining an effective low cell repertoire size. It is also worth noting that even with a total replacement of the normal behaviour during the testing phase the TAT-AIS obtained a full detection rate with a relatively low number of false positives.

During the testing phase, TCELLs clonal size changes according to whether the TCELLs recognizes normal or abnormal patterns, as depicted in Fig. 33.3a. For example the TCELL "abcd" binds only with normal patterns and its clonal size

Table 33.2 Optimized parameters and the results obtained during experiments

Run	Optimized parameters			Repertoire size			TPs		FPs	
	$\frac{K_{\max}}{P_{\max}}$	$\frac{\Phi P}{\Phi K}$	Affinity threshold	Training	Vaccine	Testing	Qty	%	Qty	%
1	1.356	0.473	0.25	135	194	255	46	100	180	0.9
2	1.348	0.138	0.50	135	258	345	46	100	394	1.9
3	1.450	0.460	0.25	212	295	454	47	100	368	1.2
4	1.438	0.484	0.50	212	380	504	47	100	220	0.7

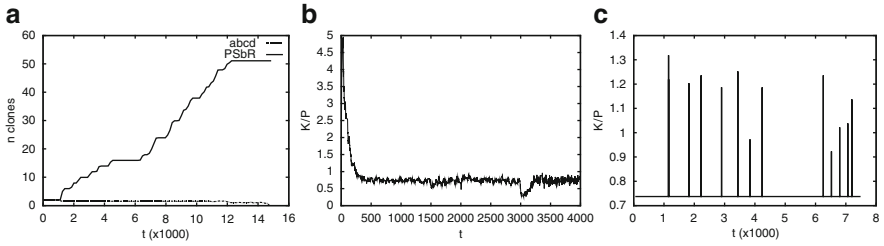


Fig. 33.3 (**a**) Clonal size dynamics; (**b**) Signal activity for TCR "abcd" recognizes a normal pattern. (**c**) TCR "PSbR" recognizes abnormal PEPT IDEs in the APCs

contributes to the system homeostasis equilibrium. In another direction, TCELL "PSbR" binds sporadically to the PEPTIDEs presented by abnormal APCs, thus its clonal size increases throughout time.

The signalling activity of both TCELLs is also depicted in Fig. 33.3b, c. For a normal TCELL like "abcd," after an initial period of time, the ratio $\frac{K}{P}$ is below one, which means the TCELL is inactive (Fig. 33.3b). Otherwise, for those associated with an abnormal pattern like "PSbR," the ratio is transiently below and above one, reflecting successive activations (Fig. 33.3c).

33.5 Discussion

We have presented some results obtained in a temporal anomaly detection AIS based on TAT theory, with two artificially generated data sets of predefined patterns, resulting from normal and abnormal behaviours. These results are in line and improved upon with some previously published results about TAT-AIS [1 3]. We have observed that TAT-AIS has interesting properties for anomaly detection, provided the following basic generic requirements is true: normal behaviour is frequent and abnormal behaviour is sporadic in time. By *frequent* we mean a pattern that repeatedly stimulates a set of TCELLs that through time, by the TAT dynamics, stabilizes its enzymatic values ($P > K$). On the other hand, by *sporadic* we mean a pattern that stimulates intermittently a set of TCELLs with such a signal that implies its activation ($K > P$). Through time, these TCELLs increase gradually its clonal size and become more reactive to further similar recognitions. In TAT-AIS, detection is also dependent on the TCELLs dynamical clonal size control mechanism surrounded by a homeostatic procedure for TCELLs that helps to recognize self patterns, allowing the abnormal TCELLs to grow up until a predefined plateau is reached.

The results thus obtained with TAT-AIS are very satisfactory, achieving a high rate of detection and a low level of false positives on the stochastic data sets we have produced. It is worth to emphasize that in these data sets, during the testing phase, normal behaviour is made to change gradually throughout time. In spite of this, TAT-AIS is able to correctly detect the new patterns and to correctly sort them into normal and anomalous sets. Moreover, based on these empirical results, we believe that TAT-AIS can compete with other approaches on the self-non-self distinction for dynamic environments that tend to change gradually throughout time their normality behaviour profile. Firstly, the clonal size and activation threshold of each TCELL is being continuously updated according to its interactions with the environment. This means that each TCELL keeps track of its historical activity and, through time, delineates a trend to detect patterns corresponding to normal or abnormal behaviours. Secondly, by taking advantage of the "vaccination" phase, TAT-AIS can learn patterns of known anomalies that may improve the detection of both known and unknown anomalous behaviours during the testing phase. We are well aware that these stochastic data sets were artificially

generated and are most certainly not completely representative of real-world phenomenon like the data sets we could obtain, for example by live network traffic collection or computer systems trace logs. We have, however, already obtained some preliminary good results with a simpler less sophisticated TAT execution model, applied to network intrusions detection with real network traffic [1].

Acknowledgements The authors acknowledge the facilities provided by the CRACS research unit, an INESC associate of the Faculty of Science, University of Porto.

References

1. Antunes M, Correia E (2008) TAT NIDS: an immune based anomaly detection architecture for network intrusion detection, Proceedings of IWPACBB'08 Advances in Soft Computing (Springer), pp 60 67
2. Antunes M, Correia E, Carneiro J (2009) Towards an immune inspired temporal anomaly detection algorithm based on tunable activation thresholds, Proceedings of International Conference of Bioinspired systems and signal processing (BIOSIGNALS), pp 357 362
3. Antunes M, Correia E (2009) An Artificial Immune System for Temporal Anomaly Detection Using Cell Activation Thresholds and Clonal Size Regulation with Homeostasis, Proceedings of International Joint Conferences on Bioinformatics, Systems Biology and Intelligent Computing (IJCBS), pp 323 326
4. Burmester G, Pezzuto A (2003) Color Atlas of Immunology. Thieme Medical Publishers, George Thieme Verlag
5. Burnet F (1959) The Clonal Selection Theory of Acquired Immunity, Vanderbilt University Press Nashville, Tennesse
6. Carneiro J, Paixao T et al (2005) Immunological self tolerance: Lessons from mathematical modeling, Journal of Computational and Applied Mathematics 184(1):77 100
7. Castro L, Timmis J (2002) Artificial Immune Systems: A New Computational Intelligence Approach. Springer, New York
8. Grossman Z and Singer A (1996), Tuning of activation thresholds explains flexibility in the selection and development of Tcells in the thymus, Proceedings of the National Academy of Sciences 93(25):14747 14752
9. Kim J, Bentley P (2001) An evaluation of negative selection in an artificial immune system for network intrusion detection, Proceedings of Genetic and Evolutionary Computation Confer ence (GECCO), 1330 1337
10. Kim J, Bentley P, Aickelin U, Greensmith J, Tedesco G, and Twycross J (2007) Immune system approaches to intrusion detection a review, Natural Computing 6(4):413 466
11. Matzinger P (2002) The danger model: a renewed sense of self, Science's STKE 296 (5566):301 305
12. Pedroso J (2007) Simple Metaheuristics Using the Simplex Algorithm for Non linear Programming, Engineering Stochastic Local Search Algorithms. Designing, Implementing and Analyzing Effective Heuristics LNCS (Springer), 4638:217 221
13. Stibor T, Timmis J and Eckert C (2005) On the appropriateness of negative selection defined over hamming shape space as a network intrusion detection system, Proceedings of IEEE congress on Evolutionary Computation (CEC) 2:995 1002
14. Vance R (2000) Cutting edge commentary: A Copernican revolution? doubts about the danger theory, The Journal of Immunology 165:1725 1728
15. van den Berg H and Rand D (2004) Dynamics of T cell activation threshold tuning, Journal of Theoretical Biology 228(3):397 416

Chapter 34
An Integrated Methodology for Mining Promiscuous Proteins: A Case Study of an Integrative Bioinformatics Approach for Hepatitis C Virus Non-structural 5a Protein

Mahmoud M. ElHefnawi, Aliaa A. Youssif, Atef Z. Ghalwash, and Wessam H. El Behaidy

Abstract A methodology for elucidation of structural, functional, and mechanistic knowledge on promiscuous proteins is proposed that constitutes a workflow of integrated bioinformatics analysis. Sequence alignments with closely related homologues can reveal conserved regions which are functionally important. Scanning protein motif databases, along with secondary and surface accessibility predictions integrated with post-translational modification sites (PTMs) prediction reveal functional and protein-binding motifs. Integrating this information about the protein with the GO, SCOP, and CATH annotations of the templates can help to formulate a 3D model with reasonable accuracy even in the case of distant sequence homology. A novel integrative model of the non-structural protein 5A of Hepatitis C virus: a hub promiscuous protein with roles in virus replication and host interactions is proposed. The 3D structure for domain II was predicted based on, the *Homo sapiens* Replication factor-A protein-1 (RPA1), as a template using consensus meta-servers results. Domain III is an intrinsically unstructured domain with a fold from the retroviral matrix protein, which conducts diverse protein interactions and is involved in viral replication and protein interactions. It also has a single-stranded DNA-binding protein motif (SSDP) signature for pyrimidine binding during viral replication. Two protein-binding motifs with high sequence conservation and disordered regions are proposed; the first corresponds to an Interleukin-8B receptor signature (IL-8R-B), while the second has a lymphotoxin beta receptor (LTβR) high local similarity. A mechanism is proposed to their contribution to NS5A Interferon signaling pathway interception. Lastly, the overlapping between LTβR and SSDP is considered as a sign for NS5A date hubs.

Keywords Bioinformatics · Hepatitis C · Promiscuous proteins · Protein 5A · Sequence alignments · SSDP

M.M. ElHefnawi (✉)
Informatics and Systems Department, National Research Center, Cairo, Egypt

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 34,

34.1 Introduction

Structural bioinformatics is used to overcome some special challenges, like *protein structure prediction* which aims to reduce the sequence-structure gap. Protein structure prediction methods are divided into three categories: homology modeling, fold recognition (or threading methods), and ab initio prediction. Recently, fold recognition becomes the most successful approach to this problem.

The fold recognition has different approaches to solve structure prediction problem. Thus the performance of meta-servers, which combine the consensus predictions from different servers, is better than individual predictors [1].

Integrative bioinformatics attempts to combine heterogeneous data sources for integrative understanding of the biology of organisms. Special attention should be given to viruses because of their small genomes that make their modeling easier and the applications of this knowledge in drug and vaccine design.

Viral proteins are typically involved in many protein protein interactions (PPIs). Such proteins are referred to as hub proteins. These proteins have two types; "date" and "party" hubs. Party hubs interact with different proteins at the same time, while date hubs interact with other proteins at different times. This ability is because of the existence of intrinsic disorder in one or both interacted proteins. Also, these intrinsic disorders are the reason in the diversity of domain architecture or the lack of unique 3D structure [2, 3].

Promiscuous domains are domains that are involved in diverse PPIs especially in various forms of signal transduction pathways. Promiscuity is a volatile and relatively fast-changing feature in evolution thus evolution of promiscuous domains seems to be constrained by the diversity of their interaction [4].

The analysis and prediction of sequences could have profound impacts on different areas of biology and medicine. Hepatitis C Virus (HCV) is a major threat to global public health. The non-structural 5A (NS5A) protein is one of its polyprotein and there are many questions to answer about its role, functions, and the structure of its last two domains [5].

An integrative approach for mining knowledge from protein sequences is introduced, and its application is shown on the NS5A HCV protein to gain a better understanding of the mechanisms and promiscuity of this date and party hub viral protein.

34.2 Methods

Our methodology for protein analysis works on two levels, as shown in Fig. 34.1: Sequence level, and Domain level.

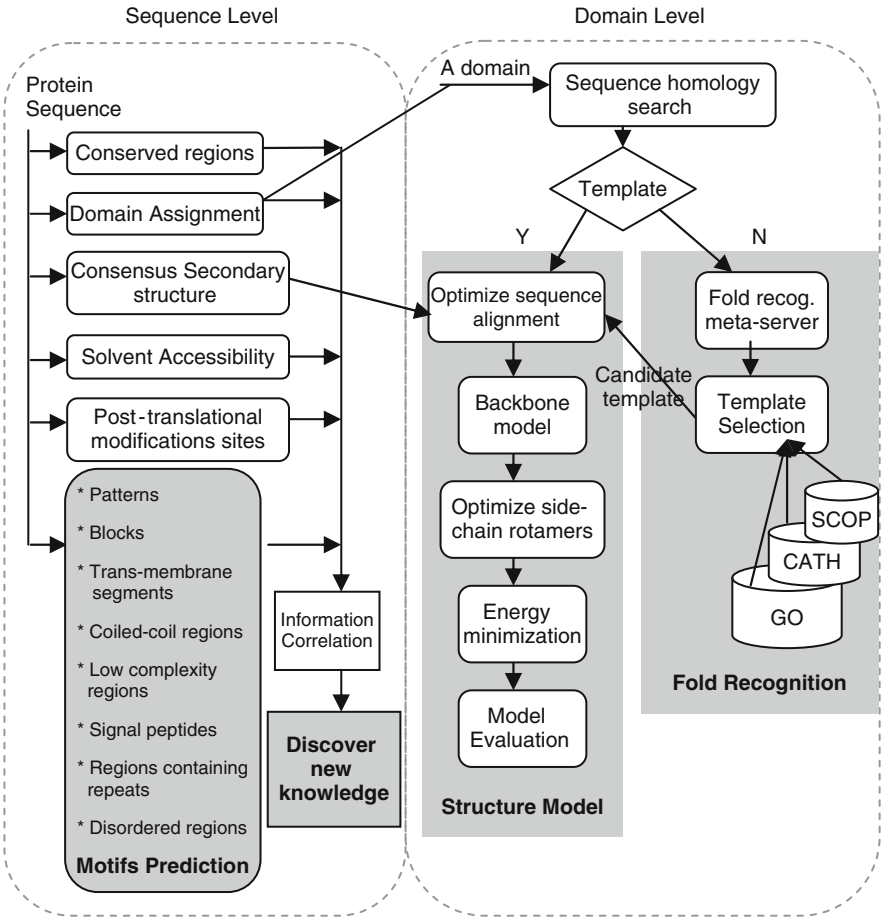


Fig. 34.1 Flowchart of integrative workflow. The protein sequence is analyzed using different servers for conserved regions, consensus secondary structures, solvent accessibility, patterns, post translational modifications (PTMs) sites, trans membrane segments, disordered regions, and functional motifs. Also, the domains are assigned, homology to PDB is checked, and fold recognition is applied for distant homologues. Templates are chosen based on consensus scores from the meta servers, and structural and functional annotations from CATH, GO, and SCOP that match biological knowledge about the structure and function of the protein. The fold recognition output is the candidate template that is used for structure model. The structure model begins by optimizing the sequence alignment based on secondary structure, then the backbone is modeled and the side chains rotamers are optimized. The energy minimization is the followed step and finally the evaluation of the model is performed

34.2.1 Sequence Level

In sequence level, the protein sequence is passed to multiple servers that outputs independent variant information that is studied and filtered according to their

relatedness to the protein's functions. The correlation between this information enables us to discover new knowledge about this protein.

The conserved regions are important to detect the regions that are common between the different genotypes or the same family of a protein. These regions highlight the locations that are important to the functionality of that protein. The PROMALS server [6] is used for the multiple sequence alignments, and then BIOEDIT software [7] was used to extract the conserved regions.

The separation of domain boundaries of a protein is a critical step for the tertiary structure prediction. The DOMAC server [8], as one of top servers in CASP7, is used for multi-domain prediction. Also, the top server in CAFASP4, SSEP server is used for two-domain prediction [9]. Each output domain from DOMAC is passed to SSEP server to be sure that it is only one domain.

The secondary structure prediction is useful to understand the protein folding, and tertiary structure prediction, and to increase the accuracy of sequence alignment. Besides, it is helpful in protein classification. To maximize the accuracy of secondary structure prediction, consensus results from five prediction servers are used: Porter [10], PROF [11], PSIPRED [12], SSPRO [13], and SAM-t02 [14].

Solvent accessibility is required to understand the protein tertiary structure, protein stability, and protein interaction analysis. Distill solvent accessibility [15] was used to find the exposed and buried regions in the protein sequence.

The motifs prediction is a key step for protein function and protein interactions prediction. Two approaches are used; regular expression approach used by PROSITE server [16], and statistical model approach used by BLOCKS [17] and SMART [18] servers. These servers are used to detect patterns, blocks, functional sites, and to predict trans-membrane segments, coiled-coil regions, low complexity regions, signal peptides, regions containing repeats, and disordered regions.

Finally, the Scansite [19] server is used to predict the post-translational modifications (PTMs) sites.

34.2.2 Domain Level

The domain level is concerned to the prediction of unknown structure of a protein. The input to this level is the output from the domain assignment step.

Sequence homology search step finds if this domain has a homologue in Protein Data Bank (PDB) with similarity greater than 30% or not. The domain sequence is passed as the query to the two major heuristic servers, BLAST [20] and FASTA [21], as well as to the ParAlign [22], and ScanPS [23] servers, which implement a modified version of the Smith Waterman algorithm. If a high similarity homologue is found, the model structure step begins.

Fold recognition prediction techniques were used for domains that don't have a direct homologue. The meta-server 3D-Jury [24] uses 12 different servers, the resulted folds and their corresponding SCOP, CATH, and GO annotation are

collected and studied. From all these information, the candidate template is specified and then used for structure modeling.

Multiple steps are required to model the specified domain. Firstly, the alignment of the candidate template, with domain sequence, is optimized using PROMALS server [6] based on the secondary structure prediction. Secondly, the UCSF Chimera program [25] is used to create an initial model structure. Thirdly, SCWRL4.0 program [26] is used to optimize side chain rotamers. Fourthly, the Chimera [25] uses the *steepest descent* method with the AMBER force field for energy minimization. Finally, the multiple verification methods in Swiss-Model structure assessment tools [27] are used to assure high accuracy of the model.

34.3 Results and Discussion

The proposed integrative methodology is applied to the NS5a HCV protein to analyze it and predict the structure of its last two domains. From each step, the results of servers are used and integrated to formulate our analysis. The detailed results are shown in our previous papers [28, 29].

The refinement of NS5A domains was the first step for structure prediction. Fold recognition techniques were used to select the appropriate fold for domain II (204-295). The functions of this candidate structure correspond with NS5A replication and interaction with transcriptional activators [30], like p53, which leads to cell growth and tumor formation [5]. Domain III (296-44) has 61% intrinsic disorder regions, which give it the ability to scaffold and recruit of different binding partners in space and time, and lacks a unique 3D structure. The biologically closest fold among the 3D-jury servers was the retroviral matrix protein, which is involved also in viral replication and protein interactions [31].

The different bioinformatics servers were used to analyze the NS5A protein and to discover protein-binding motifs relating to its hubness, promiscuity, and biological functions. This study included the prediction of family conserved regions, secondary structures, solvent accessibility, post-translational modification sites, interaction motifs, and disordered regions. The correlation of all these information demonstrates that the beta sheets are totally buried, while the alpha helices and the conserved regions are partially exposed. The intrinsic disorder regions are exposed and the three regions (217-234aa, 299-313aa, and 374-410aa) always start before the end of their corresponding alpha helices by two residues. Also, two different mechanisms were proposed by which the NS5A protein intercepts the immune system, using three predicted motifs. Others were specified for HCV replication and phosphorylation. The date hubs could be noticed between the two overlapping interacted motifs single-stranded DNA protein (399-444aa) and lymphotoxin beta receptor (418-430aa).

The future direction would be to automate the above analysis approach with information correlation and structure identification.

34.4 Conclusion

The proposed integrative methodology has been successful in understanding mechanisms of interaction, and relating the sequence to structural features and 3D structure to the many functions of the NS5A protein. Also, the NS5A protein was confirmed as a hub promiscuous protein after it was studied using the integrative approach. It was intriguing all of our predicted binding motifs as also composed of disordered regions and hot loops. Promiscuity and protein disorder were found to be two highly associated attributes through analysis of the NS5A protein.

References

1. Rychlewski L, Fischer D (2005) LiveBench 8: The large scale, continuous assessment of automated protein structure prediction. Prot Sci 14:240 245.
2. Gsponer J, Madan Babu M (2009) The rules of disorder or why disorder rules. Prog Biophys & Mol Biol. doi: 10.1016/j.pbiomolbio.2009.03.001.
3. Dunker A K, Cortese M S, Romero P, Iakoucheva L M, Uversky V N (2005) Flexible nets: The roles of intrinsic disorder in protein interaction networks. FEBS J 272(20):5129 5148.
4. Basu M K, Poliakov E, Rogozin I B (2009) Domain mobility in proteins: functional and evolutionary implications. Brief Bioinform 10(3):205 216.
5. Macdonald A, Harris M (2004) Hepatitis C virus NS5A: tales of a promiscuous protein. J Gen Virol 85:2485 2502.
6. Pei J, Grishin N V (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23(7):802 808.
7. Hall T A (1999) BioEdit: a user friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucl Acids Symp Ser 41:95 98.
8. Cheng J (2007) DOMAC: an accurate, hybrid protein domain prediction server. Nucl Acids Res 35:354 356.
9. Gewehr J E, Zimmer R (2006) SSEP domain: protein domain prediction by alignment of secondary structure elements and profiles. Bioinformatics 22(2):181 187.
10. Pollastri G, McLysaght A (2005) Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 21:19 20.
11. Rost B, Yachdav G, Liu J (2004) The PredictProtein server. Nucl Acids Res 32:W321 W326.
12. McGuffin L J, Bryson K, Jones D T (2000) The PSIPRED protein structure prediction server. Bioinformatics 16:404 405.
13. Cheng J, Randall A, Sweredoski M, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucl Acids Res 33:72 76.
14. Karplus K et al (2005) SAM T04: what is new in protein structure prediction for CASP6. Proteins 61(7):135 142.
15. Pollastri G, Martin A J M, Mooney C, Vullo A (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and struc ture information. BMC Bioinformatics 8:201.
16. Hulo N, Bairoch A, Bulliard V et al (2008) The 20 years of PROSITE. Nucl Acids Res 36: D245 D249.
17. Henikoff J G, Greene E A, Pietrokovski S, Henikoff S (2000) Increased coverage of protein families with the blocks database servers. Nucl Acids Res 28:228 230.
18. Ponting C P, et al (1999) SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucl Acids Res 27: 229 232.

19. Wan J et al (2008) Meta prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. Nucl Acids Res 36: e22.
20. Altschul S F, et al (1997) Gapped BLAST and PSI BLAST: a new generation of protein database search programs. Nucl Acids Res 25:3389 3402.
21. Pearson W R (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol 183:63 98.
22. Sæbø P E, Andersen S M, Myrseth J, Laerdahl J K, Rognes T (2005) PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. Nucl Acids Res 33:535 539.
23. Walsh T P et al (2008) SCANPS: a web server for iterative protein sequence database searching by dynamic programming, with display in a hierarchical SCOP browser. Nucl Acids Res 36:W25 W29.
24. Kaján L, Rychlewski L (2007) Evaluation of 3D Jury on CASP7 models. BMC Bioinformat ics 8:304.
25. Pettersen E F, Goddard T D, Huang C C et al (2004) UCSF Chimera a visualization system for exploratory research and analysis. J Comput Chem 25(13):1605 1612.
26. Krivov G G, Shapovalov M V, Dunbrack R L (2009) Improved prediction of protein side chain conformations with SCWRL4. Proteins doi: 10.1002/prot.22488.
27. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS MODEL Workspace: A web based environment for protein structure homology modeling. Bioinformatics 22:195 201.
28. El Hefnawi M, El Behaidy W, Youssif A, Ghalwash A, El Housseiny L A, Zada Z (2009) Natural genetic engineering of Hepatitis C virus NS5A for immune system counterattack. Nat Gen Eng and Nat Genome Editing: Ann NY Acad Sci 1178:173 185.
29. El Hefnawi M, Youssif A, Ghalwash A, El Behaidy W (2009) An integrative *in silico* model of Hepatitis C virus non structural 5a protein. Int Conf Bioinf Comput Biol, BIOCOMP'09:827 833.
30. Jacobs D M, Lipton A S, Isern N G et al (1999) Human replication protein A: global fold of the N terminal RPA 70 domain reveals a basic cleft and flexible C terminal linker. J Biomol NMR 14:321 331.
31. Christensen A M, Massiah M A, Turner B G, Sundquist W I, Summers M F (1996) Three dimensional structure of the HTLV II matrix protein and comparative analysis of matrix proteins from the different classes of pathogenic human retroviruses. J Mol Biol 264 (5):1117 1131.

Chapter 35
Enhanced Prediction of Conformational Flexibility and Phosphorylation in Proteins

Karthikeyan Swaminathan, Rafal Adamczak, Aleksey Porollo, and Jarosław Meller

Abstract Many sequence-based predictors of structural and functional properties of proteins have been developed in the past. In this study, we developed new methods for predicting measures of conformational flexibility in proteins, including X-ray structure-derived temperature (B-) factors and the variance within NMR structural ensemble, as effectively measured by the solvent accessibility standard deviations (SASDs). We further tested whether these predicted measures of conformational flexibility in crystal lattices and solution, respectively, can be used to improve the prediction of phosphorylation in proteins. The latter is an example of a common post-translational modification that modulates protein function, e.g., by affecting interactions and conformational flexibility of phosphorylated sites. Using robust epsilon-insensitive support vector regression (ε-SVR) models, we assessed two specific representations of protein sequences: one based on the position-specific scoring matrices (PSSMs) derived from multiple sequence alignments, and an augmented representation that incorporates real-valued solvent accessibility and secondary structure predictions (RSA/SS) as additional measures of local structural propensities. We showed that a combination of PSSMs and real-valued SS/RSA predictions provides systematic improvements in the accuracy of both B-factors and SASD prediction. These intermediate predictions were subsequently combined into an enhanced predictor of phosphorylation that was shown to significantly outperform methods based on PSSM alone. We would like to stress that to the best of our knowledge, this is the first example of using predicted from sequence NMR structure-based measures of conformational flexibility in solution for the prediction of other properties of proteins. Phosphorylation prediction methods typically employ a two-class classification approach with the limitation that the set of negative examples used for training may include some sites that are simply unknown to be phosphorylated. While one-class classification techniques have been considered in the past as a solution to this problem, their performance has not been

K. Swaminathan (✉)
Department of Biomedical Engineering, University of Cincinnati, Cincinnati, OH, USA

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 35,

systematically compared to two-class techniques. In this study, we developed and compared one- and two-class support vector machine (SVM)-based predictors for several commonly used sets of attributes. [These predictors are being made available at http://sable.cchmc.org/].

Keywords Phosphorylation · SVMs · Prediction · b-Factors · Solvent accessibility

35.1 Introduction

The importance of understanding protein structure and function stimulates the development of improved methods for predicting intermediate structural and functional attributes of proteins. One very successful framework for such efforts involves knowledge-based approaches that extrapolate from known (experimentally characterized) examples using statistical and machine learning methods. Successful examples of this approach include secondary structure [1, 2], solvent accessibility [2, 3], interaction sites [4, 5], phosphorylation [6 16], and other functional hotspots prediction. General principles of such an approach in the context of the problem considered here, i.e., that of predicting sites at which proteins undergo phosphorylation, consist of first identifying a set of attributes to represent the amino acid sequence, which then form the input to a machine learning-based prediction method. Standard machine learning techniques, such as Neural Networks (NNs) or SVMs, can be used to solve the underlying regression or classification problem [2, 4, 9, 17]. Both sequence-based and (when applicable) structure-based attributes (features) can be used as input for the predictor. Certain intermediate structural and functional predictions can also serve as input features to subsequent predictors of other attributes. The key to a good prediction method usually lies in the choice of appropriate representation, capable of capturing structural and functional characteristics to be predicted. One very successful and commonly used representation involves constructing an evolutionary profile of a family if related (homologous) proteins, as encoded by PSSM representation of multiple alignment [1, 2, 18]. The latter is typically computed using Psi-BLAST or related programs [19].

The three-dimensional (3D) structure of proteins can be determined using X-ray crystallography [20]. However, atoms in proteins are always in thermal motion, and this motion manifests as weakened diffraction intensities and fuzzier electron density maps [18, 20]. These uncertainties are represented by the Debye Waller or temperature factors (also referred to as B-factors) [21], which are usually provided with the X-ray structures. The atomic mobility, which is quantified by B-factors, has been shown to play a vital role in various biological processes including molecular recognition, protein protein and protein DNA interactions and catalytic activity [22 30].

Given the importance of conformational flexibility for protein dynamics and function, several methods have been proposed to predict B-factors from sequence. In particular, Yuan et al. proposed a method that employed an SVR model with PSSMs from multiple alignments as input features [31], whereas Rost et al. developed a neural network-based protocol and PSSMs augmented by three-state secondary structure and two-state solvent accessibility predictions [18]. Recently, Kurgan et al. proposed a method that employs real-valued predictions of solvent accessibility as additional feature [17]. In this study, we develop and systematically assess a similar predictor for B-factors, which employs a combination of PSSMs and probabilistic predictions of SS and RSA as input features. Toward this we present an epsilon-SVR-based method and highlight the improvement that can be obtained by combining these features. Predictions of B-factor can then provide insights into protein functions including phosphorylation, contact number, and catalysis. B-factor predictions have also been studied in the context of protein conformational disorder [17, 24, 28].

The other commonly used method for determining 3D structures of proteins is Nuclear Magnetic Resonance spectroscopy (NMR) [32]. Typically, an ensemble of NMR structures (order of 10), are determined for each chain from which equilibrium atom coordinates can be determined. Along the lines of B-factors, a parameter referred to here as Solvent Accessibility Standard Deviations (SASDs) was devised to quantify the flexibility derived using NMR structures. The objective here was to devise an easy to compute parameter to represent conformational flexibility derived from NMR structures. In general, flexible regions would be more exposed to the solvent and show more variance in exposure in an ensemble of conformations. SASDs are easily derived as described below and are by nature rotationally invariant. The DSSP tool is used to generate the true solvent accessibility of the atoms in a given NMR structure and from the ensemble; one can derive standard deviations for the same. Toward this, we have developed epsilon-insensitive SVR predictors that can predict the SASDs as derived from NMR crystallographic studies using as input features PSSMs and probabilistic predictions of SS/RSA. To the best of our knowledge, the use of SASDs and the development of a predictor for this parameter have not been studied before. The applications of these predictions are the same as B-factors and their performance over different feature representations and in comparison with B-factors has been presented here.

Phosphorylation is arguably the most common post-translational modification in proteins. Since a phosphate group has a negative charge, phosphorylation/dephosphorylation changes the charge of the protein, resulting in a change in its structure and consequently its function. These phosphorylation/dephosphorylation events trigger the activation/deactivation of many signal transduction pathways [33 38]. Further evidence of its importance is the fact that kinases represent the second most popular category of drug targets [35].

Experimental mapping of all phosphorylation sites is hampered by the transient nature of phosphorylation, relatively small change in the overall mass and other properties, and fundamental limitations of protein chemistry [39]. Consequently, computational approaches that can predict potential sites of phosphorylation or at

the least limit the number of sites to be studied experimentally are of great interest. It is known that both sequence and structural motifs around phosphorylation sites play a major role in substrate recognition by kinases [40] and on this basis, many prediction methods have been devised using attributes derived from both sequence and structure of putative targets. These methods typically cast the problem as a two-class supervised machine learning (classification) problem for differentiating phosphorylated and nonphosphorylated sites. Examples of this approach are methods by Blom et al. (NetPhos and NetPhosK) and Berry et al. that use neural networks, Kim et al. (PredPhospho) and Gnad et al. (PHOSIDA group) that use SVMs and Iakoucheva et al. that uses a linear regression based model [6 10, 41]. Thus, these methods are based on the assumption that experimentally determined phosphorylation sites, which have been assimilated in databases like Phosphobase and Phospho. ELM, represent a set of positive sites whereas all other serine/threonine/tyrosine (S/T/Y) sites can be treated as negative examples [11 13, 42 44].

The methods mentioned above are likely to be affected by what is termed as the "negative-sites problem." The set of negative sites may contain many S/T/Y sites which get phosphorylated under some conditions and simply have not been experimentally annotated as such. Such sites have the potential to affect the performance of the methods that extrapolate from both positive and negative examples to construct a decision boundary (a classifier). Methods that do not use the negative sites are also available, notably Scansite that employs a motif-based search algorithm to identify phosphorylation sites, and OCSPP that uses one-class SVMs for the same task [14 16], both kinase-specific prediction models.

One of the objectives of this study is to systematically compare a method that uses both positive and negative sites to one that uses only the positive sites, using the same representations (sets of attributes), training and validation sets, and consistent learning procedures. Toward this goal, we used two flavors of the robust, mathematical programming-based support vector machine (SVM) techniques, namely two- and one-class classifiers and their performance compared. Given that 95% of phosphorylation occurs on Serine residues [33] and that most of the experimental data is on Serine residues, this study is restricted to the predictors for Serine phosphorylation. The objective of a one-class SVM, as introduced by Schölkopf et al. is to determine a decision boundary built around the members of one class, termed "targets," that can differentiate them from nonmembers, termed "outliers," represented by the origin [45]. By varying the parameters of the SVM, it is possible to tighten/loosen the decision boundary around the training examples for better performance on the test sets. For details about one-class SVMs please refer to [16, 45 47]. Three different feature sets have been analyzed in this study with features used in the methods mentioned above in order to obtain a wider assessment of the performance. In particular, we evaluated the importance of predictions from sequence (as real values) of secondary structure and solvent accessibility, as input features for the prediction of phosphorylation. Further, we investigate the use of the B-factor and SASD predictions developed in the first part of this study as additional input features toward the prediction of phosphorylation. The idea behind the features mentioned here is that sites of phosphorylation need to be accessible to

kinases (solvent accessibility) and are mostly found in flexible regions (secondary structure, B-factors, and SASDs).

35.2 Materials and Methods

35.2.1 Prediction of B-factors and NMR Solvent Accessibility Deviations

35.2.1.1 Datasets

Data for this study was obtained from the Protein Data Bank [48]. The first step was to identify protein chains in PDB that had three-dimensional structure data from both X-ray crystallography and NMR spectroscopy (NMR). This was achieved by searching the "Method" field in the PDB data files and matching chains from X-ray data with chains from NMR structures that had at least 90% sequence homology as determined by BLAST [49]. Once the protein chains of interest were identified, the sequences were pair-wise aligned against each other with goal of removing redundancy caused by sequence homology. This exercise yielded 305 protein sequences and corresponding NMR and X-ray structures, from which the B-factors and NMR SASDs were obtained. The NMR SASDs were obtained by calculating the standard deviations of the solvent accessibility values from the different NMR structures, facilitated by DSSP [50].

35.2.1.2 Features

Three different input feature sets were employed in this study. The first feature set consisted of PSSMs generated from multiple alignment against a nonredundant database, using Psi-BLAST [19]. This feature set is based on the one employed by Yuan et al. [31], in which each residue contributes 21 features and in total W times 21 for a window size W. The second feature set is composed of the secondary structure and relative solvent accessibility predictions (SS/RSA) from SABLE [2]. SABLE has been shown to provide state-of-the-art accuracy for the prediction of SS and RSA from protein sequence [2]. This feature set is similar to the one used by Kurgan et al. [39] and contributes five features for every residue viz. the probabilistic representation of the residue to form a Helix (H), Sheet (E), or Coil (C) structure and two relative solvent accessibility features, the probabilistic representation and the value representing the confidence. In general, conformationally rigid domains in proteins tend to be relatively buried and less exposed to the solvent [30]. Hence, solvent accessibility can be considered as a parameter of conformational flexibility and as an important input feature to its prediction [17].

Solvent accessibility predictors, specifically SABLE, define RSA of an amino acid residue as the ratio of the residue's solvent-exposed surface area (SA) in a given structure, which can be calculated using DSSP, and the maximum attainable value (in an extended conformation) of the particular amino acid's solvent-exposed surface area (MSA) [2].

The third feature set is a combination of the first two feature sets with the goal of assessing its advantage. Different window sizes were studied and a size of 15 adopted for the final predictors. It should be noted that B-factor values are dependent on the experimental considerations including resolution and refinement and hence are not absolute in nature [18]. Analyzing B-factors, especially predicting them, hence necessitates employing a normalization procedure that allows one to define relative (rather than absolute) and chain-specific measures of flexibility. Therefore, prior to use the B-factors undergo a per chain normalization (PCN), done here by using *z*-scores [17] where the mean (E) and standard deviation (SD) are calculated for the chain containing the residue. In the case of SASDs, both PCN and a global normalization (GN) have been compared to assess the transferability of SASD as an absolute measure of flexibility. Both use the same formula [17], except in GN the mean and standard deviation are calculated over the entire dataset and not just the chain containing the residue. The use of the PCN technique implies that the predictions provided by these predictors are also relative measures.

35.2.1.3 Regression Protocol

The B-factor and NMR SASD predictors in this study were developed using an epsilon-insensitive Support Vector Regression (ε-SVR) model as implemented in the software tool LIBSVM [51]. Linear and nonlinear radial basis function (RBF) kernels were employed with the latter being adopted due to better performance. The parameters of this regression model viz. cost, gamma and epsilon, were varied to optimize the performance. A fivefold cross-validation scheme was devised for testing these regression models developed and their performance was evaluated using Pearson's correlation.

35.2.2 Prediction of Phosphorylation

35.2.2.1 Datasets

Phosphorylation data for this study was obtained from the Phospho.ELM database [43]. Each sequence in the database was pair-wise aligned against every other sequence using Blast, in order to reduce any bias due to homology [49]. Groups of sequences that had high homology (10^{-6}) were identified and only one member of that group was retained. However, those homologous sequences in which the sites of phosphorylation were different were retained. Since the number of negative sites

was one order higher than the number of positive sites, only negative sites in sequences that had other homologous sequences (and were excluded) were considered. In other words, only those sites in sequences that were homologous and not reported as phosphorylated in any sequence in that group were included, in an effort to identify representative negative sites. Altogether, the process yielded 868 positive and 1,600 negative sites.

35.2.2.2 Features

Three different feature sets, including features previously used in other predictors, were employed in this study. The first feature set consisted of position-specific scoring matrices (PSSMs) generated from multiple alignment against a nonredundant database, using Psi-BLAST [19]. In this feature set each residue contributes 20 features, which in total would be W times 20 for a window size W. The second feature set contained all the features in the first as well as secondary structure and relative solvent accessibility predictions (SS and RSA) from the SABLE server [2]. This kind of representation (PSSMs + real value SS & RSA) is novel in this context, although it has been considered in other contexts [4, 52], and has been independently used by Gnad et al. [10] who also exploit the predictions of SABLE. We evaluate this representation further. For every residue, five features were extracted viz. the probabilistic representation of the residue to form a Helix (H), Sheet (E), or Coil (C) structure, the probabilistic representation of RSA, and the value representing its confidence. Each residue in this set has 25 features associated with it. The third feature set consisted of averaged amino acid properties (AAA) viz. hydrophobicity (using the Kyte Doolittle scale [53]), charge, and the count of certain groups of residues viz. {P}, {G}, {S,T}, {N,Q}, {K,R,H}, and {E,D}, combined with SS and RSA features described above. The features representing the different counts of residues were also normalized by their natural frequency of occurrence [54]. Window sizes 9, 11, 13, and 15 were studied and 11 adopted for better performance.

Additionally, the predicted values of B-factors and NMR SASDs from the SVR models presented here were assessed as input features toward prediction of phosphorylation. Toward this, the mentioned SVR predictions were combined, individually and in combination, with the features from PSSMs and probabilistic predictions from SABLE to form additional feature sets and their performance compared to the other feature sets.

35.2.2.3 Classification Protocol

The predictors in this study used the supervised machine learning (classification) technique of SVM as implemented in the software tool LIBSVM [51]. For this study, both two-class and one-class types of SVMs and for each type, the linear and the nonlinear radial basis function (RBF) kernels were used. Different parameters

of the SVMs viz. cost, gamma, and nu, were varied to maximize the performance. For training and testing on the datasets used in this study, a fivefold cross-validation scheme was employed. In order to evaluate the predictors of phosphorylation, the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) were first determined. Using these values the two performance measures of Accuracy (A) and Mathews Correlation Coefficient (MCC) were calculated as shown below.

$$A = \frac{TP + TN}{TP + FP + TN + FN} \tag{35.1}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{35.2}$$

35.3 Results and Discussion

35.3.1 Prediction of X-ray B-factors and NMR SASDs

Table 35.1 contains the results generated from the testing of ε-SVR prediction methods for X-ray B-factors and NMR SASDs. In the case of B-factor prediction, the probabilistic SS/RSA representation performs slightly lower than the PSSM one and clearly both representations are surpassed by their combination. The performance of the PSSM representation matches that of the SS/RSA representation, in the case of SASD prediction using PCN as well as GN, with the combined representation again showing significant improvement. The merit of using both the feature sets together is evident in these predictions. It can also be observed that the performance of the SASD predictions (using PCN) are higher than those of B-factors, suggesting that it is relatively easier to predict SASDs, a finding that can be exploited in its applications. While comparing the two normalization schemes in SASD predictions, one can observe that the PCN does better than GN. This suggests that it is difficult to capture overall rigidity/flexibility of different types of proteins.

Table 35.1 Performance of the nonlinear ε SVR models for the prediction of X ray B factors and NMR SASDs, on different feature sets

Feature set	X ray B factor prediction (CC) using PCN	NMR SASD prediction (CC) using PCN	NMR SASD prediction (CC) using GN
PSSM	0.46 (0.02)	0.49 (0.006)	0.45 (0.007)
SS/RSA	0.45 (0.008)	0.49 (0.006)	0.45 (0.008)
PSSM + SS/RSA	0.48 (0.005)	0.51 (0.003)	0.47 (0.007)

Values are averaged over cross validation and standard deviations are in parentheses

These enhanced predictors of structural properties, especially the novel parameter of SASDs, can be applied to other predictions of protein functions, e.g., phosphorylation as discussed in the next section and protein conformational disorder, a course of future research.

35.3.2 Prediction of Phosphorylation

35.3.2.1 Comparison of One- and Two-Class SVMs

The results from the study of one and two-class SVM protocols towards the prediction of phosphorylation are presented in Table 35.2. From this table the following observations can be made. Firstly, the two-class SVM performs significantly better than its one-class counterpart across all feature sets and in both the linear and nonlinear kernel versions, including the set AAA + SS/RSA not presented here. This strongly suggests that the training of predictors of phosphorylation is facilitated by the presence of the negative class. Secondly, it can be observed across all feature sets that, while the two-class nonlinear kernel performs significantly better than the linear kernel, the one-class nonlinear kernel performs poorer than its linear version. This is a very interesting observation and suggests that the presence of the other class of training examples is more needed when employing relatively complex kernels that necessitate the training of more number of parameters. However, in the future with the availability of more experimental data it may be possible to more effectively employ one-class methods.

35.3.2.2 Significance of Conformational Flexibility Parameters

Focusing on the different feature sets in Table 35.2, one can observe that the combined feature set of PSSM and SS/RSA feature set performs significantly better than the other sets. This is consistent with the results from an independent study by Gnad et al. [10]. With this feature set as reference, the performance of prediction with the addition of the conformational flexibility parameters was studied in this

Table 35.2 Performance of one and two class SVM methods on the different feature sets and kernels

Feature set	SVM kernel	Two class SVM		One class SVM	
		Accuracy	MCC	Accuracy	MCC
PSSM	Linear	72.1 (2.09)	0.37 (0.04)	70.5 (2.50)	0.31 (0.06)
	Nonlinear	76.7 (1.93)	0.47 (0.04)	65.7 (1.10)	0.15 (0.04)
PSSM + SS/RSA	Linear	73.1 (1.71)	0.4 (0.04)	71.8 (1.17)	0.34 (0.03)
	Nonlinear	78.0 (1.48)	0.51 (0.03)	66.7 (1.72)	0.17 (0.06)

Values are averaged over cross validation and standard deviations are in parentheses

Table 35.3 Assessment of performance of a two class nonlinear phosphorylation predictor using B factors and SASDs as input features

Feature set	Accuracy	MCC
PSSM + SS/RSA	78.0 (1.48)	0.51 (0.03)
PSSM + SS/RSA + B factor + NMR SASD	82.3 (1.25)	0.51 (0.03)

Values are averaged over cross validation and standard deviations are in parentheses

paper and is presented in Table 35.3. We observed that the performance with the addition of B-factor and SASD predictions, individually, to the input space provides no improvement. However, addition of B-factors and SASD predictions in combination with the reference features significantly improves the prediction of phosphorylation as shown in Table 35.3. The use of NMR SASDs in improving the prediction of phosphorylation, as presented here, has not been previously demonstrated. These results highlight the importance of conformational flexibility, as represented by B-factors and NMR SASDs, to the identification of phosphorylation sites in proteins.

35.4 Conclusion

Prediction of structural and functional attributes of proteins has been of great interest to researchers in the last decade. The identification of the best input features for such prediction methodologies is vital to its performance. In this study, we focused on two widely used representations, namely PSSMs derived from multiple sequence alignments and real-valued secondary structure and solvent accessibility predictions (SS/RSA), with the goal of enhancing the predictions of B-factors derived from X-ray structures, SASDs derived from NMR structures, and protein phosphorylation. In the case of B-factors, we have shown that a combination of PSSMs and real-valued SS/RSA predictions can significantly improve the prediction of B-factors and have presented an epsilon-insensitive support vector regression (ε-SVR) model toward this. Similarly, we have also developed an ε-SVR predictor for SASDs, which to the best of our knowledge has not been utilized before in the context of protein structure and function prediction. This rotationally invariant and easy to compute measure can effectively capture protein conformational flexibility in solution, and as demonstrated in this work, can be predicted reliably and used to subsequently improve the prediction of phosphorylation.

In the case of prediction of phosphorylation, the use of PSSMs and SS/RSA predictions has been proposed before by others [10, 16]. Here we further assess the potential of these features, showing the improvement in performance that can be obtained by use of real-valued SS/RSA predictions. Furthermore, the novel use of B-factors and SASDs, in combination, has been shown to significantly improve the prediction of phosphorylation. These results highlight the potential for use of enhanced B-factor and SASD predictions toward various other predictors, one of

which viz. protein disorder is something we are studying currently. We have also developed and compared the performance of one- and two-class SVM-based phosphorylation predictors for several commonly used sets of input attributes. Although one-class SVMs were recently reported as a possible solution to this problem [24] they were not compared systematically to two-class strategies. In this regard, our conclusion is that the negative examples do play a significant role in the prediction process and that the two-class method performs better than the one-class in all cases studied. Also, it is evident that when using one-class strategies, the learning models (here we compared explicitly linear and nonlinear kernels within SVM framework), should be kept as simple as possible for better performance. The improvement in prediction of phosphorylation by the employment of enhanced conformational flexibility parameters, as demonstrated here, is motivation for similar investigations of other structural and functional properties in proteins.

Acknowledgments This work was supported in part by NIH grants GM067823 and P30 ES006096. Computational resources were made available by Cincinnati Children's Hospital Research Foundation and University of Cincinnati College of Medicine.

References

1. Jones DT (1999) Protein secondary structure prediction based on position specific scoring matrices. J Mol Biol. 292(2):195 202.
2. Adamczak R, Porollo A, Meller J (2004) Accurate prediction of solvent accessibility using neural networks based regression. Proteins 56(4):7537 7567.
3. Pollastri G, Baldi P et al (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins 47(2):142 153.
4. Altschul SF, Lipman DJ et al (1990) Basic local alignment search tool. J Mol Biol 215:403 410.
5. Liu B, Wang X et al (2009) Prediction of protein binding sites in protein structures using hidden Markov support vector machine. BMC Bioinformatics. 10:381.
6. Blom N, Brunak S et al (1999) Sequence and structure based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294(5):1351 1362.
7. Blom N, Brunak S et al (2004) Prediction of posttranslational glycosylation and phosphoryla tion of proteins from the amino acid sequence. Proteomics 4(6):1633 1649.
8. Berry EA, Dalby AR, Yang ZR (2004) Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algo rithms. Comput Biol Chem 28(1):75 85.
9. Kim JH, Koh I et al (2004) Prediction of phosphorylation sites using SVMs, Bioinformatics 20(17):3179 3184.
10. Gnad F, Ren S, Cox J, Olsen JV, Macek B, Oroshi M, Mann M (2007) PHOSIDA (phosphor ylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. Genome Biol 8(11):R250.
11. Zhou FF, Yao X et al (2004) GPS: a novel group based phosphorylation predicting and scoring method. Biochem Biophys Res Commun 325(4):1443 1448.
12. Xue Y, Yao X et al (2008) GPS 2.0, a tool to predict kinase specific phosphorylation sites in hierarchy. Mol Cell Proteomics 7(9):1598 1608.
13. Dang TH, Laukens K et al (2008) Prediction of kinase specific phosphorylation sites using conditional random fields. Bioinformatics 24(24):2857 2864.

14. Yaffe MB, Cantley LC et al (2001) A motif based profile scanning approach for genome wide prediction of signaling pathways. Nat Biotechnol 19(4):348 353.
15. Obenauer JC, Yaffe MB et al (2003) Scansite 2.0: Proteome wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31(13):3635 3641.
16. Li T, Fu H, Zhang X (2007) Prediction of kinase specific phosphorylation sites by one class SVMs. IEEE Int Conf Bioinform Biomed 217 222.
17. Kurgan L, Zhang H et al (2009) On the relation between residue flexibility and local solvent accessibility in proteins. Proteins 76(3):617 636.
18. Schlessinger A, Rost B (2005) Protein flexibility and rigidity predicted from sequence. Proteins 61(1):115 126.
19. Altschul SF, Lipman D et al (1997) Gapped BLAST and PSI BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389 3402.
20. Blow D (2002) Outline of crystallography for biologists. Oxford University Press, New York, p 237.
21. Debye P (1913) Interferenz von Röntgenstrahlen und Wärmebewegung (in German). Ann d Phys 348(1):49 92.
22. Dunker AK, Obradovic Z (2001) The protein trinity linking function and disorder. Nat Biotechnol 19:805 806.
23. Liu J, Tan H, Rost B (2002) Loopy proteins appear conserved in evolution. J Mol Biol 322:53 64.
24. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6:197 208.
25. Yuan Z, Wang ZX et al (2003) Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng 16:109 114.
26. Teague SJ (2003) Implications of protein flexibility for drug discovery. Nat Rev Drug Discov. 2:527 541.
27. Daniel RM, Smith JC et al (2003) The role of dynamics in enzyme activity. Annu Rev Biophys Biomol Str 32:69 92.
28. Dunker AK, Obradovic Z et al (2002) Intrinsic disorder and protein function. Biochemistry 41:6573 6582.
29. Tobi D, Bahar I (2005) Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state. Proc Natl Acad Sci USA 102:18908 18913.
30. Bhalla J, Storchan GB et al (2006) Local flexibility in molecular function paradigm. Mol Cell Proteomics 5:1212 1223.
31. Yuan Z, Bailey TL, Teasdale RD (2005) Prediction of protein B factor profiles. Proteins 58(4):905 912.
32. Grzesiek S, Sass HJ (2009) From biomolecular structure to functional understanding: new NMR developments narrow the gap. Curr Opin Struct Biol 19(5):585 595.
33. Siegel GJ, Agranoff BW et al (1998) Basic Neurochem. Molecular Cellular & Medical Aspects. LWW Publishers.
34. Structural Genomics Consortium website. http://www.sgc.ox.ac.uk/research/pds.html.
35. Cohen P (2002) Protein kinases the major drug targets of the twenty first century?. Nat Rev Drug Discov 1(4):309 315.
36. Sardari S, Nam NH et al (2003) Protein kinases and their modulation in the central nervous system. Curr Med Chem 3(4):341 364.
37. Lu KP, Zhou XZ et al (2002) Pinning down proline directed phosphorylation signaling Trends Cell Biol 12(4):164 172.
38. Secko DM. Protein Phosphorylation: http://www.bioteach.ubc.ca/CellBiology/ProteinPhosphorylation/index.htm.
39. Manning G, Whyte DB et al (2002) The protein kinase complement of the human genome. Science 298(5600):1912 1934.
40. www.neb.com/nebecomm/tech reference/protein tools/protein kinase substrate recognition. asp.

41. Iakoucheva LM, Dunker AK et al (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32(3):1037 1049.
42. Kreegipuu A, Blom N et al (1999) PhosphoBase, a database of phosphorylation sites. Nucleic Acids Res 27(1):237 239.
43. Diella F, Gibson TJ et al (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics 5(1):79.
44. Rychlewski L, Reimer U et al (2004) Target specificity analysis of the Abl kinase using peptide microarray data. J Mol Biol 336(2):307 311.
45. Schölkopf B, Smola AJ (2002) Learning with kernels. MIT Press.
46. Chen Y, Zhou X et al (2001) One class svm for learning in image retrieval. Proc Intl Conf Image Process 1:34 37.
47. Manevitz LM, Yousef M (2002) One class svms for document classification, J Mach Learn Res 2:139 154.
48. Berman HM, Bourne PE et al (2000) The Protein Data Bank. Nucleic Acids Res 235 242.
49. Altschul SF, Lipman DJ et al (1990) Basic local alignment search tool. J Mol Biol 215:403 410.
50. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers. 22(12):2577 2637.
51. Chang C C, Lin C J (2001) LIBSVM: a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
52. Cao B, Meller J et al (2006) Enhanced recognition of protein transmembrane domains with prediction based structural profiles. Bioinformatics 22(3):303 309.
53. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105.
54. Available online at: http://folding.chmcc.org/online/am acids/am acids.html.

Chapter 36
Why Are MD Simulated Protein Folding Times Wrong?

Dmitry Nerukhdn

Abstract The question of significant deviations of protein folding times simulated using molecular dynamics from experimental values is investigated. It is shown that in the framework of Markov State Model (MSM) describing the conformational dynamics of peptides and proteins, the folding time is very sensitive to the simulation model parameters, such as forcefield and temperature. Using two peptides as examples, we show that the deviations in the folding times can reach an order of magnitude for modest variations of the molecular model. We, therefore, conclude that the folding rate values obtained in molecular dynamics simulations have to be treated with care.

Keywords Molecular dynamics · MSM · Protein folding · Simulation models · VPAL

36.1 Introduction and Overall Methodology

Modern computational power is enough to simulate small proteins up to the times when they fold into their native conformations [1 4]. This is a remarkable achievement because phenomenological, relatively simple interatomic interactions built into the model lead to the molecular structures that essentially coincide with the crystallographically determined native conformations.

In contrast to the structure of proteins, the results on folding times are not as optimistic. For the majority of successfully folded proteins, there are significant discrepancies between simulated and experimental folding times [5 7]. This is taking into account that only the results obtained when the trajectories approach

D. Nerukhdn
Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, Cambridge University, Cambridge CB2 1EW, UK
e mail: 232@cam.ac.uk

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 36,

the folded conformations sufficiently close are published. In few cases, even complete failures to reach the folded state in silico in simulations significantly exceeding the experimental folding times are reported [8]. Indeed, it is well known in the modelling community how difficult it is to fold a protein ab initio, that is without introducing any information on the intermediates.

By analysing the MD trajectories of peptides in explicit water, we suggest an explanation for these discrepancies. We show that the folding rates are very sensitive to the details of the simulation model. The sensitivity is so high that the obtained values of folding times are meaningless and cannot be compared between each other and with the experiment.

We use the MSM [9 12] to describe the folding process. The configurational states are defined by clustering the MD simulated trajectories. This is done by analysing the Ramachandran plots of the residues of the peptide, Fig. 36.1. Each Ramachandran plot is clustered independently and the molecule's configurations are defined by the cluster indices from each plot. Not all possible combinations of index values are realised in the trajectory. For example, for the peptide from Fig. 36.1, the conformation B_1C_2 was very scarcely populated and was, therefore, joined with A_1C_2 into one conformation, thus resulting in five total configurations of the molecule.

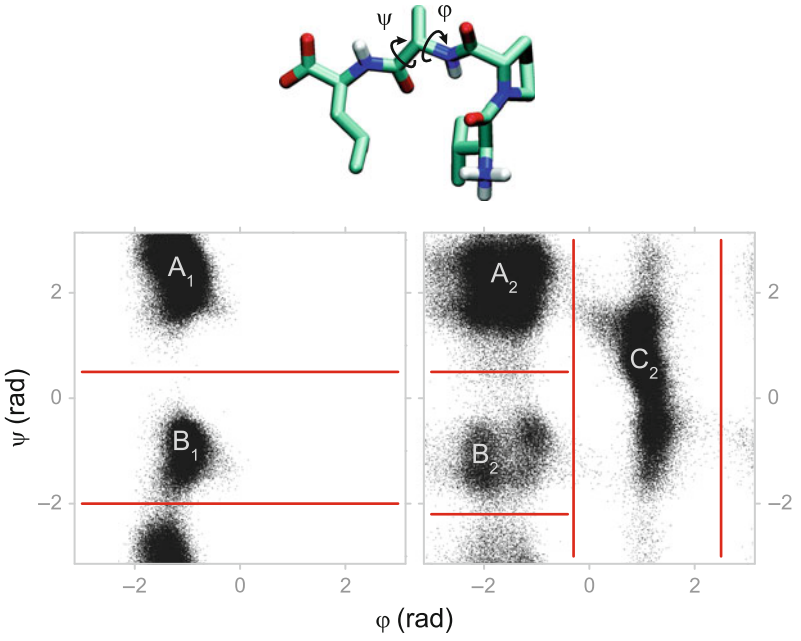


Fig. 36.1 Four residue peptide VPAL and the Ramachandran plots for the Proline (*left*) and Alanine (*right*). The clustering is marked by the boundaries that define the conformations as pairwise combinations of the indices from the sets $\{A_1, B_1\}$ and $\{A_2, B_2, C_2\}$

In the MSM framework, the model is described by a state vector v, which holds probabilities of all the configurations at a given moment of time, and a transition matrix T. The total probability of the state vector has to sum to 100% since the peptide has to be in some configuration at any time. The transition matrix holds the probability that the system is transferred from one state to another at the next time step. Because the total probability of the state vector has to be conserved, the requirement $\sum_i T_{ij} = 1$ is imposed, where i and j run over all states. Given that the system has a state vector v_t at time t, the state vector at the time $t + \Delta t$ can be calculated as $v_{t+\Delta t} = Tv_t$. The property of a transition matrix is such that its eigenvalues λ_i are in the range 0 1 with one eigenvalue being 1. In the following, it is assumed that the eigenvalues are ordered in descending order so that $\lambda_0 = 1$.

By expanding the transition matrix in terms of the left $|\lambda_i\rangle$ and right $\langle\lambda_i|$ eigenvectors, the time evolution of the system is given by

$$v_{t+n\Delta t} = T^n v_t = \sum_i \lambda_i^n |\lambda_i\rangle\langle\lambda_i|v_t. \tag{36.1}$$

This representation immediately provides information on the behaviour of the system under investigation.

First, by analysing how the eigenvalues vary with time step Δt, it can be determined whether the dynamics of the system can actually be described by a Markov model [13]. We have shown [13] that the dynamics of a four-residue peptide become Markovian (that is, the next time step conformation depends on the current conformation only) at the time scale $\Delta t \approx 50$ ps. For a larger peptide of 15 residues, our recent estimations confirm the same time scale of Markovian behaviour. From general considerations, it is reasonable to assume the same time period of "loosing memory" for the dynamics of larger peptides and small proteins.

Second, both the folded state and the folding time are readily obtained from (36.1). Indeed, at the limit $n \to \infty$ only the largest eigenvalue, equal to 1, survives while all other eigenvalues, being less than 1 tend to 0. Therefore, the eigenvector $|\lambda_0\rangle$ corresponds to the equilibrium distribution of conformations, the folded state. The speed at which the system approaches the equilibrium distribution is described by all other eigenvalues that are less than 1. Again, at $n \to \infty$ the second largest eigenvalue dominates since it describes the slowest convergence in the system, while all smaller valued eigenvalues become negligible. Thus at this limit, λ_1 defines the folding rate.

The transition matrix T can vary in the simulation due to, for example, the differences in the forcefield or the variation in the macroscopic parameters of the system: temperature, simulation box size (number of molecules), etc. The force field differences are the result of phenomenological nature of the classical molecular dynamics potentials as well as different calibration criteria. The other sources of the changes are various alterations of the interaction potentials that are aimed at speeding up the folding, and are becoming increasingly popular lately [14 17].

Similarly, temperature variations are the cornerstone of such widespread technique as Replica Exchange MD. These too can affect the transitions T.

Our goal is to investigate what happens to the folding rate if the transition matrix is changed. In our framework, it means that the effect of the variation of T to λ_1 has to be determined. Here, we assume that the matrix changes do not affect the folded state, that is, the eigenvector $|\lambda_0\rangle$ remains the same[1].

To understand the meaning of the λ_1 variation, $\delta\lambda_1$, in more intuitive terms of folding times, we introduce a "folding half time" measure as follows. Let us assume that after a large number of time steps, n, λ_1 reduces its value in half (36.1): $(\lambda_1^n/\lambda_1) = (1/2)$. Then, we designate this time as a "folding half time" that can be calculated as

$$n_{1/2} = \frac{\ln 2}{\ln \lambda_1}. \tag{36.2}$$

Suppose that we have changed the dynamics and, as a result, the eigenvalue λ_1 has changed by an amount $\delta\lambda_1$. The half time for this new eigenvalue is $n'_{1/2} = (\ln 2/\ln(\lambda_1 - \delta\lambda_1))$ (for an accelerated folding, $\delta\lambda_1$ is negative). The ratio $r = (n_{1/2}/n'_{1/2})$ gives us a representative measure of the sensitivity of the folding time to the changes in the transition matrix T. In other words, r is an amount by which the folding time is changed. Thus, the sensitivity in our description is a function of two parameters: the second largest eigenvalue λ_1 and its variations $\delta\lambda_1$ caused by the changes in the simulation model.

The rest of the paper is devoted to the estimations of the values of λ_1 and $\delta\lambda_1$ for representative protein systems simulated using MD.

Both parameters can be calculated directly from the transition matrices T obtained in the simulations of systems with varying parameters described above. We modelled the variations by performing the simulations of peptides and altering two parameters of the system: (a) scaling the masses of the atoms (corresponds to changing the forcefield) and (b) varying the temperature in the simulation (resembles the Replica Exchange MD conditions). The variation in the masses was done by the introduction of a unified parameter α, so that the new masses are αm: $(\alpha m)a = -(\delta V/\delta r)$ or $ma = -(\delta\frac{V}{\alpha}/\delta r)$. Therefore, varying the masses is equivalent to varying the potential energy, that is, changing the forcefield of the model. The changes in temperature were achieved using standard methods by simply setting the thermostat to different temperatures.

We have simulated exhaustively a four-residue peptide VPAL (Valine-Proline-Alanine-Leucine) and calculated both λ_1 and $\delta\lambda_1$ directly from the transition

[1]To the best of our knowledge, this requirement has never been checked in numerous methods aiming to accelerate the folding in MD (Accelerated Molecular Dynamics, Hyperdynamics, Replica Exchange Molecular Dynamics, etc. [14 17]). This almost inevitably leads to incorrect folded state obtained in these simulations (see [18] for details).

matrices by varying the parameters in both ways described above. We have obtained the value of λ_1 equal to 0.785. For $\delta\lambda_1$, VPAL produces the values of 0.073 when varying the scaling α in the range 0.75 1.25 and 0.161 when varying the temperature in the 280 320 K boundaries. These correspond to the ratios $r = 1.40$ for varying α and $r = 1.95$ for varying T. In other words, the folding times became almost twice as high in the different scenarios used.

It should be noted that for longer peptides, λ_1 is normally larger. This is not surprising since larger peptides have lower folding rate, that is, larger λ_1. Indeed, we have also analysed the folding trajectories of a 15-residue peptide with the sequence SESYIDPDGTWTVTE and obtained $\lambda_1 = 0.9915$. Assuming approximately the same value for $\delta\lambda_1$, say $\delta\lambda_1 = 0.1$, we obtain $r = 13.45$, that is, more than an order of magnitude increase in folding half time.

These results clearly demonstrate the high sensitivity of the folding times to the details of simulation models. It also seems reasonable to conclude that the sensitivity tends to be higher for larger peptides that fold slower. Therefore, the results on the folding times for larger realistic proteins would be even less reliable.

Since any force field is only approximately correct this means that calculated folding times are significantly inaccurate, although the folded state reached in the simulation is correct. It is therefore not meaningful to make a comparison between a simulated folding time and the one determined experimentally, especially for slowly folding proteins.

Acknowledgement The work is supported by Unilever and the European Commission (EC Contract Number 012835 EMBIO).

References

1. H. Lei and Y. Duan, Journal of Physical Chemistry B **111**, 5458 (2007), ISSN 1520 6106.
2. H. Lei and Y. Duan, Journal of Molecular Biology **370**, 196 (2007).
3. A. Suenaga, T. Narumi, N. Futatsugi, R. Yanai, Y. Ohno, N. Okimoto, and M. Taiji, Chemistry An Asian Journal **2**, 591 (2007), 10.1002/asia.200600385.
4. S. Gnanakaran, H. Nymeyer, J. Portman, K. Y. Sanbonmatsu, and A. E. García, Current Opinion in Structural Biology **13**, 168 (2003).
5. J. Kubelka, J. Hofrichter, and W. A. Eaton, Current Opinion in Structural Biology **14**, 76 (2004).
6. D. L. Ensign, P. M. Kasson, and V. S. Pande, Journal of Molecular Biology **374**, 806 (2007).
7. L. Tsai, H. Chen, T. Lin, W. Wang, and Y. Sun, Journal of Theoritical and Computational Chemistry **6**, 213 (2007).
8. P. L. Freddolino, F. Liu, M. Gruebele, and K. Schulten, Biophysical Journal **94**, L75 (2008).
9. C. Schuette, A. Fischer, W. Huisinga, and P. Deuflhard, Journal of Computational Physics **151**, 146 (1999), ISSN 0021 9991.
10. W. C. Swope, J. W. Pitera, and F. Suits, The Journal of Physical Chemistry B **108**, 6571 (2004).
11. J. D. Chodera, N. Singhal, V. S. Pande, K. A. Dill, and W. C. Swope, The Journal of Chemical Physics **126**, 155101 (2007).
12. F. Noe and S. Fischer, Current Opinion in Structural Biology **18**, 154 (2008).
13. C. H. Jensen, D. Nerukh, and R. C. Glen, The Journal of Chemical Physics **128**, 115107 (2008).
14. Y. Sugita and Y. Okamoto, Chemical Physics Letters **314**, 141 (1999).
15. X. Periole and A. E. Mark, Journal of Chemical Physics **126**, 11 (2007).

16. A. Baumketner and J. E. Shea, Theoretical Chemistry Accounts **116**, 262 (2006).
17. K. P. Ravindranathan, E. Gallicchio, R. A. Friesner, A. E. McDermott, and R. M. Levy, Journal of the American Chemical Society **128**, 5786 (2006).
18. C. H. Jensen, D. Nerukh, and R. C. Glen, The Journal of Chemical Physics **129**, 225102 (2008).

Chapter 37
Automatic TEM Image Analysis of Membranes for 2D Crystal Detection

Argyro Karathanou, Nicolas Coudray, Gilles Hermann, Jean-Luc Buessler, and Jean-Philippe Urban

Abstract TEM image processing tools are devised for the assessment of 2D-crystallization experiments. The algorithms search for the presence and assess the quality of crystalline membranes. The retained scenario emulates the decisions of a microscopist in selecting targets and assessing the sample. Crystallinity is automatically assessed through the diffraction patterns of high magnification images acquired on pertinent regions selected at lower magnifications. Further algorithms have been developed for membrane characterization. Tests on images of different samples, acquired on different microscopes led to good results.

Keywords Image processing · Automated TEM · Image segmentation · Sample characterization

37.1 Tools for Automatic Image Analysis

Tools for image analysis have gained growing interest in life sciences where the automation of repetitive tasks is mandatory. Automatic Image acquisitions and processing are important to accelerate biological research. This chapter is dedicated to a set of algorithms developed for automatic analysis of 2D biological membrane images acquired in Transmission Electron Microscopy (TEM).

These algorithms are built for images acquired at three levels of magnification at which target regions should be selected and at which the sample should be characterized:

1. At low magnification, the grid quality is evaluated (the carbon film covering the copper mesh grids can be locally broken).

G. Hermann (✉)
MIPS, Université de Haute Alsace, France
e mail: gilles.hermann@uha.fr

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 37,

2. At medium magnification, two goals are achieved: the specimen characterization (size, shape, and stacking), and Regions of Interest (ROI) selection (the largest non-staked membranes represent the potential crystalline regions).
3. At high magnification, images of potentially crystalline regions are acquired, and the study of the diffraction pattern evaluates the success of the crystallization experiments.

The next section introduces the particular context of our study. Sections 37.3 37.5 present the tools developed at the three different magnifications. For clarity, results are given within each section.

37.2 Context

Proteins are important to understand as they constitute major drug targets. Protein structural data can be obtained by several methods. The main method, 3D crystallization for X-ray crystallography, has led to the determination of several structures, and benefits from automated tools that help in finding the protocols that lead to protein 3D crystallization [1, 2]. However, membrane proteins, difficult to process using the 3D-crystallization technique, seem to be more adapted to the 2D-crystallization technique, which consists of embedding proteins within bi-lipidic layers [3].

Many ad hoc tests are required to identify the conditions that, for a given protein, would give large sheet-like crystallized membranes. The HT-3DEM (high throughput-three-dimensional electron microscopy [4]) project, in which this work is included, aims to develop a tool chain to generate and analyze a large amount of 2D-crystallization attempts automatically.

The success of 2D-crystallization experiments is assessed by observing the samples using a transmission electron microscope [5]. During manual observations, the microscopist screens the unbroken regions looking for potential crystalline membranes, and then magnifies the selected targets to check at high resolution if a diffraction pattern is visible.

Some existing software packages [6, 7] are capable of automatic microscope control (eucentric height, autofocusing, stage displacement to given targets, image acquisition, etc.), but the choice of the on-line targets and analysis of the images still rely on humans.

The algorithms presented intend to emulate the microscopist by selecting interesting targets and by characterizing the specimen.

37.3 Low Magnification Analysis

To assess the grid quality, a new technique based on histogram analysis and local background analysis has been developed [8]. The principle is to measure the mean gray-level of the background (the last histogram peak for each square): when it is

high, the carbon film is broken, and when it is low, the carbon film is not locally damaged. The objective is to identify a level T_s corresponding to the threshold separating the two types of background.

Four classes of grid squares, represented in Fig. 37.1 (left), can be distinguished. Classes A and B correspond to a good quality grid. The goal is to select grid squares of these two classes with a minimum of false positives. For class A, membranes are visible: the convex area of the background covers the whole grid square and contains holes corresponding to large and contrasted membranes. Class C represents grid squares where the carbon film is locally broken: the background covers only a small portion of the grid square. However, based on the background segmentation, we cannot differentiate grid squares of class B (membrane not visible) from those of class D (no carbon film at all).

The proposed classification is therefore composed of two steps. First, orifices from classes A and C are identified and used to extract the background gray-level statistics: the mean gray-levels and standard deviations of the background when the carbon film is present (noted $\overline{\mathrm{GL}}_{\mathrm{good}}$) and when the carbon film is absent (noted $\overline{\mathrm{GL}}_{\mathrm{broken}}$). Second, threshold T_s, used to separate grid squares of class B from those of class D, is calculated from these statistics using

$$T_s = \frac{\overline{\mathrm{GL}}_{\mathrm{good}} + \overline{\mathrm{GL}}_{\mathrm{broken}}}{2} \tag{37.1}$$

Examples of grid square selections are given in Fig. 37.1 (right). The objective is to select convenient grid squares for further analysis at medium magnification. Each grid possesses more than 100 grid squares, but only a few (around ten generally) need to be explored, the sample being homogeneous over the grid. It is more important to avoid false positives rather than trying to avoid false negatives, since exploring damaged grid squares imply a loss of time. In this optic, our algorithm performs very well since less than 1% of the classified squares are wrongly selected as interesting targets.

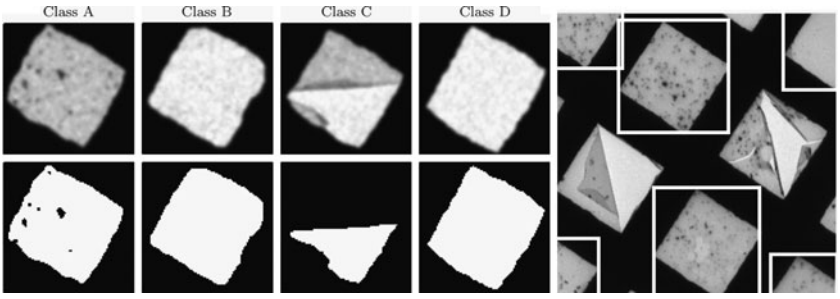


Fig. 37.1 On the *left*, four different classes of grid squares (raw 1: gray level images; raw 2: binary image after background segmentation); on the *right*, examples of grid square selections (surrounded by *white rectangle boxes*). Field of view: 360 × 360 μm

37.4 Medium Magnification Analysis

37.4.1 Membrane Characterization

Several algorithms are integrated and combined to realize the different steps leading to membrane characterization: image segmentation using a split-and-merge-like approach, background/foreground labeling, and, finally, characterization of the stacking level.

The characterization requires the contours of the membranes to be fully identified. A multi-scale edge detection algorithm has been developed [9] to split the image into regions delimiting membranes: edges are first extracted at different scales and then combined into a single image called the reconstructed gradient-like (RGL) image, which leads to a compromise between the edge detection and the precision of the edge positioning. The partition (Fig. 37.2(ii)) is obtained after applying the watershed algorithm [10] on the RGL image.

The tests done on images of different samples acquired in various conditions of illumination and focusing showed that the membranes are well delimited, even when the contrast is low. An algorithm based on statistical transition validation has been developed to validate the partition and reduce the induced over-segmentation [11]. The technique explores the profile transition between two objects of the partition and validates this transition if there is a gradient orthogonal to the contour that is statistically significant. Transitions below the threshold, defined using statistical hypotheses, are then iteratively removed using region merging (Fig. 37.2(iii)).

Following this split-and-merge segmentation, a first global semantic labeling aims to differentiate the background from the foreground regions. Figure 37.3(i) shows the background (in black) extraction applied on Fig. 37.2. Corrected partition facilitates this task based on the analysis of the size and mean gray-level of the regions [11]: we suppose that the background is represented by a single region,

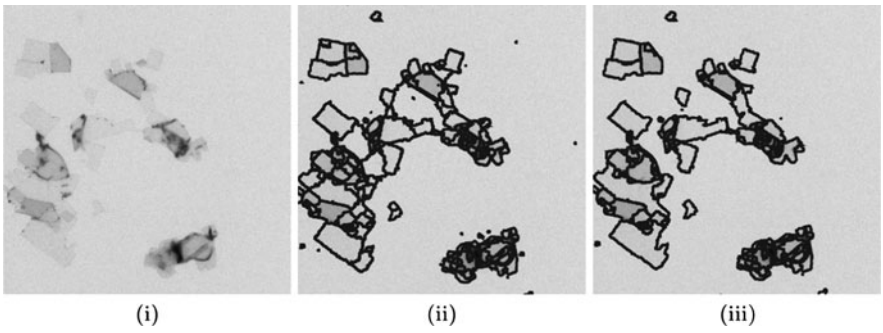
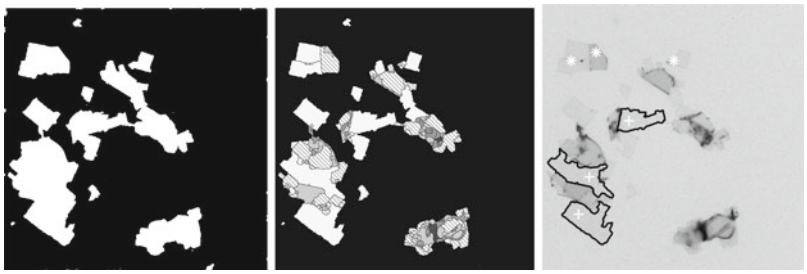


Fig. 37.2 Segmentation tools: (i) initial image; (ii) partitioned image after splitting; (iii) partitioned image after merging spurious transitions



(i) (ii) (iii)

Fig. 37.3 (i) Foreground background identification. (ii) Multi stack representation: foreground from white to gray according to the stacking (from non stacked to multi stacked membranes). (iii) ROI selection: example of selections using the PED process (*white stars*) and using the membrane characterization (*white crosses* in their centers and their contours are highlighted in *black*) based on the size and the gray level (the three largest non stacked membrane regions)

called the principal background region, associated to the biggest and brightest region. This region will be used as a reference by the stacking-level algorithm, presented below, which will then be able to identify secondary background regions for a total background extraction.

The semantic labeling was tested on various TEM images that differ by their aspect and attributes (type of microscope, acquisition conditions, sample natures etc.). The background region is well extracted in 88% of the TEM images.

Once these first steps are realized, characterization algorithms can be applied. This method aims to identify the stacking level. The method relies on the iterative measure of the local contrast between membrane regions and the nearest background area [12]. At each iteration, a gray-level quantity specific of a non-stacked membrane is defined and used to identify the stacking level of the regions. Figure 37.3(ii) illustrates typical results which, as can be seen by comparing with the initial gray-level images in Fig. 37.2(i), constitute a very good stacking representation. This classification allows now to extract the background in more images (up to 95% of TEM images) by detecting secondary background regions.

37.4.2 ROI Selection

Two algorithms are proposed to select ROI to be observed at high magnification for crystalline assessment.

The first solution, called partial edge detection (PED), avoids the necessity of an advanced characterization of the images by identifying some edges corresponding to potential low contrasted non-stacked membranes [8].

The second solution, more complete but more time consuming, is based on the algorithms presented in the previous section. It permits a more precise targeting,

based on the stacking level (non-stacked membranes are preferred), the size (the largest membranes), and/or the shape of the membrane regions. Figure 37.3(iii) shows ROI selected with the two approaches.

37.5 High Magnification Analysis

The algorithm used here aims to identify spots in the diffraction pattern of the acquired images [8].

Spot identification is made difficult by the ring-shaped noise due to the contrast transfer function (CTF) of the microscope. To tackle this problem, a three-step method is used: (1) the background generated by the CTF is evaluated using a mean radial profile; (2) this evaluation is used to correct the initial FFT image; and (3) the resulting image is segmented so that spots with signal to noise ratio above 3.5 are segmented (corresponding to a quality index of at least 2 according to the measure proposed by [13]). Figure 37.4 illustrates this method on an example. The algorithms perform efficiently, leading to about 97% of good classification (tests done on 236 diffraction patterns).

37.6 Discussion

We have presented a set of image processing algorithms developed for the automatic analysis of images of 2D-crystallization experiment samples. The algorithms run repetitive tasks for region selection and sample characterization in a fully automatic manner. The algorithms have been validated using hundreds of images acquired on different microscopes under usual conditions. Combined with a microscope control software package, the tools aimed at region targeting are applicable on-line. Interfaced with the Tecnai 12 (FEI Company) at the Biozentrum, Basel, some of these tools have already been successfully tested for automatic image acquisition of 2D crystals.

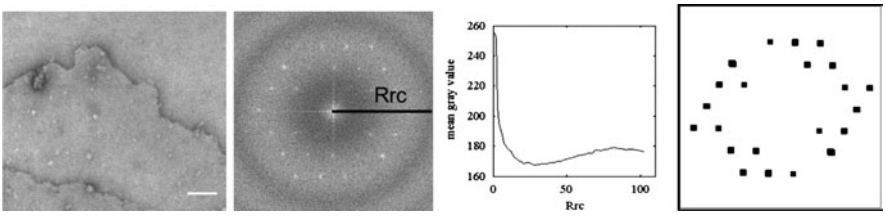


Fig. 37.4 From *left* to *right*: Image of a crystalline sheet membrane (scale bar 100 nm); Fourier Transform of the image and the radius Rrc; Mean radial profile; diffraction spots identified after the background subtraction and the thresholding steps

Thanks to the general approach upon which the ROI selection is based, the targeting algorithms could be used or easily adapted and parameterized for the automatic image acquisition of other types of specimens.

Acknowledgments This work has been supported by the EU sixth framework (HT3DEM, LSHG CT 2005 018811). We thank the Biozentrum of Basel and FEI company Eindhoven for the good collaboration and for providing the TEM images.

References

1. Wilson J (2004) Automated evaluation of crystallisation experiments. Crystallography Reviews 10(1):73 84.
2. Zhu X, Sun S, Cheng SE, Bern M (2004) Classification of protein crystallization imagery. EMBS 04, vol 3, pp 1628 1631.
3. Stahlberg H, Fotiadis D, Scheuring S, Rémigy H, Braun T, Mitsuoka K, Fujiyoshi Y, Engel A (2001) Two dimensional crystals: a powerful approach to assess structure, function and dynamics of membrane proteins. FEBS Letters 504(3):166 172.
4. HT 3DEM. http://www.ht3dem.org.
5. Signorell GA, Kaufmann TC, Kukulski W, Engel A, Rémigy H (2007) Controlled 2D crystallization of membrane proteins using methyl [beta] cyclodextrin. Journal of Structural Biology 157(2):321 328.
6. Cheng A, Leung A, Fellmann D, Quispe J, Suloway C, Pulokas J, Carragher B, Potter CS (2007) Towards automated screening of two dimensional crystals. Journal of Structural Biology 160(3):324 331.
7. Oostergetel GT, Keegstra W, Brisson A (1998) Automation of specimen selection and data acquisition for protein electron crystallography. Ultramicroscopy 74(1 2):47 59.
8. Coudray N, Buessler JL, Kihl H, Urban JP (2007) Automated image analysis for electron microscopy specimen assessment. EUSIPCO 07, Poznan, Poland, PTETiS Poznan, pp 120 124.
9. Coudray N, Buessler JL, Kihl H, Urban JP (2007) Multi scale and first derivative analysis for edge detection in tem images. ICIAR 07, Montréal, Canada, Springer LNCS, vol 4633, pp 1005 1016.
10. Meyer F (1994) Topographic distance and watershed lines. Signal Processing 38(1):113 125.
11. Karathanou A, Buessler J L, Kihl H, Urban J P (2009) Background Extraction in Electron Microscope Images of Artificial Membranes. IFIP, Thessaloniki, Greece, Springer, vol 296, pp 165 173.
12. Hermann G, Karathanou A, Buessler JL, Urban JP (2009) Evaluation of membrane stacking in electron microscope images. In Digital Imaging Sensors and Applications, Part of the Imaging Science and Technology/SPIE, 21st Annual Symposium on Electronic Imaging, San Jose, CA, USA.
13. Henderson R, Baldwin JM, Downing KH, Lepault J, Zemlin F (1986) Structure of purple membrane from halobacterium halobium: recording, measurement and evaluation of electron micrographs at 3.5 Å resolution. Ultramicroscopy 19:147 178.

Chapter 38
High Performance Computing Approaches for 3D Reconstruction of Complex Biological Specimens

M. Laura da Silva, Javier Roca-Piera, and José-Jesús Fernández

Abstract Knowledge of the structure of specimens is crucial to determine the role that they play in cellular and molecular biology. To yield the three-dimensional (3D) reconstruction by means of tomographic reconstruction algorithms, we need the use of large projection images and high processing time. Therefore, we propose the use of the high performance computing (HPC) to cope with the huge computational demands of this problem. We have implemented a HPC strategy where the distribution of tasks follows the master slave paradigm. The master processor distributes a slab of slices, a piece of the final 3D structure to reconstruct, among the slave processors and receives reconstructed slices of the volume. We have evaluated the performance of our HPC approach using different sizes of the slab. We have observed that it is possible to find out an optimal size of the slab for the number of processor used that minimize communications time while maintaining a reasonable grain of parallelism to be exploited by the set of processors.

Keywords 3D reconstruction · Parallel computing · Master-slave paradigm

38.1 Introduction

Electron microscopy together with sophisticated image processing and 3D reconstruction techniques yield quantitative structural information about the spatial structure of biological specimens [1]. The studies of complex biological specimens at subcellular levels have been possible thanks to electron tomography (ET) [2], image processing, and 3D reconstruction techniques.

ET combines electron microscopy and the principles of tomographic imaging to elucidate the 3D structure of the specimen at molecular resolution [3, 4].

M.L. da Silva (✉)
Dpto. de Arquitectura de Computadores, Universidad de Almería, 04120 Almería, Spain
e mail: laura@ace.ual.es

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 38,

Tomographic reconstruction allows 3D structure determination from a series of electron microscope images. These projection images are acquired at different orientations, by tilting the specimen around one or more axes. The 3D structure is reconstructed by means of the well-known reconstruction method in ET called weighted back projection (WBP).

Parallel computing has been widely investigated for many years as a means to provide high-performance computational facilities for large-scale and grand-challenge applications. In ET, the reconstruction files are usually large and, as a consequence, the processing time needed is considerable. Parallelization strategies with data decomposition provide solutions to this kind of problem [4 7].

The master slave paradigm has been extensively used in the high performance computing (HPC) field, mainly due to its simplicity and to its ability to balance the workload. This paradigm uses a master processor that distributes the jobs and receives the results, while the slave processors only do the processing. In our case, the tasks consist in the reconstruction of a slab of slices, that is, a piece of the final 3D structure.

The aim of this chapter is to evaluate and to find an optimal size of the slab for the number of processors to achieve the scalability of the parallel reconstruction implementation and lower communications time.

38.2 Three-Dimensional Reconstruction of Biological Specimens

In the course of the structure reconstruction of biological specimens by ET, there are two fundamental steps: the acquisition of projection images and the 3D reconstruction [8]. In the first step, the projection images from a single individual specimen are acquired by following the so-called single-axis geometry as shown

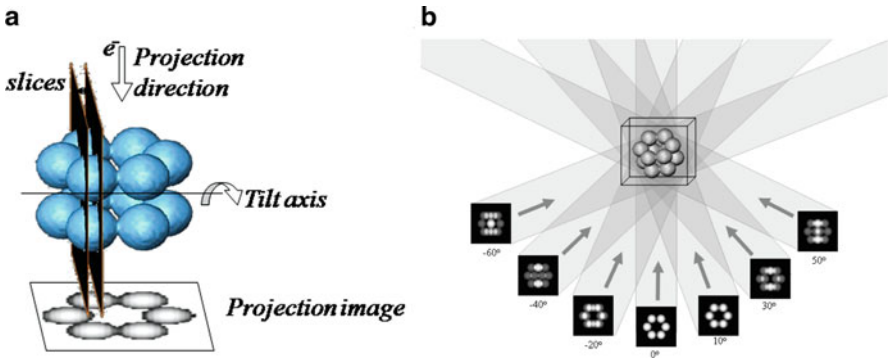


Fig. 38.1 (**a**) Acquisition of a 2D projection image while the object is tilted around an axis. (**b**) 3D reconstruction from projections using the WBP method

in Fig. 38.1a. In the microscope, the specimen is tilted over a limited range, typically from $-70°$ to $+70°$, at small tilt increments ($1°$ $2°$). An image of the same object area is then recorded at each tilt angle via, usually, CCD cameras. Those images represent projections of the specimen.

The second step, 3D reconstruction, is intended to derive the structure of the specimen from the set of projection images. The most common reconstruction method in ET is WBP. Under the assumption that projection images represent the amount of mass density encountered by imaging rays, this method simply distributes the known specimen mass present in projection images evenly over the reconstruction volume. When this process is repeated for a series of projection images recorded at different tilt angles, backprojection rays from the different images intersect and reinforce each other at the points where mass is found in the original structure (see Fig. 38.1b).

The choice of the set of basis functions to represent the object to be reconstructed influences the result of the algorithm. We have used the well-known voxel (volume element) as basis function. The voxels are capable of modeling the structure by non-overlapping density cubes.

38.3 Parallel Computing in Electron Tomography

Cluster computing turns out to be a cost-effective vehicle for supercomputing, based on the usage of commodity hardware and standard software components. The availability of application programming interfaces (APIs) such as message-passing interface (MPI) allows programmers to implement parallel applications regardless of the computing platforms.

The use of voxels allows decomposition of the global 3D problem into multiple, independent problems of 2D reconstruction of slices orthogonal to the tilt axis (Figs. 38.1b and 38.2a). This decomposition is relatively straightforward to implement on parallel computers [7] as the reconstruction of the one-voxel-thick slices orthogonal to the tilt axis (or subsets of slices, known as slabs) can be assigned to an individual node on the parallel computer (Fig. 38.2b). This computational model applied for parallel computing is called single-program multiple data (SPMD).

We have implemented a SPMD strategy based on the commonly used master slave model. In ET, this model has been previously used where the tasks consisted in the reconstruction of individual slices [2]. Instead, in our approach the tasks to be distributed consist in the reconstruction of a slab of slices. In the following, NS denotes the number of slices in a slab. Note that the set of 1D projections needed to reconstruct a 2D slice is usually called sinogram.

In the implemented algorithm, the master node receives the message "READY" from the idle slaves (Fig. 38.3(1)) and then sends a slab of sinograms to them (Fig. 38.3(2)). Each slave receives this slab of sinograms and executes the same program to reconstruct its different data subdomain (see Fig. 38.3(3)), that is, its

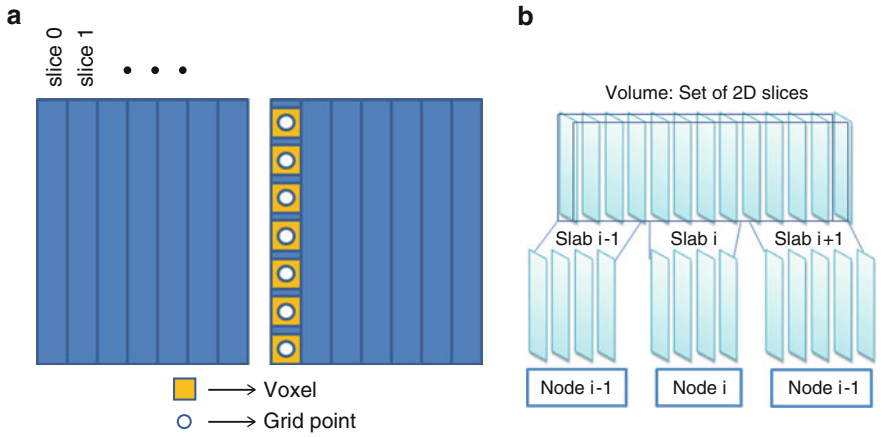


Fig. 38.2 (**a**) Decomposition of the 3D object into multiple, independent slices orthogonal to the tilt axis (see Fig. 38.1a). (**b**) Decomposition of the global 3D problem into multiple, independent reconstruction problems of slabs (i.e., subsets) of slices that are assigned to different nodes in the parallel computer. Each column represents a 2D slice orthogonal to the tilt axis of the volume to be reconstructed

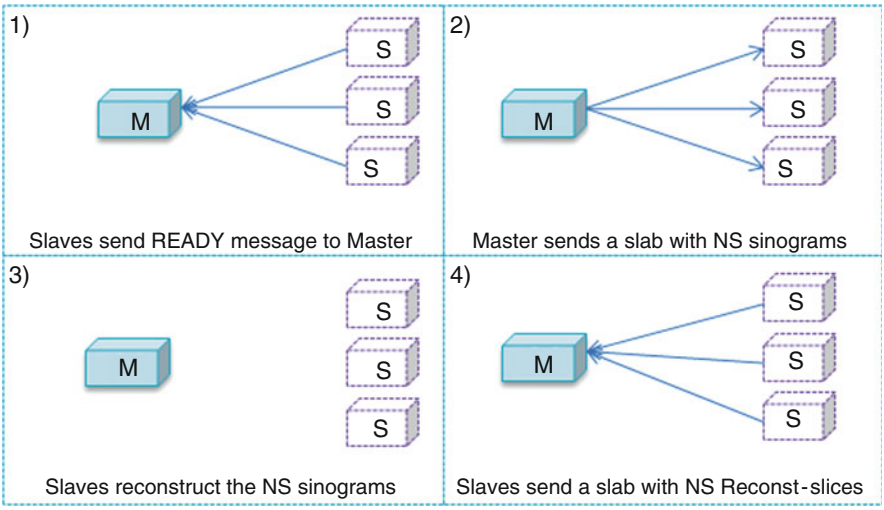


Fig. 38.3 Working diagrams for the parallel reconstruction algorithm

slab of NS sinograms. In the next step, we can see in Fig. 38.3(4) that the slave builds the reconstructed slab with the NS reconstructed slices and sends this slab to the master. This procedure is repeated while there are tasks to send and the slaves are idle.

38.4 Results

In order to quantify the influence of the slab size, we measured the times dedicated to the different stages of our parallel reconstruction algorithm, as shown in Figs. 38.4a, b. Datasets based on a synthetic mitochondrion phantom [5] were used and consisted of 180 projection images taken at a tilt range [−90°, +89°] at interval of 1°.

The dataset referred to as 128 had 128 sinograms of 180 1D projections of 128 × 128 pixels to yield a reconstruction of 128 × 128 × 128 voxels, and so forth. The values of the slab size were set up according to the size of the datasets (ND ∈ {128, 256, 512}) and the number of processors (NP ∈ {1, 2,. . .,32}) used, according to TSlab $= 2^N$ with $N = 0,\ldots, \log_2(\mathrm{ND/NP})$. Each experiment was evaluated five times and the average times for processing and communications were computed.

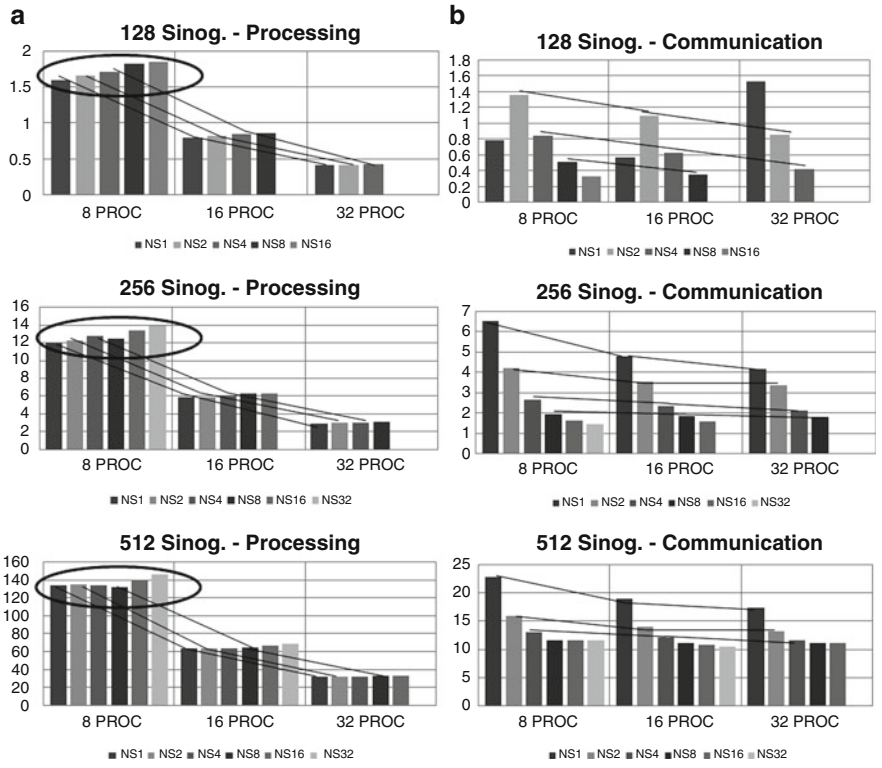


Fig. 38.4 (**a**) Processing times (seconds) for the parallel strategy and the dataset of 128, 256, and 512 sinograms, respectively. NS* denotes the slab size and *PROC denotes the number of processors. (**b**) Communication times (seconds) for the parallel strategy and the dataset of 128, 256, and 512 sinograms, respectively. NS* denotes the slab size and *PROC denotes the number of processors

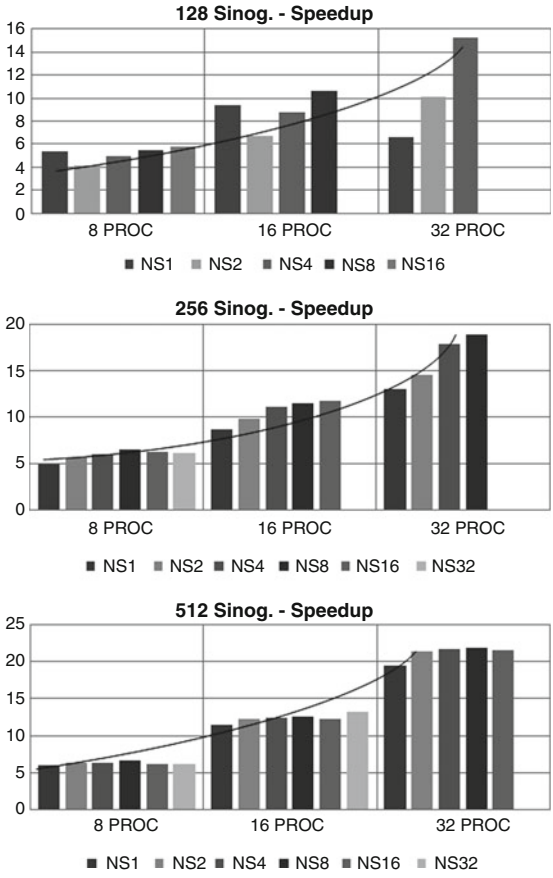


Fig. 38.5 Speedup for the parallel strategy and the dataset of 128, 256, and 512 sinograms, respectively. NS* denotes the slab size and *PROC denotes the number of processors

38.4.1 Analysis of the Processing and Communications Time

Figure 38.4a clearly shows that the processing time is scaled by the number of processors, thereby using more processors following a SPMD model, which implies a lower processing time. On the other hand, the processing time is an absolute time, that is, it is independent of the slab size. So, in principle, the processing time should remain constant for a given number of processors. However, we can observe in Fig. 38.4a that the processing time varies with the slab size. That may come from the additional tasks carried out to handle the data structures associated to the slabs, particularly the allocation of memory for the slabs of the reconstructed slices.

The communication time is a considerable part of the total time according to the number of processors. It comes from the fact that there is more communications master slave when the number of processors is higher.

The influence of slab and the processors number is remarkable in the communications. We can observe in Fig. 38.4b for all slab sizes that the larger the slab is, the lower the communications time. This is because the larger the slab, the number of slabs to be processed are less. However, we can also see that there is an inflection point from the slab size 8 approximately, where the values of the communication times keep constant for slab size of 256 and 512. This may come from the fact that the master slave communications are harder, because more data have to be transferred and all slaves need to communicate with the master, which causes a possible delayed communication time and turns the access to the network into a bottleneck.

38.4.2 Speedup

The performance of the parallel approaches is commonly evaluated in terms of the speedup, which is defined as $S = T_1/T_p$, where T_1 is the execution time of the sequential algorithm in a single cluster's processor and T_p is the execution time (processing + communication time) of the parallel algorithm with p processors.

Figure 38.5 clearly shows that the increase in the number of processors produces an increasing curve for all data sizes. On the other hand, the influence of the slab size only affects significantly for the data sizes of 128 and 256, where the increase of the slab size improves speedup. However, for the data size of 512, the variation of the slab size keeps the speedup roughly constant.

38.5 Conclusions

In this work, we have analyzed the application of parallel strategies for 3D reconstruction in electron tomography of complex biological specimens, with a special emphasis on the computational perspective. We have applied HPC techniques so as to face the high computational demands and take advantage of parallel systems. We have evaluated their performance in a cluster of workstations using datasets with different problem sizes and the standard tomographic reconstruction method.

The results that we have obtained clearly show that the combination of HPC and a changeable slab size yields solutions in reasonable computation times. The use of voxels has the main benefit of the data independence, which makes easier the parallel approach based on domain decomposition, the SPMD model, and master slave paradigm. The SPMD model provides scalability in the parallel algorithm and the master slave paradigm helps to make a balanced distribution of the tasks among the slave nodes.

Acknowledgments Work partially supported by grants MCI TIN2008 01117, JA P06 TIC01426, and CSIC PIE200920I075.

References

1. Sali, A., Glaeser, R., Earnest, T., Baumeister, W. (2003) From words to literature in structural proteomics. Nature 422:216 225
2. Lucic, V., Foerster, F., Baumeister, W. (2005) Structural studies by electron tomography: From cells to molecules. Annual Review of Biochemistry 74:833 865
3. Perkins, G.A., Renken, C.W., Song, J.Y. et al (1997) Electron tomography of large, multicom ponent biological structures. Journal of Structural Biology 120:219 227
4. Fernández, J.J., Carazo, J.M., García, I. (2004) Three dimensional reconstruction of cellular structures by electron microscope tomography and parallel computing. Journal of Parallel Distributed Computing 64:285 300
5. Fernández, J.J., Lawrence, A.F., Roca, J. et al (2002) High performance electron tomography of complex biological specimens. Journal of Structural Biology 138:6 20
6. Fernández, J.J., Gordon, D., Gordon, R. (2008) Efficient parallel implementation of iterative reconstruction algorithms for electron tomography. Journal of Parallel Distributed Computing 68:626 640
7. Fernández, J.J. (2008) High performance computing in structural determination by electro ncryomicroscopy. Journal of Structural Biology 165:1 6
8. Fernández, J.J., Sorzano, C.O.S., Marabini, R., et al (2006) Image processing and 3D reconstruction in electron microscopy. IEEE Signal Processing Magazine 23(3):84 94

Chapter 39
Protein Identification Using Receptor Arrays and Mass Spectrometry

Timothy R. Langlois, Richard W. Vachet, and Ramgopal R. Mettu

Abstract Mass spectrometry is one of the main tools for protein identification in complex mixtures. When the sequence of the protein is known, we can check to see if the known mass distribution of peptides for a given protein is present in the recorded mass distribution of the mixture being analyzed. Unfortunately, this general approach suffers from high false-positive rates, since in a complex mixture, the likelihood that we will observe any particular mass distribution is high, whether or not the protein of interest is in the mixture. In this paper, we propose a scoring methodology and algorithm for protein identification that make use of a new experimental technique, which we call *receptor arrays,* for separating a mixture based on another differentiating property of peptides called *isoelectric point* (*pI*). We perform extensive simulation experiments on several genomes and show that additional information about peptides can achieve an average 30% reduction in false-positive rates over existing methods, while achieving very high true-positive identification rates.

Keywords Receptor · Array · Mass · Spectrometry · Protein · Identification · Isoelectric · Point

39.1 Introduction

Identifying proteins in complex mixtures is important for a wide variety of proteomics applications. Currently, two general types of methods based on mass spectrometry (MS) are used to identify proteins: tandem mass spectrometry and peptide mass fingerprinting. Tandem mass spectrometry requires specialized equipment and relies on the production of large datasets. In contrast, protein identification by mass fingerprinting typically involves separation of a complex mixture of proteins, proteolytic digestion of an isolated protein(s) of interest, mass analysis of the

R.R. Mettu (✉)
University of Massachusetts, Amherst, MA, USA
e mail: mettu@ecs.umass.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 39,

resulting peptide fragments, and database searching. This method is relatively straightforward and uses readily available and more easy-to-use instrumentation (i.e., MALDI-TOF). In addition, because liquid separation techniques (e.g., HPLC) are typically not needed prior to MALDI-TOF analyses, but are required in sequence tagging approaches, method development is more straightforward with the peptide mass fingerprinting approach. The challenge of mass fingerprinting approaches, however, is to overcome a relatively high false-positive rate, along with low success rates when searching large databases.

Recently, researchers have explored the use of other physical and chemical information from peptides to improve protein identification in mass fingerprinting approaches [1, 12]. One new approach uses self-assembling polymers to fractionate simultaneously protein digest mixtures and provide search-constraining physical and chemical information so that protein identification by mass fingerprinting can be achieved without resorting to time-consuming gel-based separations [3, 8]. We call this new experimental technique a *receptor array* since it is similar in spirit to microarray experiments, in which mRNA is hybridized according to sequence. Analogously, the technique we study in this paper (see Sect. 39.3) is able to separate peptides in a mixture with chosen isoelectric point values.

In this paper, we propose a scoring function and method for protein identification for a generalized version of peptide mass fingerprinting using isoelectric point separation. We argue that this approach can achieve a significant reduction in false-positive rates while also achieving very high accuracy, thus eliminating the traditional disadvantage of mass fingerprinting approaches, while preserving the simplicity of the experimental setup. We perform extensive simulation on four different genomes, studying the effectiveness of our algorithm for protein identification in mixtures which contain up to half the genome. At the high level, incorporation of additional peptide properties such as isoelectric point shows considerable promise for making protein identification in complex mixtures a reality. For mixtures of over 1,000 proteins, our approach achieves accuracies of over 90%, which is about 25% better than using mass only. Furthermore, our approach yields false-positive rates under 10%, while the mass-only approach yields about a 30% false-positive rate.

39.2 Related Work

While de novo methods for protein identification exist [11] (i.e. in which the primary sequence of the protein of interest is not known), we limit our discussion here to database methods, since these are most directly comparable with the work described in this paper. Given an experimental spectrum, database methods use a scoring algorithm to rank proteins from the database. The particular scoring algorithm varies across different methods, but at a high level, the score generally represents the likelihood that any particular protein (or set of proteins) from the database is contained in the sample from which the spectrum was obtained [11]. There are many algorithms

based on this general approach, examples include but are not limited to SEQUEST [15], which utilizes spectral matching, or dot-product methods [11] such as TANDEM [4, 10], OMSSA [7], MASCOT [13], and Comet [9].

While our algorithm is essentially a database approach, it differs from the methods described above since it is designed to work on mass fingerprinting data instead of tandem mass spectrometry data. It is known that utilizing isoelectric point data allows greater separation of the peptides than that with mass alone, which in effect reduces false-positive peptide, and thus, protein, identification [1, 5]. However, the receptor array approach provides a simpler and more reliable method for utilizing isoelectric point information (see Sect. 39.3 for a more specific discussion). The approach of incorporating multiple data sources has been used previously; for example, MultiIdent is a program for protein identification that uses the spectra, *pI* information, and amino acid composition data to attempt to identify a protein. However, it is designed to use data obtained from 2D gels that give *pI* data for an entire protein rather than for sets of peptides.

39.2.1 Problem Definition

In a typical experiment to analyze a mixture M of proteins using MALDI-MS, we must first digest the mixture using an enzyme such as trypsin, since typical proteins are too large to be observed directly using mass spectrometry. Trypsin is a serine protease that cleaves proteins between Lysine and Arginine residues (except when either is followed by a Proline)[1] for MALDI-MS, and a spectrum that shows the abundance of peptides at particular masses is recorded. Throughout this paper, we will assume that the resulting spectrum gives an (normalized) abundance of peptides at a particular mass (in parts per million, or ppm[2]). Given such a spectrum, the problem of protein identification is to test whether a given protein, or a set of proteins (i.e., their set of constituent peptides), is actually present in the mixture.

We now introduce some notation that we will use to more formally describe the problem of protein identification. First, we assume that the mixture M is drawn from a single organism, or proteome P whose sequences are known in advance. Thus, when a mixture M of proteins is to be analyzed, it will be digested into a mixture of peptides with masses in the range $[L \ldots H]$. We will denote the error of the mass spectrometer by δ_{m}. Then, for our purposes, we will view the spectrum resulting from any mixture drawn from P as a set of $a = \frac{(H - L)}{\delta_{\mathrm{m}}}$ measurements. Each of these measurements corresponds to a mass range, which we call a *bin*, where the ith bin is given by

$$\mathrm{bin}(i) = [L + i\delta_{\mathrm{m}}, L + (i+1)\delta_{\mathrm{m}}].$$

[1]For simplicity, in our experiments we assume that trypsin cleaves ideally, since we use the same mixture for all experiments; that is, this assumption does not bias the experimental accuracies.

[2]It is typical to report mass/charge ratio in a spectrum, but we will use "mass" and "mass/charge" interchangeably in this paper.

For any set of peptides X, we let $\sigma^*(X, i) \in [0, 1]$ denote the abundance of pepides from X that fall into mass bin i. Note that this simply corresponds to an idealized version of a mass spectrum of X.

Thus the input to the protein identification problem consists of a database of possible proteins P, a set of proteins $R \subset P$ to be identified, a spectrum $\Sigma(M) = (\sigma(M, 1), \sigma(M, 2), \ldots, \sigma(M, a))$, and a mass error δ_{m}. The output simply consists of a yes/no answer to whether the set of proteins P to be identified is contained in the input mixture (i.e., as represented by $\Sigma(M)$). While experimental error (i.e., true-positive identification rate) is of some concern in mass spectrometry, as pointed out in the previous section, the primary issue with traditional mass spectrometry is mass degeneracy (i.e. false-positive identification rate). As the mass error δ_{m} of the spectrometer grows, the peptides identified by the spectrum simply all fall into the same mass bin, and thus P is more likely be reported as present in M when it is not.

39.3 Receptor Arrays for Mass Spectrometry

While other properties of peptides have been previously studied, including isoelectric point [1], we make use of a new experimental technique that allows rapid selection of peptide properties for identification. This separation protocol for selecting peptides based on their isoelectric point uses a homopolymer that self-assembles into reverse micelles in apolar solvents [3]. This reverse micelle assembly selectively extracts positively charged peptides from aqueous solution into an immiscible apolar phase. By sequentially decreasing the pH of the aqueous solution, peptides can be sequentially extracted according to their isoelectric points (*pI*). Consequently, peptides falling in any given *range* of isoelectric values can be isolated as well. Subsequent MALDI-MS analysis of each subsequent organic phase extract provides m/z measurements of the peptides (i.e. in a chosen isoelectric range). We call this method a "receptor array" since it is similar in spirit to microarray experiments, in which mRNA is hybridized according to sequence.

Using the technique described above, we can select a *pI* value and only extract from the given mixture those peptides whose *pI* is higher than the chosen *pI*. More concretely, for any chosen isoelectric point value p, the mixture M can be viewed as being split into two sets of peptides: M_{P}, the set of all peptides with isoelectric point greater than p, and $M_{\mathrm{p}} = M - M_{\mathrm{p}}$, the remaining peptides. By performing MALDI-MS on these mixtures (after separation), instead of on M, we can obtain two spectra, $\Sigma(M_{\mathrm{P}})$ and $\Sigma(M_{\mathrm{p}})$ (see Fig. 39.1). Using known sequence of the protein to be identified, and calculated *pI* values for its peptides, we can "limit" our mixture to selected peptides. Figure 39.1 illustrates how this approach can reduce the mass degeneracy at a peak in the original spectrum $\Sigma(M)$. Currently, the above experimental techniques are performed by separating a two-phase solution, and it is currently possible to separate the mixture by isoelectric point, to an accuracy of 1 *pI* unit [2]; we assume in our simulations that we are able to select *pI* ranges of 1 unit each.

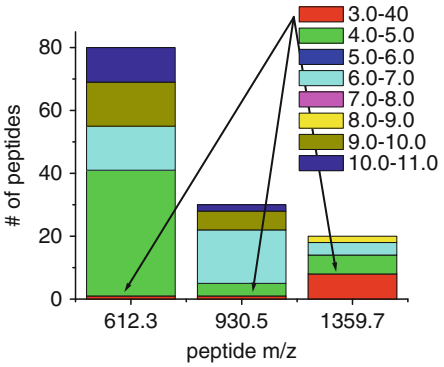


Fig. 39.1 Selecting Peptides by Isoelectric Point. Breakdown of *pI* values for peptides at 3 different *m*/*z* (from *P. aerophilum*). Choosing receptors that select for *pI* values of 3.0 4.0 would identify the individual proteins giving rise separately to *m*/*z* 612.3 and 930.5, but would not uniquely identify any of the proteins giving rise to the peptide at *m*/*z* 1359.7

39.4 Algorithm

In this section, we give an algorithm to incorporate the experimental approach described in Sect. 39.3 for protein identification using MALDI-MS. First, we describe the scoring function we use, and then describe how we utilize additional experimental data to record mass spectra. Then, we describe how we choose the threshold for our score that, along with the scoring function and recorded spectra, is used for identification.

As mentioned earlier, we assume that we have proteome P of the organism we are studying. From this database, we construct a database of the peptides, $D = \{p_1, \ldots, p_n\}$, of this organism by simulating the digestion of its proteins. For each peptide in D, we also compute the predicted isoelectric point using the EMBOSS tool [14] and mass using EXPASY [6]. For simplicity, we will assume throughout this section that we are asked to identify whether a single protein is in the given set R. Also, we will fix the mixture M to be analyzed throughout. Recall that we are also given a mass error δ_{m} along with the mass spectrum $\Sigma(M)$.

39.4.1 Scoring Function

For the given spectrum $\Sigma(M)$ and the protein R, we wish to identify, we define the *score* of R in the mixture as

$$\Theta(R) = \sum_{k \in R} \left| \frac{\sigma(M,k)}{\sum_{j \in R} \sigma(M,j)} - \frac{\sigma^*(R,k)}{\sum_{j \in R} \sigma^*(R,j)} \right|$$

where $k \in R$ and $j \in R$ denote $\{k|\sigma^*(R,k)>0\}$ and $\{j|\sigma^*(R,j)>0\}$. This scoring function captures whether the distribution of observed masses in the given spectrum

matches that of the protein of interest. This scoring function is very similar to the approach used in many statistical methods; and indeed the proposed generalization we are developing can be applied in it as well.

We note that in a situation where R and M have identical sets of peptides, the resulting score is 0, and that when R and M have no peptide masses in common, the score is 2. Thus all scores will be between 0 and 2, and the likelihood of R being in the mixture will depend on how close $\Theta(R)$ is to 0.

39.4.2 Incorporating Isoelectric Point Selectivity

Typically, δ_{m} is large enough that there can be a large number of distinct peptides in each bin. This complicates identification since we do not know which peptides account for each $\sigma(M,j)$. To assist identification methods, we can reduce the size of $\sigma(M,i), i = 1,\ldots,a$ through the use of isoelectric point (pI) separation. We first describe how the scoring function can be applied straightforwardly when we consider different isoelectric point ranges. Suppose that we chose k pI point ranges, $d_1 \ldots d_k$. Then M would be split into $k+1$ mixtures $M_1,\ldots,M_k+1$ such that

$$M_j = \{p_x \in M | d_j \leq pI(p_x) < d_{j+1}\}$$

The same scoring function also applies, except that now we can apply it to only a subset of the original mixture. Define M' as

$$M' = \bigcup_j M_j$$

for all j, where a peptide in isoelectric point range j from R contributes to the score. More precisely, the new scoring function would be

$$\hat{\Theta}(R) = \sum_{k \in R} \left| \frac{\sigma(M',k)}{\sum_{j \in R} \sigma(M',j)} - \frac{\sigma^*(R,k)}{\sum_{j \in R} \sigma^*(R,j)} \right|$$

where $k \in R$ and $j \in R$ are as before.

Note that, for any set of proteins R and mixture M, $\Theta(R) \geqslant \hat{\Theta}(R)$. This observation follows from the fact that if we do not split the mixture at all, we would have that $\Theta(R) = \hat{\Theta}(R)$. The only source of error in this inequality is the error due to the receptor array not selecting isoelectric point ranges accurately; however, it has been shown that this is possible with a high enough degree of accuracy, up to 1 isoelectric point unit [2].

Ideally we would only choose those M_i such that $\sigma(R,M_i) > 0$. To increase the accuracy of the identification, we would choose enough M_i to include all peptides

in R, while at the same time making each M_i as small as possible to reduce the number of possible proteins. In our algorithm, we simply take the peptides present in R and organize them according to isoelectric point ranges (of size 1) and save any range that contains peptides. We then score R according to the resulting spectra.

Now that we have established a method of scoring any particular set R of proteins to be identified from a mixture M, we must also define how to determine whether R is indeed in M or not. To determine the threshold, we first create a reference mixture M^* by taking half of P at random. Then, for each possible score (between 0 and 2, in 0.01 increments), we calculate false-positive, false-negative, true-positive and true-negative rates based on identifying every protein in P. Then, we choose the threshold that jointly maximizes true-positive rate and minimizes false-positive rate.

39.5 Experiments

To test the utility of our scoring scheme and algorithm for identification, we have performed a series of simulation experiments for four genomes, *P. islandicum*, *P. aerophilum*, *E. coli*, and *S. cervisiae*. For each proteome, we took the proteins listed by expasy.org as the set of proteins P. We note that we do not consider post-translational modifications in our experiments, since expasy.org does not include these as part of the genomes we considered. For each genome, we first choose at random a mixture consisting of half the genome, and then choose at random a set of 30 proteins in the mixture and not in the mixture. To construct the digested mixture, we simulated cleavage by trypsin and computed the isoelectric point using [14] and the mass of the resulting peptides using [6]. Then, we use our identification method for 100 trials using mass only, and mass and *pI*. When using *pI* for identification, we take the peptides in the set of proteins to be identified and use the isoelectric ranges that are present in those proteins. For each mixture we calculate the masses of the peptides using a standard table, but the size of the bins in the resulting spectrum is varied according to a chosen mass error. We chose to test our algorithm on four proteomes, *P. aerophilum*, *P. islandicum*, *E. coli* (K12 strain), and *S. cervisiae*. *P. aerophilum* and *P. islandicum* are hyperthermophiles that have medium-sized proteomes of 2,600 and 2,000 proteins, respectively, while *E. coli* and *S. cervisiae* have much larger proteomes, with 4,000 and 6,700 proteins, respectively.

The results of our algorithm for mixtures from these four proteomes are given in Figs. 39.2 and 39.3. For each proteome, we report average scores for each trial, as described in Sect. 39.4, for sets of proteins in the mixture and not in the mixture, given a particular mass error. A typical mass error for MALDI-MS is 50 ppm; the average accuracies and false-positive rates given below assume this mass error. The goal of our identification method is of course to determine whether the set to be identified is in the mixture or not. Thus, as we see in Figs. 39.2 and 39.3, the gap in average score indicates in some sense the amount of information provided by mass and isoelectric point. In all cases, the use of receptor arrays provides a significantly

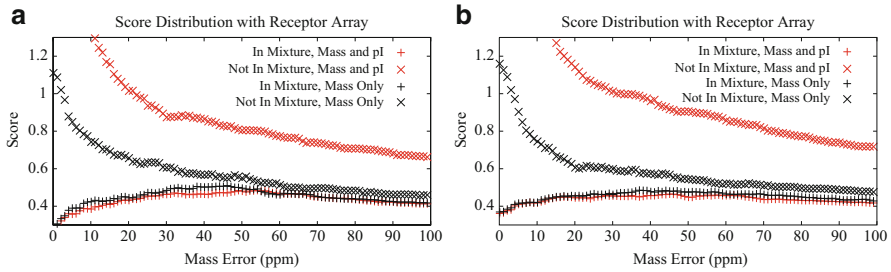


Fig. 39.2 Results for *P. aerophilum* and *P. islandicum*. (**a**) The average score distribution for protein identification with varying mass error, for mixtures of about 1,300 proteins. (**b**) The average score distribution for protein identification with varying mass error, for mixtures of about 1,000 proteins

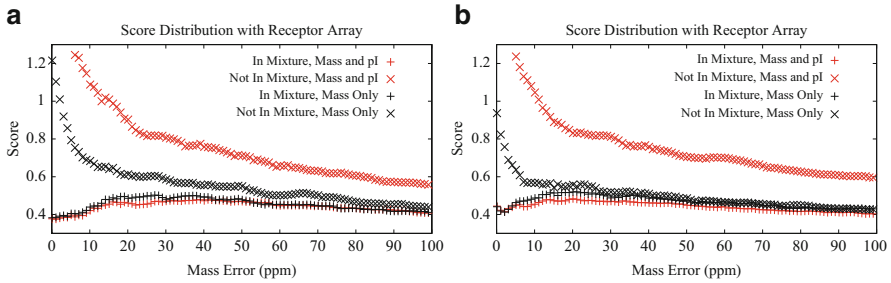


Fig. 39.3 Results for *E. coli* and *S. cervisiae*. (**a**) The average score distribution for protein identification with varying mass error, for mixtures of about 2,000 proteins. (**b**) The average score distribution for protein identification with varying mass error, for mixtures of about 3,300 proteins

larger separation in score between proteins in the mixture and proteins not in the mixture, making identification more robust.

For *P. aerophilum*, the average size of a mixture was 1,300 proteins. Using mass only, our identification method achieves an accuracy of 63% and a false-positive rate of 37%. In contrast, we were able to achieve an accuracy of 90% and a false-positive rate of 11% using isoelectric point separation. For *P. islandicum*, the average size of a mixture was 1,000 proteins. Using mass only, our identification method achieves an accuracy of 65% and a false-positive rate of 34%. With isoelectric point separation, we can achieve an accuracy of 91% and a false-positive rate of 8.2%.

For *E. coli* and *S. cervisiae*, the size of the mixture was considerably larger than that of the other two proteomes we tested, with an average mixture having 2,000 proteins and 3,300 proteins, respectively. For *E. coli*, using mass only yields an accuracy of 59% and a false-positive rate of 40%, while the addition of isoelectric point information gives an accuracy of 84% and a false-positive rate of 16%. For *S. cervisiae*, using mass only yields an accuracy of 49% and a false-positive rate of 51%, while the addition of isoelectric point information gives an accuracy of 83%

and a false-positive rate of 17%. In Figs. 39.2 and 39.3, we see that the addition of isoelectric point information alone yields a much more separable scoring trend than that for mass alone. This shows that our experimental approach does not require a particularly sensitive mass spectrometer; whereas with mass alone, identification in complex mixtures may not be possible due to the mass error of the spectrometer.

Acknowledgment S. Thayumanavan in the Department of Chemistry at UMass Amherst for useful ideas and discussions.

References

1. B. J. Cargile and J. L. Stephenson Jr. An alternative to tandem mass spectrometry: Isoelectric point and accurate mass for the identification of peptides. *Anal. Chem.*, 76(2):267 275, 2004.
2. M. Y. Combariza, E. Savariar, S. Thayumanavan, and R.W. Vachet. Isoelectric point depen dent fractionation and detection of protein digests using polymeric nanoassemblies and MALDI MS analysis. In *Proceedings from the 56th ASMS Conference on Mass Spectrometry and Allied Topics*, 2008.
3. M. Y. Combariza, E. N. Savariar, D. R. Vutukuri, S. Thayumanavan, and R. W. Vachet. Polymeric inverse micelles as selective peptide extraction agents for MALDI MS analysis. *Anal. Chem.*, 79(18):124 130, 2007.
4. R. Craig and R. C. Beavis. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.*, 17(20):2310 2316, 2003.
5. Amal S. Essader, Benjamin J. Cargile, Jonathan L. Bundy, and James L. Stephenson. A comparison of immobilized ph gradient isoelectric focusing and strong cation exchange chromatography as a first dimension in shotgun proteomics. *Proteomics*, 5:24 34, 2005.
6. E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R.D. Appel, and A. Bairoch. Expasy: The proteomics server for in depth protein knowledge and analysis., 2003. http://www.expasy.org.
7. L. Y. Geer, S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi, and S. H. Bryant. Open mass spectrometry search algorithm. *J. Proteome Res.*, 3(5):958 964, 2004.
8. A. Gomez Escudero, M. A. Azagarsamy, N. Theddu, R. W. Vachet, and S. Thayumanavan. Selective peptide binding using facially amphiphilic dendrimers. *J. Am. Chem. Soc.*, 130 (33):11156 11163, 2008.
9. A. Keller, J. Eng, N. Zhang, X. J. Li, and R. Aebersold. A uniform proteomics MS/MS analysis platform utilizing open XML file formats. *Mol. Syst. Biol.*, 1:0017, 2005.
10. B. MacLean, J. K. Eng, R. C. Beavis, and M. McIntosh. General framework for developing and evaluating database scoring algorithms using the tandem search engine. *Bioinformatics*, 22(22):2830 2832, 2006.
11. A. I. Nesvizhskii, O. Vitek, and R. Aebersold. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Meth.*, 4(10):787 797, 2007.
12. M. Palmblad, M. Ramström, K. E. Markides, P. H/a kansson, and J. Bergquist. Prediction of chromatographic retention and protein identification in liquid chromatography/mass spec trometry. *Anal. Chem.*, 74(22):5826 5830, 2002.
13. D. N. Perkins, D. J. C. Pappin, D. M. Creasy, and J. S. Cottrell. Probability based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, 20(18):3551 3567, 1999.
14. Peter Rice, Ian Longden, and Alan Bleasby. EMBOSS: The European molecular biology open software suite. *Trends Genet.*, 16(6):276 277, 2000.
15. J. R. Yates, J. K. Eng, A. L. McCormack, and D. Schieltz. Method to correlate tandem mass spectra of modified peptides to amino acid sequences in protein database. *Anal. Chem.*, 67(8):1426 1436, 1995.

Part IV
Comparative Sequence, Genome Analysis, Genome Assembly, and Genome Scale Computational Methods

Chapter 40
Assessing Consistency Between Versions of Genotype-Calling Algorithm Birdseed for the Genome-Wide Human SNP Array 6.0 Using HapMap Samples

Huixiao Hong, Lei Xu, and Weida Tong

Abstract Highly accurate and reproducible genotype calling is a key to success of genome-wide association studies (GWAS) since errors introduced by calling algorithms can lead to inflation of false associations between genotype and phenotype. The Affymetrix Genome-Wide Human SNP Array 6.0 is widely utilized and was used in the current GWAS. Birdseed, a genotype-calling algorithm for this chip, is available in two versions. It is important to know the reproducibility between the two versions. We assessed the inconsistency in genotypes called by the two versions of Birdseed and examined the propagation of the genotype inconsistency to the downstream association analysis by using the 270 HapMap samples. Our results revealed that genotypes called from version-1 and version-2 of Birdseed were slightly different and the inconsistency in genotypes propagated to the downstream association analysis.

Keywords Algorithm · Association · Genotype calling · HapMap · Reproducibility

40.1 Introduction

Genome-wide association studies (GWAS) aim to identify genetic variants, usually single nucleotide polymorphisms (SNPs), across the entire human genome that are associated with phenotypic traits, such as disease status and drug response.

The views presented in this article do not necessarily reflect those of the US Food and Drug Administration. Correspondence should be addressed to: Dr. Huixiao Hong, Division of Systems Toxicology, National Center for Toxicological Research, US Food and Drug Administration, 3900 NCTR Road, Jefferson, Arkansas 72079, USA, Tel: 870 543 7296, Fax: 870 543 7382

H. Hong (✉)
Center for Toxicoinformatics, Division of Systems Toxicology, National Center for Toxicological Research, US Food and Drug Administration, Jefferson, AR 72079, USA
e mail: Huixiao.Hong@fda.hhs.gov

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 40,

A statistically significant difference in allele frequencies of a genetic marker between cases and controls implies that the corresponding region of the human genome contains functional DNA sequence variants that make contributions to the phenotypic traits of interest. Recently, GWAS has been applied to identify common genetic variants associated with a variety of phenotypes [1 8]. These findings are valuable for scientists to elucidate the allelic architecture of complex traits in general. However, replication of GWAS showed that only a small portion of association SNPs in the initial GWAS results can be replicated in the same populations. Obviously, there are potentially false positive associations found (Type I errors) and true positive associations missed (Type II errors) in GWAS results, limiting the potential for early application to personalized medicine.

The complexity of GWAS analysis leads to the potential for introducing errors and biases which, in turn, have the potential to generate Type I and II errors. The sources that cause the Type I and II errors are quite diverse and include small sample sizes, a bias in selection of cases and controls and a population stratification. In addition, genotype calling is a potential source to cause these errors. Highly accurate and reproducible genotype-calling results are paramount since errors can lead to inflation of false associations between genotypes and phenotypes.

The Affymetrix Genome-Wide Human SNP Array 6.0 was used in the current GWAS [1 8]. Birdseed [9] was developed to call genotypes from the raw data. Birdseed is a model-based clustering algorithm that converts continuous intensity data to discrete genotype data. Currently, there are two different versions of Birdseed. Version-1 fits a Gaussian mixture model into a two-dimensional space using SNP-specific models as starting points to start the Expectation Maximization (EM) algorithm, while version-2 uses SNP-specific models in a pseudo-Bayesian fashion, limiting the possibility of arriving at a genotype clustering that is very different to the supplied models. Therefore, inconsistent genotypes may be called from the same raw intensity data using different versions of the algorithm, especially when intensity data do not fit the SNP-specific model perfectly. The inconsistency in genotypes caused by different versions of Birdseed has the potential to cause Type I and II errors. Thorough evaluation of inconsistencies between different versions of Birdseed and their effect on significantly associated SNPs has not yet been reported.

In this chapter, we analyzed the differences in genotypes called by different versions of Birdseed using the data of the 270 HapMap samples. We further assessed whether and how the variations in genotypes propagate to the significant SNPs identified in the downstream association analysis.

40.2 Materials and Methods

Raw intensity data (CEL files) were obtained from Affymetrix (3 DVDs with the compressed data were sent to us by Affymetrix). The CEL file format can be found on Affymetrix's developer pages. Three population groups composed the dataset and each group contained 90 samples: CEU had 90 samples from Utah residents

with ancestry from northern and western Europe (termed as European in this chapter); CHB+JPT had 45 samples from Han Chinese in Beijing, China, and 45 samples from Japanese in Tokyo, Japan (termed as Asian in this chapter); and YRI had 90 samples from Yoruba in Ibadan, Nigeria (termed as African in this chapter). The quality of the raw intensity data of the 270 HapMap samples was assessed using the program apt-geno-qc in the Affymetrix Power Tools (ATP) (version 1.10.0) software distributed by Affymetrix.

Genotype calling by Birdseed (both version-1 and version-2) was conducted using apt-probeset-genotype of APT. For the studies reported here, all the parameters were set to the default values recommended by Affymetrix. The chip description file (cdf) and the annotation file were downloaded from the Affymetrix Web site. The training set necessary for the use of Birdseed was also downloaded from the Affymetrix library.

The genotype calls from version-1 and version-2 of Birdseed were compared using a set of in-house programs written in C++. Overall call rates for each of the genotype-calling results, call rates for individual samples and SNPs in each of the experiments, and concordant calls were calculated and exported as tab-delimited text files using the in-house programs.

In order to study the propagation of genotype discordance to significantly associated SNPs, all resulting genotype calls were analyzed using χ^2 statistics tests for associations between the SNPs and the case control setting. Prior to the association analysis, QC of the calls was conducted to remove markers and samples of low quality. A call rate threshold of 90% was used to remove uncertain SNPs and samples. Minor allele frequency was used to filter SNPs and its cut-off was set to 0.01. Departure from Hardy Weinberg equilibrium (HWE) was checked for all SNPs. The p value for the χ^2 test for HWE was calculated at first and then the resulting p values were adjusted for multiple testing using Benjamini and Hochberg's false discovery rate (FDR). FDR of 0.01 was set as the cut-off. To mimic a "case control" study in GWAS, each of the three population groups (European, African, and Asian) was assigned as the "cases" while the other two were considered as the "controls" for each set of the resulting genotype calls. In association analysis, a 2×2 contingency table (allelic association) and a 2×3 contingency table (genotypic association) were generated for each SNP. Then χ^2 statistics were applied to the contingency tables to calculate p values for measuring the statistical significance for each of the associations. Bonferroni correction was applied to adjust the resulting p values. Lastly, a criterion of Bonferroni-corrected p value less than 0.01 was used to identify statistically significant SNPs.

40.3 Results

The genotypes from version-1 and version-2 of Birdseed were compared to measure the consistency. The overall call rates for version-1 and version-2 were 99.824% and 99.827%, respectively. The missing call rates per SNP and per sample were calculated and compared using the one-against-one comparisons of distributions of

the missing call rates. The scatter-plot in Fig. 40.1a compares the missing call rates per SNP of version-1 (x-axis) and version-2 (y-axis) of Birdseed. Each of the points represents one of the SNPs. A larger number of SNPs are not consistent in missing call rates and some behave very differently between the versions. The Pearson correlation coefficient was only 0.8493. The scatter-plot in Fig 40.1b depicts the missing call rates per sample of the version-1 (x-axis) and the version-2 (y-axis). Some of the 270 samples are not consistent in their missing call rates, but the differences are much smaller compared with those per SNP. The Pearson correlation coefficient was very high (0.9989).

The p value of paired two sample t test for comparisons of the missing call rates per SNP was 0.02264, less than 0.05, indicating that the difference is statistically significant. However, the p value for comparing the missing call rates per sample was 0.8943, much greater than 0.05, demonstrating that the difference is not statistically significant.

Three genotypes (homozygote: AA, heterozygote: AB, and variant homozygote: BB) are possible for a genotype call and one cannot be sure that the genotype calls of a SNP for a single sample using the version-1 and the version-2 are exactly same even when both are successfully called. Therefore, we determined the consistency of successful genotype calls. The overall concordance of successful genotype calls was calculated. There are a total of 27,827 genotypes (0.0114%) that were successfully determined by both versions, but the assigned genotypes were different between the two versions. We further examined concordances of the successful calls of three genotypes. It was observed that concordance of the SNPs that were assigned genotypes to AA and BB was higher than AB, for both version-1 and version-2. Moreover, discordant genotypes between heterozygote (AB) and homozygote (AA or BB) were larger than those between two homozygous types AA and BB, for both versions of Birdseed.

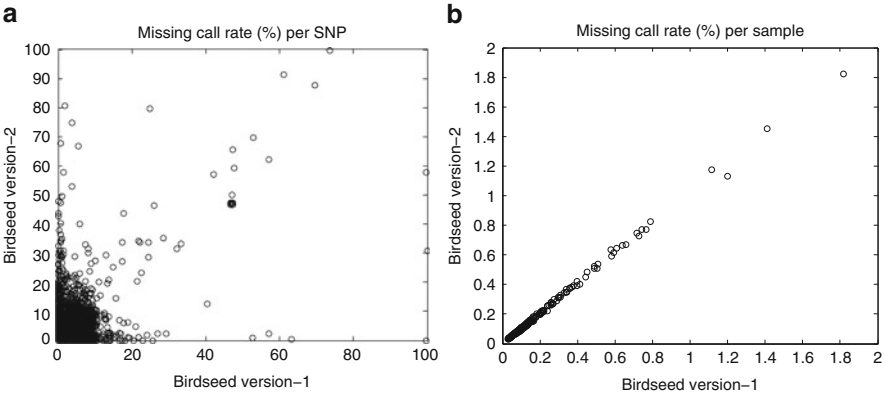


Fig. 40.1 Comparison of missing call rates per SNP (**a**) and per sample (**b**) between version 1 and version 2 of Birdseed

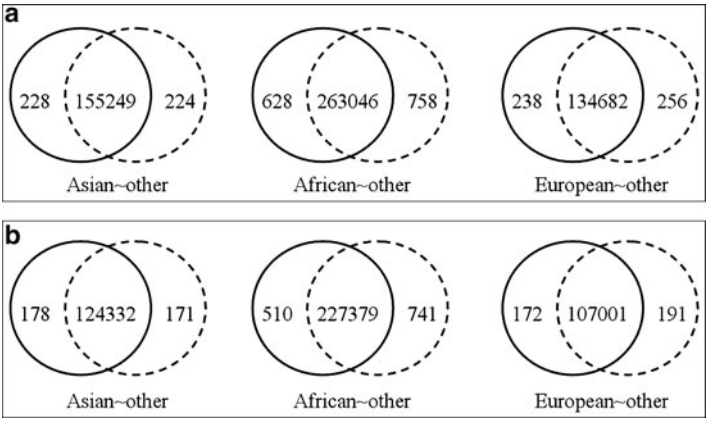


Fig. 40.2 Comparisons of the associated SNPs from version 1 and version 2. The associated SNPs identified using allelic association test (**a**) and genotypic association test (**b**) were compared between version 1 (numbers in the *solid circles*) and version 2 (numbers in the *dotted circles*)

The objective of a GWAS is to identify the genetic markers associated with a specific phenotypic trait. It is critical to assess whether and how the inconsistencies in genotypes propagate to the significant SNPs identified in the downstream association analysis. Comparisons of the significantly associated SNPs obtained from one and two degree of freedom allelic and genotypic association tests between the two versions are given in Fig. 40.2a, b, respectively. The significant SNPs from version-1 are given in the solid circles, and the ones from version-2 in the dotted circles. The "case" "control" designs are given under each Venn diagram. It is clear that, for all "case" "control" designs and for both allelic and genotypic tests, the inconsistencies in genotypes propagated into the downstream association analyses, resulting in different lists of associated SNPs.

40.4 Discussion

A very small error introduced in genotypes from genotype calling may result in inflated Type I and II error rates in the downstream association analysis. Reproducibility and robustness are as important to genotype calling as is the accuracy and call rate that are usually used to evaluate performance of genotype-calling algorithms. Birdseed was developed and is being used to make genotype calling for the Affymetrix Genome-Wide Human SNP Array 6.0, the chip used in the current GWAS studies [1 8]. There are two different versions of the algorithm. Therefore, it is important for scientists to know whether the variations in genotypes called by different versions can be a potential source causing Type I and II errors in GWAS results. Our study demonstrated that the different versions of Birdseed generate slightly inconsistent genotypes which, in turn, propagate to the downstream

association analysis. Therefore, it could be a potential source of Type I and II errors in GWAS results. However, it was found that the extent of inconsistence in genotypes and in associated SNPs is much less than what we previously found for the batch effect of calling algorithm BRLMM [10].

A heterozygous genotype carries a rare allele. Therefore, the robustness of calling heterozygous genotypes reduces Type I and II errors in GWAS. Our studies demonstrated that variation in heterozygous genotypes between the two versions was larger than variation in homozygous genotypes, consistent with our previous finding [10].

Prior to association tests, QC is needed to ensure that high quality data are used in the association analysis. In addition to evaluating the consistency of sex-linked (X-chromosome) SNPs, minor allele frequency, testing on HWE, heterozygosity checking, the QC process usually includes call rates per SNP and per sample to discard samples and markers of low quality. Our results showed that the distributions of missing call rates for SNPs and samples were different for a same dataset because of the large difference in numbers of samples (270) and SNPs (906,600). In GWAS, the number of samples is usually much smaller than the number of SNPs. The call rate per SNP is usually lower than the call rate per sample. Therefore the same cut-off value for call rate per sample and per marker for QC (as in most practices) is not the best choice. Our observation suggests that using a lower cut-off of call rate per SNP compared to per sample may avoid losing some true associated SNPs in GWAS.

References

1. Shi J, Levinson DF, Duan J et al (2009) Common variants on chromosome 6p22.1 are associated with schizophrenia. Nature 460:753 757
2. Ikram MA, Seshadri S, Bis JC et al (2009) Genomewide association studies of stroke. N Engl J Med 360:1718 1728
3. Woodward OM, Köttgen A, Coresh J et al (2009) Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. Proc Natl Acad Sci USA 106:10338 10342
4. Erdmann J, Grosshennig A, Braund PS et al (2009) New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nat Genet 41:280 282
5. Myocardial Infarction Genetics Consortium, Kathiresan S, Voight BF et al (2009) Genome wide association of early onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat Genet 41:334 341
6. Zheng W, Long J, Gao YT et al (2009) Genome wide association study identifies a new breast cancer susceptibility locus at 6q25.1. Nat Genet 41:334 341
7. Köttgen A, Glazer NL, Dehghan A et al (2009) Multiple loci associated with indices of renal function and chronic kidney disease. Nat Genet 41:712 717
8. Kanetsky PA, Mitra N, Vardhanabhuti S et al (2009) Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. Nat Genet 41:811 815
9. http://www.affymetrix.com/products/software/specific/ birdseed algorithm.affx.
10. Hong H, Su Z, Ge W et al (2008) Assessing batch effects of genotype calling algorithm BRLMM for the Affymetrix GeneChip Human Mapping 500K Array Set using 270 HapMap samples. BMC Bioinform 9:S17

Chapter 41
An Overview of the BioExtract Server: A Distributed, Web-Based System for Genomic Analysis

C.M. Lushbough and V.P. Brendel

Abstract Genome research is becoming increasingly dependent on access to multiple, distributed data sources, and bioinformatic tools. The importance of integration across distributed databases and Web services will continue to grow as the number of requisite resources expands. Use of bioinformatic workflows has seen considerable growth in recent years as scientific research becomes increasingly dependent on the analysis of large sets of data and the use of distributed resources. The BioExtract Server (http://bioextract.org) is a Web-based system designed to aid researchers in the analysis of distributed genomic data by providing a platform to facilitate the creation of bioinformatic workflows. Scientific workflows are created within the system by recording the analytic tasks preformed by researchers. These steps may include querying multiple data sources, saving query results as searchable data extracts, and executing local and Web-accessible analytic tools. The series of recorded tasks can be saved as a computational workflow simply by providing a name and description.

Keywords Database integration · Genomic analysis · Scientific provenance · Scientific workflows · Web services

41.1 Introduction

Bioinformatic workflows generally represent automated complex scientific processes involving the integration of distributed data sources and analytic tools. With the exponential growth in data and computational resources, workflows are becoming increasingly important to the achievement of scientific advances. They can improve the process of scientific experiments by making computations explicit

C.M. Lushbough (✉)
Department of Computer Science, University of South Dakota, Vermillion, SD, USA
e mail: carol.lushbough@usd.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 41,

and underscoring data flow [1]. Workflow creation frequently requires specialized expertise and time-consuming development, but once created workflows can be useful to many researchers. Furthermore, there is a growing demand for the ability to collaborate in a geographically distributed environment. Therefore, it is important to provide a system that will permit researchers to share distributed resources through Web-based collaborative groups.

The BioExtract Server is a publicly accessible Web-based distributed service that gives researchers access to a variety of heterogeneous biomolecular data sources, analytic tools, and workflows. It offers a central distribution point for uniformly formatted genomic data from a variety of resources, including Web services, community databases, and specialized datasets. The basic operations of the BioExtract Server allow researchers via their Web browsers to: select data sources; flexibly query the sources with a full range of relational operators; determine download formats for their resulting extracts; execute analytic tools; create workflows; share research results; and name and keep query results persistent for reuse. One of the primary goals of the BioExtract Server is to provide researchers with an accessible, easy-to-use platform for the development and sharing of bioinformatic workflows.

41.2 BioExtract Server Functionality

41.2.1 Querying the Data Resources

BioExtract Server data queries are performed by first selecting from a list of available data resources presented on the *Query* tab of the system (see Fig. 41.1). The data resources available to researchers are classified as either publicly available data sources or previously created, private data extracts. The BioExtract Server's public data sources represent distributed community databases, curated databases and data repositories. They are implemented as relational databases, data warehouses, and data sources accessed through Web services.

A BioExtract Server data extract is a group of records resulting from the execution of a query or analytic tool and is privately owned by a researcher. All functionality in the BioExtract Server that can be applied to data sources may also be applied to these extracts (e.g., query, export, analyze). By having the ability to save data extracts, researchers may share and subsequently analyze subsets of data. Reproducibility and verification of results is one of the primary principles of the scientific method and being able to store data persistently helps in this effort.

Within the system, queries may be applied to multiple data resources simultaneously. The system has a defined global ontology of search terms to which each data resource maps its local querying capabilities. The results of a query are displayed on the *Extracts* tab. This set of data may be filtered, saved, exported and/or used as input into analytic tools. By saving the results of a query to a data extract, the researcher is able to subsequently query that dataset or use it as an input into an analytic tool. In addition, the researcher is able to accurately report the dataset used in a specific scientific process.

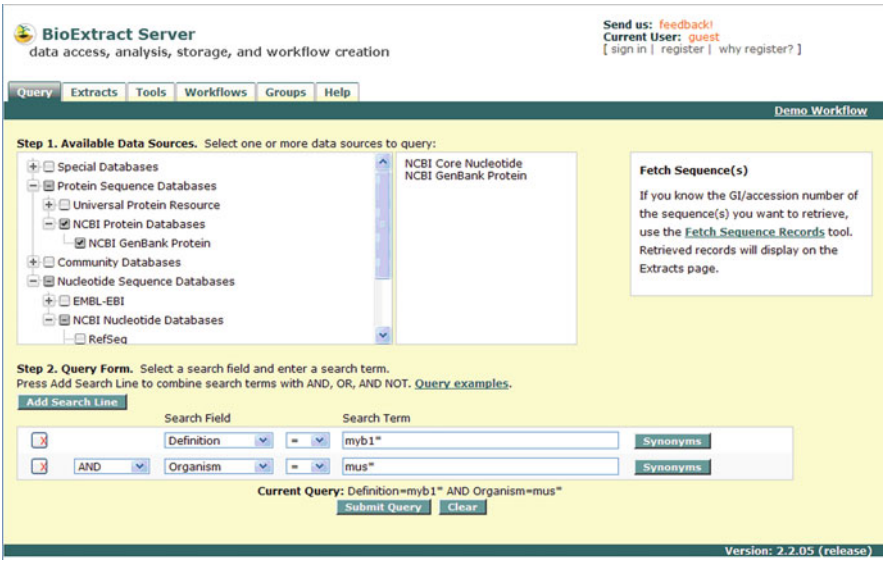


Fig. 41.1 Data sources. Data sources are available to researchers through the *Query* tab of the BioExtract Server. The GUI controls at the bottom of the screen provide an example of a Boolean query retrieving mouse records from NCBI Core nucleotide database related to myb1* genes

41.2.2 Analyzing Data

Bioinformatic analytic tools are made available to researchers on the *Tools* tab of the BioExtract Server. Researchers have the option of executing, adding, or modifying analytic tools.

41.2.2.1 Executing an Analytic Tool

The BioExtract Server incorporates a set of commonly used analytic tools allowing data to be processed against a number of algorithms to automatically identify, correct, and annotate it for many of the most common problems found in sequence misalignments or putative sequence identifications. Examples of tools included in the list are BLAST [2], ClustalW [3], and VMatch [4]. Figure 41.2 displays an example of accessing an analytic tool within the system.

To execute an analytic tool within the BioExtract Server, researchers select a tool from the list. A GUI is displayed, prompting the user for input and tool parameter settings. Depending on the analytic tool selected, the input into the tool may be entered directly, uploaded from a local file, provided by the current data extract, or may be the output from a previously executed tool. If the researcher chooses to use the current data extract as input into a tool, by default the system will convert the result set to a file in FASTA format (Pearson format). If the tool requires

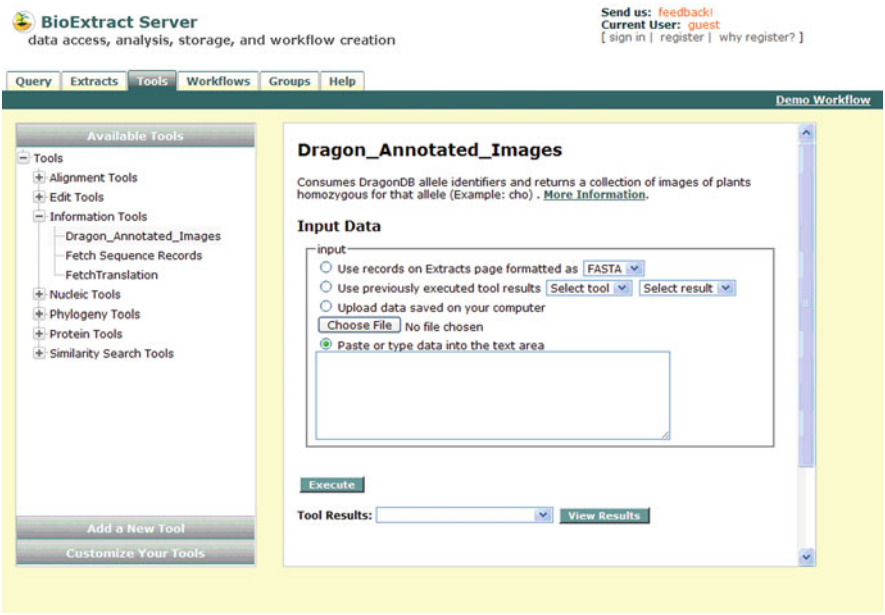


Fig. 41.2 Analytic tools. The list of available analytic tools is displayed on the *Tools* tab. Additional tools may be inserted into this tree by selected them from the list of Web services accessible through the *Add a New Tool* functionality or by adding a local tool

that the input be in a different format, the researcher must explicitly select from a list of available formats. Upon completion of tool execution, the tool's output may be viewed, saved locally, and used as input into another analytic tool.

41.2.2.2 Adding an Analytic Tool

Because of the large number of bioinformatics tools, it is important to allow researchers to have access to those tools that they deem most appropriate for their specific research. To that end, the BioExtract Server provides the researcher with the ability to select tools from a large list of Web services as well as integrate tools residing on their local machine.

Researchers may add a tool to their list of favorite tools by selecting the *Add a New Tool* bar located on the *Tools* tab. Clicking the *Save Tool* button will cause the tool to be added to the researcher's list of favorite tools under the *My Tools* heading and become available for execution. Tool attributes may be modified either before the tool is saved or after through the *Customize Your Tools* option.

The BioExtract Server provides functionality to allow researchers to add local tools to the system. Local tools are essentially command line programs that reside

on a researcher's workstation. The researcher must supply the location of the tool on the local machine as well as the command line parameter requirements. The destination of the output files must also be specified.

41.2.3 Workflows

A workshop sponsored by the National Science Foundation (NSF) dealing with the challenges of scientific workflows reported that workflows can play a valuable role in scientific work by providing a formal and declarative representation of complex scientific processes that can then be managed efficiently through their lifecycle from assembly to execution and sharing [5].

The NSF workshop report goes on to state that workflows are important because they can capture complex analysis processes at various levels of abstraction. They provide the provenance information necessary for scientific reproducibility, result publication, and result sharing among collaborators [5].

41.2.3.1 Creating a Workflow

BioExtract Server workflows are created specifically to compose and execute a series of analytic and data management steps within the genomic and proteomic domains of science. Researchers create a workflow implicitly by working with the system. Steps such as saving a data extract, executing an analytic tool, and querying a data resource are saved in the background. At any point, a workflow can be created simply by entering a name and description through the *Create and Import Workflows* node on the *Workflow* tab.

Workflows within the BioExtract Server are implemented as directed acyclic graphs (DAGs). Workflow nodes represent processes, and paths signify data flow dependencies. For example, if the execution of an analytic tool specifies that the input will be the current data extract (i.e., results of a database query) then that tool cannot execute until after the query has completed. Another example of a data dependency is when one tool's input is drawn from the output of another.

41.2.3.2 Executing a Workflow

A BioExtract Server workflow is executed on the *Workflows* tab of the system. As a workflow executes, the state of the nodes changes in order to supply feedback to the user (see Fig. 41.3). Output from a workflow step can be viewed after the step has completed execution by clicking on the circle or square located on the graph.

Researchers can interact with the workflow execution by clicking on the pause icon (⏸) in the upper right corner of a step node in the graph. For example, pausing a

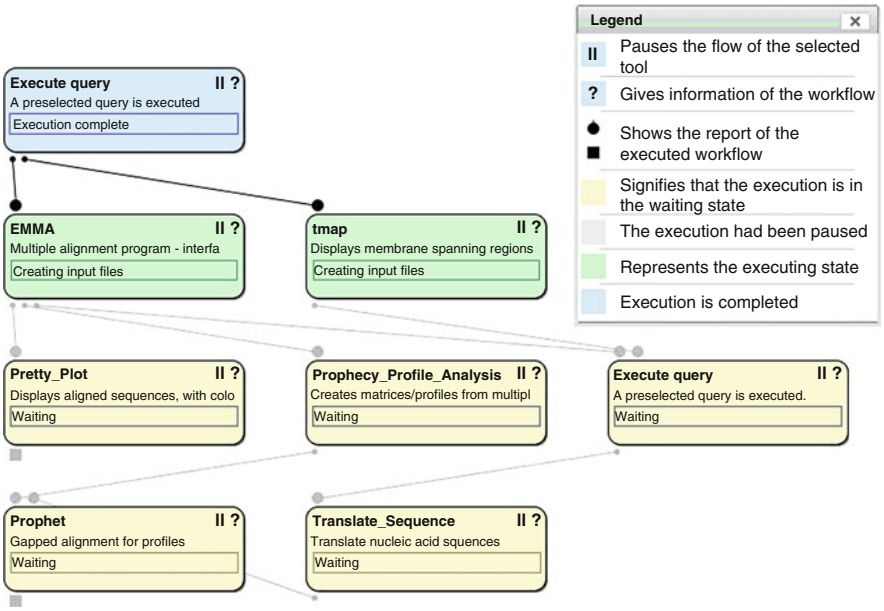


Fig. 41.3 BioExtract Server workflow. Workflows within the BioExtract Server are represented as Directed Acyclic Graphs (DAGs). As a workflow executes, changes in the state of a processes are communicated to the user both by change in color and label

step might be useful when a researcher would like to filter the results of a query before using it as input into an analytic step.

The BioExtract Server workflow reports give details about the workflow including analytic tool parameter settings, queries used in searching databases, and input/output information. The intent of these workflow reports is to keep a record of the workflow execution and to help others reproduce the work.

41.2.3.3 Modifying a Workflow

One of the primary advantages of creating a workflow is the ability to execute it numerous times, possibly with different input, parameter settings, and/or queries. There are three specific types of processes that comprise a BioExtract Server workflow: executing a query, saving a data extract, and executing an analytic tool. Query processes are modified by changing the query and/or selected data resources, whereas saving data extract processes can be altered by modifying the extract's name.

The options available for modifying an analytic tool process are partially dependent on how the tool was originally executed. If the researcher provided the input into the tool by uploading a file or entering the data into the input text box, then the input data can later be modified. If the input came from the use of the

current data extract or a previously executed tool, then the researcher does not have the option of changing the input.

41.2.3.4 Sharing Workflows

A bioinformatic workflow is a representation of a scientific process and potentially is of interest to colleagues and other researchers in the scientific community. Publishing a workflow along with the results of an experiment allows the provenance of results to be understood and reproduced [6]. Having the ability to share workflows can benefit researchers because they are able to take advantage of the effort others have put forth in their creation.

The BioExtract Server permits the sharing of scientific workflows in two ways. Functionality is built into the system that allows researchers to create groups and invite collaborators to become members. Through these groups, researchers can share workflows, analytic tools, and data extracts.

Workflows can also be shared with the research community through MyExperiment at http://www.myexperiment.org. MyExperiment is a virtual research environment designed to provide a mechanism for collaboration and sharing workflows. This was achieved by adopting a social Web approach, which is tailored to the particular needs of the scientist [7]. The BioExtract Server takes advantage of the functionality made available at MyExperiment in order to facilitate the publishing and sharing of BioExtract Server workflows.

41.3 Comparisons with Other Approaches

Many available biological data and analytic tool integration systems offer powerful, comprehensive, workflow creation functionality. Taverna [8], Kepler [9], Triana [10], and Trident (http://research.microsoft.com/en-us/collaboration/tools/trident.aspx) workflow management systems are primarily stand-alone applications requiring installation of client software. Through a "drag-and-drop" User Interface (UI), these systems allow researchers to develop complex workflows from distributed and local resources.

BioBike and Galaxy are primarily Web-based workflow management systems. BioBike (http://nostoc.stanford.edu/Docs/) offers a programming interface that enables researcher to combine tools, data, and knowledge in complex ways [11]. Galaxy (http://galaxy.psu.edu/) is an open source workflow system developed at Penn State University. It is an easy-to-use system that allows researchers to create complex workflows in an intuitive way that can be shared, cloned, and edited.

The BioExtract Server provides a potentially simpler facility for creating workflows than other systems. While researchers work with the system (e.g., perform queries, save data extracts, and/or apply analytic tools to data extracts) each of their

steps is automatically recorded. These steps are saved as a workflow when the user enters a name and description. The BioExtract Server is entirely Web based, requiring no client software installation and making it available anywhere where there is an Internet connection. It is the only Web-based system that provides researchers with the ability to choose from a large list of distributed resources or integrate their favorite local analytic tool. By having the ability to save searchable data extracts and view workflow provenance, users are able to accurately reproduce research results. Researchers are able to pause and resume processes during workflow execution permitting them to dynamically filter data input.

41.4 Conclusions and Future Direction

The BioExtract Server is a generalized, flexible, distributed system designed to consolidate and serve data subsets from accessible, heterogeneous, biomolecular databases. It offers a central distribution point for uniformly formatted data from various data sources. Some of the primary strengths of the system include: easy addition of analytic tools; the option of storing data subsets; the ability to treat subsets of data as integrated data sources; and the ability to save, execute, export, import, share, clone, and modify workflows.

The integration of provenance support promises to be a significant advancement in scientific workflow management systems for the verification, validation, and replication of scientific experiments [12]. Metadata provenance systems record such information as how, where, when, why, and by whom the research results were achieved [13]. The first provenance challenge was initially defined by Simon Miles and Luc Moreau (University of Southampton) and Mike Wilde and Ian Foster (University of Chicago/Argonne National Laboratory) in May 2006 and later became a community effort intended to help researchers understand the expressiveness of provenance representation and identify the necessary capabilities of provenance systems [14]. Through this and subsequent efforts, key provenance querying capabilities were defined.

The BioExtract Server currently offers functionality allowing researchers to generate workflow documents reporting information such as author, time of execution, database queries, analytic tools used, input/output values, and parameter setting. To expand on this module, the Open Provenance Model (OPM) will be integrated into the system to expand on workflow information. With this model, provenance objects are represented by annotated DAGs [15]. This will permit queries such as *Find the process that led to sequence AB003498 to be included in the set of unique sequences* or *Find all invocations of process blastn using a word size of 11*. Using the OPM will also promote inter-operability with other systems allowing for the exchange of provenance information.

Acknowledgments The BioExtract Server project is currently supported in part by the National Science Foundation grant DBI 0606909.

References

1. K. Verdi, H. Ellis, and M. Gryk, Conceptual level workflow modeling of scientific experi ments using NMR as a case study, BMC Bioinformatics, 8:31, 2007
2. S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, Gapped BLAST and PSI BLAST: a new generation of protein database search programs, Nucleic Acids Research, 25(17):3389 3402, 1997
3. R. Chenna, H. Sugawara, T. Koike, R. Lopez, T.J. Gibson, D.G. Higgins, and J.D. Thompson, Multiple sequence alignment with the Clustal series of programs, Nucleic Acids Research, 31(13):3497 3500, 2003
4. M.I. Abouelhoda, S. Kurtz, and E. Ohlebusch, The enhanced suffix array and its application to genome analysis, Lecture Notes in Computer Science, 2452:449 463, 2002. http://www.vmatch.de/
5. E. Deelman and Y. Gil, Workshop on the Challenges of Scientific Workflows; Sponsored by the National Science Foundation, http://vtcpc.isi.edu/wiki/images/3/3a/NSFWorkflowFinal.pdf, May 1 2, 2006
6. D. De Roure and C. Goble, Software design for empowering scientists, IEEE Software, 26(1): 88 95, 2009
7. D. De Roure, C. Goble, and R. Stevens, The design and realization of the *my*Experiment Virtual Research Environment for social sharing of workflows, Future Generation Computer Systems, 25(5):561 567, 2009. corrected proof available as: DOI http://dx.doi.org/10.1016/j.future.2008.06.010
8. D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. Pocock, P. Li, and T. Oinn, Taverna: a tool for building and running workflows of services, Nucleic Acids Research, 34(Web Server issue):W729 W732, 2006
9. B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E.A. Lee, J. Tao, and Y. Zhao, Scientific workflow management and the Kepler system, Concurrency and Computation: Practice & Experience, 18(10):1039 1065, 2006
10. A. Harrison, I. Taylor, I. Wang, and M. Shields, WS RF workflow in Triana, International Journal of High Performance Computing Applications (IJHPCA), 22(3):268 283, 2008
11. J. Elhai, A. Taton, J. Massar, J. Myers, M. Travers, J. Casey, M. Slupesky, and J. Shrager, BioBIKE: A Web based, programmable, integrated biological knowledge base, Nucleic Acids Research, 37(Web Server issue):W28 W32. doi10.1093, 2009
12. S. Bowers, T. McPhillips, B. Ludäscher, S.Cohen, and S. Davidson, A Model for user oriented data provenance in pipelined scientific workflows, Lecture Notes in Computer Science, Springer, Berlin, ISBN: 978 3 540 46302 3, pp 133 147
13. C. Goble, Position statement: musings on provenance, workflow and (semantic web) annotations for bioinformatics, Proceedings of the Workshop on Data Derivation and Provenance, 2002; http://people.cs.uchicago.edu/∼yongzh/papers/provenance workshop 3.doc
14. L. Moreau, B Ludäscher, I. Altintas, R. Barga, S. Bowers, , S. Callahan, G. Chin, B. Clifford, S. Cohen, S. Cohen Boulakia, S. Davidson, E. Deelman, L. Digiampietri, I. Foster, J. Freire, J. Frew, J. Futrelle, T. Gibson, Y. Gil, C. Goble, J. Golbeck, P. Groth, D. A. Holland, S. Jiang, J. Kim, D. Koop, A. Krenek, T. McPhillips, G. Mehta, S. Miles, D. Metzger, S. Munroe, J. Myers, B. Plale, N. Podhorszki, V. Ratnakar, E. Santos, C. Scheidegger, K. Schuchardt, M. Seltzer, Y. Simmhan, C. Silva, P. Slaughter, E. Stephan, R. Stevens, D. Turi, H. Vo, M. Wilde, J. Zhao, and Y. Zhao, The First Provenance Challenge, Concurrency and Computation: Practice & Experience, 20(5):409 418, 2008
15. L. Moreau, J. Futrelle, R. McGrath, J. Myers, and P. Pualson, The open provenance model: an overview, Lecture Notes in Computer Science, Springer, Berlin/Heidelberg, ISBN 978 3 540 89964 8, 5272:323 326, 2008

Chapter 42
A Deterministic DNA Database Search

A. Kheniche, A. Salhi, A. Harrison, and J.M. Dowden

Abstract This paper suggests a novel way for measuring the similarity between sequences of symbols from alphabets of small cardinality such as DNA and RNA sequences. The approach relies on finding one-to-one mappings between these sequences and a subset of the real numbers. Gaps in nonidentical sequences are easily detected. Computational illustrations on DNA sequences and a comparison with BLAST are included.

Keywords DNA · RNA · Similarity · Mapping · BLAST

42.1 Introduction

Demand for powerful tools to compare DNA, RNA, or protein sequences [1, 2] is increasing in step with the increase in the amount of genetic data gathered. A number of tools have been and are being developed in order to meet this demand; BLAST [3] is perhaps one of the most prominent. Although it has been proved to be both robust and efficient, there is room for improvement for the simple reason that larger and larger datasets have to be processed. Also, because it relies on stochastic search, the parameters of which have to be preset arbitrarily, there is no guarantee that it will detect what it is searching for even when the latter is present. Here, a deterministic algorithm to tackle the problem is presented. We believe that it has advantage over BLAST of being deterministic, and also that of not requiring parameter settings except for the rate of similarity the user is interested in to cut out unnecessary searches. An implementation of the suggested algorithm is compared to BLAST on DNA sequences.

A. Salhi (✉)
Department of Mathematical Sciences, The University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK
e mail: as@essex.ac.uk

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 42,

42.2 The Problem

It can be stated as follows. Given two sequences of alphanumeric symbols, how similar are they? It can be made more complex by requiring, for instance, that the comparison result flags out the places where there is no similarity. This is a standard problem in textual database management and operation; looking for a record in a textual database is indeed equivalent to the problem considered here. For simplicity of exposition, we will consider DNA sequences, i.e., sequences made up with the base symbols A, C, G, and T. In particular, we will be concerned with the following questions.

1. Given a DNA sequence q of length n and a reference sequence r of length m, $m \geq n$, does q occur in r?
2. If the answer to question 1 is "no", how much of sequence q can be found in any subsequence of r of length n and what are the positions of their similar symbols?

42.3 Solution Method

The answer to question 1 can be found if q could be turned into a real number as well as all subsequences of r of length n; equal numbers indicate similar sequences. Although the answer to question 2 is less straightforward, it is motivated by this same idea. How can this be achieved?

First, assign to each of the four bases a different digit from 0 to 9, as in $A = 1$, $C = 2$, $G = 3$, and $T = 4$, for example. Then, find an injective function f with good properties (to be stated explicitly later) that associates with any n-long DNA sequence s a *unique* real number $f(s)$. In this way the query sequence is, therefore, represented by a real number $f(q)$, and the reference sequence is assigned a vector of $m - n + 1$ real numbers, each being the value of the function f for the subsequence r_i of r that lies between the ith and the $(i + n - 1)$th base of the reference sequence. Having computed $f(q)$ and $f(r_i)$, where $i = 1, \ldots, m - n + 1$, similarity or otherwise between q and r_i can be detected by computing the difference $|f(r_i) - f(q)|$ for all i; In other words, by computing the vector $V = (|f(r_1) - f(q)|, |f(r_2) - f(q)|, \ldots, |f_{(rm\ n+1)} - f(q)|)$. Because of the property of injectivity, a zero at position i corresponds to a perfect similarity between the query sequence q and the subsequence r_i of r. But what if the difference is not zero? Could this nonzero value give some indication as to whether there is substantial similarity, little similarity, or very little similarity between the two sequences under consideration? The answer to this question will be considered later. For the time being, and before moving any further, let us establish that an injective function does exist.

42.3.1 Existence

Let a DNA sequence be encoded with the digits 1, 2, 3, and 4 standing for the bases A, C, G, and T, respectively. For example, $ATCGTA$ is encoded as (1, 4, 2, 3, 4, 1).

Lemma 3.1. *Let t be a transcendental number [4], and let $s = (a_1, a_2, \ldots, a_n)$ be an n-long DNA sequence. The function $f(s) = \sum_{i\ 1,\ldots,n} a_i t^{i\ 1}$ is injective.*

Proof. Suppose there are two sequences $s = (a_1, a_2, \ldots, a_n)$ and $s' = (a'_1, a'_2, \ldots, a'_n)$ so that $f(s) = f(s')$. This means that $\sum_{i\ 1,\ldots,n}(a_i - a_i)' t^i = 0$. Let $I = \{i / a_i \neq a'_i\}$. The polynomial $P(x) = \sum_{i \in I} (a_i - a'_i) x^{i\ 1}$ has t as a root by assumption. Since a transcendental number cannot be a root for any polynomial with nonnull rational coefficients, I is empty, i.e., for all i we have $a_i = a'_i$. Therefore, $s = s'$.

42.3.2 Note

For practical purposes and since we are dealing with powers of t, a reasonable choice of t would be something slightly bigger than 1, such as $\pi/3$, for example. However, it is difficult to characterize the nonzero differences, due to the relative difficulty in dealing with noninteger numbers. This points to giving preference to functions with integer ranges.

42.3.3 Functions with Better Properties

Lemma 3.2. *Let $t \geq 4$ be an integer. With the same notation as before, f is injective.*

Proof. Let $s = (a_1, a_2, \ldots, a_n)$ and $s' = (a'_1, a'_2, \ldots, a'_n)$ such that $s \neq s'$. Let i_0 be the largest number such that $a_{i0} \neq a'_{i0}$ and suppose, without loss of generality, that $a_{i0} > a'_{i0}$. Then, we have
$f(s) \geq t^0 + t^1 + \ldots + t^{i0\ 2} + a_{i0}\, t^{i0\ 1} + a_{i0+1} t^{i0} + \ldots + a_n t^{n\ 1}$ and
$f(s') \leq 4(t^0 + t^1 + \ldots + t^{i0\ 2}) + a'_{i0}\, t^{i0\ 1} + \ldots + a'_n t^{n\ 1}$.
Hence $f(s) - f(s') \geq (a_{i0} - a'_{i0}) t^{i0\ 1} - 3(t^0 + t^1 + \ldots + t^{i0\ 2})$.

The eliminated terms are equal (see definition of i_0). Using the formula of the sum of terms of a geometric sequence and the fact that $a_{i0} - a'_{i0} \geq 1$, we find that $f(s) - f(s') \geq t^{i0\ 1}\ \ 3(t^{i0\ 1}\ \ 1)/(t\ \ 1)$. Since $t \geq 4$, then $3/(t-1) \leq 1$ and thus $f(s) - f(s') \geq 1 > 0$. That is $f(s) \neq f(s')$.

With $t = 10$, we have noticed that the corresponding function possesses some good properties. Suppose that $f(q) - f(r_i) = x$, x being an n-digit number (when it has $< n$ digits, we append zeros to its left to make up the required length). We noticed that whenever there is a match between two bases in those sequences in a

specific position, the digit of x in the same position is either 0 or 9. The reason is that the appearance of 0 is obvious since a match corresponds to the appearance of the same digit in the two sequences. The appearance of a 9 can be understood by considering the subtraction operation when carried out by hand. For example, work out $431 - 234$ manually. Note, however, that for the same reason the observation is not reciprocal: 0 or 9 as a digit in x does not mean necessarily that there is a match at that position. Consider, for instance, $433 - 124$; neither the 0 nor the 9 means a match. This problem will be solved in the following section, but before that we should mention that by using $t = 10$, the value that this function assigns to a sequence could be obtained by just reversing the order of the digits that represent that sequence. For example, $f(2, 4, 1, 3, 4, 4, 2, 3, 3) = 332443142$. This suggests that we use the function $g(s) = \sum_{i\ 1,\ldots,n} a_i t^{n\ i}$. This function assigns to the sequence the value that corresponds to its digits without reversing the order. For example, $g(2, 4, 1, 3, 4, 4, 2, 3, 3) = 241344233$. We, therefore, no longer need this function; it is invoked implicitly when the sequence is translated into the numerical form. The whole of the q and r_i sequences considered as numbers, and the difference is calculated as usual.

42.3.4 A Better Encoding of Bases A, C, G, *and* T

We have yet to consider the question of the best choice of digits to assign to the four bases. It turns out that a choice such as 0, 2, 4, and 6, or any other four digits no two of which are consecutive, is better than the previous ones. The justification is simple: with that choice of encoding, a match between two bases in a specific position is detected if, and only if, there is a 0 or a 9 in that position in the numerical expression of x, the result of subtraction. This can be shown easily using the classical subtraction algorithm.

Example q = ATTCCG = 066224
r = GTCTCGTATGTCCG
= (462624, 626246, 262460, 624606, 246064, 460646, 606462, 064622, 646224)
V = (396400, 560022, 196236, 558382, 179840, 394422, 540238, 001602, 580000).

Here, a 0 or a 9 corresponds to a match between two bases at this position, and any other digit indicates a mismatch. A simpler approach is to consider each one of the four bases as a digit in a chosen number system such as $A = 10$, $C = 12$, $G = 14$, and $T = 16$ in the number system of base 17. This removes the need to translate the DNA sequences into numerical values, and keeps the properties of the algorithm intact. Here, however, instead of being concerned with the digits 0 and 9, we will be concerned with the 0 and the biggest digit in our number system, i.e., T in the last example. In the following algorithm, which converts the differences between the values of the query sequence q and subsequences r_i of r ($i = 1,\ldots,$

$m - n + 1$), d_i stands for the ith digit of x starting from the right, and V is a (0, 1) vector.

42.3.5 Algorithm

Let A, C, G, and T be integers in a chosen number system (where X is the digit with the largest value) such that no two of the digits are consecutive. Let Q be an n-digit query sequence, and R an m-digit reference sequence. Consider $R(i)$ to be the number composed of the ith n-digits of R, that is, the digits number i, $i + 1$,..., $i + n - 1$. The following algorithm finds the matches and mismatches between Q and $R(i)$ for $i = 1, \ldots, m - n + 1$.

```
BEGIN find match()
  FOR i   1 to m   n + 1 DO
      V (i)   |Q   R(i)|
      IF Number of V (i) digits < n THEN
         Append (n   number of V (i) digits) zeros to its right;
      ENDIF
      FOR j   1 to n DO
          IF The j^th digit of V (i) equals 0 or X THEN
               Match between the j^th base in Q and the (i + j)^th base in R;
          ELSE There is a mismatch;
          ENDIF
      ENDFOR
  ENDFOR
END
```

The above algorithm is obviously finite for finite sequences Q and R.

42.4 Computational Results

The suggested algorithm has been implemented in MatLab as DSS. We have worked with a query sequence and a reference sequence (real data) of length 30 and 189,777, respectively. We ran BLAST [3] (standalone) to compare the query sequence with the subsequences of the same length of the reference sequence several times, each time tuning program parameters such as word size (W), gap penalty (q), and expectation (e). As is known [5], BLAST gives very good results when parameter tuning is done well. This, however, is often only possible when information on the sequences involved is available. This means that the lack of information makes tuning at best arbitrary. Consequently, BLAST's results will, potentially, be not all that good. By this we mean that it may report *no sequence found* for a given similarity level, when in fact one or more sequences with that similarity level do exist in the reference sequence. In what follows, we give examples of output results from both programs.

42.4.1 BLAST with Default Parameter Values

In the following, BLAST detects only one case of significant similarity, when default parameters are used.

```
Length = 189777
Score = 22.3 bits (11), Expect = 0.70
Identities = 11/11 (100%)
Strand = Plus / Plus
Query: 16       tgcaccattcg 26
                 ||||||||||
Sbjct: 68783    tgcaccattcg 68793
Database: ref-seq.txt
Number of letters in database: 189,777
Number of sequences in database: 1
```

DSS which requires only one parameter, the rate of similarity, finds this sequence and others if the rate of similarity is set to 45%.

```
46.6667 percent match starting from base number
68768 match positions:
7 10 16 17 18 19 20 21 22 23 24 25 26 28
ttcctggccttccggtgcaccattcggtta
gaatcagggtgaatctgcaccattcgttcg
```

42.4.1.1 Word Size

Word size is 7 by default in BLAST. This means that if no matching block of contiguous bases of length 7 is found, it will not report any significant similarity even if the overall similarity is high. DSS, however, detects the similar sequence starting at 73637 position.

```
63.3333 percent match starting from base number
73637 match positions:
2 4 6 7 8 9 11 12 13 14 15 16 17 21 22 23 25 26 30
ttcctggccttccggtgcaccattcggtta
gtacaggccatccggtggtgcatgcgacaa
```

If we set the word size to 4 (the minimum identical block size), tune other parameter values such as the penalty gap (set to −3), and run BLAST again, then it will find good alignments as in the following example.

```
Length = 189777
Score = 20.6 bits (12), Expect = 1.1
Identities = 17/22 (77%)
Strand = Plus / Plus
Query: 1        ttcctggccttccggtgcacca 22
                 |||| ||||||||  | || ||
Sbjct: 131915   ttcccggccttcctctccagca 131936
```

This is also detected by DSS.

```
  60 percent match starting from base number
131915
  matching positions:
  1 2 3 4 6 7 8 9 10 11 12 13 16 18 19 21 22 30
  ttcctggccttccggtgcaccattcggtta
  ttcccggccttcctctccagcaggttccca
```

Here is an example for which BLAST fails.

```
  70 percent match starting from base number
185373
  match positions:
  1 2 3 5 6 7 9 10 11 13 14 15 17 18 19 21 22 23 25 26 27
  ttcctggccttccggtgcaccattcggtta
  ttcatgggcttgcggcgcatcatccggcac
```

Clearly, despite the good overall rate of similarity (21 of 30 bases), this is not spotted by BLAST. This is due to the default BLAST word size of 4, while the blocks of similarity in the example are at most of length 3 (as in the blocks $1-2-3$ and $5-6-7$).

42.4.2 Expectation Value

Expectation is another parameter that has an important bearing on what BLAST does and does not report. Value 10 is the default for this parameter. With this value, BLAST does not search thoroughly. If increased, it will search longer at the cost of increasing the search time, with no guarantee of finding anything even when it is present.

42.5 Conclusion

We have shown that the problem of comparing two sequences can be reduced to the simpler problem of looking for two digits in a vector of integer numbers. This then led to the design of a simple deterministic algorithm that can find all sequences similar to a given query sequence given a specified rate of similarity. This is not always possible with BLAST as it is based on stochastic search and it also relies heavily on a number of parameters, the setting of which is hard to get right. Its default parameters can be shown to be problematic, particularly when a new reference sequence is used. We can safely say that DSS is more robust and more reliable than BLAST. However, BLAST is faster for many reasons, including the fact that it is based on a nondeterministic algorithm (nondeterminism is often synonymous with high speed). DSS exists currently as a prototype written in Matlab, with about five dozen lines of code. A conversion into C++ is in progress and it is expected that the resulting implementation will run much faster.

References

1. C.A. Orengo and D.T. Jones, (2003), “Bioinformatics: Genes, Proteins & Computers”, BIOS Scientific Publishers Ltd, Oxford, UK.
2. J. Xiong, (2006), “Essential Bioinformatics”, Cambridge University Press.
3. http://www.ncbi.nlm.nih.gov/BLAST/
4. G. Birkhoff, and S. MacLane, (1977), “A Survey of Modern Algebra”, Macmillan Publishing Co., New York, 4th Ed.
5. http://www.clcbio.com/index.php?id 995

Chapter 43
Taxonomic Parsing of Bacteriophages Using Core Genes and In Silico Proteome-Based CGUG and Applications to Small Bacterial Genomes

Padmanabhan Mahadevan and Donald Seto

Abstract A combined genomics and in situ proteomics approach can be used to determine and classify the relatedness of organisms. The common set of proteins shared within a group of genomes is encoded by the "core" set of genes, which is increasingly recognized as a metric for parsing viral and bacterial species. These can be described by the concept of a "pan-genome", which consists of this "core" set and a "dispensable" set, i.e., genes found in one or more but not all organisms in the grouping. "CoreGenesUniqueGenes" (CGUG) is a web-based tool that determines this core set of proteins in a set of genomes as well as parses the dispensable set of unique proteins in a pair of viral or small bacterial genomes. This proteome-based methodology is validated using bacteriophages, aiding the reevaluation of current classifications of bacteriophages. The utility of CGUG in the analysis of small bacterial genomes and the annotation of hypothetical proteins is also presented.

Keywords Bacterial genomes · CGUG · Core genes · Genomics · Pan-genome · Proteomics approach · Web-based tool

43.1 Introduction

The continuing and predicted explosion of whole genome data is staggering, with de novo determinations of genomes from previously unsequenced genomes as well as multiple determinations of related genomes, e.g., multiple *Escherichia coli* genomes. If this can be pictured as a "tsunami" to place visually the enormity of the data flow and to understand the immense amount of data that need to be mined (and given a relatively sparse array of software tools to do so), then the parallel and earlier capture of bacteriophage and small bacterial whole genome data may be described as a

P. Mahadevan (✉)
Department of Biological Sciences, Vanderbilt University, Nashville, TN 37235, USA
e mail: padmahadevan@gmail.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 43,

"seiche," both in terms of the smaller sizes of the genomes they represent and the apparent smaller "stage" upon which they play. Nonetheless, bacteriophages are long-studied and remain valuable for their contributions to basic biology knowledge.

"CoreGenesUniqueGenes" (CGUG) is a user-friendly web-based tool that performs "on-the-fly" protein protein analyses in order to determine the "core" set of proteins of a set of small genomes. It is a redevelopment of an existing tool, CoreGenes [1], and is an enhancement based on suggestions from the wet-bench bacteriophage research community. CGUG has been validated in two different functions, as will be presented in this report. First is in the annotating and comparing of small bacterial genomes, ca. 2 Mb and less. The second is in reexamining the complex and long-standing relationships of bacteriophages. The latter is very important, given the extensive genetic and biochemical studies in the past. Integrating genomics and in silico proteomics with that data gives another dimension to the value of genome determinations and databases.

The International Committee on the Taxonomy of Viruses (ICTV) is an organization of researchers in a particular field that considers the classification of viruses in their field [2], for example, the bacteriophages. A current classification system accepted by the ICTV is based on a hierarchical system that groups viruses by the characteristics that they share. These characteristics reflect the technologies and methodologies that were available at the time, and may be limited by the same. Specifically, in the case of bacteriophages, in the past and currently, these metrics include morphology, genome size, host range, proteins (immunochemistry), and the physical characteristics of the genome, e.g., whether the genome is linear, circular, or supercoiled [3]. In response to the newly available genome and the resulting proteome data, researchers are now considering these data in the scheme of viral relationships and classifications.

As an example of this, a proteome-based approach has been used recently to reexamine and suggest a reclassification of bacteriophages by computationally building a proteome tree based on BLASTP analyses [4]. The disadvantage of this approach is that there was no readily accessible tool that performs this analysis and generates the proteome tree for inspection and analysis. Ideally, a web-based tool that performs the BLASTP analyses and produces easily interpretable output would be very useful to wet-bench biologists. CoreGenes is a tool that was developed earlier and is used currently to determine the "core" or common set of proteins in a set of genomes. CGUG is an upgrade and a modification that incorporates suggestions from several members of the ICTV who were interested in using the software for their research. Core sets of genes have been used to reconstruct ancestral genomes [5], organismal phylogenies [6], and organism classifications [7]. It can be and has been applied to the bacteriophage studies [7].

43.2 CoreGenesUniqueGenes Algorithm

CGUG is implemented in the Java programming language, using a combination of servlets and HTML. The algorithm is based on the GeneOrder algorithm to determine gene order and synteny [8]. The algorithm accepts between two and five genome

accession numbers. These genomes are then retrieved from GenBank, and the protein sequences are parsed and extracted from the GenBank files. One genome is designated as the reference genome, and the rest are the query genomes. If only two genome accession numbers are entered, all against all protein similarity analyses are performed for each protein of the query genome against the reference genome using WUBLASTP from the WUBLAST package. The results from the protein similarity analyses are parsed according to a previously specified threshold BLASTP score (default = "75"). If the scores from the protein alignments are equal to or greater than the threshold score, the protein pairs are stored and a table of proteins common to the two genomes (that is, "core" proteins) is created as the output.

In the case of more than two genomes, a consensus genome is created from the results of the similarity analysis between the reference genome and the first query genome. This consensus genome becomes the new reference genome. Protein similarity analyses are performed with the second query genome against this new reference genome using WUBLASTP. The algorithm proceeds in an iterative manner, analyzing the subsequent query genomes against the newly created reference genomes. The final output is a set of "core" proteins between a set of up to five small genomes in the form of a table. The unique proteins to a pair of genomes are also presented in tabular format below the "core" protein table. CGUG is available at http://binf.gmu.edu:8080/CoreGenes3.0 and can also be accessed at http://binf.gmu.edu/geneorder.html.

At the request of wet-bench bacteriophage researchers, a homolog count function has also been implemented. This is displayed as the sum of proteins in each column of the table. In addition, in recognition that some genomes may be newly sequenced and desired to be analyzed before submission into public databases, custom data can also be entered for CGUG analysis using the "custom data" interface.

The groups analyzed to demonstrate the function of reconfirming and verifying existing (ICTV) genera are the T7-like bacteriophages (*Escherichia* phage T7, *Yersinia* phage φA1122, *Yersinia* Berlin, *Escherichia* phage T3, *Yersinia* phage φYeO3-12, *Escherichia* phage K1F, *Vibrio* phage VP4, and *Pseudomonas* phage gh-1). Based solely on the GC content and length, it may be difficult to gain meaningful information about the bacteriophages in order to classify them. But an analysis of their proteomes using CGUG yields more informative results. All CGUG analyses are performed at the default threshold setting of "75."

43.3 Results and Discussion

43.3.1 Bacteriophage Genomes Application: Verification of Existing Genus of the Podoviridae

Bacteriophages are notoriously difficult to classify because of their genome variations due to horizontal transfers [9]. Recently, several researchers, who are members of the ICTV, have used CoreGenes to reanalyze and reclassify

bacteriophages using proteome data [7]. Their previous experiences in applying CoreGenes to earlier work gave insights as to what additional features were needed; these have been incorporated into CGUG and integrated into the later portions of their analyses of these genomes. The bacteriophage families examined to date include the Podoviridae and the Myoviridae [10]. To illustrate the usefulness of CGUG, the T7 genus reanalysis data from a recent collaborator [7] are presented here to emphasize the utility of this approach in the reclassification of the bacteriophages.

The T7-like phages constitute a genus of the Podoviridae family. Using a homologous protein cutoff of 40%, CGUG analysis reveals that the members of the T7-like phages all share greater than 40% homologous proteins with bacteriophage T7. This cutoff is used because it has been used previously to produce clear relationships between bacteriophage genera of the Podoviridae [7]. This shared protein analysis reconfirms and verifies the existing ICTV classification of these phages as belonging to the T7-like phage genus.

43.3.2 CGUG Analysis and Reclassification of T7, P22, and Lambda Bacteriophages

The tailed bacteriophages T7 and P22 are currently and traditionally classified as belonging to the Podoviridae family due to a shared presence of short tails [3]. However, P22 and lambda are more related to each other than to T7, based on the CGUG in silico proteomics analysis. Therefore, P22 should be moved to and classified in the Siphoviridae to which lambda belongs. CGUG analysis reveals that T7 and P22 share only two proteins. T7 and lambda also share only two proteins. The percent identities of these shared proteins are very low, ranging from 14 to 20%. In contrast, P22 and lambda share 19 proteins, several of which show high percent identities (>80%).

Thus, these results show that P22 is more related to lambda than to T7, given the genome and the proteome data. The whole genome percent identity between T7 and P22 is 47.5%, while the identity between T7 and lambda is approximately 46%. This nucleotide level does not provide meaningful information about the relatedness of these genomes. Similarly, the percent GC content of these genomes (between 48 and 50%) is also not informative. The CGUG analysis looks at the in silico proteome, and the relatedness of the T7, P22, and lambda phages can be assessed more meaningfully using this information.

43.3.3 Application to Niche Specific Bacterial Strains

The transcriptomes of two closely related strains of *Burkholderia cenocepacia* were recently mapped by high-throughput sequencing [11]. *B. cenocepacia* strain AU1054 is an opportunistic pathogen found in cystic fibrosis patients, while the

B. cenocepacia strain HI2424 is a soil-dwelling organism. Despite the fact that these two organisms live in very different environments, they share 99.8% nucleotide identity in their conserved genes. Even in such highly related bacterial strains, there are cases of hypothetical proteins in one strain that are not annotated with a function that is related to annotated proteins in the other strain. One example is the case of the 3-carboxy muconate cyclase-like protein found in AU1054, while the counterpart protein is annotated as hypothetical in HI2424. These two proteins share only 12.1% identity to each other. However, their lengths are not very dissimilar, and analyses using PFAM (http://pfam.sanger.ac.uk) show that the hypothetical protein appears to contain a "3-carboxy-cis,cis-muconate lactonizing enzyme" domain. Therefore, it is possible that the hypothetical protein shares a function similar to that of the annotated protein in AU1054.

In contrast, there is a case where the hypothetical protein is in AU1054, while the counterpart annotated protein is in HI2424. This annotated protein is an amidohydrolase. The percent identity between these proteins is 19.7%, and their lengths are similar as well at 319 amino acids for the amidohydrolase and 281 amino acids for the hypothetical protein. Functional prediction of this hypothetical protein using the SVMProt server [12] indicates that it belongs to the iron-binding protein family, with a probability of correct classification of 78.4%. Indeed, several enzymes in the amidohydrolase superfamily bind metal ions [13]. The catalytic activity of one amidohydrolase, cytosine deaminase, is highest with iron [14]. Analysis using the Phyre fold recognition server [15] indicates that the hypothetical protein apparently belongs to the "N-terminal nucleophile aminohydrolases". However, it must be noted that the E values from Phyre and the percent precision are not significant. Nevertheless, this taken together with the fact that CGUG puts these two proteins together suggests that the hypothetical protein may indeed be an amidohydrolase. Further wet-bench experiments are needed to confirm this prediction.

43.4 Conclusions

Whole genome and in silico proteome analysis tools are necessary to obtain meaningful information about organisms when the nucleotide data are not especially informative. CGUG is a user-friendly tool that is especially suited to wet-bench biologists with little interests in compiling code or deconstructing software in order to analyze proteomes from small genomes such as those from viruses, chloroplasts, and mitochondria, as well as small bacterial genomes. The utility of CGUG is illustrated in the verification of current classifications of bacteriophages and in the proposal of new classifications based on CGUG and other data. The current ICTV classification of the T7-like phages is verified using the whole proteome CGUG analysis.

Currently, the T7 and P22 phages are classified in the Podoviridae, while lambda phage is classified in the Siphoviridae. The in silico proteome analysis by CGUG shows that P22 is more related to lambda than to T7. That is, P22 and lambda share

more proteins with each other than they both do with T7. This means that P22 should be classified in the Siphoviridae like lambda. In this case, proteome analysis is more meaningful than the shared morphological similarity between T7 and P22.

The usefulness of CGUG in annotating hypothetical proteins is illustrated in the case of the two closely related niche-specific *Burkholderia* bacteria. The assignment of putative functions to these hypothetical proteins will provide a starting point to help wet-bench scientists confirm these predictions in the lab.

The web-based nature of CGUG makes it accessible and useful to wet-bench-based biologists. The "on-the-fly" nature of the tool avoids the limitations of precomputed data and allows the retrieval of the genomes directly from GenBank. The ability to enter custom data further enhances the tool immensely. These types of software tools allow biologists to take full advantage of the expanding genome sequence databases.

Acknowledgments We thank Drs. Andrew Kropinski and Rob Lavigne for their very helpful suggestions of features to add to CoreGenes. We thank Jason Seto for editorial comments and Chris Ryan for maintaining the server on which the software is hosted. This paper is based on Mahadevan and Seto (BIOCOMP '09).

References

1. Zafar N, Mazumder R, Seto D (2002) CoreGenes: a computational tool for identifying and cataloging "core" genes in a set of small genomes. BMC bioinformatics **3**, 12
2. Fauquet CM, Fargette D (2005) International Committee on Taxonomy of Viruses and the 3,142 unassigned species. Virology journal **2**, 64
3. Nelson D (2004) Phage taxonomy: we agree to disagree. Journal of bacteriology **186**, 7029 7031
4. Rohwer F, Edwards R (2002) The Phage Proteomic Tree: a genome based taxonomy for phage. Journal of bacteriology **184**, 4529 4535
5. Koonin EV (2003) Comparative genomics, minimal gene sets and the last universal common ancestor. Nature reviews **1**, 127 136
6. Lerat E, Daubin V, Moran NA (2003) From gene trees to organismal phylogeny in prokar yotes: the case of the gamma Proteobacteria. PLoS biology **1**, E19
7. Lavigne R, Seto D, Mahadevan P, Ackermann H W, Kropinski AM (2008) Unifying classical and molecular taxonomic classification: analysis of the Podoviridae using BLASTP based tools. Research in microbiology **159**, 406 414
8. Mazumder R, Kolaskar A, Seto D (2001) GeneOrder: comparing the order of genes in small genomes. Bioinformatics (Oxford, England) **17**, 162 166
9. Hendrix RW, Smith MC, Burns RN, Ford ME, Hatfull GF (1999) Evolutionary relationships among diverse bacteriophages and prophages: all the world's a phage. Proceedings of the National Academy of Sciences of the United States of America **96**, 2192 2197
10. Lavigne R, Darius P, Summer EJ, Seto D, Mahadevan P, Nilsson AS, Ackermann HW, Kropinski AM (2009) Classification of Myoviridae bacteriophages using protein sequence similarity. BMC microbiology **9**, 224
11. Yoder Himes DR, Chain PS, Zhu Y, Wurtzel O, Rubin EM, Tiedje JM, Sorek R (2009) Mapping the Burkholderia cenocepacia niche response via high throughput sequencing. Pro ceedings of the National Academy of Sciences of the United States of America **106**, 3976 3981

12. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ (2003) SVM Prot: web based support vector machine software for functional classification of a protein from its primary sequence. Nucleic acids research **31**, 3692 3697
13. Sadowsky MJ, Tong Z, de Souza M, Wackett LP (1998) AtzC is a new member of the amidohydrolase protein superfamily and is homologous to other atrazine metabolizing enzymes. Journal of bacteriology **180**, 152 158
14. Porter DJ, Austin EA (1993) Cytosine deaminase. The roles of divalent cations in catalysis. Journal of biological chemistry **268**, 24005 24011
15. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the web: a case study using the Phyre server. Nature protocols **4**, 363 371

Chapter 44
Analysis of Gene Translation Using a Communications Theory Approach

Mohammad Al Bataineh, Lun Huang, Maria Alonso, Nick Menhart, and Guillermo E. Atkin

Abstract Rapid advances in both genomic data acquisition and computational technology have encouraged the development and use of advanced engineering methods in the field of bioinformatics and computational genomics. Processes in molecular biology can be modeled through the use of these methods. Such processes include identification and annotation of all the functional elements in the genome, including genes and regulatory sequences, which are a fundamental challenge in genomics and computational biology. Since regulatory elements are often short and variable, their identification and discovery using computational algorithms is difficult. However, significant advances have been made in the computational methods for modeling and detection of DNA regulatory elements. This paper proposes a novel use of techniques and principles from communications engineering, coding, and information theory for modeling, identification, and analysis of genomic regulatory elements and biological sequences. The methods proposed are not only able to identify regulatory elements (REs) at their exact locations, but can also "interestingly" distinguish coding from non-coding regions. Therefore, the proposed methods can be utilized to identify genes in the mRNA sequence.

Keywords Communications engineering · Gene expression · Regulatory elements · Translation

44.1 Introduction

Communications and information theory has been proved to provide powerful tools for the analysis of genomic regulatory elements and biological sequences [1 5]. An up-to-date summary of current research can be found in [6]. The genetic

M. Al Bataineh (✉)
Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL, USA
e mail: albamoh@iit.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 44,

information of an organism is stored in the DNA, which can be seen as a digital signal of the quaternary alphabet of nucleotides $X = \{A, C, G, T\}$. An important field of interest is gene expression, the process during which this information stored in the DNA is transformed into proteins. Gene expression codes for the expression of specific proteins that carry out and regulate such processes. Gene expression takes place in two steps: transcription and translation.

This paper is organized as follows. Section 44.2 describes our previous model for the process of translation in gene expression which is compared to work done in [5]. Section 44.3 presents four new other models for the process of translation with simulation results shown in Sect. 44.4. Finally, conclusions are drawn in Sect. 44.5.

44.2 Previous Model

The process of translation in prokaryotes is triggered by detecting an RE known as the Shine-Dalgarno (SD) sequence. Physically, this detection works by homology mediated binding of the RE to the last 13 bases of the 16S rRNA in the ribosome [7]. In our work [1] and [2], we have modified this detection/recognition system introduced in [5] by designing a one-dimensional variable-length codebook and a metric. The codebook uses a variable codeword length of N between 2 and 13 using the Watson Crick complement of the last 13 bases of the 16S rRNA molecule. Hence, we obtain $(13 \quad N + 1)$ codewords; $C_i = [s_1, s_2, \ldots, s_{i+N\ 1}]$; $i \in [1, 13 - N + 1]$, where $s = [s_1, s_2, \ldots, s_{13}] = [\text{UAAGGAGGUGAUC}]$ stands for the complemented sequence of the last 13 bases. A sliding window of size N applies to the received noisy mRNA sequence to select subsequences of length N and to match them with the codewords in the codebook (see Table 44.1). The codeword that results in a minimum weighted free energy exponential metric between doublets (pair of bases) is selected as the correct codeword, and the metric value is saved. Biologically, the ribosome achieves this by means of the complementary principle. The energies involved in the rRNA mRNA interaction tell the ribosome when a signal is detected and, thus, when the start of the process of translation should take place. In our model, a modified version of the method of free energy doublets presented in [7] is adopted to calculate the energy function (see 44.1). This function represents a free energy distance metric in kcal/mol instead of

Table 44.1 16S rRNA codebook

C_1	Codeword
C_1	UAAGG
C_2	AAGGA
C_3	AGGAG
C_4	GGAGG
C_6	AGGUG
C_7	GGUGA
C_8	GUGAU
C_9	UGAUC

Table 44.2 Energy doublets

Pairs of bases energy			
AA	0.9	GA	2.3
AU	0.9	GU	2.1
UA	1.1	CA	1.8
UU	0.9	CU	1.7
AG	2.3	GG	2.9
AC	1.8	GC	3.4
UG	2.1	CG	3.4
UC	1.7	CC	2.9

minimum distance (see Tables 44.2) [5]. Our algorithm assigns weights to the doublets such that the total energy of the codeword increases with a match and decreases with a mismatch. Therefore, the total energy gets more emphasized or de-emphasized when consecutive matches or mismatches occur. The energy function has the following form:

$$E_k = \sum_{n=1}^{N} w_n E_n \delta_n, \tag{44.1}$$

where $\delta_k$means a match ($\delta_k = 1$) or a mismatch ($\delta_k = 0$) and $w_k$is the weight applied to the doublet in the kth position. The weights are given by

$$w_n = \begin{cases} \rho_n + a^{\sigma_n}, & \text{if match,} \\ \max\left\{w_{n\ 1} - \left(a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n}\right), 0\right\}, & \text{if mismatch,} \end{cases} \tag{44.2}$$

where, σ and $\tilde{\sigma}$ are the numbers of consecutive matches or mismatches and ρ is an offset variable updated as follows:

$$\rho_n = \begin{cases} \rho_{n\ 1}, & \text{if } \delta_n = 1, \\ 0, & \text{if } \delta_n = 0 \text{ and } \rho_{n\ 1} \leq a, \\ \max\left\{w_{n\ 1} - \left(a^{\tilde{\sigma}_n+1} - a^{\tilde{\sigma}_n}\right), 0\right\}, & \text{otherwise,} \end{cases} \tag{44.3}$$

Where, a is a constant that will control the exponential growth of the weighting function. The offset variable ρ updated at each step according to (44.3) is introduced for the purpose of keeping track of the growing trend that happens when a consecutive number of matches occurs followed by a mismatch. When a mismatch occurs, we increment the number of mismatches that is initialized to zero by one, reset the number of matches back to zero, calculate the current weighting factor, and finally reevaluate the offset variable to be used in the next alignment. Without the use of this offset variable, we will have several peaks when we come into a good match of the codeword in that particular alignment.

For larger values of a, the exponential will grow faster as the number of consecutive matches increases (hence increasing the likelihood that the right sequence is enhanced), making the algorithm more sensible to the correlation in

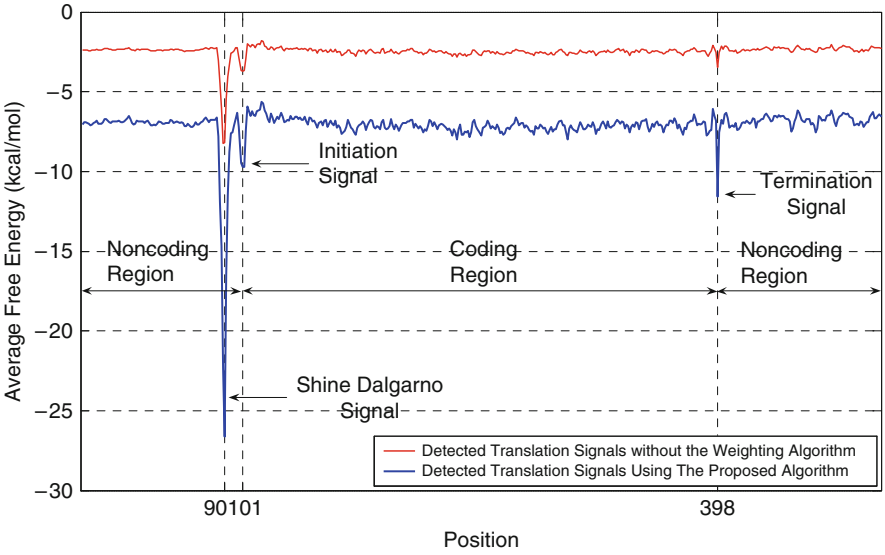


Fig. 44.1 Detection of translation signals

the sequence. Not only does this algorithm allow controlling the resolution of detection (by the choice of the parameter a), but also allows the identification of the exact position of the best match of the Shine Dalgarno signal in the genes under study.

For the analysis, sequences of the complete genome of the prokaryotic bacteria *E. coli* strains MG1655 and O157:H7 were obtained from the National Center for Biotechnology Information. Our proposed exponentially weighting algorithm was not only able to detect the translational signals (Shine Dalgarno, start codon, and stop codon) but also resulted in a much better resolution than the results obtained when using the codebook alone (without weighting). Fig. 44.1 shows the comparison of the average results for the detection of the SD, and start and stop codons to the previous work [5]. It can be seen that the proposed algorithm is able to identify the Shine Dalgarno (peak at position 90) and the start codon (peak at position 101), and the stop codon (peak at position 398). Moreover, these results support the arguments for the importance of the 16S rRNA structure in the translation process. Different mutations were tested using our algorithm and the results obtained further certified the correctness and the biological relevance of the model.

44.3 New Models

The previous model discussed in the introduction is based on coding theory (codebook). We have developed different models for the detection process that the ribosome uses to identify and locate translation signals (Shine Dalgarno,

initiation signal, and termination signal) [3]. These models are based on concepts in communications theory such as Euclidean distance (model I), matched filter (model II), free energy doublets (model III), and correlation-based exponential metric (model IV). The four models are briefly described below.

44.4 Model I. Euclidean Distance-Based Algorithm

In this model, a Euclidean distance measure can be used to detect a given binding sequence in the mRNA sequence. This measure is calculated at each single base in the mRNA sequence as described in [3]. This method is able to detect the binding sequences in their exact location and accounts for mismatches as well.

44.5 Model II. Cross Correlation (Matched Filter)

This model is based on using a matched filter of an impulse response equal to $h(n) = y(-n)$ and an input of $x(n)$, where $y(n)$ is the binding sequence and $x(n)$ is the mRNA sequence [3].

44.6 Model III. Free Energy Metric

In this method, we use the free energy table (see Tables 44.2) to calculate a free energy distance metric in kcal/mol. This metric is calculated at each alignment between the mRNA sequence and the binding sequence under study, as described in [3].

44.7 Model IV. Exponential Detection Metric

This method detects a binding sequence based on aligning it with the mRNA sequence. An exponential metric related to the total number of matches at each alignment is evaluated as follows:

1. Slide the binding sequence under study along the mRNA sequence one base at a time.
2. At the ith alignment, calculate an exponential weighting function ($W(i)$) using the equation

$$W(i) = \sum_{n=1}^{N} w(n), \tag{44.4}$$

where $w(n)$ is the weight applied to the base in the nth position and N is the length of the binding sequence under study. The weights are given by

$$w(n) = \begin{cases} a^{\sigma}, & \text{if match,} \\ 0, & \text{if mismatch,} \end{cases} \tag{44.5}$$

where a is an input parameter that controls the exponential growth of the weighting function W, and σ is the number of matches at each alignment.

3. Repeat step 2 for all alignments along the mRNA sequence to get the weighting vector W:

$$W = [w(1), w(2), \ldots, w(L - N + 1)], \tag{44.6}$$

where L is the length of the mRNA sequence.

4. Plot the weighting vector W, and detect peaks.

44.8 Simulation Results

In this section, we show the results of applying the four models described briefly in Sect. 44.3 and demonstrate their usefulness in pointing out interesting and new biological insights related to the process of translation in gene expression. Without loss of generality and since all of the four models showed similar behavior in detecting translational signals, we chose to show the results of using the Exponential Detection Metric (Model IV) as an example.

In our analysis, sequences of the complete genome of the prokaryotic bacteria *E. coli* strain MG1655 (also we used O157:H7 strain with similar results) were obtained. These sequences are available in the National Center for Biotechnology Information (NCBI) [8]. For presentation purposes and because of the fact that genes are of different lengths, all the tested sequences were selected to follow a certain structure such that they are all 500 bases long. The Shine Dalgarno was set at position 90, the initiation codon at position 101, and the termination codon at position 398. The four new models were used to detect the last 13 bases of the 16S rRNA molecule in the given mRNA sequence by averaging over all the 500-bases-long test sequences. Simulation results show that the proposed models allow the detecton of the translational signals at their exact corresponding locations as expected. Furthermore, they allow the identification of the coding regions (higher ripple region) and the non-coding regions (lower ripple region) as can be observed in Figs. 44.2 44.5. This new result suggests that the last 13-base sequence of 16S rRNA molecule has a higher correlation with coding regions compared with the non-coding regions. This suggests that the proposed models, which were originally designed for regulatory sequence identification, can help identify genes as well. The four models will be applied to other organisms.

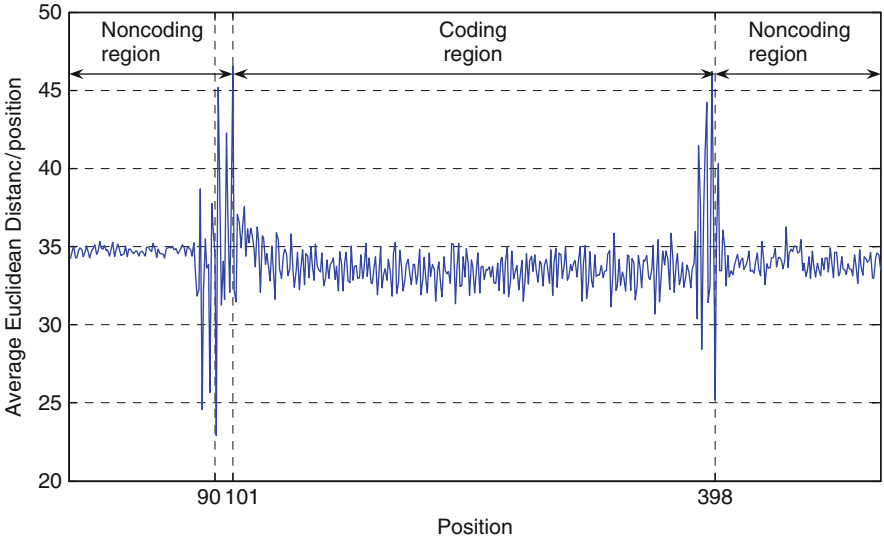


Fig. 44.2 Euclidean distance metric

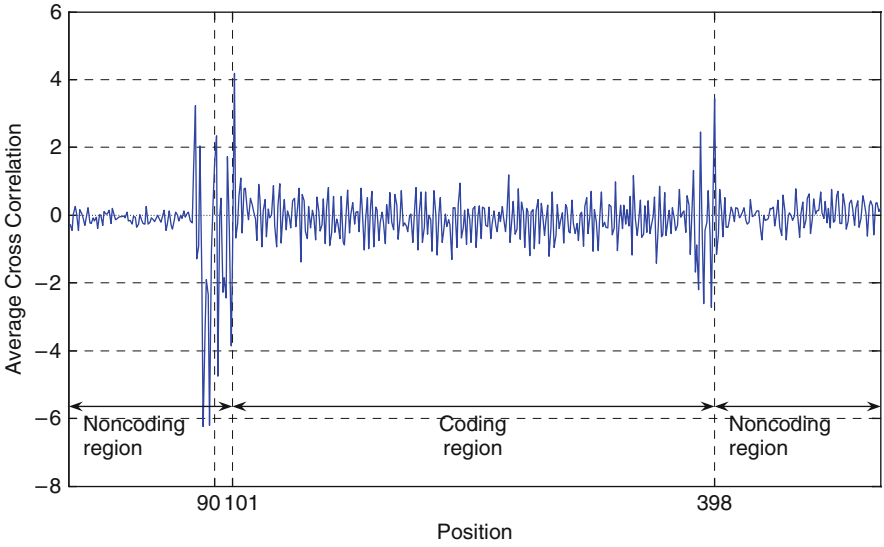


Fig. 44.3 Cross correlation (matched filter)

To study the effect of mutations on the detection of translational signals, different types of mutations were incorporated in the last 13-base sequence and then tested using the developed models. In this work, we have considered Jacob, Hui, and De Boer mutations.

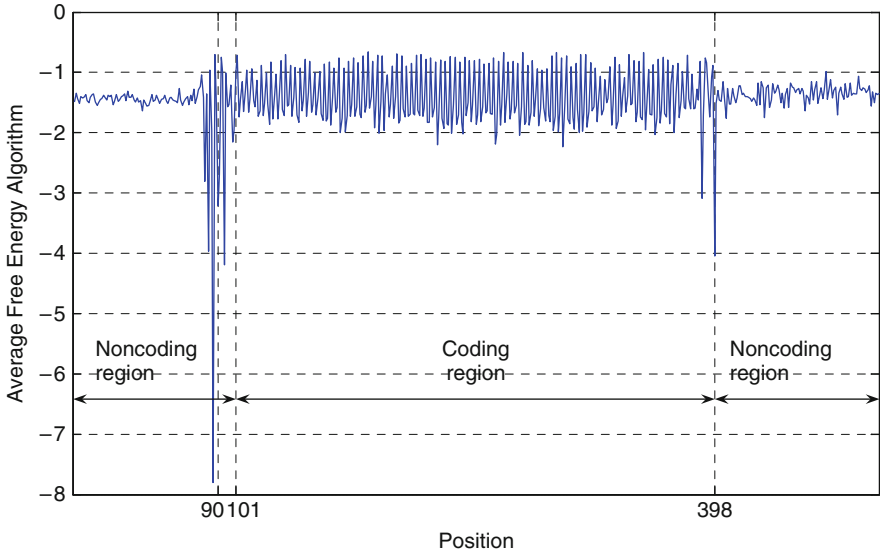


Fig. 44.4 Free energy metric

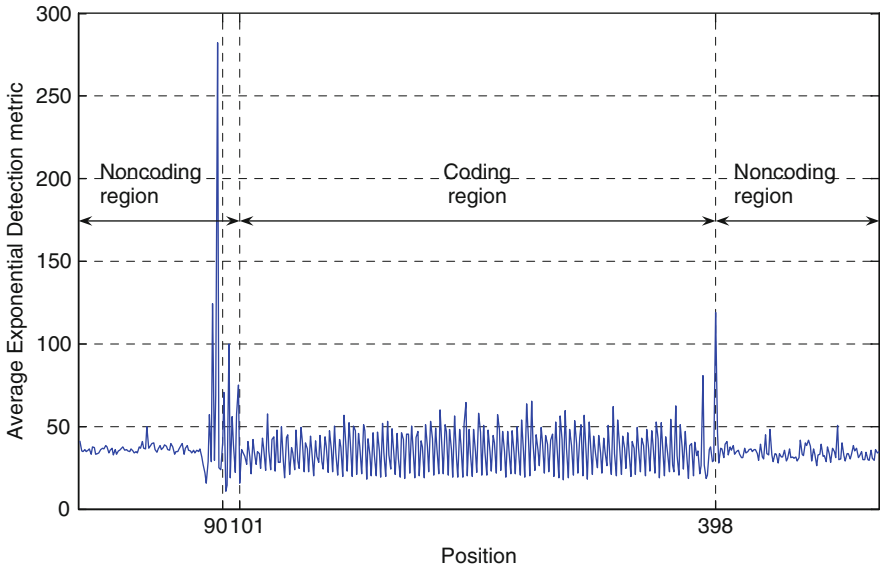


Fig. 44.5 Exponential detection

Jacob mutation, a mutation in the 5th position of the last 13 bases of 16S rRNA molecule [9], results in a reduction in the level of protein synthesis. This mutation was tested using Model IV. Simulation results in Fig. 44.6 show a reduction in the

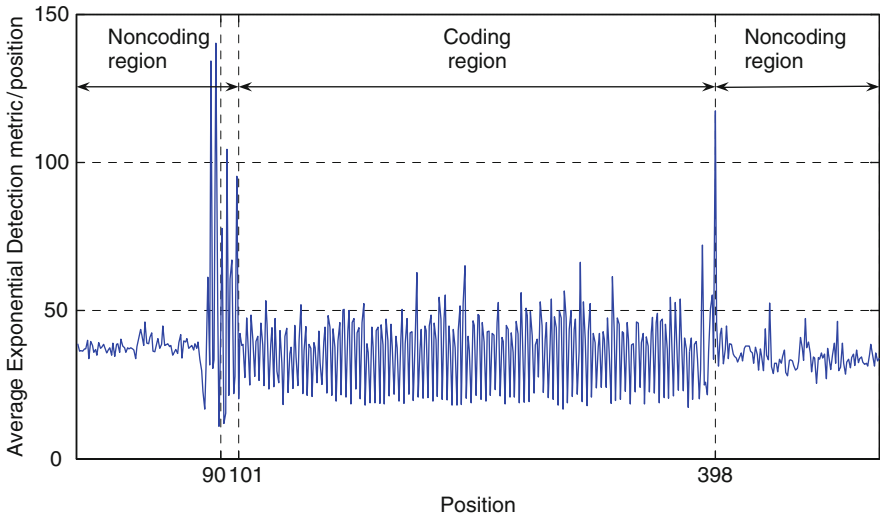


Fig. 44.6 Jacob mutation

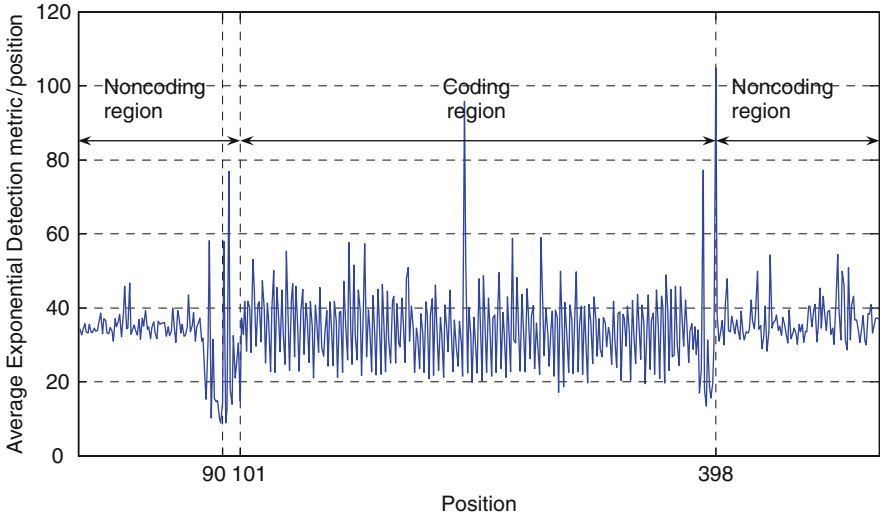


Fig. 44.7 Hui mutation

amplitude of the Shine Dalgarno signal compared to the non-mutation case in Fig. 44.5. This reduction can be interpreted as a reduction in the level of protein synthesis, i.e., the levels of protein production will be reduced but not completely stopped.

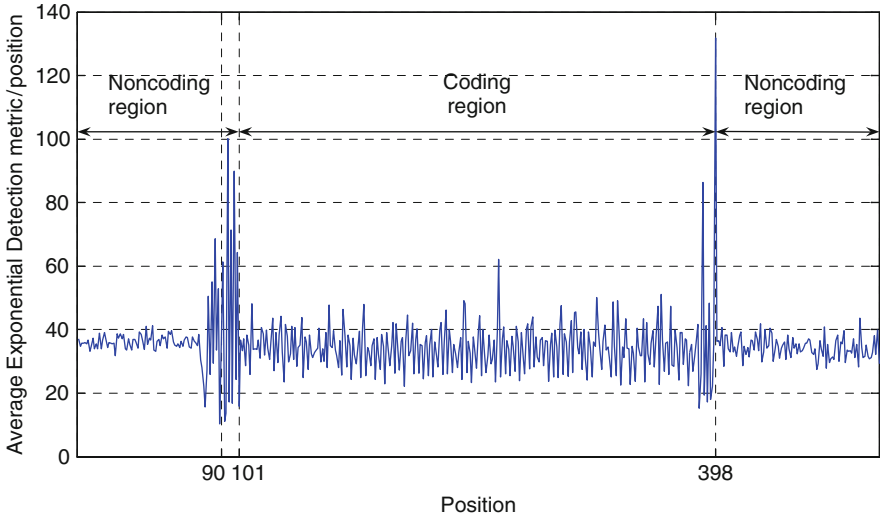


Fig. 44.8 De Boer mutation

Hui and De Boer mutations occur in positions 4 8 (GGAGG → CCUCC) and positions 5 7 (GAG → UGU) of the last 13-base sequence [10]. The results of both mutations are lethal for the organism in the sense that the production of proteins stop. Figure 44.7 shows a complete loss of the Shine Dalgarno (SD) signal (at position 90). Hence, it can be inferred that the translation will never take place Fig. 44.8.

44.9 Conclusion

The increase in genetic data during the last years has prompted the efforts to use advanced techniques for their interpretation. This paper proposes a novel application of ideas and techniques from communications and coding theory to model and analyze gene expression, and gene and regulatory sequence identification. Different models for regulatory elements identification are developed and investigated. Simulation results verify the correctness, accuracy, and biological relevance of these models in detecting regulatory sequences. Moreover, as these models are surprisingly capable of distinguishing coding from non-coding regions, they can help identify genes. Mutations in the $3'$ end of the 16S rRNA molecule were investigated. The obtained results totally agree with biological experimentations. This further supports the correctness and the biological relevance of the proposed models, and hence can serve as a way to introduce new lines of biological research.

References

1. Mohammad Al Bataineh, Maria Alonso, Siyun Wang, Wei Zhang, and G. E. Atkin, "Ribosome Binding Model Using a Codebook and Exponential Metric," *2007 IEEE Interna tional Conference on Electro/Information Technology,* pp. 438 442, 17 20 May 2007.
2. Mohammad Al Bataineh, Maria Alonso, Siyun Wang, G. Atkin, and W. Zhang, "An Opti mized Ribosome Binding Model Using Communication Theory Concepts," *Proceedings of 2007 International Conference for Bioinformatics and Computational Biology,* June 25 27, 2007.
3. Mohammad Al Bataineh, Lun Huang, Ismaeel Muhamed, Nick Menhart, and G. E. Atkin, "Gene Expression Analysis using Communications, Coding and Information Theory Based Models," *BIOCOMP'09 The 2009 International Conference on Bioinformatics & Compu tational Biology,* pp. 181 185, 13 16, 2009.
4. E. E. May, M. A. Vouk, D. L. Bitzer, and D. I. Rosnick, "Coding theory based models for protein translation initiation in prokaryotic organisms," *Biosystems,* vol. 76, pp. 249 260, Aug Oct 2004.
5. Z. Dawy, F. Gonzalez, J. Hagenauer, and J. C. Mueller, "Modeling and analysis of gene expression mechanisms: a communication theory approach," *IEEE International Conference on Communications (ICC),* vol. 2, pp. 815 819, 2005.
6. "DNA as Digital Data Communication Theory and Molecular Biology," *IEEE Engineering in Medicine and Biology,* vol. 25, January/February 2006.
7. D. Rosnick, "Free Energy Periodicity and Memory Model for Genetic Coding." vol. PhD thesis, Raleigh: North Carolina State University, 2001.
8. "NCBI: National Center for Biotechnology Information," from http://www.ncbi.nlm.nih.gov/.
9. W. F. Jacob, M. Santer, and A. E. Dahlberg, "A single base change in the Shine Dalgarno region of 16S rRNA of *Escherichia coli* affects translation of many proteins," *Proc Natl Acad Sci USA,* vol. 84, pp. 4757 4761, 1987.
10. A. Hui and H. A. de Boer, "Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosomes in *Escherichia coli*," *Proc Natl Acad Sci USA,* vol. 84, pp. 4762 4766, 1987.

Chapter 45
A Fast and Efficient Algorithm for Mapping Short Sequences to a Reference Genome

Pavlos Antoniou, Costas S. Iliopoulos, Laurent Mouchard, and Solon P. Pissis

Abstract Novel high-throughput (Deep) sequencing technology methods have redefined the way genome sequencing is performed. They are able to produce tens of millions of short sequences (reads) in a single experiment and with a much lower cost than previous sequencing methods. In this paper, we present a new algorithm for addressing the problem of efficiently mapping millions of short reads to a reference genome. In particular, we define and solve the *Massive Approximate Pattern Matching* problem for mapping short sequences to a reference genome.

Keywords DNA sequences · Genome sequencing · Mapping sequences · Pattern matching · SBS

45.1 Introduction

High-throughput, (or Deep) sequencing technologies have opened new and exciting opportunities in the use of DNA sequences. These new emerging technologies, which are based on in vitro cloning and *sequencing-by-synthesis* (SBS) [2], are able to produce tens of millions of short reads of currently, typically, 25 50 bp in a single experiment. The recent advances in high-throughput sequencing technologies, and the way they will crucially impact tomorrow's biology, are presented in [2].

Examples of these new technologies are the *Genome Analyzer*, developed by *Illumina-Solexa*, which generates millions of very short reads ranging from 25 to 50 bp in a single experiment [4]; the *ABI-SOLiD* system, which performs massively parallel sequencing by hybridization and ligation [4]; and the *Roche-454* system,

P. Antoniou (✉)
Department of Computer Science, University of Cyprus, Nicosia, Cyprus
e mail: panton@cs.ucy.ac.cy

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 45,

which generates fewer but longer sequences [3]. The common denominator of these technologies is the fact that they are able to produce a massive amount of relatively short reads.

A major issue for biologists is how to map these short reads efficiently and accurately to a reference genome. Popular alignment programs like *BLAST* or *BLAT* are not successful because they focus on the alignment of fewer and longer sequences [3]. Recently, a new thread of applications addressing the short sequences mapping problem has been devised. These applications (*ELAND* developed by Illumina-Solexa, *SeqMap* [3], or *SOAP* [4], to name a few), which are based on the pigeonhole principle [3], make use of indexing and hashing techniques.

In this paper, we address the problem of efficiently mapping millions of short reads to a reference genome. In particular, we define and solve the *Massive Approximate Pattern Matching* problem, by extending the results presented in [1]. Our approach preprocesses the genomic sequence, based on the short reads length, by using word-level parallelism and sorting. We do not hash the short reads, but instead we convert each read into a unique arithmetic value, using 2-bits-per-base encoding of the DNA alphabet, and then use the pigeonhole principle, binary search, and simple word-level operations for mapping the short reads to the genomic sequence.

45.2 Preliminaries

A *string* is a sequence of zero or more symbols from an alphabet Σ. The set of all strings over the alphabet Σ is denoted by Σ^*. In this work, we consider the finite alphabet Σ for DNA sequences, where $\Sigma = \{A, C, G, T\}$. The length of a string x is denoted by $|x|$. The ith symbol of a string x is denoted by $x[i]$. A string w is a *substring* of x if $x = uwv$, where $u, v \in \Sigma^*$. We denote by $x[i..j]$ the substring of x that starts at position i and ends at position j.

For two strings x and y, such that $|x| = |y|$, the Hamming distance $\delta_H(x, y)$ is the number of places in which the two strings differ, i.e. have different characters. Formally, $\delta_H(x, y) = \sum_{i=1}^{|x|} 1_{x[i] \neq y[i]}$, where $1_{x[i] \neq y[i]} = 1$, if $x[i] \neq y[i]$, or 0, otherwise.

45.3 Problem Definition

We denote the generated short reads as the set $p_0, p_1, ..., p_{r\ 1}$, where $r > 10^7$, and we call them *patterns*. The length of each pattern is currently, typically between 25 and 50 bp long. Without loss of generality, we denote that length as ℓ. We are given a solid genomic sequence $t = t[1..n]$, where $n > 10^8$, and a positive threshold $k \geq 0$. The case that $k > 0$ corresponds to the possibility that the pattern either contains a

sequencing error, or a small difference between a mutant and the reference genome, as explored in [6].

We define the *Massive Approximate Pattern Matching* problem for mapping short sequences to a reference genome as follows.

Problem 1. Find whether the pattern $p_i = p_i[1..\ell]$, for all $0 \leq i < r$, with $p_i \in \Sigma^*$, $\Sigma = \{A, C, G, T\}$, occurs with at most k mismatches in $t = t[1..n]$, with $t \in \Sigma^*$.

45.4 Massive Approximate Pattern Matching

In this section, we extend the results for the case that $k = 0$, presented in [1], to address Problem 1 for the case that $k \geq 0$. The idea of using the pigeonhole principle to split each read into v fragments is adopted.

Lemma 1. *Given the number of fragments v of a string* $x = \{x^1, x^2, ..., x^v\}$, *and the number of allowed mismatches k,* $k \leq v$, *any of the k mismatches cannot exist, at the same time, in at least* $v - k$ *fragments of x.*

Proof. Immediate from the pigeonhole principle.

Our aim is to preprocess t and construct v sorted lists L_j, for all $1 \leq j \leq v$. Each list L_j holds elements of type $e(z_i^j) = (i, \sigma(z_i^j), \text{prev}_i^j, \text{next}_i^j)$, where i represents the starting position of substring $z_i = t[i..i + \ell - 1]$; prev_i^j points to the element of the previous fragment in L_{j-1}, for all $2 \leq j \leq v$; and next_i^j points to the element of the next fragment in L_{j+1}, for all $1 \leq j \leq v - 1$. Let $e(z_i) = \{e(z_i^1), ..., e(z_i^v)\}$ and $c_g(\sigma(x)) = \{\sigma(x^{q_1}), ..., \sigma(x^{q_{v-k}})\}$; for all $1 \leq g \leq \binom{v}{v-k}$, the $\binom{v}{v-k}$ combinations of $\sigma(x) = \{\sigma(x^1), ..., \sigma(x^v)\}$. We define the following operations:

- $\sigma(x)$: it returns a unique arithmetic value of x, iff $x \in \Sigma^*$, in $O(\ell)$.
- *Find-all* $(c_g(\sigma(x)), L)$: it returns $A_g = \{e(z_{u_1}), ..., e(z_{u_v})\}$, such that for each $\sigma(x^{q_j}) \in c_g(\sigma(x))$, for all $1 \leq j \leq v - k$, $\sigma(x^{q_j}) \in e\left(z_{u_i}^{q_j}\right)$, for all $1 \leq i \leq v$. It uses binary search to first locate $\sigma(x^{q_1})$ in list L_{q_1} in $O(\log n)$ time, and then return $\sigma(x^{q_j}) \in c_g(\sigma(x))$, for all $2 \leq j \leq v - k$, with the use of pointers, in $O(vv)$.
- *Find-rest* $(A_g, c_g(\sigma(x)), \sigma(x))$: it returns $T_g = \{e(z_{u_1}), ..., e(z_{u_v})\}$, such that for each $\sigma(x^{q_j}) \in \{\sigma(x) - c_g(\sigma(x))\}$, for all $1 \leq j \leq k$, $\sigma(x^{q_j}) \in e\left(z_{u_i}^{q_j}\right)$, for all $1 \leq i \leq v$. Given A_g, it returns T_g in $O(vk)$.
- *Bitop* $(\sigma(x), \sigma(y))$: it returns $\delta_H(x, y)$ in constant time, given that $|x| = |y| = \alpha$ and $2\alpha \leq w$, where w is the size of the computer word.

An outline of Algorithm I is as follows:

Step 1. We partition the text into a set of substrings $z_1, z_2, ..., z_{n-\ell+1}$, where $z_i = t[i...i + \ell - 1]$, for all $1 \leq i \leq n - \ell + 1$. We compute $\sigma(z_i)$, split it into v fragments $\sigma(z_i) = \left\{\sigma\left(z_i^1\right), \sigma\left(z_i^2\right), ..., \sigma\left(z_i^v\right)\right\}$, and add $(i, \sigma\left(z_i^j\right), i, i)$ to a list L_j, for all $1 \leq j \leq v$. As soon as we compute $\sigma(z_1)$, each $\sigma(z_i)$, for all $2 \leq i \leq n - \ell + 1$, can be retrieved in constant time (using "shift"-type of operation).

Step 2. We sort the lists L_j, for all $1 \leq j \leq v$, based on the signature field $\sigma(z_i^j)$, ensuring that in a case that we swap elements, we preserve that prev_i^{j+1} and next_i^{j-1} point to the swapped element.

Main Step. Assume that we have a query $p_i[1...\ell]$ for all $0 \leq i \leq r$. We compute $\sigma(p_i)$, split it into $\sigma(p_i) = \{\sigma(p_i^1), \sigma(p_i^2), ..., \sigma(p_i^v)\}$, and compute $c_g(\sigma(p_i))$, $A_g = \text{find - all}\,(c_g(\sigma(p_i)), L)$, and $T_g = \text{find - rest}\,(A_g, c_g(\sigma(p_i)), \sigma(p_i))$, for all $1 \leq g \leq \binom{v}{v-k}$. If there exists $e(z_{u_j}) \in T_g$, for all $1 \leq j \leq v$, $1 \leq g \leq \binom{k}{v}$, such that $\sum_{\lambda=q_1}^{q_k} \text{bitop}(\sigma(p_i^\lambda), \sigma(z_{u_j}^\lambda)) \leq k$, for all $\sigma(z_{u_j}^\lambda) \in e(z_{u_j}^\lambda)$, then p_i occurs in t.

Theorem 1. *Given the text $t = t[1..n]$, the set of patterns $p_0, p_1, ..., p_{r-1}$, the length of each pattern ℓ, the number of fragments v, and the number of mismatches k, Algorithm I solves Problem 1 in $O(vn\log n + r\binom{v}{v-k}\log n)$ units of time.*

Proof. Step 1 can be done in $O(vn)$ time. In step 2, the time required for sorting the list L_j, for all $1 \leq j \leq v$, is $O(vn\log n)$. The main step runs in $O(r\binom{v}{v-k}\log n)$ time. Hence, asymptotically, the overall time is $O(vn\log n + r\binom{v}{v-k}\log n)$, which is $O(n\log n + r\binom{v}{v-k}\log n)$, in practice.

45.5 Experimental Results

In order to evaluate the correctness and efficiency of Algorithm I, we have mapped 31,116,663 Illumina 25-bp reads, taken from *RNA-Seq* experiments [5], to genomic sequences of various lengths of the mouse chromosome *X*, as well as to the whole chromosome sequence (166,650,296 bp), allowing up to 2 mismatches. The experimental results of Algorithm I are illustrated in Table 45.1.

The presented experimental results are very promising concerning both the efficiency of the proposed algorithm and the sensitivity of our approach in terms of mapping. Our program maps 1,043,120 of 31,116,663 Illumina-Solexa 25-bp reads to the mouse chromosome *X* in 23m49s, allowing up to 2 mismatches. A direct comparison, in terms of efficiency, with other mapping programs would not be reliable. The existing well-known mapping programs are very data and parameter dependent, based on whether they index the reference genome or the reads, e.g., *SeqMap* requires more than 16 GB of memory for this experiment. As illustrated in [3], *SeqMap* can map 455,384 out of 11,530,816 Illumina-Solexa 25-bp reads to the mouse chromosome *X* in 36m53s, allowing up to 2 mismatches.

Table 45.1 Mapping 31,116,663 Illumina Solexa 25 bp reads to the mouse chromosome *X*, allowing up to 2 mismatches

Genomic sequence	Running time	Mapped reads
40,000,000 bp	7m38s	455,411
80,000,000 bp	12m40s	684,934
166,650,296 bp	23m49s	1,043,120

The proposed algorithm was implemented[1] in *C++* language. The program accepts FASTA format for the reference and the short reads. The experiments were conducted on a machine with 3 GHz Intel Xeon CPU and 16 GB memory, running 64-bit Linux operating system.

45.6 Conclusion

The new high-throughput sequencing technologies produce a huge number of very short sequences, and these sequences need to be tagged and recognized as parts of a reference genome. We have presented an algorithm to address the Massive Approximate Pattern Matching problem for mapping these short sequences to a reference genome. The presented experimental results are very promising, in terms of efficiency and sensitivity on mapping, compared to more traditional approaches.

References

1. Antoniou, P., Daykin, J.W., Iliopoulos, C.S., Kourie, D., Mouchard, L. and Pissis S.P. (2009) Mapping uniquely occurring short sequences derived from high throughput technologies to a reference genome, Proceedings of the 9th International Conference on Information Technology and Applications in Biomedicine (ITAB 09), DOI: 10.1109/ITAB.2009.5394394
2. Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology, *J. Exp. Biol*., Vol. 210, No. 9, 1518 1525
3. Jiang, H. and Wong, W.H. (2008) *SeqMap*: mapping massive amount of oligonucleotides to the genome, *Bioinformatic*, Vol. 24, No. 20, 2395 2396
4. Li, R., Li, Y., Kristiansen, K. and Wang, J. (2008) SOAP: short oligonucleotide alignment program, *Bioinformatics*, Vol. 24, No. 5, 713 714
5. Mortazavi, A., Williams, B. A. A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by *RNA Seq*, *Nat Methods*, Vol. 5, No. 7, 621 628
6. Wang, Z., Gerstein, M. and Snyder, M. (2008) *RNA Seq*: a revolutionary tool for transcrip tomics, *Nat Rev Genet*, Vol. 10, No. 1, 57 63

[1]The implementation is available at a website (http://www.dcs.kcl.ac.uk/pg/pississo/), which has been set up for maintaining the source code and the documentation.

Chapter 46
Sequence Analysis and Homology Modeling *Gallus gallus* Glutathione S-transferase (Q08392)

Patchikolla Satheesh, Allam Appa Rao, G.R. Sridhar, Kudipudi Srinivas, and Chandra Sekhar Akula

Abstract Glutathione S-transferases (GST) belong to the transferase group of enzymes; GST are a family of enzymes that catalyze the addition of glutathione to endogenous or xenobiotic, often toxic electrophilic chemicals, and a major group of detoxification enzymes. We used the homology modeling technique to construct the structure of *Gallus gallus* GST. The amino acid sequence identity between the target protein and sequence of template protein 1ML6 (*Mus musculus*) was 66.2%. Based on the template structure, the protein model was constructed by using the Homology program Modeller9v1, and briefly refined by energy minimization steps; it was validated by PROCHECK. In all, 94.4% of the amino acids were in allowed regions of Ramachandran plot, showing the accuracy of the model and good stereochemical quality. Our results correlated well with the experimental data reported earlier, which proved the quality of the model. This generated model can be further used for the design and development of more potent GST inhibitors.

Keywords Glutathione S transferase · Integrated Scientific Information System · Root Mean Square Deviation · Protein Data Bank · Protein Information Resource

46.1 Introduction

Glutathione S-transferases (EC no: 2.5.1.18) constitute a large family of enzymes that catalyze the addition of glutathione to endogenous or xenobiotic, often toxic electrophilic chemicals, and a major group of detoxification enzymes [1]. All eukaryotic species possess multiple cytosolic and membrane-bound GST isoenzymes, each of which displays distinct catalytic as well as noncatalytic binding properties: the cytosolic enzymes are encoded by at least five distantly related

G.R. Sridhar (✉)
Endocrine and Diabetes Centre, 15 12 16 Krishnanagar, Visakhapatnam 530 002, AP, India
e mail: sridharvizag@gmail.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 46,

gene families (designated class alpha, mu, pi, sigma, and theta GST), whereas the membrane-bound enzymes, microsomal GST and leukotriene C4 synthetase, are encoded by single genes and both have arisen separately from the soluble. Evidence suggests that the level of expression of GST is a crucial factor in determining the sensitivity of cells to a broad spectrum of toxic chemicals.

Individual isoenzymes of GST contribute to resistance to carcinogens, antitumor drugs, environmental pollutants, and products of oxidative stress. Polymorphisms in GST genes (GST-M1, GST-T1, and GST-P1) and susceptibility to prostate cancer among male smokers [2]. Polymorphisms of Genotypes *GSTM1*, *GSTT1*, and *GSTP1* in Glutathione S-transferase susceptibility to Risk and Survival of Pancreatic Cancer [3]. GST may be an ideal target for structure-based drug design. In the current report, we describe the structure of GST of *Gallus gallus*. Homology modeling tool Modeller9v1 was utilized for the modeling of GST.

46.2 Results and Discussion

46.2.1 Choice of Template

The GST sequence closest to that of *G. gallus* comes from *M. musculus* (66.2% sequence identity). We selected template from the structure of *M. musculus* (PDB accession code 1ML6) and performed an alignment using FASTA (Fig. 46.1). Out of 221 residues, 66.2% of amino acids are conserved across these sequences.

Fig. 46.1 Ribbon drawing model of the modeled protein *Gallus gallus* Glutathione S transferase

46.2.2 *Quality of Homology Model*

The Ramachandran plot for GST shows that most nonglycin residues are within the allowed region, with 94.4% of these residues in the energetically most favored area. No residues are in the disallowed region of the plot (Fig. 46.2). The modeled Glutathione S-transferase from *G. gallus* has 0.14 A root mean square (RMS) deviation to all Cα atoms in Glutathione S-transferase from *M. musculus*, respectively.

From the Table 46.1 BL-80 was selected since it has high percent identity (66.2%), % similarities (83.1%) and low E-value (2.3e-95) when compared with other matrices i.e PAM 120, PAM 250, BL 50, and BL 62.

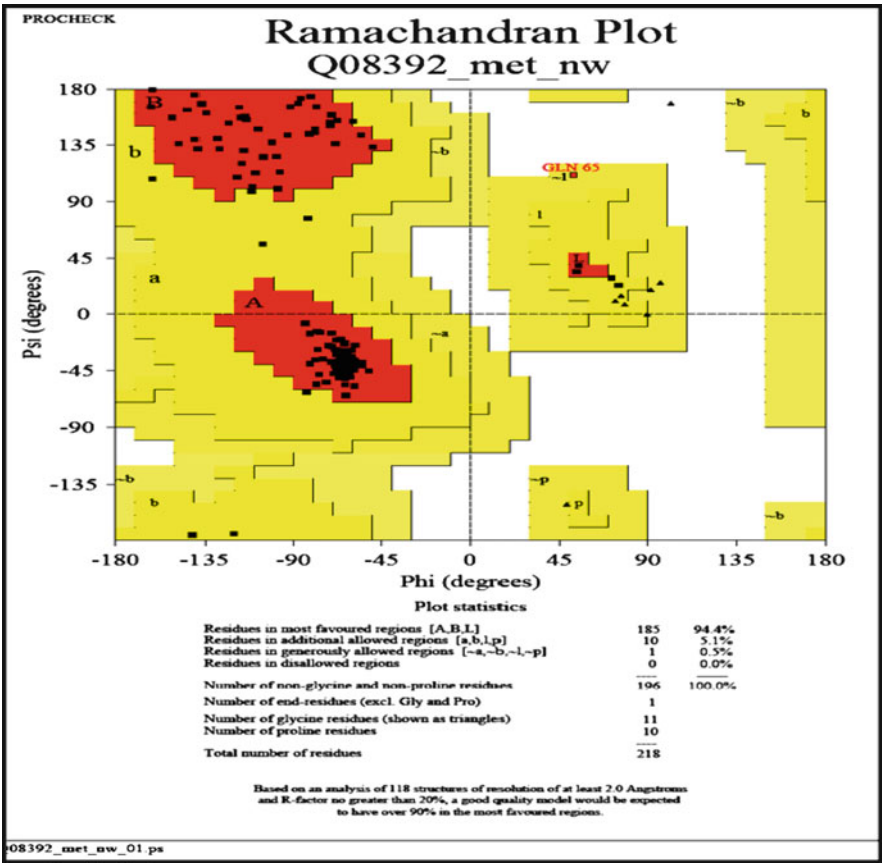


Fig. 46.2 Ramachandran plot of modeled structure

Table 46.1 FASTA analysis using five different matrices, score, E value, % identity, % similarities, gaps, and overlaps

Matrix	% Identity	% Similarity	Number of gaps	Score (bits)	E value	PDB ID	Residue overlap
BL 50	66.2	86.8	0	230	2.1e 60	1ML6	2 220:1 219
BL 62	66.2	86.3	0	276.3	4.1e 74	1ML6	2 220:1 219
BL 80	*66.2*	*83.1*	*0*	*346.9*	*2.3e 95*	*1ML6*	*2 220:1 219*
PAM 120	66.2	86.8	0	327.4	1.6e 89	1ML6	2 220:1 219
PAM 250	66.2	90.4	0	153.7	3.2e 37	1ML6	2 220:1 219

Table 46.2 The final docked energies of 1GSF and Q08392

Ligands	Energy of 1GSF	Energy of Q08392
Compound 6	147.489	106.986
Compound 7	*177.82*	*179.398*
Compound 8	159.607	28.03
Ethacrynic acid (EAA)	*113.96*	*114.986*
GABA + EA	*103.184*	*92.1078*

46.2.3 Overall Structure

The comparison of our modeled GST with *M. musculus* shows that the secondary structure and domain organization are fairly well conserved.

The above reported Bivalent inhibitors of GST that are Compounds 6, 7, 8, Ethacrynic acid (EAA), and GABA + EA [4] were drawn by using Integrated Scientific Information System (ISIS) draw, and the drawn ligands were converted to 3D by using TSAR. Our modeled protein (Q08392) and PDB protein (1GSF) were docked with bivalent inhibitors of Glutathione S-transferases (25) by using Molegro virtual docker 2007. The final docked energies are reported in the Table 46.2.

These results show that modeled protein (Q08392) and PDB (1GSF) have slightly similar final docked energies Q08392-EAA = −114.986 kcal/mol and 1GSF-EAA = −113.96 kcal/mol, and QO8392-compound-7 = −179.398 kcal/mol and 1GSF-compound-7 = −177.82 kcal/mol. This study supported the reliability of modeled protein because both the proteins exhibit similar final docked energies, and our model was reliable as it has features that are comparable with those of the existing model.

46.3 Methods

46.3.1 Choice of Template

The target enzyme sequences identified by FASTA were homologous with those in the PDB in NCBI. The chosen templates were the sequence from the latest version of the PDB with the lowest expected value and the highest score after four iterations.

46.3.2 Sequence Alignment

Alignment of sequences with their template structure was done using the alignment. malign() command in Modeller9v1 [5]. The MODELLER script was used for aligning all target sequences in the .ali format with their corresponding template structures in the PDB files. Finally, the alignment was written in two formats, Protein Information Resource (PIR). The PIR format is used by Modeller in the subsequent model-building stage.

46.3.3 Homology Modeling

A 3D model of the target sequence was constructed with the automodel class of Modeller9v1 to generate five similar iterative models of the target sequence based on its template structure and the alignment input file "filename.ali" (PIR format). The "best" model was selected by picking the model with the lowest Modeller objective function value. Quality of the homology model and the quality of the structures were analyzed with the PROCHECK program to calculate the main-chain torsional angle, i.e., a Ramachandran plot.

46.3.4 Molegro Virtual Docke

MolDock is based on a new heuristic search algorithm that combines differential evolution with a cavity prediction algorithm. The docking scoring function of MolDock is an extension of the piecewise linear potential (PLP) including new hydrogen bonding and electrostatic terms. The docking accuracy of MolDock has been evaluated by docking flexible ligands to 77 proteins. In this study, Molegro Virtual Docker was utilized for docking Bivalent inhibitors of Glutathione S-transferases. The quality of the modeled protein was analyzed with ligand-protein binding energy.

46.4 Conclusion

In order to use any sequence alignment tool with different scoring matrices, one must have to quantify scoring matrices that may likely conserve the physical and chemical properties necessary to maintain the structure and function of the protein. Pair wise sequence analysis of three different alignment programs like FASTA is employed to study the influence of matrix on clear evolutionary relationship. The analysis performed with *chick* protein sequence database has identified relevant homology with PDB protein 1ML6 (66.2% Identity, 83.1% Positives). Our results suggest that BL 80 matrix plays a major role in predicting the structure and

functional relationships to the target protein. Homology modeling initiated with Modeller9v1 program run in Windows operation system resulted in a number of models. The data were consistent with the model as it reported low RMSD (0.1374 A), so it is the best model. The Ramachandran plots identified that the probable number of residues in the most favored region increased from *93.5%* in the template structure 1ML6 to *94.4%* in the modeled protein Q08392. The number of residues in the disallowed regions is zero in both modeled and template proteins. The docking results demonstrated that final docked energies of bivalent inhibitors (GST) with Modeled protein (Q08392) and 1GSF (PDB) are slightly similar in the case of compound 7, Ethacrynic acid. These results supported the reliability of the modeled protein.

This study suggests that a fast and reliable homology model is possible by considering the sequences with profound similarity at the sequence level as the method employed is customizable and result oriented.

References

1. Sheehan D, Meade G, Foley VM, Dowd CA (2001) Structure, function and evolution of glutathione transferases: implications for classification of non mammalian members of an ancient enzyme superfamily, Biochem J 360(Pt 1):1 16
2. Kidd LC, Woodson K, Taylor PR, Albanes D, Virtamo J, Tangrea JA (2003) Polymorphisms in glutathione S transferase genes (GST M1, GST T1 and GST P1) and susceptibility to prostate cancer among male smokers of the ATBC cancer prevention study. Eur J Cancer Prev 12:317 320
3. Jiao L, Bondy ML, Hassan MM, Chang DZ, Abbruzzese JL, Evans DB, Smolensky MH, Li D (2007) Glutathione S transferase gene polymorphisms and risk and survival of pancreatic cancer, Cancer, 8 9.
4. Maeda DY, Mahajan SS, Atkins WM, Zebal JA (2006) Bivalent inhibitors of glutathione S transferase: the effect of spacer length on isozyme selectivity, Bioorg Med Chem Lett, 3780 3783.
5. Sali A, Blundell TL (1995) Comparative protein modelling by satisfaction of spatial restraints, J Mol Biol 234, 779 815.

Chapter 47
Toward Optimizing the Cache Performance of Suffix Trees for Sequence Analysis Algorithms Suffix Tree Cache Performance Optimization

Chih Lee and Chun-Hsi Huang

Abstract Efforts have been devoted to accelerating the construction of suffix trees. However, little attention has been given to post-construction operations on suffix trees. Therefore, we investigate the effects of improved spatial locality on certain post-construction operations on suffix trees. We used a maximal exact repeat finding algorithm, MERF, on which software REPuter is based, as an example, and conducted experiments on the 16 chromosomes of the yeast *Saccharomyces cerevisiae*. Two versions of suffix trees were customized for the algorithm and two variants of MERF were implemented accordingly. We showed that in all cases, the *optimal cache-oblivious* MERF is faster and displays consistently lower cache miss rates than their non-optimized counterparts.

Keywords Cache-oblivious algorithms · Maximal exact repeats · Suffix trees

47.1 Introduction

A suffix tree encodes all the suffixes of a given string using a tree-like structure. Before the advent of suffix trees, searching for k exact patterns of length m in a string of length n using the Knuth Morris Pratt algorithm took $O(k(n+m))$ time. However, this can be done in $O(n+km)$ time with a suffix tree. With the availability of complete genomes and proteomes of many organisms, it is not unusual to encounter DNA or peptide sequences of millions of nucleotides or amino acids. Therefore, suffix trees are particularly useful in dealing with biological sequences.

C. Lee (✉)
Department of Computer Science and Engineering, University of Connecticut, 06269 Storrs, CT, USA
e mail: chih.lee@uconn.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 47,

The linear construction algorithms in [4] and references therein suffer from the memory hierarchy of modern computers. Efforts have been made to deal with tree construction. Bedathur and Haritsa [2] proposed a buffer management strategy, TOP-Q, and showed that it is consistently better than others. Tata et al. [12] presented a $O(n^2)$ cache-efficient suffix tree construction algorithm, producing suffix trees with no suffix link.

Suffix trees find many applications in computational biology [4]. SMOTIF [13] searches for structured patterns and profile motifs, while STAN [8] searches for patterns described by string variable grammars. Kaderali and Schliep [6] introduced a combination of suffix trees and alignment algorithms for probe design. Stoye and Gusfield [11] developed algorithms for finding tandem repeats. REPuter [7] detects various types of maximal repeat pairs, while Bakalis et al. [1] extended this idea to locating maximal repeat tuples.

In this study, we investigate the cache performance of suffix trees in terms of cache miss rates. To understand how the cache performance of the data structure affects its applications, we use a maximal exact repeat finding (MERF) algorithm [4] on which REPuter is based, as an example, and evaluate its performance when suffix trees of different cache performance are used.

47.2 Methodology

47.2.1 Suffix Trees

In this section, we introduce two suffix tree data structures customized for MERF. Consider the suffix tree of a string S of length n, whose last character $S[n-1]$ does not appear elsewhere in the string. Our representation of suffix tree nodes closely follows the one described in [10]. A leaf node holds an index i into S, denoting the suffix $S[i : n-1]$, and the preceding character $S[i-1]$. An internal node N with path label $a\alpha$ from the root node to itself holds the following fields, where a is a letter and α is a string. (1) The depth or the length of $a\alpha$. (2) A leaf node in the subtree rooted at N. (3) A data structure holding the child nodes. (4) A suffix link to the internal node with path label α. The first two fields are common to the two suffix tree data structures. For the third field, one data structure points to a linked list of child nodes. The other uses an array of $|\Sigma|$ pointers to child nodes, where $|\Sigma|$ is the size of alphabet Σ. We refer to the two versions as the *linked-list* and *pointer array* suffix tree, respectively.

To understand the maximal performance gain due to improved spatial locality, we build an optimal suffix tree customized for MERF out of the *pointer array* suffix tree. The *optimal* one has no suffix links and its internal nodes hold N_c pointers to the child nodes, where N_c is the exact number of children, making the internal nodes smaller. All the nodes are stored in the depth-first-search order in memory. To build the *optimal* suffix tree, we construct a *pointer array* suffix tree and traverse it in a

depth-first way. Development of sophisticated algorithms for constructing near optimal suffix trees is discussed in Sect. 47.4.

47.2.2 MERF and Types of Repeat Matches

Two substrings $S[i : i + l - 1]$ and $S[j : j + l - 1]$ of S form a repeat pair if they are identical, while a repeat is maximal if $S[i - 1] \neq S[j - 1]$ and $S[i + l] \neq S[j + l]$. REPuter considers four types of matches including forward, reverse, complement, and palindromic matches. We denote the reverse, complement, and palindrome of S as S^{r}, S^{c}, and S^{p}, respectively. To find all the forward matches, a suffix tree of $S' = d_1 S d_2$ needs to be constructed, where d_i's $\notin S$ are distinct letters. To find all the reverse and/or forward matches, a suffix tree of $S' = d_1 S d_2 S^{\mathrm{r}} d_3$ is required, where d_i's $\notin S, S^{\mathrm{r}}$ are distinct letters. Similarly, if more types of matches are desired, the corresponding sequence, say S^{c}, followed by another delimiter is appended to S'.

We briefly introduce the MERF algorithm in [4] and our implementations. MERF traverses the tree from bottom up, which can be achieved by a depth-first search on the tree. The operations performed at each internal node are sketched in Fig. 47.1.

```
1:  Input: N_p
2:  Given the minimal repeat length l_min
3:  if the depth of N_p < l_min then
4:      for each internal child node N_c of N_p do
5:          FINDREPEATS(N_c)
6:      end for
7:  else
8:      L_p ← |Σ| empty linked lists
9:      for each child node N_c of N_p do
10:         if N_c is an internal node then
11:             L_c ← FINDREPEATS (N_c)
12:             for each c_1 in Σ do
13:                 for each c_2 in Σ \ {c_1} do
14:                     Match leaf nodes in Lp [c_1]
                        with those in L_c [c_2]
15:                 end for
16:             end for
17:             Append L_c to L_p
18:         else
19:             c ← the preceding letter of N_c
20:             for each c_1 in Σ \ {c} do
21:                 Match N_c with leaf nodes in
                    L_p [c_1]
22:             end for Append N_c to L_p [c]
23:         end if
24:     end for
25: end if
26: return L_p
```

Fig. 47.1 FINDREPEATS (N_{p}): Operations performed to find repeat pairs at each internal node N_{p}

Our two implementations differ from that described in Fig. 47.1 in the way that leaf nodes are stored in arrays at each internal node. For both implementations, we use one dynamic array instead of $|\Sigma|$ linked lists to store leaf nodes. For the first implementation, denoted as *MERF1* , an array A_p is maintained at each internal node, holding pointers to leaf nodes. Another array A_c is used to collect leaf nodes under the sub-tree rooted at each child node. Similarly, leaf nodes in A_c are matched with those in A_p, while the preceding letters of a repeat pair are compared to ensure its maximality. The second implementation, *MERF2*, uses only one master array to hold leaf nodes. We describe it in Fig. 47.2 for clarity.

47.2.3 Cache Simulation

We use a trace-driven cache simulator Dinero IV [5] to evaluate cache performance. We assume that the underlying processor has 128 KB unified level-1 4-way set associative cache; the cache is empty after construction, and the only memory access type is read. The genome of *Saccharomyces cerevisiae* is used. The statistics of chromosomes 1 and 4 are summarized in Table 47.1. The numbers of internal and leaf nodes in the suffix tree constructed to search for all the four types of matches are also shown.

```
1: Input: N_p
2: Given the minimal repeat length l_min
3: if the depth of N_p < l_min then
4:     A_m ← an empty array
5:     for each internal child node N_c of N_p do
6:         FINDREPEATS (N_c)
7:     end for
8: else
9:     for each child node N_c of N_p do
10:        n_p ← size of A_m
11:        if N_c is an internal node then
12:            FINDREPEATS (N_c)
13:            n_c ← size of A_m
14:            Match leaf nodes A_m [n_p:n_c −1]
               with A_m [0:n_p −1]
15:        else
16:            Match N_c with leaf nodes
               A_m [0:n_p −1]
17:            Append N_c to A_m
18:        end if
19:    end for
20: end if
```

Fig. 47.2 Our second implementation of FINDREPEATS (N_p)

Table 47.1 Statistics of chromosomes 1 and 4 of the yeast *Saccharomyces cerevisiae*

Chr.	Length	# Internal	# Leaf	Ratio
1	230,208	602,923	920,837	0.65
4	1,531,919	3,954,639	6,127,681	0.65

47.3 Results and Discussion

Because of page limit, we show only the results on chromosomes 1 and 4. To understand how $l_{\min}$ affects the cache performance of MERF, we investigate the cases $l_{\min} = 8$, 12 and 16 in searching for all the four types of matches. Cache performance of the two variants of MERF with varying $l_{\min}$ is shown in Fig. 47.3a c. One is *MERF1* with the *linked-list* suffix tree, denoted as *MERF1 + Linked-list ST*. The other is *MERF2* with the *optimal* suffix tree, denoted as *MERF2 + Optimal ST* or the *optimal cache-oblivious* MERF since it requires no knowledge of cache specifications.

Except for *MERF1 + Linked-list ST* with $l_{\min} = 8$ on chromosome 4, the larger the cache block size, the lower the cache miss rate. Moreover, *MERF2 + Optimal ST* has consistently lower cache miss rates than *MERF1 + Linked-list ST*. For $l_{\min} = 8$, the cache miss rates are very low in all the cases, and thus the difference between the two variants of MERF is not significant. However, for $l_{\min} = 12$ and $l_{\min} = 16$, the differences of cache miss rates are at least around 10 and 20%, respectively.

We now show the improvement in terms of time spent on searching for repeats. For different $l_{\min}$, the percentage of time saved

$$\frac{\text{Time saved by MERF2} + \text{Optimal ST over MERF1} + \text{Linked} - \text{list ST}}{\text{Time spent by MERF1} + \text{Linked} - \text{list ST}}$$

for chromosomes 1 and 4 is listed in Table 47.2. To better understand the extreme cases, results for $l_{\min} = 20$ and 24 are also included. We can observe that, for

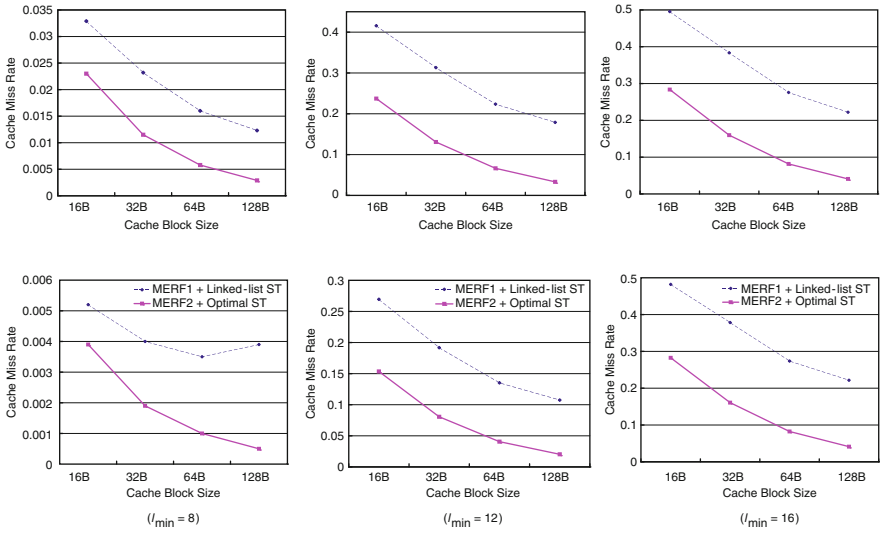


Fig. 47.3 Cache performance of the two variants of MERF. *Top row*: chromosome 1; *bottom row*: chromosome 4

Table 47.2 Improvement made by *MERF2 + Optimal ST* over *MERF1 + Linked list ST* in terms of relative search time

Chr.	Percentage of time saved				
	l_{min} 8	12	16	20	24
1	10.27%	69.23%	77.78%	76.47%	82.35%
4	3.73%	49.18%	82.43%	85.82%	86.33%

$l_{min} = 8$, the shorter chromosomes have higher percentages of improvement than the longer ones, which becomes less evident as l_{min} increases. Furthermore, it is noteworthy that *MERF2 + Optimal ST* is at least twice and five times faster than *MERF1 + Linked-list ST* for $l_{min} = 12$ and $l_{min} = 16$, respectively. As l_{min} approaches 24, however, the increases in saving become less significant.

47.4 Conclusions and Future Work

In this study, we investigated the effects of improved cache performance of suffix trees on MERF. We introduced two suffix tree data structures, the *linked-list* and *pointer array* suffix trees, and constructed the *optimal* suffix tree out of the *pointer array* one in a naïve manner. We described two implementations of MERF based on the *linked-list* and *optimal* suffix trees, respectively.

In Sect. 47.3, we demonstrated the superiority of the *optimal cache-oblivious* MERF to its non-optimized counterpart. The *optimal cache-oblivious* MERF is at least five times faster than the non-optimized one in searching for matches of 16 nucleotides or longer. The results also imply an improvement on REPuter, which depends on MERF. Furthermore, we showed and analyzed the convergence of percentage of run time saving as the minimal length of repeats approached 24.

We have shown that the impact of cache performance on suffix tree operations. To obtain performance gain, one needs to customize the suffix tree data structure for the application of interest. In the future, the cache-conscious allocation strategy [3] may be adapted for a refined algorithm. The statistics of strings may be utilized to improve an algorithm as shown in [9].

Acknowledgment This work was supported in part by NSF grant CCF 0755373.

References

1. Bakalis, A., Iliopoulos, C., Makris, C., Sioutas, S., Theodoridis, E., Tsakalidis, A., Tsichlas, K. Locating maximal multirepeats in multiple strings under various constraints. The Computer Journal **50**(2), 178 185 (2007).
2. Bedathur, S.J., Haritsa, J.R. Engineering a fast online persistent suffix tree construction. In: ICDE '04: Proceedings of the 20th International Conference on Data Engineering, pp. 720 731. IEEE Computer Society, Washington, DC, USA (2004).

3. Chilimbi, T.M., Hill, M.D., Larus, J.R. Cache conscious structure layout. SIGPLAN Not. **34**(5), 1 12 (1999).
4. Gusfield, D. Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press, New York, NY, USA (1997).
5. Hill, M.D., Edler, J. Dinero iv trace driven uniprocessor cache simulator. (1998). Software available at http://www.cs.wisc.edu/ markhill/DineroIV.
6. Kaderali, L., Schliep, A. Selecting signature oligonucleotides to identify organisms using DNA arrays. Bioinformatics **18**(10), 1340 1349 (2002).
7. Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schleiermacher, C., Stoye, J., Giegerich, R. Reputer: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research **29**(22), 4633 4642 (2001).
8. Nicolas, J., Durand, P., Ranchy, G., Tempel, S., Valin, A.S. Suffix tree analyser (stan): Looking for nucleotidic and peptidic patterns in chromosomes. Bioinformatics **21**(24), 4408 4410 (2005).
9. Puzak, T.B., Huang, C.H. An analysis of the effects of spatial locality on the cache perfor mance of binary search trees. In: ICSOFT (2), pp. 94 101 (2006).
10. Sahni, S. Data structures, algorithms and applications in java. McGraw Hill, Inc., New York, NY, USA (1999).
11. Stoye, J., Gusfield, D. Simple and flexible detection of contiguous repeats using a suffix tree. Theoretical Computer Science **270**(1 2), 843 850 (2002).
12. Tata, S., Hankins, R.A., Patel, J.M. Practical suffix tree construction. In: VLDB '04: Proceed ings of the Thirtieth International Conference on Very Large Data Bases, pp. 36 47. VLDB Endowment (2004).
13. Zhang, Y., Zaki, M. Smotif: Efficient structured pattern and profile motif search. Algorithms for Molecular Biology **1**(1), 22 (2006).

Chapter 48
Toward a Visualization of DNA Sequences

David N. Cox and Alan L. Tharp

Abstract Most biologists associate pattern discovery in DNA with finding repetitive sequences or commonalities across several sequences. However, pattern discovery is not limited to finding repetitions and commonalities. Pattern discovery also involves identifying objects and distinguishing objects from one another. Human vision is unmatched in its ability to identify and distinguish objects. Considerable research into human vision has revealed to a fair degree the visual cues that our brains use to segment an image into separate regions and entities. In this paper, we consider some of these visual cues to construct a novel graphical representation of a DNA sequence. We exploit one of these cues, proximity, to segment DNA into visibly distinct regions and structures. We also demonstrate how to manipulate proximity to identify motifs visually. Lastly, we demonstrate how an additional cue, color, can be used to visualize the Shannon entropy associated with different structures. The presence of large numbers of such regions and structures in DNA suggests that they likely play some important biological role and would be interesting targets for further research.

Keywords DNA · Visualization · Repeats · Patterns · Sequence Analysis

48.1 Introduction

Deciphering DNA is an important and open research question. Now that we have the sequences for whole genomes, the goal is to understand what these sequences represent. This goal exists partly to satisfy human curiosity most people want to understand the nature of life. But this goal also exists for practical reasons. Many if

D.N. Cox (✉)
Department of Computer Science, North Carolina State University, Raleigh, NC 27695 8206, USA
e mail: david@guffy.net

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 48,

not all diseases are genetically based. These include cancer, heart disease, and diseases that we know have specific genetic causes such as Downs syndrome and Huntington's disease. Developing treatments and cures will require understanding disease at the genetic level. Understanding how different people respond to medications will require understanding these differences at the genetic level. Developing accurate diagnostic tests also requires understanding diseases at the genetic level.

Deciphering DNA sequences requires an understanding of which subsequences of DNA are transcribed into proteins, which subsequences enhance transcription, and which subsequences inhibit transcription. It also requires understanding how DNA folds itself into compact structures which in turn influence the regulation of transcription.

The predominant approach for gaining this understanding is comparative genomics to explore the similarities and differences in DNA sequences [15, 17]. Other approaches include looking for specific patterns in DNA. These patterns might be specified with a template, might be specified statistically (e.g., using a Hidden Markov Model), or might be specified using a rule (e.g., a highly repetitive pattern). In all of these approaches, it is left to the computer to identify the subsequences of interest. The computer scores the subsequences and presents those with the highest scores to the biologist.

There have been very few attempts to exploit human visual perception as a way to find patterns in DNA. Those that have been attempted were not based on the principles of visual perception. These include dot plots [14], spectrograms [7], DNA walks [3], chaos game representations [12], color coding [20], color merging [1], and repeat graphs and pygrams [8]. Most bioinformaticists are familiar with dot plots, where diagonal lines represent regions of two DNA sequences that match. It is true that our visual system responds effortlessly to lines. However, it also responds to many other visual cues. These include proximity, symmetry, similarity, size, grouping, orientation, curvature, closure, continuity, connectedness, figure, ground, and color.

While the above techniques do produce visual patterns, none have been analyzed in terms of visual perception. Dot plots rely on the perception of lines. Color merging, color coding, pygrams, and spectrograms rely on our ability to detect differences in color. DNA walks rely on our ability to distinguish curved shapes. Chaos game representations, which are fractal-like images, rely on our ability to distinguish shapes. However, none of the authors presented their work in terms of visual perception or attempted to exploit visual perception beyond these singular cues.

Here we describe the visualization of DNA sequences based on the principles of visual perception. We begin with simple visual cues and then build on those cues to reveal more patterns. We start with simple black and white representations of DNA based on proximity. We demonstrate the kinds of patterns revealed. We demonstrate how manipulating proximity can reveal other patterns. Lastly, we demonstrate how adding other visual cues reveals additional patterns from multiple dimensions of information. The presence of large numbers of patterns in DNA

suggests that they play some important biological role. An avenue of future research is to study these patterns in detail using sophisticated tools, such as microarrays and mass spectroscopy, that are now available to the biologists.

48.2 Pattern Discovery

Hardy once wrote, "A mathematician, like a painter or poet, is a maker of patterns." [11] In his book, *Mathematics as a Science of Patterns*, Resnick states, "... in mathematics the primary subject-matter is not the individual mathematical objects but rather the structures in which they are arranged." [18] The first set of results returned by Google when searching for the term "pattern" is a set of images. The first few of these are shown here (Fig. 48.1):

In his book, "Pattern Discovery in Bioinformatics," Parida states that a pattern is a nonunique phenomenon observed in a set of input data [16]. In this definition, repetition is central and repetition is certainly seen in the examples returned by Google. Even common usage of the word, pattern, suggests that repetition is central to its definition. For example, "I think I see a pattern here" is commonly exclaimed after examining a series of events implying that something has been seen repeatedly.

Patterns, however, involve more than repetition. In his book, "Information Visualization: Perception for Design," Colin Ware refers to a pattern as something that can be visualized as a coherent whole [19]. Here are two images. Both contain about 2,000 randomly placed points. However, certain points on the right were painted white to render them invisible. There is no repetition in these images. Even

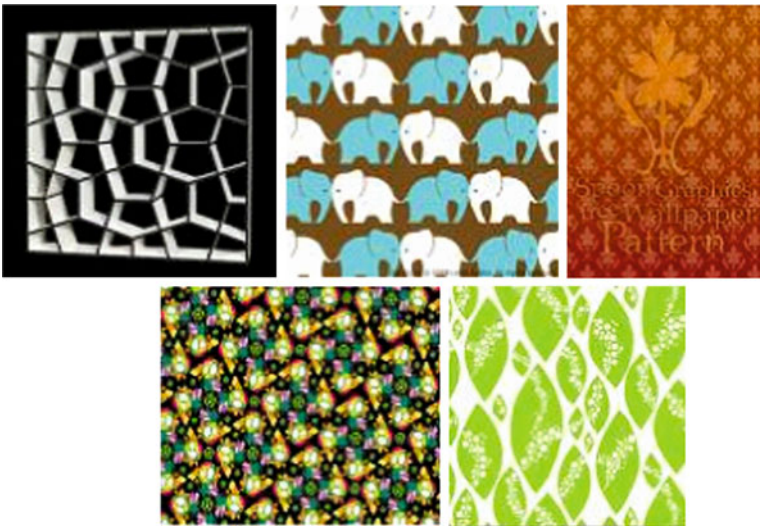


Fig. 48.1 Examples of patterns illustrating the presence of repetition. (Google Image Results [10])

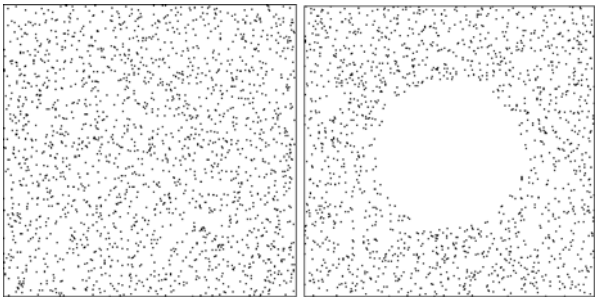


Fig. 48.2 Illustration of a pattern with no repetition in the visual elements

so, the image on the right clearly contains what most people would perceive to be a circle (Fig. 48.2).

Ware asks, "What does it take to for us to see a group? How can 2D space be divided into perceptually distinct regions? Under what conditions are two patterns recognized as similar? What constitutes a visual connection between objects?" Partitioning an image into objects is the process of pattern discovery. To answer these questions and to provide insights into pattern discovery, Ware refers to the Gestalt laws of pattern perception (indeed, the German word gestalt means "pattern").

48.2.1 Gestalt Laws

The estalt school of thought was a group of German psychologists (Max Westheimer, Kurt Koffka, and Wolfgang Kohler) founded in 1912. Although founded nearly a hundred years ago, the Gestalt laws of pattern perception remain valid even today. The laws describe several visual cues that we use to organize what we see. The visual cues are proximity, similarity, connectedness, continuity, symmetry, closure, relative size, and figure and ground.

48.2.1.1 Proximity

Our visual system forms groups for objects that are near each other. This happens automatically during an early stage of visual processing. In the following two images, the distances between the points differ slightly. In the image on the left, the rows are closer together than the columns, causing most people to see the points organized into columns. In the right image, the columns are closer together than the rows, causing most people to see the points organized into rows (Fig. 48.3).

48.2.1.2 Similarity

Similar visual elements tend to be grouped together. The objects on the left are rendered as circles and squares. The distances between the objects are the same, thus ruling out proximity as a grouping factor. Most people group the objects by their shapes and thus see rows of similar objects. Objects of similar color and texture also tend to be grouped together. The objects on the right have the same shape but the color is varied. Again, the shapes are organized into rows based on their similar colors (Fig. 48.4).

48.2.1.3 Connectedness

According to Ware, connectedness was introduced by Palmer and Rock as a fundamental Gestalt organizing principle. Here, the objects connected by lines tend to form groups (Fig. 48.5).

Fig. 48.3 An example of how proximity affects our grouping of object. On the *left,* the vertical distances between the *points* are smaller than the horizontal distances. On the *right*, the reverse is true [19]

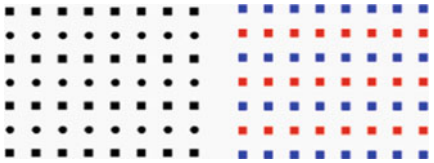


Fig. 48.4 An example of how similarity affects grouping. On the *left,* objects are grouped by similar shape. On the *right,* they are grouped by similar color [19]

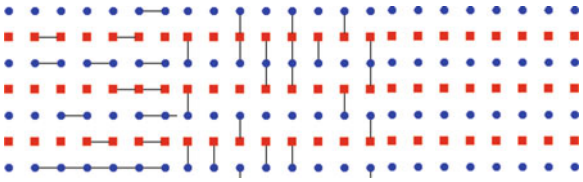


Fig. 48.5 An example of how we group objects that are connected. Despite differences in color and shape, we tend to group the objects that are connected by the *horizontal and vertical lines,* and can identify them very quickly despite the presence of distracters [19]

48.2.1.4 Continuity

In the following image, people tend to see two crossed lines rather than four individual lines that meet at one point. This illustrates the principle of continuity where we construct objects out of visual cues that are smooth and continuous (Fig. 48.6).

48.2.1.5 Symmetry

We tend to perceive two symmetrical objects as a single object. In the following figure the two lines on the left are reported to appear as two separate objects. On the right, one of the lines is flipped. The resulting symmetry causes the two lines to represent a single object (Fig. 48.7).

48.2.1.6 Closure

Closed contours tend to be seen as a single object. We also tend to see contours as closed even though they have gaps. As shown here, the blue figure is perceived to be a whole circle rather than three quarters of a pie (Fig. 48.8).

Fig. 48.6 An example of continuity. Here, we see two *crossed lines* rather than four *individual lines* that meet at a point [19]

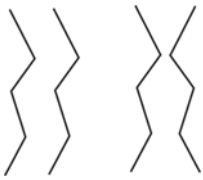


Fig. 48.7 An example of symmetry. On the *left*, we tend to perceive two separate lines. On the *right*, the *two lines* are different by symmetry. Rather than two images, we tend to perceive the *lines* on the *right* as composing a single object [19]

Fig. 48.8 An example of closure. We tend to perceive a *rectangle* in front of a *complete circle* rather than three quarters of a pie

48.2.1.7 Relative Size

Smaller objects tend to be perceived as objects in preference over larger objects. Here we see both blue and red areas. Most people see blue objects over a red background rather than red objects on a blue background (Fig. 48.9).

48.2.1.8 Figure and Ground

The following is a classic example where the image is sometimes perceived to be a single vase and sometimes perceived as two faces. What is perceived depends on how we respond to the image. Sometimes, we see parts of an image as being in the foreground (i.e., figure) and sometimes as being in the background (i.e., ground) (Fig. 48.10).

48.2.2 Preattentive Processing

The Gestalt Laws help to explain how we perceive patterns. Preattentive processing helps to explain why some patterns "pop out" and are distinguishable from other patterns. Preattentive processing occurs early in visual perception. This processing determines which visual features grab our attention.

An example is color. The red circle below is immediately noticed by our visual system. Another example given by Ware is a counting exercise. On the right are several rows of numbers. Some of the 3s are colored red. When given the task of counting the 3s, subjects were able to count the red 3s in constant time regardless of the number of other distracting numbers. When the 3s were the same color as the

Fig. 48.9 An example of relative size. We tend to perceive *small triangles* on a larger background [19]

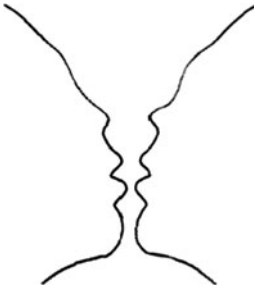


Fig. 48.10 An example of figure and ground. Is it a vase or two faces? [19]

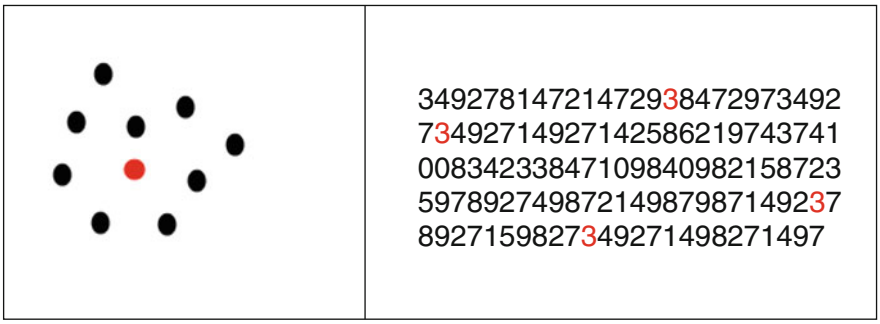


Fig. 48.11 An example of preattentive processing. We can easily find the *red circle* and the *red numbers* despite the presence of distracters [19]

distracters, the time to count them increased linearly as the number of distracters increased (Fig. 48.11).

Using such experiments, it has been determined that features that are preattentively processed fall into the following categories: form, color, motion, and spatial position. These features are as follows:

- Form
  - Line orientation
  - Line length
  - Line width
  - Line collinearity
  - Size
  - Curvature
  - Spatial grouping
  - Blur
  - Added marks
  - Numerosity
- Color
  - Hue
  - Intensity
- Motion
  - Flicker
  - Direction of motion
- Spatial Position
  - 2D position
  - Stereoscopic depth
  - Convex/concave shape from shading

By exploiting Gestalt laws and preattentive processing, it will be shown that we can exploit human visual perception to find patterns in DNA that likely represent subsequences that have specific biological functions.

48.3 Toward a Visualization of DNA

A simple experiment illustrates several aspects of visualization. Take a look at the following DNA sequence for 2 s and then cover the sequence with a piece of paper.

GTGTGTATGCACCTGTGTGTGTGTATGCACCTACGTGTGTGTGTATGC
ACCTGTGTGTGCACCTGTGTGTATGCACCTATGTGTGTGTATGCACCTA
TGTGTGCATGTACCTGTGTGTATGGACCTATGTATGTGTGTATGCGTGT
GTATGCACCTGTGTATGCACCTGTGTGTATGCACCTATGTGTGTGTGTG
TGTATGGACCTATGT

Now, describe the sequence as best you can. Some things to consider are

- Are there any repeats in the sequence?
- If so, how many repeats are there and how long are they?
- Are there any other discernable features in the sequence?
- Do any groups of nucleotides form patterns discernable from other groups of nucleotides?

Now, glance at the following figure and consider the same questions (Fig. 48.12).

The above image is a specially constructed scatter plot (referred to as a symbolic scatter plot) that represents the preceding DNA sequence. Very discernable features are visible in the image. There is a high degree of repetition in the points. There are several clusters of points. We can count the clusters and quite easily determine their lengths. The clusters are separated by other vertical arrangements of points. These vertical arrangements have similar shapes and sizes. All of these features are recognizable immediately after glancing at the image.

Figure 48.13 highlights several of these features in red (yet another visual cue!). We perceive them as similarly shaped curves of the same size and not simply as individual points. We also notice the regularity of the spacing between the shapes.

These groupings of points into recognizable and distinct shapes are the result of pattern discovery. From a variety of visual cues, we have segmented a DNA sequence into smaller entities. With assistance from the computer, we can work backwards from the image and the entities it contains to the DNA sequence. We can

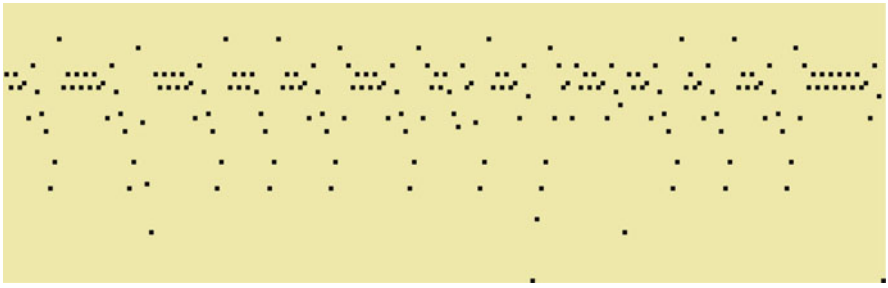


Fig. 48.12 A graphical representation of a DNA sequence that exploits several visual cues: line orientation, line length, spatial grouping, shape, proximity, similarity, and continuity

Fig. 48.13 Continuity allows us to perceive curves from collections of points (*highlighted in red*). Similarity allows us to recognize copies of these shapes. We perceive the shapes as being vertically rather than horizontally arranged. These shapes along with other features, such as *horizontal lines* and the regularity of spacing between features, allow us to segment the DNA easily, which was not possible from the text alone

easily determine which nucleotides correspond to the different visible features. Are these features and their underlying sequences biologically important? Before we delve into this question, let us first explain how symbolic scatter plots are created.

48.4 Symbolic Scatter Plots

A good number of DNA analysis algorithms such as BLAST and Tandem Repeats Finder begin by dividing a DNA sequence into small k-mers. These short strings of DNA range from 3 to 11 nucleotides in length. Subsequent steps in these algorithms rely on the statistical properties of these k-mers. Tandem Repeats Finder is a particularly good example to look at more closely.

Tandem Repeats Finder [2] begins its analysis of DNA by building a hash table for each k-mer in the sequence. Benson provides an example with $k = 5$. Consequently, the possible 5-mers are AAAAA ... TTTTT. A window of length 5 slides along the sequence and the position of the window is noted for the 5-mers residing in the window. These positions are added to the corresponding 5-mer's hash table (referred to by Benson as the 5-mer's "history list").

When a new position is added to a history list, the position is compared to other positions already in the list. The hypothesis is that each position corresponds to a possible tandem repeat. To validate the hypothesis, the distance d between two positions i and j is calculated. To determine if repeats are present at i and j, other 5-mers between i and j are examined. If these 5-mers are also found to have a copy a distance d away, then a repeat is flagged as detected. To allow for possible substitutions, insertions, and deletions, a sum of heads distribution, a random walk distribution, and an apparent size distribution are applied to evaluate the two potential repeats. If the required heuristics are met, the repeats are reported to the user. The process is depicted in Fig. 48.14.

In comparison, symbolic scatter plots are created by first determining the positions of k-mers in a sequence. Rather than k-mers of size 5, the typical symbolic scatter plot uses size 3. History lists for each 3-mer are constructed visually. The

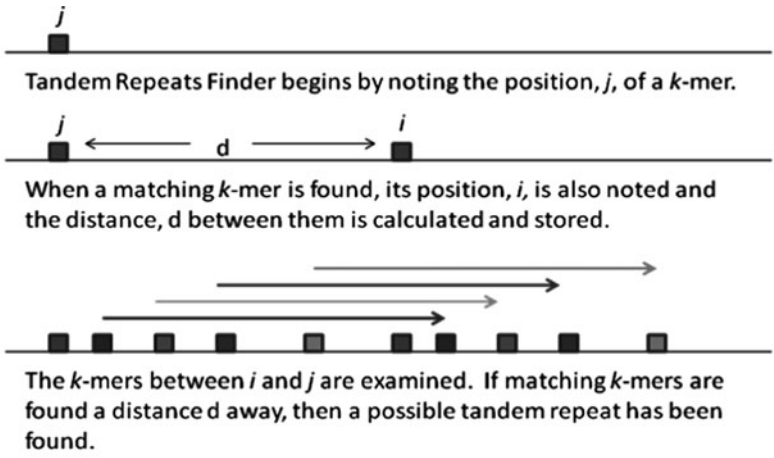
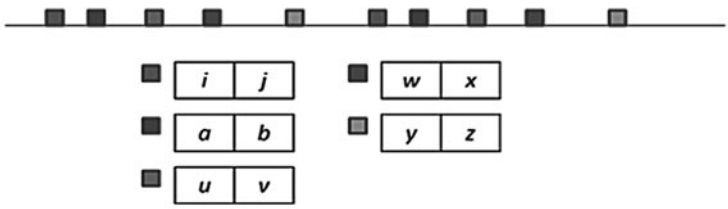


Fig. 48.14 Illustration of *k* mer processing in Tandem Repeats Finder

y-axis corresponds to the entire set of possible 3-mers where each 3-mer occupies a single row. The history list corresponds to the columns of each row. Thus, if AAA is the 100th 3-mer in the sequence, then a point is plotted in the 100th column of row AAA. The end result is a scatter plot of points that map one-to-one to the positions recorded in the history lists of Tandem Repeats Finder. These differences are illustrated in Fig. 48.15.

More simply, the *x*-coordinate of each point corresponds to the position of the 3-mer in the sequence. The *y*-coordinate for each point corresponds to a numerical value assigned to the 3-mer. Because there are four nucleotides, there are $4^3 = 64$ possible 3-mers. By numbering each from 0 to 63, we are able to create a mapping from each of the 64 possible 3-mers to a unique number. Thus, we are able to covert a sequence of characters (the As, Cs, Gs, and Ts of DNA) to a sequence of integers. Using these integers, we are able to create the symbolic scatter plot.

Figure 48.16 presents two examples of symbolic scatter plots created from the human genome assembly, specifically the Y-chromosome. Both of these scatter plots were created for the following contig: gi|89061207|ref|NT 113967.1| HsY 111686. The top image corresponds to the sequence beginning at position 2715 and the bottom image corresponds to the sequence beginning at position 6433. This first thing that strikes the viewer is the high degree of similarity in these images. This similarity is detectable from the features visible in the plots. Although highly similar, the images also present differences. A few of these differences are indicated by the arrows.



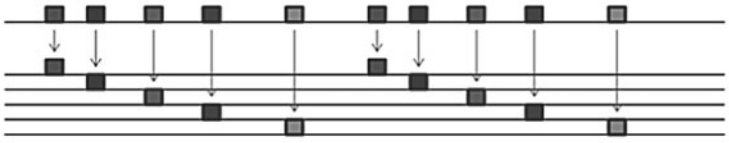


Fig. 48.15 Illustration of *k* mer processing in a symbolic *scatter plot*

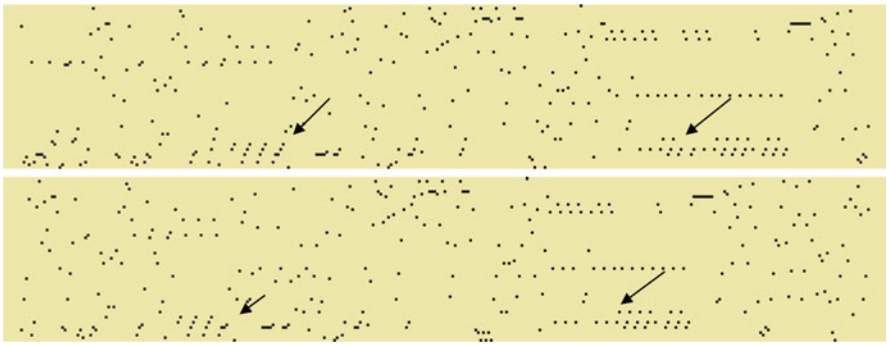


Fig. 48.16 Comparing sequences visually is easier with additional visual cues. Text visually presents a sequence. However, the As, Cs, Gs, and Ts all have the same relative size and shape. Cs and G are particularly difficult to differentiate

Now consider comparing the text of the sequences:

GACACCCCACAAAACACCCCCCCCCCCCCACAAAACACTCTATACA
CACAAAAAAAAAACACTCTCTCACAAAACACCCCTCTGTGGGGAGAAA
ATATTTTTTTTTTCTCCCCCCCACAGAGGGGAGACACCCCACAAAAAAA
CACCCCCCCACATATCTCTCTATATATATATATTTTCTCTCTGTGTGTAT
AAAACACCCCCCCGCGAGAAAAAAACACCCCTCTCTCACAAAAAAAGA
GCGCCCCTCTATAAAACACCCCCCCTCTATAAAACACCCCCCCTCTATA
AAACACCCCCCCCCCTCTATACACACAGAGTGTTTTGTGAG

CCACAGAGAGACACCCCTCTGTGGGGAGAAAATATTTTTTTTTCTC
CCCCCCACAGAGGGGAGACACCCCACAAAAAAACACCCCCCCACATAT
CTCTCTATATATATATATTTTCTCTCTGTGTGTATAAAACACCCCCCCGC
GAGAAAAAAATATCTCTCTCTCACAAAAAAAGAGCGCCCCTCTATAAA
ACACCCCCCCTCTATAAAACACCCCCCCTCTATAAAACACCCCCCCCCC
GCGAGACACACAGAGTGTTTTCTCACAGAGGGGTGTCTCCCCCCCCCCC
CCCCCGCGTGTCTCCCCCCCCCCGCGTGTATAGAGCGCTC

Although repeats of Ts, C and some As are visible in the text, the similarities and differences of various features are not as readily apparent as in the scatter plot. This is because the scatter plots take advantage of additional visual cues to make greater distinctions between different regions of the DNA sequences. These cues include the proximity of the points to each other and the shapes, sizes, and orientations of the various structures visible in the plots.

48.5 Manipulating Visual Cues

Symbolic scatter plots are extremely easy to create and have some properties that can be manipulated to enhance pattern discovery. The first of these properties is that a single point is plotted in each column of the plot. The Gestalt law of proximity tells us that we tend to perceive patterns from objects that are near each other rather than far apart. By relaxing the rule that a single point occupies a column, we can move the points in an image closer to each other. The effect is to "squish" the points horizontally.

Figure 48.17 illustrates the result. The top image is an original symbolic scatter plot that appears to display a random array of points. In the subsequent scatter plots, more than one 3-mer is plotted in a column. The second panel of Fig. 48.17 plots two points per column. The bottom two panels plot four and eight points per column, respectively. If there are copies of a 3-mer in a group, a single point in the column represents all copies. While this does represent a loss of information, the closer proximity of the entire set of points does allow one to see patterns that would not otherwise be visible. Repetitive patterns not obvious in the first panel become much more noticeable in the subsequent panels. The bottom panel even shows how one pattern ends abruptly with a random looking region of points followed by another more orderly region.

Another way to manipulate proximity is to change the mapping of 3-mers to y-coordinates. This is particularly useful for visualizing motifs. Given a motif of length n, determine the overlapping 3-mers occurring in the motif. If 3-mers repeat, then eliminate the repeats, leaving only one of each 3-mer. Using this set of unique 3-mers, assign the 3-mers consecutive y-values. For example, in the motif TATA-TATA, the 3-mer TAT is found three times, as is ATA. These two 3-mers are remapped so that they occupy consecutive y-values. Regions of DNA containing these 3-mers will appear as small diagonal lines. Mismatches will be visible as small diagonal lines with gaps.

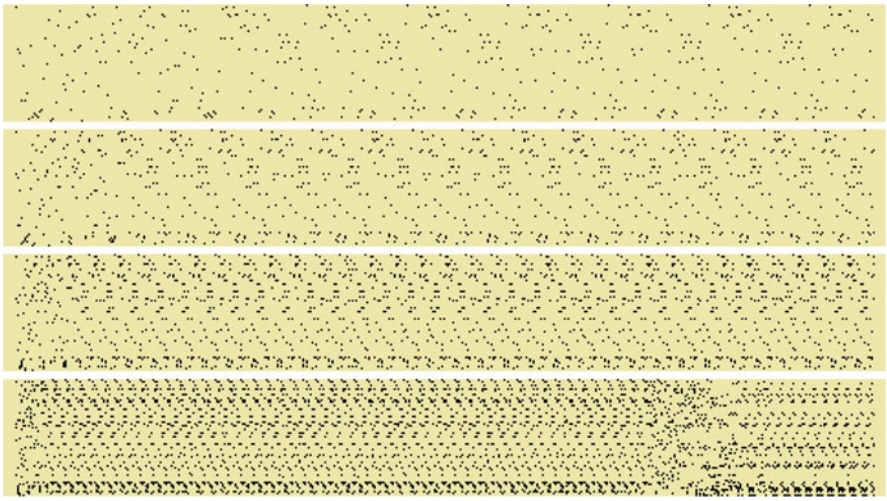


Fig. 48.17 Manipulating visual cues can render patterns more visible. Here, proximity is manipu lated by bringing the points closer together horizontally. This "squishing" of the points transforms an apparent random display of points into a highly ordered array of 3 mers

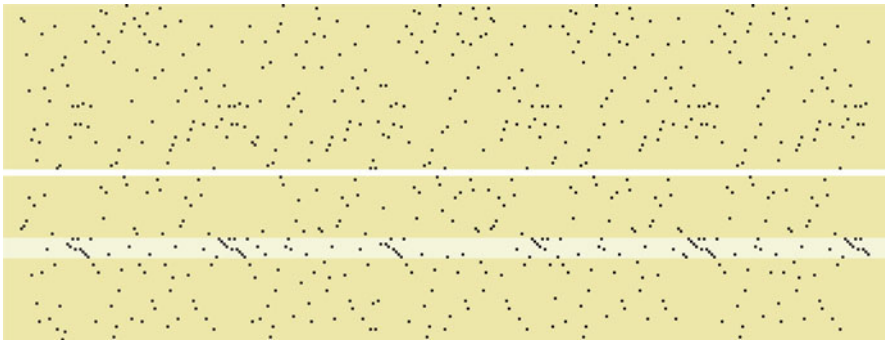


Fig. 48.18 The visual cue proximity can also be manipulated by remapping the *y* values associated with selected 3 mers. Doing so allows motifs to become visible that would otherwise be impossible to see

The top panel of Fig. 48.18 shows another apparently random set of points. A closer look does show some repetition in the plot. Using the motif AGATAGA GAGCAG, we can remap the *y*-values for the 3-mers AGA, GAT, ATA, TAG, GAG, AGC, and CAG to force them to occupy the center of the plot. The remaining 3-mers are mapped to other *y*-values in no particular order. The bottom panel of Fig. 48.18 shows the result. The middle region that these 3-mers occupy is highlighted to enhance their visibility. The diagonal clusters in this highlighted area indicate approximate matches of this motif at several locations in the sequence. By

clicking on these diagonals, one can immediately see the exact sequence and can determine to what degree they match the provided motif.

Aside from proximity, we can enhance the scatter plots with color to highlight additional dimensions of information. An example of an extra dimension of information is Shannon entropy for DNA. Shannon entropy is a measure of the minimum number of bits needed to encode information. Here we use a sliding window of length 64 that is centered on each 3-mer. As the window slides across the DNA sequence, the Shannon entropy is calculated and assigned to the 3-mer at the center of the window. Color is used to represent how close the entropy is to the highest or lowest entropies in the plot. Green represents entropy that is near the highest entropy value and red represents entropy near the lowest entropy value. Intermediate values map to shades of green and red between pure green at the high end and pure red at the low end. Figure 48.19 illustrates the result.

In some cases, regions of low and high entropy are not so obvious from the scatter plot alone. Color helps to identify these regions, while retaining the underlying patterns of 3-mers as revealed in the scatter plot. Figure 48.20 gives an example.

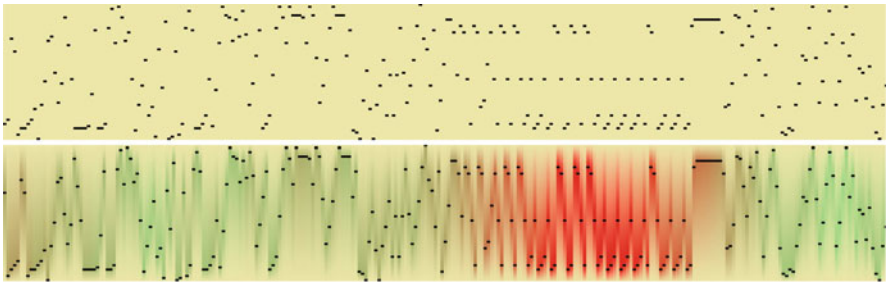


Fig. 48.19 Additional dimensions of information can be added to a symbolic *scatter plot* with additional visual cues. Here, entropy is added using color using a scale that varies from *green* for high entropy to *red* for low entropy

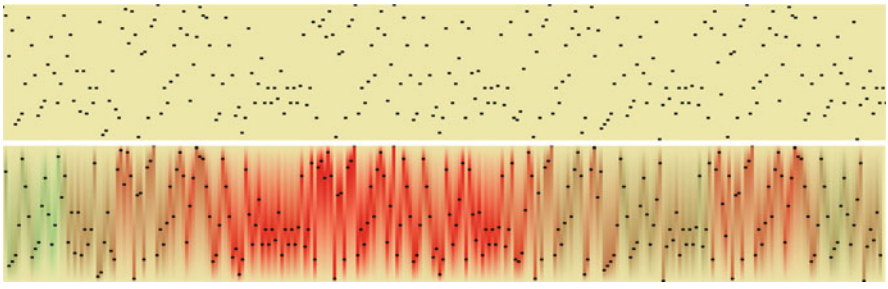


Fig. 48.20 Frequently it is not obvious in a symbolic *scatter plot* which regions have high and which have low entropy. *Color* again makes the distinction clear

48.6 Discussion

Scatter plots are nothing new, and most biologists are familiar with comparing sequences using dot plots. The predominant visual cue in dot plots is the diagonal line allowing one to recognize regions of similarity *between* two sequences. Symbolic scatter plots differ by revealing patterns *within* a single sequence. The predominant visual cue in symbolic scatter plots is proximity, allowing one to recognize visually distinct regions and structures with varying shapes and sizes within a sequence.

Proximity reveals clusters of points, horizontal and vertical lines, diagonal lines, curves, and other arrangements of points. Proximity also reveals regions that vary by density of points. In some regions, the points will be widely separated from each other. In other regions, the points will be much closer together.

By manipulating proximity, we can reveal patterns that would not otherwise be visually obvious. By compressing or "squishing" images horizontally, patterns arise by bringing points into closer proximity. By remapping 3-mers to bring them into closer proximity vertically, we can reveal clusters of points whose underlying DNA is similar to a given motif. Other visual cues such as color can be added to reveal additional dimensions of information.

Symbolic scatter plots allow us to *see* DNA as something other than a sequence of letters. Patterns that are not apparent in text become obvious when rendered with additional visual cues. What do these patterns represent? Seeing them is a first step. Now that we can see them, we can target them for further research. We can run microarray experiments to see if these patterns are active in tissue samples. We can conduct experiments to see if they enhance or inhibit transcription. Pattern discovery is not the final answer. But it is a tool to help us to find the final answer.

The software for creating symbolic scatter plots is freely available. For more information, please contact one of the authors.

References

1. Alston M, Johnson C G, & Robinson G (2003). Colour merging for the visualization of biomolecular sequence data. *Proceedings of the Seventh International Conference on Infor mation Visualization*.
2. Benson G (1998). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research, 27* (2), 573 580.
3. Berger J A, Mitra S K, Carli M, & Neri A (2004). Visualization and analysis of DNA sequences using DNA walks. *Journal of the Franklin Institute*, 37 53.
4. Buchner M, & Janjarasjitt S (2003). Detection and visualization of tandem repeats in DNA sequences. *IEEE Transactions of Signal Processing*, *51* (9).
5. Cox D N (2008). Visualizing the sieve of eratosthenes. *Notices of the American Mathematical Society, 55* (5), 579 582.
6. Cox D N, & Dagnino L (2009). An analysis of DNA sequences using symbolic scatter plots. *Accepted for publication at BIOCOMP '09*.

7. Dimitrova N, Cheung Y H, & Zhang M (2006). Analysis and visualization of DNA spectro grams: open possibilities for the genome research. *ACM Multimedia 2006 Conference*, 1017 1024.
8. Durand P, Mahe F, Valin A, & Nicolas J (2006). Browsing repeats in genomes: pygram and an application to non coding region analysis. *BMC Bioinformatics*, *7*, 477.
9. Frith M C, Fu L, Chen J, Hansen U, & Weng Z (2004). Detection of functional DNA motifs via statistical over representation. *Nucleic Acids Research*, *32*, 1372 1381.
10. *Google Image Results*. (2009). Retrieved from Google: http://www.google.com
11. Hardy G H (1940). *A Mathematician's Apology*. Cambridge: University Press.
12. Jeffrey H J (1990). Chaos game representation of gene structure. *Nucleic Acids Research*, *18* (8), 2163 2170.
13. Keich U, & Pevzner P A (2002). Finding motifs in the twilight zone. *Proceedings of the Sixth Annual International Conference on Computational Biology*, 196 204.
14. Maizel J V, & Lenk R P (1981). Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proceedings of the National Academy of Sciences USA*, *78* (12), 7665 7669.
15. Mount D W (2004). *Bioinformatics Sequence and Genome Analysis* (Second ed.). Cold Spring Harber: Cold Spring Harbor Laboratory Press.
16. Parida L (2008). *Pattern Discovery in Bioinformatics*. Boca Raton: Chapman & Hall/CRC.
17. Pevsner J (2003). *Bioinformatics and Functional Genomics*. Hoboken, NJ: John Wiley & Sons, Inc.
18. Resnick M D (1997). *Mathematics as a Science of Patterns*. Oxford: Oxford University Press.
19. Ware C (2004). *Information Visualization: Perception for Design*. San Francisco: Morgan Kaufmann Publishers.
20. Yoshida T, Obata N, & Oosawa K. (2000). Color coding reveals tandem repeats in the *Escherichia coli* genome. *Journal of Molecular Biology*, *298*, 343 349.

Notes

Sources for images returned from google:
http://www.alfredo haeberli.com/products/pattern/1.html,
http://printpattern.blogspot.com/2008 02 01 archive.html,
http://www.blog.spoongraphics.co.uk/freebies/free ornate wallpaper pattern,
http://evildesign.com/2007/04/wzrd pattern wzrd pattern.html,
http://printpattern.blogspot.com/2007 05 01 archive.html

Chapter 49
A Practical Approach for Computing the Active Site of the Ribonucleoside Hydrolase of *E. coli* Encoded by rihC

Anthony Farone, Mary Farone, Paul Kline, Terrance Quinn, and Zachariah Sinkala

Abstract We predict the potential active and catalytic sites, the transition state and how it is stabilized, and the mechanism of rihc Ribonucleoside hydrolase of *E. Coli*. Our approach is based on well-known primary sequence analysis techniques. A canonically associated extreme value distribution is used to assess the significance of the prediction. Parameters for the extreme value distribution are computed directly from data. Our practical approach is consistent with known results in the literature. We obtain BLOSUM matrices in a way that is intrinsically tied to the data base, and we employ user-friendly techniques that should be applicable to a range of medically significant scenarios.

Keywords Amino acids · BLOSUM matrices · *E. coli* · rihC · Sequence analysis

49.1 Introduction

This article extends previous work [4]. We report on recent progress toward identifying the active site within the primary sequence, for rihC ribonucleoside hydrolase of *Escherichia coli*. Using sequence analysis of rihC and inosine uridine nucleoside hydrolase (IU-NH) from *Crithidia fasciculata*, a number of amino acids have been identified as potentially important in the interaction of the enzyme with its substrate. Results for IU-NH can be found in [1, 2, 5, 10]. Sequence analysis of IU-NH from *C. fasciculata* (the most studied nucleoside hydrolase) and rihC identifies (likely) amino acids involved in the active site of the enzyme. We also

A. Farone (✉)
Middle Tennessee State University, Murfreesboro, Tennessee 37132, USA
e mail: afarone@mtsu.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 49,

obtain: a list of potential active site residues; information on the mechanism of the enzyme encoded by rihC; and identification of the transition state. See Sect. 49.3 for detailed findings.

Our approach makes use of sequence alignment methodology [3, 7, 8]. We use the primary sequence of IU-NH from *C. fasciculata* as a basis for comparison with rihC. Residuals central to the activity of an enzyme are conserved in the enzyme family. The method uses scoring functions to assess local similarity between the rihC sequence and IU-NH of *C. fasciculata* sequence. We obtain a BLOSUM r matrix in a new way so that the choice of $r\%$ is intrinsically tied to the data base. A more detailed description of our basic mathematical results follows.

49.2 Parameter Calculations

49.2.1 Visiting Negative Values

As in [7], p. 212, we look to the sets

$$E_{\ j} = \begin{cases} s = \text{the infinite sequences} = s_0 s_1 s_2 \cdots s_n\ {}_1 s_n \cdots, \\ \text{the first negative value is} -j \end{cases}$$

Using C++ code (pseudocode below), one may use the probability distribution of T to generate large sample sets of the random walks just defined. Relative actual frequencies provide approximations for $\{R_{\ j}, j = 1, ..., c\}$.

The pseudocode for computing $\{R_{\ j}, j = 1, ..., c\}$ is:

- For i from 1 to M do
- $S \leftarrow 0$
- while $S \leftarrow 0$ do
- $y \leftarrow$ select randomly from T using empirical probabilities $p_{\ c}, ..., p_d$
- $S \leftarrow S + y$
- end do
- count$[-S] \leftarrow$ count$[-S] + 1$
- end do:
- $R(-k) \leftarrow \text{count}[k]/M, k = 1, 2, \ldots, c.$

Note that by definition of random walks in [7], except for a set of random walks E_0 of measure zero, all random walks drift to $-\infty$. (See [7], p. 214.) Hence, with $R_{\ j} = \text{Prob}(E_{\ j})$, we get

$$\sum_{j=c}^{j=1} R_{\ j} = 1.$$

49.2.2 Visiting Positive Values

We also need to determine Q_k, the probability that a trajectory visits the positive integer k before visiting any other positive value. Note that because $p_c > 0$ and $p_d > 0$, we get

$$0 < Q_1 + Q_2 + \cdots + Q_d < 1.$$

As in [7], we have the following equations:

$$\sum_{k=1}^{d} Q_k \exp(k\lambda^\star) = 1 \tag{49.1}$$

$$\overline{Q} = 1 - \sum_{k=1}^{d} Q_k \tag{49.2}$$

Using a similar approach as for R_j, one may use C++ code to approximate the values for Q_k by using sampling of random walks.

The pseudocode for computing $\{Q_k, k = 1, ..., d\}$ is:

- For k from 1 to d do
- count$[k] \leftarrow 0$
- end do:
- for i from 1 to M do
- $S \leftarrow 0$
- while $S <= 0$ do
- $y \leftarrow$ select step size in T randomly with respect associated empirical probabilities $p_c, ..., p_d$
- $S \leftarrow S + y$
- count $n \leftarrow$ count $n + 1$
- if (count $n >=$ length of the random walk) break;
- end of while loop
- if $S > 0$ then count$[S] \leftarrow$ count$[S] + 1$
- end do loop
- $Q_k \leftarrow$ count$[k]/M$, $k = 1, ..., d$

49.2.3 Parameters for the Associated Extreme Value Distribution

As is well known from random walk theory, the extreme value distribution is of the form

$$P(Y_{\max} \geqslant \alpha) \approx 1 - \exp\left(-KN \exp(-\lambda\alpha)\right) \tag{49.3}$$

for all large α [7]. Let the symbol t refer to the step sizes from the set T and define $M(\lambda) = E(\exp(t)) = \sum_{j=\ c}^{d} \exp(\lambda j) p_j$, as defined above.

The parameter λ^* is easily computed as the unique positive number satisfying $M(\lambda^*) = 1$. To obtain K note that $K = \frac{C}{A}\exp(-\lambda)$ [7], p. 224, eq. 7.25, where $\lambda = \lambda^*$ and both C and A can be given explicitly in terms of $\lambda^*, R_{\ j}, Q_k$ and where as defined above, $j = -c, ..., -1$ and $k = 0, 1, ..., d$.

Explicitly

$$C = \frac{\overline{Q}\left(1 - \sum_{j=1}^{c} R_{\ j} \mathrm{e}^{\ j\lambda^*}\right)}{(1 - \mathrm{e}^{\ \lambda^*})\left(\sum_{k=1}^{d} kQ_k \mathrm{e}^{k\lambda^*}\right)}, \tag{49.4}$$

and

$$A = \frac{\sum_{j=1}^{d} jR_{\ j}}{-\sum_{j=\ c}^{d} jp_j}. \tag{49.5}$$

One way to compute C and A is to use power series approximations (see, e.g. [9]). However, as indicated above, a direct and computationally straightforward way is to run large sample sets of random walks in order to directly obtain approximations for $R_{\ j}, Q_k$. The values obtained can then be substituted into the defining equations for C and A.

To apply this structure to sequence alignments, we suppose two sequences of lengths N_1 and N_2, respectively. Let N be the length of the alignment of two sequences with gaps resulting from local alignment by the Smith and Waterman dynamic algorithm. Note that $N \geq \max\{N_1, N_2\}$. In this setting, $Y = s$ the score function and the general result [7], Eq. 7.26, p. 224 is that

$$P(s_{\max} \geq \alpha) \approx 1 - \exp\left(-KN^2 \exp(-\lambda\alpha)\right) \tag{49.6}$$

49.3 Application

Constructing a substitution matrix, we avoid Markov chain interpolation to construct substitution matrix since the conserved regions of nucleoside hydrolase are more divergent in the database protein blocks. We estimate the pattern directly from sequences in the blocks [6]. We obtain $r = 92$, meaning that we obtain *BLOSUM*92 as the substitution matrix. We obtain 55 clusters in the first block, 62 clusters in the second, and 55 in the third. For details of construction *BLOSUM*92, see [7]. In Table 49.1, we obtain background frequencies associated with *BlOSUM*92.

Table 49.1 Background frequencies

p_A	$\frac{272671}{2195680}$	0.1241852182	p_C	$\frac{2353}{219568}$	0.01071649785
p_D	$\frac{63841}{658704}$	0.09691910175	p_E	$\frac{179527}{3293520}$	0.05450915738
p_F	$\frac{4215}{109784}$	0.03839357283	p_G	$\frac{18197}{137230}$	0.1326022007
p_H	$\frac{6451}{219568}$	0.02938041973	p_I	$\frac{27203}{439136}$	0.06194664068
p_K	$\frac{10059}{439136}$	0.02290634337	p_L	$\frac{23131}{439136}$	0.05267388691
p_M	$\frac{1995}{109784}$	0.01817204693	p_N	$\frac{49321}{1097840}$	0.04492549005
p_P	$\frac{16885}{219568}$	0.07690100561	p_Q	$\frac{1873}{109784}$	0.01706077388
p_R	$\frac{11739}{439136}$	0.02673203746	p_S	$\frac{78567}{2195680}$	0.03578253662
p_T	$\frac{5527}{109784}$	0.05034431247	p_V	$\frac{3837}{54892}$	0.06990089631
p_W	$\frac{1467}{109784}$	0.01336260293	p_Y	$\frac{4959}{219568}$	0.02258525833

Next, we construct *BLOSUMr* matrix using $s(a,b) = C \times ln\left(\frac{p_{ab}}{p_a p_b}\right)$ with $C = \frac{2}{ln2}$. This gives *BLOSUM*92.

We use this BLOSUM matrix to determine the parameter K of associated extreme value distributions. K is determined by $\lambda^\star$, C, and A (see 49.5 above) for the random walk determined by the above substitution matrix.

The quantity $\lambda = \lambda^\star$ is the unique strictly positive solution of the equation $\sum_{a,b\in\sigma} p_i p_j \exp(\lambda s(a,b)) = 1$. Using the Newton Raphson method, we get that $\lambda^\star = 0.3388306068$. As defined, the quantity A depends on the set of quantities $\{R_j\}_{j=1}^{c}$, where R_j is the probability that a walk finishes at the first negative $-j$; and on the background frequencies. In a similar way, the quantity C depends on the set of quantities $\{Q_k\}_{k=1}^{d}$, where Q_k is the probability for a walk visits $k > 0$ before visiting any other positive value.

As in the general theory, the BLOCK substitution matrix can be used to generate random walks. For the matrix S obtained above, we get step sizes

$$T = \{-c, -c+1, ..., 0, 1, d-1, d\},\ c = 5 \text{ and } d = 7,$$

with respective probabilities in Table 49.2.

Observe also that conditions of a random walk are satisfied (See [7]). That is

1. $p_c > 0$ and $p_d > 0$ $c = 5, d = 7$;
2. The step size has negative mean, $\sum_{j=c}^{d} jp_j = -1.117791283 < 0$; and
3. The greatest common divisor of all positive elements $j \in T$ for which $p_j > 0$ is clearly equal to 1.

Table 49.2 Probabilities of the step sizes

p_{-5}	0.04380850588	p_{-4}	0.06481974156	p_{-3}	0.1313357057
p_{-2}	0.2107522459	p_{-1}	0.2952402369	p_0	0.08816745483
p_1	0.06261746145	p_2	0.01418994226	p_3	0.01858022449
p_4	0.04748032555	p_5	0.006422547111	p_6	0.01358925624
p_7	0.002996352045				

Table 49.3 Computation of R_{-j} for 10,000 random walks

$R_{-5} \approx 0.0501$	$R_{-4} \approx 0.0845$	$R_{-3} \approx 0.1711$	$R_{-2} \approx 0.2777$
$R_{-1} \approx 0.4166$			$\sum_{j\ 1}^{c} R_{-j} \approx 1$

Table 49.4 Computation of R_{-j} for 50,000 random walks

$R_{-5} \approx 0.05216$	$R_{-4} \approx 0.08636$	$R_{-3} \approx 0.17196$	$R_{-2} \approx 0.27714$
$R_{-1} \approx 0.41238$			$\sum_{j\ 1}^{c} R_{-j} \approx 1$

Table 49.5 Computation of Q_k for 10,000 random walks

$Q_1 \approx 0.1334$	$Q_2 \approx 0.0601$	$Q_3 \approx 0.0588$	$Q_4 \approx 0.0695$
$Q_5 \approx 0.0143$	$Q_6 \approx 0.0182$	$Q_7 \approx 0.0034$	$\sum_{k\ 1}^{d} Q_k \exp(\lambda Q_k) \approx 1$

Table 49.6 Computation of Q_k for 50,000 random walks

$Q_1 \approx 0.13082$	$Q_2 \approx 0.06486$	$Q_3 \approx 0.05664$	$Q_4 \approx 0.06882$
$Q_5 \approx 0.01588$	$Q_6 \approx 0.01762$	$Q_7 \approx 0.0034$	$\sum_{k\ 1}^{d} Q_k \exp(\lambda Q_k) \approx 1$

For approximating the probabilities $R_{-5}, \ldots, R_{-1}$ and $Q_1, \ldots, Q_7$, we used the pseudocodes from Sect. 49.2. For both R_{-j} and Q_k we ran two sets of simulations, with 10,000 and 50,000 random walks, respectively.

For R_{-j}, 10,000 random walks gave an error of 0.01. See Table 49.3.

For 50,000 random walks, the error term is 0.00442135954, and we obtained the results in Table 49.4.

In a similar way, for $Q_1, \ldots, Q_7$, 10,000 random walks gave an error term of .01, and 50,000 walks gave an error term of .005572135954. Our two sets of values for Q_k are as follows: For 10,000 walks, see Table 49.5; and for 50,000, see Table 49.6.

We obtain quantities $K = 0.09300930513$ and $K = 0.1159884173$ for 10,000 and 50,000 random walks, respectively.

Next, we identify a member of the rihC family about which active and catalytic sites and the transition state are known. IU-NH from *C. fasciculata* belongs to family and results can be found in [1, 2, 5, 10]. The enzyme catalyses the hydrolysis of all purine and pyramidine nucleosides, and the hydrolysis forms ribose and the base.

We use the Smith and Waterman algorithm with gap 10 and extension of 0.2 implemented in a C++ program, to find the local alignment between IU-NH from *C. fasciculata* and rihC ribonucleoside hydrolase of *Escherichia coli* to predict possible catalytic sites and active sites of rihC. The local alignment score is $s_0 = 433.50$.

Sequence analysis of IU-NH from *C. fasciculata* and rihC helps identify (likely) amino acids involved in the active site of the enzyme. We also obtain: a list of potential active site residues; information on the mechanism of the enzyme encoded

by rihC; and identification of the transition state. His233 in its primary sequence acts as a proton donor to activate the hypoxathine leaving group. The catalytic site contains Asp10, Asn165, and His233. The active site contains Asp10, Asp14, Asp15, and Asp234 along with Ile121, coordinate of divalent cation. Its mechanism is that His233 acts as an acid to protonate the N7 of the purine or N4 of the pyrimidine of the leaving group. The Ca^{2+} ion together with Asp10 activates a water molecule which nucleophilically attacks ribose. Asp10 accepts a proton from the water molecule. This leads to the possibility of a transition state which spontaneously dissociates to form ribose and purine or ribose and pyrimidine. We also propose that the transition state is stabilized by Asn165.

$$E_{\text{value}} = 0.2024682270 \times 10^{-59} \quad \text{and } P_{\text{value}} = 0.000$$

Since the *P*-value of the local alignment of rihC and IU-NH from *C. fasciculata* is less than say 0.05, we conclude that the above conclusions are probably biologically significant. We are very grateful to the Office of Graduate Studies at Middle Tennessee State University for helping support this research project. (Award # 2-21462).

References

1. Degano, M., Almo, S. C., Sacchettini, J. C., Schramm, V. L. Trypanosomal nucleoside. A novel mechanism from the structure with a transition state inhibitor. N ribohydrolase from *Crithidia fasciculata*. Biochemistry. **37**, 6277 6285 (1998).
2. Degano, M., Gopaul, D. N., Scapin, G., Schramm, V. L., Sacchettini, J. C. Three dimensional structure of the inosine uridine nucleoside N ribohydrolase from *Crithidia fasciculata*. Bio chemistry. **35,** 5971 5981 (1996).
3. Ewens W. J., Grant, G. R. Statistical Methods in Bioinformatics. 2nd ed., Springer, New York (2004).
4. Farone, A., Farone, M., Khaliq, A., Kline, P., Quinn, T., Sinkala, Z. Identifying the active site of ribonucleoside hydrolase of *E. coli* encoded by RihC. In: Arabnia H. R., Yang, M. Q. (eds.) Proceedings of The 2009 International Conference on Bioinformatics & Computational Biology: BIOCOMP2009 Vol. I, pp. 216 218, CSREA Press, USA (2009).
5. Gopaul, D. N., Meyer, S. L., Degano, M., Sacchettini, J. C., Schramm, V. L. Inosine uridine nucleoside hydrolase from *Crithidia fasciculata*. Genetic characterization, crystallization, and identification of histidine 241 as a catalytic site residue. Biochemistry. **35**, 5963 5970 (1996).
6. Henikoff, S., Henikoff, J. G. Amino acid substitution matrices from protein blocks. Proc. Nat. Acad. Sci. USA. **89**, 10915 10919 (1992).
7. Isaev, A. Introduction to Mathematical Methods in Bioinformatics. Springer, Berlin German (2004).
8. Karlin, S., Altschul, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl Acad. Sci. USA. **87**, 2264 2268 (1990).
9. Karlin, S., Dembo, A. Limit distributions of maximal segmental score among Markov dependent partial sums. Adv. Appl. Prob. **24**, 113 140 (1992).
10. Mazumder, D., Kahn, K., Bruice, T. C. Computer simulations of trypanosomal nucleoside hydrolase: Determination of the protonation state of the bound transition state analogue. J. Am. Chem. Soc. **124**, 8825 8833 (2003).

Part V
Drug Design, Drug Screening, and Related Topics

Chapter 50
Addressing the Docking Problem: Finding Similar 3-D Protein Envelopes for Computer-Aided Drug Design

Eric Paquet and Herna L. Viktor

Abstract Consider a protein (P_X) that has been identified, during drug design, to constitute a new breakthrough in the design of a drug for treating a terminal illness. That is, this protein has the ability to dock on active sites and mask the subsequent docking of harmful foreign proteins. Unfortunately, protein X has serious side effects and is therefore not suitable for use in drug design. Suppose another protein (P_Y) with similar outer structure (or envelope) and functionality, but without these side effects, exists. Locating and using such an alternative protein has obvious benefits. This paper introduces an approach to locate such similar protein envelopes by considering their three-dimensional (3D) shapes. We present a system which indexes and searches a large 3D protein database and illustrate its effectiveness against a very large protein repository.

Keywords Computational drug discovery · Pattern classification and recognition · Protein folding and fold recognition · Proteomics

50.1 Introduction

Over the past 10 years, the number of three-dimensional (3-D) protein structures has grown exponentially [1]. This is due mainly to the advent of high throughput systems. Consequently, molecular biologists need systems to effectively store,

E. Paquet (✉)
Visual Information Technology, National Research Council, 1200 Montreal Road, Ottawa, ON K1A 0R6, Canada
School of Information Technology and Engineering, University of Ottawa, 800 King Edward, Ottawa, ON K1N 6N5, Canada
e mail: eric.paquet@nrc cnrc.gc.ca

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 50,

manage and explore these vast repositories of 3D structures. They want to determine if an unknown structure is in fact a new one, if it has been subjected to a mutation and to which family it possibly belongs. Furthermore, they require the ability to find similar proteins in terms of functionalities. Importantly, they aim to find docking sites for drugs; the later being the binding of two proteins, the ligand and the receptor in order to form a stable complex. This similarity in functionality and specifically the task to find docking sites are related to the outer shape of the protein [2]. The outer shape (or envelope), in part, determines whether two proteins may have similar functionalities and may thus aid us to determine the location of protein binding sites. The previously introduced docking problem may be better understood from the perspective of drug design. Most diseases and drugs work on the same basic principle. When we become ill, a foreign protein docks itself on a healthy protein and modifies its functionality. Such a docking is possible if the two proteins have two sub-regions that are compatible in terms of 3D shape, a bit like two pieces of a puzzle. Drugs are designed to act in a similar way. Namely, a drug docks on the same active site and prevents the docking of foreign proteins that can potentially cause illness. Consider a protein that has shown to be successful in a prescription drug developed to treat a terminal illness. However, this protein has serious contra-indications and causes severe adverse effects in a certain sub-set of the population. Suppose a protein with similar structure and functionality, but without these serious adverse effects, may be found. The subsequent modification of the harmful drug has obvious benefits.

To address these requirements, we have designed and implemented a system which is able to index and search a very large database of proteins based on the 3D outer shape or envelope. By identifying proteins with similar envelopes, we are able to identify proteins that have similar functionalities. This implies that we are able to potentially find a non-toxic protein for the treatment of a given illness. Such computer-aided drug design has many advantages. Our system is very fast, (typically a second for a search in our case) and it provides an exhaustive screening of all proteins, thus potentially reducing the time spent in the laboratory, and the associated financial burden, by orienting the research towards more promising directions.

In recent years, a number of researchers have investigated finding similar 3-D protein structures, mainly using structure alignment [3]. Research includes the work of [4, 5], who use local approaches to calculate the similarities of protein structures, thus possibly accumulating error and potentially overlooking semantic information about the inter-relationships of the structures. Other work includes shape-based approaches such as [6 8] which typically employ a sphere, grid, pie or spherical trace transform to compare structures. Our approach differs from the above-mentioned work, in which we perform an exhaustive search in real time considering all the structures (60,000 to date) as contained in the Protein Data Bank. Furthermore, our indexes or descriptors are optimised for surface representation, as opposed to most approach in which a volumetric representation of the protein is required. Also, as will be shown, our method may be generalised to address the docking problem.

50.2 Description of the System and Calculation of the Indexes

Our system consists of three main components. The first component constitutes the calculation of the outer surface (or envelope) from a segmentation based on the position of the constituent atoms and their electronic density distribution [2]. The second component computes the index which provides a complete description of the 3D shape of the envelope. The last component is the search engine which, from a given protein structure, finds the most similar proteins based on a given metric. In the current implementation, the normalised Euclidian distance is used. Our objective is to define an index that describes a protein from a 3D shape point of view and that is translation, scale and rotation invariant [9]. The algorithm may be described as follows. Firstly, the protein envelope is triangulated into a mesh. Next, the centre of mass of the protein structure is calculated and the coordinates of its vertices are normalised relatively to the position of its centre of mass. A translation invariant representation is then achieved. Next, the tensor of inertia of the protein envelope is calculated. To address the tessellation in the computation of these quantities, we do not utilise the vertices as is, but the centres of mass of the corresponding triangles; the so-called tri-centres. In all subsequent calculations, the coordinates of each tri-centre are weighted with the area of their corresponding triangle. The latter is being normalised by the total area of the envelope. In this way, the calculation may be made robust against tessellation.

The Eigen vectors of the tensor of inertia are calculated, to achieve rotation invariance. The Jacobi method, which has been proven successful for real symmetric matrices, is used. Once normalised, the unit vectors define a unique reference frame, which is independent on the pose and the scale of the corresponding object: the so-called Eigen frame. The unit vectors are identified by their corresponding Eigen values. The descriptor is based on the concept of a cord. A cord is a vector that originates from the centre of mass of the object and that terminates on a given tri-centre. The coordinates of the cords are calculated in the Eigen reference frame in cosine coordinates. The cords are weighted in terms of the area of the corresponding triangles; the later being normalised in terms of the total area of the protein. The statistical distribution of the cords is described in terms of three histograms: one histogram for the radial distribution and two for the angular distribution of the cords. The three histograms constitute the shape index of the corresponding protein envelope. As stated earlier, the shape indexes are placed in a protein database, which are then accessed by means of our search engine. We use the normalised Euclidian distance to calculate the similarity between proteins [9].

50.3 Experimental Results: Searching for the Most Similar Protein Envelopes

This section describes the experimental evaluation of our protein indexing and retrieval system. We implemented the system using Java and ran the experiments on workstations with two dual core processors and 4 GB of memory. Our data

comes from the Protein Data Bank or PDB [1] which contain all known 3D protein structures. There are currently (21 September 2009) about 60,000 proteins in the database, with a resultant database size of around 1 Terabyte. To validate our results, we have used the Structural Classification of Proteins or SCOP Database [10] which provides a grouping of proteins in terms of family which present, in general, very similar shapes. The SCOP classification is based on the amino acid sequences. (It has been experimentally proven that the later is strongly correlated with the 3D shape of the protein [11]). These proteins have been classified both by experts and by automated approaches based on amino acid sequences. One example is presented for illustration: phage T4 lysozyme from bacteriophage T4 (entry 142l in the PDB database). Figure 50.1 shows the results from 1 80. All results belong to the same family. The distance in between the reference envelope and the obtained results is calculated for each one of them. As noted above, this distance is defined as the normalised Euclidian distance in between the triplet of histograms associated with the reference envelope and the triplet of histograms associated with a given result; a 120 dimensional space. Similarity, for retrievals 81 to 160, only two results do not belong to the same family as 142l. These exceptions are the entries 1lf3 and 1lq8 which belong to the plasmepsin (a haemoglobin-degrading enzyme) from plasmodium falciparum, plasmepsin II and protein C inhibitor from human (homo sapiens), respectively. The search is exhaustive and takes less than 1 s on a laptop computer. (This is because the descriptor has only 120 bytes which makes the search very efficient.). Typically, when searching the PDB, we obtain a minimum of 90% precision for a recall of 90%, as based on family membership; from

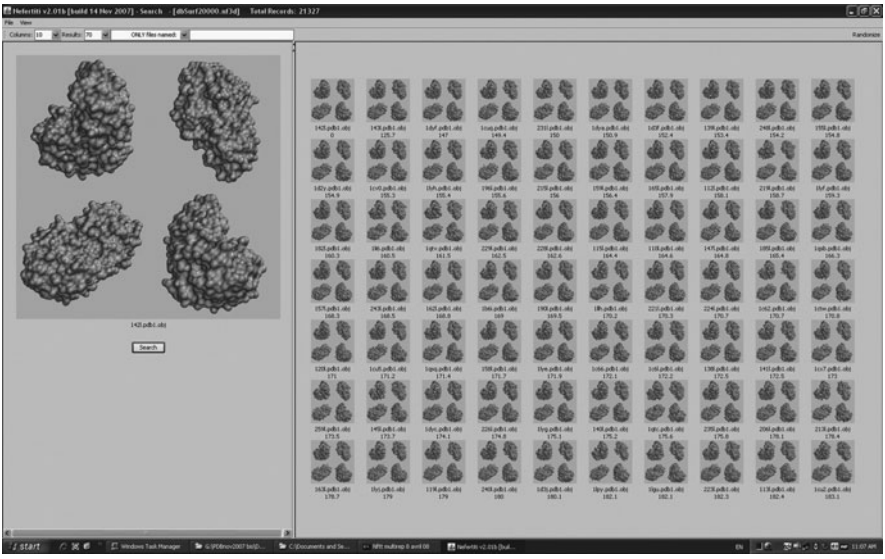


Fig. 50.1 Results 1 80 of a 3D similarity search for the envelope of protein 142l. All results belong to the same family

the SCOP point of view [9, 12, 13]. This indicates that the system is stable and reliable. What are the implications, then, for the docking problem? We have shown that our system is able to locate proteins with similar shape and have verified that these proteins do, indeed, belong to the same SCOP family. Recall that, for drug design, we may wish to replace a toxic protein A, which binds on a healthy one, with another non-toxic protein B. Here, A and B will have similar envelopes and thus similar docking sites, which our system is able to find. Our system is very efficient considering that the query is made on the large database of 1 Tbyte, with a very high precision-recall and a fast processing time.

50.4 Fragmentation Approach for the Docking Problem

In the previous sections, we described a system for the retrieval of similar protein based on the 3D shape of their entire outer envelope. The docking problem is, however, much more complex. Here, one has to find two proteins that have compatible regions, from the 3D point of view [14]. The docking problem is also complicated by the fact that proteins are not rigid entities. Rather, they are able to deform and interact locally, to adapt their contact zone to fit better [15]. This deformation and interaction issue may be further addressed by considering various configurations of the same protein. Such configurations are available from the Protein Data Bank, for a limited number of proteins [15]. Using such different configurations, thus provide us with valuable information about the various geometrical degrees of freedom of the associated protein. Currently, most approaches used to find docking regions are based on some kind of correlation based on Fourier transform and spherical harmonics among others [16]. In view of the application, these approaches present some drawbacks. The first one is that they require a volumetric representation to perform the correlation. If this would not be the case, there would be no correlation signal unless the match is absolutely perfect. Since the surface representation is bidimensional, a thickness has to be artificially introduced, to make the calculation possible. Correlations are also very sensitive which means that they provide results more in terms of exact matching than in terms of compatible matching; as we saw earlier, this is not what is required. Finally, the amount of calculation is prohibitive unless supercomputers or extensive grids are available 16.

We present early results of the extension of our algorithm to address the docking problem. The system is similar to the one described before, in the sense that the same indexing and retrieval components are used. To adapt our algorithms to the docking problem, we introduce an additional layer of pre-processing: each protein is broken into fragments, and their size and location on the surface of the protein is being determined randomly. The fragmentation is performed as follows. We select, randomly, a vertex on the surface, and we extract all the connected vertices within a pre-determined distance from the pivot vertex. The later may be either Euclidian or geodesic. The Euclidian distance is more robust to topological accidents, e.g. holes,

while the geodesic distance follows the geometry of the problem more naturally [17]. Evidently, geometrical obstructions are not taken into account by the present approach; two regions might be compatible locally without being accessible globally. Nevertheless, the obstruction problem needs only to be addressed for compatible regions which, in general, are very few. This may be accomplished, for instance, with collision detection techniques. The fragments are then indexed and searched with the algorithms described in Sect. 50.3. Given a receptor, our system locates the most compatible ligand in the fragment database.

An example of a query for the proposed approach is shown in Fig. 50.2. For this experiment, the fragment database consists of approximately 70,000 fragments generated from seven conformations of the Uracil-DNA glycosylase inhibitor protein from Bacteriophage pbs2, iugi. Each conformation is broken into 10,000 fragments, which are then indexed and searched with the system described in Sects. 50.2 and 50.3. The objective is to retrieve, given a region in a particular conformation, the corresponding regions in the other six conformations. The corresponding regions should be relatively similar, considering that they represent diverse conformations of the same protein. In this way, we can simulate, given a receptor, the search for the possible ligands in a database of 70,000 fragments. The outcome of such a query is shown in Fig. 50.2. The left image shows four views of the chosen receptor. The left images show, respectively, four views of the original receptor, which correspond to the closest match being identical, and then four views of the same region, but belonging to another conformation. One should notice that, as in the previous section, the search is performed in real time. Direct verification shows that the receptor and the ligands belong to the same region of the protein and

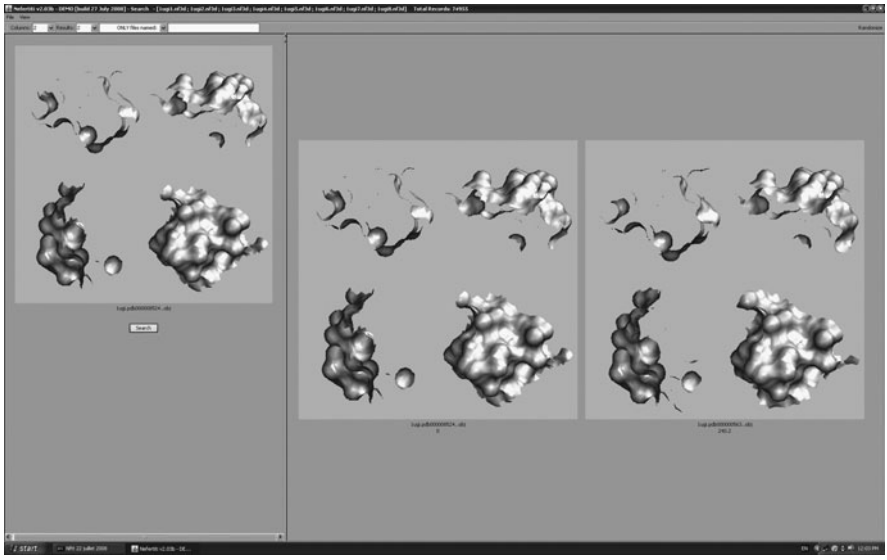


Fig. 50.2 Preliminary results for the docking problem based on the fragmentation approach

that, consequently, the system has been able to find the best matches within a database of 70,000 fragments.

50.5 Conclusions

We have shown that it is possible to search a very large database of protein envelopes and to retrieve, based on 3D envelope, the most similar structures using a fast exhaustive query. The indexation of the envelopes in entirely automatic and only the envelope of the reference protein is needed for the query. All proteins may be accessed and visualised directly in 3D and the results and the indexes can be exported to another analysis package as a flat file (common representation to exchange data). This work constitutes a first step towards computer-aided design of drugs. The current implementation has many applications in structural proteomics and may be used to facilitate the design of drugs, due to the strong correlation in between local similarity and global similarity for many proteins. That is, we have shown that it is possible to find proteins which have very similar envelopes in the PDB database. In addition, by adding a pre-processing step of fragmentation, it is possible to use the same framework to find similar sub-regions between two docking proteins.

References

1. Berman, H. M. et al. (2000) The Protein Data Bank. Nucleic Acids Research 28:235 242
2. Humphrey, W. et al. (1996) VMD visual molecular dynamics. Journal of Molecular Graphics 14:33 38
3. Akbar, S., Kung, J., Wagner, R. (2006) Exploiting Geometrical Properties of Protein Similar ity Search. Proceeding of the 17th International Conference on Database and Expert Systems Applications (DEXA'06), Krakow, Poland, pp. 228 234
4. Huang, Z. et al. (2006) 3D Protein Structure Matching by Patch Signatures. DEXA 2006, LNCS 4080, Springer, Berlin, pp. 528 537
5. Park, S. H., Park, S. J., Park, S.H. (2005) A protein structure retrieval system using 3D edge histogram. Key Engineering Materials 324 330
6. Abeysinghe, S., Tao, J., Baker, M. L., Wah, C. (2008) Shape modeling and matching in identifying 3D protein structures. Computer Aided Design 40 (6):708 720
7. Ying, Z.; Kaixing, Z., Yuankui, M. (2008) 3D Protein Structure Similarity Comparison Using a Shape Distribution Method. 5th International Conference on Information Technology and Applications in Biomedicine in Conjunction with 2nd International Symposium & Summer School on Biomedical and Health Engineering, Shenzhen, China, pp. 233 236
8. Zaki, M. J., Bystroff, C. (2008) Protein Structure Prediction. Humana Press, Totowa, NJ
9. Paquet, E., Viktor, H.L. (2007) CAPRI Content based Analysis of Protein Structure for Retrieval and Indexing. VLDB 2007 Workshop on Bioinformatics, Vienna: Austria, VLDB Press, p. 10
10. Andreeva, A. et al. (2008) Data growth and its impact on the SCOP database: New develop ments. Nucleic Acids Research 36:D419 D425

11. Guerra, C., Istrail, S. (eds.) (2003) Mathematical Methods for Protein Structure Analysis and Design. Springer, New York
12. Paquet, E., Viktor, H.L. (2008) CAPRI/MR: Exploring Protein Databases from a Structural and Physicochemical Point of View. Very Large Data Bases, Auckland, New Zealand, pp. 1504 1507
13. Paquet, E., Viktor, H.L. (2009) Finding Protein Family Similarities in Real Time through Multiple 3D and 2D Representations. Indexing and Exhaustive Searching, ISTICC, ACM SIGMIS International Conference on Knowledge Discovery and Information Retrieval KDIR, Madeira, Portugal, pp. 127 133
14. Ray, N. et al. (2005) Interf: Dynamic interface between proteins. Journal of Molecular Graphics and Modelling 23:347 354
15. Martin, C., Gherbi, R. (2006) Toward a user oriented immersive interaction for protein protein docking. The Mediterranean Journal of Computers and Networks 2(3):108 117
16. Nukada, A. et al. (2007) High Performance 3D Convolution for Protein Docking on IBM Blue Gene. International Symposium on Parallel and Distributed Processing and Applications ISPA 2007, LNCS 4742, Niagara Falls, Canada, pp. 958 969
17. Tenenbaum, T. et al. (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290:2319 2323

Chapter 51
Simultaneous Pathogen Detection and Antibiotic Resistance Characterization Using SNP-Based Multiplexed Oligonucleotide Ligation-PCR (MOL-PCR)

Jian Song, Po-E Li, Jason Gans, Momchilo Vuyisich, Alina Deshpande, Murray Wolinsky, and P. Scott White

Abstract Extensive use of antibiotics in both public health and animal husbandry has resulted in rapid emergence of antibiotic resistance in almost all human pathogens, including biothreat pathogens. Antibiotic resistance has thus become a major concern for both public health and national security. We developed multiplexed assays for rapid, simultaneous pathogen detection and characterization of ciprofloxacin and doxycycline resistance in *Bacillus anthracis, Yersinia pestis,* and *Francisella tularensis*. These assays are SNP-based and use Multiplexed Oligonucleotide Ligation-PCR (MOL-PCR). The MOL-PCR assay chemistry and MOLigo probe design process are presented. A web-based tool MOLigoDesigner (http://MOLigoDesigner.lanl.gov) was developed to facilitate the probe design. All probes were experimentally validated individually and in multiplexed assays, and minimal sets of multiplexed MOLigo probes were identified for simultaneous pathogen detection and antibiotic resistance characterization.

Keywords Antibiotic resistance · MOL-PCR · Multiplexed · Pathogen · SNP-based detection

51.1 Introduction

Extensive use of antibiotics in public health and animal husbandry have resulted in rapid emergence of antibiotic resistance in almost all human pathogens, many of which are resistant to multiple antibiotics. As most pathogenic bacteria become

J. Song (✉) and P.S. White (✉)
Biosecurity and Public Health (B 7), Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87544, USA
e mail: jian@lanl.gov; scott white@lanl.gov

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 51,

resistant to many commonly used antibiotics, physicians often need to try different antibiotics until the right antibiotic that works is found. However, using the wrong kind of antibiotics not only does not work, but can also aggravate the resistance problem. Exposure to one antibiotic can allow bacteria to become resistant to other antibiotics, leading to the emergence of multidrug resistance in the same bacteria. The best way to minimize the misuse and improper use of antibiotics in the treatment is to know whether sickness is due to a bacterial infection and to which antibiotics the causative pathogen is still susceptible. So the successful treatment of disease requires not only timely and accurate diagnosis, but also full characterization of the antibiotic resistance profiles of the infectious agents. This will require simultaneous pathogen detection and antibiotic resistance characterization. This is particularly true with threat pathogens *Bacillus anthracis*, *Yersinia pestis*, and *Francisella tularensis* which cause acute, often fatal diseases and thus prompt diagnosis and antibiotic treatment is of critical importance for the infected individuals to survive [1].

To successfully accomplish such a difficult task, we will need to develop rapid, accurate, and robust surveillance systems that are not only capable of detecting these pathogens, but also able to characterize their antibiotic resistance profiles well before clinical symptoms are observed in the affected population. Most of the molecular diagnostic tests that have been developed so far are either for detection of the pathogens or for detection of a few resistance markers, but not both, and are not high-throughput [1 3]. Both capabilities will be essential for a surveillance system and in response to any bioterrorist attacks and disease pandemics. Investigators at Los Alamos National Laboratory have developed a powerful new type of assay called Multiplexed Oligonucleotide Ligation-PCR (MOL-PCR) [4] and readout is performed on a flow cytometry platform (e.g., Luminex) (Fig. 51.1) It is ideally suited for high-throughput analysis and simultaneous detection and antibiotic resistance profile determination of multiple pathogens. Here, we report the development of multiplexed assays for rapid pathogen detection and characterization of ciprofloxacin and doxycycline resistance in *B. anthracis, Y. pestis*, *F. tularensis*.

51.2 Identification of Antibiotic Resistance Determinants

A detailed understanding of the genetic determinants for antibiotic resistance is essential for designing a complete set of molecular signatures for detection and characterization of antibiotic resistance in the target pathogens. A thorough literature review on ciprofloxacin and doxycycline resistance was performed for the three target pathogens and all known resistance determinants (*data not shown*) were identified for signature design.

51.2.1 Ciprofloxacin Resistance

Ciprofloxacin is a fluoroquinolone antibiotic that inhibits bacterial type II topoisomerases, DNA gyrase and topoisomerase IV and thus blocks DNA replication

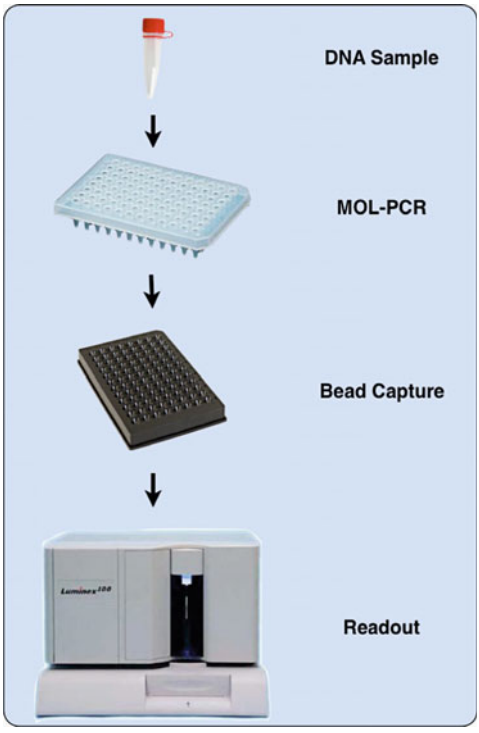


Fig. 51.1 MOL PCR assay process. Samples are analyzed for up to 100 markers each, providing an equivalent of 100 assays per sample. On a 96 well plate, an equivalent of 9,600 assays can be obtained in one run of MOL PCR, all of which can be performed within 6 h

[5]. Gyrase, an A2B2 complex encoded by the *gyrA* and *gyrB* genes, catalyzes ATP-dependent negative supercoiling of DNA and is involved in DNA replication, recombination, and transcription. Topoisomerase IV, a C2E2 complex encoded by the *parC* and *parE* genes, is essential for chromosome partitioning [6]. Resistance to fluoroquinolones is due to specific chromosomal mutations that occur in a discrete 39aa region called the quinolone resistance determinant region (QRDR) in these four genes. These mutations often occur in a stepwise fashion with each additional mutation leading to a higher level of resistance.

51.2.2 Doxycycline Resistance

Doxycycline is one of the semisynthetic tetracyclines that inhibit bacterial growth by stopping protein biosynthesis. Tetracyclines have been used for the past 50 years as one of the major antibiotics in treating diseases in both humans and animals. They have also been widely used as additives to livestock feed to stimulate weight

gain in some domestic animals and to improve the health and promote the growth of fish in commercial fisheries. As a result, tetracycline resistance is becoming widespread. An excellent web resource on tetracycline resistance genes is provided by Dr. Marilyn Roberts (http://faculty.washington.edu/marilynr).

51.3 MOL-PCR Assay and MOLigo Probe Design

51.3.1 MOL-PCR Assay

The MOL-PCR was developed based on a novel assay chemistry (Fig. 51.2) and uses a flow cytometry detection platform for readout. It allows rapid, multiplexed detection of target DNA and SNPs. The versatility of this assay allows simultaneous interrogation of a variety of molecular variations such as indels, SNPs, and the presence/absence of specific signatures through multiplexing of various specific probes.

The MOL-PCR can accomplish a higher level of multiplexing more easily than RT-PCR because it employs a pre-PCR ligation step that accomplishes target detection. This step is then followed by a singleplex PCR with a pair of universal primers that will amplify all the resulting products from the ligation step. The PCR step uses a fluorescently labeled universal primer that will allow the detection of the ligated product. The PCR products are then hybridized to microspheres and

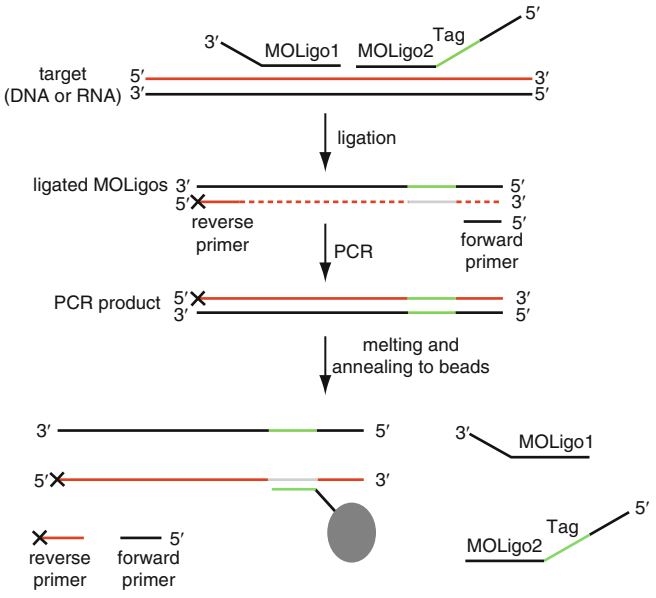


Fig. 51.2 A depiction of MOL PCR assay

analyzed by flow cytometry. Up to 100 genetic markers can be analyzed in each assay. Because it uses 96-well plates, an equivalent of 9,600 assays can be obtained in a single run of MOL-PCR, which can be performed in <6 h. Parallel processing and the use of 384-well plates will drive throughput even higher.

The MOL-PCR requires two oligonucleotide probes, designated as MOLigo1 and MOLigo2. MOLigo1 contains two components: the sequence that is complementary to the target sequence and the sequence of a universal reverse primer (RPS). MOLigo2 contains three modular components: a sequence complementary to a universal forward primer (FPCS), a tag unique to the MOLigo probe (also allowing detection of a SNP base) that will allow capture of the ligated product onto a microsphere, and the last component is the sequence that is complementary to the target sequence (specific region). Different tags can be used on MOLigo2 probes for different SNP bases to allow allelic discrimination. MOLigo1 and MOLigo2 are annealed to the target sequence around the SNP base when the target is present. If the matching target, including the SNP base, DNA ligase will covalently link MOLigo1 and MOLigo2 together creating a MOLigo1+MOLigo2 complex. The ligation will occur if the terminal bases of MOLigo1 and MOLigo2 at the junction site are completely complementary to the target sequence.

After ligation, all ligated products undergo singleplex PCR with a pair of universal PCR primers. The RPSr is also tagged with a fluorescent dye on the 5′ end (depicted as a "x"), which allows for the detection of the ligated product. The amplified products are then hybridized to Luminex xTAG® microsphere array via the tags that are covalently linked to the microspheres (beads). A microsphere will show a fluorescence signal only if it is linked to a labeled RPS on MOLigo1, which can only occur if ligation was successful. A more detailed description of the MOL-PCR process can be found in [4].

51.3.2 MOLigo Probe Design

To fully realize all the advantages and detection potential of MOL-PCR, it is necessary to develop a robust design process that is capable of producing a multiplexed set of specific robust signatures. The multiplex MOLigo probe set is capable of both accurately detecting the pathogen, and fully characterizing all the known antibiotic resistance determinants for the target pathogen. Because of its novel chemistry, MOL-PCR presents several challenges: (1) designing SNP-specific MOLigo probes the probes have to be designed around the targeted SNPs, which will reduce the sequence space where MOLigo probes can be designed; (2) using capture tags and universal primers in the MOLigo probes will increase the chance of secondary structure formation; (3) multiple target SNPs close to each other the MOLigo probes for those SNPs may interfere with each other by competing for the same target site if they are combined in a multiplexed reaction; and (4) multiplexing a large set of long MOLigo probes longer MOLigos are more likely to form heterodimers. To address these challenges, we have developed an

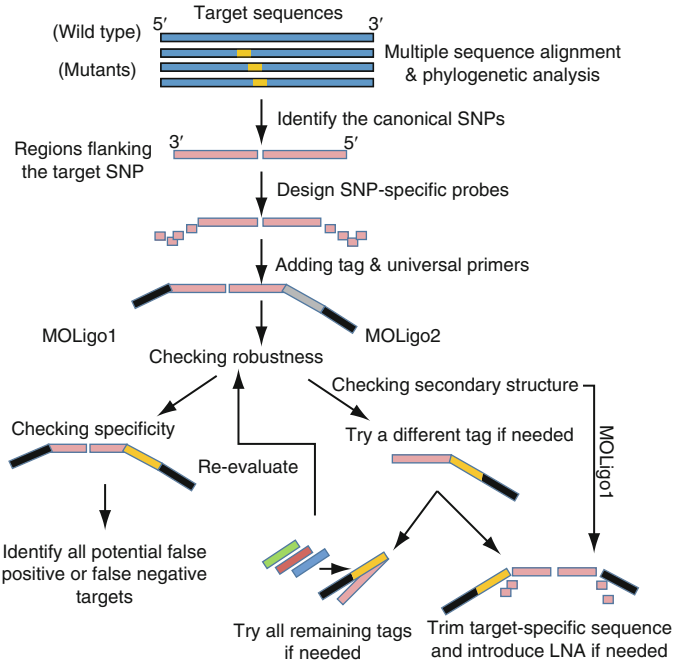


Fig. 51.3 MOLigo probe design process

integrated MOL-PCR probe design pipeline. The major steps of this pipeline are shown in Fig. 51.3 and are described in detail elsewhere (http://MOLigoDesigner.lanl.gov/).

51.3.3 Development of the MOLigoDesigner

A web-based MOLigo design tool, MOLigoDesigner (http://MOLigoDesigner.lanl.gov), was developed to facilitate the MOLigo design. It integrated the several major steps in our MOL-PCR probe design pipeline and allows users to design MOLigo probes for a single target SNP and perform quality check of all designed MOLigo probes.

51.4 Experimental Validation

Experimental evaluation of all MOLigo probe sets for specificity included testing them individually (not multiplexed) against synthetic targets. The sensitivity was tested using a range of target copy number (10^1 10^6) to determine the detection limit. A signal to noise ratio of 4 was used as a cut off for identifying "positive"

signals in the initial screen. Further testing against wild-type pathogen DNAs, as well as near neighbor and diversity panels provided information about the performance of each probe set, and pointed out any potential interference that could lead to false positive and/or false negative assay results. Any probes that did not perform satisfactorily were redesigned and retested. After screening all MOLigo probe sets individually, they were combined into two multiplexed assays to test the specificity and multiplex capability.

51.4.1 Pathogen Detection

To validate the specificity of "pathogen detection" assay, the MOLigo probes were tested against three validation panels, including one panel of near neighbors that are close phylogenetic relatives of each of the three pathogens, and two diversity panels that included taxa that were more distantly related to the target pathogens. They were successfully tested for specificity against all three different validation panels.

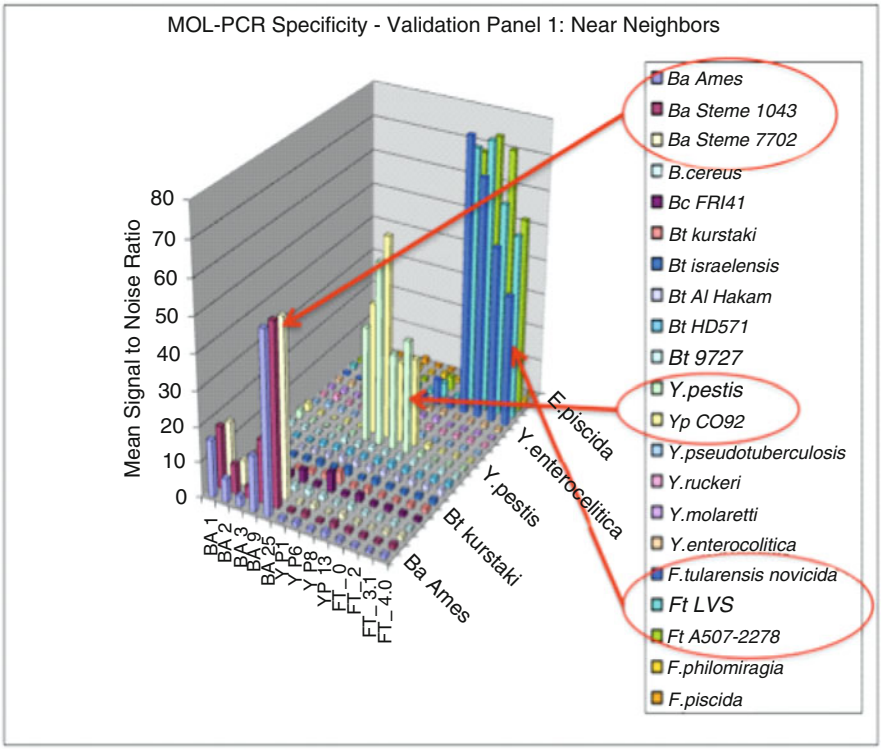


Fig. 51.4 Testing the specificity of the pathogen detection MOLigo probes against the near neighbors

The results from validation against the "Near Neighbors" panel are shown in Fig. 51.4.

51.4.2 Characterization of Antibiotic Resistance

The minimal set was selected to allow a better chance of multiplexing. In the minimal set, one MOLigo probe pair may be able to detect multiple resistant strains if these strains share a common resistance determinant, even though they may have

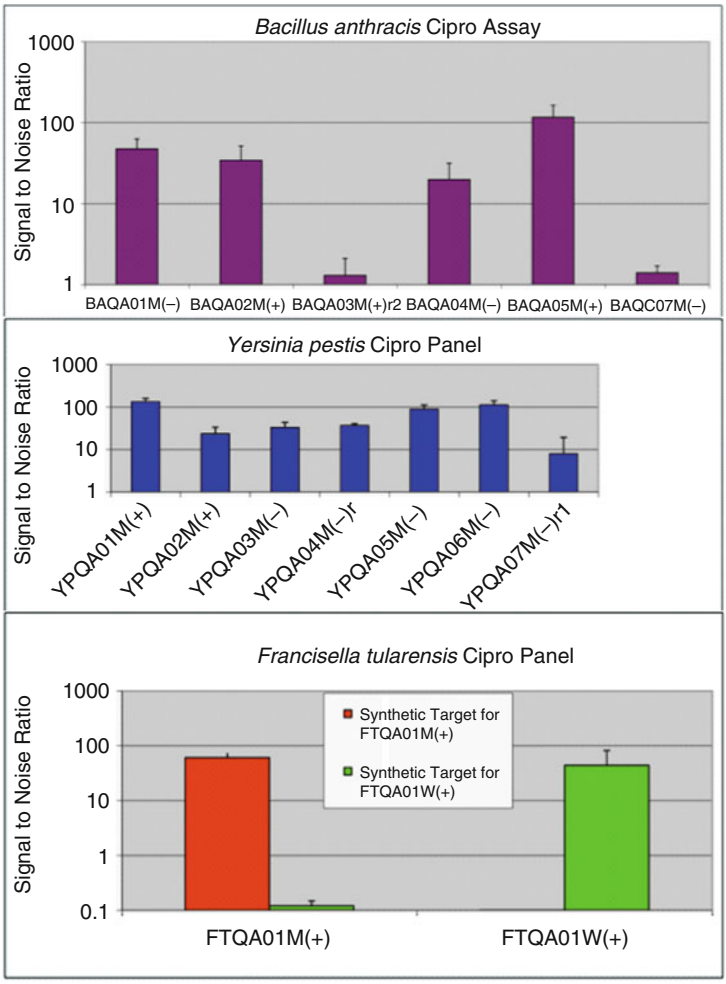


Fig. 51.5 Cipro resistance probes validation for B. anthracis (*top*), Y. pestis (*middle*), and F. tularensis (*bottom*). Results represent triplicate experiments against synthetic targets. Signal to noise was calculated by using the wild type sequences

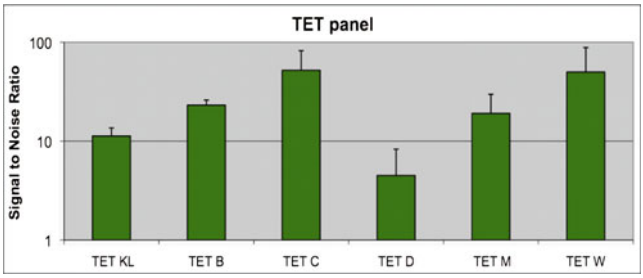


Fig. 51.6 Tetracycline resistance probes validation. Results represent triplicate experiments against synthetic targets. Signal to noise values were calculated by using the wildtype sequences

many other additional resistance determinants. To detect a particular resistant strain and possibly understand the level of the resistance (particularly for ciprofloxacin resistance), all contributing resistance determinants need to be characterized. The minimal set of MOLigo probes for each pathogen was screened for specificity and multiplex capability (Fig. 51.5). A few probes did not perform well when multiplexed as a part of the minimal set. After further evaluation, we concluded that those probes failed in multiplexed assays overlap in their gene-specific sequences, which led to interference with each other. This type of interference will happen any time multiple target SNPs are closely situated, which is the case for SNPs that confer ciprofloxacin resistance.

For tetracycline resistance detection, we designed a minimal set of 6 MOLigo probe pairs that cover 7 tetracycline resistant gene families. They all performed well when multiplexed in the minimal set (Fig. 51.6).

51.5 Discussion

We have designed and tested MOL-PCR probes for both pathogen detection and ciprofloxacin (fluoroquinolone) and doxycycline (tetracycline) resistance in *B. anthracis*, *Y. pestis*, and *F. tularensis*. All probes performed well when tested individually against synthetic targets. When assayed in multiplexed minimal sets, some of the probes failed but were redesigned. We have successfully identified all the probes needed for detection of the target pathogen and characterization of all the known resistance determinants in three threat pathogens against synthetic targets. But these probes have to be tested using clinical samples or clinical samples spiked with the target pathogens before the assay can be used for pathogen detection and surveillance. Multiplexing is still an issue. Several probes performed well when assayed individually, but failed when multiplexed with other probes. This may be an inherent limitation for SNP-based assays when multiple SNPs are too close to each other, and the probes will likely interfere and compete against each other for the same binding site. In fact, some MOLigo probes could conceivably serve as

target for other probes in multiplexed assays. But given the high throughput capability, many probes that could not be multiplexed can be assayed separately on the same 96-well plate. This large assay capacity also leaves room for many more assays to be added as new targets are identified.

Acknowledgment This work was supported by the Department of Homeland Security. Los Alamos National Laboratory is operated by Los Alamos National Security, LLC for the U.S. Department of Energy's NNSA.

References

1. Lindler, L.E. and W. Fan (2003) Development of a 5' nuclease assay to detect ciprofloxacin resistant isolates of the biowarfare agent Yersinia pestis. Mol Cell Probes 17(1):41 47.
2. Woodford, N. and A. Sundsfjord (2005) Molecular detection of antibiotic resistance: when and where? J Antimicrob Chemother 56(2):259 261.
3. Perreten, V., L. Vorlet Fawer, et al. (2005l) Microarray based detection of 90 antibiotic resistance genes of gram positive bacteria. J Clin Microbiol 43(5):2291 2302.
4. Mark, J., L. Green, A. Deshpande, and P. White. (2007). System integration and development for biological warfare agent surveillance. Proceedings of SPIE, Optics and Photonics in Global Homeland Security III 6540:65401D.
5. Bast, D.J. and J.C. de Azavedo (2001l) Quinolone resistance: older concepts and newer developments. Curr Infect Dis Rep 3(1):20 28.
6. Gonzalez, I., M. Georgiou, et al. (1998) Fluoroquinolone resistance mutations in the parC, parE, and gyrA genes of clinical isolates of viridans group streptococci. Antimicrob Agents Chemother 42(11):2792 8.

Chapter 52
Specification and Verification of Pharmacokinetic Models

YoungMin Kwon and Eunhee Kim

Abstract A model checking technique to specify and verify temporal properties of drug disposition changes is proposed. In pharmacokinetics and pharmaceutics, drug kinetics is often modeled as single or multiple compartment models. In this paper, a probabilistic temporal logic, called *iLTL*, is introduced to specify many interesting properties of drug kinetics. Given a specification, a computerized technique, called *model checking* [1], is used to check whether all drug disposition changes of a compartment model comply with the specification.

Keywords Computer-based medical systems · Computational systems biology

52.1 Introduction

Compartment models [5] have long been used as a mathematical model to describe the drug concentration level changes in our bodies. These models help us understand the relationship between drugs and their clinical effects. Thus, many compartment models exist for many types of drugs. However, compared to their importance, systematic evaluation methods on them are not well developed, most notably, they are manually examined by drawing graphs.

To address this problem, we propose to use a computerized systematic evaluation method based on iLTL model checking [2]. Specifically, the model checker searches for a trajectory of drug disposition changes that would violate the

Y. Kwon (✉)
Microsoft Corp. Redmond, Redmond, WA, USA
e mail: ykwon4@cs.uiuc.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 52,

specification. Since the search completely explores every possible combination of doses, one can determine the existence of a satisfiable dose. Also, by checking the negated specification, a desirable dose can be found as a counterexample.

In the compartment model, the amount of drug leaving from one compartment to another is proportional to the amount of drug present in the first compartment. This relation makes the *memoryless* property: future drug dispositions will depend only on its current disposition. Because of this memoryless property, we can transform the compartment models to *Continuous Time Markov Chains* (CTMCs), and convert them again to *Discrete Time Markov Chains* (DTMCs) [3, 7], which are the formal models of iLTL. After a slight modification, the DTMCs can describe the changes of physical quantities instead of probabilities.

Throughout this paper, we explain our techniques with a three compartment model of insulin [6]. However, application to other compartment models should be straightforward.

52.2 Model

Compartment models are composed of one or more compartments that represent a group of tissues with similar blood flow and drug affinity and drug transition rates between the compartments [5]. As the compartment models have the *memoryless property*, they can be naturally converted to Markov processes. The conversion steps to CTMCs are:

- The states of the CTMC are the compartments and a fresh sink state for the cleared drug.
- The transition probability rates between states are (1) the fractional turnover rates between the corresponding compartments and (2) the fractional drug elimination rates from the corresponding compartment to the fresh sink state.

With this representation, the probability that a CTMC in a certain state is equal to the fraction of the drug in the corresponding compartment. Figure 52.1 shows a CTMC model for a compartment model of insulin-^{131}I. This compartment model is obtained by extending the three compartment model of Silvers et al. [6]. This CTMC has a set of states $\{Pl, IF, Ut, Cl, FD, SD, Re\}$. Among the seven states, *Pl*, *IF*, and *Ut* states and the transition rates between them are from the original compartment model. *Cl* state is the fresh sink state. To compute the dose later in the example section, we extended the model with two more compartments: *FD* for unabsorbed fast acting drug and *SD* for unabsorbed slow acting drug. We choose the rates from *FD* and *SD* to *Pl* such that the drug concentration at *Ut* reaches its maximum at about 2.5 and 4.5 h, respectively. *Re* state is introduced to make the specification in physical units instead of probability.

Given a CTMC C, one can compute a DTMC D whose *probability mass function* (pmf) changes are equal to the sampled pmfs of C. Let $\mathbf{R} \in \Re^{n \times n}$ be an infinitesimal generator matrix with $\mathbf{R}_{i,j}$ being the rate from state s_j to state s_i. Using a probability

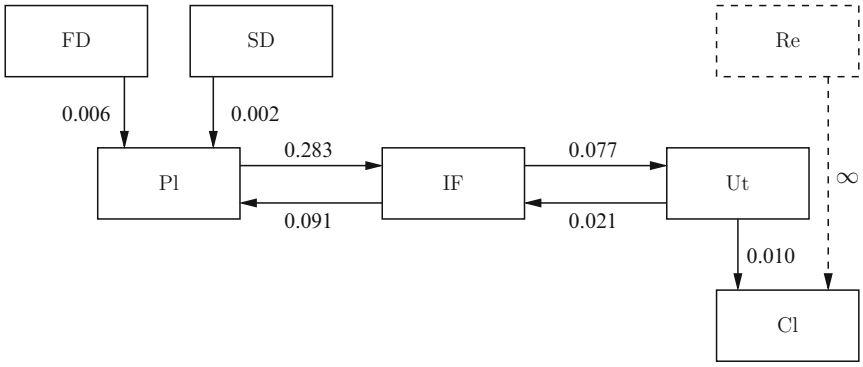


Fig. 52.1 CTMC model of an insulin compartment model. The boxes are the states and the numbers are the transition probability rates. The states represent: *FD* unabsorbed fast acting drug, *SD* unabsorbed slow acting drug, *Cl* cleared drug, *Pl* plasma, *IF* interstitial fluid, *Ut* site of utilization and degradation. *Re* is an additional state to make the specification in physical units

vector function $\mathbf{x}$: $\Re \rightarrow \Re^n$ with $\mathbf{x}(t)_i = \mathrm{P}[C(t) = s_i]$, we can simply write $\mathbf{x}(t) = \mathrm{e}^{\mathbf{R}\cdot t}\cdot\mathbf{x}(0)$. Periodically sampling C with a period T results in a DTMC whose probability transition matrix is $\mathbf{M} = \mathrm{e}^{\mathbf{R}\cdot T}$. Let a probability vector function $\mathbf{y}$: $\aleph \rightarrow \Re^n$ be $\mathbf{y}(k)_i = \mathrm{P}[D(k) = s_i]$, and let $t = k{\cdot}T$, then $\mathbf{x}(t) = \mathrm{e}^{\mathbf{R}\cdot t}\cdot\mathbf{x}(0) = \mathbf{M}^k\cdot\mathbf{y}(0) = \mathbf{y}(k)$. Observe that $\mathbf{M}_{i,j} = \mathrm{P}[D(k+1) = s_i | D(k) = s_j]$, and D satisfies the Chapman-Kolmogrov equation: $\mathbf{y}(k+1) = \mathbf{M}\cdot\mathbf{y}(k)$.

Now, let us consider using physical units in the specification. The linearity of C and D plays a crucial role here. If we disregard the fact that $\mathbf{x}(t)$ and $\mathbf{y}(t)$ are pmfs, whose elements add up to one, and scale their initial pmfs, then their trailing pmfs would scale by the same amount because of the linearity. For example, if d (mg) of slow acting drug is administrated, then its initial state is $\mathbf{x}(0) = d{\cdot}[0, 0, 0, 0, 0, 1]^{\mathrm{T}}$ and $\mathbf{x}(t)_i$ is the fraction of the d (mg) of the drug in s_i at time t.

However, if we consider the fact that $\mathbf{x}(0)$ and $\mathbf{y}(0)$ are pmfs, then we cannot scale them arbitrarily: their sum must add up to one. To address this problem, we introduced the *Re* state of Fig. 52.1. *Re* state is instantly sinked to *Cl* state without interacting with other states. To introduce physical units in the specification, one can simply choose large units so that the physical amounts can be fit in the probability range [0, 1], and put the remaining probability in *Re* state. As an example, let us consider a combined dosage of 10 mg of intravenous injection, 20 mg of fast acting drug, and 30 mg of slow acting drug. This dosage is equivalent to 0.01, 0.02, and 0.03 g of the drugs, respectively, and the corresponding pmf is $\mathbf{x}(0) = [0.01, 0, 0, 0, 0.02, 0.03, 0.94]^{\mathrm{T}}$. Handling the infinite rate from *Re* state to *Cl* state could be problematic in the CTMC; however, it simply becomes one in the corresponding DTMC. The extended probability transition matrix of D that has *Re* state is $\mathbf{M}'_{i,j} = \mathbf{M}_{i,j}$ for $1 \leq i,j \leq 6$, $\mathbf{M}'_{4,7} = 1$, and $\mathbf{M}'_{i,j} = 0$ in other cases.

52.3 Logic

In this section, we briefly describe the syntax and an informal semantics of iLTL. For detailed description about the logic, please refer to [2].

The *syntax of iLTL* formula Ψ is as follows:

$$\begin{gathered}\Psi ::= \mathrm{T}|\mathrm{F}|ap|(\Psi)| \\ \sim \Psi|\Psi \wedge \Psi|\Psi \vee \Psi|\Psi \rightarrow \Psi|\Psi \leftrightarrow \Psi \\ \mathrm{X}\Psi|\mathrm{G}\Psi|\mathrm{F}\Psi|\Psi \,\mathrm{U}\, \Psi|\Psi \,\mathrm{R}\, \Psi\end{gathered}$$

An *atomic proposition* (ap) is an equality or an inequality about an *expected reward* [4] of a DTMC. Let $\{s_1,\ldots,s_n\}$ be the set of states of a DTMC D, then the atomic propositions are defined as follows:

$$ap ::= r_1 \cdot \mathrm{P}[D(t_1) = s_1] + \cdots + r_n \cdot \mathrm{P}[D(t_n) = s_n] \Diamond r,$$

where $t_i \in \aleph$ is a time offset, $r_i \in \Re$ is a reward associated with the state s_i, and $\Diamond$ is one of $<$, $\leq$, $=$, $\geq$, or $>$.

The meaning of ap is $r_1 \cdot \mathrm{P}[D(t_1) = s_1] + \cdots + r_n \cdot \mathrm{P}[D(t_n) = s_n] \Diamond r$ at time t is true *if and only if* (iff) $r_1 \cdot \mathrm{P}[D(t + t_1) = s_1] + \cdots + r_n \cdot \mathrm{P}[D(t + t_n) = s_n] \Diamond r$. The meaning of *logical operators* are as usual and the meaning of *temporal operators* are: X Ψ is true at t iff Ψ is true at $t + 1$, G Ψ is true at t iff Ψ is always true from t, and F Ψ is true at t iff eventually Ψ becomes true at some time $t_1 \geq t$. Ψ U Φ is true at t iff there is a time $t_1 \geq t$ when Φ is true and Ψ is true at t_2 for $t \leq t_2 < t_1$. Ψ R Φ is true at t iff Φ is true while Ψ is false from t and up to the moment when Ψ becomes true.

52.4 Examples

Finally, in this section, we demonstrate the usefulness of our specification and verification techniques through three drug administration examples. We use the three compartment model of Fig. 52.1 sampled at a 10 min interval. Throughout this section, we assumed that the body weight is 60 kg, the volume of display of the *Ut* compartment is 15.8% of the body weight.

As a first example, we compute a dose for an oral drug administration that could satisfy: (1) the onset time is no later than 1.5 h, (2) the active duration is at least 6 h, (3) the *Minimum Effective Concentration* (MEC) is 1.4 μg/ml, and (4) the *Minimum Toxic Concentration* (MTC) is 2.1 μg/ml. Based on these parameters, the mass of the drug in the *Ut* compartment at the MEC and at the MTC are $mem = 0.019908$ g and $mtm = 0.013272$ g, respectively.

Let us specify the desired onset time of 1.5 h. Because the sampling period is 10 min, the drug concentration level at *Ut* should be larger than the MEC at the ninth step. Using the time offset, this condition can be simply expressed as:

$$\Psi_{\text{onset}} : \mathrm{P}[D(9) = Ut] > mem.$$

The condition about the active duration can be specified similarly using the time offset. However, the 6 h duration and the 10 min sampling period require 37 different inequalities. We reduced the number of atomic propositions to 10 using the *next* operator X (we can reduce the number to the square root of the consecutive steps). The 10 atomic propositions are e_i: P[$D(i) = Ut$] > mem for $i = 0, 4, 8, \ldots, 36$. Let Ψ_{quarter} be $e_0 \wedge e_4 \wedge e_8 \wedge e_{12} \wedge e_{16} \wedge e_{20} \wedge e_{24} \wedge e_{28} \wedge e_{32}$, then the 6-h duration can be written as:

$$\Psi_{\text{dur}}: \Psi_{\text{quarter}} \wedge \text{X}\ \Psi_{\text{quarter}} \wedge \text{XX}\ \Psi_{\text{quarter}} \wedge \text{XXX}\ \Psi_{\text{quarter}} \wedge e_{36}.$$

Since this active duration will not start immediately after the drug is administrated, we wrote the specification as F Ψ_{dur}, meaning that the active duration should eventually occur.

The third condition is about the MTC: the concentration level at the *Ut* compartment should never exceed the MTC. This condition can be easily specified using the *always* operator G as follows:

$$\Psi_{\text{mtc}}: \text{G}\ (\text{P}[D = Ut] > mtm).$$

The drug administration options can be specified as a precondition about the initial condition. The oral drug administration option makes the precondition as Ψ_{ia}: P[$D = SD$] + P[$D = FD$] + P[$D = Re$] = 1. That is, all drugs are at these three states initially.

To sum up, a desirable dose can be found by model checking the combined specification:

$$\Psi_{\text{a}}: \Psi_{\text{ia}} \rightarrow\sim (\Psi_{\text{onset}} \wedge \text{F}\Psi_{\text{dur}} \wedge \Psi_{\text{mtc}}).$$

Observe that we negated the required conditions in the specification. Thus, any counterexample would satisfy $\Psi_{\text{ia}} \wedge \Psi_{\text{onset}} \wedge \text{F}\ \Psi_{\text{dur}} \wedge \Psi_{\text{mtc}}$.

Model checking Ψ_{a} showed that a combined dose of 47.845 mg of fast acting drug and 74.432 mg of slow acting drug could achieve the goal. Figure 52.2 shows the drug concentration level change for this dosage. The dashed line and the dotted line in this graph are the drug concentration due to the fast acting drug and the slow acting drug, respectively. The solid line is their combined effect. From the graph, we can check that all the three requirements are satisfied. We further discovered that the requirements cannot be satisfied by the fast acting drug alone or the slow acting drug alone.

As a second example, we compute a dose for the multidosage regimen. Specifically, we look for a repeatable state with 6-h period that could satisfy the MTC and the MEC conditions during the transition. Once such a state is reached, the drug concentration level can be maintained in the band between the MTC and the MEC by simply taking the same amount of the drug at 6-h intervals.

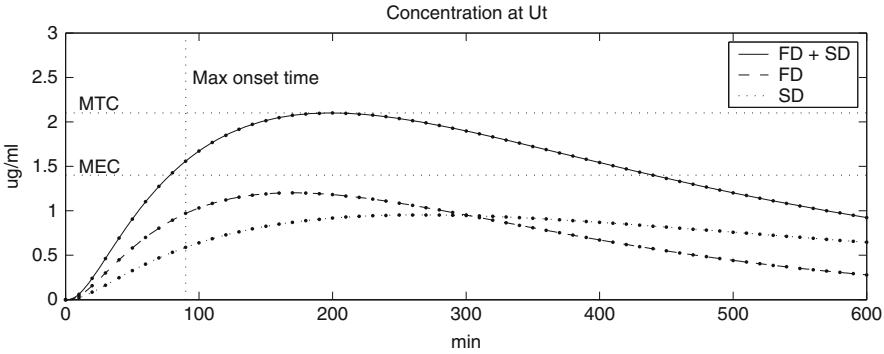


Fig. 52.2 Drug disposition changes for fast acting drug (*dashed line*), slow acting drug (*dotted line*), and their combined effect (*solid line*)

Among the seven states of the DTMC D, the *FD* and the *SD* states are our control variables, and the *Cl* and the *Re* states are for the noninteracting drugs. Thus, we want to have the *Pl*, the *IF*, and the *Ut* states repeated. One concern here is that $\lim_{t\to\infty} \mathrm{P}[D(t) = s] = \mathrm{P}[D(t + 36) = s]$ for $s \in \{Pl, IF, Ut\}$, which violates one of the *completeness condition* of iLTL model checking that the RHS of an atomic proposition is not equal to its LHS (to prevent the transient modes keep changing the truth value of atomic propositions). To avoid this problem, we replaced the equality with a small interval. Let a parameterized formula $\Psi^{\mathrm{s}}{}_{\mathrm{rep}}$ be $(\mathrm{P}[D = s] < \mathrm{P}[D(36) = s] + 10^{-9}) \wedge (\mathrm{P}[D = s] > \mathrm{P}[D(36) = s] - 10^{-9})$, then the specification can be written as:

$$\Psi_{\mathrm{b}} : \sim \left(\mathrm{G}\Psi_{\mathrm{mtc}} \wedge \Psi_{\mathrm{dur}} \wedge \Psi^{\mathrm{P1}}_{\mathrm{rep}} \wedge \Psi^{\mathrm{IF}}_{\mathrm{rep}} \wedge \Psi^{\mathrm{Ut}}_{\mathrm{rep}}\right).$$

Model checking the negated specification, Ψ_{b}, reported a pmf vector $[1.862, 4.877, 12.272, 888.540, 40.600, 50.850, 0] \cdot 10^{-3}$ as a counter example. Interpreting the pmf vector, if 1.862, 4.877, and 13.272 *mg* of the drug were in the *Pl*, the *IF*, and the *Ut* compartments respectively, then the same amount of drug will be found in these compartments 6 h later if 40.6 *mg* of fast acting drug and 50.85 *mg* of slow acting drug were in their unabsorbed states.

Figure 52.3 shows the concentration level change for this multidosage regimen. From 100 min, the 6-h cycle begins. The first graph shows the concentration level change of the *Ut* compartment, and the second graph shows the amount of unabsorbed fast acting drug (solid line) and slow acting drug (dashed line). The last two jumps in this graph are the required amount of drug to maintain the cycles. They are 35.918 *mg* of the fast acting drug and 26.099 *mg* of the slow acting drug. The first jump is different from the other two. We will explain the difference along with the first 100 min of the graph in the next example.

As a final example, let us find how to get to this repeatable state. In this example, we assume that the available drug administration options are the IV bolus, fast

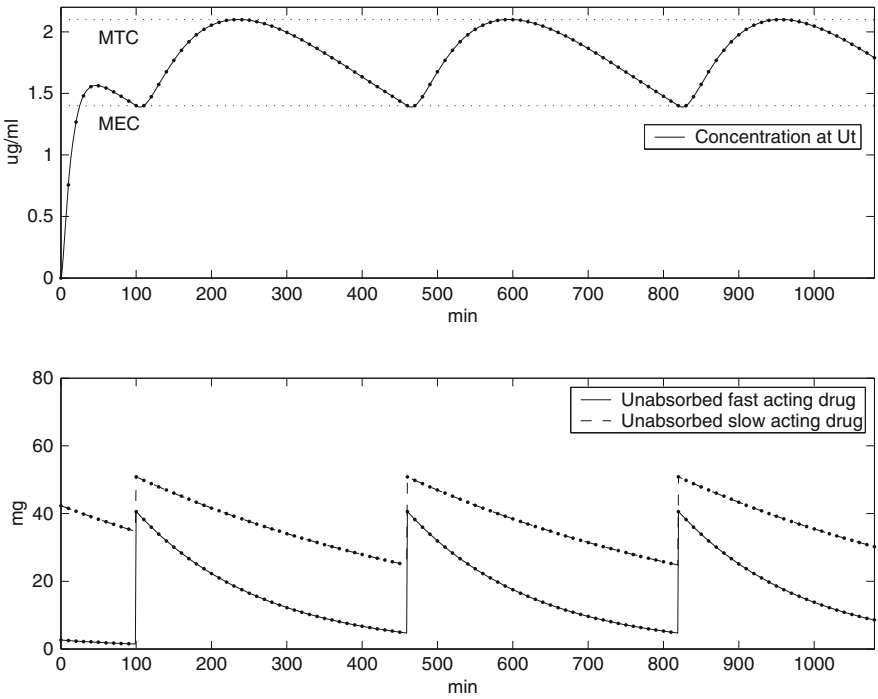


Fig. 52.3 The drug concentration level changes at the *Ut* compartment from a multidosage regimen (*top*), and the amount unabsorbed fast acting drug (*solid line*), and slow acting drug (*dotted line*) (*bottom*). The jumps are the required dose

acting drug, and slow acting drug. These available options make the initial condition Ψ_{ic}: $\text{P}[D = Pl] + \text{P}[D = SD] + \text{P}[D = FD] + \text{P}[D = Re] = 1$. The condition that the repeatable state is arrived at is:

$$\Psi_{\text{arr}}: \text{P}[D = Pl] = 1.862\text{e}^{\ -3} \wedge \text{P}[D = IF] = 4.877\text{e}^{\ -3} \wedge \text{P}[D = Ut] = 13.272\text{e}^{\ -3}.$$

The numbers in Ψ_{arr} are from the counterexample of the previous example. Because we want to have this condition occur eventually, we write the requirement as $\text{F}\ \Psi_{\text{arr}}$. Combining the conditions, the whole specification to check is as follows:

$$\Psi_{\text{c}}: \Psi_{\text{ic}} \rightarrow\sim (\text{G}\ \Psi_{\text{mtc}} \wedge \text{F}\ \Psi_{\text{arr}}).$$

Model checking Ψ_{c} showed that if 23.78 mg of the drug is intravenously injected and 2.624 mg of fast acting drug and 42.343 mg of slow acting drug are orally administered, then the repeatable state of the previous example will be reached. The initial part of the graphs in Fig. 52.3 confirms it. The formula Ψ_{arr} does not include any conditions about the *FD* or the *SD* states. Thus, when the repeating state is reached, the amount of drug in these states might be different from the other cycles.

Hence, the first jump in the second graph is different from the others. The required dose at 100 min is 39.160 mg of fast acting drug and 16.182 mg of slow acting drug.

52.5 Conclusions

In this paper, we demonstrated the usefulness of *iLTL* in specifying and verifying many interesting properties about drug kinetics. This computerized model checking technique not only proves certain drug disposition properties, but also computes a dose that could satisfy complicated requirements. The DTMC model of the logic can also be directly obtained from the compartment models after a simple conversion. We demonstrated all the steps from end to end using a three compartment model of insulin.

References

1. Clarke, E., Grumberg, O., Peled, D. Model checking. MIT Press, Cambridge, MA (2000)
2. Kwon, Y., Agha, G. Linear inequality LTL (iLTL): A model checker for discrete time markov chains. In: International conference on formal engineering methods, pp 194 208. LNCS 3308 (2004)
3. Papoulis, A. Probability, Random variables, and stochastic processes, 3rd edn. McGraw Hill (1991)
4. Russell, S., Norvig, P. Artificial intelligence a modern approach. Prentice Hall (1995)
5. Shargel, L., B.C., A. Applied biopharmaceutics and pharmacokinetics. Appleton Century Crofts (1985)
6. Silvers, A., Swenson, R.S., Farquhar, J.W., Reaven, G.M. Derivation of a three compartment model describing disappearance of plasma insulin 131i in man. The Journal of Clinical Investigation (1969)
7. Start, H., Woods, J.W. Probability and random processes with applications to signal processing, 3rd edn. Prentice Hall (2002)

Chapter 53
Dehydron Analysis: Quantifying the Effect of Hydrophobic Groups on the Strength and Stability of Hydrogen Bonds

Christopher M. Fraser, Ariel Fernández, and L. Ridgway Scott

Abstract In the past decade, research has demonstrated that defectively packed hydrogen bonds, or "dehydrons," play an important role in protein-ligand interactions and a host of other biochemical phenomena. These results are due in large part to the development of computational techniques to identify and analyze the hydrophobic microenvironments surrounding hydrogen bonds in protein structures. Here, we provide an introduction to the dehydron and the computational techniques that have been used to uncover its biological and biomedical significance. We then illustrate how dehydron-based computational analysis can be used as a basis for reengineering pharmaceutical compounds to improve their binding specificities.

Keywords Dehydron · Drug design · Hydrogen bond · Hydrophobic · Solvation

53.1 The Dehydron

Hydrogen bonds and hydrophobic interactions have long been recognized as key determinants of protein structure and function [43, 45, 47 49]. Perhaps less well known is that hydrophobic/nonpolar groups (NPGs) (i.e., CH_n, $n = 0, 1, 2, 3$) can enhance the formation of hydrogen bonds by protecting them from solvation [2, 14, 18, 28 30, 34, 35, 39, 42, 50]. Recent research has shown that solvent-exposed backbone hydrogen bonds (i.e., carbonyl-amide hydrogen bonds not protected by hydrophobic groups) influence a wide array of biological phenomena [11 13, 16 20, 22, 23, 28, 31, 32, 38]. Moreover, it has been demonstrated that these unprotected bonds can guide the redesign of ligands to improve specificity [4 6, 15, 24, 26, 27, 33]. The use of computational techniques to quantitatively assess the effect of hydrophobes

L. Ridgway Scott (✉)
Institute for Biophysical Dynamics, The Computation Institute and Departments of Computer Science and Mathematics, University of Chicago, 1100 E. 58th St., Chicago, IL 60637, USA
e mail: ridg@cs.uchicago.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 53,

on hydrogen bonds has been key to many of these developments. We describe herein the basis for these techniques and their use in bioengineering.

The propensity for clustered hydrophobes to produce a localized region conducive to hydrogen bond formation has emerged in several different lines of research, and has been referred to alternatively as "blocking" [2], "shielding" [39], and "wrapping" [34], with each term denoting the insulation of an electrostatic interaction (the hydrogen bond) from a dissipative dielectric (water). We use the term "wrapping" henceforth to describe the desolvation of a hydrogen bond by hydrophobic groups. A hydrogen bond formed within a hydrophobic microenvironment may be thought of as "well-wrapped," while a hydrogen bond in a solvent-exposed microenvironment may be thought of as "under-wrapped."

In [2] the authors estimate that a hydrogen bond is stabilized by 0.5 kJ/mol for each residue close to the bond. Consistent with this estimate, the analysis presented in [29] determines the average decrease in Coulomb energy associated with the complete desolvation of a backbone hydrogen bond to be 3.91 ± 0.67 kJ/mol. Under-wrapped hydrogen bonds (UWHBs) are thermodynamically stabilized by desolvation because the enthalpic benefit associated with water removal offsets the entropic cost due to hydrophobe immobilization [2, 14, 29, 30, 34]. Furthermore, this energetic favorability results in a measurable, attractive force between UWHBs and hydrophobes [14, 18, 29]. It is this attractive force, or "stickiness" that has led to the classification of the UWHB as a new biochemical motif: the "dehydron" [30].

To define a dehydron, we first need a precise way of describing the microenvironment surrounding a hydrogen bond. The term "desolvation domain" is used to describe this microenvironment, or more precisely, the region in which the presence of water may have an adverse effect on hydrogen bond formation [23, 28, 34]. A simple and effective approach to define the desolvation domain consists of designating spheres surrounding the donor and acceptor atoms of a backbone hydrogen bond, as in Fig. 53.1. The desolvation domain was chosen in [28] to be the union of two intersecting 6.5 Å-radius spheres centered at the α-carbons of the residues paired by the hydrogen bond. The 6.5 Å-radius represents a typical cutoff distance when evaluating interactions between nearby residues [28, 40], and is the radius at which the packing of side chain residues appears to achieve a maximum density [40].

Given the definition of a desolvation domain, we can now provide a detailed account of what it means to "wrap" a hydrogen bond and, conversely, what it means for a hydrogen bond to be under-wrapped. For a protein structure to persist in water, its electrostatic bonds must be protected from solvation [28, 34, 44]. This protection can be achieved via the placement of NPGs in the vicinity of electrostatic bonds. Simply put, amide and carbonyl partners in backbone hydrogen bonds can become separated temporarily due to thermal fluctuations or other movements of a protein; if such groups remain insulated from water, or well-wrapped, they are protected from hydration and more easily return to the bonded state [2, 14, 18, 28 30, 34, 35, 39, 42, 50].

The most effective way to quantify wrapping is by looking below the level of residue abstraction and counting all NPGs, independent of the type of side chain they inhabit. Under this "group-counting" schema initially proposed in [28], the

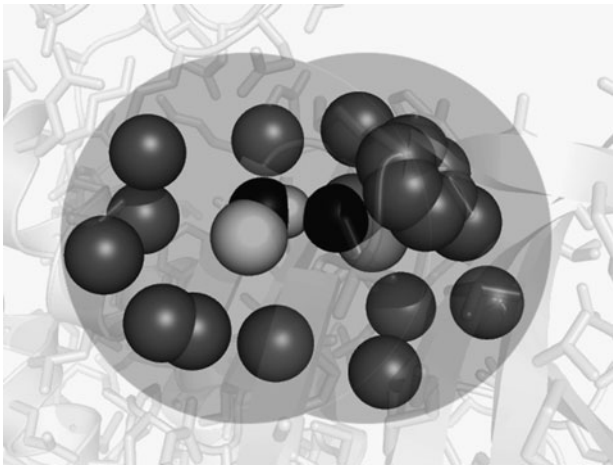


Fig. 53.1 An under wrapped backbone hydrogen bond, or "dehydron." The two large gray transparent spheres centered at the α carbons (light gray) of the donor nitrogen and acceptor oxygen (both in black) illustrate the desolvation domain. The hydrogen bond is wrapped by side chain nonpolar groups (dark gray). With the exception of the amide hydrogen (light gray), only heavy atoms are shown. PDB file: 1T45 [41] (rendered in PyMOL [10])

extent of intramolecular desolvation of a hydrogen bond is determined by the number of side chain NPGs in the desolvation domain, regardless of side-chain classification (Fig. 53.1). A statistically based definition of a "candidate dehydron" (an UWHB that has not been experimentally verified) can then be derived by taking a representative sample from the Protein Data Bank (PDB) and estimating the mean number of NPGs that wrap a hydrogen bond [28] (defining a hydrogen bond via geometric criteria, as described in [3]). One may then define a candidate dehydron as a hydrogen bond whose number of wrappers, ρ_G, is one standard deviation below the sample mean. Research has shown that $\rho_G = 19$ serves as a reliable cutoff between well-wrapped and under-wrapped backbone hydrogen bonds for a desolvation domain radius of 6.5 Å [15, 19, 24, 36].

53.2 Computational Analysis of the Dehydron

When combined with (1) a simple algorithm to predict hydrogen positions and (2) a geometric definition of the hydrogen bond, the group-counting method is able to reliably identify UWHBs in PDB structures [4, 13, 15, 16, 23 25, 27, 28, 31, 32]. The first computational biology software to implement this simple method was YAPView [4], a molecular viewing program designed to identify and display candidate dehydrons in individual protein structures.

YAPView is similar to widely used molecular viewers such as RASMOL [46] and PyMOL [10] in that it provides a means of visually inspecting graphically

rendered PDB files. YAPView employs a simple algorithm to find potential hydrogen bonds and count the number of NPGs within their desolvation domains. The software then highlights the associated α-carbons if the number of NPGs is below a user-specified threshold [4]. YAPView has been used to locate protein protein interaction sites [16, 32], to undertake evolutionary studies [13, 19, 38], and to guide pharmaceutical reengineering [4, 15, 24, 33]. YAPView has also served as motivation to design new dehydron data mining software (now under development) that provides a flexible and scalable platform for high-throughput studies of the PDB [36].

A more sophisticated approach to analyzing dehydrons involves testing hydrogen bonds *in silico* for "de-wetting propensity" [7 9, 21, 26, 27]. In this case, molecular dynamics (MD) simulations are used to estimate the residence times of water molecules in proximity of candidate UWHBs. Hydrogen bonds whose hydration shells are only weakly formed (and so characterized by low residence times for water molecules) are defined as candidate dehydrons [21, 26, 27]. As with the simple method exemplified by YAPView, the MD-based method of dehydron identification has been used to guide experimental work [26, 27]. Researchers have also begun to implement MD-based dehydron analysis to examine the potential impact of UWHBs on protein structures [7 9].

53.3 Dehydron-Guided Ligand Reengineering

The computational techniques described above have already been used to improve the binding specificities of pharmaceutical compounds [4 6, 15, 24, 26, 27, 33]. A general protocol for dehydron-guided ligand reengineering is as follows. First, a given protein and its most similar paralogs (to which a ligand binds) are aligned. Next, candidate UWHBs located within the binding sites of each protein are compared for proximity, and the dehydrons unique to the protein target of interest are determined. These candidate wrapping sites are then used to guide the synthesis of a modified form of the ligand. That is, if a dehydron is found in the ligand binding site of the protein of interest but not in the binding sites of the protein's paralogs, a methylated variant of the ligand that is likely to wrap the unique dehydron is synthesized.

The effectiveness of dehydron-guided ligand reengineering is illustrated by the redesign of the anticancer drug imatinib (STI-571/Gleevec/Glivec) [15, 26, 27]. Imatinib has proven to be successful in treating several types of cancer [41], and is particularly effective at combating gastrointestinal tumors through inhibition of the c-Kit tyrosine kinase [1]. However, imatinib's affinity for the Abl kinase can lead to cardiotoxic side effects [37]. To eliminate these side effects, an imatinib variant (WBZ 4) methylated at the C2 position was synthesized, with the additional NPG intended to wrap the Cys673-Gly676 (acceptor donor) dehydron unique to c-Kit (Fig. 53.2) [26]. WBZ 4 was shown to inhibit peptide phosphorylation by c-Kit while having a substantially reduced inhibitory effect on a wide array of paralogous

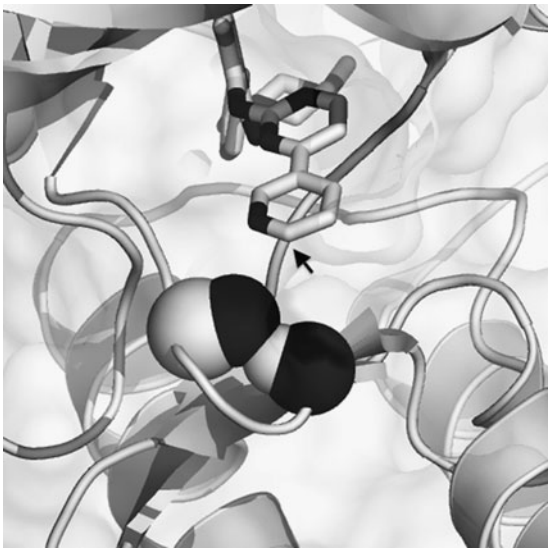


Fig. 53.2 Imatinib binding to the c Kit kinase. The C2 position (arrow) was identified in [26] as a potential methylation site to wrap the Cys673 Gly676 dehydron (black and white spheres in the lower portion of the figure). The C2 methylated form of imatinib (WBZ 4) exhibits a higher specificity for c Kit relative to its specificity for Abl, Lck, Pdk1, and Chk1. PDB file: 1T46 [41] (rendered in PyMOL [10])

kinases (Abl, Lck, Pdk1, and Chk1 being the most similar in structure and dehydron distribution to c-Kit) [26]. The effectiveness of WBZ 4 as an improved anticancer agent has also been tested in animal models; not only is it equally effective at reducing gastrointestinal tumor growth compared to the original imatinib, but it fails to exhibit the cardiotoxicity of imatinib as well [26]. Dehydron-guided ligand reengineering has also been used to produce other imatinib variants [15, 27], including one that shows promise in treating patients who are resistant to the original drug [27].

Dehydron analysis thus offers a means of enhancing the clinical safety and efficacy of pharmaceuticals. However, this is only one of its many applications. Dehydron analysis is becoming an important tool for basic scientific research and biomolecular engineering alike, and shows how computational techniques can be employed to refine hypotheses and accelerate the discovery process in the life sciences.

Acknowledgments Ridgway Scott and Christopher Fraser would like to thank the Institute for Mathematics and Its Applications for its support. Ridgway Scott is also supported by the NSF grant DMS 0920960. Ariel Fernández is supported in part through NIH grant R01 GM072614 from the National Institute of General Medical Sciences.

References

1. Attoub S, Rivat C, Rodrigues S et al (2002) The c kit tyrosine kinase inhibitor STI571 for colorectal cancer therapy. Cancer Res 62:4879 4883
2. Bai Y, Englander S (1994) Hydrogen bond strength and β sheet propensities: The role of a side chain blocking effect. Proteins 18:262 266

3. Baker EN, Hubbard RE (1984) Hydrogen bonding in globular proteins. Prog Biophys Mol Biol 44:97 179
4. Chen J, Zhang X, Fernández A (2007) Molecular basis for specificity in the druggable kinome: sequence based analysis. Bioinformatics 23:563 572
5. Crespo A, Fernández A (2008) Induced disorder in protein ligand complexes as a drug design strategy. Mol Pharm 5:430 437
6. Crespo A, Zhang X, Fernández A (2008) Redesigning kinase inhibitors to enhance specificity. J Med Chem 51:4890 4898
7. De Simone A, Dodson GG, Fraternali F et al (2006) Water molecules as structural determi nants among prions of low sequence identity. FEBS Lett 580:2488 2494
8. De Simone A, Dodson GG, Verma CS et al (2005) Prion and water: Tight and dynamical hydration sites have a key role in structural stability. Proc Natl Acad Sci USA 102: 7535 7540
9. De Simone A, Zagari A, Derreumaux P (2007) Structural and hydration properties of the partially unfolded states of the prion protein. Biophys J 93:1284 1292
10. DeLano W (2002) The PyMOL molecular graphics system. DeLano Scientific, Palo Alto, CA
11. Despa F, Fernández A, Scott LR et al (2008) Hydration profiles of amyloidogenic molecular structures. J Biol Phys 34:577 590
12. Fernández A (2002) Insufficient hydrogen bond desolvation and prion related disease. Eur J Biochem 269:4165 4168
13. Fernández A (2004) Functionality of wrapping defects in soluble proteins: What cannot be kept dry must be conserved. J Mol Biol 337:477 483
14. Fernández A (2005) Direct nanoscale dehydration of hydrogen bonds. J Phys D Appl Phys 38:2928 2932
15. Fernández A (2005) Incomplete protein packing as a selectivity filter in drug design. Struc ture 13:1829 1836
16. Fernández A (2005) What factor drives the fibrillogenic association of beta sheets? FEBS Lett 579:6635 6640
17. Fernández A, Berry RS (2002) Extent of hydrogen bond protection in folded proteins: a constraint on packing architectures. Biophys J 83:2475 2481
18. Fernández A, Berry RS (2003) Proteins with H bond packing defects are highly interactive with lipid bilayers: Implications for amyloidogenesis. Proc Natl Acad Sci USA 100:2391 2396
19. Fernández A, Berry RS (2004) Molecular dimension explored in evolution to promote proteomic complexity. Proc Natl Acad Sci USA 101:13460 13465
20. Fernández A, Boland M (2002) Solvent environment conducive to protein aggregation. FEBS Lett 529:298 302
21. Fernández A, Chen J, Crespo A (2007) Solvent exposed backbone loosens the hydration shell of soluble folded proteins. J Chem Phys 126:245103
22. Fernández A, Colubri A, Berry RS (2002) Three body correlations in protein folding: the origin of cooperativity. Physica A 307:235 259
23. Fernández A, Kardos J, Scott LR et al (2003) Structural defects and the diagnosis of amyloidogenic propensity. Proc Natl Acad Sci USA 100:6446 6451
24. Fernández A, Maddipati S (2006) A priori inference of cross reactivity for drug targeted kinases. J Med Chem 49:3092 3100
25. Fernández A, Plazonic KR, Scott LR et al (2004) Inhibitor design by wrapping packing defects in HIV 1 proteins. Proc Natl Acad Sci USA 101:11640 11645
26. Fernández A, Sanguino A, Peng Z et al (2007) An anticancer C Kit kinase inhibitor is reengineered to make it more active and less cardiotoxic. J Clin Invest 117:4044 4054
27. Fernández A, Sanguino A, Peng Z et al (2007) Rational drug redesign to overcome drug resistance in cancer therapy: Imatinib moving target. Cancer Res 67:4028 4033
28. Fernández A, Scheraga HA (2003) Insufficiently dehydrated hydrogen bonds as determinants of protein interactions. Proc Natl Acad Sci USA 100:113 118

29. Fernández A, Scott LR (2003) Adherence of packing defects in soluble proteins. Phys Rev Lett 91:18102
30. Fernández A, Scott LR (2003) Dehydron: A structurally encoded signal for protein interac tion. Biophys J 85:1914 1928
31. Fernández A, Scott LR (2003) Under wrapped soluble proteins as signals triggering mem brane morphology. J Chem Phys 119:6911 6915
32. Fernández A, Scott LR, Berry RS (2004) The nonconserved wrapping of conserved protein folds reveals a trend towards increasing connectivity in proteomic networks. Proc Natl Acad Sci USA 101:2823 2827
33. Fernández A, Sessel S (2009) Selective antagonism of anticancer drugs for side effect removal. Trends Pharmacol Sci 30:403 410
34. Fernández A, Sosnick TR, Colubri A (2002) Dynamics of hydrogen bond desolvation in protein folding. J Mol Biol 321:659 675
35. Franzen J, Stephens R (1963) The effect of a dipolar solvent system on interamide hydrogen bonds. Biochemistry 2:1321 1327
36. Fraser CM, Tran TVD, Fernández A et al (2009) WRAPPA: Web based Residue Analysis Program for Phobicity Assessment. Poster: 17th Annual International Conference on Intelli gent Systems for Molecular Biology (ISMB) & 8th European Conference on Computational Biology (ECCB). Stockholm, Sweden.
37. Kerkelä R, Grazette L, Yacobi R et al (2006) Cardiotoxicity of the cancer therapeutic agent imatinib mesylate. Nat Med 12:908 916
38. Liang H, Plazonic KR, Chen J et al (2008) Protein under wrapping causes dosage sensitivity and decreases gene duplicability. PLoS Genet 4:e11
39. Luo P, Baldwin RL (1999) Interaction between water and polar groups of the helix backbone: An important determinant of helix propensities. Proc Natl Acad Sci USA 96: 4930 4935
40. Miyazawa S, Jernigan RL (1985) Estimation of effective interresidue contact energies from protein crystal structures: Quasi chemical approximation. Macromolecules 18:534 552
41. Mol CD, Dougan DR, Schneider TR et al (2004) Structural basis for the autoinhibition and STI 571 inhibition of c Kit tyrosine kinase. J Biol Chem 279:31655 31663
42. Némethy G, Steinberg IZ, Scheraga HA (1963) Influence of water structure and of hydropho bic interactions on the strength of side chain hydrogen bonds in proteins. Biopolymers 1:43 69
43. Pace CN (1992) Contribution of the hydrophobic effect to globular protein stability. J Mol Biol 226:29 35
44. Petrey D, Honig B (2000) Free energy determinants of tertiary structure and the evaluation of protein models. Protein Sci 9:2181 2191
45. Rose GD, Wolfenden R (1993) Hydrogen bonding, hydrophobicity, packing, and protein folding. Annu Rev Bioph Biom 22:381 415
46. Sayle RA, Milner White EJ (1995) RASMOL: Biomolecular graphics for all. Trends Biochem Sci 20:374 376
47. Scheraga HA, Némethy G, Steinberg IZ (1962) The contribution of hydrophobic bonds to the thermal stability of protein conformations. J Biol Chem 237:2506 2508
48. Stickle D, Presta L, Dill K et al (1992) Hydrogen bonding in globular proteins. J Mol Biol 226:1143 1159
49. Tanford C (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J Am Chem Soc 84:4240 4247
50. Vila JA, Ripoll DR, Scheraga HA (2000) Physical reasons for the unusual alpha helix stabilization afforded by charged or neutral polar residues in alanine rich peptides. Proc Natl Acad Sci USA 97:13075 13079

Chapter 54
Docking to Large Allosteric Binding Sites on Protein Surfaces

Ursula D. Ramirez, Faina Myachina, Linda Stith, and Eileen K. Jaffe

Abstract The inactive porphobilinogen synthase (PBGS) hexamer has an oligomer-specific and phylogenetically variable surface cavity that is not present in the active octamer. The octamer and hexamer are components of a dynamic quaternary structure equilibrium characteristic of morpheeins. Small molecules that bind to the hexamer-specific surface cavity, which is at the interface of three subunits, are predicted to act as allosteric inhibitors that function by drawing the oligomeric equilibrium toward the hexamer. We used GLIDE as a tool to enrich a 250,000 molecule library for molecules with enhanced probability of acting as hexamer-stabilizing allosteric inhibitors of PBGS from *Yersinia enterocolitica*. Eighty-six compounds were tested *in vitro* and five showed hexamer stabilization. We discuss the application of computational docking to surface cavities as an approach to find allosteric modulators of protein function with specific reference to morpheeins that function as an equilibrium of non-additive quaternary structure assemblies.

Keywords Computational docking · GLIDE · Hexamer · PBGS · Protein surfaces

Small molecule docking to protein targets was developed as a drug discovery tool at a time when drug discovery was focused predominantly on enzyme active sites rather than allosteric sites. However, as early as 1963, Monod and coworkers astutely pointed out that protein function can be regulated allosterically by small molecules that need not be structurally related to a protein's target (e.g., substrate) because they bind somewhere other than the active site [1]. Some allosteric sites are similar to active sites, which are often located in deep grooves, somewhat buried in the protein structure, and contain a limited set of residues that provide a

E.K. Jaffe (✉)
Fox Chase Cancer Center, 333 Cottman Ave, Philadelphia, PA 19111, USA
e mail: eileen.jaffe@fccc.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 54,

well-defined binding site. Other allosteric sites are more like protein protein interfaces (PPIs), which have less rigid binding requirements, frequently have species-specific variations in composition, and have overall greater structural flexibility.

Application of computational docking to allosteric sites or PPIs has been done successfully [2 5] despite the fact that most related software has been developed and benchmarked with active sites in mind [6]. Oftentimes, targeting PPIs is undertaken to find small molecules that will interfere with oligomer assembly. Alternatively, one can target a specific cleft formed at a PPI with the notion of stabilizing a particular assembly. We have undertaken a series of docking studies that address allosteric sites that occur at PPIs where small molecule binding modulates an equilibrium of functionally distinct alternate quaternary structure assemblies [4, 5]. Proteins with such assemblies have been called morpheeins, and the equilibrium can be illustrated using a morphing dice schematic (Fig. 54.1). The distinguishing characteristic of proteins that function as morpheeins is that there exist alternate assemblies, and each oligomeric assembly may have surface clefts that are assembly-specific. These clefts do not have the evolutionary requirement for conservation that is characteristic of active sites. The utility of docking to clefts in PPIs has been established [7, 8]. Oligomer-stabilizing, small molecule binding to one oligomeric form of a morpheein equilibrium can be schematically represented in the dice analogy by a tetramer-specific wedge whose binding draws the equilibrium toward that oligomer (Fig. 54.1).

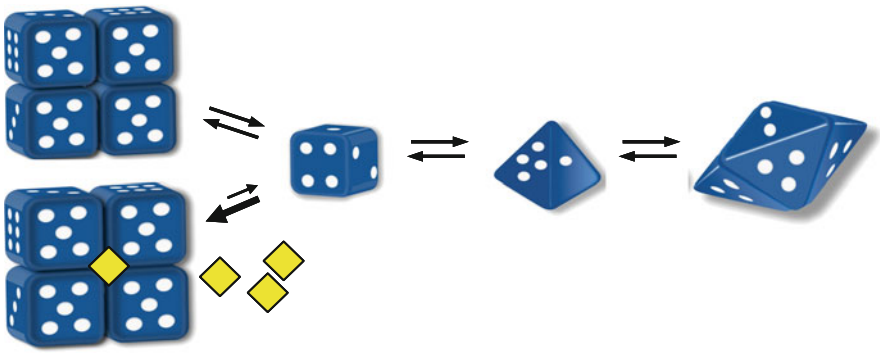


Fig. 54.1 *Cubic and tetrahedral dice illustrate the behavior of proteins that function as mor pheeins.* The quaternary structure dynamic characteristic of morpheeins is illustrated using a dice analogy where one die can morph between cubic and tetrahedral shapes. The illustrated assemblies associate the die face with one spot with the die face with four spots, such that the cubic die assembles into a tetramer and the tetrahedral die assembles into a pentamer. The analogy is that two different conformations of a protein subunit can each dictate assembly to a different oligomer. All dice in one assembly must be of the same shape. Thus, for example, the tetramer must come apart, and its component dice must change shape to a pyramid before participating in assembly into a pentamer. A diamond shape is used to illustrate a tetramer specific stabilizing agent that can bind to the largest face of the tetramer. Binding of this shape would hinder tetramer dissociation, and thereby pull the equilibrium toward the tetramer

54.1 PBGS: The prototype Morpheein Drug Discovery Target

Porphobilinogen synthase (PBGS, EC 4.2.1.24), also known as 5-aminolevulinic acid dehydratase (or ALAD), catalyzes the first common step in the biosynthesis of the tetrapyrrole pigments (for example heme, chlorophyll, and vitamin B_{12}). As such, PBGS is an essential enzyme for most organisms. The residues of the PBGS active site are highly conserved [9], and thus do not provide sufficient structural variation to yield species-selective inhibitors through computational docking. However, PBGS from some species are established to exist in a dynamic equilibrium of oligomeric forms, including a high-activity octamer, and a low-activity hexamer, the interconversion of which occurs at the level of a dimer whose conformation dictates the stoichiometry and architecture of the higher assembly state [4, 5, 10 14]. One dimer conformation is competent for assembly to the octamer and the other for assembly to the hexamer. The percentage of each component in the equilibrium of PBGS quaternary structure assemblies is dependent on protein sequence and responds to protein concentration, pH, and ligand binding at the active site or the allosteric site [13 15]. The physiological relevance of the equilibrium of quaternary structure assemblies for human PBGS is established through the relationship between this equilibrium and the disease ALAD porphyria [16]. The physiological relevance of the quaternary structure assemblies for plant PBGS is established by the existence of a naturally occurring allosteric activator of the plant protein [10]. Of importance to a discussion of docking is that a hexamer-specific cavity exists, which is not phylogenetically conserved. Therefore, this cavity can be targeted for the development of species-selective inhibitors as lead compounds for antimicrobials, herbicides, or pesticides. The structural basis for this binding site is described below.

The first crystal structures of PBGS revealed each subunit to have two domains, an αβ-barrel and an extended N-terminal arm [17]. Crystal structures of the octameric assembly have been solved for PBGS from multiple species [18]. In each dimeric unit cell of octameric PBGS, the arm of one subunit wraps around the αβ-barrel of the neighboring subunit, making a hugging-dimer (Fig. 54.2) [17].

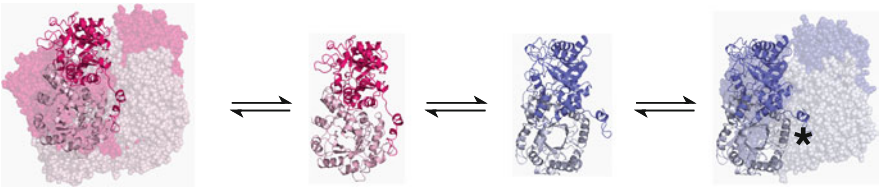


Fig. 54.2 *The PBGS quaternary structure equilibrium.* The crystallographic asymmetric units of PBGS octamers and hexamers are hugging dimer (*middle left*) and detached dimer (*middle right*), respectively. These dimers assemble around a central axis to form the larger oligomers (90° rotation for the octamer (*left*) and 120° rotation for the hexamer (*right*)). The dimers are shown as ribbons with one subunit in a light shade and the other in a dark shade. Each dimer is shown in the context of its oligomer with the remaining subunits shown as spheres. The star illustrates the location of the hexamer specific surface cavity

There is only one structure of a PBGS hexamer that is of the naturally occurring (and porphyria-associated) human PBGS variant F12L [10]. The asymmetric unit of the hexameric PBGS crystal structure is a dimer with the N-terminal arms extended ("detached-dimer") (Fig. 54.2). The conformation of the N-terminal arm determines which assembly can be formed (Fig. 54.2). The hexamer is less active than the octamer because it lacks an arm-to-barrel interaction that stabilizes the closed conformation of the active site lid [19].

The conformational differences between the subunits comprising the hexamers versus octamers define an oligomer-specific target region that can be utilized to find small molecules that preferentially bind only to the hexamer. We have targeted a hexamer-specific surface cavity to find small molecules that will trap the low-activity hexameric form of human and pea PBGS with some success. In the case of human PBGS, we docked to the crystal structure of the hexamer [4, 10]; in the case of pea PBGS, we docked to a homology model that was prepared using multiple template structures [5]. In that case, the human hexamer structure (PDB code 1PV8) was used alongside the octameric *Pseudomonas aeruginosa* PBGS (*Pa*PBGS) structure (PDB code 1GZG [20]) to form a model of the *Pa*PBGS hexamer. The *Pa*PBGS model was used as the template to prepare homology models of pea and *Yersinia enterocolitica* PBGS (*Ye*PBGS) hexamers. Here we report the results of docking to the *Ye*PBGS homology model. We discuss the challenges and solutions we have found for these docking studies.

54.2 Computational Docking Studies for Discovery of Compounds that Stabilize the Inactive PBGS Hexamer

The goal of these studies was to find small molecules that would preferentially bind to and stabilize the inactive PBGS hexamer. In the case of the human protein, such compounds would inhibit the protein and potentiate diseases related to low PBGS activity. These diseases are ALAD porphyria and lead poisoning. In the case of the prototype plant pea PBGS, such compounds, if species selective, could form the starting point for the development of a herbicide. In the case of *Ye*PBGS, such compounds provide starting points for a new class of antibiotic agents. In all cases, we used the same version of the docking software GLIDE with libraries of small molecules from Life Chemicals, Inc. that the company had purported to be drug-like with a molecular weight of approximately 350 500 Da. Against pea and human PBGS, we used only a 65,000 compound subset of the company's library, whereas for *Ye*PBGS, we used their entire "in stock" library (250,000 compounds). Here we report results for *Ye*PBGS and compare the hit compounds to those obtained from docking to the human and pea proteins.

A large docking site (a cubic region with 25 Å dimensions into which the ligand must fit and a central cubic region with 14 Å dimensions in which the center of the ligand must lie) was defined at the interface of three of the subunits that comprise

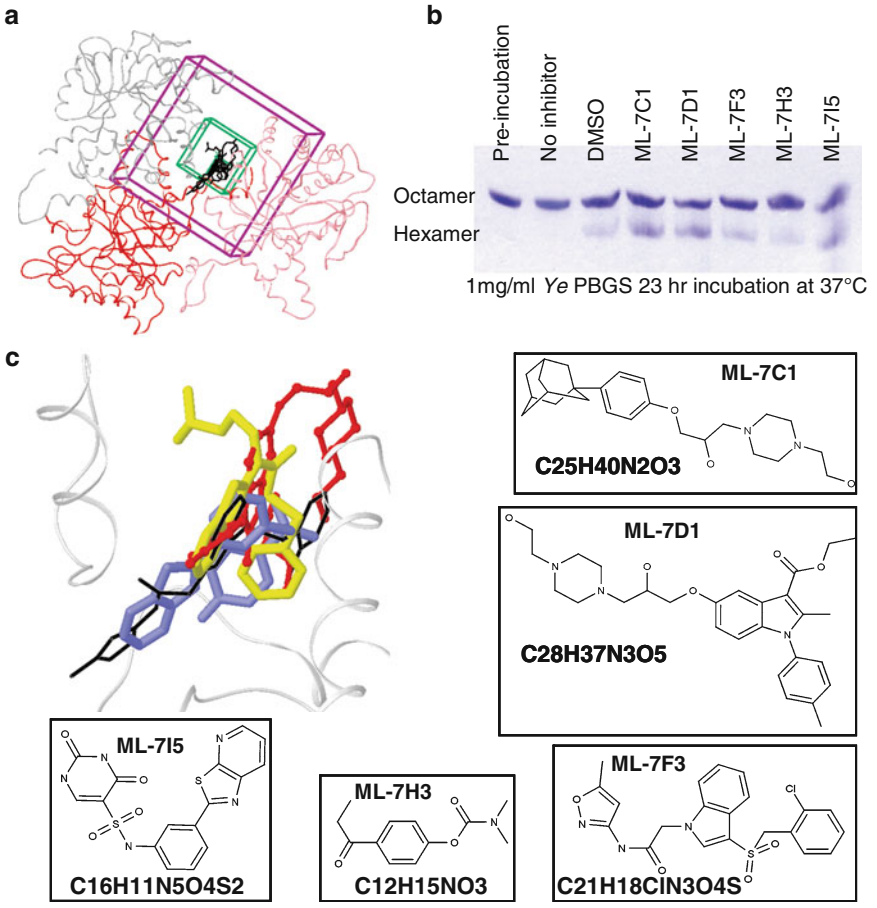


Fig. 54.3 *YePBGS hexamer stabilizing compounds.* (**a**) Ribbon diagram of the three *Ye*PBGS hexamer subunits comprising the docking target; inner and outer docking boxes are indicated. The docked poses of compounds ML 7D1, ML 7F3, ML 7H3, and ML 7I5 are shown as sticks. (**b**) *Ye*PBGS at 1 mg/ml was not incubated (pre incubation lane) or incubated for 23 h at 37°C alone, with DMSO only or with one of the five *Ye*PBGS hexamer stabilizing compounds (2 mM in DMSO). Samples were resolved on 12.5% polyacrylamide native PhastGels. (**c**) Close up of (**a**) with docked poses of compounds ML 7D1 in ball and stick, ML 7F3 in black wire, ML 7H3 light shade tube, and ML 7I5 in dark shaded tube. Two dimensional structural representations and molecular formulas of all five *Ye*PBGS hexamer stabilizing compounds are shown

the hexamer (Fig. 54.3a). Because the asymmetry in the PBGS dimer in the crystallographic asymmetric units impacts residues in the docking site, two different docking boxes were used as described previously [5]. We docked to equivalent sites in the human PBGS hexamer crystal structure and the pea PBGS and *Ye*PBGS hexamer homology models. Two-dimensional structures supplied by Life Chemicals, Inc. were prepared for docking using LIGPREP (Schrödinger, Inc.) to limit molecules to those comprised of only C, H, N, O, S, P, and halogens and then to

generate three-dimensional representations of all stereoisomers, tautomers, and ionic states within a narrow near-neutral pH range for each structure.

The same docking strategies were applied to all targets. Docking was done first by Standard Precision Mode (SP) of GLIDE version 3.5 docking software from Schrodinger, Inc [21]. The top 10% of compounds identified by SP Glidescore (a proprietary modified Chemscore) were then docked using Extra Precision (XP) Mode of GLIDE version 3.5. The top 10% of compounds identified by XP Glidescore (or about 1% of our starting library) were then further analyzed. These compounds were evaluated for solubility by calculating Log S solubility estimate using QIKPROP (Schrödinger, Inc.), and compounds with Log S less than 6 were eliminated from further consideration since compounds need to be soluble for testing in *in vitro* assays. Further consideration limited the remaining docked structures to those that were within proper distance to make van der Waals contacts or hydrogen bonds to at least one atom from each of the three subunits that comprise the docking site. This criterion limited compounds selected to those with the highest potential for stabilizing the hexamer. Finally, the remaining compound poses were manually sorted to provide the broadest possible assortment of chemical diversity and binding locations within the docking region.

Because we do not know of naturally occurring ligands that bind to the targeted site, we have no *a priori* information about the chemical structure of the molecules for which we were searching. Thus, our approach was designed to maximize sampling of experimental space. Using the above approaches and a final manual by-eye inspection, approximately 100 compounds were selected for purchase and *in vitro* testing against each target. Based on supplier availability, 76 compounds for pea PBGS, 77 for human PBGS, and 86 for *Ye*PBGS were obtained. Purchased compounds were tested using both native polyacrylamide gel electrophoresis to separate oligomeric forms of PBGS, and activity assays to quantify inhibition of PBGS. As reported previously, of twelve compounds seen to stabilize the human PBGS hexamer, chemical characterization yielded two compounds that significantly shifted the equilibrium of human PBGS, but not pea PBGS, toward hexamer [4]. In the case of pea PBGS, of 10 compounds that each provided some stabilization of the hexamer, only one was a potent inhibitor that shifted pea PBGS, but not human PBGS toward hexamer [5]. These results demonstrated that both homology models and crystal structures can be used to identify compounds using our methods. Here we present our results for *Ye*PBGS.

54.3 Hexamer-Stabilizing Inhibitors of *Ye*PBGS

A total of 86 compounds were individually incubated at 37°C with *Ye*PBGS (1 mg/ml), and native gel electrophoresis was used to evaluate the percentage of hexamer and octamer present relative to protein that was incubated with solvent DMSO alone. This method determined whether the compounds caused an increased mole fraction of hexamer in the equilibrium of *Ye*PBGS oligomers. Incubations were

done for 1, 2, 4, 6, and 23 h prior to loading gels. Five of the compounds (referred to here as ML-7C1, ML-7D1, ML-7H3, ML-7F3, and ML-7I5) showed some increase in the mole fraction of hexamer and a decrease in the mole fraction of octamer (Fig. 54.3b). Compound ML-7C1 was selected from docking to one box, and the other four compounds from the other docking box. The docked poses of the four compounds that docked to the same box are shown in Fig. 54.3c along with the structures and chemical formulas of all five hexamer-stabilizing agents. Compound ML-7D1 showed maximal mole fraction hexamer by the end of the 2-h incubation, compound ML-7C1 showed maximal mole fraction hexamer after 6 h, and the other three reached their maxima by 23 h (data are shown for 23-h incubation only). Inhibition assay data were inconclusive for these compounds because the protein's activity was impaired by the extended incubation time with and without inhibitors (data not shown). Further characterization of these molecules requires that we overcome this experimental limitation.

The current work is significant in three ways. First, the hexamer-stabilizing compounds reported here offer a starting point for species-specific antibiotics. Although inhibition by these compounds could not be quantified, three of them (ML-7C1, ML-7D1, and ML-7H3) were shown to inhibit *Y. enterocolitica* growth in an *in vivo* disk zone inhibition assay (data not shown). Second, the results demonstrate the utility of docking to PPIs on homology models. Although PPIs have a lower phylogenetic conservation than active sites, PPIs also have significant structural flexibility. Residues in PPIs may sample many conformations of which the template crystal structure is only one. Thus, a homology model of a PPI is likely to represent a conformation that is populated in solution. Third, oligomer-stabilizing compounds may provide tools to assist in crystal structure determination, which in turn will allow a more critical evaluation of docking to homology models. The quaternary structure dynamic characteristic of morpheeins has interfered with the generation of crystals that diffract to high resolution. Co-crystal structures of PBGS with oligomer-stabilizing compounds are expected to allow a better understanding of docking to PPIs of homology models.

Acknowledgments The authors acknowledge Susan Slechta for *in vivo* testing of inhibitors, the Fox Chase Cancer Center High Performance Computing Cluster, and grant support from the National Institutes of Health grants R01 ES003654 (EKJ), R21 AI063324 (EKJ), P30 CA006927 (FCCC), and T32 CA009035 (Institute for Cancer Research, a component of FCCC).

References

1. Monod, J, et al., J Mol Biol, 1963. **6**:306 329.
2. Bond, CJ, et al., Biochemistry, 2000. **39**(50): 15333 15343.
3. Gonzalez Ruiz, D, Gohlke, H, Curr Med Chem, 2006. **13**(22):2607 2625.
4. Lawrence, SH, et al., J Biol Chem, 2009. **284**:35807 35817.
5. Lawrence, SH, et al., Chem Biol, 2008. **15**(6):586 596.

6. Alvarez, J, Shoichet, B, *Virtual Screening in Drug Discovery*. 1st ed. 2005, Boca Raton, FL: CRC Press. 470.
7. Arkin, MR, et al., Proc Natl Acad Sci USA, 2003. **100**(4):1603 1608.
8. Li, Y, et al., Structure, 2005. **13**(2): 297 307.
9. Jaffe, EK, Chem Biol, 2003. **10**(1):25 34.
10. Breinig, S, et al., Nat Struct Biol, 2003. **10**(9): 757 763.
11. Kokona, B, et al., Biochemistry, 2008. **47**(40): 10649 10656.
12. Selwood, T, et al., Biochemistry, 2008. **47**(10):3245 3257.
13. Tang, L, et al., J Biol Chem, 2006. **281**(10): 6682 6690.
14. Tang, L, et al., J Biol Chem, 2005. **280**(16):15786 15793.
15. Jaffe, EK, Trends Biochem Sci, 2005. **30**(9):490 497.
16. Jaffe, EK and Stith, L, Am J Hum Genet, 2007. **80**(2):329 337.
17. Erskine, PT, et al., Nat Struct Biol, 1997. **4**(12):1025 1031.
18. Berman, HM, et al., Nucleic Acids Res, 2000. **28**(1):235 242.
19. Jaffe, EK, Bioorg Chem, 2004. **32**(5): 316 325.
20. Frere, F, et al., J Mol Biol, 2002. **320**(2): 237 247.
21. Halgren, TA, et al., J. Med Chem, 2004. 47: 1750 1759.

Chapter 55
Modeling of ATP-Sensitive Inward Rectifier Potassium Channel 11 and Inhibition Mechanism of the Natural Ligand, Ellagic Acid, Using Molecular Docking

Alex J Mathew, Nixon N Raj, M Sugappriya, and Sangeetha M Priyadarshini

Abstract Diabetes mellitus is a disorder in which blood sugar (glucose) levels are abnormally high because the body does not produce enough insulin to meet its needs. Post-prandial hyperglycemia (PPHG) is an independent risk factor for the development of macro vascular complications. It is now recognized that normalizing post-prandial blood glucose is more difficult than normalizing fasting glucose. Potassium channels are the most widely distributed type of ion channel and are found in virtually all living organisms. The function of KATP channels is best understood in pancreatic beta cells, the membrane potential of which is responsive to external glucose concentration. Beta cells show a remarkably complex electrical bursting behavior in response to an increase in glucose level. Nateglinide and Glimepiride are a class of insulin secretagog agents that lowers blood glucose levels by stimulating insulin secretion from the pancreas. These compounds interact with the ATP-sensitive potassium (K+ATP) channel in pancreatic beta cells. However, the side effects of these drugs overpass their uses, and the need to identify compounds with less adverse effects is exigent. In our research study, we used the natural compound ellagic acid, which is an already proven anti-carcinogen, anti-mutagen, and anticancer initiator, for its anti-diabetic activity in comparison to the two commercial drugs (Nateglinide and Glimepiride). The drugs and the compounds were docked to the ATP-dependent potassium channel and their energy value showed that the compound had higher binding value than the commercial drugs. Then an ADME/Tox analysis for the compound was carried out which showed that ellagic can be a possible lead molecule.

Keywords Beta cells · Diabetes mellitus · KATP · Molecular docking · PPHG

A.J. Mathew (✉)
Department of Bioinformatics, Sathyabama University, Chennai 600119, Tamil Nadu, India
e mail: mathponz@yahoo.co.in

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 55,

55.1 Introduction

The main challenge in the management of patients with diabetes mellitus is to maintain blood glucose levels as close to normal as possible. In general, normalizing post-prandial blood glucose is more difficult than normalizing fasting hyperglycemia. The molecular biology of ATP-sensitive K+ channels, or KATP channels, which are present in pancreatic b-cells in the islets of Langerhans where they play a key role in stimulus-secretion coupling by providing a link between changes in metabolism and membrane electrical activity [1]. Thus these potassium channels prove to be an effective target for controlling the post-prandial glucose level in the body. Drugs like nateglinide (N-[(trans-4-isopropylcyclohexyl)-carbonyl]-D-phenylalanine and Glimepiride trigger insulin release by direct action on the KATP channel of the pancreatic beta cells [2, 3]. The primary side effects of these drugs may result in the development of severe hypoglycemic reactions with coma, seizure, or other neurological impairment. Ellagic acid is a polyphenol antioxidant found in numerous fruits and vegetables including raspberries, strawberries, cranberries, walnuts, pecans, pomegranates, and other plant foods. The anti-proliferative and antioxidant properties of ellagic acid [4] have spurred preliminary research into the potential health benefits of ellagic acid consumption.

In the current study, it was found that ellagic acid possesses anti-diabetic property, which was analyzed by molecular interaction studies. Computational Biology and bioinformatics have the potential not only of speeding up the drug discovery process thus reducing the costs, but also of changing the way drugs are designed. Rational Drug Design (RDD) helps to facilitate and speedup the drug designing process, which involves a variety of methods to identify novel compounds. One such method is the docking of the drug molecule with the receptor (target). The site of drug action, which is ultimately responsible for the pharmaceutical effect, is a receptor.

55.2 Methodolgy

Computer-Aided Drug Design (CADD) is a specialized discipline that uses computational methods to simulate drug receptor interactions. CADD methods are heavily dependent on bioinformatics tools, applications, and databases.

The homology modeling of ATP-sensitive inward rectifier Potassium channel 11 was carried out using Modeller 9v1. MODELLER is used for homology or comparative modeling of protein three-dimensional structures [5, 6]. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints [7, 8], and can perform many additional tasks.

The critical first step in homology modeling is the identification of the best template structure; the simplest method of template identification relies on serial pairwise sequence alignments aided by database search techniques such as FASTA

and BLAST [9]. Blast search was implied, which gave the PDB ID: 1U4E, Chain A, Crystal Structure of Cytoplasmic Domains of Girk1 Channel as the most appropriate template with 49% identity and 1% gap. A model was constructed based on the target-template alignment using Modeller.

The Modeled Structure is validated using the Procheck server where the stereochemical quality of the modeled protein structure is analyzed [10]. The stereochemical validation of model structures of proteins is an important part of the comparative molecular modeling process [11]. The CE (Combinatorial Extension) Algorithm was employed for calculating the pairwise structure alignment of the template and the modeled protein [12].

The Modeled structure of Potassium channel 11 was then docked with ellagic acid, Glimepiride, and Nateglinide using Autodock 4.0. Understanding the ruling principles whereby protein receptors recognize, interact, and associate with molecular substrates and inhibitors is of paramount importance in drug discovery efforts. Protein ligand docking aims to predict and rank the structure(s) arising from the association between a given ligand and a target protein of known 3D structure [13]. AutoDock is a suite of automated docking tools. It is designed to predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure [14]. After the docking process, the resulting complexes were viewed in Accelrys Discovery Studio Visualizer. The DS Visualizer provides functionality for visualizing, analyzing, and sharing biological and chemical data. The number of hydrogen bonds formed between the protein and ligand was identified and the bond distance was calculated for all the complexes.

A significant bottleneck that remains in the drug discovery process is the analysis of the ADME and toxicity properties of drug candidates. PreADMET is a web-based application for predicting ADME data and building drug-like library using the in silico method [15]. The ADMET properties of ellagic acid were analyzed using the PreADMET web server (Figs. 55.1 and 55.2).

Fig. 55.1 Structure of modeled potassium channel 11

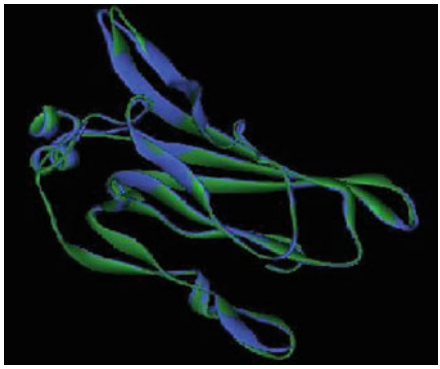


Fig. 55.2 Structural alignment of the protein (*blue*) and the template 1U4E (*green*) by CE algorithm

55.3 Result

55.3.1 Homology Modeling

The Structure of potassium channel 11 was modeled with Modeller 9v1 with 1U4E as template. The quality of the Modeled protein was evaluated using Procheck suite. On the analysis of the Ramachandran plot, it was observed that 94.8% of residues were present in the most favored regions. From CE, the Structural alignment of the modeled protein and the template was found to have RMSD value = 0.4 A°.

55.3.2 Active Site Prediction

The Active sites of the potassium channel 11 receptor was identified using the q-site finder. The Q-site finder works by binding hydrophobic probes to the protein, and finding clusters of probes with the most favorable binding energy [16]. The active sites predicted were found to be VAL43, ARG45, THR47, PRO90, LEU91, TYR92, ASP93, LEU94, ALA95, ASP98, HIS102, ASP104, LEU105, and ILE107.

55.3.3 Molecular Docking

The drugs like nateglinide, glimipiride, and the natural ligand ellagic acid were docked with the modeled potassium channel 11. The pdb files of the drugs were retrieved from Drug bank. Using Corina 3D, the pdb file for ellagic acid was generated. The docking results tabulated between the potassium channel 11 receptor and ellagic acid as well as with the conventional drugs and showed the binding affinities Nateglinide (−4.93), Glimepiride (−5.8), and Ellagic Acid (−6.58).

55.3.4 ADMET Analysis

ADME means absorption, distribution, metabolism, and excretion, which are major parts of pharmacokinetics. Numerous in vitro methods have been used in the drug selection process for assessing the intestinal absorption of drug candidates. Among them, Caco-2 cell model has been recommended as a reliable in vitro model for the prediction of oral drug absorption [17]. The PCaco-2 (nm/s) for ellagic acid was 20.48. It was also found that ellagic acid satisfied the Lipinski's rule of five [18]. One of the most crucial reasons why a drug discovery fails is the toxicity of the drug candidates. It means that designing drugs byconsidering their toxicity is very important. PreADMET predicts mutagenicity and carcinogenicity of a compound, helping you to avoid toxic compound. The Ames test [19] TA 100 (+S9), TA1535 (+S9), TA1535 (−S9), and TA98 (+S9), TA98 (−S9) showed negative results. The carcinogenicity tests which showed positive, where it indicates a positive prediction, inferring that No evidence of carcinogenic activity.

55.4 Discussion

The Modeller results suggested that the model constructed is of good quality, and the RMSD value was 0.4 Ao and the Ramachandran plot showed that 94.8% residues were inside the most favorable region. From the Q-site finder, the most appropriate active site residues were retrieved. The autodock also showed that the binding energy values of ellagic acid (−6.58) and the receptor were high compared to those of the commercial drugs nateglinide (−4.93) & glimepiride (−5.8). The number of hydrogen bonds formed between the receptor and the ligand (ellagic acid) was comparatively high (5 bonds) than that in the commercial drugs,where only one bond was formed in each, respectively showing that the ligand ellagic acid & receptor protein interaction was high compared to that in the drugs. The ADME/T studies suggested that ellagic acid could be a possible lead molecule for diabetes. The safety assessment of ellagic acid as a food additive in F344 rats showed that no mortality or treatment-related clinical signs were observed throughout the experimental period [20].

55.5 Conclusion

Modern drug discovery is characterized by the production of vast quantities of compounds and the need to examine these huge libraries in short periods of time. CADD represents computational methods and resources that are used to facilitate the design and discovery of new therapeutic solutions. Digital repositories, containing detailed information on drugs and other useful compounds, are gold mines for

the study of chemical reactions' capabilities. The Modeling of the Potassium channel receptor and its macromolecular interaction studies with the drug compounds and ellagic acid showed that ellagic acid may possess antidiabetic property along with its other therapeutic uses.

References

1. L. Aguilar Bryan, J. Bryan. Molecular biology of adenosine triphosphate sensitive potassium channels. Endocr. Rev. 20(2):101 135
2. D. .K. Song, M. Fances. Ashcroft,,Glimepiride block of cloned b cell, cardiac and smooth muscle katp channels. Br. J. Pharmacol. 133:193 199, 2001.
3. M. Chachin, M. Vamada, A. Fujita, T. Matsuoka, K. Matsushita, Y. Kurachi. Nateglinide, a d phenylalanine derivative lacking either a sulfonylurea or benzamido moiety, specifically inhibits pancreatic cell type katp channels. J Pharmacol. Exp. Ther. 304(3)
4. H. S. Aiyer, M. V. Vadhanam, R. Stoyanova, G.D. Caprio, M. L. Clapper, R. C. Gupta. Dietary berries and ellagic acid prevent oxidative dna damage and modulate expression of dna repair genes. Int. J. Mol. Sci. 9:327 341, 2008.
5. N. Eswar, M. A. Marti renom, B. Webb, M. S. Madhusudhan, D. Eramian, M. Shen, U. Pieper, A. Sali. Comparative protein structure modeling with modeller. Current protocols in bioinfor matics. Wiley, New York, supplement 15, 5.6.1 5.6.30, 2006.
6. M. A. Marti renom, A. Stuart, A. Fiser, R. Sánchez, F. Melo, A. Sali. Comparative protein structure modeling of genes and genomes. Annu. Rev. Biophys. Biomol. Struct. 29:291 325, 2000.
7. A. Sali, T. L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234:779 815, 1993.
8. A. Fiser, R. K. Do, A. Sali. Modeling of loops in protein structures. Protein Sci 9:1753 1773, 2000.
9. Guillaume launay and thomas simonson. Homology modelling of protein protein complexes: a simple method and its possibilities and limitations. BMC Bioinformatics 9:427, 2008.
10. R. A. Laskowski, M. W. Macarthur, D. S. Moss, J. M. Thornton Procheck: a program to check the stereochemical quality of protein structures. J. Appl. Cryst. 26:283 291, 1993.
11. A. L. Morris, M. W. Macarthur, E. G. Hutchinson, J. M. Thornton. Stereo chemical quality of protein structure coordinates. Proteins 12:345 364, 1992.
12. Ilya n.shindyalov,philip e.bourne.protein structural alignment by incremental combinatorial extension(ce) of the optimal path. Protein Eng. 11(9):739 7747, 1998.
13. Sérgio filipe sousa, pedro alexandrino fernandes, maria joao ramos .protein ligand docking: current status and future challenges.
14. G. M. Morris, D. S. Goodsell, R. S. Halliday 2, R. Huey, W. E. Hart, R. K. Belew, J. Arthur. Olson Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function.
15. S. K. Lee, I. H. Lee, H. J. Kim, G. S. Chang, J. E. Chung, K. T. No. The PreADME Approach: Web based program for rapid prediction of physico chemical, drug absorption and drug like properties. EuroQSAR 2002 Designing drugs and crop protectants: processes, problems and solutions. Blackwell Publishing, Boston, MA, pp. 418 420, 2003.
16. T. Alasdair, R. Laurie, R. M. Jackson. Q SiteFinder: an energy based method for the predic tion of protein ligand binding sites
17. S. Miret et al. Comparison of in vitro models for the prediction of compound absorption across the human intestinal mucosa. J. Biomol. Screen. 9(7):598 606, 2004, DOI: 10.1177/ 1087057104267162.
18. C. A. Lipinski et al., Adv. Drug Del. Rev. 23:3, 1997, [PMID:11259830].

19. B. N. Ames, E. G. Gurney, J. A. Miller, H. Artsch. Carcinogens as frameshift mutagens: metabolites and derivatives of 2 acetylaminofluorene and other aromatic amine carcinogens (aromatic nitroso carcinogens/carcinogen detection with Salmonella/DNA intercalation)", Proc. Natl. Acad. Sci. USA 69(11):3128 3132.
20. M. Tasaki, T. Umemura, M. Maeda, Y. Ishii, T. Okamura, T. Inoue, Y. Kuroiwa, M. Hirose, A. Nishikawa. Safety assessment of ellagic acid, a food additive, in a subchronic toxicity study using F344 rats. Food Chem. Toxicol. 46(3):1119 24, 2008.

Chapter 56
GPU Acceleration of Dock6's Amber Scoring Computation

Hailong Yang, Qiongqiong Zhou, Bo Li, Yongjian Wang, Zhongzhi Luan, Depei Qian, and Hanlu Li

Abstract Dressing the problem of virtual screening is a long-term goal in the drug discovery field, which if properly solved, can significantly shorten new drugs' R&D cycle. The scoring functionality that evaluates the fitness of the docking result is one of the major challenges in virtual screening. In general, scoring functionality in docking requires a large amount of floating-point calculations, which usually takes several weeks or even months to be finished. This time-consuming procedure is unacceptable, especially when highly fatal and infectious virus arises such as SARS and H1N1, which forces the scoring task to be done in a limited time. This paper presents how to leverage the computational power of GPU to accelerate Dock6's (http://dock.compbio.ucsf.edu/DOCK 6/) Amber (J. Comput. Chem. 25: 1157 1174, 2004) scoring with NVIDIA CUDA (NVIDIA Corporation Technical Staff, Compute Unified Device Architecture Programming Guide, NVIDIA Corporation, 2008) (Compute Unified Device Architecture) platform. We also discuss many factors that will greatly influence the performance after porting the Amber scoring to GPU, including thread management, data transfer, and divergence hidden. Our experiments show that the GPU-accelerated Amber scoring achieves a 6.5× speedup with respect to the original version running on AMD dual-core CPU for the same problem size. This acceleration makes the Amber scoring more competitive and efficient for large-scale virtual screening problems.

Keywords Virtual screen · Dock · Amber scoring · GPU · CUDA

H. Yang (✉)
Department of Computer Science and Engineering, Sino German Joint Software Institute, Beihang University, 100191 Beijing, China
e mail: hailong.yang@jsi.buaa.edu.cn

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 56,

56.1 Introduction

One early stage of new drug discovery is focused on finding pharmacologically active compounds from the vast number of chemical compounds. In order to reduce the amount of wet-lab experiments, virtual screening is developed to search chemical compounds database for potentially effective compounds. Computer-assisted virtual drug screening is a definite shortcut to develop new drugs. It can reduce the amount of candidate compounds for biological experiments by thousands of times and is very prospective in exploiting new drug candidates [4].

Identifying the interactions between molecules is critical both to understanding the structure of the proteins and to discovering new drugs. Small molecules or segments of proteins whose structures are already known and stored in database are called ligands, while macromolecules or proteins associated with a disease are called receptors [5]. The final goal is to find out whether the given ligand and receptor can form a favorable complex and how appropriate the complex is, which may inhibit a disease's function and thus act as a drug.

Virtual screening can usually be roughly divided into two parts functionally:

- Routines determining the orientation of a ligand relative to the receptor, which are known as docking;
- Routines evaluating the orientation, which are known as scoring.

Docking is the first step in virtual screening that needs a potential site of interest on the receptor to be identified, which is also known as the active site. Within this site, points are identified where ligand atoms may be located. In dock6, a program routine called sphere center is to identify these points by generating spheres filling the site. To orient a ligand within the active site, some of the spheres are "paired" with ligand atoms, which are also called "matched" in docking.

Scoring is the step after docking, which is involved in evaluating the fitness for docked molecules and ranking them. A set of sphere-atom pairs will be on behalf of an orientation in receptor and evaluated with a scoring function on three-dimensional grids. At each grid point, interaction values are to be summed to form a final score. These processes need to be repeated for all possible translations and rotations. There are many kinds of existing scoring algorithms, while amber scoring is prevalent due to its fast speed and considerable high accuracy. The advantage of amber scoring is that both ligand and active sites of the receptor can be flexible during the evaluation, which allows small structural rearrangements to reproduce the so-called induced fit. While the disadvantage is also obvious, it brings tremendous intensive floating-point computations. When performing amber scoring, it calculates the internal energy of the ligand, receptor, and the complex, which can be broken down into three steps:

- minimization,
- molecule dynamics (MD) simulation,
- more minimization using solvents.

The computational complexity of amber scoring is very huge, especially in the MD simulation stage. Three grids that individually have three dimension coordinates are used to represent the molecule during the orientation such as geometry, gradient, and velocity. In each grid, at least 128 elements are required to sustain the accuracy of the final score. During the simulation, scores are calculated in three nested loops, each of which walks through one of the three grids. Derived from vast practical experiences, the MD simulation is supposed to be performed 3,000 times until a preferable result is obtained. All above commonly means that the problem size can reach as large as $3{,}000 \times 128^3 \times 128^3 \times 128^3$, which is computationally infeasible in one single computer.

While many virtual screening tools such as GasDock [6], FTDock [7], and Dock6 can utilize multi-CPUs to parallel the computations, the incapacity of CPU in processing floating-point computations still remains untouched. GPU has been widely used for general purpose computations because of its high floating-point computation capability [8]. Compared with CPU, GPU has the advantages of computational power and memory bandwidth. For example, a GeForce 9800 GT can reach 345 GFLOPS at peak rate and has an 86.4 GB/s memory bandwidth, whereas an Intel Core 2 Extreme quad core processor at 2.66 GHz has a theoretical 21.3 peak GFLOPS and 10.7 GB/s memory bandwidth. Another important factor why GPU is becoming widely used is that it is more cost effective than CPU.

Our contributions in this paper include porting the original Dock6 amber scoring to GPU using CUDA, which can archive a 6.5× speedup. We analyze the different memory access patterns in GPU which can lead to a significant divergence in performance. Discussions on how to hide the computation divergence on GPU are made. We also conduct experiments to see the performance improvement.

The rest of the paper is organized as follows. In Sect. 56.2, an overview of Dock6's amber scoring and analysis of the bottleneck is given. In Sect. 56.3, we present the main idea and implementation of the amber scoring on GPU with CUDA, and details of considerations about performance are made. Then we give the results, including performance comparisons among various GPU versions. Finally, we conclude with discussion and future work.

56.2 Analysis of the Amber Scoring in DOCK6

56.2.1 Overview

A primary design philosophy of amber scoring is allowing both the atoms in the ligand and the spheres in the receptor to be flexible during the virtual screening process, generating small structural rearrangements, which is much like the actual situation and gives more accuracy. As a result, a large number of docked orientations need to be analyzed and ranked in order to determine the best fit set of the matched atom-sphere pairs.

In the subsection, we will describe the amber scoring program flow and profile the performance bottleneck of the original amber scoring, which can be perfectly accelerated on GPU.

56.2.2 Program Flow and Performance Analysis

Figure 56.1 shows the steps to score the fitness for possible ligand receptor pairs in amber. The program first performs conjugate gradient minimization, MD simulation, and more minimization with solvation on the individual ligand, the individual receptor, and the ligand receptor complex, and then calculates the score as follows:

$$\text{Ebinding} = \text{Ecomplex} - (\text{Ereceptor} - \text{Eligand})$$

The docked molecules are described using three-dimension intensive grids containing the geometry, gradient, or velocity coordinate's information. The order of magnitude of these grids is usually very large. Data in these grids are represented using floating-point, which has little or no interactions during the computation.

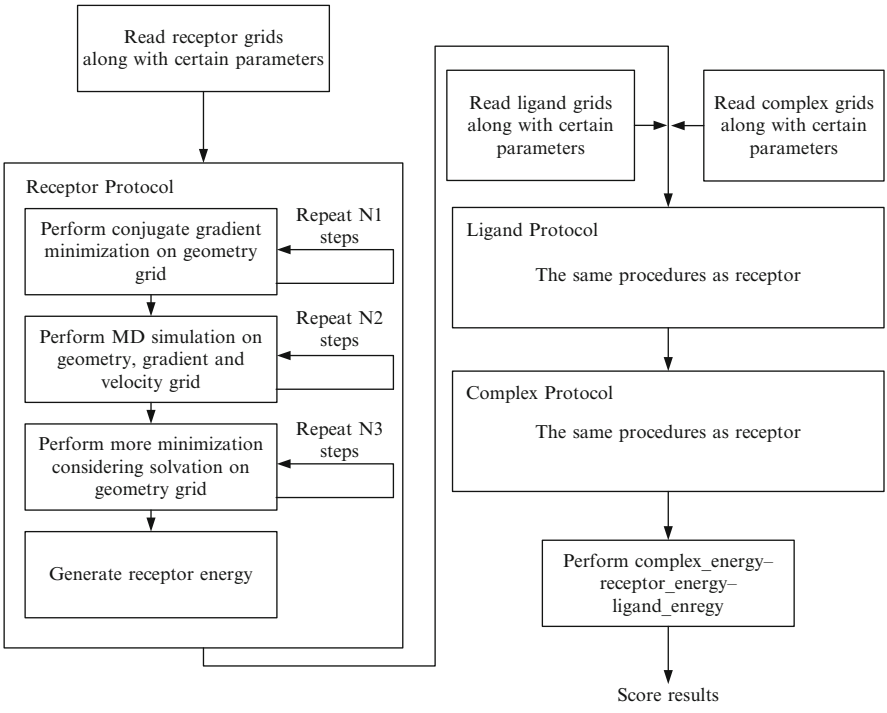


Fig. 56.1 Program flow of amber scoring

In order to archive higher accuracy, the scoring operation will be performed repeatedly, perhaps hundreds or thousands times.

Due to the characteristics of the amber scoring such as data independency and high arithmetic intensity, which are exactly the sweet spots of computing on GPU, it can be perfectly paralleled to leverage the computing power of GPU and gain preferable speedup.

56.3 Porting Amber Scoring to GPU

56.3.1 Overview

To determine the critical path of amber scoring, we conduct an experiment to make statistics about the cost of each step as shown in Table 56.1. We see that the time spent in processing a ligand is negligible, because ligand in docking always refers to small molecules or segments of protein whose information grids are small and can be calculated quickly. We also observe, however, that MD simulation on receptor and complex is the most time-consuming part, which takes up to 96.25% of the total time. Either on receptor or on complex, MD simulation performs the same functionality. Therefore, in our GPU-accelerated version, we focus on how to port the MD simulation to GPU, which could accelerate the bulk of the work.

For simplicity and efficiency, we take advantage of the Compute Unified Device Architecture (CUDA), a new parallel programming model that leverages the computational power in NVIDIA G80-based GPUs. We find that the key issue to utilize GPU fully is the high ratio of arithmetic operations to memory operations, which can be achieved through refined utilization of memory model, data transfer pattern, parallel thread management, and branch hidden.

Table 56.1 Runtime statistics for each step of amber scoring 100 cycles are performed for minimization steps and 3,000 cycles for md simulation step

Stage		Run time (s)	Ratio of total (%)
Receptor protocol	Gradient minimization	1.62	0.33
	MD simulation	226.41	45.49
	Minimization solvation	0.83	0.17
	Energy calculation	2.22	0.45
Ligand protocol	Gradient minimization	≈0	0
	MD simulation	0.31	0.06
	Minimization solvation	≈0	0
	Energy calculation	≈0	0
Complex protocol	Gradient minimization	8.69	1.75
	MD simulation	252.65	50.76
	Minimization solvation	2.69	0.54
	Energy calculation	2.22	0.45
Total		497.64	100

56.3.2 CUDA Programming Model Highlights

At the core of CUDA programming model are three key abstractions a hierarchy of thread groups, shared memories, and barrier synchronization, which provide fine-gained data parallelism, thread parallelism, and task parallelism. CUDA defines GPU as coprocessors to CPU that can run a large number of light-weight threads concurrently. The programming language of CUDA is a minimal set of C language extensions based on a few low learning carve abstractions for parallel computing. Threads are manipulated by kernels representing independent tasks mapped over each sub-problem. Kernels contain the processing assignments that each thread must carry out during the runtime. More specifically, same instruction sets are applied on different partitions of the original domain by the threads in SPMD fashion.

In order to process on the GPU, data should be prepared by copying it to the graphic board memory first. Actually, the problem domain is defined in the form of a 2D *grid* of 3D *thread blocks*. The significance of a thread block primitive is that it is the smallest granularity of work unit to be assigned to a single *streaming multiprocessor* (SM). Each SM is composed of eight *scalar processors* (SP) that indeed run threads on the hardware in a time-slice manner. Every 32 threads within a thread block are grouped into warps. At any time, there is only one warp active on the SM and it will proceed to run until it has to stop and wait for something to happen such as I/O operations. The hardware scheduler on the SM selects the next warp to execute. Within a warp the executions are in order, while beyond the warp the executions are out of order. Therefore, it does not matter if there are divergent threads among different warps. However, if threads within a warp follow divergent paths, only threads on the same path can be executed simultaneously.

In addition to global memory, each thread block has its own private shared memory that is only accessible to threads within that thread block. Threads within a thread block could cooperate by sharing data among shared memory with low latency. Among thread blocks, synchronization can be achieved by finishing a kernel and starting a new one. It is important to point out that the order of thread blocks assigned to the SMs is arbitrary and non-deterministic. Therefore, sequential semantics should not be fulfilled depending on the execution orders of the thread blocks, which may lead to race condition or deadlock.

56.3.3 Parallel Thread Management

To carry out the MD simulation on GPU, a kernel needs to be written, which is launched from the host (CPU) and executed on the device (GPU). A kernel is the same instruction set that will be performed by multiple threads simultaneously. This parallelism is implemented through the GPU hardware called Streaming Multi-processor (SM). By default, all the threads are distributed onto the same SM, which

cannot fully explore the computational power of the GPU or may cause launch failure if the threads are more than what one SM can hold. In order to utilize the SMs more efficiently, thread management must be taken into account.

We divide the threads into multiple blocks, and each block can hold the same number of threads. These blocks will be distributed among SMs. There are two kinds of IDs in CUDA named blockId and threadId, which are used to simplify the memory addressing among threads. Blocks have their own blockId during the kernel lifetime, and threads within the same block can be identified by threadId. Therefore, we can take advantage of these two kinds of IDs to issue threads computing on different partitions of the grids concurrently.

In the geometry, gradient, and velocity grids, 3D coordinates of atoms are stored sequentially and the size of the grid usually reaches as large as 7,000. Calculation works are assigned to blocks on different SMs; each thread within the blocks computes the energy of one atom, respectively, and is independent of the rest (see Fig. 56.2). We compose N threads into a block ($N = 512$ is the maximum number of threads per block in GeForce 9800 GT), which calculates N independent

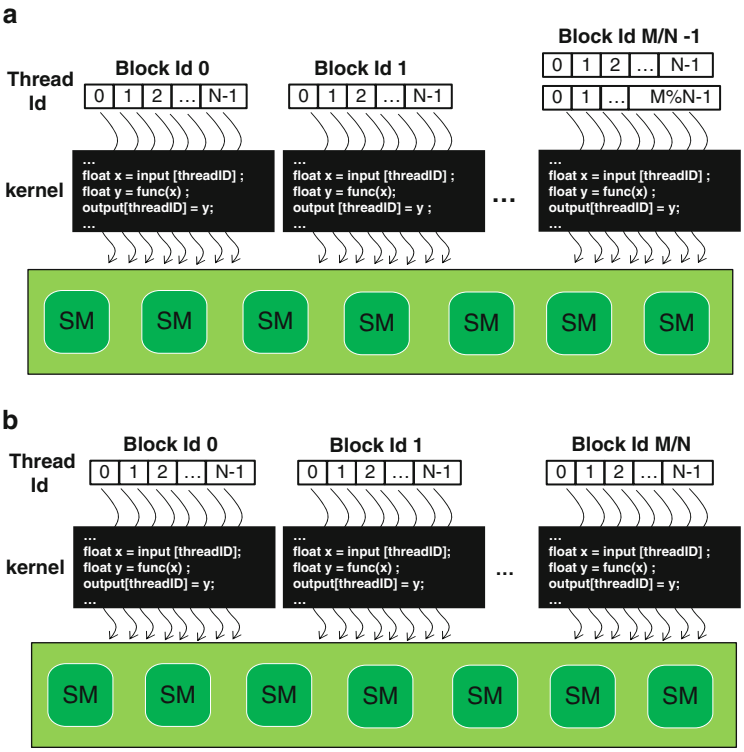


Fig. 56.2 Threads and blocks management about processing molecule grids on GPU: (**a**) blocks whose threads in the last block may calculate two atoms each (**b**) blocks whose threads in the last block may have nothing to do

atoms in the grids. Assuming the grid size is M and M is divisible by N, there will happen to be M/N blocks.

While in most cases the grid size M is not divisible by N, we designed two schemes dealing with this situation. In the first scheme, there will be M/N blocks. Since there is $M\%N$ atoms left without threads to calculate, we will rearrange the atoms evenly to the threads in the last block. One more atom will be added to the threads in the last block until no atoms are left, which is ordered by ascending thread ID. The second scheme is to construct $M/N + 1$ blocks. Each thread in the blocks still calculates one atom; however, the last block may contain threads with nothing to do.

Our experiment proves that the former scheme obtained better performance. This is caused by underutilized SM resources and branch cost in the second scheme. When there is a branch divergence, all the threads must wait until they reach the same instructions again. Synchronization instructions are generated by the CUDA complier automatically, which is time consuming. Furthermore, the redundant threads have to wait for all the calculations to be done, while they take up the SM computing cycles which cannot be utilized by other working threads and which cause a waste of resources.

56.3.4 Memory Model and Data Transfer Pattern

The first step to perform GPU computations is to transfer data from host (CPU) memory to device (GPU) memory since the receptor, ligand, and complex grids need to be accessible by all SMs during the calculations. There are two kinds of memory that can be used to hold these grids. One is the constant memory, which can be read and written by the host but can only be read by the device. The other is the global memory, which can be read and written by both the host and device. One important distinction between the two memories is the access latency. SMs can get access to the constant memory magnitude order faster than to the global memory. While the disadvantage of the constant memory is also obvious, it is much smaller, which is usually 64 KB compared to 512 MB global memory. Thus, a trade-off has to be made on how to store these grids.

During each MD simulation cycle, the gradient and velocity grids are read and updated. Therefore, they should be stored in global memory. While once entered in the MD simulation process, the geometry grids are never changed by the kernel. Hence, they can be stored in constant memory (see Fig. 56.3). Considering the out-bound danger due to the limited capacity of the constant memory, we observed the size of each geometry grid. The receptor and complex geometry grids usually contain no more than 2,000 atoms each, while the ligand geometry grid contains 700 atoms, which totally requires $2{,}000 \times 3 \times 4 \times 2 + 700 \times 3 \times 4$ bytes (56.4 KB) memory to store them. Since it is smaller than 64 KB, the geometry grids shall never go out-bound of the constant memory.

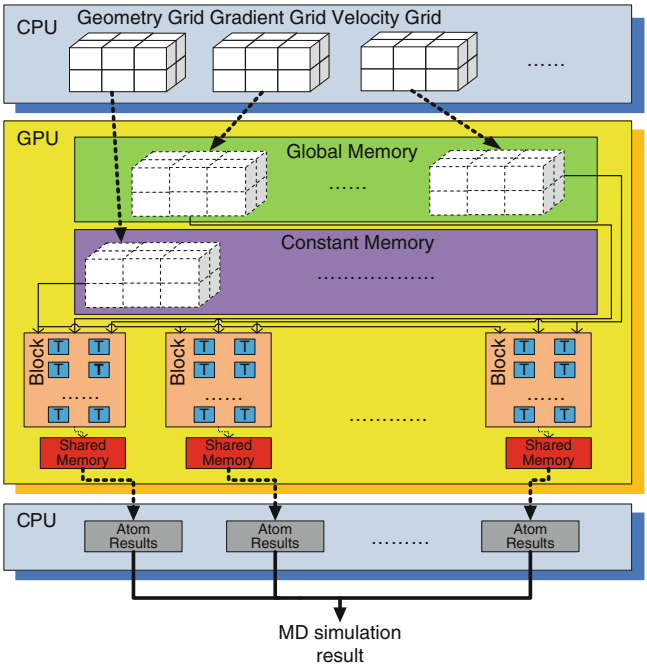


Fig. 56.3 Memory model and data transfer pattern during the MD simulation cycles. Grids are transferred only once before the simulation. Atom results are first accumulated in the shared memory within the block. Then the accumulations per block are transferred into the host memory and summed up

The time to transfer molecule grids from the host to their corresponding GPU memory is likewise a critical issue, which may degrade the benefit archived from the parallel execution if not considered carefully [9]. For each MD simulation cycle, we could transfer one single atom 3D coordinates in the geometry, gradient, and velocity grids to device memory when they are required by the SMs. The other solution is to transfer the entire grids into the GPU memory before the MD simulation stage. When the simulation starts, these grids are already stored in device memory, which can be accessed by simulation cycles performed on SMs.

Based on our experiment, we noticed that there was a significant performance divergence between the two schemes. The former version turned out to be not speedup but obvious slowdown. Before data transformation, a certain time is required to get the PCIE bus and device ready. The data representing one atom only takes up 12 bytes, which is very tiny compared to the 8-GB bandwidth of the PCIE bus. Thus the time spent on real data transformation can be neglected, and most of the time is wasted in frequent bus communications.

Significant performance improvements are obtained from the second scheme since the molecule grids are transferred only once for all before the MD simulation. Therefore, the SMs donot have to halt and wait for the grids to be prepared.

Generally, at least 3,000 MD simulation cycles are required for each molecular stage to maintain accuracy, which means highly intensive floating-point calculations are performed on the same molecule grids with updated values in each atom. Thus, the ratio of memory access and floating-point calculations should be pretty high, which obviously speeds up the parallel execution of the MD simulations by fully utilizing the SMs.

The MD simulations are executed parallelly on different SMs, and threads within the different blocks are responsible for the calculation of their assigned atoms of the grids. The traditional approach is to transfer all the atom results back to the host memory where the accumulation is performed. In practice, this may be inefficient since shared memory in GPU can be utilized to cut down the communication cost between the device and host. However, the limitation is that synchronization can only be applied within the block (see Fig. 56.3). Our solution is to synchronize threads within the blocks, which generates atom results separately. Then a transformation is performed to store the atom results from shared memory to host memory in a result array. The molecule result shall be achieved by adding up all the elements in the array without synchronization. The experiment has showed that this solution has a significant impact on performance improvement as the simulation size scales compared to the original approach without shared memory synchronization.

56.3.5 Divergence Hidden

Another important factor that dramatically impacts the benefits achieved by performing MD simulation on GPU is the branches. Original MD simulation procedure involves a bunch of nested control logics such as bonds of Van der Waals force and constrains of molecule energy. When the parallel threads computing on different atoms in the grids come to a divergence, a barrier will be generated and all the threads will wait until they reach the same instruction set again. The above situation can be time consuming and outweigh the benefits of parallel execution; thus divergence must be hidden to the minimum.

We extract the calculations out of the control logic. Each branch result of the atom calculation is stored in a register variable. Inside the nested control logic, only value assignments are performed, which means the divergence among all the threads will be much smaller; thus the same instruction sets can be reached with no extra calculation latency. Although this scheme will waste some computational power of the SMs since only a few branch results are useful in the end, it brings tremendous improvements in performance. These improvements can be attributed to that, in most cases, the computational power we required during the MD simulation is much less than the maximum capacity of the SMs. Hence, the extra calculations only consume vacant resources, which in turn speed up the executions. The feasibility and efficiency of our scheme have been demonstrated in our experiment.

56.4 Results

The performance of our acceleration result is evaluated for two configurations:

- Two cores of a dual core CPU
- GPU accelerated.

The base system is a 2.7-GHz dual core AMD Athlon processor. GPU results were generated using an NVIDIA GeForce 9800 GT GPU card.

We referred to the Dockv6.2 as the original code, which was somewhat optimized in amber scoring. We also used the CUDAv2.1, whose specifications support 512 threads per block, 64 KB constant memory, 16 KB shared memory, and 512 MB global memory. Since double precision floating point was not supported in our GPU card, transformation to single precision floating point was performed before the kernel was launched. With small precision losses, the amber scoring results were slightly different between CPU version and GPU version, which can be acceptable.

Table 56.2 compares the original CPU version with the GPU-accelerated version in runtime for various stages. The MD simulation performed comprises 3,000 cycles in each molecular stage, which clearly afford very high speedups due to the utilizations of multi-blocks, one time data transfer pattern, shared memory and divergence hidden. The overall speedup achieved for the entire amber scoring is over 6.5×. One interesting phenomenon we noticed is not all the minimizations are speeded up but ligand is slowed down. We find a reasonable explanation that the ligand is generally a small molecule requiring a negligible amount of floating-point calculations. When mapped on GPU, these calculations are insufficient to hide the latency of data transformation and the time consumed to initial the device.

Figure 56.4 depicts the total speedups of different GPU schemes with respect to the range of increasing MD simulation cycles. As mentioned in Sect. 56.3.3, the GPU version with only one block did not speedup the process of MD simulation but slowed

Table 56.2 CPU times, GPU times, and speedups with respect to 3,000 MD simulation cycles per molecule protocol

Stage		CPU	GPU	Speedup
Receptor protocol	Gradient minimization	1.62	0.89	1.82
	MD simulation	226.41	31.32	7.23
	Minimization solvation	0.83	0.15	5.53
	Energy calculation	2.22	1.21	1.83
Ligand protocol	Gradient minimization	≈0	0.02	
	MD simulation	0.31	0.60	
	Minimization solvation	≈0	0.03	
	Energy calculation	≈0	≈0	0
Complex protocol	Gradient minimization	8.69	2.88	3.02
	MD simulation	252.65	34.79	7.26
	Minimization solvation	2.69	2.05	1.31
	Energy calculation	2.22	1.47	1.51
Total		497.64	75.41	6.5

The CPU version was performed using dual core, while GPU version with all superior scheme

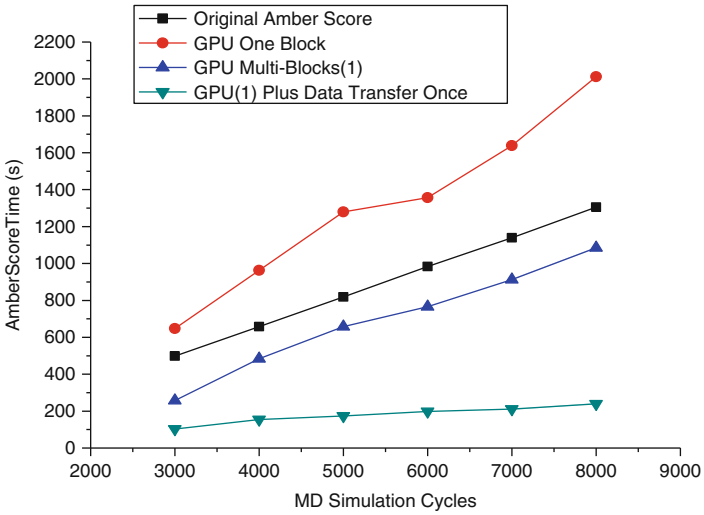


Fig. 56.4 Shown is a comparison of amber scoring time between original amber and different GPU versions whose speedup varies significantly as the MD simulation cycles increase from 3,000 to 8,000

down, which should be attributed to the poor management of threads since each block has a boundary of maximum active threads. The most significant performance improvements are achieved by transferring the molecule grids only once during the MD simulation in addition to the usage of multi-blocks. This scheme can greatly diminish the overhead produced by duplicate transferring the molecule grids from CPU to GPU, which dominates the time consumed during the MD simulation.

Figure 56.5 depicts the second speedup in performance obtained from the utilizations of divergence hidden and synchronization on shared memory. Since the branch calculations are extracted from the control logic and stored in temporary variables, only one single instruction will be performed which assigns corresponding values into the final result when divergences occur. While threads within a block will accumulate atom simulation values into a partial result of a molecule on shared memory, the result array transferred back to the host is very small. Performance improvements are obtained when summing up the elements in the array to form the molecule simulation result. We also notice that as the MD simulation cycles scale up, the speedup becomes more considerable in our best GPU version.

56.5 Related Work

Studies on utilizing GPU to accelerate molecule docking and scoring problems are rare, the only work that we find more related to our concern is in the paper of Bharat Sukhwani [10]. The author described a GPU-accelerated production docking code,

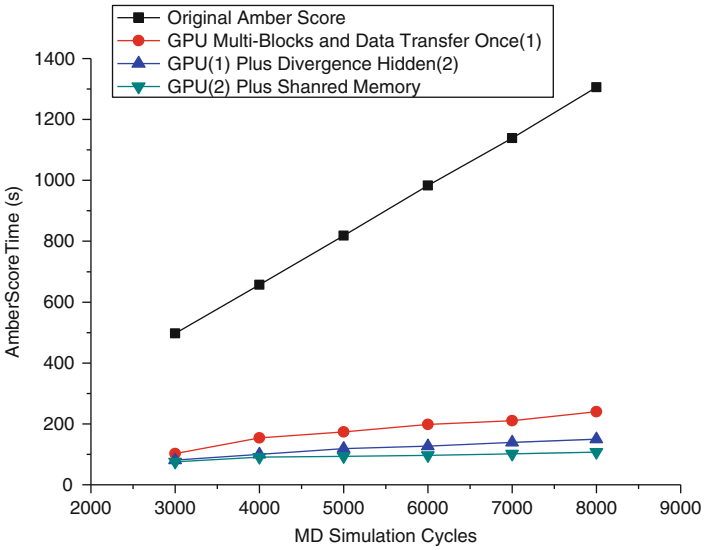


Fig. 56.5 Comparison of speedups among different GPU versions based on Fig. 56.4 in addition to divergence hidden and shared memory

PIPER [11], which achieves an end-to-end speedup of at least 17.7× with respect to a single core. Our contribution is different from the former study in two aspects. First, we focus our energy on flexible docking such as amber scoring, while the previous study mainly focuses on rigid docking using FFT. Thus our work is more complex and competitive in the real world. Second, we noticed that the logic branches in the parallel threads on GPU degraded the entire performance sharply. We also described the divergence hidden scheme and represented the comparison on speedup with and without our scheme.

Another attractive work that needs to be mentioned is that by Michael Showerman and Jeremy Enos [12] in which they developed and deployed a heterogeneous multi-accelerator cluster at NCSA. They also migrated some existing legacy codes to this system and measured the performance speedup, such as the famous molecular dynamics code called NAMD [13, 14]. However, the overall speedup they achieved was limited to 5.5×.

56.6 Conclusions and Future Works

In this paper, we present a GPU-accelerated amber score in Dock6.2, which achieves an end-to-end speedup of at least 6.5× with respect to 3,000 cycles during MD simulation compared to that of a dual core CPU. We find that thread management utilizing multiple blocks and single transferring of the molecule grids

dominates the performance improvements on GPU. Furthermore, dealing with the latency attributed to thread synchronization, divergence hidden, and shared memory can lead to elegant solutions, which will additionally double the speedup of the MD simulation. Unfortunately, the speedup of Amber scoring cannot go much higher due to Amdahl's law. The limitations are as follows:

- With the kernel running faster because of GPU acceleration, the rest of the Amber scoring takes a higher percentage of the run time.
- Partitioning the work among SMs will eventually decrease the individual job size to a point where the overhead of initializing an SP dominates the application execution time.

The work we presented in this paper only shows a kick-off stage of our exploration in GPGPU computation. We will proceed to use CUDA acceleration various applications with different data structures and memory access patterns, and hope to be able to work out general strategies about how to use GPU more efficiently. With greater divergences in architectural designs of CPU and GPU, our goal is to find a parallel programming model to leverage the computation power of CPU and GPU simultaneously.

Acknowledgment Many thanks to Ting Chen for thoughtful discussions and comments about our implementation and paper work. This work was supported by the National High Technology Research and Development Program of China under the grant No. 2007AA01A127.

References

1. Dock6: http://dock.compbio.ucsf.edu/DOCK_6/.
2. Wang, J., Wolf, R.M., Caldwell, J.W., Kollman, P.A. and Case, D.A. Development and testing of a general Amber force field. Journal of Computational Chemistry 25, Pages: 1157 1174, 2004.
3. NVIDIA Corporation Technical Staff, Compute Unified Device Architecture Programming Guide, NVIDIA Corporation, 2008.
4. Sukhwani, B. and Herbordt, M. Acceleration of a production rigid molecule docking code. In Proceedings of the IEEE Conference on Field Programmable Logic and Applications Pages: 341 346, 2008.
5. Kuntz, I., Blaney, J., Oatley, S., Langridge, R. and Ferrin, T. A geometric approach to macromolecule ligand interactions. Journal of Molecular Biology 161, Pages: 269 288, 1982.
6. Honglin Lia, Chunlian Lia, Chunshan Guib, Xiaomin Luob and Hualiang Jiangb. GAsDock: a new approach for rapid flexible docking based on an improved multi population genetic algorithm. Bioorganic & Medicinal Chemistry Letters 14(18), Pages: 4671 4676, 2004.
7. Servat, H., Gonzalez, C., Aguilar, X., Cabrera, D. and Jimenez, D. Drug design on the cell broadband engine. Parallel Architecture and Compilation Techniques, Pages: 16:425 425, 2007.
8. Krüger, J., Westermann, R. Linear algebra operators for GPU implementation of numerical algorithms. ACM Transactions on Graphics 22(3) Pages: 908 916, 2003.
9. Govindaraju, N.K., Gray, J., Kumar, R. and Manocha, D. GPUTeraSort: High performance graphics coprocessor sorting for large database management. Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data.

10. Bharat Sukhwani and Martin C. Herbordt. GPU acceleration of a production molecular docking code. In Proceedings of 2nd Workshop on General Purpose Processing on GPUs, Pages: 19 27, 2009.
11. PIPER: http://structure.bu.edu/index.html
12. Michael Showerman, Wen Mei Hwu, Jeremy Enos, Avneesh Pant, Volodymyr Kindratenko, Craig Steffen and Robert Pennington. QP: A Heterogeneous Multi Accelerator Cluster. In 10th LCI International Conference on High Performance Clustered Computing, 2009.
13. NAMD: http://www.ks.uiuc.edu/Research/namd/.
14. James C. Phillips, Gengbin Zheng, Sameer Kumar and Laxmikant V. Kalé. NAMD: Biomo lecular Simulation on Thousands of Processors, Conference on High Performance Networking and Computing, Pages: 1 18, 2002.

Part VI
Computational Methods and Diagnostic Tools in Biomedical

Chapter 57
Discriminative Distance Functions and the Patient Neighborhood Graph for Clinical Decision Support

Alexey Tsymbal, Martin Huber, and Shaohua Kevin Zhou

Abstract There are two essential reasons for the slow progress in the acceptance of clinical similarity search-based decision support systems (DSSs); the especial complexity of biomedical data making it difficult to define a meaningful and effective distance function and the lack of transparency and explanation ability in many existing DSSs. In this chapter, we address these two problems by introducing a novel technique for visualizing patient similarity with neighborhood graphs and by considering two techniques for learning discriminative distance functions. We present an experimental study and discuss our implementation of similarity visualization within a clinical DSS.

Keywords Clinical Decision Support System · Similarity Search · Distance Learning · Neighborhood Graph · Case Retrieval

57.1 Introduction

There is growing interest in the use of clinical decision support systems (DSSs) to reduce medical errors and to increase health care quality and efficiency [2]. Clinical DSSs vary greatly in design, functionality, and use. According to the reasoning method used in clinical DSSs, one important subclass is that of Case-Based Reasoning (CBR) systems systems which have reasoning by similarity as the central element of decision support [2, 8].

One reason for the slow acceptance of CBR systems in biomedical practice is the especial complexity of clinical data and the resulting difficulty in defining a

A. Tsymbal (✉)
Corporate Technology Div.,, Siemens AG, Erlangen, Germany
e mail: alexey.tsymbal@siemens.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 57,

meaningful distance function and adapting the final solution [10]. Another commonly reported reason is the lack of transparency and explanation in clinical CBR.

To solve the problems described above, we introduce a novel technique for visualizing patient similarity, which is the central concept in any clinical CBR system. Moreover, we consider two techniques for learning discriminative distance functions, which when used in combination with the neighborhood graph, can make it a powerful and flexible tool for clinical knowledge discovery and decision support in different classification contexts. We present an experimental study with the two distance learning techniques and discuss our solutions in the implementation of presented distance learning and similarity visualization techniques within the DSS Health-e-Child CaseReasoner, which we develop in the framework of the EU-funded project Health-e-Child.

This chapter is organized as follows. Section 57.2 gives a review of basic concepts and related work in discriminative distance learning, then introduces our experimental framework, and presents most important results with a selection of biomedical data sets. In Sect. 57.3, a technique for visualizing patient similarity based on neighborhood graphs is introduced and our implementation of it is discussed. We finish in Sect. 57.4 with a brief summary.

57.2 Discriminative Distance Learning

57.2.1 Related Work

While historically research on distance learning has started from supervised learning of distance functions for nearest neighbor classification in the original "feature vector-class label" representation, today by far the most commonly used representation in this context is the one based on so called *equivalence constraints* [5].

Usually, equivalence constraints are represented using triplets (x_1, x_2, y), where x_1, x_2 are data points in the original space and $y \in \{+1, \ \ 1\}$ is a label indicating whether the two points are similar (from the same class) or dissimilar. Learning from these triples is often called learning in the *product space*. An alternative is to learn in the *difference space*, the space of vector differences. The difference space is normally used with homogeneous high-dimensional data, such as pixel intensities or their Principal Component Analysis (PCA) coefficients in imaging.

There are two essential reasons that motivate the use of equivalence constraints in learning distance functions; their natural availability in some learning contexts and the fact that they are a natural input for optimal distance function learning [1]. It can be shown that the optimal distance function for classification is of the form $p(y_i \neq y_j | x_i, x_j)$. Under the *i.i.d.* assumption the optimal distance measure can be expressed in terms of generative models $p(x/y)$ for each class as follows [7]:

$$p(y_i \neq y_j | x_i, x_j) = \sum\nolimits_y p(y|x_i)(1 - p(y|x_j)) \tag{57.1}$$

For a Random Forest (RF) learnt for a certain classification task, the proportion of the trees, where two cases appear together in the same leaves can be used as a measure of their similarity [4]. The cases are propagated down all K trees within the forest and their terminal positions z in each tree are recorded. The similarity between two cases is then equal to (I is the indicator function):

$$S(x_1, x_2) = \frac{1}{K} \sum_{i=1}^{K} I(z_{1i} = z_{2i}) \tag{57.2}$$

Similarity (2) is successfully used in [11] for hierarchical clustering of tissue microarray data. It is interesting that using this similarity for the most immediate task, nearest neighbor classification, is rather uncommon, comparing to its use for clustering. In one of the related works, [9] use it for protein protein interaction prediction, and the results compare favorably with all previously suggested methods.

57.2.2 Distance Learning: Experimental Framework

Data under study include four clinical data sets; one data set with cardiac aortic valve meshes, *Meshes*, including 63 3D meshes representing healthy and diseased valves, and three data sets from the UCI repository, *Liver*, *Thyroid,* and *Heart* [3]. The number of features varies from 5 in Thyroid and 6 in Liver to 13 in Heart and 6,000 in Meshes. For more details about the valve data the reader is referred to [6]. Moreover, experiments were conducted on five other data sets, four gene expression samples (*Lymphoma*, *Embryonal_Tumors*, *Colon*, and *Leukemia*)[1], and one mass spectrometry data set for cancer identification from the NIPS 2003 challenge (*Arcene*). The data sets represent binary classification tasks and the number of features varies from 2,000 in Colon to 10,000 in Arcene.

Leave-one-out cross validation was used with the Meshes data and with all genetic data sets (Lymphoma, Embr Tumors, Colon, and Leukemia), and three runs of tenfold cross-validation were used with the rest. The distance functions are evaluated based on the accuracy of k-nearest neighbor classification (*k-NN)*, with $k = 7$, and weighting the votes of the neighbors inversely proportional to the distance. All ensemble models (*AdaBoost* and *RF*) included 50 component trees.

Among the considered algorithms, only RF-based learning algorithms are able to successfully deal with high-dimensional data, such as the four gene expression data sets, Arcene and Meshes in their raw form, as long as RFs incorporate an explicit feature selection process at each node in the training phase. In order to get rid of this limitation and put each algorithm in the same initial conditions, at each

[1]The microarray gene expression data sets were taken from and are available at www.upo.es/eps/aguilar/datasets.html.

cross-validation run, 200 features were preselected according to their Information Gain, for all experiments with the high-dimensional data sets.

The learning algorithms mentioned above and the experimental framework for our study were implemented on the basis of machine learning library *WEKA 3.4* [14]. All learning algorithms used default settings besides the ones already mentioned.

57.2.3 Distance Learning: Experimental Results

Table 57.1 includes classification accuracies for plain learning, such as *k-NN*, AdaBoost (*AB*), and *RF*; for *k*-NN with distance learning, with AB and RF in product and difference spaces (*AB-prod*, *AB-diff*, *RF-prod*, and *RF-diff*); and for *k*-NN with the intrinsic RF distance (*RF_dist*); for the nine data sets and on average. The best accuracies for every data set are given in bold, and statistically significant differences according to the McNemar's test (this happens for the Embr tumors and Arcene data sets only) are given in italic. From the table, one may see that *k*-NN with learning distance functions results in accuracies no worse or better than the plain techniques. These results are in line with conclusions made in similar previously reported studies.

There is no clear advantage of neither product nor difference space, same for AB and RF. Selection of a proper space and learner seems to be context-specific and validation in advance can be recommended, given that there is enough data. Theoretically, representation with the product space is richer, however, it is easier to be overfitted; one cure for that is the difference space, although its representative ability is not good enough for some tasks. A little surprising observation is the performance of the intrinsic RF distance. It exhibits the most robust behavior, leading to the best average accuracy.

To better understand the two types of distance learning and to validate the hypothesis regarding the higher sensitivity to overfitting for equivalence constraints, we have conducted a separate series of experiments with RFs with different numbers of trees; 3, 7, 15, 31, and 63 (as a series of powers of 2 1).

Table 57.1 Classification accuracies for plain learners and learning discriminative distance

Data set	*k* NN	AB	RF	AB prod	RF prod	AB diff	RF diff	RF dist
Mesh	0.889	0.905	0.905	0.921	0.937	0.905	0.905	0.921
Lymphoma	0.978	0.911	0.956	1	1	1	1	0.978
Embr Tumors	0.717	0.733	0.767	0.767	0.783	0.750	0.75	0.800
Colon	0.790	0.839	0.855	0.855	0.871	0.871	0.871	0.871
Leukemia	0.944	0.931	0.958	0.958	0.972	0.972	0.972	0.972
Arcene	0.837	0.873	0.858	0.865	0.863	0.862	0.880	0.898
Liver	0.646	0.703	0.700	0.721	0.711	0.658	0.706	0.698
Thyroid	0.924	0.934	0.944	0.967	0.947	0.964	0.962	0.969
Heart	0.818	0.816	0.811	0.811	0.821	0.824	0.824	0.829
Average	0.838	0.849	0.861	0.874	0.878	0.867	0.874	0.882

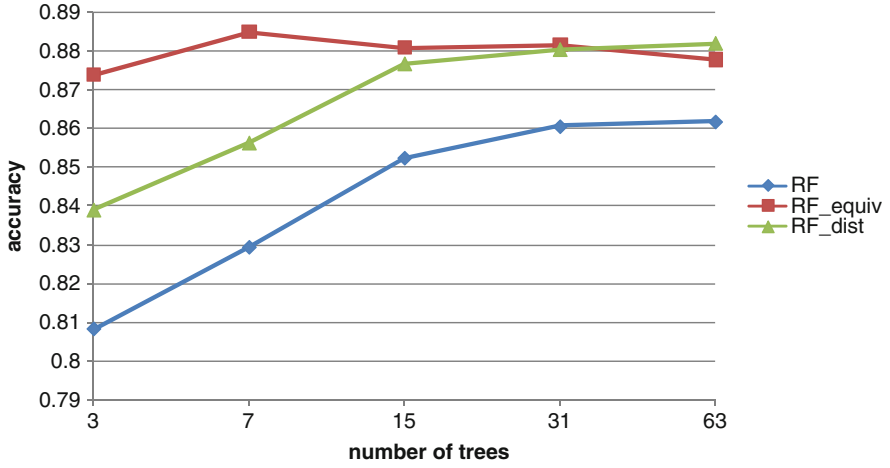


Fig. 57.1 Classification accuracy for RF, RF with equivalence constraints (*RF equiv*), and the intrinsic RF distance (*RF dist*), for different numbers of component trees

In Fig. 57.1, classification accuracy is shown for plain RF, RF with equivalence constraints (*RF_equiv*), and the intrinsic RF distance (*RF_dist*), for different numbers of component trees, averaged over the nine data sets. For *RF_equiv*, the best space is preselected for every data set to represent equivalence constraints.

The behavior of plain RF and intrinsic RF distance is not surprising; the accuracy plateaus with the increase in the number of trees. The behavior of learning from equivalence constraints is rather unexpected. The peak of accuracy is achieved already with seven trees and then accuracy decreases slightly, supporting our hypothesis regarding the greater risk of overfitting with equivalence constraints. A similar phenomenon was observed with AB as well.

57.3 Neighborhood Graph for Visualizing Patient Similarity

The neighborhood graph provides an intuitive way of visualizing patient similarity with a node-link entity-relationship representation. There can be distinguished three basic types of neighborhood graphs that can be used to visualize object proximity in DSSs [13]; (1) relative neighborhood graph (RNG), (2) distance threshold graph, and (3) k-NN graph. These graphs are studied and applied in different contexts; in particular, as data visualization tools the threshold and k-NN graphs are often used for the analysis of gene expression data in bioinformatics.

In a *relative neighborhood graph*, two vertices corresponding to two cases A and B in a data set are connected with an edge, if there is no other case C which is closer to both A and B with respect to a certain distance function d [12]. An important benefit of RNG comparing to the other two is the fact that it is always connected

with nodes having a reasonable small degree; it is often planar or close to planar. RNG is an expressive way of presenting relevant information. From the graph, one may easily comprehend patient distribution according to the studied similarity context and see patient groupings, identify outliers, easy to classify cases and the borderline cases classification for which is likely to be uncertain.

In our toolbox for visualization, navigation and management of the three neighborhood graphs, which are being developed as a part of the clinical DSS Case-Reasoner, we implemented the following functionality:

- node coloring, to represent numeric and nominal features;
- node filtering, according to feature values in the patient record;
- edge coloring and filtering, according to the underlying distance;
- graph (hierarchical) clustering into an arbitrary number of components;
- reconfigurable tooltips displaying clinical data and images;
- nearest neighbor classification and regression performance visualization for each node, for a selected class feature and the given similarity context;
- image visualization within the nodes of the graph.

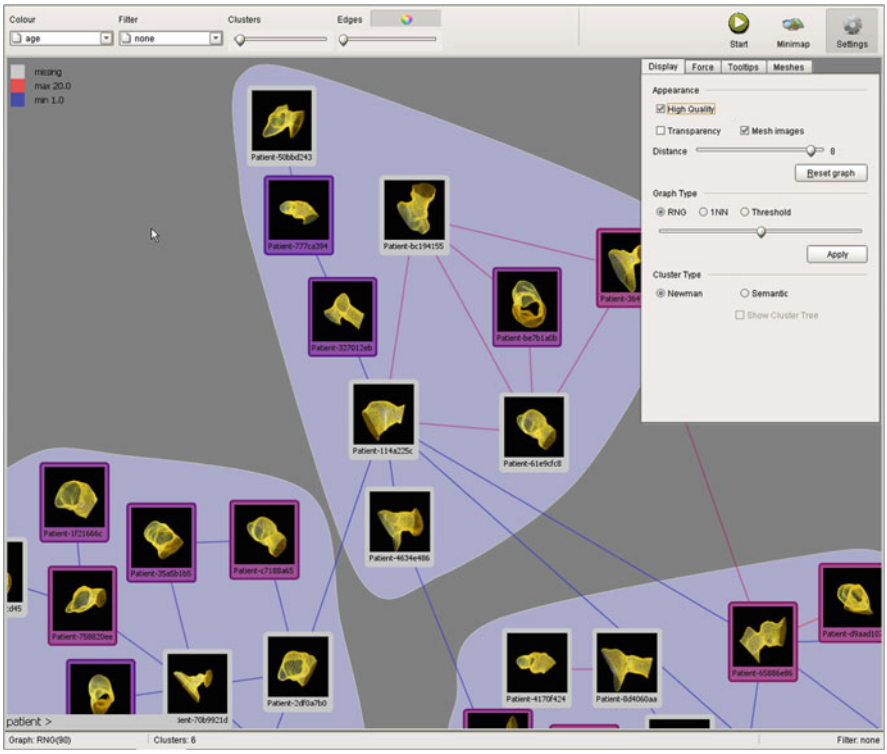


Fig. 57.2 Toolbox GUI and image visualization within the neighborhood graph

Besides clinical data and patient similarities, the neighborhood graphs are nicely suitable for displaying images corresponding to patients. While the existing radiology software normally visualizes a set of images corresponding to a single patient, in this visualization images corresponding to a cohort of patients can be displayed together with their relationship.

In Fig. 57.2, meshes corresponding to the pulmonary trunks of the patients are displayed within the nodes of RNG displaying a cohort of Health-e-Child cardiac patients, and a sketch of GUI of the neighorhood graph module is shown, including the toolbar with access to the basic operations on the graph, a pop-up graph settings control panel, and a status bar with basic information about the currently displayed graph.

57.4 Conclusions

In this chapter, two techniques for learning discriminative distance were considered; learning from equivalence constraints and the intrinsic RF distance. The techniques are different in their nature but have an important commonality that they are *discriminative* distance functions, or functions optimized for retrieval in a certain classification context. Our experiments confirm that both techniques are competitive with respect to plain learning; they help to combine the power of strong learning algorithms with the transparency of case retrieval. Future work includes studying various applications to other subject domains with complex data. Moreover, we introduced a novel technique for visualizing patient similarity with neighborhood graphs, which, especially when used in combination with distance learning, can be helpful in clinical knowledge discovery and decision support.

Acknowledgments This work has been partially funded by the EU project Health e Child (IST 2004 027749).

References

1. Bar Hillel A (2006) Learning from weak representations using distance functions and gener ative models. PhD Thesis, The Hebrew University of Jerusalem
2. Berlin A, Sorani M, Sim I (2006) A taxonomic description of computer based clinical decision support systems. Journal of Biomedical Informatics 39 (6): 656 667
3. Blake CL, Keogh E, Merz CJ (1999) UCI repository of machine learning databases. Depart ment of Information and Computer Science, University of California at Irvine
4. Breiman L (2001) Random forests. Machine Learning 45(1):5 32
5. Hertz T (2006) Learning distance functions: algorithms and applications. PhD Thesis, The Hebrew University of Jerusalem
6. Ionasec RI, Tsymbal A, Vitanovski D et al (2009) Shape based diagnosis of the aortic valve. In: Proc Int Conf SPIE Medical Imaging *SPIE Medical Imaging*

7. Mahamud S, Hebert M (2003) The optimal distance measure for object detection. In: Proc Int Conf Computer Vision and Pattern Recognition CVPR'03
8. Nilsson M, Sollenborn M (2004) Advancements and trends in medical case based reasoning: an overview of systems and system development. In: Proc Int FLAIRS Conf. on AI
9. Qi Y, Klein Seetharaman J, Bar Joseph Z (2005). Random Forest similarity for protein protein interaction prediction from multiple sources. In: Proc Pacific Symposium on Biocomputing
10. Schmidt R, Vorobieva O (2005) Adaptation and medical case based reasoning focusing on endocrine therapy support. In: Proc Int Conf AI in Medicine, LNCS 3581, Springer
11. Shi T, Horvath S (2006) Unsupervised learning with Random Forest predictors. Computa tional and Graphical Statistics 15(1):118 138
12. Toussaint GT (1980) The relative neighborhood graph of a finite planar set. Pattern Recogni tion 12(4):261 268
13. Tsymbal A, Zhou SK, Huber M (2009) Neighborhood graph and learning discriminative distance functions for clinical decision support. In: Proc Annual Int Conf of IEEE Engineering in Medicine and Biology Society EMBC'09
14. Witten I, Frank E (2005) Data mining: practical machine learning tools with Java implemen tations. Morgan Kaufmann, San Francisco

Chapter 58
A Scalable and Integrative System for Pathway Bioinformatics and Systems Biology

Behnam Compani, Trent Su, Ivan Chang, Jianlin Cheng, Kandarp H. Shah, Thomas Whisenant, Yimeng Dou, Adriel Bergmann, Raymond Cheong, Barbara Wold, Lee Bardwell, Andre Levchenko, Pierre Baldi, and Eric Mjolsness

Abstract Motivation: Progress in systems biology depends on developing scalable informatics tools to predictively model, visualize, and flexibly store information about complex biological systems. Scalability of these tools, as well as their ability to integrate within larger frameworks of evolving tools, is critical to address the multi-scale and size complexity of biological systems.

Results: Using current software technology, such as self-generation of database and object code from UML schemas, facilitates rapid updating of a scalable expert assistance system for modeling biological pathways. Distribution of key components along with connectivity to external data sources and analysis tools is achieved via a web service interface.

Availability: All sigmoid modeling software components and supplementary information are available through: http://www.igb.uci.edu/servers/sb.html.

Keywords Bioinformatics · Biosynthetic · Database · Metabolic · Modeling · Signal transduction · Simulation · Systems biology

58.1 Introduction

Here, we describe Sigmoid, a generative, scalable software infrastructure for systems biology designed to facilitate global modeling of biological systems. If deciphered as an acronym, SIGMOID would translate to SIGnal MOdeling Interface and Database. Here, the term Signal, in a biological sense, would be broadly interpreted. Sigmoid supports the process of cycling between model

E. Mjolsness (✉)
Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697, USA
School of Information and Computer Sciences, University of California, Irvine, CA 92697, USA
e mail: emj@ics.uci.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 58,

building, hypothesis generation, biological experimentation, and data gathering, by integrating the hypothesis and discovery phases of the research process.

In Sigmoid, we address the problem of creating a scalable expert assistance system for modeling biological pathways, using current software technology to decrease the difficulty and cost of building the system. The reason for building such a system is to provide computational support to biologists and computational scientists who need to create and explore predictive dynamical models of complex biological systems such as metabolic, gene regulation, or signal transduction pathways in living cells [1].

58.1.1 Overview of the Software Infrastructure

The Sigmoid modeling system core consists of distributed modules implementing: (1) pathway/cell model generation and simulation (Cellerator; [2]), (2) a pathway modeling database (Sigmoid Proper), (3) a Web service-oriented middleware, (4) a World Wide Web model browser, and (5) a graphical user interface (Sigmoid Model Explorer (SME)) friendly to a biologist user. From there, other components have been integrated into the system such as a parameter optimization module and functional connections to compatible external data sources. These modules are organized in a classical three-tier architecture (Fig. 58.1). The back end currently consists of the database, the simulator, and other model manipulators. The GUI front end does not access the back-end modules directly but rather via a Web

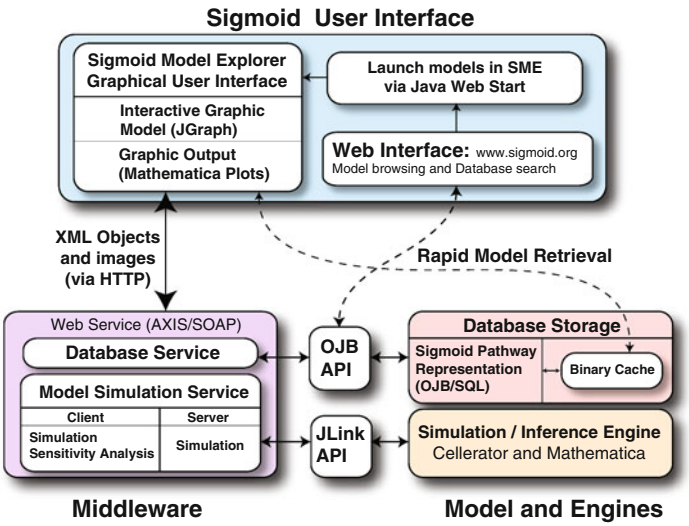


Fig. 58.1 Sigmoid three tier architecture. Separation of modules into a communicating distributed system increases scalability of the architecture. Our simulator is the xCellerator model generator/simulator; the database is Sigmoid (autogenerated from a UML schema); user interface is the Sigmoid Model Explorer (SME)

service middleware module. The extra development overhead introduced by the middle layer is more than compensated by the advantages in terms of distributed computing, performance, flexibility, and scalability. With the exception of rapid model retrieval, the middleware layer brokers all communications between the GUI and the back-end components and also among the back-end components themselves. We have found that storing binary instances of models in a database cache can provide significant improvements in model retrieval times in comparison to full model reconstruction and retrieval through the middleware layer. In the event that the rapid model retrieval interface is not accessible, the system will shift access to the database through the middleware. This infrastructure was created by a close collaboration between bioinformaticians and biologists by having the design of many of the essential software objects and their relationships be visible as implementation proceeded.

We have coordinated the development of various software modules in Sigmoid by using the Universal Modeling Language (UML) to diagram the most important biological objects notably reactions and molecular reactants. This UML diagram is used as a template to automatically generate several parts of Sigmoid, in particular a realization of the Sigmoid pathway modeling database (in SQL) and the corresponding Java object hierarchy along with support files for facilitating the object-relational mapping and end-user documentation. Also the Graphical User Interface (GUI) relies heavily on the Java reflection utility to automatically discover much of what it needs to know about the Sigmoid schema. Thus, there is a guarantee that the software actually implements something very close to the UML construction of biological objects. In addition, coding time for different modules of the system is reduced.

To keep the infrastructure flexible and manageable as it grows, we have resorted to a "generative" approach that seeks to partially automate the generation of both executable code and mathematical models. We have applied this approach to as many of the modules in Fig. 58.1 as possible, starting from high-level inputs such as UML diagrams and reaction notations understandable to noncomputer scientists.

58.2 Methods

58.2.1 Model Generation and Simulation: xCellerator

In order to facilitate the modeling of biochemical reactions, a library of reusable reaction models that can be expressed in a simple higher-level language that specifies the molecular species and the type of reaction is required. Cellerator [2] code is implemented as a Mathematica notebook and designed to facilitate biological modeling via automated equation generation. Sigmoid now supports xCellerator [3], the most recent version of Cellerator.

Many models of molecular interactions have been implemented in xCellerator using different formalisms, such as differential equations or stochastic molecular

simulation formalism ranging from the law of mass action and simple Michaelis-Menten models to more complex models of enzyme reactions (e.g., the Monod Wyman Changeaux or MWC model for allosteric enzymes [4]) and gene regulation [5]. The list of reaction models continues to expand along with the library of actual pathway models comprising sets of coordinated reactions with parameters derived from the literature whenever possible. In addition, an extended set of enzyme mechanism models for single and multi-substrate, positively and negatively regulated, and allosteric enzymes, called kMech, has been written for xCellerator and continues to develop [6]. Sigmoid currently supports all the available xCellerator and kMech reaction models.

To illustrate xCellerator utility, consider the example of a three-stage catalytic model. This reaction is a composite representation of three reversible reactions; substrate enzyme complex formation, the conversion of the substrate to product within the complex, and subsequent disassociation of the enzyme product complex into free enzyme and product. When presented with the correct input notation, xCellerator will translate the symbolic reaction to differential equations. The resulting differential equations and variable definitions are passed to Mathematica where they are solved by the numeric solver function (NDSolve) and time plots generated. See example in Fig. 58.2. The parameters for this enzyme mechanism

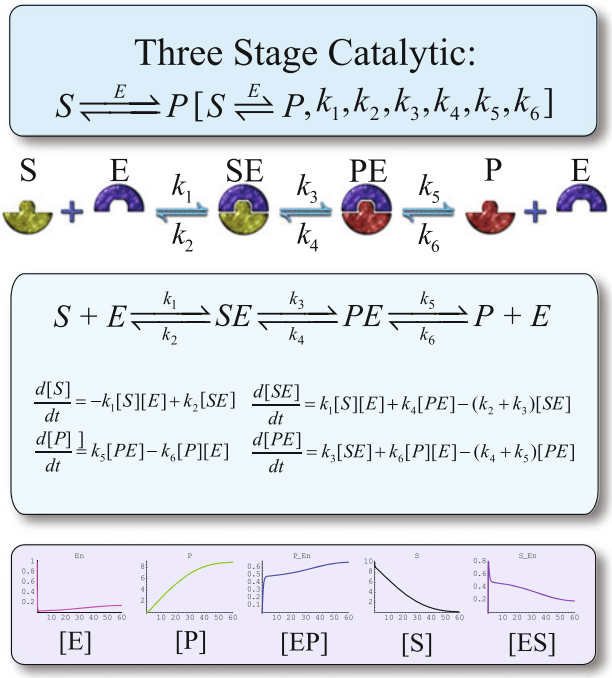


Fig. 58.2 Sigmoid Three Stage Catalytic model. From *Top* to *bottom*. xCellerator input notation, reaction cartoon, resulting differential equations and an example of numerical output

are stored in the Sigmoid Pathways Database. In short, xCellerator converts symbolic reactions to mathematical equations and solves the corresponding equations.

58.2.2 Sigmoid Pathway Database

The pathway model database is defined by a UML schema. Comprehensive UML class diagrams of the Sigmoid Schema can be found at http://www.sigmoid.org. The schema is organized into four main diagrams. The first diagram consists of the various top-level container classes such as the Model Class and the Gene Ontology source class. The first diagram also contains the parameter set hierarchy, classes for graphical layout in SME, and various classes to handle units and measures. The three remaining diagrams consist, respectively, of three major class hierarchies: Reactions, Reactants, and Knowledge Sources. Reactions utilize Reactants for their products, substrates, and enzymes; Models are composed of parameterized Reactions, and these three class hierarchies utilize Knowledge Sources to reference external information about themselves.

While initial versions of the Sigmoid database were implemented by hand, we wished to automatically transform the class descriptions contained in the high-level UML diagram of this hierarchy into a set of instantiable objects upon which applications may be built. Our current approach to the process of auto-generating software components from a master UML diagram relies on the capabilities of several existing open-source projects [1]. These pre-existing projects remove much of the core software development responsibilities and allow us to focus on tying them together to produce the specific software products needed for our own use. Object-relational database code autogeneration from UML is itself a contribution of potentially general interest in database software engineering. The current version of Sigmoid is implemented using PostgreSQL, the main Open Source database software.

An essential function of Sigmoid is to assist in the translation of biological knowledge into mathematical form. The representation of Reactions in Sigmoid is aimed at this goal. Sigmoid Reactions represent biochemical processes that transform molecular or other biological objects that are represented as Sigmoid Reactants. A major design feature of Sigmoid is to support translation of biology into mathematics. Reactions are defined in two subhierarchies: Biological Reactions and Mathematical Reactions. The Biological Reaction hierarchy is intended to provide biologically oriented users with symbolic representations of a biochemical reaction or process. Attributes that represent the basic reactants with primary roles are included. The kinetics of the reaction are abstracted out and delegated to Mathematical Reactions. Mathematical Reactions constitute a type hierarchy of mathematical models of reactions or other processes in the Sigmoid schema. Such representations include particular rate laws, as well as the translation of compound reactions into a subnetwork of more elementary reactions, each of which has a more elementary mathematical model. Most Mathematical Reactions currently have direct xCellerator/kMech implementation functions associated with them.

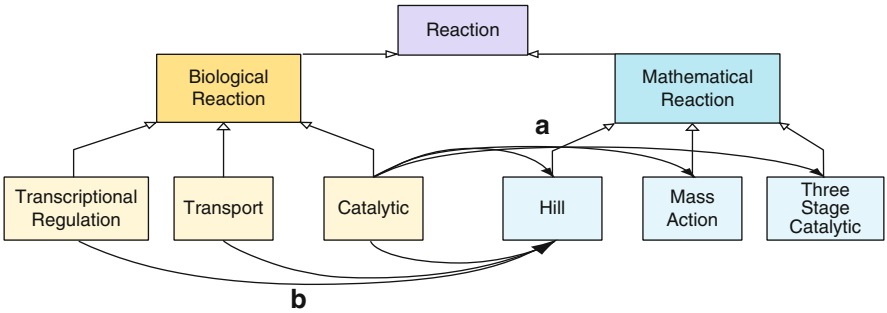


Fig. 58.3 Simplified version of the Sigmoid Schema Reaction hierarchy. (**a**) There may exist one to many relations between a particular biological reaction and potential functions (Mathematical Reactions) that may be assigned to model the kinetics of the interaction. For instance, numerous mathematical functions can be assigned to model a catalytic process. (**b**) In reverse, the functional application of a particular set of differential equations may be conserved over a variety of biological phenomena, so there also may be one to many associations between a particular mathematical function (Reaction) and the biological scenarios it may be applied to. For instance, a hill equation may provide useful in modeling a catalytic reaction, transcriptional regulation, or even a transport process

Numerical parameters associated with each reaction are contained by reference, which enables key reaction parameters to be shared within a Mathematical Reaction or across a full reaction network.

This way, the Sigmoid architecture can offer explicit support for the translation of biological processes into mathematical process models. Each type of biological reaction may, in principle, be translated into several alternative mathematical reaction models, and each mathematical reaction model can serve as the translation of several different biological reactions. An example of the importance of many-to-many reaction translations is shown in Fig. 58.3.

58.2.3 Sigmoid Web Middleware for Distributed Computing and Web Services

A new distributed Web middleware layer was built which accesses the Sigmoid database and translates reaction sets into the input language of the xCellerator cell model generator. It then calls xCellerator with requests for model generation and simulation and receives output plots in response. All these functions are exposed as Web services available to Java application programs and/or other clients. In addition to load balance and security management, the middleware provides a gateway between the front end and the back end of the architecture, allowing each one to evolve independently as long as the interface to the middleware is properly maintained. Furthermore, the middleware allows scalability in terms of the number of users who can be served simultaneously simply by increasing the computational and database server resources [1].

58.2.4 The Graphical User Interface: Sigmoid Model Explorer User Interface

The last component of the system to be initiated, and the most recent to achieve functional maturity, is the SME Web-compatible Graphical User Interface. The GUI allows the user to visualize, design, edit, and store pathway models, parameters, and initial conditions and their properties, to simulate the models by calling the simulator through the middleware, and to view and compare the properties of simulated models by viewing the temporal evolution of the concentration of chemical species under different conditions. The GUI runs from any Web browser as a Webstart or as a local client program.

Recent enhancements to SME are: (1) For model creation: there exists a new mechanism to create biological models completely from within SME and save them locally or commit them to the database. To facilitate the construction of more complex biological processes, one to many mathematical reactions can be assigned to each biological reaction. Also, there are utilities to facilitate the use of webpages as source of information for data input and perform queries to the Gene Ontology database from within SME. Gene Ontology entities can be either used to tag Sigmoid objects or instantiated directly as Sigmoid objects, i.e., Reactants or Biological reactions. (2) Numerous enhanced display features. (3) Model translation: SME can preform local translation of Sigmoid models to xCellerator code and perform translation of SBML 1.0 to Mathematica code. (4) Model simulation: SME supports simulation through a local Mathematica license using the JLink library as well as through the remote server, and there is an option to retrieve and display the output graphs for intermediate complexes generated by xCellerator/kMech reaction types. (5) Connectivity: SME now supports the Web Services Description Language (WSDL), which is an XML grammar for describing network services. Supporting WSDL expedites adoption of supplementary datasets and functionalities from other systems that support this standard.

58.3 Results

58.3.1 Sigmoid Database Population

The generative version of Sigmoid has been successfully populated with over 20 published models that range from simple molecular interactions to complex cell-fate decision networks. A majority of the models in the database focus on virtual representation of intracellular pathways that include examples in signaling, metabolism, the cell cycle, and gene regulation. Large-scale models of the signaling pathways include the mammalian Epidermal Growth Factor Receptor (EGFR) pathway [7] and the yeast pheromone response pathway [8], while other models represent common aspects of metabolism that include the anabolic Calvin cycle in

plants [9], two models of branched chain amino acid biosynthesis in bacteria [10, 11], and catabolic glycolysis [12]. Furthermore, a simple model of the circadian clock [13] and two models of intracellular calcium flux [14] demonstrate oscillating outputs. Separate models of the NFkB [15], calcineurin [16], and the p53 [17] regulatory networks demonstrate how transcription factors and their ability to activate or inhibit gene expression are regulated. Lastly, some models in the database represent diverse processes, including the mechanism of degradation of enzymes during industrial food processing [18] and the cell-fate decisions of protists in the presence of far-red light under starvation conditions [19].

Finally, computational models of the mitogen-activated protein kinase (MAPK) cascade are also present in the Sigmoid database. Several models derived from [20] examine the same MAPK cascade with two separate mechanisms, mass action and Michaelis-Menten, for each of the phosphorylation and dephosphorylation events. For each of these mechanisms, the models increase in complexity as the site and order of phosphorylation are taken into account in the set of reactions. In contrast to these models, Huang 1996 MAPK and its xCellerator notebook "MAPK cascade: Huang and Ferrell 1996," present the celebrated (1996) model that demonstrates the connection between a nonprocessive, two-collision dual-phosphorylation mechanism of kinase activation and an ultrasensitive, switch-like response. The model Bardwell 2007 MAPK Variable Feedback and the corresponding notebook "MAPK Cascade with Variable Feedback" extend this model to include a simple feedback phosphorylation of an upstream kinase by the MAPK (Fig. 58.4). The effects of the feedback loop on the system depend upon the nature of the feedback: if feedback phosphorylation increases the activity of the upstream kinase (positive feedback), a bistable, all-or-none response may result [21]. In contrast, if feedback phosphorylation decreases the activity of the upstream kinase (negative feedback), then the result may be damped or sustained oscillation of the activity of the kinases in the cascade [22]. The notebook contains examples of parameter values that will generate either of these outcomes, illustrating how complex, diverse, and

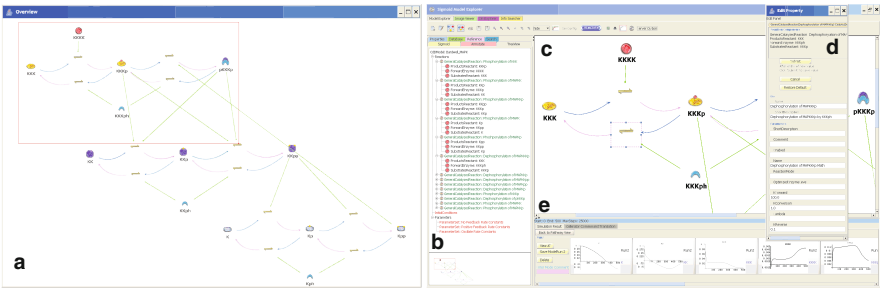


Fig. 58.4 Sigmoid Model Explorer showing portion of MAPK pathway. (**a**) Global Network View; (**b**) TreeView of compositional hierarchy; (**c**) network layout visualization; (**d**) parameter editing panel; (**e**) output plot preview panel. Along the top are various action buttons for saving and running the model, and for switching the main panel to view output plots. User can select reaction icons.

biologically useful behaviors can emerge from the combination of an ultrasensitive cascade architecture and a simple feedback loop.

Since the flexible but comprehensive schema of the Sigmoid database allows us to easily leverage other databases, we are developing "populator" programs which capture community input from diverse sources and make it available to a biologist end-user in an integrated manner. For example, without much effort we were able to populate Sigmoid with the yeast GOnet database [23], which contains information about yeast ORFs and their annotations, gene ontology (GO), and protein protein interactions. The GOnet database itself is periodically updated and integrates information from three different sources: (1) ORFs (description, mutant phenotype, gene product, etc.) from the *Saccharomyces* Genome DataBase (SGD); (2) GO term annotation from the Gene Ontology Consortium arranged in the three categories of Molecular Function, Biological Process, and Cellular Component; and (3) genetic and physical interactions information from the General Repository for Interaction Datasets (GRID).

58.3.2 Parameter Optimization

A Simulated Annealing Optimizer [24] has been integrated into Sigmoid through the web services interface. It uses a global optimization technique and Lam Delosme schedule to make the optimization process faster and more efficient when compared with other general schedules available [25]. It aims to reverse engineer model parameters (e.g., kinetic rate constants) given both the model structure (represented as ordinary differential equations) and empirical system dynamics as expressed by time series experimental data.

58.3.3 Parameter Analysis

The Parameter Analysis routine in Sigmoid allows one to quickly sample the parameter space of a particular model and quantify the diversity of model outputs resulting from variation of the parameters in specified ranges. First, free parameters are defined within the model that will be part of the analysis. Then, a simulation function is defined that accepts a particular parameter variation and returns the model's output. Users have options to select Sigmoid output functions, such as the temporal sequence of a particular state variable. The output variation is measured using preset or user-defined metrics aimed at focusing on particular aspects of output behavior. For example, one can measure the difference between the obtained output and some reference time state or determine the time points at which the output might have peaks or troughs in an oscillatory response. The value of the metric might reflect on how sensitive a certain model is to simultaneous variation of any number of parameters, from one to all. This information can then

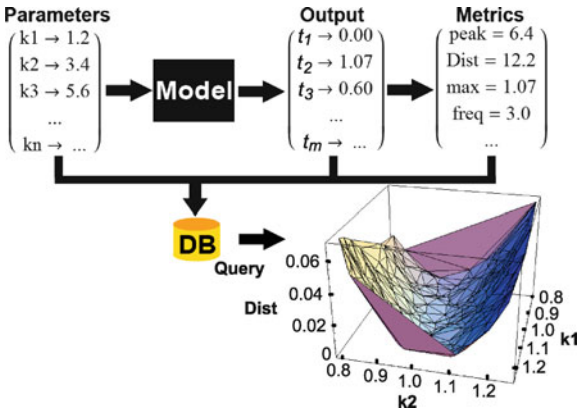


Fig. 58.5 Sensitivity of model output to parameter variations is handled by a set of operations integrated into the Sigmoid environment. These functions or their user defined variants can allow fast and efficient generation of a set of solutions corresponding to variation of any parameter number from one to all and storage of these solutions in a database that can be queried to form various metrics of model performance. The results can be used to analyze the robustness of various models of a specific biochemical system of interest

be used in investigation of robustness of the model and the corresponding biological process. The values of the varied parameters, model output, and resulting metrics are stored in a database table using Mathematica's Database Link package. Using a database provides a convenient method for storing the vast amounts of tabular data and allows for rapid remote access. Since model evaluations are independent, the procedure is easily parallelized. The same notebook can run on multiple computers simultaneously, as long as all can connect to the same database. Lastly, Mathematica's powerful visualization and analysis features can be used to observe correlations between parameter values and associated metrics. (See Fig. 58.5.)

58.4 Conclusions

We have described the Sigmoid intelligent software infrastructure for systems biology. A version of each of the main components is available today, and there are clear signs that the infrastructure can already be used to yield biologically relevant results. Since Sigmoid is based upon a computer algebra representation tool, it stands poised to serve as a formidable engine in model analysis. For instance, the *Escherichia coli* metabolic pathway model correctly predicts the effect of certain mutations, and the MAP Kinase cascade model shows that, depending on the parameter sets and initial conditions chosen, it can generate a switch-like or graded input output relationship, or even produce oscillatory behavior.

Development and expansion of Sigmoid continues at all levels. As the mediator of the user experience with Sigmoid, the GUI and web interface are bound to attract

the largest number of feature requests from users. Because the overall architecture is now functional, many of these requests can be met at reasonable levels of effort and cost. Depending on their accessibility to software agents, an opportunity exists to import relevant data from other sources such as KEGG, Systems Biology Workbench, SiBML/GeneNet, SabioRK, Biomodels etc. There is also a need for new reaction types in xCellerator to deal with various kinds of (nontranscriptional) feedback. Other reaction types already in xCellerator and kMech (such as various enzymatic models, GMWC, GRN, etc.) will need to be exposed for further pathway modeling. An essential aspect of the scale-up of Sigmoid will be expert curation of the allowed and suggested mappings from biological reaction mechanisms to mathematical reaction models.

Sigmoid capitalizes on the robust mathematical software tools and the problem-solving environment that Mathematica offers (along with the xCellerator/kMech packages designed to facilitate biological modeling via automated equation generation). Sigmoid implements the web services framework [1] to create a truly distributed system. This flexible, scalable architecture offers powerful modularity that, in conjunction with the generative nature of the Sigmoid coding cycle, allows for manageable, cost-effective adoption of new system components (new simulation engines, analysis tools, and data structures) while opening the ability to play within yet larger bioinformatics frameworks.

Acknowledgment This work has been supported by NSF grant EIA 0321390 and NIH grant T15 LM007443 to PB, a Laurel Wilkening faculty innovation award to PB, a UC Systemwide Biotechnology Research and Education Program 2002 2006 award to PB, NIH grant GM069013 to EM, and NCI Director's Challenge support to Children's Hospital Los Angeles for EM. B.C. was supported by NIH grant T15LM07443 from the National Library of Medicine; A.L. was supported by NIH grants: GM69013 and GM072024, KS and TW were supported by NIH P50 grant GM76516, NASA Intelligent Systems Program support of EM, and by the Institute for Genomics and Bioinformatics at UCI. We would like to thank students, programmers, and colleagues who have provided us with valuable feedback or have helped implement particular components of the infrastructure. They include Ben Bornstein, G. Wesley Hatfield, Peter Hebden, Elliot Meyerowitz, Kirill Petrov, Lucas Scharenbroich, Tarek Najdi, Li Zhang, Bruce Shapiro, Diane Trout, and Chin ran Yang.

References

1. J. Cheng, L. Scharenbroich, P. Baldi, and E. Mjolsness. Sigmoid: Towards a generative, scalable software infrastructure for pathway bioinformatics and systems biology. *IEEE Intelligent Systems*, 20(3):68 75, 2005.
2. B. E. Shapiro, A. Levchenko, E. M. Meyerowitz, B. J. Wold, and E. D. Mjolsness. Cellerator: Extending a computer algebra system to include biochemical arrows for signal transduction simulations. *Bioinformatics*, 19(5):677 678, 2003.
3. B. E. Shapiro, J. Lu, M. Hucka, E. Mjolsness. Mathematica platforms for modeling in systems biology: Recent developments in MathSBML and Cellerator. Poster Abstract G24, Eighth International Conference on Systems Biology (ICSB), Long Beach, 2007. URL http://icsb 2007.org/ , last accessed on June 29, 2010.

4. T. S. Najdi, C. R. Yang, B. E. Shapiro, G. Wesley Hatfield, and E. D. Mjolsness. The generalized Monod, Wyman, Changeux model for mathematical modeling of metabolic enzymes with allosteric regulation. In *Proceedings of the IEEE Computational Systems Bioinformatics Conference*, Stanford University, Stanford, CA, 2005.
5. I. H. Segel. *Enzyme Kinetics. Behavior and Analysis of Rapid Equilibrium and Steady State Enzyme Systems*. Wiley, New York, NY, 1992.
6. C. R. Yang, B. E. Shapiro, E. D. Mjolsness, and G. W. Hatfield. An enzyme mechanism language for the mathematical modeling of metabolic pathways. *Bioinformatics*, 21:774 780, 2005.
7. B. N. Kholodenko, O. V. Demin, G. Moehren, and J. B. Hoek. Quantification of short term signaling by the epidermal growth factor receptor. *J Biol Chem*, 274(42):30169 30181, 1999.
8. B. Kofahl and E. Klipp. Modelling the dynamics of the yeast pheromone pathway. *Yeast*, 21(10):831 850, 2004.
9. M. G. Poolman, H. E. Assmus, and D. A. Fell. Applications of metabolic modelling to plant metabolism. *J Exp Bot*, 55(400):1177 1186, 2004.
10. T. S. Najdi, C. R. Yang, B. E. Shapiro, G. W. Hatfield, and E. D. Mjolsness. Application of a generalized MWC model for the mathematical simulation of metabolic pathways regulated by allosteric enzymes. *J Bioinform Comput Biol*, 4(2):335 355, 2006.
11. C. R. Yang, B. E. Shapiro, S. P. Hung, E. D. Mjolsness, and G. W. Hatfield. A mathematical model for the branched chain amino acid biosynthetic pathways of *Escherichia* coli k12. *J Biol Chem*, 280(12):11224 11232, 2005.
12. K. Nielsen, P. G. Sarensen, F. Hynne, and H. G. Busse. Sustained oscillations in glycolysis: An experimental and theoretical study of chaotic and complex periodic behavior and of quenching of simple oscillations. *Biophys Chem*, 72(1 2):49 62, 1998.
13. J. J. Tyson, C. I. Hong, C. D. Thron, and B. Novak. A simple model of circadian rhythms based on dimerization and proteolysis of PER and TIM. *Biophys J*, 77(5):2411 2417, 1999.
14. J. M. Borghans, G. Dupont, and A. Goldbeter. Complex intracellular calcium oscillations. A theoretical exploration of possible mechanisms. *Biophys Chem*, 66(1):25 41, 1997.
15. A. Hoffmann, A. Levchenko, M. L. Scott, and D. Baltimore. The IkappaB NF kappaB signaling module: temporal control and selective gene activation. *Science*, 298(5596):1241 1245, 2002.
16. Z. Hilioti, D. A. Gallagher, S. T. Low Nam, P. Ramaswamy, P. Gajer, T. J. Kingsbury, C. J. Birchwood, A. Levchenko, and K. W. Cunningham. GSK 3 kinases enhance calcineurin signaling by phosphorylation of RCNS. *Genes Dev*, 18(1):35 47, 2004.
17. A. N. Bullock and A. R. Fersht. Rescuing the function of mutant p53. *Nat Rev Cancer*, 1(1):68 76, 2001.
18. C. M. Brands and M. A. van Boekel. Kinetic modeling of reactions in heated monosaccharide casein systems. *J Agric Food Chem*, 50(23):6725 6739, 2002.
19. W. Marwan. Theory of time resolved somatic complementation and its use to explore the sporulation control network in *Physarum* polycephalum. *Genetics*, 164(1):105 115, 2003.
20. N. I. Markevich, J. B. Hoek, and B. N. Kholodenko. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *J Cell Biol*, 164(3):353 359, 2004.
21. J. E. Ferrell and E. M. Machleder. The biochemical basis of an all or none cell fate switch in Xenopus oocytes. *Science*, 280:895 898, 1998.
22. B. N. Kholodenko. Negative feedback and ultrasensitivity can bring about oscillations in the mitogen activated protein kinase cascades. *Eur J Biochem*, 267:1583 1588, 2000.
23. B. Irwin, M. Aye, P. Baldi, N. Beliakova Bethell, H. Cheng, Y. Dou, W. Liou, and S. Sandmeyer. Retroviruses and yeast retrotransposons use overlapping sets of host genes. *Genome Res*, 15:641 654, 2005.
24. L. Zhang. *Dynamic Biological Signaling Pathway Modeling and Parameter Estimation Through Optimization*. PhD thesis, Information and Computer Science: University of California, Irvine, 2008. LD 791.9 I5 2008 Z43, OCLC:276454918.
25. J. Lam and J. Delosme, Performance of a New Annealing Schedule. Proc. of the 25th ACM/ IEEE DAC, pp. 306 311, 1988.

Chapter 59
Registration of In Vivo Fluorescence Endomicroscopy Images Based on Feature Detection

Feng Zhao, Lee Sing Cheong, Feng Lin, Kemao Qian, Hock Soon Seah, and Sun-Yuan Kung

Abstract The confocal fluorescence endomicroscopy is an emerging technology for imaging the living subjects inside the animals and human bodies. However, the acquired images vary, due to two degrees of freedom tissue movement and tissue expansion/contraction. This makes the 3D reconstruction of them difficult and thus limits the clinic applications. In this chapter, we propose a feature-based registration algorithm to correct the distortions between these fluorescence images. The good alignment enables us to reconstruct and visualize the 3D structure of the living cells and tissues in real time, which provides the opportunity for the clinicians to diagnose various diseases, including the early-stage cancers. Experimental results on a collection of more than 300 confocal fluorescence images of the gerbil brain microvasculature clearly demonstrate the effectiveness and accuracy of our method.

Keywords Confocal fluorescence image · Biomedical image processing · Image alignment · Endomicroscopic imaging · Computational system

59.1 Introduction

Traditionally, the gold standard for the detection of cancer is the ex vivo histological examination of the hematoxylin and eosin (H&E) stained biopsy samples under white light microscope. However, the invasive nature of tissue extraction carries the risk of causing bleeding, infection, perforation, or mechanical agitation that may

F. Zhao (✉)
School of Computer Engineering, Nanyang Technological University, Singapore 639798

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 59,

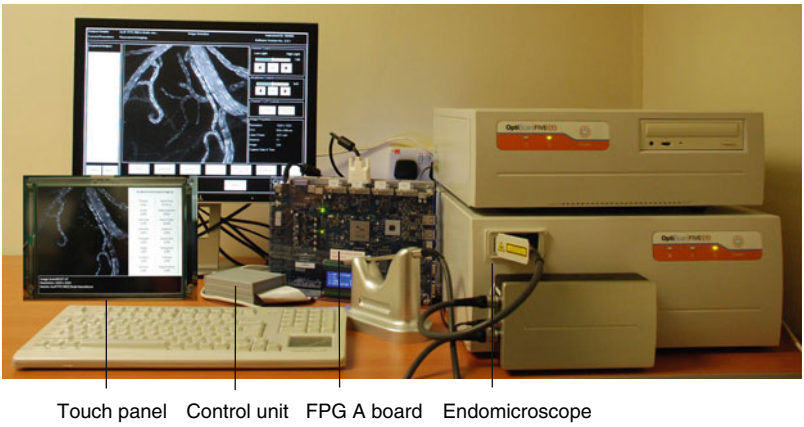


Fig. 59.1 The proposed *in vivo* cellular imaging system

lead to the spread of tumor cells through the blood and lymphatic vessels [1]. In contrast, the molecular imaging allows the visualization of the expression and activity of specific key molecular targets and host responses associated with early events in carcinogenesis, in addition to eliciting the location of the tumor [2].

The optical fluorescence imaging is a typical molecular imaging modality, which transmits an incident light within the absorption spectrum of the fluorochrome and images the emitted light of a different spectrum. The frequently used imaging probes in cancer imaging contain fluorescence proteins [3] such as green fluorescence protein (GFP) and red fluorescence protein (RFP), and photosensitizers [4] such as 5-aminolevulinic acid (5-ALA) and 5-ALA esters. The available systems include the fluorescence endoscopy [5], the fluorescence laser confocal scanning microscopy [6], and the newly emerging confocal fluorescence endomicroscopy [7, 8].

As illustrated in Fig. 59.1, our system is a powerful handheld fluorescence endomicroscope specifically designed for *in vivo* imaging of living subjects such as animal tissues. It consists of four main components: (1) the imaging instrument that is a confocal fluorescence endomicroscope (OptiScan FIVE 1); (2) the computing device that is a field programmable gate array (FPGA) board; (3) the control unit that is interfaced between the FPGA computing system and the imaging probe to make the entire 3D image stack acquired automatically instead of manual adjustment of the z-depth; and (4) the display device that is a touch panel screen allowing the visualization of the output. Being noninvasive, it enables direct observation of molecular mechanisms by continuously scanning the surface and subsurface tissue structures without removing tissues or sacrificing animals.

Our system provides unique submicron resolution *in vivo* 3D microscopic volume of slice images with a 475 $\times$ 475 μm field of view (FOV) and a 4 μm z-depth increment. Figure 59.2 shows some confocal fluorescence images of the microvasculature in the gerbil brain [9, 10]. We can see that such endomicroscopy

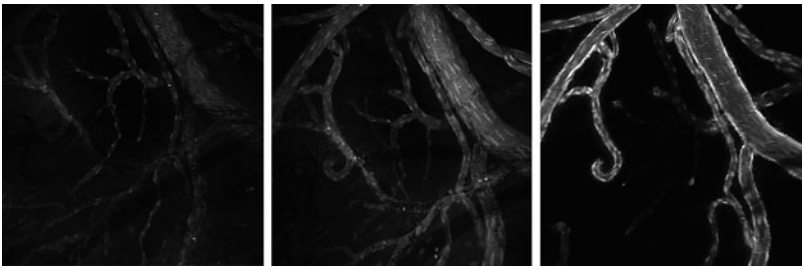


Fig. 59.2 Fluorescence images of mouse brain microvasculature at different z depths

images have several inherent characteristics: (1) the images are molecular imaging of the living animal tissue, (2) they are largely magnified by the microscopic probe, (3) the images are labeled with photosensitizers that selectively accumulate in the tissue, (4) the fluorescence images are much noisier, (5) the images are unevenly illuminated, and (6) the images are translation-, rotation-, and scale-variant.

Combined with new photosensitizers and pharmaceutical-grade penetration enhancers selectively accumulate in abnormal tissues (e.g. neoplasm and cancerous cells), the 3D microstructures can be reconstructed and visualized [11 14]. However, the major problem is the automated registration of consecutive fluorescence images. The challenges are twofold. On the one hand, each of the 2D arbitrarily taken slice images suffers from possible movement and tissue expansion or contraction, which means that the exact coordinates misalignment and geometrical distortion of the 2D slice images are unknown. The reconstruction of 3D spatial information from these 2D misaligned and distorted slice images is an immerse difficulty. On the other hand, beyond a certain time frame, the 3D volumetric images may be different due to physiological changes. Thus, the deciding factor of whether to reconstruct certain slice into the 4th dimension (temporal) as opposed to using it to refine the existing 3D spatial volume needs to be identified and considered. The slow rate of image acquisition and low signal to noise ratio (SNR) further complicate the matter. To solve the problems, we propose a feature-based image registration algorithm, which first detects a few salient points from the endomicroscopy images as features and then aligns every two consecutive slice images based on the remaining feature points after postprocessing. The good alignment between slice images will help us obtain a better reconstruction and visualization of the 3D microstructures of the cells and tissues. Thus, it will further assist the clinicians to understand the mechanisms of various diseases [15, 16].

The remainder of this chapter is organized as follows. Section 59.2 presents the skeleton-based feature extraction method, including both the preprocessing steps and the postprocessing techniques. The feature-based image registration algorithm is described in Sect. 59.3. Section 59.4 details the implementation process on the FPGA board. Experimental results and discussions are reported in Sect. 59.5. Section 59.6 concludes this chapter.

59.2 Skeleton-based Feature Detection

From the images shown in Fig. 59.2, we note some salient structures existing in the fork regions and ending regions of the blood vessels, which can be represented by two types of feature points: bifurcations and endpoints, respectively. The feature extraction algorithm generally consists of four main steps: (1) use an adaptive thresholding algorithm to compute the binary image from the input gray scale image; (2) use a thinning algorithm to compute the skeleton image from the binary image; (3) use Rutovitz crossing number to extract the feature points from the skeleton (thinning) image; and (4) postprocess the feature set according to some heuristic rules for the elimination of false features.

To obtain a better skeleton image of the blood vessels, we propose several preprocessing techniques before thinning of the binary image: (1) identify the bright areas (corresponding to the nuclei of the circular smooth-muscle cells and the vascular endothelial cells lining the lumen of the arteriole) and replace their gray scale intensities with the average intensity values in their respective neighborhoods; (2) filter the gray scale image using the Gaussian kernel function for noise removal; (3) apply the morphological top-hat filtering on the gray scale image using a disk structure element to correct uneven illumination; and (4) apply some morphological operations on the binary image to fill the holes and remove the small isolated areas. Figure 59.3 shows the effect of the preprocessing steps, where we can see that the thinning image after preprocessing roughly reflects the structures of the blood vessels.

After the thinning image is obtained, we scan the whole image and extract all the feature points, including bifurcation and ending points according to the Rutovitz crossing number (*CN*). The definition of crossing number for a pixel P is: $CN = \frac{1}{2}\sum_{i=1}^{8} |P_i - P_{i+1}|$, where P_i is the binary pixel value in the neighborhood of the center pixel P with $P_i = (0 \text{ or } 1)$ and $P_1 = P_9$. Figure 59.4 illustrates the properties of *CN,* and Fig. 59.5 shows an example to demonstrate the feature extraction results. Among the extracted feature points, some are true features we

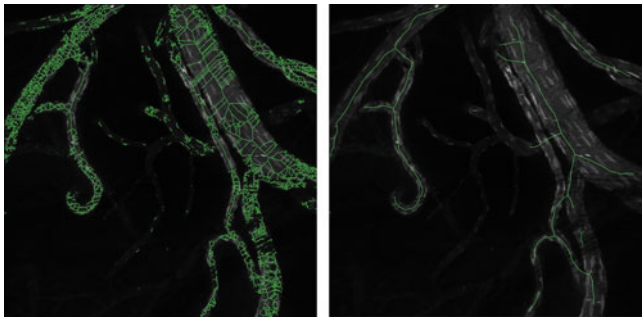


Fig. 59.3 An example to show the effect of the preprocessing steps by overlaying the thinning image on the original gray scale image (*left*: without preprocessing; *right*: with preprocessing)

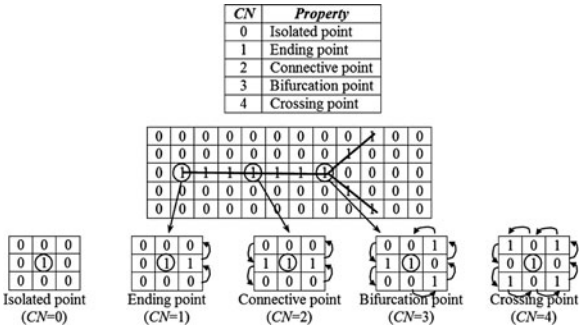


Fig. 59.4 Properties of Rutovitz crossing number (1: skeleton pixels in the thinning image)

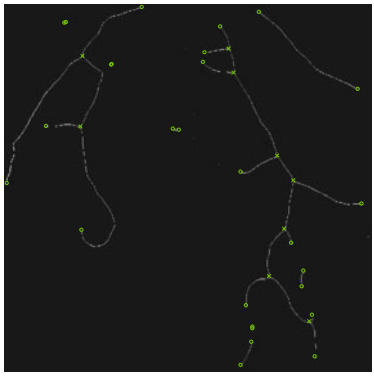


Fig. 59.5 Feature extraction results overlaying on the thinning image (o: endpoints; x: bifurcations)

prefer, while some are false features due to noises. The false ones will be eliminated in the postprocessing stage.

Due to various noises in the endomicroscopy images, the feature extraction algorithm produces a number of spurious feature points such as break, spur, dot, merge, and island, as shown in Fig. 59.6. Therefore, reliably differentiating spurious features from genuine features in the postprocessing stage is crucial for the later image registration stage. The more spurious features are eliminated, the better the registration performance will be. Moreover, the registration speed will be significantly improved. This is very important since the computation time is a critical parameter for an online system.

To speed up the online postprocessing process, the extracted feature points are handled in an efficient way. First, we remove the two endpoints of a break according to the following conditions: (1) the two endpoints are not connected with each other, and the distance between them is below a threshold T_1; (2) the difference between the orientation angles of the two endpoints (Ang_1, Ang_2) is within an interval of $[\theta_1, \theta_2]$; and (3) the difference between the orientation angle of the line connecting the two endpoints (Ang_3) and Ang_1 (or Ang_2) is within an interval of

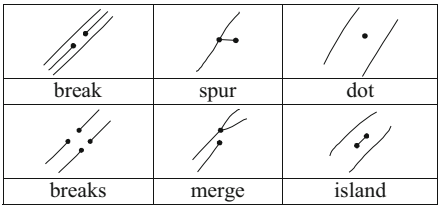


Fig. 59.6 Illustration of false feature points (*black dots*)

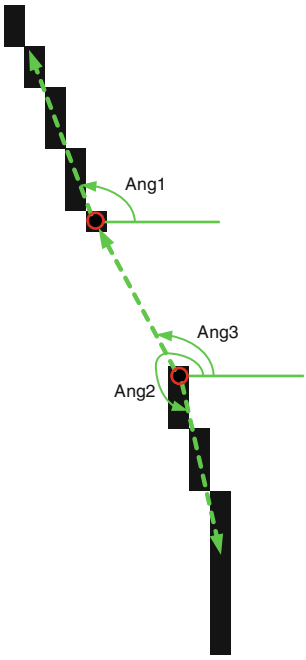


Fig. 59.7 Elimination of the break (o: endpoints)

$[\theta_3, \theta_4]$. To calculate the orientation angle of an endpoint, we look for the 8-connected neighbors around the endpoint. The orientation angle is determined by the endpoint and the last neighbor with respect to the horizontal axis in counterclockwise direction, as shown in Fig. 59.7.

Secondly, we identify the spur consisting of a bifurcation and an endpoint if the two feature points are both globally and locally connected with each other and their distance is below a threshold T_2. In this work, we first label the connected pixels in the skeleton image. If the distance (d) between a bifurcation point and an endpoint is below the threshold T_2 and their labels are the same, we again label the connected pixels within a small circular window of radius d centered around the endpoint or the bifurcation point. If their labels are still the same, we remove

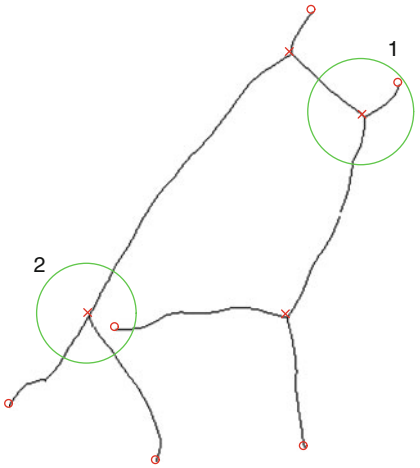


Fig. 59.8 Elimination of the spur. After relabeling the connected pixels within the circles, the bifurcation and the endpoint inside window 1 still have the same labels, while those within window 2 have different labels (o: endpoints; x: bifurcations)

both of them; otherwise they will be preserved. Figure 59.8 demonstrates the procedure.

After that, the dot and island are removed by labeling the connected components and counting their areas (number of pixels) in the skeleton image, since dots are isolated pixels and islands are short lines consisting of a small number of pixels. Finally, we eliminate those features that are too close to each other and all the points within a certain distance threshold T_3 from the image boundary. After postprocessing, a large percentage of the spurious feature points are eliminated, the remaining ones are treated as true features that are used for the later image registration.

59.3 Image Registration Using Feature Pairs

As shown in Fig. 59.9, the extracted feature points (bifurcations and endpoints) after postprocessing are represented by the attributes: $\vec{f} = [x, y, z, \theta_i]^T$, where i=1 (for an endpoint) or 1,2,3 (for a bifurcation), x and y are the coordinates of the origin point of the one (endpoint) or three (bifurcation) associated branches in the X Y plane, z is the z-depth in the scanning direction of the FIVE 1 microscopic probe, and θ_i denotes the local orientation of the ith associated branch.

The image registration is applied for every two consecutive slice images, which involves four basic steps: feature detection, feature matching, mapping function design, and image transformation and resampling. Assume that the image scanning is in the direction from the surface layer to the subsurface layers of the living tissue. Let $F^A = \{x^m, y^m, z^m, \theta_i^m\}_{m=1}^M$ and $F^B = \{x^n, y^n, z^n, \theta_i^n\}_{n=1}^N$ denote the feature vectors of two slice images with M and N extracted feature points (image B is the next

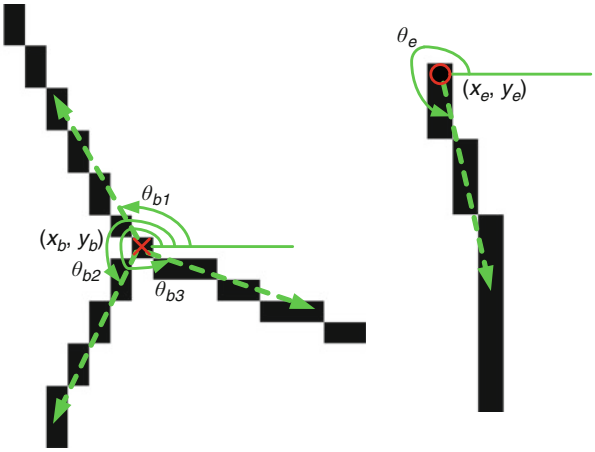


Fig. 59.9 Feature vector attributes of a bifurcation (*left*) and an endpoint (*right*)

layer of image A, i.e. $z^n = z^m + 1$), the feature-based image registration algorithm is designed as follows.

1. For each feature point $\bar{f}^B$ in image B, compare it with every feature point $\bar{f}^A$ of the same type within a specified neighborhood of $\bar{f}^B$ in image A. Assume that the two feature points are matched.
2. Estimate the translation vector $[\Delta x, \Delta y]^T$ and the rotation angle $\Delta\theta$ by

$$\begin{pmatrix} \Delta x \\ \Delta y \\ \Delta\theta \end{pmatrix} = \begin{pmatrix} x_0^B \\ y_0^B \\ \theta_0^B \end{pmatrix} - \begin{pmatrix} x_0^A \\ y_0^A \\ \theta_0^A \end{pmatrix}, \tag{59.1}$$

   where $[x_0^B, y_0^B, \theta_0^B]^T$ and $[x_0^A, y_0^A, \theta_0^A]^T$ represent the two corresponding reference feature points, based on which the transformation parameters are estimated.
3. Align all the feature points in image B with respect to the reference feature point $[x_0^B, y_0^B, \theta_0^B]^T$ by

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} \cos\Delta\theta & -\sin\Delta\theta \\ \sin\Delta\theta & \cos\Delta\theta \end{pmatrix} \begin{pmatrix} x - x_0^B \\ y - y_0^B \end{pmatrix} + \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix}, \tag{59.2}$$
$$\theta'_i = \theta_i + \Delta\theta$$

   where $[x, y, \theta_i]^T$ represents a feature point in image B and $[x', y', \theta'_i]^T$ represents the corresponding aligned feature point.
4. Compare the locations and orientations of the aligned feature points in image B with those in image A by calculating the Euclidean distance between the corresponding origin points and the average orientation difference by

$$\begin{aligned} \Delta d_{m,n} &= \sqrt{(x^{n\prime} - x^m)^2 + (y^{n\prime} - y^m)^2}\,, \\ \Delta\theta_{m,n} &= mean(\theta_i^{n\prime} - \theta_i^m) \end{aligned} \tag{59.3}$$

where $[x^{n\prime}, y^{n\prime}, \theta_i^{n\prime}]^T$ and $[x^m, y^m, \theta_i^m]^T$ represent the nth aligned feature point in image B and the mth feature point in image A, respectively. The two feature points are said to be matched if $\Delta d_{m,n}$ and $\Delta\theta_{m,n}$ are smaller than their specified thresholds.

5. Store the matched feature pairs and count the total number of them. Then, go to Step 1 for a new feature alignment and matching process using another pair of reference feature points. The one that produces a maximal number of matched feature pairs is considered as the best matching between image B and image A.
6. Register image B with image A based on all the matched feature pairs that correspond to the best matching, and resample the registered image to its original resolution.

59.4 Accelerated Implementation by Fine-grained Parallelism

In the system, the real-time video stream captured by the FIVE 1 endomicroscope is directly transferred to the FPGA computing system through the digital visual interface (DVI). The video data is first displayed on the touch panel screen frame by frame by means of the static random access memory (SRAM, 8 MB) on the FPGA board, and then transferred to the synchronous dynamic random access memory (SDRAM, 256 MB) on the FPGA board. Every video frame is saved in the SDRAM and the processing of these stored slice images is accomplished in the SDRAM as well. Finally, the processing results will be displayed on the touch panel screen by transferring data back to the SRAM. Figure 59.10 demonstrates the whole process.

Using the hardware descriptive languages (Handel-C and Verilog) [17, 18] and the primary developmental environments (Celoxica DK and Xilinx ISE), the image processing algorithms are optimized and implemented on the FPGA computing device (Celoxica RC340). As illustrated in Fig. 59.11, we have developed a graphical user interface with the touch panel screen that allows the users to apply various functions and filters (e.g. video setup, image processing, measurement, and magnification) in real time with a simple touch on the screen. The interactive touch-screen panel is designed with 16 menu buttons placed on the right, which can be further extended. The processed image is displayed on the top left panel and the

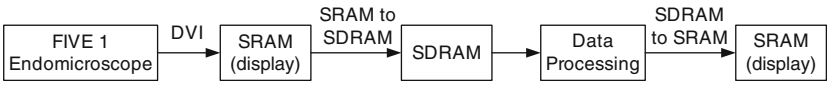


Fig. 59.10 The flowchart of the video data transferring and processing

Fig. 59.11 Snapshot of the reconfigurable real time computing system

bottom informative panel is used to display the statistical measurements of the entire image or the selected region(s) of interest.

The implementation process involves four major steps: (1) redesign each of the algorithms in Handel-C and refine the code for parallel processing of the pixels with minimal clock cycles, and then optimize the logic block arrangement and pipelining by using the registers in Verilog code, (2) synthesize the logic of the design and translate it to the gate level design, and then map the gate design to the gate primitives available in the target FPGA chip, (3) place these assigned gate primitives onto the physical gate positions on the chip, route the wires to link the gates, and generate the FPGA configuration instructions consisting of the physical layout, and (4) load these configuration instructions onto the FPGA chip. Each FPGA chip contains several millions of logic gates that are arranged as array logic elements. Based on the reconfigurable instructions loaded, the FPGA computing device will be configured with the logic and routing resources taking a particular state, allowing the design to be implemented in a fine-grained parallel and optimized manner.

To demonstrate how fine-grained parallelization of the algorithms works for the real-time processing, we present a denoising filter (3×3) implemented on the FPGA board (see Fig. 59.12). It corrects the counting statistic image acquisition defect, which is inherent to all the fluorescence images due to the low photon count. It is an optimized incomplete sorting algorithm that obtains the median value. The logic block is designed in such a way that it is pipelined into eight stages. Each of the stages involves both the basic functions of comparing two input values and copying the input value. The results from the comparing and copying are stored in nine registers, thus incurring one clock cycle for each stage, and the latency of the filter is only eight clock cycles.

In the first stage, eight of the nine neighborhood pixel values ($P_{i\ \ 1,\ j\ \ 1}$, $P_{i\ \ 1,\ j}$, $P_{i\ \ 1,\ j+1}$, $P_{i,\ j\ \ 1}$, $P_{i,\ j}$, $P_{i,\ j+1}$, $P_{i+1,\ j\ \ 1}$, $P_{i+1,\ j}$, $P_{i+1,\ j+1}$) undergo comparison through the use of four comparators, which produces four pairs of sorted values being held in registers: ($L_{1\_0} < L_{1\_1}$), ($L_{1\_2} < L_{1\_3}$), ($L_{1\_4} < L_{1\_5}$), ($L_{1\_6} < L_{1\_7}$). In the second stage, the four pairs of sorted values are further sorted using another four

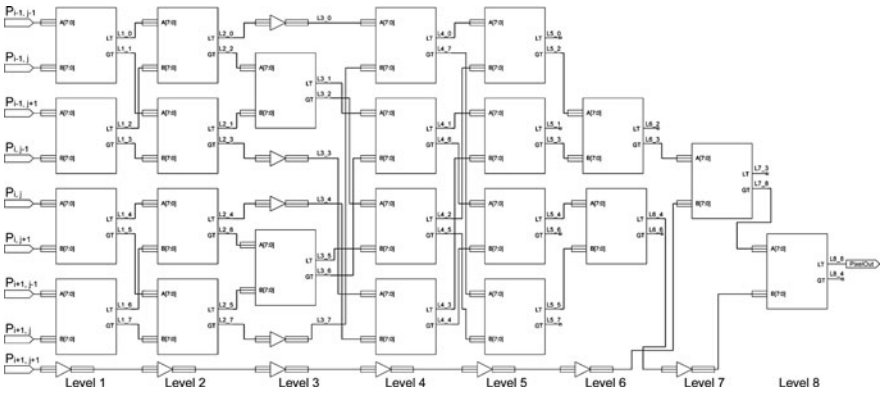


Fig. 59.12 A denoising filter implemented on the FPGA board

comparators such that the values are grouped into two groups, where the smallest and largest values of each group are found: ($L_{2\_0} < (L_{2\_1}, L_{2\_2}) < L_{2\_3}$), ($L_{2\_4} < (L_{2\_5}, L_{2\_6}) < L_{2\_7}$). In the third stage, the two unsorted middle values in each group are sorted through the use of two comparators: ($L_{3\_0} < L_{3\_1} < L_{3\_2} < L_{3\_3}$), ($L_{3\_4} < L_{3\_5} < L_{3\_6} < L_{3\_7}$). In the fourth stage, four comparators are used to combine the sorting of the two groups into one such that the four smallest values are placed in the first four registers, while the other four largest values are placed in the next four registers: ($L_{4\_0}, L_{4\_1}, L_{4\_2}, L_{4\_3}$) < ($L_{4\_4}, L_{4\_5}, L_{4\_6}, L_{4\_7}$). In the fifth stage, using four comparators, both the first four and next four registers are arranged into two groups of sorted values: (($L_{5\_0} < L_{5\_2}$), ($L_{5\_1} < L_{5\_3}$)) < (($L_{5\_4} < L_{5\_6}$), ($L_{5\_5} < L_{5\_7}$)). In the sixth stage, using two comparators, the largest value for the first four registers and the smallest value for the next four registers are obtained: ($L_{6\_0}, L_{6\_1}, L_{6\_2}$) < L_{6_3} < L_{6_4} < ($L_{6\_5}, L_{6\_6}, L_{6\_7}$). In the seventh stage, we ignore the smallest three values from the first set of four registers, and the largest three values from the next set of four registers. In this way, we optimize the comparison, as unnecessary comparison is not performed. Taking the center two values along with the last pixel value ($P_{i+1, j+1}$), which is copied in the previous six stages, we sort out the smallest value to the register L_{7_3}: $L_{7\_3} <$ (L_{7_4}, L_{7_8}). In the last stage, using a comparator, the median value is placed into the register L_{8_8} after comparison between the two register values L_{7_4} and L_{7_8}. This register value is output as the median value: $L_{8\_3} < L_{8\_8} < L_{8\_4}$.

The latency of the median denoising filter is eight clock cycles, since each of the eight levels contributes one clock cycle to the latency. Such a hardware-based solution's latency is much shorter than that of the software-based solution, which requires 22 clock cycles as all the comparisons are performed in sequential order. Furthermore, the output speed for the hardware-based solution is at one output per clock cycle, as the median denoising circuit is pipelined with registers between each of the levels. In contrast, the software-based solution can only output at the same speed as its latency, which is 22 clock cycles.

59.5 Experimental Results

To evaluate the performance of the proposed algorithms, we perform a series of experiments on a collection of 342 confocal fluorescence endomicroscopy images of the gerbil brain microvasculature. The images are captured by the FIVE 1 system at different z-depths with a submicron resolution of 1,024 $\times$ 1,024 pixels after using three fluorescent agents (nuclear labeling with acriflavine, plasma staining with FITC-dextran, and injection of labeled red blood cells) simultaneously. Figure 59.2 shows some typical examples.

For every input gray scale image, we first apply the binarization-thinning process to obtain the skeleton image and extract some salient points, including bifurcations and endpoints, as features from the skeletons according to the Rutovitz crossing number. The extracted feature points are then purified using our postprocessing techniques. Figure 59.13 shows some examples to demonstrate the postprocessing results, where different colors represent the removed feature points by respective methods (green: remaining feature points; red: eliminated breaks; yellow: eliminated spurs; cyan: eliminated dots, islands and close features; and magenta: eliminated boundary features). From the results, we can see that the remaining feature points after purification are capable of representing the structures of the blood vessels and reflecting the transformation of them between consecutive slice images. Thus, we can use these feature points for the registration of endomicroscopy images.

Figure 59.14 gives some registration results between consecutive slice images. As can be seen from the results, our image alignment algorithm based on the extracted feature points shows a good capability of correcting the distortions between endomicroscopy images, which will greatly benefit the subsequent reconstruction and visualization of the 3D structures of the blood vessels. To display the image alignment results in a better way, we pad the aligned image with "white" pixels before truncating it to the same size of the base image, and add a red box to indicate the image border. Actually, the blank areas in every aligned image will be padded with "0," without affecting the 3D reconstruction and visualization results.

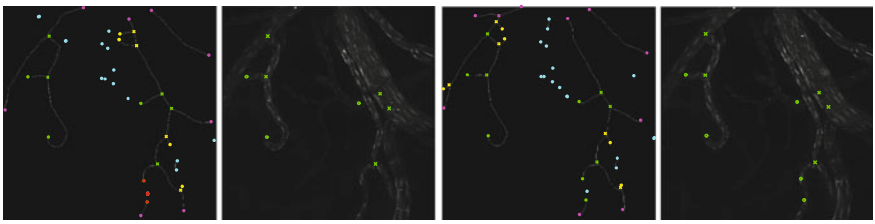


Fig. 59.13 Postprocessing results overlaying on the thinning image (*odd columns*) and the remaining feature points overlaying on the original gray scale endomicroscopy image (*even columns*)

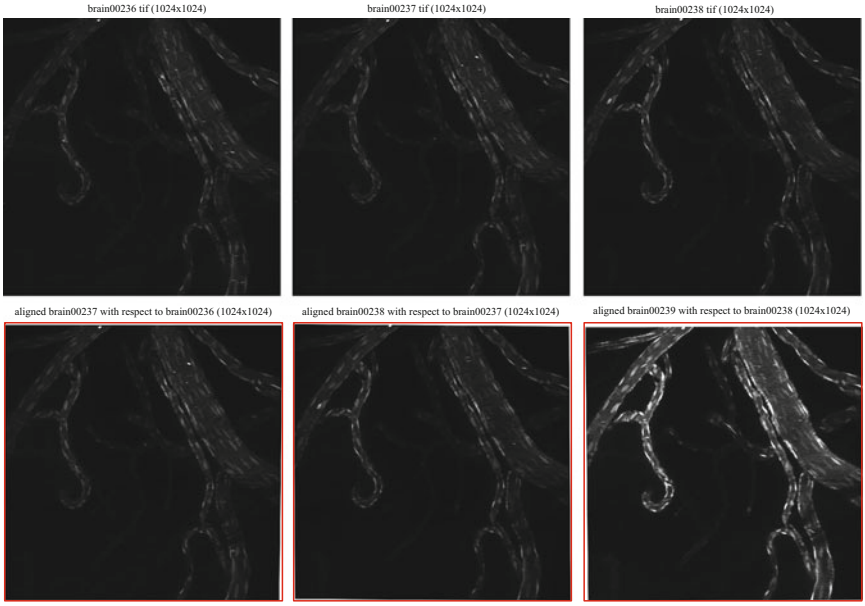


Fig. 59.14 Examples of the endomicroscopy images after registration (*bottom row*) with their corresponding base images (*top row*)

59.6 Conclusions

In this work, we have carried out a pilot study on the emerging confocal fluorescence endomicroscopy images and developed a feasible image registration algorithm based on feature detection. The experimental results conducted on a set of fluorescence images reveal that the extracted feature points of high-quality enable a good alignment between every two consecutive slice images, which will directly benefit the 3D reconstruction and visualization result of the living cells and tissues. It helps the clinicians for the diagnosis of various diseases such as the early-stage cancers in an *in vivo* noninvasive way, which will offer profound health benefits to the society by improving the survival rate and lowering the economical cost and emotional burden. In addition, it will enable further advancement in the field of basic cell biology, aid our understanding of the mechanism of disease progression, and allow the monitoring of drug effects at the cellular level.

Acknowledgments This work was supported by two grants: SBIC RP C 010/2006 from A Star Biomedical Research Council, Singapore, and AcRF/RGM 35/06 from Ministry of Education, Singapore. The authors would like to thank M. Goetz, C. Schneider, et al. (University of Mainz, Germany), who provided the fluorescence images of mouse microvasculature for this study. We are also thankful to S. Thomas (Optiscan Pty. Ltd., Australia) and our clinic partners, Prof. Soo Khee Chee, A/P Malini Olivo and Dr Patricia Thong (National Cancer Centre Singapore).

References

1. Koenig F, Knittel J, Stepp H (2001) Diagnosing cancer in vivo. Science 292:1401 1403
2. Massoud TF, Gambhir SS (2003) Molecular imaging in living subjects: seeing fundamental biological processes in a new light. Genes Dev 17:545 580
3. Hoffman RM (2005) The multiple uses of fluorescent proteins to visualize cancer in vivo. Nat Rev Cancer 5:796 806
4. Berg K et al (2005) Porphyrin related photosensitizers for cancer imaging and therapeutic applications. J Microsc 218:133 147
5. Zheng W et al (2002) Detection of neoplasms in the oral cavity by digitized endoscopic imaging of 5 aminolevulinic acid induced protoporphyrin IX fluorescence. Int J Oncol 21:763 768
6. Gerger A et al (2005) Diagnostic applicability of in vivo confocal laser scanning microscopy in melanocytic skin tumors. J Invest Dermatol 124:493 498
7. Thong PSP et al (2007) Laser confocal endomicroscopy as a novel technique for fluorescence diagnostic imaging of the oral cavity. J Biomedical Optics 12(1):014007.1 8
8. Kiesslich R et al (eds) (2008) Atlas of endomicroscopy. Springer Medizin Verlag, Heidelberg
9. Goetz M et al (2007) In vivo confocal real time mini microscopy in animal models of human inflammatory and neoplastic diseases. Endoscopy 39:350 356
10. Goetz M et al (2008) Dynamic imaging of microvasculature and perfusion by miniaturised confocal laser microscopy. Eur Surg Res 41:290 297
11. O'Conor K, Voorheis HP, O'Sullivan C (2004) 3D visualisation of confocal fluorescence microscopy data. In: Proc Fifth Irish Workshop Computer Graphics. pp 49 54
12. Wang Q, Sun Y, Rajwa B, Robinson JP (2006) Interactive volume visualization of cellular structures. In: Proc SPIE. Volume 6065. 60651D.1 12
13. Wang Q, Sun Y, Robinson JP (2007) GPU based visualization techniques for 3D microscopic imaging data. In: Proc SPIE. Volume 6498. 64981H.1 12
14. Mosaliganti K et al (2008) Reconstruction of cellular biological structures from optical microscopy data. IEEE Trans Vis Comput Graph 14(4):863 876
15. Ohtake T et al (2001) Computer assisted complete three dimensional reconstruction of the mammary ductal/lobular systems. Cancer 91(12):2263 2272
16. Almsherqi ZA, Kohlwein SD, Deng Y (2006) Cubic membranes: a legend beyond the flatland of cell membrane organization. Cell Biol 173(6):839 844
17. Celoxica Ltd (2005) Handel C language reference manual
18. Palnitkar S (2003) Verilog HDL: a guide to digital design and synthesis. Prentice Hall PTR

Chapter 60
Kinetic Models for Cancer Imaging

V.J. Schmidvolker

Abstract As tumors have distinctly different blood flow compared to that of normal tissue, the kinetic in cancerous tissue is of importance in cancer diagnosis and in assessing the efficacy of treatment. To this end, dynamic cancer imaging provides a noninvasive way for early detection of tumors and subsequent treatment planning. This paper provides an overview of currently available imaging modalities and compares different kinetic models used for analyzing tumor scans. Specific research issues that arise when analyzing dynamic imaging scans are examined and current developments in the field are highlighted.

Keywords Cancer imaging · Detection of tumors · Imaging modalities · Kinetic models · MRI · PET · SPECT

60.1 Introduction

In the last 25 years, several medical imaging modalities that assess the kinetics of tissue in vivo have been developed. The kinetics in cancerous tissue is significantly different from the surrounding structures: Tumors initiate angiogenesis and destroy the integrity of normal tissue, leading to increased blood flow. To image the kinetic processes, dynamic imaging is an important clinical tool. A tracer is typically transported via the blood stream, and so from the time curve of tracer concentration one can derive the blood flow in the tissue in question. Dynamic imaging provides a noninvasive way of not only detecting tumors, but also evaluating the status of the tumor and tracking or quantifying the effects of treatment.

The aim of quantitative analysis of medical images is to estimate biologically meaningful parameters which summarize the physical processes in the tissue, and

V.J. Schmidvolker
Department of Statistics, Ludwig Maximilians University, Munich, Germany
e mail: schmid@stat.uni muenchen.de

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 60,

can be compared across scans, patients, and patient groups. Quantitative analysis is usually performed by fitting a kinetic model to the time curve of tracer concentration, either per pixel or based on average tracer concentration in a region of interest (ROI). In this review, we discuss ways to improve the robustness of kinetic parameter estimation and how to quantify the accuracy of kinetic model fitting.

60.2 Imaging Modalities

Positron Emission Tomography (PET) has been used to study tumor kinetics for over 25 years [1]. PET uses a radionuclide as tracer and a ring of Gamma detectors to collect the photon emissions. Similarly, Single Photon Emission Computed Tomography (SPECT) uses a radioligand as tracer. Reconstruction techniques for PET and SPECT are difficult and spatial resolution is low, however, temporal resolution is high. PET is used frequently in cancer imaging, and SPECT is also been used in oncology, especially in lung and breast cancer [2].

In contrast, imaging modalities based on Magnetic Resonance Imaging (MRI) do not involve ionizing radiation. They can be used in different imaging settings to assess the kinetic processes in the tissue. MRI is widely available and data acquisition is fast. Furthermore, signal-to-noise ratio (SNR), and, hence, contrast-to-noise ratio (CNR), are relatively high.

By using a low molecular weight contrast agent (less than 1,000 Da in molecular weight, often a Gadolinium complex such as Gd-DTPA, Gd-DOTA, or Gd-HP-DO3A [3]), Dynamic Contrast-Enhanced (DCE-)MRI relies on the reduction in T1 relaxation time caused by a contrast agent [4]. Large molecular agents, known as macromolecular contrast media (MMDC) or blood-pool agents, are in preclinical development and not yet approved for human use, but have shown potential for applications in cancer imaging [3].

Contrast agent concentration C_t are computed from the T1 signal by converting the signal into T1 relaxation time values using proton density weighted images and data from calibration phantoms with known T1 relaxation times. Baseline T1 values can be computed as mean value of the dynamic images before contrast arrival, or more accurately, using a sequence with different flip angles or TR.

DCE-MRI has been proven to be useful in imaging a range of different tumor types. Probably, most work in DCE-MRI is done in breast cancer. Several studies showed that early DCE-MRI can be used as a biomarker for changes in tumor angiogenesis, and hence for success or failure of a treatment [5].

For Dynamic Susceptibility Contrast (DSC-)MRI, the effect of magnetic inhomogeneities caused by a bolus of contrast agent passing through a capillary bed is exploited [6]. Typically, this is observed via T2*-weighted sequences, but T1-weighted sequences can also be used. Scans derived from this technique typically have a lower spatial resolution in favor of a higher time resolution Arterial Spin Labeling (ASL) uses magnetically labeled water protons as tracers for MRI scans. Therefore, this technique is the only completely noninvasive method for

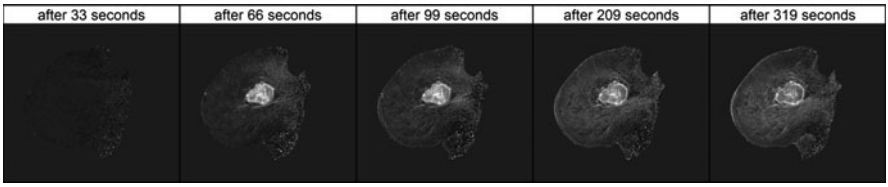


Fig. 60.1 Time series of DCE MR images from a breast cancer patient. Fast enhancement of the tumor is clearly visible

imaging perfusion in vivo. The recent availability of 3T and higher field MRI scanners permit clinical applications of ASL [7], however, to date applications of ASL are mainly focused on neuroimaging.

A relatively new imaging technique is Dynamic Contrast-Enhanced Computer Tomography (DCE-CT), also known as Perfusion-CT or Multi-Detector row CT (MDCT). It uses contrast agents similar to DCE-MRI, but scans are acquired via CT. DCE-CT allows a much higher spatio-temporal resolution, but at a cost of exposing patients to ionizing radiation. Initial clinical applications of DCE-CT demonstrate high potential in oncology, but its clinical values have yet to be established [8] (Fig. 60.1).

60.3 Kinetic Models

Kinetic models are typically based on the idea of one or more compartments in the tissue, which exchange the tracer via perfusion. The exchange between the compartments can be described via a set of differential equations.

The simplest compartment model is the Kety model, in which the tracer is assumed to be washed in via the vascular compartment, and exchange between the vascular compartment and the main compartment is driven by two parameters, the influx rate K^{trans} and the efflux rate k_{ep} (standardized quantities for DCE-MRI as defined by Tofts et al. [9]). Under the restriction $C_t(0) = 0$, the solution of the system of differential equations is

$$C_t(t) = C_p(t) \otimes K^{\text{trans}} \exp(-k_{\text{ep}} t),$$

where $\otimes$ denotes convolution and C_p is the arterial input function (AIF). This model was developed first by Kety in 1960 [10]. Later, Koepe et al. described a Kety model for PET [11], and for DCE-MRI, Larsson and Tofts independently developed similar models [12].

The physiologic interpretation of the transfer constant K^{trans} depends on the capillary permeability and blood flow in the tissue. In case of low permeability, the transfer constant is $K^{\text{trans}} = PS\rho$, with P the permeability of the capillary wall, S the surface area, and ρ the density of tissue. In case of high permeability, influx is limited by flow and $K^{\text{trans}} = F\rho(1 - \text{Hct})$, with F the blood flow and Hct the Hematocrit.

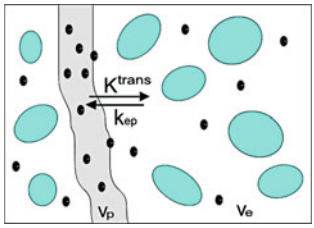


Fig. 60.2 Compartmental model of kinetics in the tissue. Tracer is delivered via the vascular space, perfuses into the extracellular extravascular space with rate K^{trans}, perfuses back to vascular space with rate k_{ep}, and is eventually washed out

The single compartment Kety model is often not complex enough to give a good fit to the data. A natural extension is to assume more compartments in the tissue. In MRI studies, typically only an additional vascular compartment is assumed (extended Kety model). In PET studies, however, often two or three compartments or even an unknown number of compartments are used [13].

Compartment models are intrinsically a simplification of the true kinetic processes in the tissue. For example, they assume that the tracer is instantaneously well mixed in each of the compartments. In contrast, Distributed Parameter Models (DPM) like the adiabatic approximation to the tissue homogeneity (AATH) overcome these assumptions [14]. However, these models can typically only be used for imaging modalities with high temporal resolution like DSC-MRI and DCE-CT. In contrast-based MRI studies, the tracer concentration is observed indirectly via the change of magnetization of H_2O. To account for water diffusing between compartments thermodynamic pharmacokinetic models can be used [15], but due to the complexity of these models the water exchange effect is often neglected (Fig. 60.2).

Kinetic models are always an approximation to the true physiological processes in the tissue and the fit between model and observed data is often poor. Empirical Mathematical Models (EMM) are an alternative and often allow for a better model fit in practice [16]. However, parameters in empirical models typically do not have a proper biological interpretation. Spline-based models are even nonparametric, but although they do not provide kinetic parameters, they can be used for denoising before model fitting [17].

60.4 Numerical Optimization

Numerical optimization of kinetic models is typically based on least-square techniques, i.e., the minimization of the sum of squared errors between fitted model and observed data. The intrinsic assumption of these techniques is that the distribution of the noise in the observed data is symmetric around the mean. However, the distribution of the observation error is typically unknown. In MRI, the observation error of the T1 signal is known to have a Ricean distribution [18], which

is approximately Gaussian for high SNR, but tracer concentration is a nonlinear transformation of the T1 signal. Hence, the error distribution cannot be determined analytically.

As all kinetic models are nonlinear models, optimization is difficult and often sensitive to starting properties of the optimization algorithm. Levenberg Marquardt, MINPACK-1, Simplex minimization, and quasi-Newton bounded minimization algorithms are used for optimization. No algorithm is superior, and multiple search start point algorithms are necessary to gain reliable estimates [19]. In theory, deconvolution of the input function before optimization can help to ease the optimization. Early approaches used deconvolution with Fourier methods, which however is unstable and very sensitive to noise [20].

The Bayes framework is an alternative technique for optimization. With this approach, prior information about the kinetic parameters is used, for example, a biological reasonable range for the kinetic parameters. This allows for a more robust estimation of kinetic parameters and significantly reduces fit failures and avoids biological unrealistic parameters 21.

However, Bayesian models can include not only local, pixel-wise information, but also the information about the context of a kinetic model. For example, spatial dependencies between pixels can be used as prior information. Here, adaptive estimation of smoothing parameters is essential, as this allows both the existence of locally homogeneous regions and quite sharp boundaries between drastically different tissue types, for example, normal tissue versus tumor, but also sharp features in the tumor 21.

For the initial peak of the concentration series, one can assume that the kinetic process is only driven by the inflow of the tracer, but not yet by the outflow. Therefore, Patlak et al. [22] proposed to neglect the washout parameter and developed a linear model by dividing the extended Kety model by the input function. In general, a kinetic model can always be reduced to a linear problem by disregarding the continuous property of both CTC and AIF [23]. However, this method only works for data with high time resolution. As alternative a set of precomputed basis functions representing the most important components of a set of standard kinetic curves [24] or representing a B-Spline basis [17] can be used. With the latter approach, more complex models can be fitted to the deconvolved and denoised data even with low temporal resolution.

So far, little work has been done in investigating the ways to quantify the estimation error of a given scan. More often, theoretical errors in the estimation of kinetic parameters are assessed via reproducibility or simulation studies [25]. However, the precision of the parameter estimation is an important issue, not only for automatic quality control, but also for assessing the accordance of model and actual imaging data.

Bayesian approaches implicitly permit the assessment of the accuracy of parameter estimation. Inference on kinetic parameters is based on the probability density function (PDF) a posteriori. The posterior PDF naturally provides information about the uncertainty of parameter estimation and allows for the computation of estimation errors or credible intervals for kinetic parameters 21. In addition, tumor masks can

be derived from the uncertainty information. As an alternative, bootstrap methods can be utilized to investigate the precision of kinetic parameter estimation [26].

60.5 Input Function and Onset

The input function is a major element in all kinetic models. An inaccurate AIF has direct implications on the kinetic parameters, in particular on K^{trans} and v_e [27]. In images where the feeding vessel is in the field of view, the AIF can be measured directly from the vessel. Horsfield and Morgan [20] compared three different ways of handling the discrete samples of the AIF: Approximation as Dirac impulse, piecewise constant, or piecewise linear function. In their study, the piecewise constant representation performed best.

To reduce noise, a function can be fitted to the observed AIF per subject or averaged over the whole study population. A parametric function described by Tofts and Kermode [28] is now used in many studies. This technique benefits from the denoising of the observed AIF. Henderson et al. [29] pointed out that a temporal resolution of one scan per second is necessary to sample the AIF properly. However, Rijpkema et al. [30] showed that individual AIFs can substantially reduce the variation between successive measurements. The selection of arterial voxels can be done in an automatic algorithm, and such techniques have been proposed for DCE-MRI and DSC-MRI. When the AIF is measured, one has to be careful to choose a vessel near the tumor. For example, in liver cancer DCE-MRI reproducibility is significantly improved when using an input function measured in the spleen, rather than in the aorta [31]. Reference region (RR) techniques use dynamic data from a region with known kinetic parameters, typically a muscle [32]. More commonly, the AIF is taken from literature, using the results of Weinman [33] or Fritz-Hansen et al. [34].

The time between the injection of tracer and the arrival in the tissue under study is generally unknown. The time of arrival can be determined from the first enhancement of the tracer time curve in the tissue. Similar to the input function, the enhancement onset time is an essential parameter for an accurate estimation of kinetic parameters [35].

Cheong et al. [36] proposed to fit a quadratic function to the first few data points for all possible onset times. The time point, which minimizes the sum of squared errors, is chosen as the onset time. Alternatively, the onset time can be handled as an unknown parameter in the optimization of the kinetic model, however, the numerical estimation of the parameter is not trivial [37].

60.6 Registration

For all imaging techniques described above, a dynamic series of 3D images is acquired over a time period of up to 10 min. So for all techniques, motion is a big issue. Motion is mainly caused by two reasons: body movement of the patient

and respiratory-induced organ deformation. The latter is only important for tumors in the torso, especially in the lungs and liver. Scans in this area are often done with a breath hold protocol to minimize motion artifacts. Some groups choose to acquire dynamic images during breath-hold only. This, however, results in a contrast concentration time series with large gaps, which is hard to fit. Navigator techniques can help to adjust for breathing and heart activity, but fail with large movements [38].

Nonrigid registration using free-form deformations can be applied to DCE-MRI scans of the breast [39]. However, registration of the dynamic images can be problematic due to the large differences in image intensity, especially in the area of interest surrounding the tumor mass. In a recent approach, Buonaccorsi et al. [40] proposed to register and model dynamic series simultaneously in an iterative algorithm. This allows using the information from the dynamic modeling for a better registration of the images. However, so far, the proposed algorithms are not guaranteed to converge and further work in this area is needed.

60.7 Conclusion

Complex kinetic models beyond the simple Kety model based on multicompartmental or distributed parameter models are generally more appropriate for the kinetic processes in the tissue. They also provide further insights into tumor angiogenesis, microvessel structure and function. Technical advances in medical imaging have resulted in increased temporal and spatial resolution, but even with these improved data, more robust optimization techniques are essential.

Recent developments focus on connecting different steps in kinetic model optimization and try to improve optimization by using information from more than one source in contextual kinetic models. Simultaneous registration and modeling of tracer concentration curves allows for the use of information from kinetic modeling for a better registration of the images and vice versa. Multiple reference tissue methods utilize the data acquired in different areas of normal tissue to gain a robust estimate of the input function. Spatial frameworks are based on the fact that spatially adjacent pixels may be similar in their kinetic properties and they allow for a large reduction in estimation errors compared to a pixel-wise algorithm. For clinical drug trial studies, recent work proposes to analyze all scans from all patients simultaneously in a global kinetic model. In this way, the estimation of treatment effect is expected to be much more robust [41].

A range of different modalities is available to evaluate tissue kinetics with dynamic imaging. Although each modality has specific advantages and disadvantages, they all assess the same kinetic processes in the tissue. Technical advances like the introduction of high-field MR scanners, improvements in scanning protocols and new optimization approaches, including contextual kinetic models, will provide improved understanding of the initial onset and the progression of disease, as well as the effect of different treatment regimes.

References

1. R. A. Hawkins and M. E. Phelps, “PET in clinical oncology,” *Cancer Metastasis Rev*, vol. 7, no. 2, pp. 119 142, 1988.
2. K. Yutani, *et al*., “Comparison of FDG PET with MIBI SPECT in the detection of breast cancer and axillary lymph node metastasis,” *J Comput Assist Tomogr*, vol. 24, no. 2, pp. 274 280, 2000.
3. H. Gries, “Extracellular MRI contrast agents based on gadolinium,” In *Contrast Agents I*, Berlin, Springer, pp. 1 24, 2002.
4. G. J. M. Parker and A. R. Padhani, “T1 w DCE MRI: T1 weighted dynamic contrast enhanced MRI,” In *Quantitative MRI of the Brain*, P. Tofts (Ed.), Chichester, Wiley, pp. 341 364, 2003.
5. J. P. Delille, *et al*., “Invasive ductal breast carcinoma response to neoadjuvant chemotherapy: noninvasive monitoring with functional MR imaging Pilot study,” *Radiology*, vol. 228, pp. 63 69, 2003.
6. N. S. Akella, *et al*., “Assessment of brain tumor angiogenesis inhibitors using perfusion magnetic resonance imaging: quality and analysis results of a phase I trial,” *J Magn Reson Imaging*, vol. 20, no. 6, pp. 913 922, 2004.
7. E. T. Petersen, *et al*., “Non invasive measurement of perfusion: a critical review of arterial spin labelling techniques,” *Br J Radiol*, vol. 79, no. 944, pp. 688 701, 2006.
8. V. Goh and A. Padhani, “Imaging tumor angiogenesis: functional assessment using MDCT or MRI?,” *Abdominal Imaging*, vol. 31, no. 2, pp. 194 199, 2006.
9. P. S. Tofts, *et al*., “Estimating kinetic parameters from dynamic contrast enhanced T1 weighted MRI of a diffusable tracer,” *J Magn Reson Imaging*, vol. 10, pp. 223 232, 1999.
10. S. Kety, “Blood tissue exchange methods. Theory of blood tissue exchange and its applica tions to measurement of blood flow,” *Methods Med Res*, vol. 8, pp. 223 227, 1960.
11. R. A. Koeppe, J. E. Holden, and W. R. Ip, “Performance comparison of parameter estimation techniques for the quantitation of local cerebral blood flow by dynamic positron computed tomography,” *J Cereb Blood Flow Metab*, vol. 5, no. 2, pp. 224 234, 1985.
12. H. B. Larsson and P. S. Tofts, “Measurement of the blood brain barrier permeability and leakage space using dynamic Gd DTPA scanning a comparison of methods,” *Magn Reson Med*, vol. 24, no. 1, pp. 174 176, 1992.
13. V. J. Cunningham and T. Jones, “Spectral analysis of dynamic PET studies,” *J Cereb Blood Flow Metab*, vol. 13, pp. 15 23, 1993.
14. K. S. St. Lawrence and T Y. Lee, “An adiabatic approximation to the tissue homogeneity model for water exchange in the brain: I. Theoretical derivation,” *J Cereb Blood Flow Metab*, vol. 18, pp. 1365 1377, 1998.
15. C. Landis, *et al*., “Determination of the MRI contrast agent concentration time course in vivo following bolus injection: effect of equilibrium transcytolemmal water exchange,” *Magn Reson Med*, vol. 44, pp. 567 574, 2000.
16. X. Fan, *et al*., “New model for analysis of dynamic contrast enhanced MRI data distinguishes metastatic from nonmetastatic transplanted rodent prostate tumors,” *Magn Reson Med*, vol. 51, no. 3, pp. 487 494, 2004.
17. V. J. Schmid, *et al*., “Quantitative analysis of dynamic contrast enhanced MR images based on Bayesian P splines,” *IEEE Trans Med Imaging*, vol. 28, pp. 789 798, 2009.
18. H. Gudbjartsson and S. Patz, “The Rician distribution of noisy MRI data,” *Magn Reson Med*, vol. 34, no. 6, pp. 910 914, 1995.
19. T. S. Ahearn, *et al*., “The use of the Levenberg Marquardt curve fitting algorithm in phar macokinetic modelling of DCE MRI,” *Phys Med Biol*, vol. 50, pp. N85 N92, 2005.
20. M. A. Horsfield and B. Morgan, “Algorithms for calculation of kinetic parameters from T1 weighted dynamic contrast enhanced magnetic resonance imaging,” *J Magn Reson Imag ing*, vol. 20, pp. 723 729, 2004.

21. V. J. Schmid, *et al*., "Bayesian methods for pharmacokinetic models in dynamic contrast enhanced magnetic resonance imaging," *IEEE Trans Med Imaging*, vol. 25, pp. 1627 1636, 2006.
22. C. S. Patlak, *et al*., "Graphical evaluation of blood to brain transfer constants from multiple time uptake data," *J Cereb Blood Flow Metab*, vol. 3, pp. 1 7, 1983.
23. K. Murase, "Efficient method for calculating kinetic parameters using T1 weighted dynamic contrast enhanced magnetic resonance imaging," *Magn Reson Med*, vol. 51, pp. 858 862, 2004.
24. A. L. Martel, "A fast method of generating pharmacokinetic maps from dynamic contrast enhanced images of the breast," In *MICCAI 2006*, R, Larsen, M. Nielsen, J. Sporring (Eds.), Berlin, Springer, pp. 101 108.
25. D. L. Buckley, "Uncertainty in the analysis of tracer kinetics using dynamic contrast enhanced T1 weighted MRI," *Magn Reson Med*, vol. 47, no. 3, pp. 601 606, 2002.
26. L. E. Kershaw and D. L. Buckley, "Precision in measurements of perfusion and microvascular permeability with T1 weighted DCE MRI," *Magn Reson Med*, vol. 56, pp. 986 992, 2006.
27. R. Walker, J. Paratz, and A. E. Holland, "Reproducibility of the negative expiratory pressure technique in COPD," *Chest*, vol. 132, no. 2, pp. 471 476, 2007.
28. P. S. Tofts and A. G. Kermode, "Measurement of the blood brain barrier permeability and leakage space using dynamic MR imaging," *Magn Reson Med*, vol. 17, pp. 357 367, 1991.
29. E. Henderson, B. K. Rutt, and T. Y. Lee, "Temporal sampling requirements for the tracer kinetics modeling of breast disease," *Magn Reson Imaging*, vol. 16, no. 9, pp. 1057 1073, 1998.
30. M. Rijpkema, *et al*., "Method for quantitative mapping of dynamic MRI contrast agent uptake in human tumors," *J Magn Reson Imaging*, vol. 14, no. 4, pp. 457 463, 2001.
31. H. W. van Laarhoven, *et al*., "Method for quantitation of dynamic MRI contrast agent uptake in colorectal liver metastases," *J Magn Reson Imaging*, vol. 18, no. 3, pp. 315 320, 2003.
32. C. Yang, *et al*., "Multiple reference tissue method for contrast agent arterial input function estimation," *Magn Reson Med*, vol. 58, no. 6, pp. 1266 1275, 2007.
33. H. J. Weinmann, *et al*., "Pharmokinetics of Gd DTPA/Dimeglumine after intravenous injec tion into healthy volunteers," *Physiol Chem Physics Med NMR*, vol. 16, pp. 167 172, 1984.
34. T. Fritz Hansen, *et al*., "Measurement of the Arterial Concentration of Gd DTPA Using MRI: A step toward Quantitative Perfusion Imaging," *Magn Reson Med*, vol. 36, pp. 225 231, 1996.
35. F. Calamante, D. G. Gadian, and A. Connelly, "Delay and dispersion effects in dynamic susceptibility contrast MRI: simulations using singular value decomposition," *Magn Reson Med*, vol. 44, pp. 466 473, 2000.
36. L. H. Cheong, T. S. Koh, and Z. Hou, "An automatic approach for estimating bolus arrival time in dynamic contrast MRI using piecewise continuous regression models," *Phys Med Biol*, vol. 48, no. 5, pp. N83 N88, 2003.
37. M. R. Orton, *et al*., "Bayesian estimation of pharmacokinetic parameters for DCE MRI with a robust treatment of enhancement onset time," *Phys Med Biol*, vol. 52, pp. 2393 2408, 2007.
38. H. W. Korin, *et al*., "Adaptive technique for three dimensional MR imaging of moving structures," *Radiology*, vol. 177, no. 1, pp. 217 221, 1990.
39. E. R. Denton, *et al*., "Comparison and evaluation of rigid, affine, and nonrigid registration of breast MR images," *J Comput Assist Tomogr*, vol. 23, no. 5, pp. 800 805, 1999.
40. G. A. Buonaccorsi, *et al*., "Comparison of the performance of tracer kinetic model driven registration for dynamic contrast enhanced MRI using different models of contrast enhance ment," *Acad Radiol*, vol. 13, no. 9, pp. 1112 1123, 2006.
41. V. J. Schmid, *et al*., "A Bayesian hierarchical model for the analysis of a longitudinal DCE MRI oncology study," *Magn Reson Med*, vol. 61, pp. 163 174, 2009.

Chapter 61
Using Web and Social Media for Influenza Surveillance

Courtney D. Corley, Diane J. Cook, Armin R. Mikler, and Karan P. Singh

Abstract Analysis of Google influenza-like-illness (ILI) search queries has shown a strongly correlated pattern with Centers for Disease Control (CDC) and Prevention seasonal ILI reporting data. Web and social media provide another resource to detect increases in ILI. This paper evaluates trends in blog posts that discuss influenza. Our key finding is that from 5th October 2008 to 31st January 2009, a high correlation exists between the frequency of posts, containing influenza keywords, per week and CDC influenza-like-illness surveillance data.

Keywords Health informatics · Disease surveillance · Public health epidemiology · Information retrieval · Social media analytics

61.1 Introduction

Influenza diagnosis based solely on the presentation of symptoms is limited as these symptoms may be associated with many other diseases. Serologic and antigen tests require that a patient with influenza-like-illness (ILI) be examined by a physician who can either conduct a rapid diagnosis test or take blood samples for laboratory testing. This suggests that many cases of influenza remain undiagnosed. While the presence of influenza in an individual can be confirmed through specific diagnostic tests, the influenza prevalence in the population at any given time is unknown and can only be estimated. In the past, such estimates have relied solely on the extrapolation of diagnosed cases, making it difficult to identify the various phases of seasonal influenza, or the identification of a more serious manifestation of a flu epidemic.

C.D. Corley (✉)
Pacific Northwest National Laboratory, Richland, WA, USA
e mail: court@pnl.gov

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 61,

Web and social media (WSM) provide a resource to detect increases in ILI. This paper evaluates blog posts that discuss influenza; analysis show a significant correlation with the US 2008 2009 seasonal influenza epidemic. We briefly discuss a history of infectious disease outbreaks, and recent approaches in online public health surveillance of influenza are discussed with regard to outbreak responses. Next, the dataset used in our analysis is presented, and the methodology for information extraction and trend analysis is outlined posting trends. Through discovery and verification of trends in influenza-related blogs, we verify a correlation to Centers for Disease Control and Prevention (CDC) influenza-like-illness patients reporting at sentinel healthcare providers.

61.1.1 Background

Epidemics of infectious diseases have plagued humankind since historical times. There are accounts of epidemics dating back to the times of Hippocrates (459 377 BC) and the ancient Greeks [1]. Fourteenth century Europe lost a quarter of its 100 million people to Black Death. The fall of the Aztec empire in 1521 was due to smallpox that eradicated half of its 3.5 million population. The pandemic influenza of 1918 caused over 20 million excess deaths in 12 months. More recently, the severe acute respiratory syndrome (SARS) outbreak of 2003 highlighted the rapid spread of an epidemic at the global level. The outbreak, emanating from a small Guangzhou province in China, spread around the world requiring a concerted response from public health administrations around the world and the World Health Organization (WHO) to curtail the epidemic [5]. The WHO and CDC [2] actively engage in worldwide surveillance of infectious diseases, and prioritize prevention and control measures at the root cause of epidemics.

The pervasiveness and ubiquity of Internet and World Wide Web resources provide individuals with access to many information sources that facilitate self-diagnosis; one can combine specific disease symptoms to form search queries. The results of such search queries often lead to sites that may help diagnose the illness and offer medical advice. Recently, Google has addressed this issue by capturing the keywords of queries and identifying specific searches that involve search terms that indicate influenza-like-illness (ILI) [4]. Published research on influenza Internet surveillance also includes search "advertisement click-through" [3] using a set of Yahoo search queries containing the words *flu* or *influenza* [9] and health website access logs [6, 7]. Other information sources, such as telephone triage services, can be useful to ILI detection. The findings in Yih et al. [11] show that telephone triage service is not a reliable measure for influenza surveillance due to service coverage; however, it may be beneficial in certain situations where other surveillance measures are inadequate.

61.2 Data and Methodology

Spinn3r (http://www.spinn3r.com) is a Web and social media indexing service that conducts real-time indexing of all blogs, with a throughput of over 100,000 new blogs indexed per hour. Blog posts are accessed through an open source Java application-programming interface (API). Metadata available with this dataset includes the following (if reported): blog title, blog URL, post title, post URL, date posted (accurate to seconds), description, full HTML encoded content, subject tags annotated by author, and language.

Data are selected from an arbitrary time period of 20 weeks, beginning on 5th October 2008 and ending on 31st January 2009 (total 97,955,349; weblogs 69,627,831; forums 1,986,656; mainstream media 21,543,027; others 4,797,835). Weblog, micro-blog, and mainstream media items containing characteristic keywords are extracted from Web and online social media published at the same time frame. Characteristic keywords include flu, influenza, H5N1, H3N1, and other keywords relevant to this task. This paper defines the *blog-world* to be English language, nonspam, blog posts. We also consider the following terms to be equivalent: blog post & blog item and blogger & blog site. Indexing, parsing, and link extraction code were written in Python, parallelized using pyMPI, and executed on a cluster at the Center for Computational Epidemiology and Response Analysis at the University of North Texas. This computed resource has eight nodes (2.66 GHz Quad Core Xeon processors), 64 core, 256 GB memory, and 30 TB of network storage [8, 10].

In our analysis, we extract English language items from the blog-world index when a lexical match exists to the terms *influenza* and *flu* anywhere in its content (misspellings and synonyms are not included). The blog items are grouped by month, week (Sunday to Saturday), and by day of week. The extracted blog items containing influenza keywords are termed flu-content posts or *FC posts*. FC post trends can be monitored using the social media mining methodology presented in this paper. This methodology facilitates identification of outbreaks or increases of influenza infection in the population. This paper's most significant finding is a strong correlation between the frequency of FC posts per week and Centers for Disease Control and Prevention influenza-like-illness surveillance data.

61.3 Results

We hypothesize that the frequency of blog-world flu posts correlates with a patient reporting of influenza-like-illness during the US flu-season. To verify this statement, we compare our data with Centers for Disease Control and Prevention surveillance reports from sentinel healthcare providers. The CDC website states the Outpatient Influenza-like-illness Surveillance Network (ILINet) consists of about 2,400 healthcare providers in 50 states reporting approximately 16 million

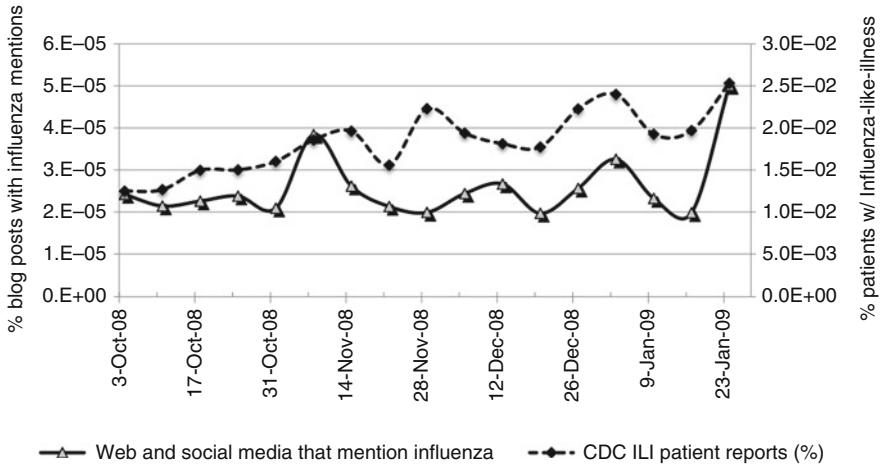


Fig. 61.1 CDC ILINet vs. normalized FC post frequency per week. Each FC posts per week data point is normalized by the corresponding blog world posts per week count. Pearson correlation 0.626, 95% confidence

patient visits each year. Each provider reports data to CDC on the total number of patients seen and the number of those patients with influenza-like-illness (ILI) by age group. For this system, ILI is defined as fever (temperature of 100°F [37.8°C] or greater) and a cough and/or a sore throat in the absence of a known cause other than influenza [2].

The CDC ILINet surveillance and FC post per week data are plotted in Fig. 61.1. CDC influenza-like-illness symptoms per visit at sentinel US healthcare providers labels the primary *Y*-axis. The secondary *Y*-axis marks the FC post per week frequency normalized by the corresponding blog-world week post count. Correlation between the two data series was measured with a Pearson correlation coefficient, r. Evidence to support our hypothesis (a correlation exists between CDC ILINet reports and Web and social media mined FC-post frequency) is the Pearson's correlation coefficient evaluated between the two data series. The Pearson correlation evaluates to unity if the two data series are exactly matching, $r = 1$. If no correlation exists between the data series, the Pearson correlation evaluates to zero, $r = 0$.

In our analysis, the 20 ILI and FC-post data points correlate strongly with a high Pearson correlation, $r = 0.626$, and the correlation is significant with 95% confidence.

61.4 Future Work

Once FC posts have been extracted, one can further monitor influenza outbreaks by evaluating the perspective of blog authors. Bloggers having a direct knowledge of influenza infection are more valuable to disease surveillance than those who author

objective or opinion items. Bloggers who persistently author FC posts are less likely to be infected with influenza and more likely to be writing about avian influenza (bird flu). The following post excerpts demonstrate influenza-content author perspective.

> **Self Identified**: "What began as an irritating cold became what I think might be the flu last night. I woke up in bed around three this morning with sore muscles, congested lungs/nose and chills running throughout my body."
> **Secondhand**: "According to ESPN.com, Ravens quarterback Troy Smith has lost 'a considerable amount of weight' while being hospitalized with tonsillitis and flu like symptoms. Smith and veteran Kyle Boller likely won't play in Sunday's season opener, leaving the workload to rookie Joe Flacco and Joey Harrington, who was signed Monday."
> **Objective (or opinion)**: "Domesticated birds may become infected with avian influenza virus through direct contact with infected waterfowl or other infected poultry, or through contact with surfaces or materials like that of water or feed that have been contaminated with the virus."

Identifying the perspective of influenza keyword posts facilitates determining its contribution to disease surveillance; three author perspectives have been identified. A FC post is either a self-identification of having ILI symptoms, secondhand (or by proxy) of another individual having ILI or the post is an opinion or objective article containing ILI keywords. Secondhand knowledge can be writing about a friend, schoolmate, family member or co-worker, but a blogger could also post details on a famous individual such as a sports player. The season opening of American football coincides with the data and many FC posts identify athletes who are unable to play because of an ILI. Automatic classification of the influenza-post author's perspective is ongoing research.

61.5 Conclusion

Web and social media provide a novel disease surveillance resource. We presented a method that evaluates blog posts containing keywords influenza or flu, and the results from analysis show strong co-occurrence with the US 2008 2009 flu season. This paper's key finding is that from 5th October 2008 to 31st January 2009, a high correlation exists between the frequency of posts, containing influenza keywords per week, and Centers for Disease Control and Prevention influenza-like-illness surveillance data.

Acknowledgments We would like to thank the National Science Foundation (NSF) for partial support under grant NSF IIS 0505819 and the Technosocial Predictive Analytics Initiative, part of the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory (PNNL). PNNL is operated by Battelle for DOE under contract DE ACO5 76RLO 1830. The contents of this publication are the responsibility of the authors and do not necessarily represent the official views of the NSF.

References

1. Bailey, N.: The Mathematical Theory of Epidemics. Griffin, London (1957)
2. CDC Website: Influenza surveillance reports. Website (accessed 25 July 2009). http://www.cdc.gov/flu/weekly/fluactivity.htm
3. Eysenbach, G.: Infodemiology: tracking flu related searches on the web for syndromic surveillance. AMIA Annual Symposium proceedings, 244 248 (2006)
4. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L.: Detecting influenza epidemics using search engine query data. Nature 457(7232), 1012 1014 (2009) DOI 10.1038/nature07634
5. Heymann, D., Rodier, G.: Global surveillance, national surveillance, and sars. Emerging Infectious Diseases 10(2), 173 175 (2004)
6. Hulth, A., Rydevik, G., Linde, A., Montgomery, J.: Web queries as a source for syndromic surveillance. PLoS ONE 4(2), e4378 (2009)
7. Johnson, H.A.,Wagner, M.M., Hogan,W.R., Chapman,W., Olszewski, R.T., Dowling, J., Barnas, G.: Analysis of web access logs for surveillance of influenza. Studies in Health Technology and Informatics 107(Pt 2), 1202 1206 (2004)
8. Miller, P.: pyMPI An introduction to parallel python using MPI. Livermore National Laboratories, Livermore (2002). URL http://www.llnl.gov/computing/develop/python/pyMPI.pdf
9. Polgreen, P.M., Chen, Y., Pennock, D.M., Nelson, F.D.: Using internet searches for influenza surveillance. Clinical Infectious Diseases 47(11), 1443 1448 (2008) DOI 10.1086/593098
10. Rossum, G.V., Drake, F.: Python language reference. Network Theory Ltd (2003). URL http://www.altaway.com/resources/python/reference.pdf
11. Yih, W., Teates, K., Abrams, A., Kleinman, K., Kulldorff, M., Pinner, R., Harmon, R., Wang, S., Platt, R., Montgomery, J.: Telephone triage service data for detection of influenza like illness. PLoS ONE 4(4), e5260 (2009)

Chapter 62
CodeSlinger: A Case Study in Domain-Driven Interactive Tool Design for Biomedical Coding Scheme Exploration and Use

Natalie L. Flowers

Abstract CodeSlinger is a desktop application that was developed to aid medical professionals in the intertranslation, exploration, and use of biomedical coding schemes. The application was designed to provide a highly intuitive, easy-to-use interface that simplifies a complex business problem: a set of time-consuming, laborious tasks that were regularly performed by a group of medical professionals involving manually searching coding books, searching the Internet, and checking documentation references. A workplace observation session with a target user revealed the details of the current process and a clear understanding of the business goals of the target user group. These goals drove the design of the application's interface, which centers on searches for medical conditions and displays the codes found in the application's database that represent those conditions. The interface also allows the exploration of complex conceptual relationships across multiple coding schemes.

Keywords Biomedical coding · Desktop application · Tool design · Top-down design approach

62.1 Introduction

We were approached by a small group of physician epidemiologists within our company with a unique business problem. They needed an easy-to-use tool to aid them in finding medical codes relating to specific medical conditions. They were using a variety of time-consuming methods to achieve this goal, but these methods

N.L. Flowers
GlaxoSmithKline, Research Triangle Park, NC 27709, USA
e mail: natalie.l.flowers@gsk.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 62,

did not provide a high degree of confidence that they found all the codes they needed, nor were the techniques very time efficient. The tool they were seeking would have to make use of multiple biomedical coding schemes (ontologies) and have the ability to translate from one to another. Under their current practices, the epidemiologist relied upon their medical knowledge to make these sorts of associations, but electronic resources are available that can be harnessed to aid them in this endeavor are available. Kleiner, Painter, and Merrill have demonstrated how to take advantage of a medical terminology metathesaurus to map codes between coding schemes [1].

62.2 Background

Epidemiologists are tasked with performing studies that require finding links between drugs and medical conditions. These relationships might involve negative associations like adverse events or positive associations like finding additional markets for currently approved drugs. Different demographics could also be key factors in determining positive and negative associations. To find these links, the scientists must extract from medical records databases all relevant data about the conditions and drugs of interest.

In most medical records databases, medical concepts such as drugs, devices, symptoms, conditions, and so on are represented by a variety of coding schemes. Each of these coding schemes was created to serve a specific purpose such as billing to health insurance, adverse event reporting, or tracking patient records. If the codes are known for a specific medical concept, then all relevant data relating to that concept can be extracted from the databases. The extracted data are then used for study analysis. To achieve maximum code coverage of the concept, the epidemiologists may search multiple coding schemes to reveal relationships, even if the study only needs to focus on one or two coding schemes.

The technology available to our epidemiologists had not kept up with their workplace demands. Their process for gathering all of the potentially relevant codes was time-consuming. It involved a combination of manual searches through code books, Web-based searches, and a number of other steps, none of which could be relied upon to produce a complete code set. This was the business problem that was brought to the attention of our group.

Our application, CodeSlinger, sought to address the problems of the current code collection methods with a highly interactive, sophisticated interface aimed at medical professionals who may not be well versed in computer usage. CodeSlinger began with the primary purpose of intertranslation of biomedical coding schemes, since that is what the previous code gathering procedure lacked. The goal was to collect a complete code set that would provide the best coverage of the condition or drug being researched.

62.3 Method

The decision was made to take a "top-down" approach whereby the interface was designed entirely around the needs of the user and the underlying architecture was built to support the interface. Cooper discusses how important this top-down approach is, looking at the big picture, instead of focusing on "individual widgets or specific interactions" the user may need in a piecemeal fashion [2]. An interface can get very cumbersome and overcomplicated when the design is applied after the fact with little regard to the needs of the user. Wiklund and Wilcox claim that a failure to understand the needs of the user can lead to "use error, device misuse, and adverse events" [3]. It was necessary for the application to deliver a meaningful, user-friendly interface supported by a solid architectural structure, which in turn gave the designer free reign to design around the users' needs and not be restricted by what was or was not available from the database at any given time.

Aside from the interface decisions, we decided to harness the Unified Medical Language System (UMLS) [4] which provides conceptual relationships between codes in over 150 coding schemes that are used around the world. From within the UMLS, we then needed to determine the coding schemes that we would provide for searching. We decided to incorporate the terminologies used most frequently first and add on less-used terminologies at a later date. Based on the client's needs we loaded the International Statistical Classification of Diseases and Related Health Problems (ICD-9 [5], for ninth generation, and ICD-10 [6] for tenth generation), Medical Dictionary for Regulatory Activities (MedDRA) [7], Current Procedural Terminology (CPT) [8], and Read codes [9]. Each of these coding schemes has a unique way of building its classification hierarchy, and so building an application with an interface that can support the disparate nature of these schemes was a challenge.

We sat down with our primary user and observed her performing searches step-by-step so that we could fully understand what dilemmas and challenges the epidemiologists faced. Preece, Rogers, and Sharp put an emphasis on the involvement of users early in the design process, allowing the designer to observe the tasks performed by users in their usual setting [10]. Johnson supports this approach when describing the process of a "task analysis" early in the design process through discussion and observations with the intended users [11]. In this case, one shadow session with one user was sufficient. Logic suggests that by compromising among many target users the product will meet a wider audience, but Cooper explains how designing for a single user actually will create a more successful product with happier users [12]. Because our target group of epidemiologists was very small, one session provided invaluable information for the interface designer. Had there been a larger user base, multiple observation sessions might have been more appropriate.

On the day of our session the user's goal was to collect all codes related to the condition "congestive heart failure". Some coding schemes distribute books that list all of their codes and respective health care terms. During our observation, the code collection process began by searching these books for all codes that might relate to

"congestive heart failure". The best code candidates who are found are jotted down on paper. Because each coding scheme is independent, any references found will provide only leads within an individual volume. For our epidemiologists, it would be desirable to have references that point to equivalent codes between volumes. Following the search by hand, electronic tools were employed. These included Google™, an internal tool for searching publications, and a large electronic health-records database. All of these methods are error-prone. The user must employ all of these tools because no single one provides good coverage or confidence in the code set she has collected.

Based on this observation, we were able to come up with these basic requirements:

- Search for term or code in one or more coding scheme
- Save codes of interest at any point of search process
- Browse a hierarchy of codes to reveal relationships
- Map codes between multiple coding schemes

Now, armed with a better grasp of how the user performed daily tasks, we designed the interface. This interface went through several iterations before a design was created that met all of our usability requirements. These included minimal need for training, fast searching, and a logical intuitive layout. Early iterations tried to pack features that sacrificed ease of use. The risk of "feature creep" is present with any application, especially when developers want to showcase their skills. Taking an interface-first approach and reducing application functionality can be painful for developers but truly rewarding for the users who benefit from an interface with a clear purpose that requires little explanation. The last iteration stripped out many features in favor of streamlining the application and generalizing the intended purpose.

62.4 Results

CodeSlinger seeks to automate the specific tasks we observed during the shadowing of the primary user. While previously the user had to flip through books of codes, now these codes are loaded into a database for easy search and retrieval. Before, the user had to look up a term in the index to see what page it was on. Now, she can search for the term and see a list of closest possible matches. For example, suppose the user wants to look for "idiopathic thrombocytopenia" in one of the coding schemes, ICD-9. She would first check the boxes for the sources of interest (in this case, ICD-9). Then, she would select whether she wants a verbatim search or a code search. A verbatim search is made up of one or more keywords, such as "heart fail" or "trisomy 21". An alternate way of searching is by code, which allows for one or more codes to be included in the search (e.g., "223.1 223.2 223.3"). Depending on how strict the user wants the search results to be, she can select "Contains" and/or "Fuzzy match". "Contains," the default

selection, will return only exact lexical matches. "Fuzzy match" uses a more complex search algorithm involving stemming and bigram matching [13]. This method can be useful for variance in spellings between British and American English, as well as making allowance for different word endings and misspellings. Selecting both options will cast the widest net of returned results, which in this case is what the user prefers to do. The decision to enable the selection of both options for a "wide-net" search stemmed from our medically knowledgeable users expressing a preference to see everything and weed through the results versus being returned too little and missing key codes.

As we observed in our shadow session, after flipping through the books the user would have written down the codes she thought were good matches for her search. In CodeSlinger, with the search results displayed on the screen, the user merely has to check the boxes for those codes that best meet the original criteria of the search. In this case, the user decides "287.31 Immune Thrombocytopenic Purpura" is a good match, along with a couple of its alternate terms (displayed in the window next to the results). The user decides that it will be useful to see the codes that may be related to 287.31, as it will have been displayed in a hierarchical structure within the coding book. This can be achieved within CodeSlinger by just double-clicking the code. A window opens on the right side of the screen that displays the code in its respective hierarchy in the coding scheme (see Fig. 62.1). The user then sees that some of the sibling codes of 287.31 are also good candidates, and so she checks the codes 287.30, 287.32, and 287.39. Some coding schemes allow a code to belong to multiple parent codes. These parent codes appear below the browser and can be double-clicked to view the code's alternate location. Once a code set is deemed complete, the user can save this set of codes to Excel.

Previously, there was no way of mapping codes between code sources (such as ICD-9, ICD-10, or CPT). The user had to rely on his or her medical knowledge to find related codes. Now, any of the result codes can be clicked to view possibly related concepts that span multiple sources. For example, if the user is looking for "histiocytoma" in ICD-10 but feels that there could be other matches than those displayed for ICD-10, she can broaden the search into other code sources to reveal other possibilities. When multiple sources are selected, the results are broken down into tabs, each tab representing a different source. In Fig. 62.2, the results for the code source General Practice Research Database [14], or GPRD, are displayed. The user is curious if there are any related codes in other sources for the term "reticulohistiocytoma" and clicks on it. The window to the right of the results displays all concepts that may be related. In this case, there is a code in the ICD-10 code source, "M14.3 Lipoid dermatoarthritis." Because "Lipoid dermatoarthritis" has no lexical relationship to "histiocytoma", initial searches did not reveal this term, but through conceptual relationships between sources in the database derived from the UMLS, alternative medical terminology can be revealed. Thus, mapping to other sources can unveil relationships or term differences that may or may not immediately occur to the user.

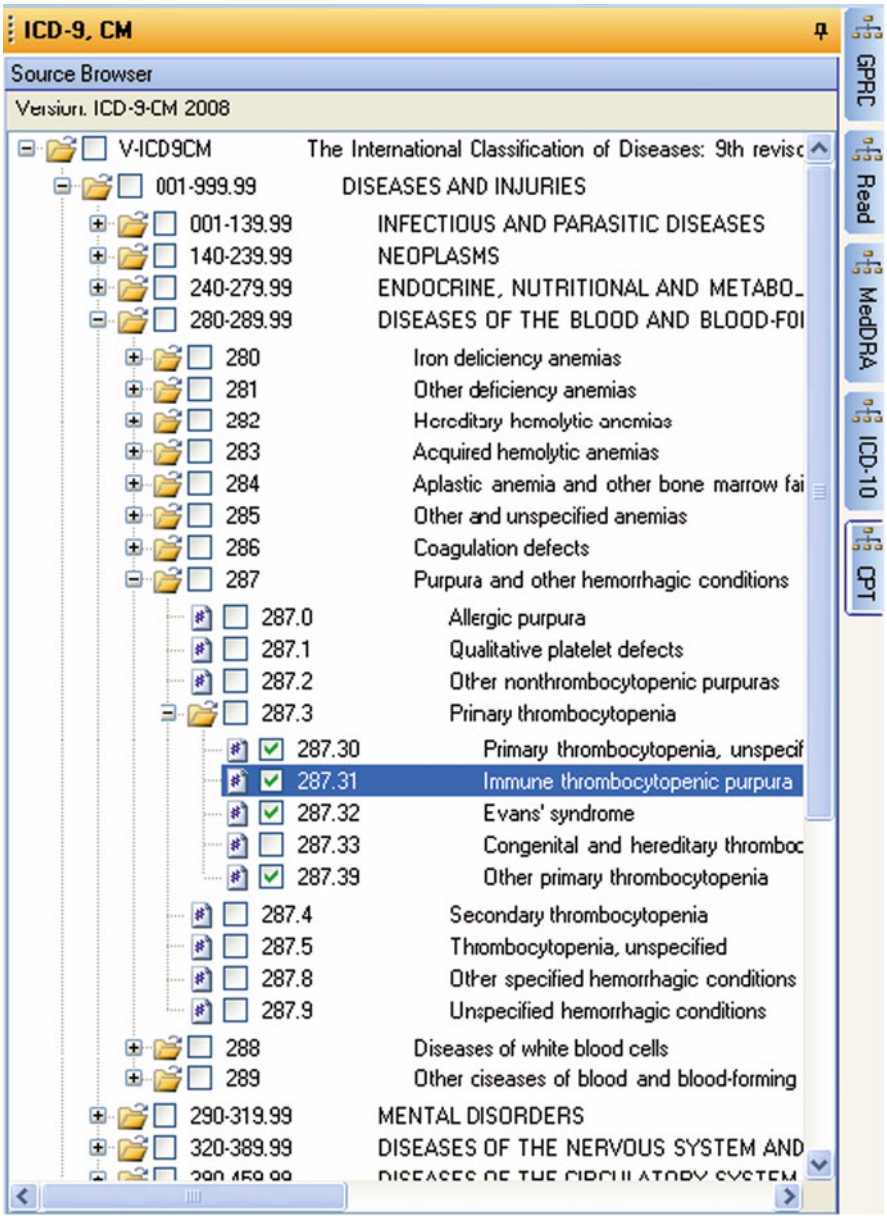


Fig. 62.1 "Immune thrombocytopenic purpora" is revealed in the coding scheme browser which opens on the right side of the interface

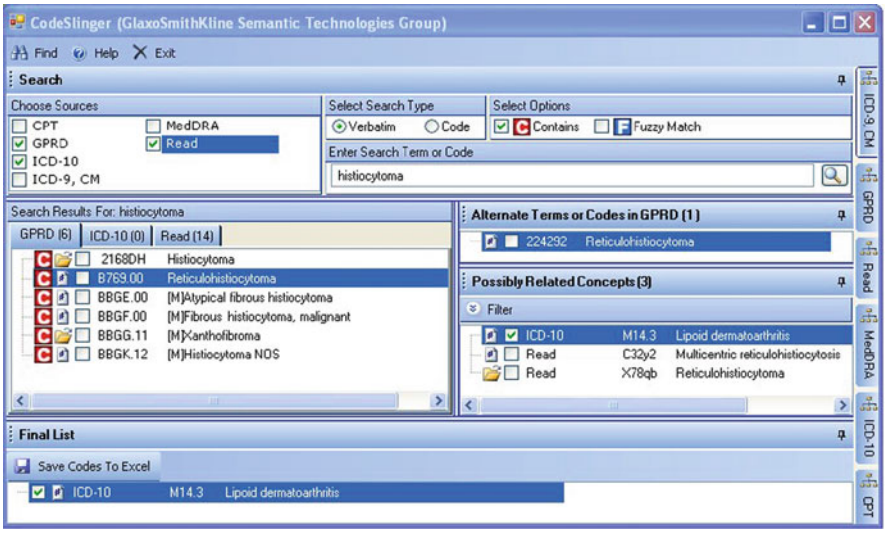


Fig. 62.2 The search result "Reticulohistiocytoma" is selected to display possibly related concepts on the right, revealing a valuable ICD 10 code

62.5 Conclusion

CodeSlinger's success is directly related to our strategy for solving the business problem. Our top-down design approach gives users access to the robust and flexible database architecture through an interface that is intuitive and easy to use. By shadowing a user as she performed her tasks in her workplace, the goals of the client were made clearly visible to the designer. The result is an interface that is well suited to the target audience. When the product was distributed to other users, the functionality and ease of use proved to be beneficial to a wider range of target users. The overwhelming positive response led us to make CodeSlinger publicly available, where it can now be found at http://www.biometrics.com.

References

1. Kleiner K, Merrill G, Painter J (2006) Inter translation of Biomedical Coding Schemes Using UMLS. 2006 AAAI Fall Symposium on Semantic Technologies Proceedings
2. Cooper A, Reimann R, Cronin D (2007) About Face 3: The Essentials of Interaction Design. Wiley, Indianapolis
3. Wiklund M, Wilcox S (2004) Designing Usability into Medical Products. CRC Press, Voca Raton
4. NLM (2009) The Unified Medical Language System. US National Library of Medicine http://www.nlm.nih.gov/research/umls

5. US Department of Health and Human Services, Centers for Medicare & Medicaid Services (2008) International Classification of Diseases, Ninth Revision, Clinical Modification. Baltimore, MD
6. World Health Organization (1998) International Classification of Diseases and Related Health Problems (ICD 10) 10th Revision. Geneva, Switzerland
7. MedDRA Maintenance and Support Services Organization (2007) Introductory Guide, MedDRA Version 10.1. Reston, VA
8. American Medical Association (2007) Current Procedural Terminology (CPT), 4th Edition, Chicago, IL
9. National Health Service National Coding and Classification Centre (1999) Clinical Terms Version 3 (Read Codes)
10. Preece J, Rogers Y, Sharp H (2002) Interaction Design. John Wiley and Sons, New York
11. Johnson J (2008) GUI Bloopers 2.0 Common User Interface Design Do's and Don'ts. Morgan Kaufmann, Boston
12. Cooper A (2004) The Inmates are Running the Asylum. SAMS, Indianapolis
13. El Nasan A, Veeramachaneni S, Nagy G (2001) Word Discrimination Based on Bigram Co occurrences. pp. 149 153, ICDAR'01
14. National Health Service Information Authority (UK) (2009) General Practice Research Database. http://www.gprd.com

Chapter 63
DigitalLung: Application of High-Performance Computing to Biological System Simulation

Greg W. Burgreen, Robert Hester, Bela Soni, David Thompson, D. Keith Walters, and Keisha Walters

Abstract The DigitalLung project represents an attempt to develop a multi-scale capability for simulating human respiration with application to predicting the effects of inhaled particulate matter. To accomplish this objective, DigitalLung integrates macroscale models of integrative human physiology, meso-to-microscale computational fluid dynamics simulations of a breathing human lung, meso-to-nanoscale particle transport and deposition models, and micro-to-nanoscale physical and chemical characterizations of particulate and their mass transfer through the mucosal layer to the epithelium. This chapter describes preliminary results and areas of ongoing research.

Keywords Computational systems biology · High performance computing and applications in biology · Biomedical computing

63.1 Introduction

Exposure to particles such as asbestos, coal dust, and diesel exhaust has been shown to induce both chronic and acute deleterious effects on respiratory and cardiovascular function. However, the mechanisms by which inhalation of particulate matter (PM) leads to pathological events are not clear. Similarly, inhalation methods for drug delivery can be significantly improved by developing capabilities to

D. Thompson (✉)
Department of Aerospace Engineering, Mississippi State University, Mississippi State, MS, USA
e mail: dst@ae.msstate.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 63,

understand and predict their efficacy. Understanding the mechanisms responsible for the effects induced by aerosol and particle inhalation is nontrivial and requires a quantitative and integrated multiscale analysis of physiology, fluid dynamics, and mass transport. A first logical step to this end is the development of human lung models that accurately relate patient-specific physiological details and ventilation patterns to inhaled particle deposition.

To this end, we have undertaken to employ high-performance computing to simulate the respiratory aspects of particle inhalation based on realistic physiological dynamics. The resulting effort, which we have christened DigitalLung, involves the development of a detailed multiscale model of respiratory physiology in order to understand particle deposition and mass transfer following PM inhalation. The DigitalLung project is truly *interdisciplinary* in that our team has significant experience in the areas of whole-body integrative physiology, high-performance computational fluid dynamics (CFD) of lung fluid dynamics, and physiochemical energetics related to particle deposition/adhesion. Our approach is *multiscale* in that it tightly integrates (a) a macroscale model of integrative human physiology (DigitalHuman) to determine ventilation parameters associated with gender/age/activity/altitude variables; (b) meso-to-microscale CFD simulation of the complete unsteady "breathing" airway system including oral cavity, trachea, upper bronchi, and multiple discrete airway paths of lower bronchioles to terminal alveoli; (c) meso-to-nanoscale particle transport and deposition models to determine particle deposition fractions and dosimetry throughout the entire lung; and (d) micro-to-nanoscale chemical and physical characterization of particles and their mass transfer through the mucosal layer to the epithelium. The resulting modeling methodology will be systematically validated and made available to the scientific community for research, teaching, and clinical applications in the near future.

63.2 Background

Inhaled particulate matter is often incidental and heavily dependent on an individual's work and residential environments. Recent evidence suggests that short-term exposure to PM can have acute cardiovascular effects [1]. Long-term elevated PM exposure has been linked to increased cardiovascular events and cardiovascular mortality, and is a causative factor for atherosclerosis and reduced life expectancy.

Compared to subcutaneous injection, noninvasive inhalation therapies offer direct circulatory uptake and quicker effect, lower systemic bioavailabilities, increased patient compliance, and lower cost due to lower required dosages [2]. A major challenge for pulmonary delivery is deep lung deposition. Aerosol formulations have not been efficient at deep lung delivery where the alveolar epithelium and thinned mucus layer present less of a barrier to circulatory uptake. Ventilation patterns significantly affect aerosol deposition. Aerosolized medications have been shown to preferentially deposit in various depths of the lungs depending on particle size [2].

63.2.1 Modeling of Integrative Human Physiology

Mathematical simulation of physiological processes has become an important tool for understanding normal and pathophysiological processes within the body. There are an extensive number of publications in the literature describing mathematical simulations of individual organ systems, but there are minimal studies demonstrating integration across systems. We have developed a state-of-the-art model of the human body, DigitalHuman, that is unique in its detail of physiological processes and predictive ability. The DigitalHuman model is comprised of ~5,000 variables that are predictive of human physiology, including cardiovascular, respiratory, neural, renal, and metabolic systems. The roots of DigitalHuman began in the late 1960s when Drs. Arthur Guyton and Thomas Coleman [3,4] of the University of Mississippi Medical Center used computer simulations to develop and test hypotheses concerning physiological systems. In a series of computer simulation studies supported by experimental observations, they were able to identify correctly a dominant role for the kidney in long-term arterial pressure control many years before it became common wisdom. Versions of the current mathematical model have been used for studies funded by NIH/NHLBI [5], NASA [6], and the EPA [7].

63.2.2 Simulation of Lung Respiration

There have been numerous simulations of lung respiration reported in the literature. Most may be divided into two categories: reduced order models that rely on relatively simple 1D empirical or semiempirical modeling methods; or CFD-based approaches that seek to solve numerically the equations governing fluid motion. Most of the CFD-based studies have focused on subsets of the bronchial tree or components of the upper airways due to the prohibitive computational expense associated with simulating the complete problem. To date, no CFD-based study has simultaneously modeled the entire lung flow field with 3D Navier Stokes methods, even for steady inspiratory flow. For brevity, we mention only two representative, state-of-the-art simulations of lung respiration. Ma and Lutchen [8] combined fully resolved CFD simulation of the upper airways up to generation six with a 1D transmission line model of the impedance of the small-scale airways. This coupled 3D/1D approach represents the current state of the art for "full lung" simulations of respiration and particle deposition. Using a more direct approach, Gemci et al. [9] reported the simulation of 17 generations of the human lung based on the anatomical model of Schmidt et al. [10]. The geometry was only partially resolved, containing 1,453 bronchi as opposed to approximately 2^{17} for a fully resolved model. Also, an equal constant pressure condition was employed at all flow outlets, which is nonphysical and does not provide a realistic coupling between the large- and small-scale regions.

63.2.3 Simulation of Particle Deposition

Particle deposition simulations have been reported in the literature for numerous idealized symmetric and asymmetric representations of segments of the bronchial tree. Recent efforts have focused on the importance of including the effects of realistic geometries generated by medical imaging techniques, such as multigenerational bronchial tube models [11,12] and orotracheal geometries [13,14]. The simulation of realistic bronchial branches including the presence of the laryngeal jet significantly increases deposition in the trachea and decreases deposition in the first few bronchi, which underscores the need to model upper airways as realistically as possible. A greater challenge is modeling the full branching structure of the entire human lung. One approach to answer this challenge is the use of complex 1D models [15,16], but this approach inherently relies on various approximations such as simplified algebraic approximations of airflow, rigid airways, a priori specification of model parameters such as deposition efficiency within specific segments, and empirical calibration constants. A more preferable option would be highly resolved CFD simulation of the entire lung. Very few CFD-based studies have sought to simulate flow and particle deposition in either the entire conducting zone or the entire lung. For the sake of brevity, we mention only two efforts in this area. Nowak et al. [17] performed a series of simulations in progressively smaller 3.5-generation symmetric branching segments for inspiratory flow only, and Zhang et al. [18] documented a similar methodology for the prediction of nanoparticle deposition throughout 16 generations. Notably, both authors state that limitations of their modeling approach included the lack of geometric realism for the upper airways and the lack of physiological outlet boundary conditions. For both studies, the solution method did not allow a full coupling of the flow in the lung at all scales. Zhang et al. [18] point out that this limits the ability to extend the methodology to unsteady (breathing cycle) simulations.

63.2.4 Models of Particle Transport

A significant majority of CFD-based particle deposition studies reported in the literature are based on Lagrangian particle tracking methods. In contrast, a relatively small number have used an Eulerian approach to solve both the fluid flow and particle transport equations (e.g., [19,20]). It is generally recognized that the Lagrangian approach is well suited for investigations of localized particle deposition in small-scale geometries, but requires prohibitive computational expense when investigating large-scale geometries such as the whole lung. Almost all of the Eulerian studies use the one-fluid model, in which the bulk convective velocity of the primary phase (air) and secondary phase (particles) is assumed to be identical. This approach is only valid for ultrafine particles and can lead to significant error for particles as small

as 70 nm [21]. Apart from the study by Kunz et al. [22], very few studies have used a two-fluid model, that solves separate momentum equations for each phase.

63.3 Preliminary Results

Specific advances must be made in order to provide accurate and computationally efficient tools for next-generation respiration and deposition modeling. These include (1) physically realistic geometry models, including patient-specific models of the upper airways and statistically accurate morphological descriptions of the lower airways; (2) physiologically realistic ventilation patterns under a wide range of environmental conditions and stressors; (3) fully coupled 3D Navier Stokes modeling of the airflow through the entire lung airway from the trachea to the alveoli; (4) accurate modeling of volumetric changes during the breathing cycle, particularly in the pulmonary and alveolar regions; and (5) the use of two-fluid Eulerian models for particle transport modeling, including particle impaction, and transport models to predict particle diffusion rates through the mucosal layer to the epithelial wall. We have made significant progress regarding several of the components necessary for the realization of DigitalLung.

63.3.1 CFD Simulations of Idealized Bronchial Tubes

Soni et al. [23] have simulated flow through asymmetric three-generation, planar, and nonplanar idealized bronchial tube models based on the fundamental unit of Hammersley and Olson [24]. Figure 63.1 illustrates the effects of simultaneous asymmetric and nonplanar branching on the primary and secondary flows. In the case of the planar geometry, the symmetry of the geometry produces symmetric secondary flows, even though the branching is asymmetric. In the case of the

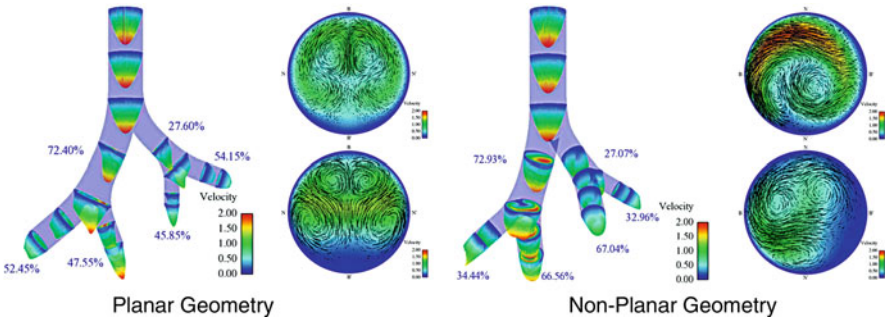


Fig. 63.1 Primary and secondary flows are significantly affected by nonplanar, asymmetric branching (from [23])

nonplanar geometry, the plane of symmetry is lost, and the resulting secondary velocities are no longer symmetric. This work demonstrates the importance of including realistic details such as asymmetry and out-of-plane branching in models of the bronchial tubes.

63.3.2 CFD Simulations of Particle Deposition

Soni et al. [25] have developed a technique based on the finite-time Lyapunov exponent (FTLE) [26] to investigate the effects of asymmetry and nonplanarity on particle deposition. A Lagrangian approach is employed to compute the final positions of particles introduced at the inlet of the model. The destination map shows the region in which a particle at the inlet is deposited, while the FTLE map shows the local rate of dispersal of the particles. Figure 63.2 highlights the importance of including both asymmetric planar and nonplanar branching in the geometrical model employed for particle deposition computations. The loss of symmetry in the nonplanar model is evidenced in the particle deposition in the third-generation tubes.

63.3.3 Flow Path Ensemble Model for Large-Scale Simulation of Lower Airways

Walters and Luke [27] have developed flow path ensemble (FPE) models and demonstrated their effectiveness in providing an accurate representation of the flow in geometrically complete, large-scale lung structures. An idealized model of generations 4 12 based on the symmetric Weibel [28] morphology with out-of-plane branching and the intragenerational geometric parameters of Hammersley and Olson [24] was used for these computational studies. Simulations were

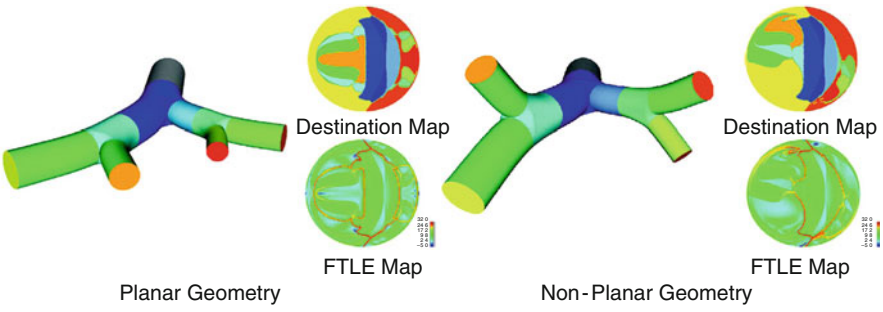


Fig. 63.2 The deposition of 10 μm water droplets is significantly affected by nonplanar, asym metric branching. The destination map shows the zones in which particles located at the inlet are deposited. The FTLE map illustrates the rate of dispersal of particles at the inlet (from [25] ©2008 IEEE)

performed for steady inspiration with an applied uniform inlet velocity corresponding to three physiologically realistic ventilation rates. Three companion FPE approximations to the eight-generation geometry were obtained by random selection of 4, 8, and 16 distinct flow paths, respectively. The unresolved outlet conditions were resolved using two separate methods: a stochastic coupling approach developed by the authors and a constant pressure condition identical to that used by Gemci et al. [9] in their 17-generation model.

Pressure contours on the lung wall are shown in Fig. 63.3 for the full geometry and the 4- and 16-path FPE models using stochastically coupled boundary conditions. The results of the test cases and the comparison between FPE models and the fully resolved model are shown in Table 63.1. Using FPE models and stochastically coupled outlet conditions, relative errors in predicted impedance and average outlet mass flow rates are remarkably small compared to the fully resolved simulation. In contrast, use of uniform outlet pressures yields errors ranging from 58.6% (4-path) to 37.8% (16-path). Computational cost of the FPE method compared to the fully resolved simulation is reduced by approximately 90%. These results highlight the utility of such a method for developing accurate "full lung" computational models encompassing all generations of the bronchopulmonary tree.

63.3.4 DigitalHuman Whole Body Model

DigitalHuman is driven by a valid XML schema that describes the physiological processes, solution control, and display of results. In this way, the functional details of the model are open source and universally available. DigitalHuman simulations

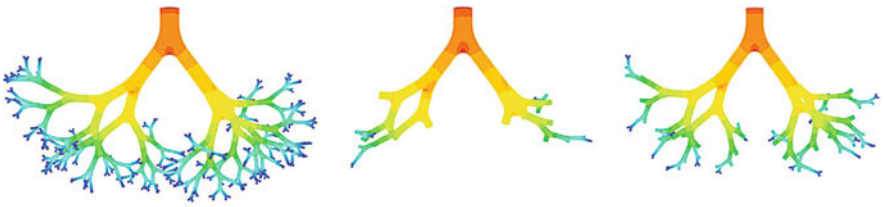


Fig. 63.3 Static pressure contours for the fully resolved eight generation lung airway model (*left*), the 4 path FPE (*middle*), and the 16 path FPE model (*right*) during steady inhalation (higher pressure at inlet, lower pressures at exits)

Table 63.1 Performance of FPE method with stochastic coupling for airflow prediction

Model	% CPU memory	Impedance (Pa) (% error)	Outlet flow rate (mg/s) (% error)
Full geometry	100	5.99 (0%)	9.92×10^{-2} (0%)
4 Path	6	6.07 (1.32%)	10.12×10^{-2} (1.91%)
8 Path	11	6.01 (0.30%)	9.97×10^{-2} (0.40%)
16 Path	19	5.99 (0.08%)	9.93×10^{-2} (0.14%)

can predict the subtle cardiovascular responses related to simple actions such as standing up after a period of lying down or the more complex physiological responses related to exercise. Increases in blood pressure, heart rate, and respiratory rate can all be predicted in an integrated approach. Recently, detail and body size scaling have been added to DigitalHuman to simulate female physiology and respiratory responses under extreme exercise conditions.

63.3.5 Particle–Surface Interaction Characterization for Modeling of Deposition Dynamics

Efforts have focused on investigating the (1) physical and chemical surface properties of lung tissue and (2) interfacial tension data on drug-doped solutions using pendant and static contact angle measurements. Cryotomed lung tissue sections were examined to estimate surface roughness, bronchiole geometry, and mucus layer and epithelium thicknesses. The bronchiole geometry was found to be non-uniform with surface roughness decreasing from proximate to distal regions of the lung. Static contact angles (SCA) have been measured on bovine lung tissue and control samples for the following probe liquids: pH 7 phosphate buffer solution, HPLC water, and two artificial saliva solutions doped with potassium iodide, salbutamol, or theophylline. A contact angle goniometer with an environmental chamber to control temperature and relative humidity was used to capture digital images and video of microliter-volume sessile and pendant drop contact angles. KRÜSS drop shape analysis software was used to determine the contact angles using several drop shape fitting models, which revealed rapid decreases in SCA during the initial 3 5 min postdeposition.

63.4 Current Efforts

Building on our preliminary results, we are currently pursuing development efforts in three main technical areas that are mutually complementary within the context of our DigitalLung project.

63.4.1 Development of a Multiscale, Dynamic Moving-Wall CFD Model of Respiration Fluid Dynamics for the Entire Lung Airway from the Extrathoracic Regions to the Terminal Alveoli

Our objective is to develop the capability to predict unsteady respiration fluid dynamics during cyclic breathing using a fully coupled, dynamic CFD model of

the human airways. Due to the associated computational cost, it is currently not feasible to perform high-fidelity CFD simulations of flow and particle deposition throughout the $\sim 2^{23}$ individual flow segments of the entire lung. However, the FPE method recently developed by our group has shown the capability to model flow through the small-scale multibranched regions effectively using 3-D Navier Stokes simulation of a finite number of randomly selected flow paths [27]. We will couple this method for the lower airways, with accurately resolved geometries for the upper airways. The upper airway models will be generated with varying degrees of realism ranging from ideal geometries to physiologically accurate models obtained from patient-specific medical imaging techniques. The result will be a fully coupled CFD simulation model of the entire flow network, down to and including the alveolar sacs. Time-dependent volume changes in the pulmonary and alveolar regions will drive inspiration and exhalation in a realistic manner. This approach offers substantial improvement relative to existing approaches because the effects of realistic geometries, time-varying flow volumes, and breathing patterns can be included in computations for the entire lung.

63.4.2 Development of One-Way Coupling Methods Between DigitalHuman-Based Human Ventilation Models and Unsteady CFD Analysis with the Goal of Simulating Realistic Fluid Dynamics in a Lung Model

This effort focuses on the development of a human lung model that accurately relates patient-specific physiological details and ventilation patterns. We will use our baseline model of integrative human physiology (DigitalHuman) to establish physiologically based ventilation parameters related to resting/working conditions, gender/size differences, and geographical altitude variations that will serve as input data to the high-fidelity CFD models. There are several areas that need improvement as this content is essential to the accurate understanding of respiratory flows and the deposition of particulate matter. One area currently under development is an accurate time-dependent description of tidal volumes in the lung during ventilation, including inhalation and exhalation modeled as a submaximal flow-volume loop. Additionally, increases in ventilation, as with exercise, produce dramatic increases in tidal volume and rate, air velocity, and particle deposition rate, and a decrease in time of inspiration. We will add detail to simulate accurately breath-by-breath inhalation and exhalation responses, including airway pressures and velocities, under basal and exercise conditions. The respiratory variables will be passed to the CFD simulation, thus providing a physiologically accurate basis for pulmonary physiology. Our goal is to innovatively couple DigitalHuman with CFD to yield a first-of-its-kind multiscale simulation capability.

63.4.3 Development of Particle Deposition and Transport Models Including Experimentally Determined Physiochemical Effects

We seek to incorporate accurate and efficient models of particle deposition and transport into the full lung CFD model described above. The overall strategy can be conceptually divided into three distinct but related parts: multiphase transport modeling of particulates and aerosols in the airway; impaction modeling of deposition efficiencies for particles striking the mucus layer coating airway surfaces; and diffusion modeling to determine particle transport rates through the mucous layer to the epithelial wall. In order to provide necessary inputs to the models, this effort will necessarily include experiments to determine parameters such as particle surface adhesion and diffusivity coefficients. The particle transport models developed will be directly integrated into the CFD models.

Currently, our group uses a Lagrangian particle tracking methodology to determine particle deposition in CFD simulations of relatively small airway segments. This approach corresponds to the standard practice found in the literature, but is impractical for use in whole lung simulations due to the number of particles that must be tracked to provide meaningful statistical data. We will therefore use the alternate Eulerian approach, in which PM is modeled as a dispersed second phase and air the primary phase. Well-developed Eulerian methods currently exist [22]; however, one open question concerns the wall boundary conditions for such models. We will investigate and extend existing models of particle impact and adhesion (deposition), and incorporate them into the CFD simulations to provide physically realistic boundary conditions. General transport models are being developed for a wide range of particle and fluid systems. In addition to the inertial, gravitational, and viscous drag forces, the model will also incorporate particle nonsphericity, size, and aggregation; external magnetic fields; particle fluid interactions; and particle particle/surface interactions (e.g., magnetic, electrostatic, and van der Waals forces).

Pulmonary epithelium is a tightly packed, single-cell layer that serves as a barrier for transport into the lung tissue. The alveolar epithelium is important in cellular/circulatory uptake partially due to a large number of alveoli ($\sim$0.5 billion), correspondingly high surface area (>100 m^2 in adults), thin overlying mucus layer, and nonciliated cells. To close the modeling loop, a mass transport model is being developed to account for the diffusion of particulate material through the mucosal lining of the lung, thus providing full model resolution of particle transfer from the ambient environment to the pulmonary epithelium. The deposition rate obtained from the CFD simulation will provide input to the model in the form of a concentration boundary condition at the air mucus interface. The model output will be local transfer rates to the respiratory and alveolar epithelium.

Acknowledgments This work has been partially funded by NSF (ITR/NGS 0326386, EPS 0556308) and NIH (HL 51971).

References

1. Higenbottam T, Siddons T, Demoncheaux E. The direct and indirect action of inhaled agents on the lung and its circulation: Lessons for clinical science. Environ Health Perspect 2001, 109 (4):559 562.
2. Carveth HJ, Kanner RE. Optimizing deposition of aerosolized drug in the lung: A Review. MedGenMed 1999, 1(3).
3. Guyton AC, Coleman TG. Quantitative analysis of the pathophysiology of hypertension. Circ Res 1969, 24:1 19.
4. Guyton AC, Coleman TG, Granger HJ. Circulation: Overall regulation. Annu Rev Physiol 1972, 34:13 46.
5. Lohmeier TE, Hildebrandt DA, Warren S, May PJ, Cunningham JT. Recent insights into the interactions between the baroreflex and the kidneys in hypertension. Am J Physiol Regul Integr Comp Physiol 2005, 288:R828 R836.
6. Summers RL, Martin DS, Meck JV, Coleman TG. Computer systems analysis of spaceflight induced changes in left ventricular mass. Comput Biol Med 2007, 37:358 363.
7. Benignus VA, Coleman T, Eklund C, Kenyon E. A general physiological and toxicokinetic (GPAT) model for simulating complex toluene exposure scenarios in humans. Toxicol Mech Methods 2006, 16:27 36.
8. Ma B, Lutchen KR. An anatomically based hybrid computational model of the human lung and its application to low frequency oscillatory mechanics. Ann Biomed Eng 2006, 34 (11):1691 1704.
9. Gemci T, Ponyavin V, Chen Y, Chen H, Collins R. Computational model of airflow in upper 17 generations of human respiratory tract. J Biomech 2008, 41:2047 2054.
10. Schmidt A, Zidowitz S, Kriete A, Denhard T, Krass S, Pietgen H O. A digital reference model of the human bronchial tree. Comput Med Imaging Graph 2004, 28:203 211.
11. De Backer JW, Vos WG, Gorle CD, Germonpre P, Partoens B, Wuyts FL, Parizel PM, De Backer W. Flow analyses in the lower airways: Patient specific model and boundary conditions. Med Eng Phys 2008, 30:872 879.
12. Ma B, Lutchen KR. CFD simulation of aerosol deposition in an anatomically based human large medium airway model. Ann Biomed Eng 2009, 37:271 285.
13. Lin C L, Tawhai MH, McLennan G, Hoffman EA. Characteristics of the turbulent laryngeal jet and its effect on airflow in the human intra thoracic airways. Respir Physiol Neurobiol 2007, 157:295 309.
14. Xi J, Longest PW, Martonen TB. Effects of the laryngeal jet on nano and microparticle transport and deposition in an approximate model of the upper tracheobronchial airways. J Appl Physiol 2008, 104:1761 1777.
15. Martonen TB, Schroeter JD, Hwang D, Fleming JS, Conway JH. Human lung morphology models for particle deposition studies. Inhal Toxicol 2000, 12(Suppl 4):109 121.
16. Asgharian B, Price OT, Hofmann W. Prediction of particle deposition in the human lung using realistic models of lung ventilation. Aerosol Sci 2006, 37:1209 1221.
17. Nowak N, Kakade PP, Annapragada AV. Computational fluid dynamics simulation of airflow and aerosol deposition in human lungs. Ann Biomed Eng 2003, 31:374 390.
18. Zhang Z, Kleinstreuer C, Kim C. Airflow and nanoparticle deposition in a 16 generation tracheobronchial airway model. Ann Biomed Eng 2008, 36:2095 2110.
19. Shi H, Kleinstreuer C, Zhang Z, Kim C. Nanoparticle transport and deposition in bifurcating tubes with different inlet conditions. Phys Fluids 2004, 16:2199 2213.
20. Wang J, Lai A. A new drift flux model for particle transport and deposition in human airways. J Biomech Eng 2006, 128:97 105.
21. Longest P, Xi J. Computational investigation of particle inertia effects on submicron aerosol deposition in the respiratory tract. J Aerosol Sci 2007, 38:111 130.

22. Kunz R, Haworth D, Leemhuis L, Davison A, Zidowitz S, Kriete A. Eulerian multiphase CFD analysis of particle transport and deposition in the human lung. Presented at BIOMEDICINE 2003, April 2 4, 2003, Ljubljana, Slovenia.
23. Soni B, Lindley C, Thompson D. The combined effects of non planarity and asymmetry on primary and secondary flows in the small bronchial tubes. Int J Num Meth Fluids 2009, 59:117 146.
24. Hammersley J, Olson D. Physical models of the smaller pulmonary airways. J Appl Physiol 1992, 72:2402 2414.
25. Soni B, Thompson D, Machiraju R. Visualizing particle/flow structure interactions in the small bronchial tubes. IEEE Trans Vis Comput Graph 2008, 14:1412 1419.
26. Haller G. Distinguished material surfaces and coherent structures in three dimensional fluid flows. Physica D 2001, 149:248 277.
27. Walters DK, Luke W. 3 D Navier Stokes simulation of large scale regions of the broncho pulmonary tree. Proceedings of ASME International Mechanical Engineering Congress and Exposition 2009, Nov 13 19, 2009, Lake Buena Vista, FL.
28. Weibel ER. Morphology of the human lung. Academic, New York, 1963.

Chapter 64
Consideration of Indices to Evaluate Age-Related Muscle Performance by Using Surface Electromyography

Hiroki Takada, Tomoki Shiozawa, Masaru Miyao, Yasuyuki Matsuura, and Masumi Takada

Abstract Recently, there has been an increasing focus on the rapid reduction of muscles that are required for the bending of the hip joint during walking (flexor muscles around the hip joint) with age. The flexor muscles around the hip joint include femoral rectus and abdominal muscles. These muscles have been implicated in falling in the elderly people. In this study, we examined the smoothed surface electromyography (sEMG) of femoral rectus muscles during biofeedback training (BFT) of the dominant leg. To this end, we developed parameters for the measurement of shapes in the smoothed sEMG, and evaluated the changes in these parameters in the muscles with age. Reduction of the muscular regulation capacity due to aging can be detected by performing sEMG during BFT by using a parameter in the muscles.

Keywords Aging · Biofeedback training (BFT) · Double-Wayland algorithm · Surface electromyography (sEMG) · Stability

64.1 Introduction

Currently, several electromyographic methods are used, and needle electromyography (nEMG) and surface electromyography (sEMG) are most often applied. To physiologically evaluate electromyographic wave patterns for the detection of abnormalities, wave patterns obtained by nEMG or sEMG are macroscopically examined and subjectively judged by physicians.

In nEMG, findings are used for the evaluation of whether a disorder is neurogenic or myogenic, and if it is both neurogenic and myogenic, they provide

H. Takada (✉)
Department of Human and Artificial Intelligent Systems, Graduate School of Engineering, University of Fukui, 3 9 1 Bunkyo, Fukui, Japan
e mail: takada@u fukui.ac.jp

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 64,

important information about whether it is acute, subacute, or chronic [1]. However, the probe is a needle electrode that is percutaneously inserted into muscular tissues.

In sEMG, findings are used for various evaluations, such as classification of trembling for the diagnosis of involuntary motion, the diagnosis or differential diagnosis of dystonia and spasm, and identification of involuntary constrictor muscles [2]. sEMG is further used for the determination of electric potential by nerve conduction examination (evoked EMG). In evoked EMG, electrostimulation of peripheral nerves is percutaneously performed [1].

The examination methods, except for the sEMG, are invasive and cause severe pain in patients. Generally, "smoothing" and "integration" refer to two ways of quantifying EMG energy over time; Smoothing refers to averaging continuously the peaks and valleys of a changing electrical signal. On the other hand, integration refers to measuring the area under a curve over a time period. These are used for the examination of the relative degree of muscular contraction, and are also employed as a parameter for the evaluation of muscular training conditions [3]. However, their results were affected by the location of the measuring electrodes, and the shape and size of the probes. That is, EMG findings are macroscopically and subjectively evaluated, as described above, and no algorithm for the quantification of the degree of muscular abnormalities or recovery has been established. In this study, we apply and discuss measurement parameters that have been developed for evaluating the smoothed sEMG data obtained from perineal muscles during biofeedback training (BFT) for the treatment of dysuria [4], and we evaluate the effects of this training [5].

Recently, the rapid reduction of muscles for the bending of the hip joint during walking (flexor muscles around the hip joint) with age has drawn attention. The flexor muscles around the hip joint consist of femoral rectus and abdominal muscles. It has been indicated that these muscles are involved in falling of the elderly people. In this study, we examined the smoothed sEMG of femoral rectus muscles performed during BFT of the dominant leg, using the above measurement parameters, and evaluated their changes with age.

64.2 Materials and Methods

Temporal data are obtained by sEMG, and here, they are expressed as $\{x(t)\}$. Generally, sEMG data are recorded in a computer at 1 kHz. Here, integral calculation is performed every 0.1 s using the following equation:

$$y(t) = \sum_{k=0}^{99} |x(t + 0.001k)|, \tag{64.1}$$

and smoothed sEMG $\{y(t)\}$ is calculated in real time, and outputted. The subject observes the outputted wave patterns and rectangular waves $f(t)$ of a 10-s cycle

superimposed on the same display, and performs intermittent continuous contraction of femoral rectus muscles corresponding to the patterns (BFT).

64.2.1 Experimental Procedure

The subjects were 31 healthy adults aged 20 73 years (mean 44.3 ± 19.9 years); they performed BFT for 2 min. All subjects gave consent in writing after sufficient explanation of this study. The subject sat back on a four-legged stool, and electromyographic electrodes were applied at an interval of several centimeters to the venter of femoral rectus muscles in the dominant left or right leg [6]. The subjects were instructed to kick a fixed belt with the bottom of the lower leg forward (kicking motion). A special electromyographic transformation box (AP-U027, TEAC Co.) was connected to a commercially available portable and versatile amplifier and recorder (Polymate AP1532, TEAC Co.), and electromyographic electrodes (bipolar) with a preamplifier were used.

First, electromyographic wave patterns obtained during the kicking motion at the maximum effort (maximum voluntary contraction, MVC [7]) for several seconds were integrated in real time using a computer, and the smoothed sEMG on the display was shown to the subject. Second, the threshold line at 75% of the mean smoothed sEMG (mV) during the muscular contraction period was shown to the subject, who was requested to perform muscular training aiming at the threshold line for 1 min 20 s. In other words, BFT was performed at 75% of the MVC. During BFT, data were recorded in a notebook computer (AP Monitor, NoruPro) at a sampling rate of 2 kHz. The high- and low-frequency cut-off filters were used at 100 and 16 Hz, respectively, and an alternating current-eliminating filter was also used.

64.2.2 Calculation Procedure

Of the sEMG data recorded over 1 min 20 s, the initial 20-s data were excluded, because the subjects may not have adjusted to the training. sEMG data of the following 6-cycle rectangular waves (target value) $f(t)$ and the smoothed sEMG were analyzed in accordance with our mathematical algorithm of the sensor output signal evaluation system [5]. Taking a mean smoothed sEMG as a threshold H to determine continuous muscular contractions, time sequences above the threshold H were regarded as continuous muscular contractions. Based on the sign of differences such as $y(t) - y(t - 0.1) > 0$ and $(y(t + 0.1) - y(t))(y(t) - y(t - 0.1)) < 0$, maximal series during the continuous muscular contractions were extracted:

(a) The mean value of the smoothed sEMG during the muscular relaxation period (x^a) and the following measurement parameters [8], indicating the shape of the

smoothed sEMG, were determined in every cycle, and the smoothed sEMG obtained from the femoral rectus muscles were evaluated.

(b) Maximum amplitude (x^b): The maximum value was examined and recorded.

(c) Duration of continuous muscular contraction (x^c): The duration between the first and last maximal values exceeding the mean smoothed sEMG in a cycle was measured.

(d) Time constant of the exponential decay curve fit to maximal points during the continuous muscular contraction period (x^d): All maximal values between the first and last maximal values over the mean smoothed sEMG in a cycle were extracted as $\{y_m(t)\}$ and fit to the exponential decay curve $\hat{y}_m(t) = C\exp[-x^d t]$. On a semi-log graph, the time constant (x^d) was estimated by the mean least-square method.

Numerical sequences of the four measurement parameters were determined at a repetition number of 6:

(1) The relationship between the age (z) of the subjects who had undergone sEMG and the value $x^i(z)$ ($i = a, b, c, d$) estimated in the fifth cycle was statistically examined to evaluate correlations between each measurement parameter and age.

(2) Since there were differences in not only the unit but also numerical order between the parameters, they were normalized using the intermediate values x^i for each cycle, and the reproducibility (stability) of measurements was evaluated using the standard deviation $\sigma[x^i/x^i]$. The normalized value is 1 when the measurement is equal to the intermediate value. When the reproducibility (stability) of measurements is high by repeated measurement, there are only small variations around this value, and the standard deviation is close to 0.

64.3 Results

(1) Numerical sequences of each of the measurement parameters at a repetition number of 6 were obtained by sEMG performed during BFT. The relationship between the age (z) of the subjects who had undergone sEMG and the value $x^i(z)$ ($i = a, b, c, d$) was examined in the fifth cycle. Figure 64.1 shows the $x^i(z)$ in all 50 subjects. The linear regression analysis by the least-square method demonstrated that the coefficient ($\hat{b}$) by which age (z) was multiplied was 0.071, −0.268, −0.006, and −0.010 for $i = a, b, c$, and d, respectively, and the parameters, except for (a), decreased with age. Since the linear regression coefficients varied with measurement parameters, correlations between the parameters and age could not be judged only using the coefficients by which age (z) was multiplied. In the t-test for the evaluation of the null hypothesis ($\hat{b} = 0$) for the regression coefficient, the test value was 1.105, 0.238, 1.621, and 3.245 for $i = a, b, c$, and d, respectively, and the only parameter exceeding $t_{48}(0.975)$ was the time constant of the exponential decay curve fit to the maximal points during the continuous muscular contraction period (x^d).

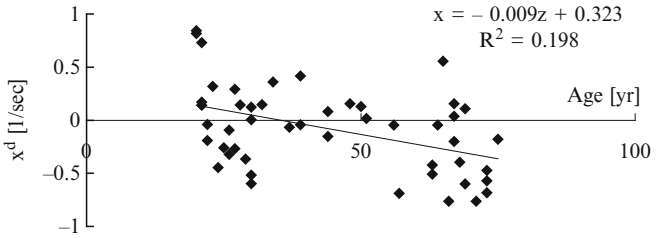


Fig. 64.1 Relationships between a measurement parameter of smoothed sEMG (time constant of the exponential curve fit to maximal points during the continuous muscular contraction period x^d) and age, and its linear regression. R^2 shows the coefficient of determination [6]

Table 64.1 Standard deviations of normalized indices x/x [8]

Age group	N	Mean during relaxation	Maximal amplitude	Duration of continuous muscular contraction	Time constant
≤25	8	0.24	0.13	0.06	1.51
≤45	9	0.21	0.14	0.08	0.55
≤65	6	0.33	0.14	0.08	1.84
65<	8	0.07	0.06	0.06	0.89

N expresses the number of subjects in each age group

(2) The intermediate values of parameters and standard deviations (σ) of normalized measurements were determined in each subject, and medians of σ in the age groups were compared in Table 64.1. The duration of continuous muscular contraction (x^c) alone showed $\sigma < 0.1$ for any age group.

64.4 Discussion

We examined correlations between parameters of smoothed sEMG and age. Regression $x^i = \hat{a} + \hat{b}z$ of each parameter ($i = a, b, c, d$) was determined, and the null hypothesis ($\hat{b} = 0$) for the regression coefficient ($\hat{b}$) was examined by the t-test. Since the test value [9] was larger than the two-sided 5% point $t_{48}(0.975)$ in the t distribution with a latitude of 48, the null hypothesis was rejected in the case of $i = d$. Therefore, the time constant of the exponential decay curve fit to the maximal points during the continuous muscular contraction period (x^d) significantly depends on age ($p < 0.05$).

Myopotentials are induced by changes in the firing pattern of nerve impulses. In sEMG, a very large number of action potential waves in the motor unit (MU) are superimposed. The state of activity of whole muscles is observed by this sEMG [10]. Therefore, it should be considered that sEMG signals are nonlinear, or more generally, sEMG shows a time series produced by stochastic processes. Recently, sEMG data have been recognized to be examined by nonlinear analytic methods, such as the recurrence plot and Wayland algorithm [10,11]. However, fast Fourier

transformation (FFT) generally performed in the previous studies and the measurement parameters proposed in this study are used as a linear analytic method of sEMG. We herein discuss the reason why we have succeeded in finding the linear index showing a correlation with age.

Complexity of the biosignal or degree of visible determinism involved in generator of those signals can be measured by our Double-Wayland algorithm [12]. In each embedding space, the Wayland algorithm estimates a parameter called translation error (E_{trans}) to measure smoothness of flow in an attractor, which is assumed to generate the time-series data [13]. In general, the threshold of the translation error for classifying the time-series data as deterministic or stochastic is 0.5, which is half of the translation error resulting from a random walk [14]. The abovementioned E_{trans} is compared with the translation error (E'_{trans}) estimated from sequences of temporal differences of the time-series data (differenced time series). E'_{trans} would be less than E_{trans} if the degree of determinism involved in the generator were reduced. Using the Double-Wayland algorithm, translation errors E_{trans} and E'_{trans} were estimated from sEMG as shown in Fig. 64.2. Intermittent muscle contraction during the BFT could decrease the translation errors that were estimated from the differenced sEMG. The form of rectangular wave as a teacher signal might reduce randomness or nonlinearity involved in the generator of sEMG. Moreover, we will employ time-series analysis such as surrogate method to ascertain the cause of the correlation between age and a linear index of smoothed sEMG.

The only parameter rejecting the null hypothesis ($\hat{b} = 0$) for the regression coefficient was herein the time constant of the exponential decay curve fit to the maximal points during the continuous muscular contraction period (x^d) ($p < 0.05$). Statistically, the only parameter showing a correlation with age was the time constant (x^d), which decreased with age and became negative over 40 years of age (Fig. 64.1). This strongly suggested that the smoothed sEMG, which should be maintained as constant during muscular contraction in the BFT, was not flat at an age of more than 40 years, indicating that the smoothed sEMG, which should be flat, gradually increased, because of poor muscular regulation function.

The reproducibility of the duration of continuous muscular contraction, which slightly decreased with age, was highest, and this parameter was not correlated with

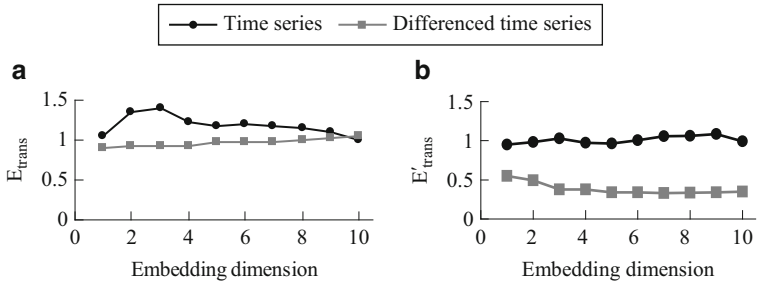


Fig. 64.2 Calculation results involved in sEMG with the use of the Double Wayland algorithm. Translation errors E_{trans} and E'_{trans} were estimated from sEMG measured during a muscle contrac tion period for 3 s (**a**) and the BFT for 3 s (**b**)

age. Although changes in the σ value caused by aging were small in the remaining parameters, it was slightly lower in the group over 60 years than that in the remaining age groups (Table 64.1). The σ value may have been decreased by mechanical output due to the reduction of muscular regulation function.

We showed that the time constant (x^d) was necessary for the evaluation of changes with age using the smoothed EMG during BFT. In other words, reduction of the muscular regulation capacity by aging can be detected via sEMG during BFT using the time constant (x^d). However, even the coefficient of determination involved in the time constants was lower than 0.2 (Fig. 64.1). Using this parameter alone, an evaluation of age-related changes in muscle control might be difficult. A meaningful combination of this parameter with other parameters should be proposed in the next step.

Acknowledgment A part of this study was supported by Hori Foundation.

References

1. Kimura, J.: Electrodiagnosis in diseases of nerve and muscles, 2nd ed., pp. 209 304. FA Davis, Philadelphia (1989)
2. Kizuka, T., Masuda, T., Kiryu, T., Sadoyama, T.: Practical usage of surface electromyogra phy, pp. 65 92. Tokyo Denki University Press, Tokyo (2006)
3. Aukee, P., Penttinen, J., Immonen, P., Airaksinen, O.: Intravaginal surface EMG probe design test for urinary incontinence patients. Acupunt. Electro Ther. Res. Int. J. **27**, 37 44 (2002)
4. Tries, J., Eisman, E.: Urinary incontinence Evaluation and biofeedback treatment. In: Schwartz, M.S., Andrasik, F. (eds.) Biofeedback, pp. 597 629. Guilford, New York (1995)
5. Shiozawa, T., Takada, H., Miyao, M.: Sensor output signal evaluation system. Japan Patent P2006 111387, Apr. 13, 2006.
6. Takada, H., Shiozawa, T., Takada, M., Miyao, M., Kawasaki, H.: Propositions of evaluating indices of muscle performances detected by using surface electromyography and the aging. Bull. Gifu Univ. Med. Sci. **1**, 91 95 (2007)
7. Carlo, J., DeLuca, C.J.: The use of surface electromyography in biomechanics. J. Appl. Biomech. **13**, 135 163 (1997)
8. Takada, H., Shiozawa, T., Takada, M., Iwase, S., Miyao, M.: Evaluating indices of age related muscle performance by using surface electromyography. Proceedings of the 31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 6271 6275 (2006)
9. Shimizu, Y., Takada, H.: Verification of air temperature variation with form of potential. Forma **16**(4), 339 356 (2001)
10. Yhoshida, H., Ujiie, H., Ishimura, K., Wada, M.: The estimation of muscle fatigue using chaos analysis. J. Soc. Biomech. **28**(4), 201 212 (2004)
11. Takada, H., Shiozawa, T., Miyao, M., Nakayama, M., Kawasaki, H.: Theoretical consider ation to set the amplitude of teacher signal in the biofeedback training. Proceedings of the 21st Symposium on Biological and Physiological Engineering, pp. 463 466 (2006)
12. Takada, H., Morimoto, T., Tsunashima, H., Yamazaki, T., Hoshina, H., Miyao, M.: Applica tions of Double Wayland algorithm to detect anomalous signals. Forma **21**(2), 159 167 (2006)
13. Wayland, R., Bromley, D., Pickett, D., Passamante, A.: Recognizing determinism in a time series. Phys. Rev. Lett. **70**, 530 582 (1993)
14. Matsumoto, T., Tokunaga, R., Miyano, T., Tokuda, I.: Chaos and time series, pp. 49 64. Baihukan, Tokyo (2002)

Chapter 65
A Study on Discrete Wavelet-Based Noise Removal from EEG Signals

K. Asaduzzaman, M.B.I. Reaz, F. Mohd-Yasin, K.S. Sim, and M.S. Hussain

Abstract Electroencephalogram (EEG) serves as an extremely valuable tool for clinicians and researchers to study the activity of the brain in a non-invasive manner. It has long been used for the diagnosis of various central nervous system disorders like seizures, epilepsy, and brain damage and for categorizing sleep stages in patients. The artifacts caused by various factors such as Electrooculogram (EOG), eye blink, and Electromyogram (EMG) in EEG signal increases the difficulty in analyzing them. Discrete wavelet transform has been applied in this research for removing noise from the EEG signal. The effectiveness of the noise removal is quantitatively measured using Root Mean Square (RMS) Difference. This paper reports on the effectiveness of wavelet transform applied to the EEG signal as a means of removing noise to retrieve important information related to both healthy and epileptic patients. Wavelet-based noise removal on the EEG signal of both healthy and epileptic subjects was performed using four discrete wavelet functions. With the appropriate choice of the wavelet function (WF), it is possible to remove noise effectively to analyze EEG significantly. Result of this study shows that WF Daubechies 8 (db8) provides the best noise removal from the raw EEG signal of healthy patients, while WF orthogonal Meyer does the same for epileptic patients. This algorithm is intended for FPGA implementation of portable biomedical equipments to detect different brain state in different circumstances.

Keywords Discrete wavelet transform · De-noising · EEG

K. Asaduzzaman (✉)
Faculty of Engineering, Multimedia University, 63100 Cyberjaya, Selangor, Malaysia
e mail: k.asaduzzaman@mmu.edu.my

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 65,

65.1 Introduction

The EEG signal is defined as an electrical activity of an alternating type recorded from the scalp surface after being picked up by metal electrodes and conductive media [1]. EEGs have long been used for the diagnosis of various central nervous system disorders like seizures, epilepsy, and brain damage and for categorizing sleep stages in patients. As an effective and common method for investigating the brain activities, EEG signal has been playing a very important role in scientific research and clinical application for many years [2].

The existing de-noising techniques that are based on frequency-selective filtering lead to a substantial loss of the EEG data. Prohibiting the subjects from doing their natural works is not a plausible solution, and in fact this effort of the subject in ensuring that he/she does not perform the aforementioned actions can have a significant impact on the recorded EEG. Due to these factors, frequency-selective filtering methods for removing noise from EEG signal are considered as a major challenge today [3]. Wavelet-based filtering is an attractive alternative considering its ability to study both the time and the frequency maps simultaneously [4]. An approach based on stationary wavelet transform (SWT) is used to de-noise the EEG by Zikov et al. [5]. As the artifacts related to the recorded EEG signal are significantly uncorrelated, the reconstructed signal in the SWT method is often not a very good approximation of the original EEG. Ramanan et al. [6] described another approach of de-noising EEG signal using HAAR wavelet of higher order. But this method is only applicable for removing noise related to eye movements.

A discrete wavelet-based noise removal is performed in this research to remove artifacts from EEG signal. Wavelet de-noising (noise removal) has been found to be effective in de-noising a number of physiological signals [7]. It is preferred over signal frequency domain filtering because it tends to preserve signal characteristics even while minimizing noise. This is because a number of threshold strategies are available, allowing reconstruction based on selected coefficients [8].

For this particular research, wavelet functions (WF) Daubechies (db2, db6, db8) and Meyer (dmey) are used during the wavelet transform for noise removal. These WFs are chosen based on the shapes of the mother wavelet, which are similar to that of EEG signal [7, 9]. RMS difference was calculated to measure the effectiveness of the noise removal using these wavelets.

Results show that WFs db8 provides the best noise removal from the raw EEG signal of healthy patients. For the case of epileptic patients, WF orthogonal Meyer presents high RMS difference compared to the other three WFs, resulting in better noise removal for the EEG signal.

65.2 Methodology

65.2.1 Data Selection and Recording Techniques

For this experiment, four sets (denoted A D) each containing 20 single channel EEG segments of 23.6 s duration were composed. These data sets have been collected from the database of the Department of Epileptology of University of Bonn, Germany. Sets A and B consist of segments taken from surface EEG recording carried out on five healthy volunteers using standardized electrode placement scheme. Data for set A were recorded by keeping the volunteers in an awake state with eyes open, while set B was recorded by keeping the eyes closed. Data for sets C and D were collected from five patients, all of whom had achieved complete seizure control. Segment in set C contains only activity measured during seizure-free intervals, while set D contains the recordings of seizure activity. All EEG signals were recorded with the same 128-channel amplifier system using as average common reference. After 12-bit analog-to-digital conversion, the data were stored continuously at a sampling rate of 173.61 Hz [10].

These EEG signals were de-noised using discrete wavelet transform (DWT) and a threshold method. The DWT and threshold-based de-noising were implemented using MATLAB Wavelet toolbox. Figure 65.1 below shows the flow of the algorithm.

65.2.2 Wavelet De-noising

Wavelets commonly used for de-noising biomedical signals include the Daubechies "db2", "db8", and "db6" wavelets and orthogonal Meyer wavelet. The wavelets are generally chosen whose shapes are similar to those of the EEG signal [7]. Figure 65.2 gives the process of noise removal using wavelet transform.

Fig. 65.1 Algorithm flow

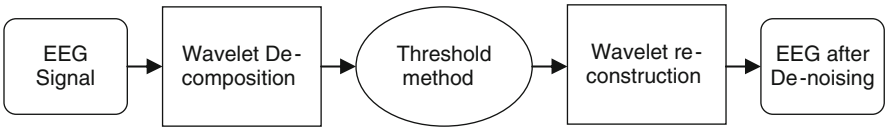


Fig. 65.2 Wavelet de noising method

65.2.2.1 Wavelet Decomposition

The WT decomposes a signal into several multi-resolution components according to a basic function called the wavelet function. Filters are one of the most widely used signal processing functions. The resolution of the signal, which is a measure of the amount of detailed information in the signal, is determined by the filtering operations, and the scale is determined by upsampling and downsampling (subsampling) operations. The DWT is computed by successive low pass and high pass filtering of the discrete time-domain signal, as shown in Fig. 65.3. In the figure, the signal is denoted by the sequence $x[n]$, where n is an integer. The low pass filter is denoted by G_0, while the high pass filter is denoted by H_0. At each level, the high pass filter produces detailed information $d[n]$, while the low pass filter associated with scaling function produces coarse approximations $a[n]$. With this approach, the time resolution becomes arbitrarily good at high frequencies, while the frequency resolution becomes arbitrarily good at low frequencies.

65.2.2.2 Threshold Method

Suppose that the noisy EEG signal f equals the original EEG signal s plus the noise n. The threshold method is applied as follows:

1. The energy of the original signal s is effectively captured, to a high percentage, by transform values whose magnitude is all greater than a threshold, $T_s > 0$.
2. The noise signal's transform values all have the magnitudes that lie below a noise threshold T_n, satisfying $T_n < T_s$.

Then the noise in f can be removed by thresholding its transform. All values of its transform whose magnitude lies below the noise threshold T_n are set equal to 0.

65.2.2.3 Signal Reconstruction

An inverse transform is performed, providing a good approximation of f. The reconstruction is the reverse process of decomposition. The approximation and

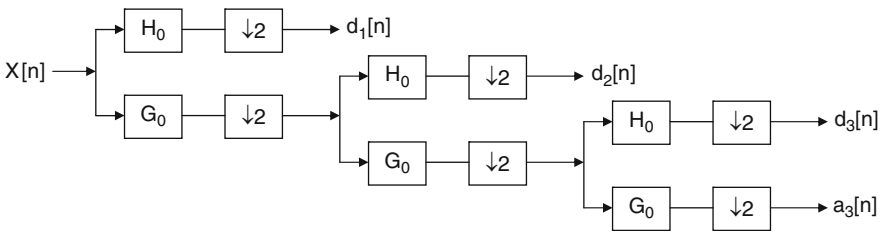


Fig. 65.3 Three level wavelet decomposition tree

detail coefficients at every level are upsampled by two, passed through the low pass and high pass synthesis filters, and then added. This process is continued through the same number of levels as in the decomposition process to obtain the original signal.

65.2.3 RMS Difference Calculation

The RMS difference of the noisy EEG signal f compared with the noise-free EEG signal s is defined by (65.1)

$$\text{RMS difference} = \sqrt{\frac{(f_1 - s_1)^2 + (f_2 - s_2)^2 + \cdots + (f_N - s_N)^2}{N}}. \tag{65.1}$$

The RMS difference was calculated for the four WFs, where f is the noisy EEG signal and s is the signal after de-noising. N is the number of samples.

65.3 Results and Discussion

Wavelet de-noising method is applied to noisy EEG signal at different physical conditions (eyes open, eyes closed, after seizure, and during seizure). RMS difference was calculated for each of the WFs (db2, db6, db8, and dmey) during all physical states. The results of the calculation are listed in Table 65.1.

From the RMS difference values listed in Table 65.1, it can be observed that for the case of healthy patients (with their eyes open and closed), WFs db4, db6, and dmey give similar kind of RMS difference, while RMS difference is high using db8 WF. This means that it is more effective during the noise removal process for the EEG signal of healthy patients. For the case of epileptic patients, the best result was obtained by using the orthogonal Meyer wavelet function. Figure 65.4 illustrates the result of de-noising method using different WFs with four levels of decomposition for a sample EEG signal.

Table 65.1 RMS difference of EEG signal at various physical conditions for four WFs

Physical condition	db4	db6	db8	dmey
Eyes open (set A)	26.2475	26.2209	26.3206	26.248
Eyes closed (set B)	56.6761	56.8987	56.8963	56.257
After seizure (set C)	22.4526	22.9973	22.9584	23.3889
During seizure (set D)	248.8886	254.1931	255.585	255.8324

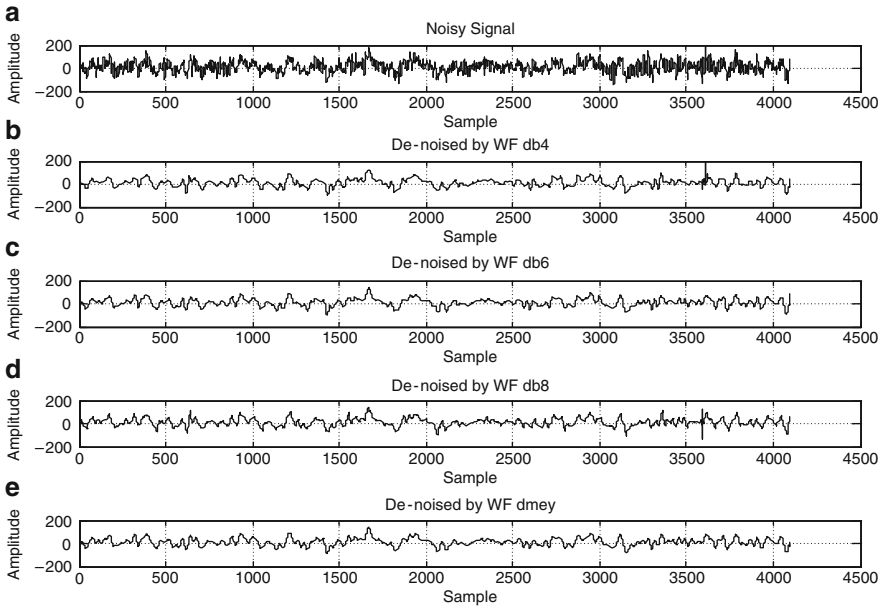


Fig. 65.4 (**a**) Noisy EEG signal of a healthy patient, result of wavelet de noising performed using (**b**) "db4" wavelet, (**c**) "db6" wavelet, (**d**) "db8" wavelet, and (**e**) "dmey" wavelet with four levels of decomposition

65.4 Conclusion

Wavelet-based noise removal has the added advantage that it is fast and easy to implement. Wavelet theory has already enjoyed great success in other biomedical signal processing, and is expected to provide a powerful complement to conventional noise-removal techniques (such as stationary wavelet transform and frequency-selective filtering, as mentioned earlier) for EEG signals. All four WFs can effectively remove noise from EEG signals but according to this research, WF db8 is found to be the most efficient for removing noise from EEG signal of healthy patients, while WF orthogonal Meyer is found to be the most effective for epileptic patients. The wavelet-based noise removal technique proposed in this research can be used for the analysis and characterization of EEG signal to different brain activity.

References

1. Niedermeyer E, Silva FH (1993) Electroencephalography: Basic principles, clinical applica tions and related fields. Lippincott, Williams & Wilkins, Philadelphia.
2. Holmes GL, Lombroso CT (1993) Prognostic value of background patterns in the neonatal EEG. J Clin Neurophysiol 10:323 352.

3. Croft RJ, Barry RJ (2000) Removal of ocular artifact from the EEG: A review. Clin Neurophysiol 30:5 19.
4. Unser M, Aldroubi A (1996) A review of wavelet in biomedical applications. IEEE Trans Biomed Eng 84:626 638.
5. Zikov T, Bibian S et al (2002) A wavelet based de noising technique for ocular artifact correction of the electroencephalogram. Proceedings of the Second Joint EMBS/BMES Conference, pp. 98 105.
6. Ramanan SV, Kalpakam NV et al (2004) A novel wavelet based technique for detection and de noising of ocular artifact in normal and epileptic electroencephalogram. Proceedings of International Conference on Communication, Circuits and Signals 2:1027 1031.
7. Carre P, Leman H et al (1998) Denoising of the uterine EHG by an undecimated wavelet transform. IEEE Trans Biomed Eng 45:1104 1114.
8. Mallat S (1998) A wavelet tour of signal processing. Academic, New York.
9. Wachowiak MP, Rash GS et al (2000) Wavelet based noise removal for biomechanical signals: A comparative study. IEEE Trans Biomed Eng 47:360 368.
10. Andrzejak RG, Lehnertz K et al (2001) Indications of nonlinear deterministic and finite dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. Phys Rev E 64:061907.

Chapter 66
Enhancing the Communication Flow Between Alzheimer Patients, Caregivers, and Neuropsychologists

Abraham Rodriguez-Rodriguez, Leidia Martel-Monagas, and Aaron Lopez-Rodriguez

Abstract It is estimated that by 2050 over 100 million people will be affected by the Alzheimer's disease (AD). It not only affects the patient, but also the whole family and specially the caregiver, who is continuously under great stress conditions. We propose a software environment, called *Mnemosine*, designed to improve the quality of life of both Alzheimer patients and caregivers and to enhance the communication with the neuropsychologists. *Mnemosine* will provide them with resources that make easier the disease monitoring and facilitates the patient's daily activities. *Mnemosine* also includes alarms and agenda modules to manage the patient's daily routine, a GPS route guidance and location system, and a reports utility to evaluate disease progression.

Keywords Alzheimer's disease · Caregiver · Monitoring · Software environment

66.1 Introduction

Alzheimer's disease (AD) is a progressive neurological illness of the brain that leads to the irreversible loss of neurons and dementia. Currently, AD is the most common cause of dementia and the third cause of mortality, just behind cardiovascular diseases and cancer. According to Alzheimer's Disease International [1], there are currently 30 million people with dementia in the world, with 4.6 million new cases annually. The number of people affected will be over 100 million by 2050 [2]. Studies have shown that only in the United States AD prevalence is estimated to be 1.6% in the 65 76 age group, 19% in the 75 84 group, and 42% in the greater than 84 group [3].

A. Rodriguez Rodriguez
Departamento de Informática y Sistemas, University of Las Palmas de Gran Canaria, Canary Islands, Spain
e mail: arodriguez@dis.ulpgc.es

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 66,

Clinical studies have shown that the mean life expectancy after diagnosis is approximately 7 years [4], and although there is no cure for AD, with appropriate treatment the mean life expectancy can be much longer [5]. AD may be among the most costly diseases for society in Europe and United States, with costs as high as $160 billion all over the world [6]. For that reason, any therapy that slows cognitive decline, thus delaying institutionalization of the patients and reducing caregivers' efforts will have economic benefits [7].

Additionally, it is important to realize that AD not only affects the patient, but also the whole family and specially the caregiver. The physical and emotional stress they have to bear is enormous [8]. Over 65% of the relatives who take care of Alzheimer patients will suffer a significant change in their lives and an important leak in their physical or psychical health, resulting about the 20% of them in a clinical profile known as "burn-out caregiver" [9]. As AD pushes a great burden on caregivers who must take care of patients continuously, sometimes even leaving their jobs in order to take care of the patient. It is also important to improve caregivers' life conditions [10, 11].

The main parts involved in the daily care of an Alzheimer patient are the patient itself, the caregivers (may be more than one), and the neuropsychologist. In this context, we propose a technological solution aimed at ease to their tasks by enforcing the communication flow between them, thus enabling all of them to be informed at any moment about the state and evolution of the patient's disease.

66.2 Mnemosine

Our proposal is a software development, called *Mnemosine*, designed to improve the quality of life of both Alzheimer patients and caregivers, and to enhance the communication flow with the neuropsychologists. It is essentially focused on Alzheimer patients in the two initial stages of disease because they still keep most of their independence [12]. *Mnemosine* can help people with AD and their caregivers with techniques based on improving the strengths and abilities of patients. This allows those people with AD to keep their self-esteem and self-confidence through their illness [13]. This framework can help caregivers and neuropsychologists to develop plans that address the activities of daily living that maximize independence, improve function, and minimize the need for support. One of the main characteristics of *Mnemosine* is its adaptation to every individual patient, as it has been shown that the treatment techniques that take into account the personal history, character, and the individuality of the person with AD have a positive impact on the progress of the disease.

It is known that caregivers of AD patients are under stress conditions whose effect depends not only on the patient's deficits, but also on the caregivers' own characteristics [14]. No matter the reasons that cause the caregiver burden, with *Mnemosine* the caregiver is partially freed of been continuously paying attention to the patient with his common tasks. It provides the patient with some autonomy

while preserving the quality of the control exerted over him. Besides, it also eases the scheduling of the patient's daily activities and the elaboration of diverse personalized materials that can be used in the patient's cognitive exercises.

From the neuropsychologist point of view, this framework eases the communication flow with the patient and his environment. It also allows him to enhance the assessment of the disease evolution, as he could access the daily log of the patient activity. All this flexibility can be further exploited by reusing positive experiences that are carried out on similar patients.

66.2.1 Capabilities

All the previously depicted characteristics of *Mnemosine* are possible by the smooth integration of the following capabilities:

- *Personal memory*: This tool allows the edition of text and/or several multimedia resources (video, pictures, audio, or even a graphical representation of relative's relations), depicting past lived experiences (a child's birth, wedding, studies, etc.), or the social circle of the patient (relatives, friend, colleagues, etc.).
- *Cognitive exercises*: It allows the neuropsychologist to manage cognitive exercises, including creation and edition. On the patient's side it allows the resolution of exercises to stimulate the cognitive capabilities.
- *Scheduling and monitoring the daily activity of the patient*: If we assume that daily routine is a key factor in Alzheimer's patients, this tool makes possible the scheduling of a limited set of activities for a certain period of time, including multimedia demonstrations of how he could perform them.
- *Patient's positioning system*: *Mnemosine* allows the definition of a geographic route, such as a walk around the neighborhood, or just going out to the grocer's for milk, including the definition of a set of alerts in case the patient walks outside a secure area, or does not come back inside the scheduled time. Among the types of alarms that *Mnemosine* can trigger, there are phone calls autonomously made by the patient's PDA device to relatives or caregivers, or SMSs sent to caregivers showing the GPS coordinates and the event that caused the warning.
- *Analysis of the disease progression*: Logs with the everyday activity of the patient are recorded on the mobile device and thus can be easily uploaded to the neuropsychologist's computer. All this information will be later presented to him through predefined reports which constitute a very useful tool to analyze the evolution of the disease.

66.2.2 Information Flow in Mnemosine

The daily activity of an Alzheimer's patient is dotted with several events that, if known, would be of interest to both the caregiver and neuropsychologist. These events could help the caregiver to interpret the patient's current status and

anticipate potential problems. All this information could be also critical in helping the neuropsychologist to analyze the disease evolution. However, there is currently no methodology or tool that efficiently records all the activities that an Alzheimer's patient does during the day, even though some of these events should be communicated urgently to the caregiver (as would be the case of a patient that gets disoriented during a walk).

As technology can play a significant role to overcome this problem, *Mnemosine* was conceived considering all the interactions between the people involved with the AD. The neuropsychologist is responsible for designing and scheduling the cognitive exercises for the patient. He also collects the activity log from the AD personal device to process his daily activity, such as walks, social events, answers to exercises, etc. The caregiver set up the patient's agenda and manages his life book, updating the information as needed. The caregiver must coordinate with the medical specialist when preparing the agenda so that every requirement and recommendation indicated by the doctor can be included on it. *Mnemosine* also implements the information flow between the caregiver and the medical specialist, and so the caregiver can receive personalized messages from the doctor and, on the other part, he is also able to send the doctor anything he may consider relevant.

66.2.3 Components

We have distributed the capabilities described in the previous section into two platforms which will be used by patients, caregivers, and neuropsychologists:

- *Mnemosine*-Mobile: It is integrated into a mobile device incorporating a GPS, such as a PDA-cell phone or Smartphone. It will be used by the patient during his everyday activities (demos, walks, exercises), to remember relevant past experiences, or just to recall some information about relatives or his social circle. At the same time, it works as a positioning device useful to locate the patient in case of an emergency.
- *Mnemosine*-Home: This module runs on a personal computer and it can be used both by the caregiver and the neuropsychologist for creating demos, schedule activities, design controlled walk, describe past lived experiences, or input data into the patient's life book. The neuropsychologist will use it to design and manage cognitive exercises, import user's activity logs, and generate several reports with the patient's activity.

All the functionalities previously described have been implemented in several modules. Figure 66.1 (left) shows a piece of the patient's life book with a graphical representation of the social network of the patient. It is also possible to add new events describing past relevant experiences and group them by social affinity (relatives, friends, works, and studies) using any combination of text and multimedia resources. It is also possible to browse the life book using the mobile device as shown in Fig. 66.1 (right).

Fig. 66.1 (*Left*) Mnemosine Home: Patient's life book showing some social relations. (*Right*) Mnemosine Mobile. Basic browsing of patient's life book

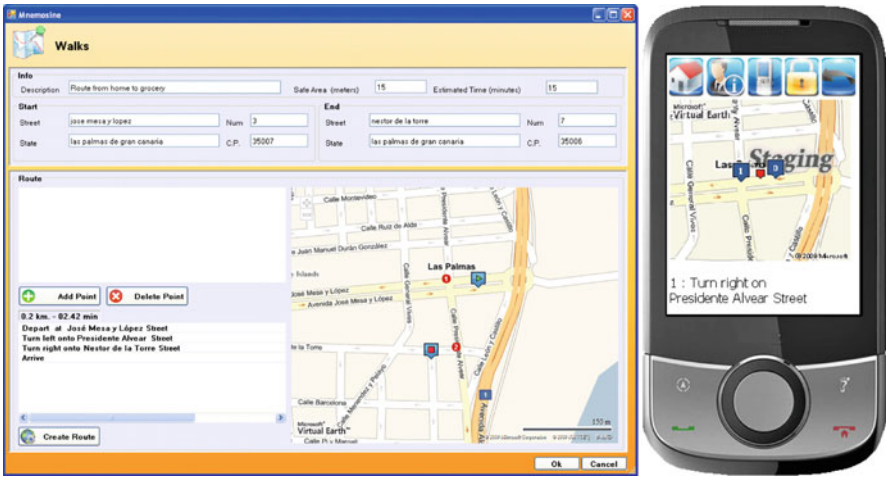


Fig. 66.2 Figure 5. (*Left*) Mnemosine Home: Route definition interface. (*Right*) Mnemosine Mobile: Route activated showing next instruction

The management of the cognitive exercises made through the *Mnemosine*-Home interface. The neuropsychologist can create, edit, or assign exercises to any of his patients (assuming no personal information is reused between patients). This module can import the information and resources from the patient's life book when designing new cognitive exercises. The use of personal (private) elements from the patient life can act as a motivational element for solving those exercises.

All the exercises will be solved by the patient using the *Mnemosine*-Mobile interface.

The scheduling of the activities that the patient will have to perform during the day can be organized by the caregiver with the interface developed for *Mnemosine*-Home. The user can browse the scheduled activities in the mobile device. An alarm policy has been implemented, and so the caregiver will be warned in case the patient forgets, or does not complete, any of the scheduled activities.

Routes can be easily defined by the caregiver using the interface shown in Fig. 66.2 (left). The patient will have the GPS activated and the mobile device will send SMS, or even make a phone call to the predefined phone numbers, in case the patient walks out a defined secure path or the time elapsed exceeds the prescribed period made by the caregiver.

66.3 Conclusions

AD is the most common cause of dementia nowadays. It mainly affects older people, and it is estimated that by 2050 over 100 million people will be affected. AD not only affects the patient, but also the whole family and specially the caregiver, who is continuously under great stress conditions. Although there is no cure to AD, some treatments, including memory interventions, are recommended to slow cognitive decline of the patient, thus delaying institutionalization and reducing caregivers' efforts. Moreover, caregiving of Alzheimer patients has significant cost implications.

We propose a software development, called *Mnemosine*, designed to improve the quality of life of both Alzheimer patients and caregivers and to enhance the communication flow with the neuropsychologists. We stimulate the brain and cognitive capabilities of the patient as well as we provide the neuropsychologist and the caregiver with resources that ease monitoring the disease evolution and controlling the patient's daily activity. This allows a better coordination and integration of all the actions comprising the patient's daily care.

Mnemosine capabilities include functions to manage the life book of the patient, where different relevant experiences of the patient are stored in any multimedia format. It also includes all the management functions for adding, modifying, and carrying out cognitive exercises. With *Mnemosine* the caregiver can schedule and monitor the daily activities of the patient, including the physical monitoring of the patient's situation with the GPS positioning system. It also allows making reports about disease evolution that can help the neuropsychologist to schedule new therapies.

Our proposal has the advantage that it is portable for the patient, which can improve its self-confidence, individuality, and reduces the need for support, as he can be assessed in his activities everywhere. Other advantages include the alarm system and the agenda for the patient's daily routine. Moreover, this tool can help patients who cannot attend Alzheimer associations or the neuropsychologist

because of a lack of mobility, center distance, shame (which is quite common in famous people affected), or initial states of the disease.

We strongly believe that this initial architecture is the base for including more functionalities, thus extending it to other types of dementia, or including characteristics as the integration of set-of-box devices (i.e., TDT) or domotics, or even the addition of social profiles coming from the internet.

References

1. Alzheimer's Disease International (2008) *The Prevalence of Dementia Worldwide*. Available at: http://www.alz.co.uk/adi/pdf/prevalence.pdf. Accessed 1 December 2009
2. Ferri CP, Prince M, Brayne C (2005) *Global Prevalence of Dementia: A Delphi Consensus Study*. Lancet 366 (9503): 2112 2117
3. Hebert LE, Scherr PA, Bienias JL, Bennett DA, Evans DA (2003) *Alzheimer Disease in the US Population: Prevalence Estimates Using the 2000 Census*. Archives of Neurology 60 (8): 1119 1122
4. Mölsä PK, Marttila RJ, Rinne UK (1986) *Survival and Cause of Death in Alzheimer's Disease and Multi Infarct Dementia*. Acta Neurologica Scandinavica 74 (2): 103 107
5. Fratiglioni L, Launer LJ, Andersen K, Breteler MM,Copeland JR, Dartigues JF (2000) *Incidence of Dementia and Major Subtypes in Europe: A Collaborative Study of Population based Cohorts. Neurologic Diseases in the Elderly Research Group*. Neurology 54 (11 Suppl 5): S10 S15
6. Wimo A, Jonsson L, Winblad B (2006) *An Estimate of the Worldwide Prevalence and Direct Costs of Dementia in 2003*. Dementia and Geriatric Cognitive Disorders 21 (3): 175 181
7. Meek PD, McKeithan K, Schumock GT (1998) *Economic Considerations in Alzheimer's Disease*. Pharmacotherapy 18 (2 Pt 2): 68 73; discussion 79 82
8. Bertolote JM (1994) *Ayuda para cuidadores de personas con demencia*. World Health Organization, p. 23, Available at http://www.alz.co.uk/adi/pdf/eshelpforcaregivers.pdf
9. Rodríguez A, *Sobrecarga psicofísica en familiares cuidadores de enfermos de Alzheimer*, Fundación SPF de Neurociencias. Available at http://www.psicologia online.com/colaboradores/delalamo/alzheimer.shtml. Accessed 1 December 2009
10. Thompson CA, Spilsbury K, Hall J, Birks Y, Barnes C, Adamson J (2007) *Systematic Review of Information and Support Interventions for Caregivers of People with Dementia*. BMC Geriatrics 7: 1
11. Schneider J, Murray J, Banerjee S, Mann A (1999) *EUROCARE: A Cross national Study of Co resident Spouse Carers for People with Alzheimer's Disease: I Factors Associated with Carer Burden*. International Journal of Geriatric Psychiatry 14 (8): 651 661
12. Fundación Antidemencia Al Andalus (2009) *Consejos para familiares y enfermos de Alzheimer*. Portal Alzheimer Online Available at http://www.alzheimer online.org/. Accessed 1 December 2009
13. Dementia: *A Quick Reference Guide* (2006) National Institute for Health and Clinical Excellence
14. Donaldson C, Burns A (1999) *Burden of Alzheimer's Disease: Helping the Patient and Caregiver*. Journal of Geriatric Psychiatry and Neurology 12 (1): 21 28

Chapter 67
An Improved 1-D Gel Electrophoresis Image Analysis System

Yassin Labyed, Naima Kaabouch, Richard R. Schultz, Brij B. Singh, and Barry Milavetz

Abstract Images obtained through the gel electrophoresis technique contain important genetic information. However, due to degradations and abnormalities from which these images suffer, extracting this information can be a tedious task and may lead to reproducibility issues. Image processing techniques that are commonly used to analyze gel electrophoresis images require three main steps: band detection, band matching, and quantification. Although several techniques were proposed to automate all steps fully, gel image analysis still requires researchers to extract information manually. This type of extraction is time consuming and subject to human errors. This paper proposes a fully automated system to analyze the gel electrophoresis images. This system involves four main steps: lane separation, lane segmentation, band detection, and data quantification.

Keywords 1-D Gel electrophoresis · Band detection · Band matching · Data quantification · Image analysis · Lane segmentation

67.1 Introduction

Gel electrophoresis is a technique for separating DNA, RNA, or protein fragments according to their molecular weights by forcing them to migrate through a substrate, such as polyacrylamide gel, under the influence of an electric field. After the process of electrophoresis, the gel is stained and then visualized on a UV transilluminator and captured by a camera. The resulting gel image consists of

N. Kaabouch (✉)
Department of Electrical Engineering, School of Engineering and Mines, University of North Dakota, Grand Forks, ND 58202 7165, USA
e mail: naimakaabouch@mail.und.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 67,

vertical lanes containing a number of horizontal bands (fragments of DNA or protein).

The amount of substance in each band is estimated by calculating the area of the band, and the molecular weight of each band is estimated by considering the position relative to a predefined reference band. Therefore, obtaining accurate genetic information from gel images depends on several parameters, including the quality of the bands isolated. Because of experimental errors and the qualities of the gel images, the images are not exploitable in many situations. Moreover, researchers spend a great deal of time manually extracting data from gel images, which leads to reproducibility issues.

Image processing techniques commonly used to analyze gel electrophoresis images require three main steps: band detection, band matching, and quantification. Several software systems have been developed to analyze the electrophoresis gel images automatically [1 7]. Some of these systems are semiautomatic and perform band detection by segmenting the image into lanes and locating the peaks of the 1D mean profile [4] or of the cumulative row difference profile of each lane [5]. However, these methods have major disadvantages because they require the user to select the region of interest and adjust different parameters manually [4, 5]. Other software systems identify bands by extracting the variance and mean variance of the 1D mean profile of the lane and classifying the valleys of the profiles as either noise or bands. Nevertheless, these methods cannot generally locate faint bands, and they sometimes detect false bands due to noise.

In previous work [7], an automated system to analyze and quantify simple and duplex DNA images automatically was proposed. The assessment of this system on other types of gel electrophoresis images revealed some limitations. Based on the analysis of these limitations, an improved system was developed that integrates a better thresholding technique, a more efficient band detection technique, additional data quantification functions, and a graphical user interface. This system is described below.

67.2 Methodology

The proposed system involves four steps:

1. Lane separation. This step consists of separating the images into lanes.
2. Lane segmentation. This step consists of applying an appropriate automatic thresholding technique in order to separate the bands from the noisy background of the lanes.
3. Band detection. This step consists of automatically detecting the location of each band in the lane.
4. Data quantification. This step consists of computing the amount of the substance in each band and its molecular weight.

67.2.1 Lane Separation

To simplify the processing of the image and to increase the efficiency of the thresholding technique, the image is automatically divided into lanes. The algorithm used is based on the derivative of the horizontal intensity projection of the gel image. The horizontal intensity projection is as follows:

$$S(i) = \sum_{j=1}^{M} I(i,j) \tag{67.1}$$

Here, $S(i)$ is the sum of the pixel intensities of column i, M is the number of rows in the image, i and j are the coordinates of the pixel located at the row i and column j.

Because this profile contains peaks that indicate the location of lanes, a first approach is to use the locations of these peaks to divide the image into lanes. However, this technique requires the users to set local threshold levels that depend on the quality of the gel image. A better approach is to use the variations of this profile to identify the gaps between lanes. In these gaps, the background levels change slowly compared to the levels of the lanes, while the left and right edges of each lane show rapid changes. Thus, by locating these rapid changes, one can locate the edges of each lane and divide the image into lanes.

67.2.2 Lane Segmentation

The thresholding technique described in [7] works well on single and duplex DNA. However, this technique does not provide good results on images corresponding to multiple DNA or protein bands per lane, such as the images in Figs. 67.1 and 67.2. Because of the number of bands, these types of images present additional

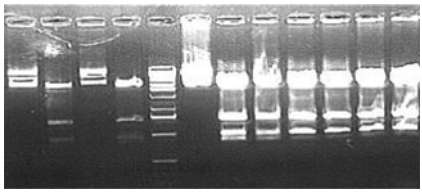


Fig. 67.1 Example of a poor quality gel image

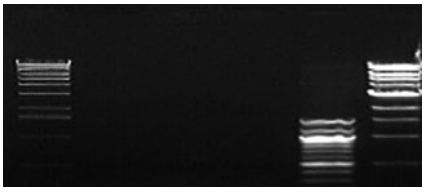


Fig. 67.2 Example of a good quality gel image

challenges. First, the images are noisier and their backgrounds are strongly nonuniform. Second, the bands can be very close to each other and may overlap. Finally, some of the bands are faint and their intensity levels are low compared to the background levels.

The proposed thresholding technique is based on estimating the background intensity level of the gel image lane from the profile of the vertical intensity projection of that lane. This profile is obtained by adding the pixel intensities of every row of the lane. From this profile, the background levels are approximated by the equations of the lines that pass through the successive minima of the intensity projection profile, as shown in Fig. 67.5. The local minima of the profile are denoted by dots. Every two minima of the profile are connected to form a line. The steps for the background removal are described as follows:

1. Obtain the intensity projection profile of row pixel intensities of the lane.
2. Find the row indices of the local minima of this profile and assign the total number of local minima to T. Assuming that the curve joining two successive local minima of the profile is linear, we obtain an equation of the form

$$y_i(x) = m_i x + c_i, \tag{67.2}$$

where m is the slope of the line joining the two successive minima, c is the constant of the line equation, i is the equation index and varies from 1 to $(T - 1)$, x is the row index of the lane, and y is the background intensity value.
3. Find the coefficients m_i and c_i for every line $y(x) = m_i x + c_i$.
4. For every pixel intensity of row x of the lane, we subtract the corresponding background intensity value given by $y(x)$.

67.2.3 Band Detection

Following the separation of the gel image lanes, a band detection technique is applied to each lane to locate the bands present within [8]. This band detection approach is similar to the one proposed to divide the gel image into lanes. First, the profile of the vertical intensity projection of the lane is calculated, and the derivative is from this profile. From the variations of this profile, the locations of the bands are detected by finding the local maxima and minima.

67.2.4 Quantification

Two types of genetic data are computed by this system for each band, the molecular weight and the amount of substance. The molecular weight is calculated using the band matching technique. This technique compares the vertical location of a band I

in the lane J to the vertical locations of the reference lane bands. The molecular weight of the band will be equal to the molecular weight of the closest band in the reference lane.

The second type of genetic data, the amount of substance, is estimated by calculating the area of each band. This information is stored in a file along with all separated lanes and the results of segmented images for further work. The system also presents the flexibility to delete false bands or to add missing bands.

67.3 Results

Figures 67.1 and 67.2 show two examples of poor and good quality electrophoresis gel images, respectively. As observed, the poor quality image contains nonuniform background and noisy stains. Additionally, some of the bands are very bright with long-tailed shapes, while others are very faint. The image of Fig. 67.2, on the other hand, has a more uniform background; however, some of the bands are very faint, and additionally, some of them are also very close to each other.

Figure 67.3 shows the profile of the horizontal intensity projection corresponding to the image of Fig. 67.1. Figure 67.4 shows the variations of the profile in Fig. 67.3 obtained by computing the derivative of this horizontal intensity projection. As can be seen, each lane is demarcated by a maximum and a minimum.

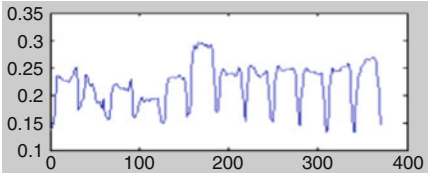


Fig. 67.3 Horizontal intensity projection profile corresponding to Fig. 67.1

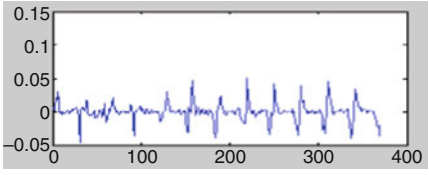


Fig. 67.4 Derivative profile of Fig. 67.3 signal

Fig. 67.5 Example of a separated lane from the image of Fig. 67.1

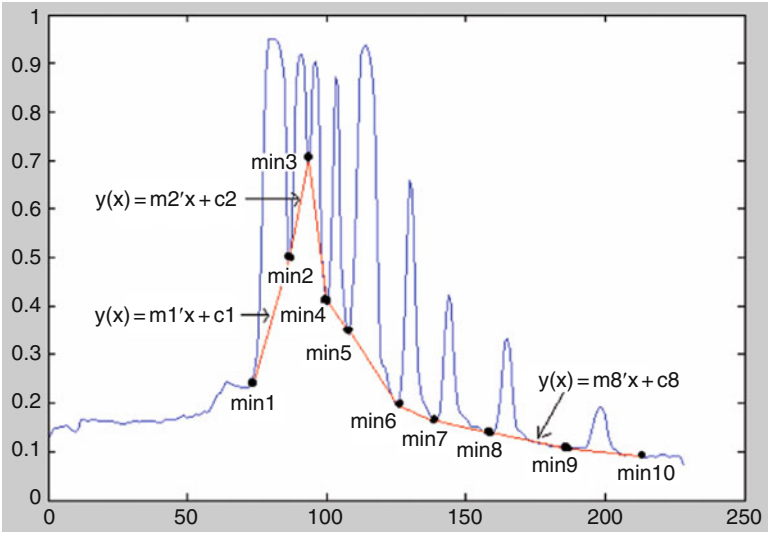


Fig. 67.6 Profile of Fig. 67.5 showing the local minima and the lines connecting them

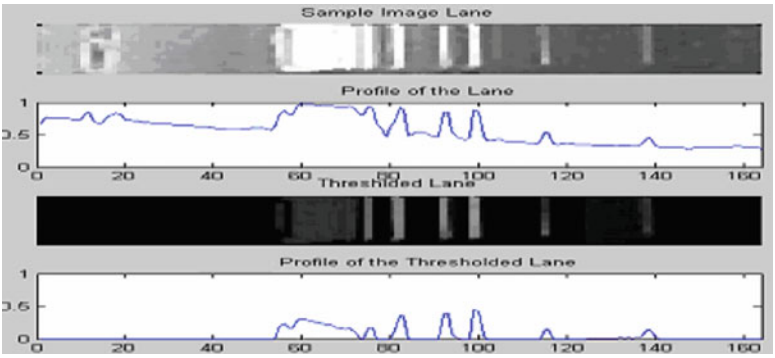


Fig. 67.7 *Lane 5* of Fig. 67.1 and its output after thresholding as well as their corresponding profiles

Figure 67.6 represents a typical intensity projection profile corresponding to the separated lane of Fig. 67.5. This Fig. 67.6 also shows the technique used to approximate the background levels, as was previously explained in Sect. 67.2.2. Figures 67.7 and 67.8 give two examples of lanes and their corresponding outputs after thresholding. As can be seen from the profiles of the thresholded lanes, the proposed technique is able to segment the bands from the background.

Figures 67.9 and 67.10 give examples of results after applying the band detection technique previously explained in Sect. 67.2.3. As can be observed, unlike the

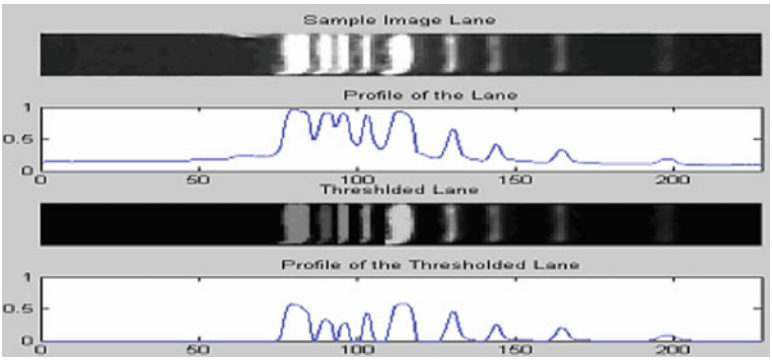


Fig. 67.8 *Lane 3* of Fig. 67.2 and its output after thresholding as well as their corresponding profiles

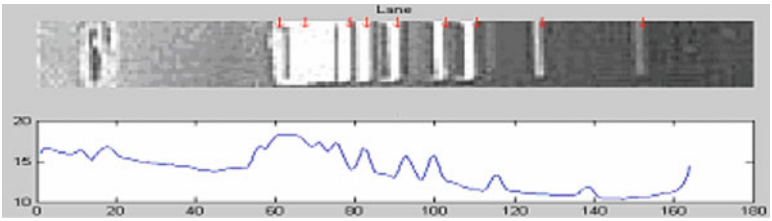


Fig. 67.9 Results after band detection using the intensity projection approach corresponding to *lane 5* of Fig. 67.1 image

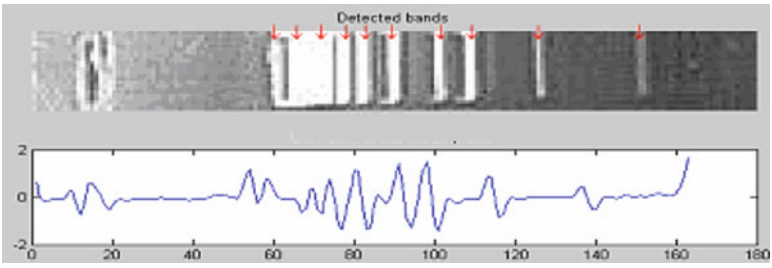


Fig. 67.10 Results after band detection using the proposed technique corresponding to *lane 5* of Fig. 67.1 image

intensity projection technique, the proposed band detection technique is able to identify all bands including overlapping bands.

Figure 67.11 shows the developed graphical user interface of the proposed system. This system allows one to select an image, to segment and divide the image into lanes, to specify the reference lane, and to analyze the bands lane by lane. This system also provides the users with the flexibility of adding a missing

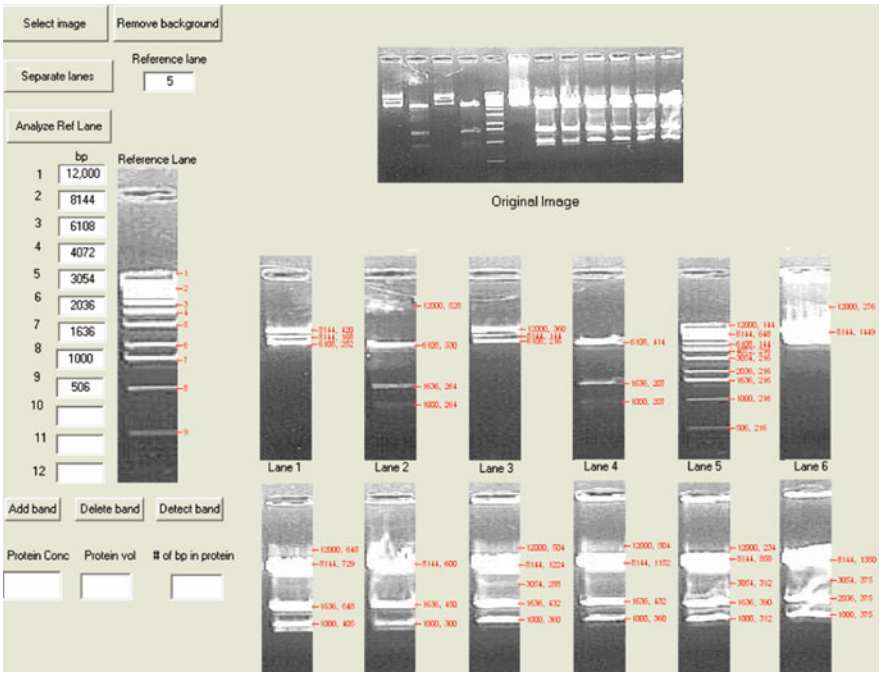


Fig. 67.11 Graphical user interface of the proposed system

Table 67.1 Results of Figs. 67.1 and 67.2 assessments

Figure	Technique	Detected bands	Missing bands	False bands
Figure 67.1	Intensity projection	34	6	12
	Proposed technique	35	0	7
Figure 67.2	Intensity projection	29	3	4
	Proposed technique	27	1	0

band or deleting a false band in a lane. Additionally, this system computes the most important data for each band, including the molecular weight and the amount of the substance, which are automatically saved in a folder that can be used for further work.

To test the performance and accuracy of the proposed system, we used a set of 80 gel electrophoresis images. Of these images, 29 contain simplex DNA bands, 17 images contain duplex DNA bands, and the remaining images correspond to DNA or protein with multiple bands in each lane. The assessment of the efficiency of the proposed system is based on the numbers of detected bands, missing bands, and false bands in each gel image. Examples of results are shown in Table 67.1, corresponding to Figs. 67.1 and 67.2. As can be observed, the proposed system

presents fewer missing and false bands than the intensity projection-based approach does. Although these numbers change from image to image, the proposed system shows greater efficiency than other techniques in thresholding, detecting, and analyzing the gel electrophoresis image bands.

67.4 Conclusion

A fully automatic system for analyzing gel electrophoresis images is proposed. This system involves four main steps: lane separation, lane segmentation, band detection, and quantification. At every step in this system, several techniques are implemented and their efficiencies are compared. The evaluation, using a large set of gel images, shows that the proposed system performs better than some existing techniques in detecting and analyzing bands in the gel electrophoresis images.

References

1. A. Machado, F. Campos, A. Siqueira, S. F. De carvalho, "An iterative algorithm for segmenting lanes in gel electrophoresis images" Proceedings of the 1997 10th Brazilian Symposium of Computer Graphic and Image Processing, SIBGRAPI 97, pp. 140 146, 1997.
2. Y. Xiangyun, C. Y. Suen, M. Cheriet, and E. Wang, "A recent development in image analysis of electrophoresis gels" Vision Interface, Canada, pp. 432 438, 1999.
3. A. Akbari, F. Albregtsen, "Automatic lane detection and separation in one dimensional gel images" Fourth International Conference on Bioinformatics of Genome Regulation and Struc ture, pp. 41 46, 2004.
4. P. S. U. Adiga and A. Bhomra, "Automatic analysis of agarose gel images," Bioinformatics, 17(11), pp. 1084 1090, 2001.
5. I. Bajla, I. Hollander, and K. Burg, "Improvement of electrophoretic gel image analysis," Measurement Science Review, 1(1), pp. 5 10, 2000.
6. C. Y. Lin, Y. T. Ching, and Y. L. Yang "Automatic method to compare lanes in gel electro phoresis images," IEEE Transactions on information technology in biomedicine, 11(2), pp. 179 189, 2007.
7. N. Kaabouch, R. R. Schultz, and B. Milavetz, "A novel automated analysis system for DNA gel electrophoresis images," Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition, pp. 36 41, July 2007.
8. Y. Labyed, N. Kaabouch, R. R. Schultz, B. Singh, "Gel electrophoresis image segmentation and band detection based on the derivative of the standard deviation," Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition, pp. 31 35, 2007.

Chapter 68
A Fuzzy Approach for Contrast Enhancement of Mammography Breast Images

Farhang Sahba and Anastasios Venetsanopoulos

Abstract This chapter presents a fuzzy-based method for contrast enhancement of mammography images. The selection of appropriate parameters for the required transformations is performed based on image-specific characteristics. The extraction of the breast border is the first step in this method. Images are then transformed to the fuzzy domain using a specific function. Next, an algorithm is applied for intensity adaptation where based on the amount of ambiguity, the proposed technique identifies the suitable form of modifications to enhance the image. Experimental results prove our method to be effective and hence of potential for use in computer-aided diagnosis systems.

Keywords Adaptive enhancement · Contrast enhancement · Fuzzy sets · Mammography images

68.1 Introduction

Using traditional radiology screening techniques, visually analyzing medical images is laborious, time consuming, and expensive. In addition, each individual scan is prone to interpretation error [1]. Mammography images are among the most difficult medical images to interpret since the features that indicate disease are typically very small. Also, they are difficult to analyze due to a wide variation in anatomical patterns. Furthermore, each individual scan is also prone to interpretation error. Therefore, there is a growing interest in incorporating automated techniques to analyze these images. As the first step for this task, image enhancement is essential for a reliable interpretation, as well as providing a faster

F. Sahba (✉)
Ryerson University, Toronto, ON, Canada
e mail: sahbafarhang@gmail.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 68,

diagnosis procedure. Over the past years, many attempts have been made to help radiologists in the detection of breast lesions [2 7]. In this chapter, we present a method to enhance the breast images. The selection of appropriate parameters for the required transformations is performed based on image-specific characteristics. One of the objectives of this work is to provide radiologists with a computer-aided diagnosis system aimed at studying the risk of developing breast cancer. Experimental results prove our method to have the potential to be used in computer-aided diagnosis systems.

68.2 Methodology

The proposed model is shown in Fig. 68.1, and the following subsections detail these algorithms.

68.2.1 Extraction of the Breast Region

The goal of this step is to exclude the image background and eliminate irrelevant data for further image analysis. Among many methods introduced in the literature, the approach presented in [8] is found to be more effective for histogram-based global thresholding of breast images. For locating of the final breast boundary, a locally weighted smoothing algorithm based on robust regression is proposed (Fig. 68.2). Details can be found in [9].

68.2.2 Adaptive Contrast Enhancement

68.2.2.1 Fuzzification

An image I of size $M \times N$ can be considered as an array of fuzzy singletons [10]. For the transformation of spatial space to the fuzzy property space, a membership function of the Gaussian type suggested in [11] is used:

$$\mu(g_{mn}) = \exp[-(g_{\text{Max}} - g_{mn})^2 / 2\upsilon^2], \tag{68.1}$$

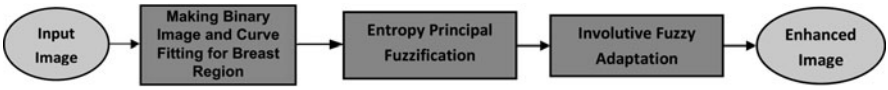


Fig. 68.1 Flow chart of the proposed method for mammogram image enhancement

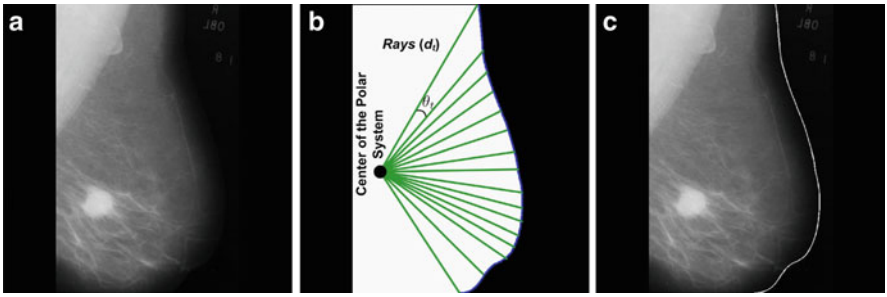


Fig. 68.2 (**a**) A sample breast image from MIAS database, (**b**) breast area with locating of the final breast boundary using smoothing algorithm, and (**c**) detected breast border

where g_{mn}, $\mu(g)$, and g_{Max} and are the intensity, single fuzzifier, and the maximum gray level present in the breast area, respectively. As can be seen, the values of membership lie in the range of $[a, 1]$ with $a = \exp[-g_{\text{Max}}^2/2\upsilon^2]$ [11]. By changing the value of υ, one can observe that higher values of υ result for a brighter image. According to information theory, a larger value of the entropy of a system indicates more information in that system. The entropy of our fuzzy membership set can be used as an important measure of information of the image and is calculated as follows [12, 13]:

$$E_\mu(I) = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} S(\mu(g_{mn})), \tag{68.2}$$

where S is Shannon function [12, 13]. The selection of parameter υ is based on the maximum fuzzy entropy principle so that we have the maximum available information represented by the membership (as suggested in [12, 13]). Therefore, the value of optimum υ_{opt} is determined such that

$$E_{\max}(I, \upsilon_{\text{opt}}) = \max\{E(I; \upsilon) | \upsilon > 0\} \tag{68.3}$$

After parameter υ_{opt} is determined, the image I can be fuzzified to generate membership values. It is interesting to note that it is better to calculate the parameter in some small sub-images in the area of breast (e.g., with size of 128 $\times$ 128) and use it for the entire image. As an implementation issue, it can increase the speed without losing the information.

68.2.2.2 Involutive Fuzzy Complements and Membership Enhancement

Intensity adaptation is performed based on involutive fuzzy complements and a measure of fuzziness as follows.

Involutive Membership Function

The complement of the membership function$\mu(g)$ is defined as $\mu(g) = 1 - \mu(g)$. Sugeno introduced a class of involutive fuzzy complements called λ-complements [14]:

$$\mu_\lambda(g) = \frac{1 - \mu(g)}{1 + \lambda\mu(g)}, \tag{68.4}$$

Where λ belongs to the interval $(-1, \infty)$. These classes of fuzzy complements can intuitively define a new class of membership functions as introduced in [15]:

$$\mu^*_\lambda(g) = 1 - \mu(g) = \frac{\mu(g) \cdot (1 + \lambda)}{1 + \lambda\mu(g)}, \tag{68.5}$$

where $\mu(g)$ is the initial membership value of g. For instance, if we define the membership as a monotone function for brightness, then we can generate images with different levels of brightness if we vary λ in this equation. According to this definition, the result would be absolutely dark (black) for $\lambda \to -1$ and absolutely bright (white) for$\lambda \to \infty$ [15]. For the extreme values of λ, the amount of ambiguity in the image property is very low because in both cases there exist a lot of pixels with very low and very high membership values. This leads to the fact that the images obtained from the middle values are more suitable for human perception. As a result, one can expect that one of the middle images (obtained from the middle values of λ) is the best one. According to this fact, we can expect a high fuzziness from middle images and apply parameter-dependent membership functions to detect one of them with maximal fuzziness [5, 15].

Intensity Adaptation Using the Enhancement of Membership Values

The grayness ambiguity (or fuzziness) of intensity values can be used as a quality measure for enhancement procedure [5, 15]. The index of fuzziness is defined as

$$\gamma = \frac{4}{MN} \sum_{g=0}^{L} {}^{1} h(g) \cdot \min(\mu^*_\lambda, 1 - \mu^*_\lambda), \tag{68.6}$$

where $h(g)$ is the frequency of g in the image. In fact, it considers the intersection of a set and its complement. To calculate the derivation regarding to λ and find the extreme values, a new measure of fuzziness using the algebraic product instead of minimum operator is introduced in [15]:

$$\gamma = \frac{4}{MN} \sum_{g=0}^{L} {}^{1} h(g) \cdot \mu^*_\lambda \cdot [1 - \mu^*_\lambda] \tag{68.7}$$

Then the point corresponding to the optimal value can be calculated using the following equation:

$$\gamma_{\max} = \gamma\{\upsilon = \upsilon_{\text{opt}}, \lambda = \lambda_{\text{opt}}\} \tag{68.8}$$

Therefore a new membership value for each pixel can be calculated as follows:

$$\mu^*_{\text{new}}(g) = \frac{\mu_{\upsilon_{\text{opt}}}(g).(1 + \lambda_{\text{opt}})}{1 + \lambda_{\text{opt}}\mu_{\upsilon_{\text{opt}}}(g)} \tag{68.9}$$

Involutive membership function generates images with different brightness levels. As mentioned, the images with higher ambiguity values seem to be more suitable for human perception. This method can greatly enhance the contrast and avoids common issues such as under-enhancement and over-enhancement.

68.3 Results

We evaluated 95 mammography images from two data sets to verify the algorithm. Figure 68.3 shows two original images and the results of applying the proposed approach. For comparison, the results of histogram stretching and histogram equalization are also demonstrated. As it can be seen, the proposed method is able to deliver good results. In fact, it can enhance the contrast without changing the image texture (as an important feature in such images). This figure also shows a close view for two different lesions, to demonstrate how well the proposed method performs. For a quantitative measure, we used the criterion presented in [16]. This criterion is calculated according to the Target-to-Background Contrast measurement (TBC) based on standard deviation. This measure indicates the difference between background and target mean gray level, where the homogeneity of the object is considered for a better visualization. This measure is obtained using the ratio of the standard deviation of the grayscales within the target before and after the enhancement as [16]:

$$\text{TBC} = \frac{(\mu_T^E/\mu_B^E) - (\mu_T^O/\mu_B^O)}{(\sigma_T^E/\sigma_T^O)} \tag{68.10}$$

where O and E specify the Original region of interest and the Enhanced region of interest, respectively. Also, T and B refer to the Target (object) and Background, respectively. In 90% of cases, the criterion (TBC) for the proposed method was maximum compared to that in the other methods, histogram stretching and histogram equalization, as well as the original image itself. To show the effectiveness of the algorithm, we applied a method based on a simple level set segmentation on a part of mammogram image containing a lesion in both original and enhanced

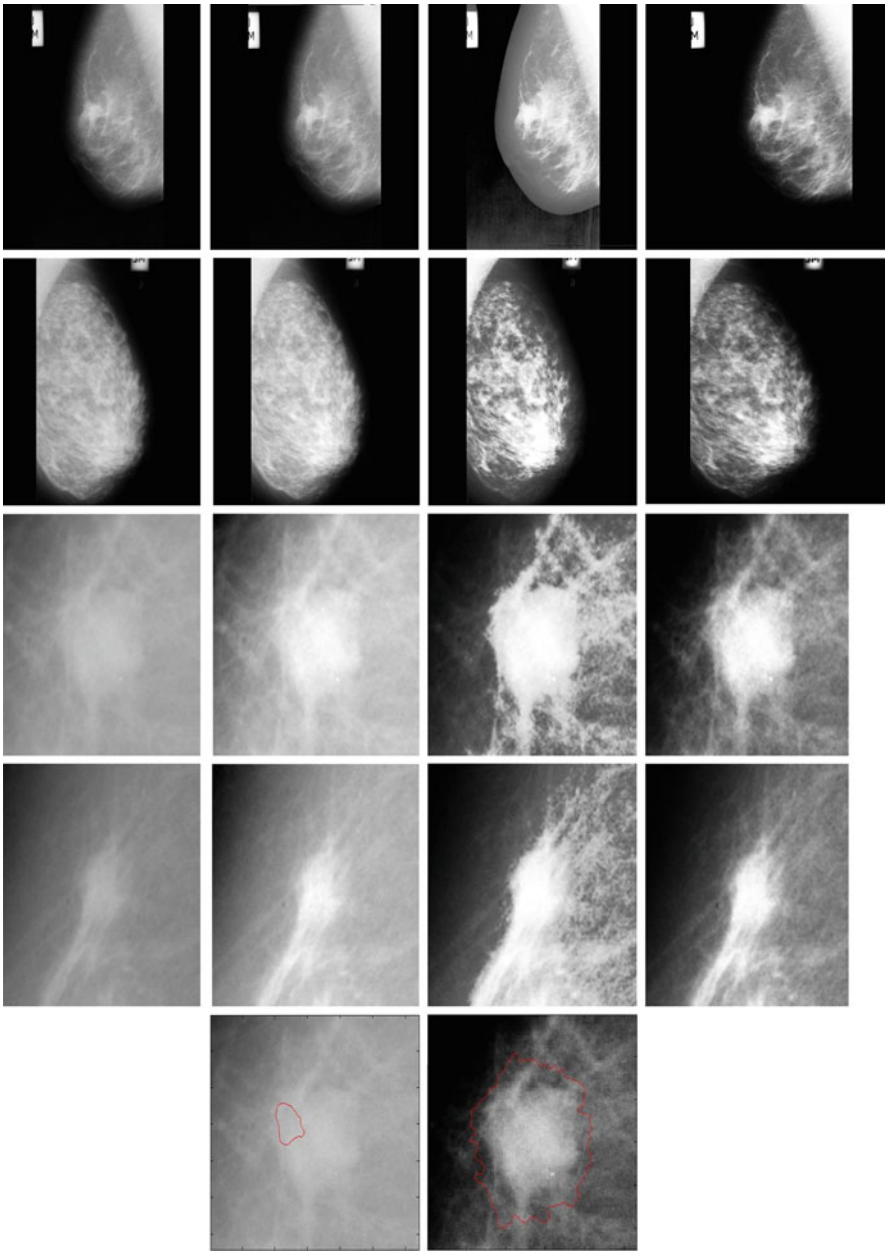


Fig. 68.3 *First and second rows*: The original breast images; *Third and forth rows*: Region of interest; *First column*: Original image, *Second column*: Histogram stretching, *Third column*: Histogram equalization, *Forth column*: Proposed approach. *Fifth row*: Applying the level set segmentation on the region of interest for the original image (*left*) and enhanced image (*right*)

versions. The results are again shown in Fig. 68.3. This sample shows how much better this segmentation method can work on the enhanced image. By applying such segmentation method, we can examine the interior of the lesion for texture or other analysis for any CAD system.

68.4 Conclusion

We have presented an approach for the contrast enhancement of mammography images. In this method, the breast border is first detected to eliminate irrelevant data. Then, using an entropy-based function, the image is transformed to the fuzzy space where an algorithm for intensity adaptation is applied. The algorithm detects the suitable form of the modification as well as appropriate parameter selection based on the specific image characteristics. The experimental results support the idea of using the proposed approach for a CAD system.

For the future of this research and as an improvement of the studied method, we will consider the following works:

Applying the method to a larger data set;
Applying a method to better optimize the parameters used in this method;
Using other criteria for evaluation of the results;
Applying a powerful segmentation method based on shape attributes to capture all of the extensions of the lesion.

Reference

1. Joseph K. T. Lee, MD.: Quality a radiology imperative: interpretation accuracy and pertinence, American College of Radiology, 4, pp. 162 165, 2007.
2. Cheng H. D., Shi X. J., Min R., Hu L. M., Cai X. P., Du H. N.: Approaches for automated detection and classification of masses in mammograms, Pattern Recognition, 39(4), pp. 646 668, 2006.
3. Guliato D., Rangayyan R. M., Carvalho J. D., Santiago S. A.: Polygonal modeling of contours of breast tumors with the preservation of spicules, IEEE Transactions on Biomedical Engineering, 55(1), pp. 14 20, 2008.
4. Rangayyan R. M., Ayres F. J., Desautels J. E. L.: A review of computer aided diagnosis of breast cancer: Toward the detection of early signs, Journal of the Franklin Institute, 344(3 4), pp. 312 348, 2007.
5. Sahba F., Venetsanopoulos A.: Contrast enhancement of mammography images using a fuzzy approach. IEEE Eng. in Medicine and Biology Conference, 2008, pp. 2201 2204, 2008.
6. Sahiner B., Petrick N., Chan H. P., Hadjiiski L. M., Paramagul C., Helvie M. A., Gurcan M. N.: Computer aided characterization of mammographic masses: accuracy of mass segmentation and its effects on characterization, IEEE Transactions on Medical Imaging, 20(12), pp. 1275 1284, 2001.

7. Tang J., Rangayyan R. M., Xu J., El Naqa I., Yang Y.: Computer aided detection and diagnosis of breast cancer with mammography: recent advances, IEEE Transaction on Infor mation Technology in biomedicine, 13(2), 236 251, 2009.
8. Ojala T., Nappi J., Nevalainen O.: Accurate segmentation of the breast region from digitized mammograms, Computerized Medical Imaging and Graphics, 25(1), pp. 47 59, 2001.
9. Sahba F., Venetsanopoulos A.: A New Fuzzy Approach to Mammographic Breast Mass Segmentation, International Conference on Image Processing, Computer Vision, and Pattern Recognition, pp. 852 858, 2009.
10. Zadeh, L. A.: Fuzzy sets, Information and Control, vol. 8, pp. 338 353, 1965.
11. Hanmandlu M., Jha D., Sharma R.: Color image enhancement by fuzzy intensification, Pattern Recognition Letters, 24, pp. 81 87, 2003.
12. Cheng H. D., Wang J., Shi X.: Microcalcification detection using fuzzy logic and scale space approaches, Pattern Recognition, 37(2), pp. 363 375, 2004.
13. Cheng H. D., Chen J. R., Li J.: Threshold selection based on fuzzy c partition entropy approach, Pattern Recognition, 31(7), pp. 857 870, 1998.
14. Sugeno M.: Theory of fuzzy integrals and its applications. Dissertation, Tokyo Institute of Technology, Japan, 1974.
15. Tizhoosh H. R., Krell G., Michaelis B.: Lambda Enhancement: Contrast adaptation based on Optimization of Image Fuzziness, FUZZ IEEE'98, pp.1548 1553.
16. Singh S., Bovis K.: An evaluation of contrast enhancement techniques for mammographic breast masses, IEEE Transaction on Information Technology in Biomedicine, 9(1), pp. 109 119, 2005.

Chapter 69
Computational Modeling of a New Thrombectomy Device for the Extraction of Blood Clots

G. Romero, I. Higuera, M.L. Martinez, G. Pearce, N. Perkinson, C. Roffe, and J. Wong

Abstract Thrombectomy devices have been developed as an alternative means for clot removal. A number of devices using a variety of methods to remove the clot are now available. This chapter covers the analysis and research into a device recently developed in the UK, called a "GP" thrombus aspiration device (TAD). Presented in this work is the development of a model of this device, as well as its simulation and interpretation of the results obtained with the potential for helping in optimizing its operation for future use. The simulation model that is presented can be used in showing the potential performance of the "GP" TAD device under different conditions of blood flow and size of blood clot, obtaining the minimum pressure necessary to extract the clot and to check that both this pressure and the time required to complete the operation are reasonable for potential use in clinical situations patients, and are in line with experimentally obtained data.

Keywords Biomedical engineering · Tools and methods for computational biology

69.1 Introduction

Cerebrovascular infarction, or stroke, occurs as a result of ischemic or hemorrhagic vascular disease. Ischemic stroke has the potential for damage in the penumbral area and in the core, and treatment aims to remove the clot. Stroke is a major cause of morbidity and mortality globally. In the UK alone, there are 130,000 strokes each year [1]. Approximately, 85% of strokes are caused by a blood clot.

G. Romero (✉)
Universidad Politecnica de Madrid, C. Jose Gutierrez Abascal 2, 28006 Madrid, Spain
e mail: gregorio.romero@upm.es

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 69,

During the last decade, mechanical thrombectomy devices (MTDs) have become more widely used. Thrombectomy devices have been developed as an alternative means for clot removal. A number of devices using a variety of methods to remove the clot are now available. These include the MERCI clot retriever and, more recently, the penumbral device; other types of devices include angiography catheters, rheolytic catheters (Angiojet), Basket-style devices, and microsnaring devices. Thrombectomy may be associated with risks, such as breakage of moving parts, penetration of the arterial wall, and downstream embolization caused by clot dislodgment [2]. Studies suggest that mechanical embolectomy is most effective in large volume proximal occlusions [3]. Other interventional surgical treatments include endarterectomy which involves surgically removing clots in the carotid arteries. This treatment has proved successful [4] but carries a risk of the clot becoming dislodged during the procedure.

The need to study new medical devices like the one described here means that computer premodeling may have the potential to help in the optimization and fine-tuning of such devices.

69.2 "GP" TAD Device

The "GP" thrombus aspiration device (TAD) device (Fig. 69.1) [5] consists of a pump that provides the necessary suction pressure for the operation, joined to a very long catheter; the "GP" TAD is located at the end of this catheter. The proposed procedure for using this device would be to introduce it into an artery in close proximity to the occluding blood clot, and position it at a distance of approximately 3 mm from it. Then the suction would begin until the clot is extracted. The clot would cross the 3 mm that separates it from the "GP" TAD and clot capture would occur and the device would be removed from the body.

It is currently being developed as a potential TAD through a series of in vitro studies. This device has the potential to be used in relatively small arteries. It has no moving parts and therefore should reduce the risk of breakage in a vessel. Since it does not touch the clot itself, it has the potential to also reduce the risk of clot disruption and downstream embolization. The internal surface has been mathematically designed. It is also associated with low forces at the periphery of the device which may therefore reduce the risk of arterial collapse during aspiration of the clot [6,7].

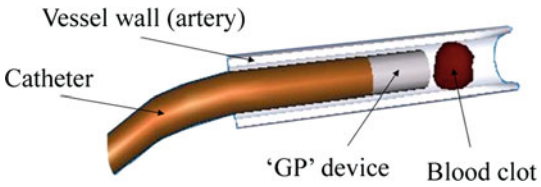


Fig. 69.1 GP thrombus aspiration device

69.3 Modeling the "GP" TAD Device

The method chosen for the representation and simulation of this model is the Bond Graph technique [8], which allows assimilating the model to a scheme made up of resistances, capacitances, and inductances.

The pump can be represented by a variable pressure source (Fig. 69.2) whose value will increase from zero to a nondetermined value (0 [30, 60] kPa) suitable for carrying out the extraction and it will be obtained from the optimization of the developed model. The time taken to reach the maximum value of depressure has been obtained from experience and must be about 3 s, after which time the depressure provided by the pump remains constant.

After the pump is the catheter, a 110-cm long 1-mm diameter hollow cylindrical tube is joined to the "GP" cylinder of the same diameter and a length of 20 mm. To represent both elements, they are considered as several pipe sections bearing in mind the different phenomena that take place in their interior: load and inertia loss, and fluid compressibility:

$$R = \frac{128\eta L}{\pi D^4}, \tag{69.1}$$

$$I = \frac{\rho L}{\pi (D/2)^2}, \tag{69.2}$$

$$K = 4B/\pi D^2 L. \tag{69.3}$$

Linear load loss (R) is due to the friction between the liquid particles and the pipe walls. Due to their being straight pipes, only linear load losses are taken into account. It can be represented by a resistance and if we assume that the blood flow is laminar due to the Reynolds number being approximately 1,000, the equation that governs its behavior can be determined by (69.1).

Secondly, the flow inertia (I) to be overcome in its movement is taken into account, and considering a section with circular geometry it can be modeled with (69.2).

Lastly, the blood compressibility (K) must be included due to it acts as a spring producing a decrease in volume when the pressure required for compression is

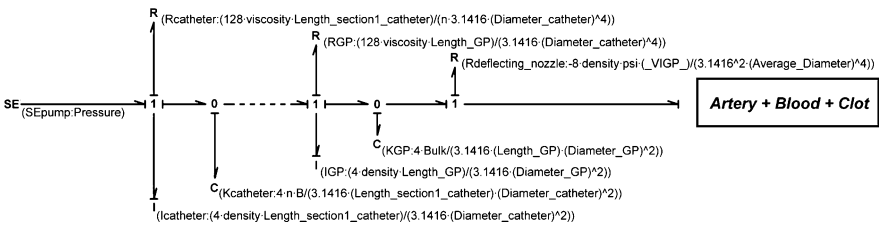


Fig. 69.2 Modeling the catheter and "GP" device components by Bond Graph technique

increased; this behavior is dependent on Bulk's blood coefficient (2.2×10^9 N/m) and it can be defined as a capacitance of K value shown in (69.3). In the previous expressions, η is the dynamic viscosity of the blood flow (0.0035 Pa s), ρ is the blood density (1,060 kg/m^3), L is the length of the pipe section, and D is its diameter.

In the model to simulate, due to the great length of the catheter, it must be partitioned in identical 10 sections and it can be represented by 10 submodels that include the three previously described phenomena. Thanks to this representation, it is possible to study the evolution of the pressure loss along the catheter.

Later, the "GP" device must be positioned and it can be represented by the same three previous phenomena (R, I, K) and with the corresponding values.

In addition, due to the artery being located at the end of the "GP" device, it is necessary to consider the transition between both elements as a secondary load loss caused by the difference in diameter of the "GP" device and the artery, respectively, and the subsequent variations in flow. These load losses can be represented as a resistance and can be calculated with the following expression:

$$R = 8\rho\xi\frac{Q}{\pi^2 D^4}, \tag{69.4}$$

where Q is the flow which circulates in the section between the end of the "GP" device and the artery (represented in Fig. 69.2 by " VIGP "), D is the mean diameter between the cylinder and the artery (2 mm), and ξ is the load loss coefficient. The load loss coefficient ξ is a dimensionless parameter that quantifies the loss produced and depends on the geometry of the junction; since this is a narrowing, this value is 0.4.

The artery located between the end of the "GP" device and the clot can be included in the model as another section of a pipe with a length of 3 mm, similar to the catheter and the "GP" device and it must be defined by the loss of linear load (R), the inertia (I), and the compressibility of the blood (K).

In addition, it is necessary to insert a parameter that represents the compressibility of the artery (see (69.5)), where E is its Young's modulus (2.8×10^9 N/m), h is the thickness of the artery (0.1 mm), V_0 is the artery initial volume, and r_0 is the artery initial radius (1.5 mm):

$$K = \frac{Eh}{V_0 2r_0}. \tag{69.5}$$

Once all the elements are defined by fluid mechanics, it is necessary to change from the domain of hydraulics to mechanics, to be able to evaluate the movements and efforts in the clot, as well as to define the physical friction between the clot and the artery. This domain change is carried out by a *Transformer* (TF) element and it should have the value of the inverse of the artery's area.

Accurately defining the clot model in order to model, it is the most complex part of the model. A clot is a cylindrically shaped element of 3 5 cm long, and of a mass

that falls between 0.5 and 1 g. This element has usually formed in another location, usually in a vascular artery, and has become dislodged remaining trapped in smaller diameter arteries, such as the cerebral artery considered in this work. Therefore, due to this difference in diameter, the clot becomes attached to the wall by a force that needs to be overcome in order to begin its movement for removal. In addition, the relative movement between the clot and the artery presents static and dynamic friction, which needs to be taken into account. If a correct approximation to reality is to be achieved, all these phenomena, as well as the circumstances restricting clot movement, need to be considered. Experimental data indicate that the clot begins its movement when this force is equal to 0.01 N.

In order to decide when to begin the movement of the obstructive element, firstly we need to insert a spring (K_{eq}) into the model to measure the force supported by the beginning of the clot and its deformation. Therefore, only when the spring undergoes this displacement should the clot be allowed to move; to the contrary, it would be prevented.

Secondly, it is necessary to insert the resistance that represents the friction between the clot and the arterial wall. The value of this parameter must be variable depending on whether the clot has not begun its movement (static friction) or if it is already in movement (dynamic friction), so that when the clot begins to move the friction value will drop considerably. This value is obtained starting from the Stokes equation and can be given a value of 2.5×10^{6} N s/m for the static friction and an order of magnitude lower than for the dynamic friction.

In addition to the spring and the resistance inserted, the model must have an element Inertance that represents the mass of the clot (1 g).

In Fig. 69.4, to ensure that the clot remains at rest while the force existing at its beginning is less than 0.01 N, a spring damper system joined to a wall (zero flow source) must be used. In this system, while the clot does not receive the force of minimum suction, it will remain attached to the wall, thereby preventing any movement. On the contrary, if the force exceeds this figure, the model will cause the bond imposed by the spring damper system to be eliminated with the clot becoming free and moving in accordance with the suction pressure acting on it from the system, letting it be removed.

To obtain the simulation of the developed model (Figs. 69.2 and 69.3), *Bondin* © software will be used.

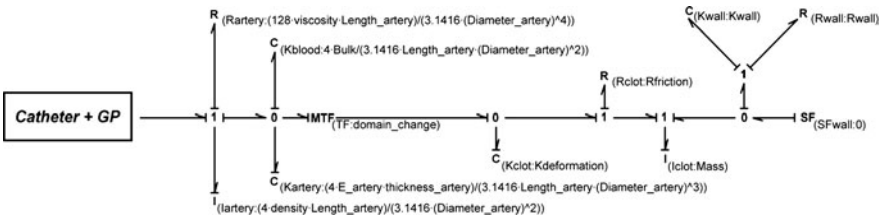


Fig. 69.3 Modeling artery, blood, and clot components by Bond Graph technique

69.4 Results

The fundamental object of this study consists in determining and optimizing the minimum pressure required for the extraction of a blood clot depending on the clot size. To do this, by varying the values of the pressure source and clot length, the movement of the clot and the time required for its extraction are measured, thereby obtaining the optimum minimum pressure.

The results of these parameters are evaluated for the sizes of the clots of 5 and 3 cm. For each pressure and size of the clot, the necessary times to reach the necessary force at the end of the clot will be obtained so that its movement begins, and is picked up, as well as the time that the extraction operation lasts.

In the following figure, the times to begin clot movement obtained for both sizes are shown. The final time to end the extraction is about 1.0 1.5% more time.

It is observed that for a longer length of clot, more time is needed with the same pressure to reach the necessary force so that the movement begins and for the subsequent extraction. Also, for the same size of the obstructive element, when the suction pressure increases, the time needed to complete the operation diminishes. These results are coherent with what would need to be obtained.

It is necessary to highlight that for a size of 5 cm long and a pressure of 30 kPa, no time value is attached. No movement of the clot can be appreciated for any time when using this suction pressure in the simulation due to the fact that pressure is insufficient to be able to create a force of 0.01 N before the obstructive element; therefore, it remains indefinitely at rest. It can be appreciated that for a length of 5 cm, the necessary time for the extraction is in the range of 60 120 s. We could use any pressure but 40 kPa is enough. However, for a length of 3 cm, it is observed that the greater the pressure, the faster the clot moves. So with a pressure of between 30 and 40 kPa, the clot can be extracted.

Analyzing what happened into the "GP" device, the loss of pressure appears mainly in the catheter part due to the length of this component and also how the loss of pressure is the same in each division. After this loss of pressure, the pressure is

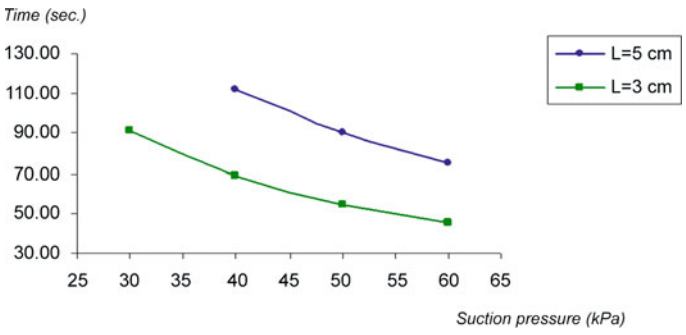


Fig. 69.4 Time to begin clot movement with different sizes and suction pressures

99% stabilized in the "GP" device (the loss of pressure is lower than in the catheter). Finally, a small percentage drop of the depressure pump removes the clot.

69.5 Conclusions

In vitro experimental data indicate that the extraction of the blood clot takes place in an interval of time of 60 120 s. In Fig. 69.4, it can be seen that the simulation produces the extraction in 92 or 114 s, a figure that is coherent with experience. The developed model indicates that the minimum value of pressure for the extraction of a 5-cm long 1-g clot is 40 kPa, while a pressure of 30 kPa is enough for the suction of a 3 cm length.

On the other hand, studies demonstrate that the artery has a resistance of 750 mmHg, which is equivalent to 100 kPa. This has been analyzed in the model and the pressure in this area rises to 1.41 kPa, so there is a very wide danger margin before rupture of the artery might occur.

Finally, it is shown that an important pressure loss takes place in the catheter joining the pump to the "GP" device. These values obtained can be used to optimize its geometry.

This work is a first attempt to model the operation of the "GP" TAD device, with the potential for helping in optimizing it for its future applicability and compatibility with the experimental data obtained. The subsequent lines of work could focus on developing a highly accurate model including studying the rechanneling of the blood flow after clot removal.

References

1. Stroke association website: http://www.stroke.org.uk/document.rm?id 330. Last accessed 30 December 2008.
2. Broderick JP (2008). Endovascular therapy for acute ischemic stroke. Stroke 40:103 106.
3. Thomassen L, Bakke SJ (2007). Endovascular reperfusion therapy in acute ischaemic stroke. Acta Neurol Scand Suppl 187:22 29.
4. Rothwell PM, Eliasziw M, Gutnikov SA, Warlow CP, Barnett HJM (2004). Endarterectomy for symptomatic carotid stenosis in relation to clinical subgroups and timing of surgery. Lancet 363:915 924.
5. Pearce G, Perkinson ND (2006). "Biomechanical Probe". International Patent Corporate Treatise (WO2006120464) published 2006 11 16; European patent (EP1893195 (A2)) pub lished 2008 03 05; Japanese patent (JP2008639924 (T)) published 2008 11 20; Chinese patent (CN101208049 (A)) published 2008 06 25.
6. Pearce G, Patrick JH, Perkinson ND (2007). A new device for the treatment of thromboembolic strokes. J Stroke Cerebrovasc Dis 16(4):167 172.
7. Pearce G, Jaegle F, Gwatkin L, Wong J, Perkinson ND, Spence J (2009). An investigation of fluid flow through a modified design for the 'GP' device. 11th International Conference on Computer Modelling and Simulation, Cambridge, UK, pp. 191 195.
8. Karnopp DC, Margolis DL, Rosemberg RC (1990). System dynamics: A unified approach, 2nd edition. Wiley, New York.

Chapter 70
NEURONSESSIONS: A Web-Based Collaborative Tool to Create Brain Computational Models

Ana Porto, Guillermo Rodríguez, Julián Dorado, and Alejandro Pazos

Abstract We have developed a collaborative web tool for computational biology by using open-source technologies. It allows the cooperative construction of computational models with NEURON. NEURON is a powerful local environment for modeling and simulating the nervous system. Our web tool facilitates researchers who are located far apart to build computational models of the brain, and share knowledge and opinions. The portal integrates all the necessary tools in just one. It allows the creation and participation in work sessions with NEURON, and synchronous and asynchronous file sharing. Moreover, it allows the analysis of the changes introduced in the models by the users, by means of a version control system, as well as real-time comments about each step in the development of each model. It only uses an Internet browser and minimum bandwidth consumption, thanks to the simplified data exchange process. In this paper, we present the tool NEURONSESSIONS, whose cooperative sessions also allow a virtual community to emerge for advancing in Neuroscience.

Keywords Brain computational models · Computational neuroscience · Collaborative work · Cooperative sessions

70.1 Brain Computational Models

A computational model for experimentation and simulation is understood as the representation of a real system for its study in a computer. The computational

A. Porto
Department of Information and Communications Technologies, University of A Coruna Campus de Elvina s/n, 15071 A Coruna, Spain
e mail: anuska@udc.es

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 70,

modeling process in Neuroscience is characterized by their extreme complexity. The models depend on thousands of parameters and hundreds of simulations.

The brain modeling process typically requires the collaboration of a huge variety of experts in different fields: neuroscientist, physicians, computer scientists, etc. They are usually located at universities and research centers geographically far away. Therefore, we developed a single and user-friendly tool to build cooperatively computational models through Internet. Everyone has to be understood by everyone and the existence of an exhaustive, fast, simple, and automatic tool for the cooperative creation of the model would be extremely useful.

The development of a brain computational model is usually tackled locally in a computer. Besides, it must be in line with the performance of electrophysiological experiments, usually in vitro. They are essential in order to study how brain cells or cell networks contribute to the functions of the various brain regions. Computer scientists, physicians, and neuroscientists are forced to travel, hold numerous meetings, and communicate via e-mail or telephone to work through the model. This is the case of the work carried out by our computing research team together with neurophysiologists located at the other side of our country. We create models with NEURON [7], a simulation environment that is used for building and using computational models of neurons and networks of neurons. Given the distance between our teams, we needed to bring together all the advantages of using instant messaging, e-mail, or telephone when we could not meet. Due to the high complexity of the simulations made, it was necessary for all the researchers to be able to check the model in real time, to analyze the generated data files and graphics, to observe previous versions, to see the changes introduced by each researcher, etc.

It should be noted that there are different local and individual brain modeling tools yielding good results, such as, e.g., NEURON [7] or GENESIS [4]. NEURON was chosen for two main reasons: on the one hand, it is the modeling environment used by our computer science research team when collaborating with the researchers of the Ramón y Cajal Institute of Neurobiology of the Higher Council of Scientific Research (CSIC) in Madrid [2, 9]. On the other, it is one of the most widely used tools by the best-known neuroscientists, thanks to its versatility and efficacy for studying the nervous system. Therefore, our new tool emerges from the comprehension of the characteristics of interdependent research teams.

Although there are very interesting "collaboratory" projects [3, 8], we have not found a collaborative web-based tool for building brain computational models. A web software tool was developed using free open-source technologies. The cooperative sessions to be established with this tool from now on allow the sharing of information and knowledge among scientists located far apart, as well as their interaction when developing a computational model. They also enable a rigorous monitoring of the changes made in the model files, thus participating in their development.

70.2 NEURONSESSIONS

We have oriented our tool to the use the NEURON environment. NEURON does not have a collaborative interface. However, the good part is that it works with small size text files. It is not necessary to have a huge bandwidth in the network, since small zipped text files are sent. Therefore, we decided to create a Web portal to manage those text files, integrating everything that is needed to facilitate work and communication.

We have intended to facilitate the work of NEURON computational model developers. This has given rise to a modular tool with a wholly reusable code. It facilitates to a great extent the integration of new functions and their application to multiple environments.

Our tool does not require additional software in the client computer and is oriented to modelization because it permits to make an on-line and off-line monitoring to know who has done every change in a model and why. Our tool permits the users to store complete computational models, to share the automatic model's history file, and to share concepts and ideas on real time.

It does not depend on a software license, is multiplatform, and only requires an Internet browser to be used. It is always operative. Although new versions emerge, this will not impact the user's computer. It allows registration in the application of researchers from different centers. They should enter their background details and interests in Neuroscience models, thus enhancing relationships among scientists working on similar topics. It allows the creation of different collaborative work sessions. The sessions are monitored, so that the cooperative work is carried out in an organized manner and every user has the chance to lead the process of building a model.

NEURONSESSIONS integrates real-time communication among session participants by means of instant messaging. It also allows the download, test, and storage of models made with NEURON. A MySQL [6] database stores the versions of the models made during the various work sessions, together with detailed information about how and why different versions emerge. It integrates an automatic version control. The fact of having information centralized in the server enables the changes made by any user to be automatically available to every user on-line.

Moreover, and regardless of the date on which a model was made, it is possible to know its author, who took part in its creation, and to access its different versions and the reports about how it has been made and modified. The said reports contain crucial and necessary information for the development of neurobiological computational models, given that they enable a thorough and detailed step-by-step comprehension.

It has been implemented with JAVA technology in an Ubuntu Linux 7.10 personal computer, using the J2EE platform [5] (Java 2 Enterprise Edition) and Jakarta Struts [10], applying the MVC [1] (Model View Controller) architectural pattern. The portal is running in a Tomcat [11] applications server and accesses a relational MySQL database.

70.3 Creating Brain Models Through Internet

Each researcher will carry out the work with NEURON locally in his computer. Users can create their models with the NEURON environment by incorporating their parameters through text files with code and data. They can also do it by means of menus and windows where values can be established and model elements can be designed.

The contribution of our tool lies in a personal identification that allows an authorized researcher to create or access securely a session showing a shared environment. This environment allows the treatment of the aspects of the model made with NEURON that might be dubious or significant. The model authors who are going to create a session with our tool must provide a zipped file (*.zip, *.rar, *.tar or *.gz) containing all the files integrating the model, together with a "readme.txt" file containing instructions on how to run it.

The cooperative work sessions are established following an elaborate protocol with regard to the control of the role played by each user in the creation and modification of the models. We have decided that only one researcher will lead one session at any given time, while many active sessions may coexist simultaneously in the system. Anyhow, a user can only be in one session, i.e., a user cannot take part in several sessions at the same time, even though they log on to the web application through different computers.

The user established as leader will have control over the session. Therefore, all his or her actions will be immediately seen by every participant in that session. The system users can browse at any point in time every active session and every participant in them. All of the information required to establish the sessions shall be saved in a centralized database, which will be constantly updated in order to ensure that the information is consistent. That database will also store the actions performed by each researcher. These actions constitute the history of the session, which may be browsed by any researcher in order to understand the steps and changes involved in building the model elaborated in that session.

70.3.1 Researcher Actions

Any researcher intending to use our tool has to log in and be accepted by the webmaster. The researchers must log in so as to browse models, create sessions, or take part in already active ones.

The functions related to the registered users are as follows:

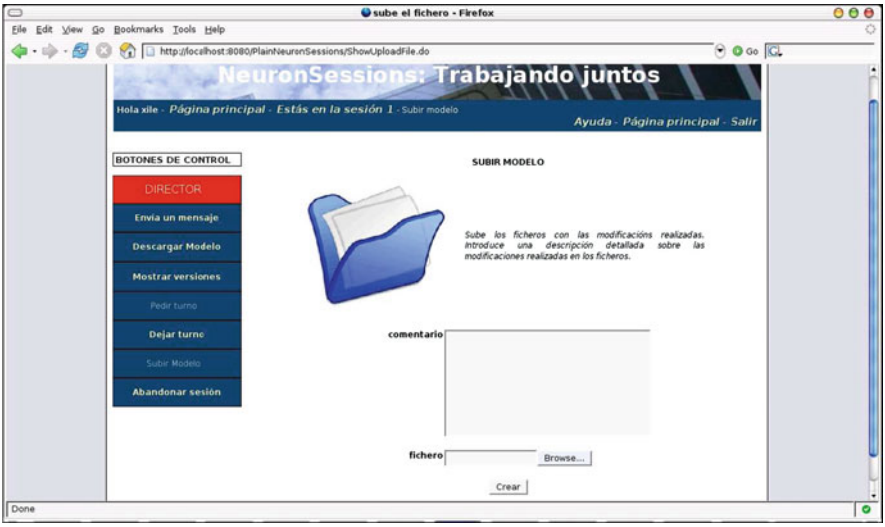


Fig. 70.1 Browsing for the model files to create a cooperative session

- Browsing active sessions and models previously saved in the database, together with all of their associated details (associated comment, creation date, different versions, etc.).
- Creating a session from the files of a model used to work collaboratively (Fig. 70.1).
- Logging on to a session to start working with the participants.
- Browsing a previously saved model, constituting a new session from it or simply downloading it in order to work individually, without sharing.
- Browsing versions and history of changes in sessions and stored models, as well as their modification reports.
- Browsing help in order to use the application and to get to know its options or possibilities.

Besides, once inside a session, the user can do the following:

- Download the latest version of the files integrating a model, which are zipped and previously saved in the database by the session leader.
- Browse versions of models in active session, as well as reports on the changes made to the session files.
- Take turns being the session leader and upload to the system a zipped file with changes in the files of the model under development in that session. When the user applies to be the leader, he or she joins a queue of applications and is assigned an order number, depending on when he or she has applied to take turns (priority is FIFO, First In First Out).

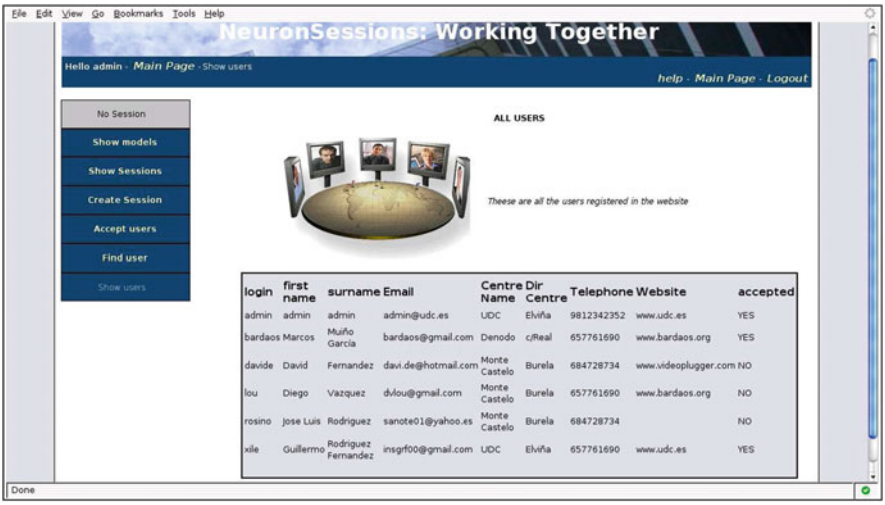


Fig. 70.2 Registered users' list. *Left menu*: administrator actions

- Leave the shift in case he or she is the leader and has completed the upload of his or her version of the model.
- Read and send messages so as to communicate in real time with the participants logged on to a session, and read the control messages sent by the system about the events happening in the current session.
- The system closes inactive sessions after 5 min.

The Web application provides special options for the administrator (Fig. 70.2). These options are "accept users", "show registered users", and "find user".

70.3.2 Test Performed

NEURONSESSIONS was evaluated in two separate set of tests, local and remote. Local tests were carried out in our Computing laboratory by creating ten sessions at the same time from several computers. The purpose of local test was to evaluate the validity of version control; the efficiency of file transfer and instant messaging; and the capacity of the system to support concurrent sessions. Remote tests were carried out in a real scenario with four pairs of neuroscientists located in different cities. The purpose of remote test was to check the tool efficiency regardless of the network's bandwidth and, especially, the reliability and usability for neuroscientists.

The results from local tests showed that the average of file transfer time was 4 seconds (s), depending on zip file size. The instant messaging average time was

3 s. The results from remote tests showed that transfer speed changed only an average of 2 s with regard to the local tests. At the end of remote tests, neuroscientists filled out a questionnaire to show their satisfaction. They provided helpful comments for interface design, such as the introduction of "breadcrumbs" in the web pages for controlling their position, and the establishment of a small status area to alert when the current user is in session (and the session number) and when the user is currently the session leader.

The tests were made with Internet Explorer (v7 and 8) and Mozilla Firefox (v2 and 3) browsers and from computers with Ubuntu Linux and Windows XP operative systems.

70.4 Conclusion

We have presented NEURONSESSIONS, a new tool for computational neuroscience emerging from the comprehension of our interdependent research team, composed of neuroscientists, physicians, and computer scientists situated geographically far away. Having understood their characteristics, we have designed and implemented a multiplatform, modular, fast, and very simple web-based system that allows to build and share computational models. This is a unique environment that is extremely useful for the remote development of brain computational models with NEURON through Internet.

This web-based collaborative tool is very useful for research groups that build brain computational models all over the world, models where every change introduced and the reason for it are crucial in order to understand them. It is not only useful for expert researchers but is also very appropriate for the learning of students and novel researchers. They will be able to participate in cooperative sessions and in emergent social networks.

This system unleashes organizational effects for saving time in the research process, saving money by avoiding journeys, and increasing the quality of the resultant models because the researchers can participate: free, through the Internet, and without any additional software. This environment can be easily extended to other areas because it enables the launch of very well-organized sessions, providing a centralized access to all the information required. It is a free-access meeting point for working together in a synchronous or asynchronous manner.

70.5 Future Trends

We have showed the initial application of our tool to brain computational models with NEURON. But we are testing this tool to build models with other simulators and in other scientific areas. For example, to develop Neural Networks in Artificial Intelligence.

Acknowledgments This work was partially supported by Grants from Spanish Ministerio de Ciencia e Innovación (REF: TIN2009 07707); Bioinformatics Galician Network (N° Exp: 2007/144); the Cancer Galician Network (N° Exp: 2006/60); and the COMBIOMED Spanish Network (RB07/0067/0005).

References

1. Buschmann F, Meunier R, Rohnert H, Sommerlad P, Stal M (1996) Pattern Oriented. Soft ware Architecture: A System of Patterns. John Wiley & Sons Ltd. Baffins Lane, Chichester, West Sussex PO19 1UD, England
2. Cajal Institute. CSIC. Spain. http://www.cajal.csic.es. Accessed 20 January 2009
3. Collaboratories. http://www.scienceofcollaboratories.org/Resources/colisting.php. Accessed 20 January 2009
4. GENESIS home page. http://www.genesis sim.org/GENESIS. Accessed 11 November 2008
5. Java 2 platform, enterprise edition. http://java.sun.com/developer/technicalArticles/J2EE/. Accessed 10 June 2009
6. MySQL home page, http://www.mysql.com. Accessed 10 June 2009
7. NEURON website. http://www.neuron.yale.edu/neuron. Accessed 2 April 2009
8. Olson GM, Teasley S, Bietz MJ, Cogburn DL (2002) Collaboratories to Support Distributed Science: The Example of International HIV/AIDS Research. Proceedings of SAICSIT 44 51
9. Porto A, Araque A, Rabunal J, Dorado J, Pazos A (2007) A New Hybrid Evolutionary Mechanism Based on Unsupervised Learning for Connectionist Systems. Neurocomputing 70: 2799 2808
10. Struts framework home page. http://struts.apache.org. Accessed 10 June 2009
11. Tomcat home page. http://tomcat.apache.org. Accessed 10 June 2009

Part VII
General Topics in Bioinformatics

Chapter 71
Toward Automating an Inference Model on Unstructured Terminologies: OXMIS Case Study

Jeffery L. Painter

Abstract Most modern biomedical vocabularies employ some hierarchical representation that provides a "broader/narrower" meaning relationship among the "codes" or "concepts" found within them. Often, however, we may find within the clinical setting the creation and curation of unstructured custom vocabularies used in the everyday practice of classifying and categorizing clinical data and findings.

A significant and widely used example of this lies in the General Practice Research Database which makes use of the Oxford Medical Information Systems (OXMIS) coding scheme to represent drugs and medical conditions. This scheme is intrinsically unstructured, is generally regarded as disorganized, and is not amenable to comparison with other hierarchically structured medical coding schemes. To improve processes of data analysis and extraction, we define a semantically meaningful representation of the OXMIS codes by way of the Unified Medical Language System (UMLS) Metathesaurus. A structure-imposing ontology mapping is created, and this process provides a complete illustration of a general semantic mapping technique applicable to unstructured biomedical terminologies.

Keywords Imposed hierarchy · Ontology mapping · OXMIS · UMLS · Vocabulary matching

71.1 Introduction

The construction of any ontology is in itself a grand challenge. Research relating to automating and proceduralizing this task continues to play a large role in the areas of ontology development, schema mapping, and the alignment of medical term

J.L. Painter
GlaxoSmithKline, Research Triangle Park, 27709, NC, USA
e mail: jeffery.l.painter@gsk.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 71,

systems. Our approach allows mapping the Oxford Medical Information Systems (OXMIS) [1] codes (used in the General Purpose Research Database,[1] or GPRD) to multiple coding schemes and is facilitated in large part by the Unified Medical Language System (UMLS) Metathesaurus.[2] The result enables the OXMIS codes to be viewed from the same navigational structure defined in existing and familiar coding schemes. This is of substantial benefit because the GPRD is heavily used in epidemiological and health outcome studies.

The GPRD originally employed OXMIS for coding data, but was later augmented by the use of Read codes. Identifying the correct set of codes representing a single medical concept in the GPRD is particularly problematic because of the mixture of the coding schemes that has evolved. As a result, some studies choose to ignore the portions of the GPRD data to take a uniform approach toward analyzing patient records:

"OXMIS codes used in the earlier years of the database are not hierarchically organized and do not map readily to equivalent Read codes. We therefore omitted practices which used OXMIS codes by selecting the 123 practices whose records included at least 100% Read codes in each year from 1987 to 2000." [2]

If the researchers working on this study had access to a knowledge representation of the Read-OXMIS codes[3] which could account for the partitioning of the GPRD data, they may have captured a more accurate account of patients' longitudinal records, regardless of which scheme the records were coded in. Unfortunately, with the mixed coding one finds in the GPRD, this is *not* currently possible without much manual work and the identification of corresponding codes between the Read and OXMIS coding schemes.

71.2 Methods

Our goal is to create a meaningful hierarchical structure of the Read-OXMIS codes as the basis for an informed retrieval model. The techniques employed aim for a high degree of meaning association among the codes. An additional goal is to enable the mapping of the Read-OXMIS codes to other coding schemes found within the UMLS. By making use of the UMLS Metathesaurus, we are able to create this structure and associate a concept hierarchy with the Read-OXMIS codes which will facilitate future mappings of Read-OXMIS to additional coding schemes.

---

[1]General Practice Research Database (GPRD) is maintained by the (UK) National Health Service Information Authority.

[2]UMLS Metathesaurus is a project of the (US) National Library of Medicine, Department of Health and Human Services. Available at: http://www.nlm.nih.org/research/umls/.

[3]From now on, we will refer to the collective set of codes found in the GPRD as Read OXMIS. The designation refers to the combination (OXMIS and Read version 2) of coding schemes found in this particular database's medical records.

The immediate problem with which we were presented was to decide on which UMLS source should serve as the "target" to which the Read-OXMIS code set would be mapped. For this case study, we restricted our attention to only those sources most often referenced by our epidemiologists (i.e., ICD-9, ICD-10, CPT, MedDRA, and CTV 3).[4]

Rahm and Bernstein [3] help illuminate the potential mechanisms by which one might automate the mapping of one schema into another. We approach the process of ontology mapping similarly by first attempting schema integration in their sense. The integration, even when schemata model similar domains (as in our case), first involves a matching process [3]. For our mapping process, we are interested in moving from one coding scheme (the "base") to another (the "target"). Our methodology aids in (1) identifying the appropriate target, and (2) generating an abstraction of the base which allows for imposing the hierarchical structure of the target.

One difficulty in mapping OXMIS codes to any other coding scheme is that there appears to be no comprehensive source of the OXMIS code set on electronic media. We extracted our Read-OXMIS code set using a dictionary listing of all the codes which appear directly in the GPRD data. The majority of these codes are linked to a verbatim string (the term) which assigns some meaning representation to the code itself. However, a few of the codes have no associated string, and this constitutes a problem in mapping them to another coding scheme.

We define two methods in our matching process. The first is a direct method employing exact string matching while the second takes an associative (or indirect) approach to mapping the Read-OXMIS codes. The associative mapping is a process using a mathematical (probabilistic based) calculation to associate code/term pairs with one or more candidate referents in the target coding scheme.

Both methods convert Read-OXMIS code/term pairs into a concept node related to a concept found in the selected target scheme. We call this process "reification of the concept" represented by the code. The reification of the concept bears a certain similarity to the idea of semantic ascent as addressed by Willard [4]. The process thereby allows us to abstract the code/term pair associations to one or more concept unique identifiers (CUIs) in the UMLS Metathesaurus.

The concept nodes are then used in formalizing the structure-imposing mapping of the Read-OXMIS codes. However, not all codes in the base coding scheme can be successfully mapped by using our current methods, and we reserve an unclassified "dummy" node category for these entries.

Note that we no longer look at the GPRD Read and OXMIS codes as separate coding schemes (as most users of GPRD previously have), but rather as a single

[4]SNOMED CT is copyrighted by the International Health Terminology Standards Organization (IHTSDO). ICD 9 refers to ICD 9, CM the International Classification of Diseases, 9th Revision, Clinical Modification. ICD 10 is copyrighted by the World Health Organization and developed by the National Center for Health Statistics. Current Procedural Terminology (CPT) is copyright the American Medical Association. The Clinical Terms Version 3 (Read Codes) are maintained by the (UK) National Health Service Information Authority.

Read-OXMIS vocabulary. This approach allows for the creation of a simpler yet powerful model, and ultimately aims to create a single view of GPRD which will improve data extraction and analysis.

71.2.1 Direct Mapping and Target Selection

In general, if two codes originating from two separate coding schemes are associated with the same term (modulo case differences), then it seems logical to assume that in fact those two codes are representations of the same concept which is in keeping with the method of lexical alignment as demonstrated by Zhang et al. [5].

We refer to this method as "direct mapping", and by using it we found that we could easily determine which potential coding scheme provided the greatest level of coverage. Clinical Terms Version 3 (CTV3) also known as Read version 3 was chosen as the target model in order to provide the basis for our hierarchy-imposing representation with direct coverage near 68%.

71.2.2 Associative Mapping

The direct mapping approach is not sufficient for a complete integration of Read-OXMIS into the concept framework, and we therefore enhance it with a less direct approach of associative mapping. These additional maps allow for lexical variations between source and target terms increasing the likelihood of concept identification within the UMLS Metathesaurus.

The associative mapping procedure attempts to identify candidate strings in the target model that have a probability of semantic similarity to the code/term pairs found within the Read-OXMIS coding scheme. It first preprocesses all of the verbatim strings from the Metathesaurus by using our customized string normalization process similar to the approach described by Bodenreider [6].

1. Remove case differences, parenthetical plurals, and contractions
2. Apply a standard stemming algorithm
3. Remove stopwords (customized for our domain)

As noted in Mork and Bernstein [7], similarity of the normalized string form is appropriate for lexical matching of this sort. However, we deviate from their metric of similarity (which they confess was based on personal choice) in favor of bigram comparison. We chose the use of bigrams since it provides a higher granularity of lexical comparison. A simple Bayesian calculation determines a probability of similarity between any two strings. By adjusting an arbitrary limit (which we define as the minimum match probability) these calculations must meet or exceed, and we are able to balance between precision and specificity through reiteration of our

process. Although this method is computationally intense, Jensen and Martinez [8] outline clear advantages to it over more simplistic matching techniques.

After normalizing both the Read-OXMIS and target terms, we employ a bigram matching algorithm to generate candidate term matches in the target coding scheme. Typically, bigram matching yields not one, but multiple candidate target strings for any particular Read-OXMIS code entry. The target terms meeting the minimum match probability are then collected into a match list. The resulting match list is similar to the match matrix described by El-Nasan et al. [9] used for word discrimination. Only the highest ranking match list item is selected for annotation by the Read-OXMIS code/term pair.

The idea is that given a certain level of probability in semantic similarity, lexically distinct terms should fall within the same or similar concept categories. The minimum match probability was set to 0.75 (based on observation) for this Read-OXMIS case. Mapping other coding schemes may require some adjustment to the minimum match threshold. Thus, associative mapping is characterized as a reproducible (and tunable) process that compares normalized versions of the base and target terms using bigram matching for the metric of similarity; and then reifying the base code to one or more concepts related to the highest ranking target term.

Adding the associative mapping procedure allowed us to map the Read-OXMIS coding scheme to the target (CTV3) vocabulary achieving 93% code coverage of the original Read-OXMIS codes (leaving only 8,665 Read-OXMIS codes with "dummy" nodes to be placed in our unclassified category).

71.3 Imposing Hierarchical Structure

The mapping accomplished in the previous steps gave us a method to annotate the existing CTV3 hierarchy. By annotation, we mean the association of a base Read-Oxmis code/term pair with a node in the target CTV3 hierarchy, where they share a common concept unique identifier (CUI) found in the Metathesaurus. A Read-OXMIS code can therefore annotate one or more CTV3 nodes via the mappings described above.

An illustration of the annotation of the target (CTV3) hierarchy appears in Fig. 71.1. This simplified example demonstrates several of the issues we faced and the decisions we made to support imposing the structure on Read-OXMIS. In this example, we assume there is one node (CUI-6) in the hierarchy which contained a CTV3 term annotated by a Read-OXMIS code. The code "J690.17 Partial villous atrophy" is then mapped into the hierarchical representation at this node (or category). If a Read-OXMIS code does not annotate any concept in the target representation, then it is placed under a "dummy" node for unclassified codes.

The principle we follow in imposing a foreign structure onto an otherwise unstructured coding scheme is that we keep only those nodes which have annotations

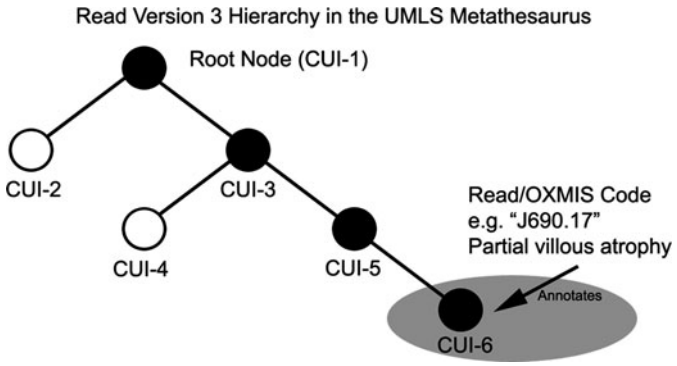


Fig. 71.1 Annotated CTV3 hierarchy

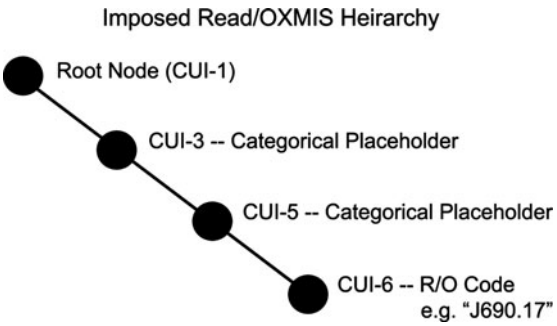


Fig. 71.2 The Read OXMIS hierarchy after purging unannotated nodes

in their downward ancestral chain. Therefore, the complete ancestral chain leading from the root to CUI-6 is preserved (denoted by solid circles). Since the nodes labeled CUI-2 and CUI-4 (open circles) fail to meet this criteria, they are subject to purging, thereby increasing the efficiency of the representation. We are then left with a single view of the Read-OXMIS codes as shown in Fig. 71.2.

The resulting structure-imposing translation now exhibits more information relevant to the coding scheme than would be possible with a rudimentary (nonstructure-imposing) translation. Just as the example above demonstrates, the actual placement of the code "J690.17" is located two levels deep from the root node, under the categorical placeholders of "Clinical findings" and "Morphology findings".

By retaining nodes from the more comprehensive target scheme as categorical placeholders, we associate contextually relevant information with the Read-OXMIS codes themselves. This is sufficient reason for preserving the complete ancestral path, since (for example) it allows a broader set of searches to succeed. In the resulting hierarchy enriched with the additional categories from CTV3, a search for "morphology" will succeed. The user may then discover the code "J690.17" in this context, and such scenarios illustrate the exploratory paradigm we are seeking to support.

71.4 Conclusion

Our process creates a view of Read-OXMIS through Clinical Terms Version 3 colored glasses. To bring structure to the Read-OXMIS codes, we borrow from the Metathesaurus-based CTV3 hierarchy to provide a template for the placement of the Read-OXMIS codes within a broad and medically meaningful context.

The additional "knowledge" this model provides by way of the semantic structures leveraged from the UMLS concept model, is now imposed on our previously deprived list of code/term pairs providing a richer environment for data retrieval and analysis.

References

1. Perry J ed: OXMIS Problem Codes for Primary Medical Care. Oxford, Headington, 1978.
2. Jones R, Latinovic R, Charlton J, Gulliford M. Physical and Psychological Co morbidity in Irritable Bowel Syndrome: A Matched Cohort Study Using the General Practice Research Database. Alimentary Pharmacology & Therapeutics 24 (5), 2006, pp. 879 886.
3. Rahm E, Bernstein PA. A Survey of Approaches to Automatic Schema Matching. The VDLB Journal 10, 2001, pp. 334 350.
4. Willard Dallas. Why Semantic Ascent Fails. Metaphilosophy 14 (3 & 4), July/October 1983, pp. 276 290.
5. Soriano, Maier, Visick, Pride. Validation of General Practitioner Diagnosed COPD in the UK General Practiec Research Database. European Journal of Epidemiology 17 (12), 2001, pp. 1075 1080.
6. Bodenreider O. Using UMLS semantics for classification purposes. Proceedings of AMIA Symposium, 2000, pp. 86 90.
7. Mork P, Bernstein PA. Adapting a Generic Match Algorithm to Align Ontologies of Human Anatomy. In: 20th International Conference on Data Engineering; 2004 March 30 April 2; Boston, MA: IEEE, 2004.
8. Jensen LS, Martinez T. Improving text classification by using conceptual and contextual features. KDD 2000 Workshop on Text Mining, Boston, 2000, pp. 101 102.
9. El Nasan A, Veeramachaneni S, Nagy G. Word Discrimination Based on Bigram Co Occurrences. ICDAR, p. 0149, Sixth International Conference on Document Analysis and Recognition (ICDAR'01), 2001.

Chapter 72
Semantic Content-Based Recommendations Using Semantic Graphs

Weisen Guo and Steven B. Kraines

Abstract Recommender systems (RSs) can be useful for suggesting items that might be of interest to specific users. Most existing content-based recommendation (CBR) systems are designed to recommend items based on text content, and the items in these systems are usually described with keywords. However, similarity evaluations based on keywords suffer from the ambiguity of natural languages. We present a semantic CBR method that uses Semantic Web technologies to recommend items that are more similar semantically with the items that the user prefers. We use semantic graphs to represent the items and we calculate the similarity scores for each pair of semantic graphs using an inverse graph frequency algorithm. The items having higher similarity scores to the items that are known to be preferred by the user are recommended.

Keywords Information retrieval · Ontology · Semantic graph · Content-based recommendation · Semantic matching

72.1 Introduction

Recommender systems (RSs) have become an important research area since the appearance of the first papers on collaborative filtering in the 1990s [1]. RS is usually classified into three categories, based on how recommendations are made: collaborative filtering recommendations (CFRs), content-based recommendations (CBRs), and hybrid approaches [2]. The CBR method, which has its roots in information retrieval (IR) research, recommends items that are similar to the ones the user is known to have preferred in the past [3]. Many current CBR systems focus

W. Guo (✉)
Science Integration Program (Human), Department of Frontier Sciences and Science Integration, Division of Project Coordination, The University of Tokyo, 277 8568 Kashiwa, Japan
e mail: gws@scint.dpc.u tokyo.ac.jp

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 72,

on recommending items containing textual information, such as Web sites [2, 4], documents [5], scientific papers [6], and news feeds.

Generally, a CBR system computes a profile for each item to be considered for recommendation by extracting a set of features from that item. That profile is then used to determine whether or not to recommend the item to a particular user by evaluating the similarity of the profile to profiles of items known to be preferred by that user. In CBR systems that base recommendations on text content, the features used to create the profile are usually keywords. However, due mainly to the ambiguity of natural languages [7, 8], a comparison of profiles based on keywords does not always result in accurate estimates of the semantic similarity of two items with respect to the intended meanings of the items. Inaccuracies can arise from ambiguities both in meanings of individual keywords and in the specific relationships between different keywords.

Semantic Web technologies offer the possibility for realizing a semantic-based model for IR systems. This chapter presents a new approach for estimating the similarity of items for CBR by using Semantic Web technologies. We call such CBR systems "semantic CBR." In semantic CBR, an item is represented by a profile in the form of a computer-interpretable semantic descriptor, giving the concepts that define the item together with the specific relationships that hold between those concepts [9]. An inference engine can then be used to estimate the similarity of the profiles for a pair of items. Because the similarity can be estimated based on logic and rule-based inference, we can get a similarity estimation that is more semantically accurate than that of text-based methods.

This chapter is organized as follows. In Sect. 72.2, we review the background of this work and introduce EKOSS, a system for creating and utilizing semantic graphs that describe knowledge resources using ontologies. In Sect. 72.3, we describe our semantic CBR approach. The inverse graph frequency (IGF) algorithm is developed to calculate the similarity scores, and we present some experimental results from applying the approach to a corpus of papers in biomedical sciences obtained from MEDLINE. Finally, we present the main conclusions of this work.

72.2 Background

In most CBR systems, items are described by keywords. For example, the Fab system [2], which recommends Web pages to users, represents Web page content with the 100 words that are judged to be most important. Similarly, the Syskill & Webert system [5] represents documents with 128 keywords. These CBR systems make recommendations by evaluating the similarities between the keyword profiles of items selected by the user previously and those of the items that have not been selected yet. The importance of each keyword for an item can be determined through weights defined using methods such as term frequency inverse document frequency (TF IDF). However, the calculation of similarities between items based on keywords suffers from problems related to the ambiguities of natural language,

such as polysemy and synonymy. In particular, the similarity based on text matching may be quite different from the similarity based on actual meaning or semantics, which is more likely to be of interest to the users.

EKOSS [10] is a Web-based knowledge sharing system that provides a set of intuitive tools for creating descriptors of shared knowledge resources with computer-interpretable semantics, called *semantic statements*, through the use of domain ontologies based on a description logic. EKOSS then uses a reasoner to provide knowledge sharing services based on interpretation of the semantic statements describing the shared knowledge resources at the semantic sentence level.

The semantic statement that has been created to describe a knowledge resource forms a semantic graph, which is composed of nodes representing instances of classes from the domain ontology used, together with arcs representing relationships between the instances. Relationship types are specified by properties

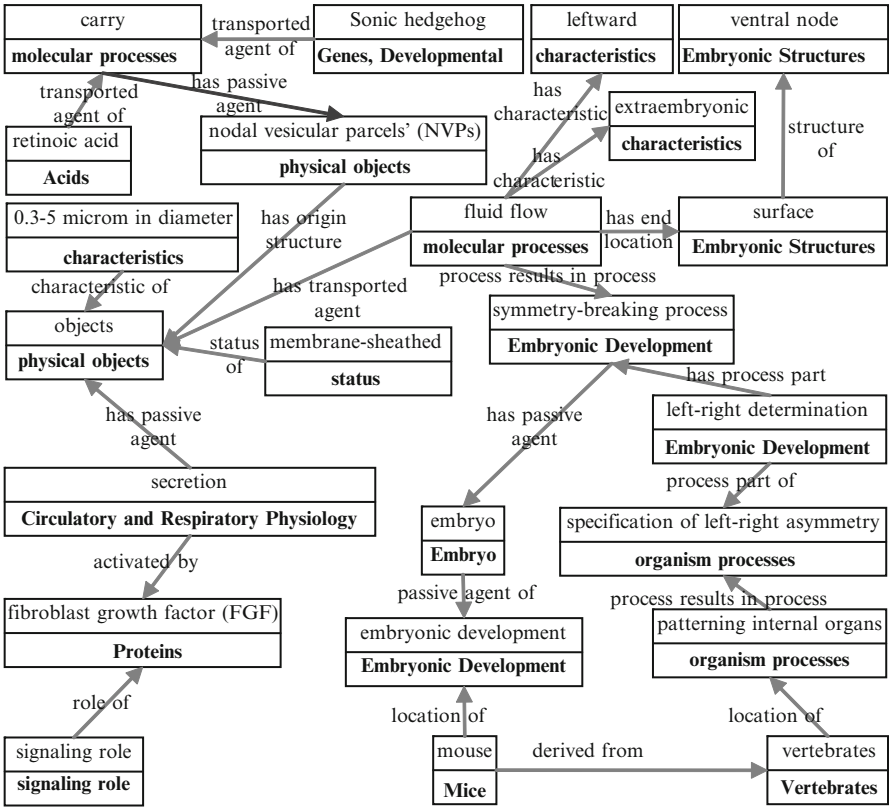


Fig. 72.1 The first semantic graph example *graph1*. *Boxes* show instances of classes from the domain ontology. The *text above the line in a box* is the instance label. The *text in bold type below the line in a box* is the class name of that instance. *Arrows* show properties, which express the relationships between instances

that are also defined in the ontology. Each instance can have a descriptive text label [10].

To illustrate how our semantic CBR system works, we will use three examples of semantic graphs that have been created for papers from the MEDLINE database using an ontology we have developed for the domain of life sciences.

Figure 72.1 shows the first semantic graph example, *graph1*, that represents a MEDLINE paper, *paper1*, entitled "FGF-induced vesicular release of Sonic hedgehog and retinoic acid in leftward nodal flow is critical for left right determination." The second example, *graph2*, represents a paper, *paper2*, entitled "Nodal flow and the generation of left right asymmetry," and the third example, *graph3*, represents a paper, *paper3*, entitled "Rapid modulation of long-term depression and spinogenesis via synaptic estrogen receptors in hippocampal principal neurons." These graphs are not shown due to space limitations.

A reasoner has been developed for evaluating matches between queries and graphs using both logical reasoning and rule-based reasoning. The logic is built into the ontology using formalisms provided by the description logic that is supported by the ontology specification we used (OWL-DL). The rules are predefined for a particular ontology by domain experts. Details are given in [10, 11].

72.3 Semantic Content-Based Recommendations

The first step of the semantic CBR method is to create a profile for each item using a descriptor with computer-interpretable semantics. In the EKOSS system, an item is a knowledge resource and a profile is a semantic graph. Therefore, we use the EKOSS system as explained in Sect. 72.2 to represent knowledge resources by semantic graphs. The second step is to measure the similarity of each pair of item profiles. We do this by calculating a similarity score from their semantic graphs using the IGF algorithm described next. We repeat the process until all pairs of semantic graphs are evaluated. Then, given a knowledge resource that is known to be preferred by a user, we can find the knowledge resources having high similarity scores with that preferred resource and recommend them to that user.

Matching semantic graphs is different from matching text strings. We use the EKOSS reasoner to determine the match between two semantic graphs based on a combination of logic and rule-based inference [11]. First, we add one semantic graph, called the *target graph*, to the reasoner's knowledge base. Then, we convert the second semantic graph, called the *search graph*, into a set of semantic triple queries. A semantic graph can be considered as a set of triples, each of which includes two instances and one relationship (property) between them. Therefore, there will be one triple for each property in the semantic graph.

Next, for each of the triple queries, we ask the EKOSS reasoner to find a pair of instances in the target semantic graph meeting the class and relationship constraints of the triple query. If such a pair of instances exists, then the triple query is said to match with the semantic graph. By using inference based on both logical reasoning and rule-based reasoning, we can get matching results that are implied semantically

because the reasoner can infer relationships between instances that are not explicitly stated in the semantic graph. The fraction of matching queries gives the degree of semantic match or similarity between the two graphs.

The algorithm for calculating the similarity scores is of central importance. Like terms in text documents, the triples in semantic graphs can be considered to have differing importance: commonly occurring triples should be less important than relatively rare triples. Thus, the importance of a triple can be modeled by the inverse frequency of the triple in the set of semantic graphs in the same way that keyword weights are evaluated using the TF IDF method. We have developed an IGF algorithm based on this model to provide weights for semantic queries for calculating the similarity scores.

The IGF algorithm is applied after the evaluation of matches between all pairs of semantic graphs is finished. We denote the ith semantic graph as G_i, $i \leq n$, where n is the total number of semantic graphs. We denote the jth query of G_i as $Q_{i,j}$, $j \leq m$, where m is the number of triple queries in G_i. The inverse graph frequency $\text{igf}_{i,j}$ of $Q_{i,j}$ is given by (72.1):

$$\text{igf}_{i,j} = \ln \frac{n}{\text{total semantic graphs matching with } Q_{i,j}}. \tag{72.1}$$

Then for a search semantic graph G_i and a target semantic graph G_j, we calculate the similarity score $\text{score}_{i,j}$ using (72.2):

$$\text{score}_{i,j} = \frac{\sum_{k=0} \text{igf}_{i,k}(\text{match}(i,k,j))}{\sum_{k=0} \text{igf}_{i,k}} \times 100. \tag{72.2}$$

Here, the function match(i, k, j) evaluates to 1 if query $Q_{i,k}$ matches with G_j and 0 otherwise. The similarity score ranges from 0 to 100, where a value of 0 means that no part of the search semantic graph matches with the target semantic graph and a value 100 means that the search semantic graph is the same or entirely contained in the target semantic graph at a semantic level.

Using (72.2), the similarity scores of the search semantic graph, *graph1*, against the target semantic graphs, *graph2* and *graph3*, are 27 and 2, respectively. An expert in the area of life sciences addressed by the three papers confirmed that the paper represented by *graph2*, *paper2*, is in fact more similar to *paper1*, represented by *graph1* than *paper3*, represented by *graph3*.

We can treat the Medical Subject Headings (MeSH) terms of the MEDLINE papers as keywords for comparison. Both *paper2* and *paper3* have two MeSH terms that are also used in *paper1*. We also can consider the class names used in each semantic graph as keywords. *Graph2* has four classes that are the same or are subsumed by classes in *graph1*, while *graph3* and *graph1* share three classes. So the large difference in the similarity of *paper1* with *paper2* and *paper3* could not be detected simply by comparing MeSH terms or individual classes, apparently because the differences in content between *paper2* and *paper3* only emerged when the relationships between the concepts in *graph2* and *graph3* were considered.

We used the semantic graph matching method to calculate the similarity scores of the MEDLINE paper represented by *graph1*, *paper1*, against a corpus we have created of semantic graphs for 392 MEDLINE papers. We also determined the numbers of MeSH terms that are shared by *paper1* with each of the 392 papers in the corpus. Figure 72.2 shows the comparison of the two similarity measures. Many MEDLINE papers, shown at the low percentiles, have high similarity scores with *paper1* as estimated by semantic graph-based method even though they do not have any MeSH terms in common. On the other hand, some papers sharing two MeSH terms with *paper1*, the papers at the highest percentiles, have low similarity scores according to the semantic graph-based method.

We can explain the second conclusion by considering that the semantic graph-based method uses ontology and logic inference technologies to account for relationships between concepts or entities when evaluating the similarity of papers. Even if two papers share several MeSH terms, if the relationships between them are different, the two papers will be identified as being dissimilar using the semantic graph-based method. One possible explanation for the first conclusion is that often MEDLINE papers have just a few MeSH terms that do not cover the entire content of the paper. Compared to the MeSH term-based method, the semantic graph-based method covers the content of the papers more thoroughly. Furthermore, by using inference based on class subsumption, two papers can be estimated as being similar even if they do not contain instances of the exact same classes.

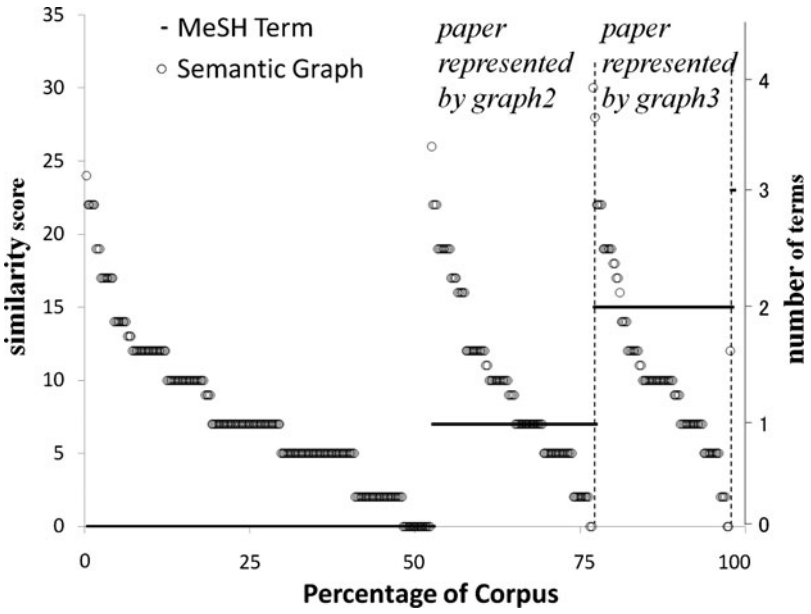


Fig. 72.2 The comparison of similarity scores calculated by MeSH term matching method and semantic graph based method for the paper represented by semantic graph *graph1* against 392 MEDLINE papers. The papers represented by *graph2* and *graph3* are indicated

72.4 Conclusions

Recommender systems have attracted attention for their ability to make useful suggestions to users. Content-based recommendation (CBR) is one type of recommender systems. Most CBR systems are designed to recommend textual items, usually described with keywords. However, the ambiguity of natural languages reduces the effectiveness of methods based on keywords. We present a semantic CBR method to recommend items that are more similar semantically with the items that the user prefers. Specifically, we use semantic graphs to represent the items and we estimate the similarity scores for each pair of semantic graphs using a semantic matching method. We use an IGF algorithm that we developed to calculate the similarity scores. We then recommend items having higher similarity scores to the items that are known to be preferred by the user. We illustrate the effectiveness of the semantic CBR method with an implementation of our approach that we have created using the EKOSS system for a corpus of scientific papers from MEDLINE.

References

1. Goldberg D, Nichols D, Oki BM, Terry D (1992) Using collaborative filtering to weave an information tapestry. Communication of the ACM 35(12):61 70.
2. Balabanovic M, Shoham Y (1997) Fab: Content based, collaborative recommendation. Com munication of the ACM 40(3):66 72.
3. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: A survey of the state of the art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6):734 749.
4. Li J, Zaiane OR (2004) Combining usage, content, and structure data to improve Web site recommendation. In: Proceedings of the 5th International Conference on Electronic Commerce and Web Technologies (EC Web'04), pp. 305 315.
5. Pazzani M, Billsus D (1997) Learning and revising user profiles: The identification of interesting Web sites. Machine Learning 27:313 331.
6. Lin J, Wilbur WJ (2007) PubMed related articles: A probabilistic topic based model for content similarity. BMC Bioinformatics 8:423. DOI 10.1186/1471 2105 8 423.
7. Rindflesch TC, Fiszman M (2003) The interaction of domain knowledge and linguistic structure in natural language processing: Interpreting hypernymic propositions in biomedical text. Journal of Biomedical Informatics 36(6):462 477.
8. Jensen LJ, Saric J, Bork P (2006) Literature mining for the biologist: From information retrieval to biological discovery. Nature Reviews Genetics 7:119 129.
9. Rzhetsky A, Seringhaus M, Gerstein M (2008) Seeking a new biology through text mining. Cell 134:9 13.
10. Kraines SB, Guo W, Kemper B, Nakamura Y (2006) EKOSS: A knowledge user centered approach to knowledge sharing, discovery, and integration on the Semantic Web. In: Cruz I, et al. (eds) ISWC 2006, LNCS 4273. pp. 833 846, Springer, Heidelberg.
11. Guo W, Kraines S (2008) Explicit scientific knowledge comparison based on semantic description matching. In: American Society for Information Science and Technology 2008 Annual Meeting, Columbus, OH.

Chapter 73
Modeling Membrane Localization: Case Study of a Ras Signaling Model

Edward C. Stites

Abstract Modeling a biological system requires the careful integration of experimental data. It is unclear how best to incorporate rate constants measured in three-dimensional solution for reactions that physiologically occur between reactants confined to the two-dimensional cell membrane. One method adjusts second order rate constants by a factor that is the ratio of the cytoplasmic volume to the volume of a shell which membrane bound proteins can access. The value for this factor has been estimated to be 250. We have previously used this method in our model of the Ras signaling network that made several experimentally confirmed predictions. Here, we investigate if the value of this parameter affects model based predictions. We find that many of our results are robust to the value used. Two predictions appear to be sensitive to the value of the parameter: predicted levels of WT RasGTP after transfection with WT Ras and the experimentally observed increased levels of WT RasGTP when a GTPase Accelerating Protein (GAP) insensitive Ras mutant is present. For these predictions that are sensitive to the value of the membrane localization parameter, we find that the theoretically derived value of 250 results in model predictions that most closely match experimental observations.

Keywords Cancer · Cell signaling · GTPases · Membrane localization · Physicochemical modeling · Ras

E.C. Stites
Medical Scientist Training Program, University of Virginia, MR6 Rm 3708, 801386 Charlottesville VA 22908, USA
e mail: ecs4a@virginia.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 73,

Abbreviations

G12D	glycine to aspartic acid at codon 12
G12V	glycine to valine at codon 12
GAP	GTPase activating protein
GDP	guanosine diphosphate
GEF	guanine nucleotide exchange factor
GTP	guanosine triphosphate
*K*m	Michaelis constant
M	molar
NF1	neurofibromin
WT	wild type

73.1 Introduction

Mathematical modeling has become an important tool for studying cell signaling networks [1, 2]. Mathematical models can be limited, however, by the quality of the information that goes into the model. It is still unclear how well rate constants measured in vitro apply to in vivo systems. It is even more unclear for membrane localized reactions. The restriction of biomolecules to diffusion in a two-dimensional membrane rather than the three-dimensional cytoplasmic volume can have large impact on system behavior [3, 4]. One theoretical analysis estimates that membrane localization results in an apparent increase in second order rate constants by approximately 10 1,000-fold [3], a separate analysis also estimates a maximum increase by a factor of 1,000 [5].

Different approaches have been used to model membrane bound reactions. One approach has simply been to use in vitro rate constants from measurements made in three-dimensional solution. This approach essentially ignores the known effects of membrane localization. An alternative approach is to fit a model to experimental data, essentially ignoring the existing quantitative data about the individual reactions in vitro. An approach in between these two extremes would seem to offer many benefits. The in vitro data likely reflects trends that will be maintained in vivo (e.g., differentiating slow reactions from fast reactions). A method for approximating in vivo rate constants from in vitro data could also reduce the number of parameters needing to be fit. One such method was used in the work of Markevich et al. in their study of receptor tyrosine kinase-induced Ras signaling kinetics [6]. Their study adjusted second order rate constants between two membrane localized reactants by a factor equal to the ratio of the cytoplasmic volume to the volume of a shell that a membrane bound protein could access; earlier, theoretical analysis estimated this value to be 250 [5]. The approach led to a model with good correlation with experimental data.

We used this method in our model of the Ras signaling network [7]. The resulting model had good agreement with existing quantitative data about Ras signal intensity. Model-based investigations yielded several interesting hypotheses; subsequent experimental work identified the patterns of Ras activation consistent with the hypotheses. It is possible that the success of our Ras model was due to the validity of the method used to correct for membrane localization. Alternatively, model predictions could be robust to the value of the parameter used for membrane localization, just as many of the model-based predictions were robust to changes in rate constants [7]. Here, we investigate how the choice of membrane localization parameter may have influenced model predictions.

73.2 Methods

Ras is a small GTPase that is central to our model (Fig. 73.1). The C-terminus of Ras localizes Ras to the cell membrane. We consider Ras and the classes of proteins that directly interact with Ras to regulate Ras signals. As a GTPase, Ras binds guanine nucleotides GTP and GDP and can hydrolyze GTP to GDP. The GTP bound form of Ras is considered the "active" form for transmitting signals (such as those involved in cell proliferation). Many proteins in the cell specifically interact with RasGTP; these proteins are termed Ras effectors. The RasGTP-effector complex contributes to effector activation and the transmission of Ras signals downstream. Nucleotides can freely dissociate from and associate with Ras. Guanine nucleotide Exchange Factors (GEFs) also facilitate nucleotide exchange. GTPase Accelerating Proteins (GAPs) catalyze nucleotide hydrolysis. These reactions were described with mass action kinetics [7]. The GAP and GEF reactions were simplified to the irreversible and reversible Michaelis Menten enzymatic reactions, respectively.

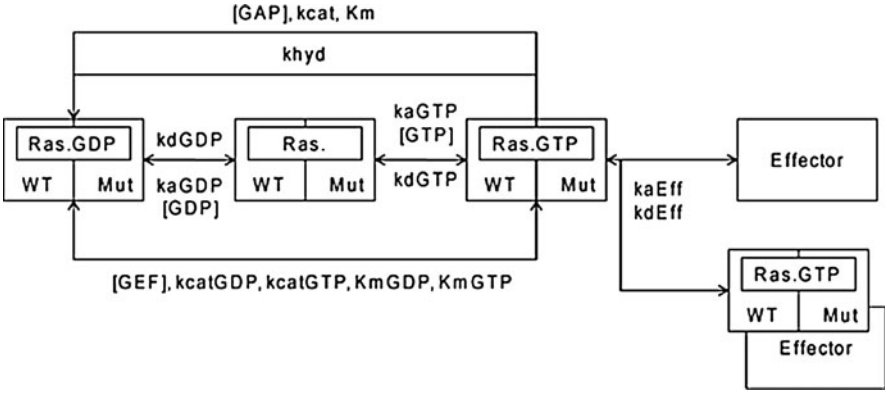


Fig. 73.1 Schematic of the modeled Ras GTPase cell signaling network

The majority of needed rate constants were found in the literature. As rate constants in the literature were measured in three-dimensional solutions rather than the two-dimensional membrane environment, we used the method described earlier to estimate membrane bound rate constants. Our original model adjusted the Km for the GAP and GEF reactions by a factor of 250 with good results [7]. This parameter is referred to as D. To remove consideration of membrane localization, we use the measured values of Km, equivalent to setting parameter D to value 1. To consider alternative levels of the membrane localization correction parameter D, we consider the values of 10, 25, 50, 100, and 500 in addition to the value of 1 mentioned above and the value of 250 used originally. The parameters for the quantity of basally active GEF, [GEF], and basally active GAP, [GAP], are the only two parameters in the model that are fit, and the fitting is done after parameter D has been applied to the Km. To find the values for these two parameters, we perform simulations with different [GEF] and [GAP] and determine the level of basal RasGTP and the rate of Ras nucleotide exchange at basal, unstimulated conditions. We find the minimum root mean square error for the difference between predictions and experimentally determined values as described previously [7]. We do this to one significant figure of precision for each parameter. This results in seven sets of {D; [GEF], [GAP]} (Table 73.1). We evaluate Ras signal intensity by using simulations to find the total amount of RasGTP (free RasGTP and RasGTP bound effector) for a set of parameters and conditions as we did in our original work. When Ras mutants are included in the model, one half of total Ras is modeled as mutant Ras and the remaining half as wild-type Ras (RasWT). Parameters of mutant RasG12V and RasG12D are as specified previously [7]. To simulate the GAP deficient state that occurs when both copies of Ras GAP neurofibromin are lost, we eliminate all GAPs in the model.

73.3 Results

Previously published experiments have measured the fraction of total Ras bound to GTP after transfection with RasWT, RasG12V, or RasG12D [8, 9]. For each of the considered membrane localization parameters, D, with dependent [GEF] and [GAP] (Table 73.1), we use our model to simulate similar conditions (Fig. 73.2).

Table 73.1 Values of membrane localization parameter and dependent parameters GEF and GAP used in the investigations of membrane localization parameter D

D	GEF [M]	GAP [M]	RasGTP
1	4×10^{-08}	6×10^{-09}	2%
10	4×10^{-09}	6×10^{-10}	2%
25	2×10^{-09}	3×10^{-10}	2%
50	1×10^{-09}	2×10^{-10}	2%
100	4×10^{-10}	9×10^{-11}	2%
250	2×10^{-10}	6×10^{-11}	2%
500	1×10^{-10}	4×10^{-11}	2%

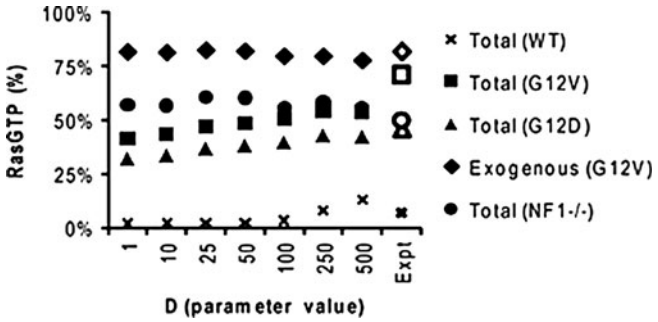


Fig. 73.2 Percentage of Ras bound to GTP for simulated conditions of transfection with a Ras construct for total Ras or exogenous Ras, and also nontransfected cells in the NF1 / GAP deficient state. Values of the membrane localization parameter used are indicated on the x axis. Expt, experimental values [8 11] for comparison with model predictions

In general, the higher value of D tends to result in a slightly higher fraction of total Ras bound to GTP. The case of RasWT transfection results in the greatest range of predicted fraction of total Ras bound to GTP, with an approximate sixfold change from $D = 1$ to $D = 500$. Previously published experiments have also measured the fraction of exogenous Ras transfected into a cell that is bound to GTP for both RasG12V and RasWT [10]. We use our simulations to find the fraction of exogenous Ras in the network bound to GTP under similar conditions (Fig. 73.2). For these conditions, the predicted fraction of RasG12V bound to GTP is relatively insensitive to the value of the membrane localization parameter and tends to match quite well with experimental observations. Experimental measurement in Ras GAP neurofibromin deficient (NF1$-/-$) cancer cells find approximately 30 50% of Ras bound to GTP [11]. Simulations of the GAP deficient state across the different membrane localization parameters are relatively insensitive to the choice of membrane localization parameter (Fig. 73.2).

One unanticipated prediction of our original model was that the presence of a GAP insensitive Ras mutant would cause an increase in the fraction of RasWT bound to GTP. This increase resulted from the nonproductive interaction between Ras GAP and GAP insensitive Ras mutants, like RasG12V and RasG12D. This interaction results in a competitive inhibition of the GAP enzymatic domain, thus preventing GAP from performing its negative regulation on RasWT. Our previous experiments have found increased levels of RasWT bound to RasGTP when a Ras mutant is present [7]. To see if this prediction would be maintained when different values of membrane localization parameter D were used, we performed simulations to find the amount of GTP bound RasWT when RasG12V or RasG12D was also present. Data (Table 73.2) show when membrane localization is ignored there is essentially no competitive inhibition as 2% of RasWT is bound to GTP. This effect slowly increases; when $D = 50$, 10% of RasWT is bound to GTP when RasG12V is present, and when $D = 500$, 26% of RasWT is bound to GTP when RasG12V is present.

Table 73.2 Levels of RasGTP for WT and mutant Ras proteins in a network with both WT and mutant Ras for different values of membrane localization parameter D. WT/G12V, network with both WT and G12V Ras. WT/G12D, network with both WT and G12D Ras (WT/G12D)

D	WT (WT/G12V)	G12V (WT/G12V)	WT (WT/G12D)	G12D (WT/G12D)
1	2%	83%	2%	66%
10	4%	83%	4%	66%
25	7%	85%	6%	68%
50	10%	85%	8%	68%
100	15%	83%	12%	65%
250	22%	84%	19%	66%
500	26%	83%	23%	64%

73.4 Discussion

This analysis was performed to investigate a method for considering membrane localization. We here investigated different values of parameter D to see how choice of parameter value might have influenced these model predictions. Many predictions were robust to the value of D considered, including $D = 1$ (ignoring membrane localization). Levels of RasGTP for mutants RasG12V, RasG12D, and the GAP deficient state were largely unaffected by the value used for D. Levels of RasGTP for WT Ras overexpression were more divergent, but the range was consistent with published values. Arguably, the values for $D = 250$ perform better, but with the imprecise nature of biological data and uncertainty in the extent of Ras WT overexpressed in the original experiments [9], we would feel uncomfortable suggesting this value was better than others from this data alone.

One of our key experimentally validated predictions (increased WT RasGTP from the competitive inhibition of Ras GAPs by GAP insensitive Ras mutants) would not have been made if membrane localization had not been considered. This suggests that modelers may miss important biological processes if they do not include membrane localization. A model without membrane localization predicted essentially no competitive inhibition. As the value for parameter D increased, competitive inhibition became an increasingly more important process. The amount of RasGTP observed experimentally with a GAP insensitive mutant present was approximately an order of magnitude greater than that observed when a GAP insensitive mutant was not present [7]. This would correlate well with the value of 250 used, where 22% of wild-type Ras was bound to GTP. Overall, the value of 250 used previously by others [6] and by us [7] seemed to work well. This value is in the range determined by different theoretical approaches [3, 4]. A value of 250 for the membrane localization parameter may therefore serve as a good first approximation for iterative model development.

References

1. Papin JA, Hunter T, Palsson BO et al. (2005) Reconstruction of cellular signalling networks and analysis of their properties. Nat Rev Mol Cell Biol 6:99 111.
2. Aldridge BB, Burke JM, Lauffenburger DA et al. (2006) Physicochemical modelling of cell signalling pathways. Nat Cell Biol 8:1195 1203.
3. Haugh JM and Lauffenburger DA (1997) Physical modulation of intracellular signaling processes by locational regulation. Biophys J 72:2014 2031.
4. Kholodenko BN (2003) Four dimensional organization of protein kinase signaling cascades: the roles of diffusion, endocytosis and molecular motors. J Exp Biol 206:2073 2082.
5. Kholodenko BN, Hoek JB and Westerhoff HV (2000) Why cytoplasmic signalling proteins should be recruited to cell membranes. Trends Cell Biol 10:173 178.
6. Markevich NI, Moehren G, Demin OV et al. (2004) Signal processing at the Ras circuit: what shapes Ras activation patterns? Syst Biol (Stevenage) 1:104 113.
7. Stites EC, Trampont PC, Ma Z et al. (2007) Network analysis of oncogenic Ras activation in cancer. Science 318:463 467.
8. Bollag G, Adler F, elMasry N et al. (1996) Biochemical characterization of a novel KRAS insertion mutation from a human leukemia. J Biol Chem 271:32491 32494.
9. Gibbs JB, Marshall MS, Scolnick EM et al. (1990) Modulation of guanine nucleotides bound to Ras in NIH3T3 cells by oncogenes, growth factors, and the GTPase activating protein (GAP). J Biol Chem 265:20437 20442.
10. Boykevisch S, Zhao C, Sondermann H et al. (2006) Regulation of Ras signaling dynamics by Sos mediated positive feedback. Curr Biol 16:2173 2179.
11. Basu TN, Gutmann DH, Fletcher JA et al. (1992) Aberrant regulation of Ras proteins in malignant tumour cells from type 1 neurofibromatosis patients. Nature 356:713 715.

Chapter 74
A Study on Distributed PACS

Aylin Kantarcı and Tolga Utku Onbay

Abstract Advances in information technologies in the past decades expanded the medical information systems outside the boundaries of healthcare institutions. Currently, sharing of medical data among medical institutions for collaborative, research, training, and diagnostic purposes is a challenging topic in medicine and computer sciences. This study focuses on how virtual organizations can be created for sharing PACS archives of geographically distributed organizations. First, we introduce DIPACS 1.0 framework that we implemented for small- and medium-scale distributed organizations. Then, we propose a grid-based framework to expand DIPACS 1.0 to encompass large-scale distributed environments. We believe that grids that integrate computing, storage, and network resources of dynamic and geographically disperse organizations will provide secure, reliable, flexible, and high-performance infrastructures for next generation health information systems.

Keywords DICOM · Distributed system · Grid · Medical image management · PACS

74.1 Introduction

Sharing of radiological images for diagnostic, collaborative, and administrative purposes is one of the challenging issues for both academia and industry. Over the last decade, digital imaging systems have become more powerful and less

A. Kantarcı (✉)
Computer Engineering Department, Ege University, Izmir, Turkey
e mail: aylin.kantarci@ege.edu.tr

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 74 ,

expensive. Hence, they have received wide acceptance in healthcare institutions. With the increase in the number of new modalities, the volume and uses of medical data have increased tremendously [1]. Consequently, Picture Archiving and Communication System (PACS) and Digital Imaging and Communications in Medicine (DICOM) concepts have been introduced and they have become the standards for accessing multimedia patient data and for setting up environments for applications such as telemedicine, collaborative work, etc. Currently, medical imaging equipments output data in the form of DICOM objects, and PACS servers store DICOM files. A DICOM object contains several descriptive information blocks in addition to pixel data. This metadata includes patient information, modality, acquisition parameters, image resolution and measurements of clinical trial studies, etc. [2, 3].

Initially, PACS was designed as a standalone system for departmental or hospital use only. Early PACS applications had simple point-to-point architectures. Advancements in networking technologies provided inexpensive communication environments with high bandwidth such as the Internet. In the literature, there are many studies that describe browsers for accessing PACS services remotely [4, 5]. However, the storage and control were still central. Although a centralized system could be used in some cases to host all data, financial, logistic, and administrative constraints make the centralized approach difficult. The high volume of data and a wide range of usage requirements from diagnosis to collaborative work made multicenter image management inevitable. Current developments in Distributed Systems field contributed to the introduction of distributed PACS systems. While initial PACS concept covered only point-to-point architectures, today it encompasses several technologies that include hardware and software for acquisition, storage, distribution, and analysis of digital images in distributed environments [2]. Current trend is to develop nation-wide PACS applications on grid-enabled middleware infrastructures [6 8].

In this study, we designed and implemented a distributed PACS, DIPACS 1.0, which enables querying and retrieval of multimedia patient data stored in the archives of multiple health centers in a transparent way. DIPACS 1.0 can easily be adopted by the already available PACS infrastructures of health centers without extra cost.

The initial design goal of DIPACS 1.0 was to share radiological data of geographically distributed branches of a private health institution in Turkey. Although DIPACS 1.0 is scalable, for the reasons that will be explained in later sections, its nationwide employment may suffer from some performance drawbacks. Therefore, we think that DIPACS 1.0 is suitable for small- and medium-sized networked environments. For a nationwide, large, distributed health system, we designed DIPACS 2.0. Due to its grid-based infrastructure, higher level of transparency and scalability can be achieved with DIPACS 2.0. In Sect. 74.2, we will introduce the architecture and services of DIPACS 1.0. In Sect. 74.3, we will give information on the design of DIPACS 2.0 together with some explanation on grid systems. Finally, conclusions will be drawn in Sect. 74.4.

74.2 The Developed System: DIPACS 1.0

Components of a DIPACS 1.0 environment are a Nameserver, DIPACS gateways, PACS servers, and workstations in health centers. Connections among these components are given in Fig. 74.1. PACS servers and workstations in a health center are connected to the Internet via a distributed PACS (DIPACS) gateway. DIPACS gateways of different health institutions are informed of each other via a Nameserver, an independent computer connected to the Internet.

Prior to system startup, access points to DIPACS gateways in DIPACS 1.0 environment are configured to the Nameserver via XML configuration files. Similarly, PACS servers and workstations in a health center are introduced to the DIPACS gateway via XML configuration files. For two-way communication within the health center, DIPACS gateways are also introduced to workstations and PACS servers through GUIs. When a new domain joins DIPACS 1.0 environment when the system is under operation, it is introduced to the Nameserver via XML files. The Nameserver then notifies other domains about the new domain.

After the configuration phase, the system is started up. During system operation, a workstation in a health center can send queries to DIPACS 1.0 environment over its DIPACS gateway. Similarly, any query/response received from a DIPACS gateway is sent to the related PACS server(s)/workstation. Communication within health centers relies on DICOM standard. For security reasons, DICOM communication among health centers is encrypted with Transport Layer Security (TLS)/Secure Socket Layer (SSL), which are cryptographic protocols that provide security for communications over networks such as the Internet [9].

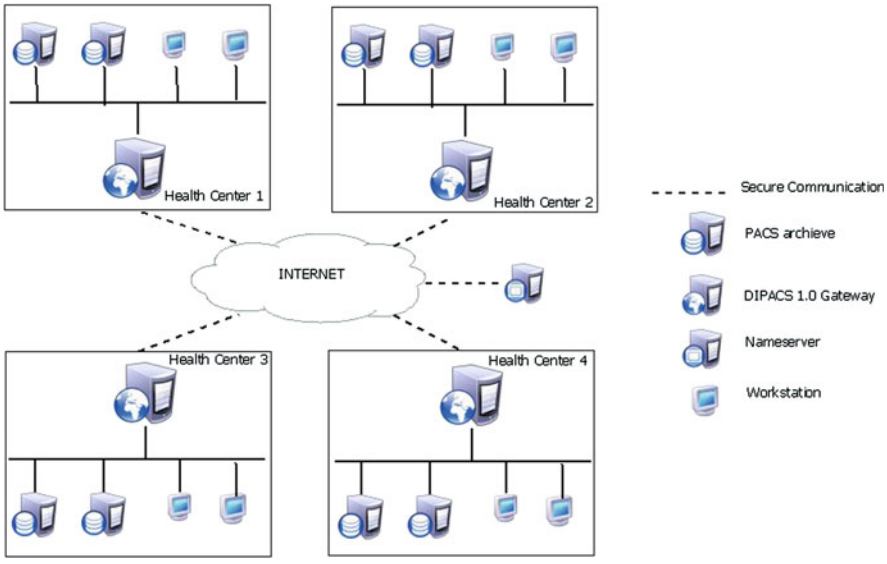


Fig. 74.1 System architecture for DIPACS 1.0

The software architecture of DIPACS 1.0 is given in Fig. 74.2. First, we implemented DIPACS 1.0 with Java Technologies. Java is an object-oriented environment with many facilities for implementing distributed applications. For example, Java Run Time environment provides interoperability with different computer system architectures. Therefore, any health center can easily integrate into DIPACS 1.0 environment. Second, Java Remote Method Invocation (RMI) technology allows implementing distributed objects whose methods can be invoked remotely. In DIPACS 1.0, we used RMI technology to implement communication between the Nameserver and DIPACS gateways. DIPACS gateways invoke the getRemoteGateways() method of the Nameserver to obtain the access points to other DIPACS gateways. Security facilities of RMI have a very important role in providing secure communication with the Nameserver. Multithreading features of RMI enable parallel processing and serving multiple requests at the same time [9].

Each DIPACS gateway's software consists of four packages, namely, DICOM, REMOTE, COMMUNICATION, and COMMON packages. DICOM package provides the exchange of DICOM objects between PACS components and DIPACS gateways over dcm4che2 library [10]. dcm4che2 is an open source Java implementation of DICOM Standard. C-Echo, C-Find, C-Move, and C-Store services of DICOM package are registered to ApplicationEntity class, which inherits NetworkApplicationEntity from the dcm4che2 library. Besides registration of DICOM services, the ApplicationEntity class provides authentication of DIPACS gateways and keeps access information of DIPACS gateways and local PACS components. COMMUNICATION package has access to DICOM and REMOTE packages. COMMUNICATION package processes messages incoming from and outgoing to the local and distributed network. DICOM package conveys the messages it receives from the COMMUNICATION package to other DIPACS gateways over the dcm4che2 library. COMMUNICATION package contacts the REMOTE package to obtain the access information of DIPACS gateways from the Nameserver. As previously mentioned, java.rmi package is used for communication with the

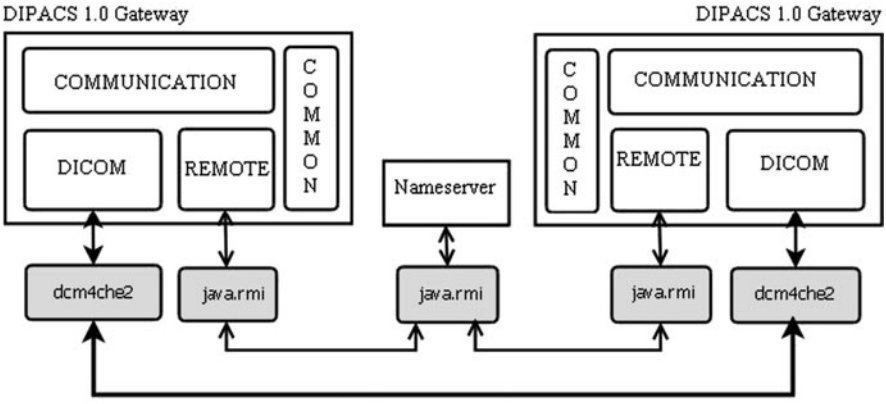


Fig. 74.2 Software Architecture for DIPACS 1.0

Nameserver. Finally, COMMON package contains the classes and interfaces used by the DIPACS gateway and the Nameserver. Further details on the implementation of DIPACS 1.0 can be found in [9].

74.3 A Grid-Based Design: DIPACS 2.0

Although DIPACS 1.0 meets the expectations of small- or medium-sized organizations, it has important inefficiencies for large medical networks. First, from the viewpoint of scalability, problems arise when new health centers join the network. Since a query is sent to each DIPACS 1.0 gateway, system load increases tremendously as new centers join DICOM 1.0. Second, there is no replication capability in DICOM 1.0. Accessing a single copy increases the load around it, resulting in high delay and low performance. In a system with replication, the nearest copy is accessed to decrease delay and system load. Inexistence of replication facility also renders the system vulnerable to failures. A problem in a PACS archive or a DIPACS 1.0 gateway makes the related images inaccessible. The Nameserver is itself a central point of failure. A fault in the Nameserver during initialization renders the operation of the system impossible. If a failure occurs after start up, new DIPACS domains cannot be introduced to existing domains, making their PACS servers inaccessible. Third, security structure of DIPACS 1.0 should be empowered. Commonly, in a networked environment, Secure Socket Layer (SSL) or Transport Layer Security (TLS) for encryption is combined with password or X.509 certificate authentication to form a Virtual Private Network (VPN) connection between end pairs. Lack of dynamic resource management, consistent authentication, and fine grain authorization mechanisms are major limitations for employing VPN-based solutions in medical image transmission systems. Confidentiality must be supported across organizational boundaries. Anonymosity is a necessity, especially in research and educational PACS in which content-based query mechanisms are employed [11].

What is needed for a large-scale medical image workflow is an architecture with high speed reliable transport, enterprise level security, large-scale data management and replication, publication, discovery, and resource management services. Grid technology, an informatics approach to federate securely independent computing, storage, and data management resources over public networks, fits this structure [12]. A grid is a high-performance hardware and software infrastructure providing scalable, dependable, and secure access to the applications utilizing the grid. The hardware is heterogeneous computers and network systems. Unlike clustered systems, participating computer systems maintain administrative autonomy. The software is a middleware with services for enabling efficient access to such a system in a transparent way. Grids spawn a virtual organization (VO) over networks between resource providers and users [7, 12, 13]. During the evolution of grid technology, diversity of grid implementations and lack of interoperability became a major bottleneck. To overcome this bottleneck, Open Grid Service Architecture

(OGSA) was introduced as a standardization effort. In its current state, OGSA's building blocks are based on Web Services Resource Framework (WSRF). Being a services-based framework built on top of existing web standards (XML, SOAP, and WSDL), OGSA enables secure exchange of data objects and messages through the grid. Web service-oriented architecture of OGSA makes application development easier with the usage of widely adopted approaches and hence enables faster transition of organizations into grid environments. The most major OGSA-based grid toolkit available today is Globus Toolkit 4 [14, 15].

Grids were initially used for computing intensive fields such as climate modeling and physics. Recently, grid technology has been adopted in many other areas such as life sciences, in particular, medicine [14]. In the literature, there are example grid systems for medical image management [6 8, 14, 16].

One problem in adopting GT4 in medical image management stems from the fact that it is designed for general purpose grids, not for a particular field. GT4 is based on OGSA services relying on web services to exchange data objects and messages securely. Therefore, grid services are based on HTTP and in the existing studies, DICOM images are converted to XML format and passed onto the grid service. Similarly, received images from the grid are also in XML format and converted to DICOM format. The problem with HTTP is that it relies on TCP/IP, which is a synchronous request/reply protocol. When HTTP-based protocols are employed in medical grids, the user has to wait for a status response until the complete series of images are transferred. This property of HTTP protocol is in conflict with the asynchronous nature of DICOM protocol. During transfer of large volumes of data, timeouts due to large end-to-end delays terminate the connection and halt the operation. The other problem is that HTTP is a text-based protocol, whereas DICOM images are encoded in binary. It is debatable whether Web services provide equal performance with binary transmission protocols for large binary datasets. For these reasons, it is questionable whether web services are suitable for medical grids [14]. In [14], the problems with web services and the commonly employed transport protocol in grid systems, GridFTP, are discussed in detail, and a new protocol called Grid-DICOM is proposed. Unlike GridFTP, Grid-DICOM does not perform modifications on the data that come with the costs of buffering and exporting images prior to transfer to grid. Grid-DICOM preserves the asynchronous nature of DICOM communication through direct communication based on dcm4che2 toolkit. The authors of [14] also developed a router that acts as a proxy and translates between the DICOM domain and grid protocol. This router has two interfaces and it is multithreaded to enable full duplex communication. DICOM messages are processed in memory buffers as they are streamed. Therefore, memory usage is minimized, allowing the transfer of very large-sized image datasets.

In the light of the existing research, we plan to follow the approach outlined in [14] in implementing DIPACS 2.0 gateway, which will transparently connect the local DICOM to the grid. The DIPACS 2.0 gateway will be developed for TR-GRID infrastructure, which is based on EGEE-2 grid relying on Globus [17]. We will implement DIPACS 2.0 gateway with GT4 and develop a new transport

protocol similar to Grid-DICOM by using dcm4che2 library as in DIPACS 1.0. Since DIPACS 2.0 will be based on grid infrastructure, there is no need for the Nameserver as in DIPACS 1.0. In this architecture, DIPACS 2.0 gateway transparently connects the local DICOM to the grid. In addition to data transfer facilities, DIPACS 2.0 gateway will include components for cataloging, replication, and security. We also plan to incorporate content-based query mechanisms based on ontologies to DIPACS 2.0.

74.4 Conclusion

In this study, we implemented a distributed PACS, DIPACS 1.0, to share DICOM objects between the domains of small-/medium-scale organizations. Considering the requirements of large-scale distributed medical image systems, we proposed a grid-based architecture, DIPACS 2.0. Currently, our research is ongoing to put DIPACS 2.0 into real life. As we progress in the implementation of DIPACS 2.0, we will share our experiences with the research community.

References

1. Tohme W G, Choi I, Vasilescu E, Mun S K (2009) The Evolution of distributed diagnsis: Teleradiology as a case study. Proc IEEE the 1st Distributed Diagnosis and Home Healthcare Conference 113 115.
2. Costa C, Fretias F, Pereira M, Silva A, Oliveria J L (2009) Indexing and retrieving DICOM data in disperse and unstructures archieves. Int J Cars 4:71 77.
3. Nema Committee (2008) Digital Imaging and Communications in Medicine (DICOM) Standard Version 3.0, National Electrical Manufactureres Association, PS3.1, PS3.3, PS3.4, PS3.6, PS3.7, PS3.8.
4. Regt D, Weinbrger E (2004) MyFreePACS: A free web based radiology image storage and viewing tool. AJR Am J Roentgenology 183:535 538.
5. Khludov S, Meinel C, Noelle G (2000) Internet oriented medical information system for DICOM data transfer, visualization and revision. Proc IEEE 13th Computer Based Medical Systems, CBMS 2000:293 296.
6. Blanquer I, Hernandez V, Mas F, Segrelles D (2004) Middleware grid for storing, retrieving and processing DICOM medical images. Workshop on Distributed Databases and Processing in Medical Image Computing (DIDAMIC).
7. Montagnat J, Glatard T, Lingrand D, Texier R (2006) Exploiting production grid infrastruc tures for medical images analysis. Proc the 1st Singaporean French Biomedical Imaging Workshop (SFBI'06), Singapore.
8. Sharma A, Pan T, Cambazoglu B, Gurcan M, Kurc T and Saltz J (2009) VirtualPACS A federating gateway to access remote image data resources over the grid. J Digit Imaging 22:1 10.
9. Onbay T U (2009) Access to medical images over distributed PACS systems, MSc. Thesis, Ege University, Turkey.
10. dcm4che2, dcm4che2 DICOM Toolkit, http://ww.dcm4che.org. Accessed 2 January 2009.

11. Erberich S G, Bhandekar M, Nelson M D, Chernevak A, Kesselman C (2006) DICOM grid interface service for clinical and research PACS: A Globus Toolkit web service for medical data grids. Int J Cars 1:87 105.
12. Erberich S G, Silverstein J C, Chervenak A, Schuler R, Nelson M D, Kesselman C (2007) Globus MEDICUS Federation of DICOM medical imaging devices into healthcare grids. Proc Stud Health Technol Inform 269 278.
13. Liu B J, Zhou M Z, Documet J (2005) Utilizing data grid architecture for the backup and recovery of clinical image data. Comput Med Imag Grap 29:95 102.
14. Vossberg M, Tolxdorff T, Krefting D (2008) DICOM image communication in Globus based medical grids. IEEE T Inf Technol B 12:145 153.
15. Mogoules F, Nguyen T M (2009) Grid resource management: Towards virtual and services compliant grid computing. Chapman & Hall/CRC, USA.
16. Krefting D, Bart J, Berenov K, Dzhimova O, Falkner J, Hartung M, Hoheisel A, Knoch T, Lingner T, Mohammed Y, Peter K, Rahm E, Sax U, Sommerfeld D, Steinke T, Tolxdorff T, Vossberg M, Viezens F, Weisbecker A (2009) MediGRID: Towards a user friendly secured grid infrastructure. Future Gener Comp SY 25:326 336.
17. TR Grid, http://www.grid.org.tr. Accessed 5 September 2009.

Chapter 75
Epileptic EEG: A Comprehensive Study of Nonlinear Behavior

Moayed Daneshyari, L. Lily Kamkar, and Matin Daneshyari

Abstract In this study, the nonlinear properties of the electroencephalograph (EEG) signals are investigated by comparing two sets of EEG, one set for epileptic and another set for healthy brain activities. Adopting measures of nonlinear theory such as Lyapunov exponent, correlation dimension, Hurst exponent, fractal dimension, and Kolmogorov entropy, the chaotic behavior of these two sets is quantitatively computed. The statistics for the two groups of all measures demonstrate the differences between the normal healthy group and epileptic one. The statistical results along with phase-space diagram verify that brain under epileptic seizures possess limited trajectory in the state space than in healthy normal state, consequently behaves less chaotically compared to normal condition.

Keywords EEG · Chaos · Brain activity · Epilepsy · Nonlinearity

75.1 Introduction

Electroencephalograph (EEG) is superposition signal of many action potential of neurons in cortex and a representation of the macroscopic behavior of brain. EEG is considered as an irregular time series and many research works have been conducted on investigating nonlinear analysis on EEG signals. Researchers have investigated the nonlinear effect on EEG of sleep cycles [1], perception [2 4], olfactory system [5, 6], epilepsy [7, 8], musical observation [9], encephalopathy [10], and Alzheimer disease [11].

M. Daneshyari (✉)
Department of Technology, Elizabeth City State University, Elizabeth City, NC 27909, USA
e mail: mdaneshyari@mail.ecsu.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 75,

In this study, we have extensively investigated the chaotic measures for normal healthy EEG and patient under epileptic seizures using five different chaotic nonlinear measures to understand the underline differences of epileptic EEG with the healthy one.

75.2 Methods of Analysis

We have analyzed the EEG data using different nonlinear measures to compare the healthy individual's EEG against epileptic EEG.

75.2.1 Lyapunov Exponent

Lyapunov exponent (LE) measures the sensitive dependence on the initial conditions by defining the rate of divergence of two nearby trajectories. A positive LE concludes the orbits are on a chaotic attractor.

Given time series of $v(t)$, and a point in the m-dimensional phase space, let us define the distance between the current and initial points $\{v(t), v(t+\ \Delta t), \ldots, v(t+(m-1)\ \Delta t)\}$ and $\{v(t_0), v(t_0+\ \Delta t), \ldots, v(t_0+(m-1)\ \Delta t)\}$ as $L(t_0)$. At a later time, t_k, initial length evolves to $L(t_k)$, LE is [12]

$$\mathrm{LE} = \frac{1}{t_M - t_0} \sum_{k=1}^{M} \log_2 \left(\frac{L(t_k)}{L(t_{k\ 1})} \right). \tag{75.1}$$

75.2.2 Correlation Dimension

If probability of two arbitrary points in the state space whose distance is less than ε is $P(\varepsilon)$, then correlation dimension (CD) is [13]

$$\mathrm{CD} = \lim_{\varepsilon \to 0} \frac{\log(P(\varepsilon))}{\log(\varepsilon)}. \tag{75.2}$$

75.2.3 Hurst Exponent

Hurst exponent (HE) measures the smoothness of the chaotic time series which is based upon asymptotic behavior of the rescaled range of the process. HE is defined as $\mathrm{HE} = \log(R/S)/\log(T)$, where T is duration of the data and R/S corresponds to rescaled range. As Hurst exponent increases the complexity of the system decreases resulting a more synchronized system.

75.2.4 Fractal Dimension

Fractal dimension (FD) is the simplest type of dimension Assume that $N(\varepsilon)$is the number of volume elements, that is, sphere, cube, etc., with diameter ε, which cover the attractor in the phase space. As $\varepsilon \to 0$, the sum of volume elements approaches the volume of the attractor. For a D-dimensional manifold, $N(\varepsilon) = k\varepsilon^{\ D}$. Hence FD is defined as [14]

$$\mathrm{FD} = \lim_{\varepsilon \to 0} \frac{\log(N(\varepsilon))}{\log(1/\varepsilon)}. \tag{75.3}$$

75.2.5 Kolmogorov Entropy

Entropy demonstrates the amount of information stored in a more general probability distribution. Kolmogorov entropy (KE) is computed [15] by evolving points very close to each other in the phase space to observe how fast they move apart. The time it takes for the points to move apart, T, defines KE, as $T = 2^{\ \mathrm{KE}\,t}$. Higher KE implies less predictability and higher chaotic behavior.

75.3 Results of Analysis

In this study, EEG was taken from a healthy individual (Group H) and an epileptic person (Group E). Both groups of signals contain 40 single channel segments. Duration about 5 s was selected from multichannel EEG recordings after visually inspecting for any artifacts. Signals in Group H are segments taken from surface EEG of the healthy individual when is relaxed with open eyes. The recording is done using standard 24-channel electrode placement scheme [16]. Signals in Group E are segments which was taken from an epileptic individual in his diagnostic period containing the seizure activities. The recording devices include an amplifier system, analog-to-digital convertor and the computer system.

The time series of four sample EEG signals for the healthy individual (Group H) and epileptic patients (Group E) are demonstrated in Fig. 75.1. Notice that the voltage range in epileptic EEG are from −0.5 to 1.0 mV, but for healthy EEG is from 0.05 to 0.1 mV, that is, the EEG amplitude is 10 times greater when is under epileptic seizures.

In Fig. 75.2, the strange attractors are shown by depicting the state space graph among the electrodes' voltage. It shows that the strange attractor in healthy EEG is more comprehensive and covers more areas of the space compared to epileptic EEG. Using more quantified measures, we now compare the state space of the two groups of EEG.

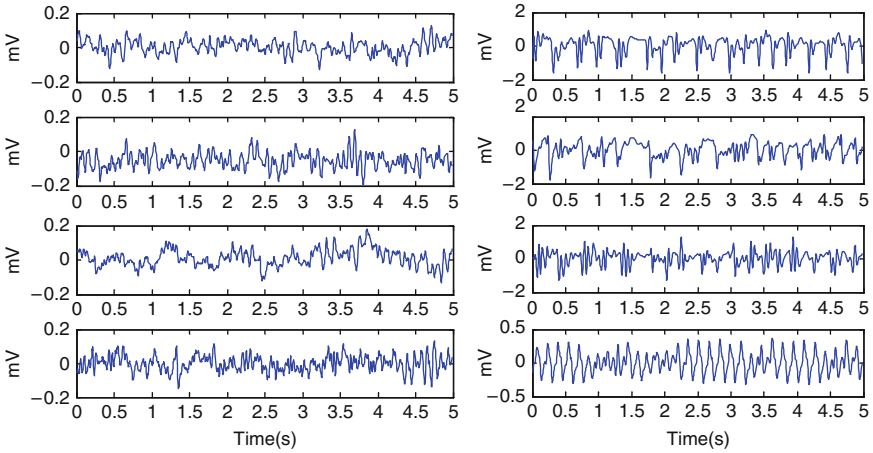


Fig. 75.1 EEG time series for the healthy person as control (*left*) and epileptic patient (*right*) from different electrodes, E1 4 (*from top to bottom*). The voltage range in epileptic EEG is 10 times larger than the voltage range in the healthy EEG

Table 75.1 demonstrates different measures for both healthy and epileptic EEG. For each measure, the highest, lowest, mean, and standard deviation for each group consisting 40 EEG signals is demonstrated. The result of the comparison of largest LE shows that the healthy group of EEG possesses higher value for the LE, implying to behave more chaotically. The CD is also compared in this table. To compute the more accurate CD, optimal embedded dimension is first computed as explained in [17]. CD increases as the embedding dimension of the system increases till it saturates at embedding dimension of 8 as seen in Fig. 75.3. Therefore, the optimal embedding dimension of 9 is considered.

The results demonstrate CD for epileptic EEG is less than that of healthy EEG. Applying the statistical *t*-test implies that the statistical results are significantly different. It verifies that epileptic seizures are low-dimensional chaotic state that emerges from nonepileptic activity which means that in the period of seizure the number of independent variables to describe the system decreases compared to the regular healthy brain activity.

Comparing the HE in the table shows that the brain acts less chaotically during the epileptic seizures. Statistical *t*-test shows that *p*-value for the two set of data to be significantly different is $p = 0.041$, which shows 96% confidence on significant difference between the two set of data. Computed FD demonstrate a drop when the epileptic seizures happen which is an indication of less activity for the brain during the seizures. Conducted statistical *t*-test shows a significant difference between the two set of data ($p = 0.033$). Finally, KE comparison shows the healthy EEG has higher KE than the epileptic one implying less activity and chaotic behavior of the brain during the epileptic seizures. Using the *t*-test, it implies two sets as significantly different data sets.

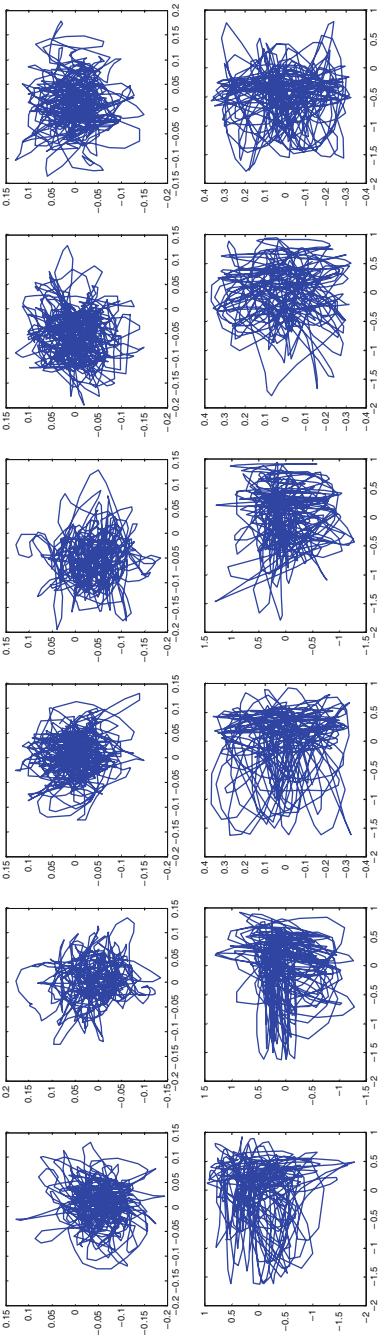


Fig. 75.2 Phase space for different electrodes for healthy (*top*) and epileptic (*bottom*) individual: (*from left to right*) E1 vs. E2, E1 vs. E3, E1 vs. E4, E2 vs. E3, E2 vs. E4, E3 vs. E4

Table 75.1 Different measures for both healthy EEG group (H) and epileptic EEG group (E): Largest Lyapunov exponent (p 0.078), correlation dimension (p 0.001), Hurst exponent (p 0.041), fractal dimension (p 0.033), and Kolmogorov entropy (p 0.001)

Measures		Highest	Lowest	Mean	Standard deviation
Largest LE	Healthy	0.249	0.182	0.213	0.024
	Epileptic	0.239	0.176	0.204	0.021
Correlation dimension	Healthy	5.046	4.707	4.895	0.127
	Epileptic	3.651	3.120	3.391	0.201
Hurst exponent	Healthy	0.384	0.249	0.309	0.043
	Epileptic	0.417	0.296	0.359	0.042
Fractal dimension	Healthy	1.982	1.688	1.865	0.019
	Epileptic	1.510	1.205	1.329	0.011
Kolmogorov entropy	Healthy	0.648	0.554	0.609	0.031
	Epileptic	0.501	0.423	0.455	0.041

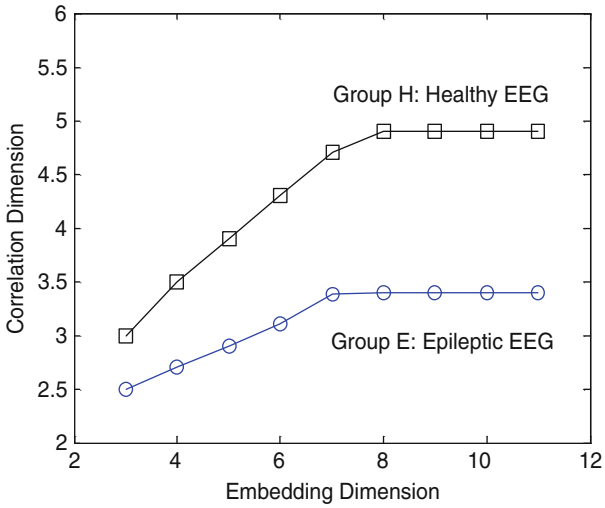


Fig. 75.3 Relationship between CD and embedding dimension: CD saturates at embedding dimension 8

75.4 Conclusion

In this study, we have conducted nonlinear measures such as largest Lyapunov exponent, correlation dimension, Hurst exponent, fractal dimension, and Kolmogorov entropy to understand the differences between the healthy brain activity and the activity of epileptic patient during the epileptic seizures. The nonlinear measures are found by conducting several algorithms on all EEG signals for each group

and the statistics are compared for the two set of data. The results demonstrate that during the epileptic seizures brain falls into a less activity period with low-dimensional chaos.

References

1. Babloyantz A, Salazar JM, Nicolis C (1985) Evidence of chaotic dynamics of brain activity during the sleep cycle. Phys Lett A 111:152 156.
2. Freeman WJ (1991) The physiology of perception. Sci Am 264(2):78 85.
3. Freeman WJ (1992) Tutorial on neurobiology: From single neurons to brain chaos. Int J Bifurc Chaos 2:451 482.
4. Guevara MR, Glass L, Mackey MC, Shrier A (1983) Chaos in neurobiology. IEEE Trans Syst Man Cybern 13:790 798.
5. Aradi I, Barna G, Erdi P, Grobler T (1995) Chaos and learning in the olfactory bulb. Int J Intell Syst 10:89 117.
6. Baird B (1986) Nonlinear dynamics of pattern formation and pattern recognition in the rabbit olfactory bulb. Phys D 22:150 175.
7. Babloyantz A, Destexhe A (1986) Low dimensional chaos in an instance of epilepsy. Proc Natl Acad Sci USA 83:3513 3517.
8. Iasemidis LD, Sackellares JC (1996) Chaos theory and epilepsy. Neuroscience 2:118 126.
9. Birbaumer N, Lutzenberger W, Rau H, Braun C, Mayer Kress G (1996) Perception of music and dimensional complexity of brain activity. Int J Bifurc Chaos 6:267 278.
10. Stam CJ, van der Leij HE, Keunen RW, Tavy DL (1999) Nonlinear EEG changes in postanoxic encephalopathy. Theory Biosci 118:209 218.
11. Adeli H, Ghosh dastidar S, Dadmehr N (2008) A spatio temporal wavelet chaos methodology for EEG based diagnosis of Alzheimer's disease. Neurosci Lett 444:190 194.
12. Wolf A, Swift JB, Swinney LH, Vastano JA (1985) Determining Lyapunov exponent from a time series. Phys D 16:285 317.
13. Grassberger P, Procaccia I (1983) Characterization of strange attractors. Phys Rev Lett 50:346 349.
14. Parker TS, Chua LO (1987) Chaos: A tutorial for engineers. Proceedings of the IEEE, Special Issue on Chaotic Systems, pp. 982 1008.
15. Kantz H, Schreiber T (1997) Nonlinear time series analysis. Cambridge University Press, Cambridge.
16. Daneshyari M, Kamkar LL, Daneshyari M (2009) Extracting low dimension nonlinearity in EEG of epilepsy. BIOCOMP 422 426.
17. Takens F (1981) Detecting strange attractors in turbulence. In: Dand DA, Young LS (eds), Dynamical systems and turbulence. Springer, Berlin.

Chapter 76
Computational Energetic Model of Morphogenesis Based on Multi-agent Cellular Potts Model

Sébastien Tripodi, Pascal Ballet, and Vincent Rodin

Abstract The Cellular Potts Model (CPM) is a cellular automaton (CA), developed by Glazier and Graner in 1992, to model the morphogenesis. In this model, the entities are the cells. It has already been improved in many ways; however, a key point in biological systems, not defined in CPM, is energetic exchange between entities. We integrate this energetic concept inside the CPM. We simulate a cell differentiation inside a growing cell tissue. The results are the emergence of dynamic patterns coming from the consumption and production of energy. A model described by CA is less scalable than one described by a multi-agent system (MAS). We have developed a MAS based on the CPM, where a cell agent is implemented from the cell of CPM together with several behaviours, in particular the consumption and production of energy from the consumption of molecules.

Keywords Morphogenesis · Cellular potts model · Multi-agents systems

76.1 Introduction

The Cellular Potts Model (CPM) is a cellular automaton (CA) developed by Gla-zier and Graner [6] to model different phenomena which occur during the morphogenesis [3, 9]. The dynamics of CPM is based on a minimisation of energy. The entities of this system are called cells and are characterised by a volume, a surface and a type. They are in interaction via contact energies and via the restricted access to grid sites.

S. Tripodi (✉)
European University of Brittany UEB UBO, EA 3883 LISyC (in virtuo) 20 av Le Gorgeu CS, 93837 29238 Brest Cedex, France
e mail: sebastien.tripodi@univ brest.fr

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 76,

The CPM can be improved to model the morphogenesis in a more realistic way [1]. Usually, the CPM is defined by a CA. In this paper, we describe the CPM as a multi-agent system (MAS), where the entities are reified. This allows to enhance the scalability of CPM. The dynamics in MAS is not given by a global function but via the interactions of the entity behaviours executed according to a scheduler.

The multi-agent approach eases the implementation of the following specific cell behaviours: secretion and consumption of molecules, transformation of molecules into energy, migration over a gradient of molecules, cell division, cell differentiation and cell death. The closest work to our approach is probably Com-pucell3D [3], a software which implements the CPM and other behaviours to model the morphogenesis. However, the notion of energy from the consumption of molecules used for the cell maintenance and division is not present in the CPM.

This paper is organised as follows. A multi-agent view of CPM is given in Sect. 2. In Sect. 3, we describe the MorphoPotts agent which represents a cell defined in the CPM to which we add the previously mentioned behaviours. Using this, we simulate a model of embryogenesis based on a Darwin theory at cellular level [7], and we observe the emergence of relevant dynamic patterns in Sect. 4. Finally, we conclude in Sect. 5.

76.2 Cellular Potts Model

In this section, we firstly present the CPM described by Graner and Glazier [6]. Secondly, we propose an optimised implementation of CPM based on a multi-agent approach.

To describe the CPM, we present in this order the notations, the state of system and the transition function:

- A grid is denoted by Sx, a site of this grid by (i,j) and the value of a site by $sx_{i,j}$. A cell is denoted by $C\sigma p$ with $\sigma \in [1,N]$, where N is the number of cells and p the type of cell. $C\sigma p$ has a target volume (resp. surface) $V\sigma t$ (resp. $S\sigma t$) and a current volume $V\sigma$ (resp. $S\sigma$). The contact energies are recorded in a matrix T such that $T_{\sigma,\sigma'}$(or $T_{p,p'}$) is the contact energy between $C\sigma p$ and $C\sigma' p'$.
- A state of system is a grid[1] Sx of D dimensions (here $D = 2$), where each site (i,j) is filled by a particle of the cell $C\sigma p$, i.e $sx_{i,j}$ is equal to σ.
- The transition function between the state Sa and Sb is verified if Sb is the state Sa, where the value of one site has been replaced by the value of a neighbouring[2] site and if the probability of transition between the state Sa and Sb is accepted. The probability of transition (a Monte Carlo probability) and the energy function (depends on the volume and the surface of each cell, and the contact energies between the cells) used here, are described in [6].

[1]The continuous case is also defined [5]

[2]Here the neighbours of a site (i,j) are the sites $(i + 1,j)$, $(i,j + 1)$, $(i \quad 1,j)$, $(i,j \quad 1)$

A multi-agent view of CPM is given by using the vowel approach [4], i.e. MAS = Agent + Environment + Interaction + Organisation (here, no specific organisation is imposed). We describe these concepts in this order:

- Here, an agent is a cell $C\sigma p$ which can modify its local environment by changing the value of its neighbouring sites by σ. This is implemented by the method replace site (see Table 76.1). Each cell knows:
- Its membrane $M\sigma$, where $M\sigma = \{<(i,j),L> \mid sx_{i,j} = \sigma \ \wedge \ L = \{(i',j') \in neighbour(i,j) \mid sx_{i',j'} \neq \sigma\} \neq \emptyset\}$, i.e. the set of pair $<(i,j),L>$,where the site (i,j) is on the membrane of the cell and L the set of sites outside the cell.
- Its target and current volume, its target and current surface, and so its energy volume (Ev_σ) and surface (Es_σ) like the difference power two between the target and current value.
- Its contact energy (Ec_σ) between the different cells.
- The environment is a grid, it is the background where the interactions (direct or not) between cells occur. The environment is always initialised by one cell $C0p$ which models the medium.
- The interactions between cells are of two types. Firstly, some indirect interactions due to the concurrence for the available places on the environment. Secondly, the cells are in direct interactions *via* the contact energy between them.

The simulation step is the following:

1. We compute the energy (Ea) of the current state (Sa).
2. We choose a random site (i,j) and a random neighbouring site (i',j') of (i,j) (the sites chosen are on the membranes of two different cells).
3. $sa_{i',j'}$ is saved into σ and the method replace site is called from the cell $Csa_{i,j}p'$ with the site (i',j').
4. We compute the energy (Eb) of the current state (Sb).
5. If $Ea \leq Eb$ and if the probability of transition (see transition function defined previously) is not accepted, then the method replace site is called from the cell $C\sigma p$ with the site (i',j').

76.3 MorphoPotts

A MorphoPotts agent keeps the properties of the cell defined in the CPM, but it also has an internal energy E which results of the consumption of molecules. This agent is very close to MorphoBlock [2] but the core of MorphoBlock is a pixel. In this section, we describe the MorphoPotts and the simulation step of couple CPM MorphoPotts. The abilities of a MorphoPotts $C\sigma p$ are given by the following methods:

- $secr(arg,Y)$ secrets a gradient of molecules Y of radius arg at gravity center of $C\sigma p$.
- $cons(\{max,arg\},Y)$ consumes molecules Y if the energy of $C\sigma p$ is lower than max. $cons(\{max,arg\},Y)$ has an inverse effect compared to $secr(arg,Y)$.

Table 76.1 Definition of the method replace_site

Require: (i, j) the site to replace by the cell $C\sigma p$			*Ensure*: void
/**Deleting of the value of site (i, j) and updating of the energies about the site (i, j)*/			
1	$\sigma' = sx_{i,j}$; $sx_{i,j} = null$;	$V\sigma' = V\sigma' - 1$;	$Ev_{\sigma'} = (V\sigma' - V\sigma'_t)^2$;
2	$S\sigma' = S\sigma' - \#L$ *where* $< (i, j), L > \in M\sigma'$;	$Es_{\sigma'} = (S\sigma' - S\sigma'_t)^2$;	
3	$Ec_{\sigma'} = Ec_{\sigma'} - \sum_{(i',j') \in L} T_{\sigma', sxi', j'}$;	$M\sigma' = M\sigma' - <(i, j), L>$;	
/* Updating of the value of site (i, j), and updating of the energies about the site (i, j) and the cell $C\sigma p$*/			
4	$neighbour = \{(i', j') \in neighbour(i, j)\, sx_{i',j'} \neq \sigma\}$;	$sx_{i,j} = \sigma$;	$V\sigma = V\sigma + 1$;
5	$Ev_\sigma = (V\sigma - V\sigma_t)^2$;	$S\sigma = S\sigma + \#\, neighbour$;	$Es_\sigma = (S\sigma - S\sigma_t)^2$;
6	$Ec_\sigma = Ec_\sigma + \sum_{(i',j') \in neighbour} T_{\sigma, sxi', j'}$;		
7	**if** *(*$\#\, neighbour > 0$*)* **then** $M\sigma = M\sigma + <(i, j), neighbour>$; **end if**		
/**Updating of the energies and the membrane of all neighbouring cells about the site (i, j)*/			
8	**for** $(i', j') \in neighbour(i, j)$ **do**		
9	$neighbour' = \{(i'', j'') \in neighbour(i', j')\, sx_{i',j} \neq sx_{i'',j''}\}$;		$\sigma' = sx_{i',j'}$;
10	**if** $(<(i', j'), L> \in M\sigma')$ **then**	$S\sigma = S\sigma' - \#L$;	$Es_{\sigma'} = (S\sigma' - S\sigma'_t)^2$;
11	$Ec_{\sigma'} = Ec_{\sigma'} - \sum_{(i'',j'') \in L} T_{\sigma', sxi'', j''}$	$M\sigma' = M\sigma'$-$<(i', j'), L>$;	**end if**
12	**if** $(\#\, neighbour' > 0)$ **then**	$M\sigma' = M\sigma' + <(i', j'), neighbour'>$;	
13	$Ec_{\sigma'} = Ec_{\sigma'} + \sum_{(i'',j'') \in neighbour'} T_{\sigma', sxi'', j''}$;	$S\sigma' = S\sigma' + \#\, neighbour'$;	$Es_{\sigma'} = (S\sigma' - S\sigma'_t)^2$;
14	**end if**	**end for**	

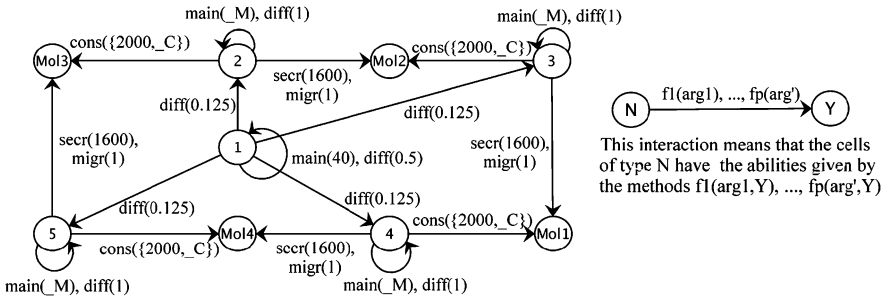


Fig. 76.1 Interaction graph used in the two simulations. In the simulation 1: M 10 and C 400. In the simulation 2: M 30 and C 150. The migration methods defined in this graph favour MorphoPotts of same type to cluster

- *migr*(*arg*,*Y*) gives the ability to MorphoPotts to migrate towards the molecules *Y* according to *arg*. For this, a new energy (E_{migr}) is added to CPM.
- $E_{migr} = -arg * \sum_{(i,j) \in M\sigma} nbMol((i,j),Y)$, where *nbMol*((*i*,*j*),*Y*) is the number of molecules *Y* on the site (*i*,*j*).
- *trans*(*arg*,*Y*) transforms the consumed molecules in energy. In this paper, for each consumed molecule the energy is incremented of 1.
- *diff*(*arg*,*Y*) associates a probability of differentiation *arg* to a cell type *Y*.
- *div*({*b*,*En*,*cost*},*t*) gives the ability to the cell to divide. If *b* is true $C\sigma p$ divide along an axis (vertical or horizontal), $C\sigma' p'$ is created according to the probabilities of differentiation. The energy of the cell $C\sigma' p'$ is equal to e' such that $(t',e') \in En$ and the energy of the cell $C\sigma p$ is equal to $E - e' - cost$.
- *main*(*arg*,*t*) decrements of *arg* the energy (maintenance).
- *die*() implies that the MorphoPotts lost all its abilities and does not generate energy in the CPM.

The simulation step of couple CPM MorphoPotts is as follows:

1. Let *i* be equal to 0 and let *n* be equal to membrane size of all MorphoPotts.
2. While *i* is lower than *n*
   (a) A step of CPM is run. For the two MorphoPotts chosen in the step of CPM, their method of division is called. *i* is incremented by 1.
3. All cells execute their method of maintenance, their method of secretion and their method of consumption.
4. For each cell, if the energy is lower than 0, the cell executes its method of death.

76.4 Simulation: Darwin Theory at Cellular Level

In this section, we present two simulations which use the Darwin theory at cellular level [7]. A model of this theory can be found in [8], where a cell (represented by 1 pixel) secrets and consumes molecules in an environment of dimension 50 × 50.

The cells can differentiate (in two types) with a probability which depends of the neighbouring cells. The cells can divide if the quantity of consumed molecules is sufficient. The first type secrets a molecule, which the second type consumes and vice versa for the second type. This inter-proliferation according to different parameters leads to: finite growth, growing cancer. In our simulations, the stochastic cell differentiation is modelled by initialising the environment by one MorphoPotts of type 1, which divides and can differentiate in four different cell types. The natural selection is modelled by the notion of energy. If a MorphoPotts finds molecules, it can increase and it can divide (via the energy), otherwise it will die.

Two simulations[3] are presented in an environment 1,000 $\times$ 1,000, with an cycle inter-dependence and with the same CPM parameters (the cells tend to square 7 $\times$ 7, and the energy of contact between two cells of different types is 0, otherwise 10). They are different by two parameters (see Fig. 76.1). The MorphoPotts can also be divided (not defined in Fig. 76.1), if their current volume is higher than 80% of the target volume and if their energy is higher than 2,000. The energy of MorphoPotts that divides is equally distributed with the MorphoPotts created. In these simulations, the MorphoPotts of type 1 (resp. 2, 3, 4, 5) is red (resp. green, cyan, yellow, blue).

The results of the simulation 1 and 2 are given in the Fig. 76.2. We can see three steps in this morphogenesis. The two first steps are the same in the two simulations. The first step (see Fig. 76.2a,b) is the cell differentiation and the natural selection. The MorphoPotts of type 1 divides and randomly differentiates in four types. This leads to the formation of tissues. The second step is the sorting (see Fig. 76.2b,c), the tissues are sorted by the simple fact of the death (the cells do not find molecules) and the division (the cells find molecules).

The third step in the simulation 1 is the proliferation and the emergence of pattern (see Fig. 76.2c). Here, a spiral proliferation emerges (not imposed in the description). We can see that the tissue renewal is continuous, i.e. the tissue is not static (see the comment written on the Fig. 76.2c). In this simulation, the proliferation seems infinite.

The third step in the simulation 2 is a finite growing (an important criterion in the embryogenesis) and the emergence of pattern (see Fig. 76.2d f). This can be explained by the fact that the consumption of molecules has been reduced and the cost of maintenance has been increased compared to the simulation 1 (see Fig. 76.1). The number of cells by tissue varies between two thresholds (see Fig. 76.2e,f). So an equilibrium between the cell death and the cell division emerges. This shows that a stochastic cell differentiation and a natural selection can be sufficient to generate a finite growing of a dynamic cell tissue, and so to model the embryogenesis.

[3]The pc used is an intel core2 quad 2.83 GHz with 3 GB of memory. The videos of these simulations are available at http://pagesperso.univ brest.fr/~tripodi/private/springer/videos.html

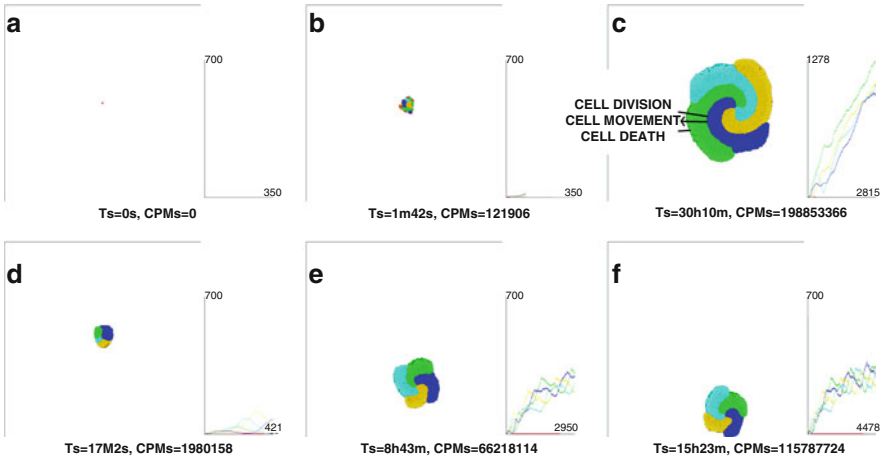


Fig. 76.2 Spiral proliferation with an infinite and finite proliferation. *Ts* is the time of simulation and CPMs is the number of CPM steps. Figure 2(**a**) (**c**) shows the simulation 1 (an infinite proliferation), and Fig. 2(**d**) (**f**) the simulation 2 (finite proliferation). In these figures, the cells whose colour is dark are dead. The graphs on the right represent the number of cells as a function of the number of simulation step

76.5 Conclusion

In this paper, we have described how adding the energetic exchanges between cells in CPM (via the MorphoPotts agent defined in this paper) improve the realism of simulations. By adding this key point happening inside multi-cellular organisms, we have shown that morphogenesis is strongly influenced by energetic exchange especially for the tissue renewal, the organism stability and the robustness of developmental patterns.

The MorphoPotts has been tested thanks to simulations of a growing cell tissue using a cellular Darwinian theory. The results show the emergence of patterns, some with a finite but dynamic growing tissue. For the same model, we observe very similar emerging patterns in different simulations. This indicates that even for a stochastic cell differentiation and a natural selection, the organism global structure remains the same.

Acknowledgments We thank the Region Bretagne for its financial contributions.

References

1. Anderson ARA, Chaplain MAJ, Rejniak KA (2007) Single Cell Based Models in Biology and Medicine. Birkhauser

2. Ballet P, Tripodi S, Rodin V (2009) Morphoblock programming: a way to model and simulate morphogenesis of multicellular organisms. J Biol Phys Chem 9:37 44
3. Cickovski T, Aras K, et al. (2007) From genes to organisms via the cell: a problem solving environment for multicellular development. Comput Sci Eng 9:50 60
4. Demazeau Y (1995) From interactions to collective behaviour in agent based systems. Proc Eur Conf Cogn Sci, St Malo France 117 132
5. Glazier JA, Graner F (1993) Simulation of the differential adhesion driven rearrangement of biological cells. Phys Rev E 47:2128 2154
6. Graner F, Glazier JA (1992) Simulation of biological cell sorting using a two dimensional extended Potts model. Phys Rev Lett 69:2013 2016
7. Kupiec JJ (1997) A Darwinian theory for the origin of cellular differentiation. Mol Gen Genet 255:201 208
8. Laforge B, Guez D, Martinez M, Kupiec JJ (2005) Modeling embryogenesis and cancer: an approach based on an equilibrium between the autostabilization of stochastic gene expression and the interdependence of cells for proliferation. Prog Biophys Mol Biol 89:93 120
9. Marée S (2000) From pattern formation to morphogenesis. PhD thesis, Utrecht University

Chapter 77
Standardizing the Next Generation of Bioinformatics Software Development with BioHDF (HDF5)

Christopher E. Mason, Paul Zumbo, Stephan Sanders, Mike Folk, Dana Robinson, Ruth Aydt, Martin Gollery, Mark Welsh, N. Eric Olson, and Todd M. Smith

Abstract Next Generation Sequencing technologies are limited by the lack of standard bioinformatics infrastructures that can reduce data storage, increase data processing performance, and integrate diverse information. HDF technologies address these requirements and have a long history of use in data-intensive science communities. They include general data file formats, libraries, and tools for working with the data. Compared to emerging standards, such as the SAM/BAM formats, HDF5-based systems demonstrate significantly better scalability, can support multiple indexes, store multiple data types, and are self-describing. For these reasons, HDF5 and its BioHDF extension are well suited for implementing data models to support the next generation of bioinformatics applications.

Keywords DNA sequencing · Bioinformatics · Data management, · highpeformance computing · HDF

77.1 Introduction

"Next Generation Sequencing (NGS)" technologies are clearly powerful. However, adoption is severely limited by the massive amounts of data produced combined with the complex and intensive multistep data processing needed to convert images into sequences, align the sequences, and convert the resulting data into quantitative and qualitative information that can be compared between different samples [1, 2]. The NGS community is recognizing the need to use more structured binary file formats to improve data storage, access, and computational efficiency. SAMtools

T.M. Smith (✉)
Geospiza Inc, 100 West Harrison N Tower 330, Seattle, WA 98119, USA
e mail: todd@geospiza.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 77,

(Sequence, Alignment, and Map) and the BAM (binary equivalent) file format [3] were developed to address this need. These formats, however, do not adequately meet NGS data handling requirements. For example, the SAMtools standards, while laudable, cannot store the alignment of one read relative to multiple references, and BAM simply replaces the SAM text file with a binary file and fails to address issues related to data redundancy and defining complex relationships in the data and between samples.

An ideal data format for NGS would be one that could reap the benefits of a structured binary file yet remain sufficiently general and extensible in order to store an expanding variety of genomic and other data, allowing hierarchical parallelized analysis for a broad set of applications. Such a format should have long-term stability and be developed by data management experts. HDF5 (Hierarchical Data Format [4]) meets these requirements and, therefore, we propose that HDF5 be used as a standard for implementing NGS data models in binary formats.

We have developed BioHDF, a first-draft data model, and tools to address common data storage, integration, and access requirements based on HDF5. We have measured the performance of loading data into the model and accessing portions of datasets for viewing summarized results and drill-down views of details. With this implementation and tool set, we are able to compare alignments between samples, alignments generated from multiple algorithms, and annotations from multiple sources using an example system from whole transcriptome (WT) analysis. To support the concept of using HDF5 as a standard bioinformatics format, we have also prototyped an integration with the Bowtie alignment program [5].

77.2 Methods

77.2.1 HDF5 Data Organization and Access

The HDF5 data model is simple. Data are organized in *groups* and *datasets*. HDF5 groups provide a mechanism for organizing collections. HDF5 datasets are multidimensional arrays of simple or complex data types. Groups and datasets may have associated attributes, which contain user-specified metadata. HDF5 can also store large arrays in chunks that can be compressed and accessed randomly, resulting in fast on-the-fly access to portions of an array while minimizing data storage requirements. HDF5 data can also be organized into multiple files that can be accessed through the HDF5 library as if they were a single file. Thus, data organization can be optimized by separating data in different ways without major changes to the surrounding software.

In the initial BioHDF model (Fig. 77.1), *Sequence Data* are stored in two datasets within the “sequences” group. Data are stored as a table with the bases in

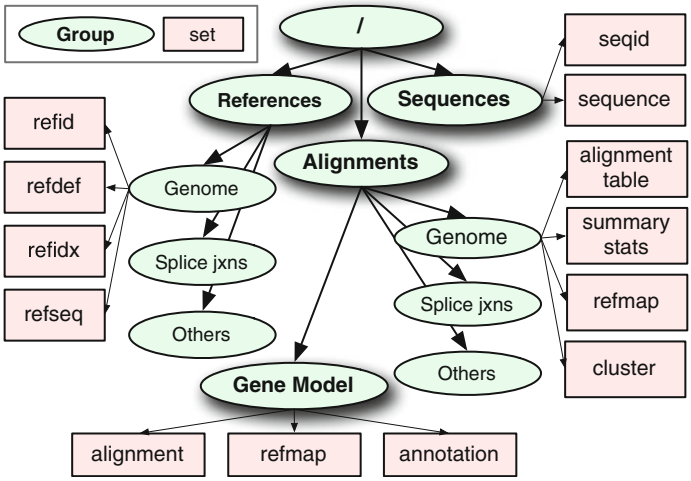


Fig. 77.1 The BioHDF Data Model. The sequences group (*oval*) holds datasets (*rectangles*) that store identifiers and read data. Reference subgroups hold datasets to manage identifiers (refid), definition lines (refdef), an index (refidx), and sequences as single character strings (refseq). Alignment subgroups, one per alignment process, hold datasets that include an alignment table, summary data, an index (refmap), and read clustering, to support retrieving reads by alignment position and other queries. Three datasets in the Gene Model integrate alignment data with reference positions (refmap) and annotations

columns and reads in the rows. This format is suitable for NGS reads. These data can be of either fixed (AB SOLiD and Illumina GA) or variable (454 and Helicos) length. *Reference Sequence Databases* are stored together as very large concatenated character strings with an index to access individual sequences. Each reference dataset is contained in a subgroup within the "references" group. This model stores several reference databases in a common HDF5 file with other data, but a particular reference sequence database could be written once to a separate BioHDF file if needed. *Alignment Data* are stored in the same HDF5 file as subgroups, named for the particular alignment operation, within the "alignments" group. This organization allows a user to determine if each read has been aligned to any number of reference sequences. *Annotations* that describe known exons, coding regions, and untranslated regions from databases, like RefSeq Genes, can be imported into a single BioHDF file from the GFF3 format [6], and indexed for fast retrieval and visualization alongside aligned reads.

Extensive C-based libraries that include a new BioHDF library and libraries provided with the HDF5 distribution support BioHDF I/O. These libraries are accessed through command-line tools that support the general data import, export, and query functions. In the case of WT analysis, specific computing steps are executed in a pipeline that combines the command-line tools together with open-source sequence alignment algorithms and visualization tools.

77.2.2 Performance Tests

HDF5 storage and I/O performance were compared to the SAM/BAM system, under a commonly used WT data analysis pipeline scenario, to evaluate the two different binary methods for storing and accessing NGS data. Three files containing tenfold increment of data ranging from 1 to 100 million Illumina reads (fastq format) were aligned independently to the human genome and RefSeq transcripts [7]. The resulting SAM files were loaded into BioHDF and BAM using BioHDF tools and SAMtools, respectively, to measure relative import/export times, compression ratios, and overall scalability. Additionally, the Bowtie algorithm (0.10.0.2) was modified to write alignments directly to BioHDF to determine the feasibility of using HDF5 as a native file format for bioinformatics applications.

Test data included data from an RNA-Seq study [8], reads from standard reference RNA samples obtained from the NCBI short-read archive (SRA002355.1), and reads obtained from the SEQC (Sequencing Quality Control) Consortium (www.fda.gov). In all cases, the same input files were used for comparisons. Time measurements were calculated as the clock time using the UNIX "time" command on a quad-core Xeon 3.06 GHz processor with 5GB of SDRAM.

77.3 Results

77.3.1 System Scalability

The first test, comparing a single alignment operation producing a single SAM file, demonstrated small performance advantages for the BioHDF system. Compression ratios were comparable and the BioHDF system averaged 1.6 times faster data import and export speeds over the SAM/BAM system. When tested under a more realistic scenario that involved aligning reads to the human genome, followed by aligning the same data to RefSeq, the BioHDF tools imported the data twice as fast into a single BioHDF file as SAMtools imported the same data into two separate BAM files, which combined were double the size of the BioHDF file. In BioHDF, reads are efficiently separated from alignment values, whereas the SAM/BAM system stores copies of entire sets of reads with each independent alignment operation. As more comparisons to individual reference databases are made, BioHDF scales incrementally, but SAM/BAM scales by at least a factor of the total number of reads plus alignment values. We were unable to compare the performance of including alignments to splice junction databases, because unlike BioHDF, SAMtools and BAM were unable to deal with the large number of short reference sequences contained in the database.

The HDF-enabled Bowtie demonstrated an overall 5% improvement in performance when compared to writing alignments to an intermediate file and then

importing data into HDF. Because data are written directly to HDF, additional file parsing and data-loading steps are eliminated, which can save significant time and reduce overall system load maintenance burden because fewer terabyte-sized files need to be managed as datasets grow to a billion or more reads.

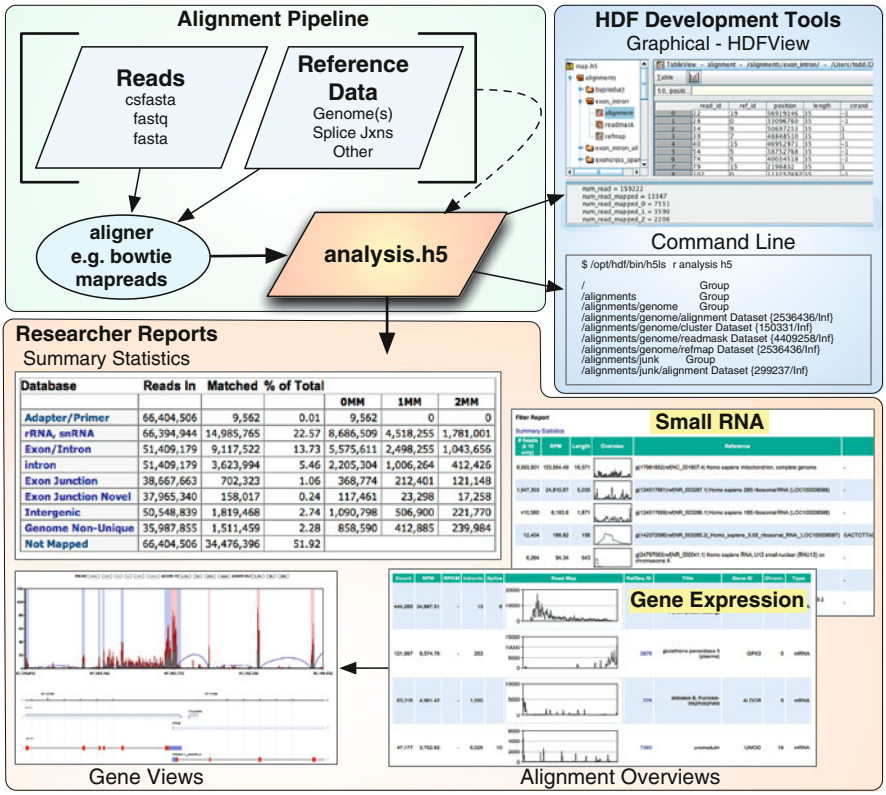
77.3.2 Data Integration

Performance-optimized alignment tools will not help biologists do their work unless these tools are supported by software systems that can integrate computation with rapid and easy data access. In WT analysis, it is common to observe alignments for 20,000 transcripts or more. In these alignments, read density can reveal the transcript's exon and intron structure, and when alignments to databases of splice junctions are layered in, novel splice forms (isoforms) can be identified [8]. As the goal of such experiments is to compare results across multiple samples, researchers also need to discover which isoforms and alleles are expressed under different conditions. Thus, summary displays and interactive views require that high-level alignment data be integrated with original read data and computed summary statistics. Alignments must be displayed with different levels of detail and it must be possible to combine alignment information from multiple samples along with annotated information about known gene structure and patterns of variation from multiple sources.

HDF5's data integration capabilities were demonstrated by aligning WT reads to the human genome, RefSeq transcripts, and a database of exon exon splice junctions taken from AceView [9]. All alignments and RefSeq annotations were imported into a single HDF5 file. The number of reads aligning to each database and the average number of mismatches observed between the reads and reference are summarized in HTML reports that link to drill-down views and external resources (Fig. 77.2). In these reports, read densities showing exon structure are displayed as thumbnail graphics, which link to dynamic reports that combine alignment data with additional annotations. Each time a user zooms or scrolls the viewer, indexed data stored in the HDF5 file are accessed and formatted for display. Using these reports, biologists can quickly learn about the overall quality of their data, and observe global and fine-grained views that show transcript isoforms and variation in the data. This system also supports additional computations such as normalizing alignment counts for entire transcripts and for individual exons to quantify gene expression and alternative splicing between samples.

77.4 Discussion

This work demonstrates five important attributes that make HDF5 well-suited for managing and working with bioinformatics data: (1) the HDF platform cleanly separates data modeling from implementation, (2) HDF5 and the HDF I/O library



Database	Reads In	Matched	% of Total	0MM	1MM	2MM
Adapter/Primer	66,404,506	9,562	0.01	9,562	0	0
rRNA, snRNA	66,394,944	14,985,765	22.57	8,686,509	4,518,255	1,781,001
Exon/Intron	51,409,179	9,117,522	13.73	5,575,611	2,498,255	1,043,656
intron	51,409,179	3,623,994	5.46	2,205,304	1,006,264	412,426
Exon Junction	38,667,663	702,323	1.06	368,774	212,401	121,148
Exon Junction Novel	37,965,340	158,017	0.24	117,461	23,298	17,258
Intergenic	50,548,839	1,819,468	2.74	1,090,798	506,900	221,770
Genome Non-Unique	35,987,855	1,511,459	2.28	858,590	412,885	239,984
Not Mapped	66,404,506	34,476,396	51.92			

Fig. 77.2 BioHDF based analysis system. Existing HDF tools can be used to view data (*upper right panel*). HDFView is used to view graphically a model's hierarchy and the data within each dataset. Command line tools display the data in a familiar UNIX style directory fashion. In Geospiza's GeneSifer software, BioHDF provides the infrastructure to create reports and combine alignment data with known annotations in using third party tools like GenomeTools graphics [15]

create a platform for building high-performance data management systems, (3) HDF5 and the HDF tools make data models self-describing, (4) systems developed on the HDF platform are extensible, and (5) HDF technologies shorten development time and decrease the maintenance burden.

Software performance is the primary motivation for using binary file formats and BAM is one step toward a binary standard in NGS. However, the SAM/BAM data model is inefficient and unable to support a typical NGS data analysis pipeline where sets of reads are independently aligned against several reference databases and the resulting data are combined with annotations. Because BAM stores the reads and alignments for each alignment operation in independent files, BAM and its APIs would need to be redesigned in order to meet the broad set of application requirements and performance levels already met by BioHDF.

Beyond performance, HDF technologies overcome many of the software development challenges associated with building scalable data processing systems. Because HDF5 separates the data model and its implementation, datasets are organized in a hierarchy of groups, and a well-defined data model specifies this general structure. All information about an experiment can be stored within this data model, instantiated either as a single HDF5 file or as a set of related files that are accessed together. Software built to interact with the HDF5 data model is not concerned with low-level file formats. Moreover, unlike most binary file formats that are described either through an API or external documentation, data stored in HDF5 are self-describing and can be read using command line tools and GUI tools, like HDFView, so groups adopting this format can realize a significant reduction in maintenance risk and burden. Thus, systems built on HDF5 will be easier to extend and require less maintenance as new NGS applications emerge.

The bioinformatics community is at the start of a new journey in scalable high-performance computing and has an important "buy versus build" decision to make. The BAM file has demonstrated the value of binary formats for NGS computing. However, support needs will grow and the groups developing and using BAM will need to decide how to best support their technology. HDF technologies have stood the test of time and considerable resources and experience exist to support new applications. The HDF Group, a nonprofit organization, supports the HDF5 file format and access libraries. HDF5 is used widely in the physical sciences community (EOS, CGNS, NeXus [10 12]), supports petabyte scale data management projects, and is the underlying format in popular programming environments (netCDF, MATLAB [13, 14]).

Based on this work and predicted future bioinformatics needs, the bioinformatics community would benefit by adopting existing HPC technologies, like HDF5, and we recommend that HDF5 and BioHDF be used as a standard implementation layer to support bioinformatics applications.

Acknowledgments We are grateful to Sandra Porter and Eric Smith for reviewing the manuscript and providing feedback. Award Number R42HG003792 from the National Human Genome Research Institute supported this work. BioHDF, tools, and additional details can be obtained from the BioHDF project site (http://www.biohdf.org).

References

1. McPherson JD (2009) Next Generation gap. Nat Methods 6:S2 5.
2. Blake JA, Bult CJ (2006) Beyond the data deluge: Data integration and bio ontologies. J Biomed Inform 39:314 20.
3. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078 9.
4. HDF hierarchical data format. http://www.hdfgroup.org/.
5. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory efficient align ment of short DNA sequences to the human genome. Genome Biol 10:R25.

6. GFF3. http://www.sequenceontology.org/gff3.shtml.
7. Pruitt KD, Maglott DR (2001) RefSeq and LocusLink: NCBI gene centered resources. Nucleic Acids Res 29:137 40.
8. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470 6.
9. Thierry Mieg D, Thierry Mieg J (2006) AceView: A comprehensive cDNA supported gene and transcripts annotation. Genome Biol 7 Suppl 1:S12.1 14.
10. HDF EOS interface based on HDF5, volume 1: Overview and examples. http://www.hdfeos.org/reference/Info docs/HDF EOS/guides/HDFEOS 5 Lib User Guides/HDFEOS5 ug vol1 may 2004.pdf.
11. CFD general notation system switch to HDF5. http://cgns.sourceforge.net/hdf5.html.
12. Nexus. http://www.nexusformat.org.
13. MATLAB. http://www.mathworks.com.
14. Netcdf. http://www.unidata.ucar.edu/software/netcdf/.
15. Steinbiss S, Gremme G, Schärfer C, Mader M, Kurtz S (2009) Annotationsketch: A genome annotation drawing library. Bioinformatics 25:533 4.

Chapter 78
Multisensor Smart System on a Chip

Louiza Sellami and Robert W. Newcomb

Abstract Sensors are becoming of considerable importance in several areas, particularly in health care. Therefore, the development of inexpensive and miniaturized sensors that are highly selective and sensitive, and for which control and analysis is present all on one chip is very desirable. These types of sensors can be implemented with microelectromechanical systems (MEMS), and because they are fabricated on a semiconductor substrate, additional signal processing circuitry can easily be integrated into the chip, thereby readily providing additional functions, such as multiplexing and analog-to-digital conversion. Here, we present a general framework for the design of a multisensor system on a chip, which includes intelligent signal processing, as well as a built-in self-test and parameter adjustment units. Specifically, we outline the system architecture and develop a transistorized bridge biosensor for monitoring changes in the dielectric constant of a fluid, which could be used for in-home monitoring of kidney function of patients with renal failure.

Keywords Fluid Biosensors · MEMS · Sensors · Smart systems · System on a chip

78.1 Introduction

In a number of areas, it would be useful to have available smart sensors which can determine the properties of a fluid and from those make a reasoned decision. Among such areas of interest might be ecology, food processing, and health care. For example, in ecology, it is important to preserve the quality of water for which a number of parameters are of importance, including physical properties such as

L. Sellami (✉)
Electrical and Computer Engineering Department, US Naval Academy, Annapolis, MD 21401, USA
e mail: sellami@usna.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 78,

color, odor, pH, as well as up to 40 inorganic chemical properties and numerous organic ones [1, pp. 363 366]. Therefore, to determine the quality of water, it would be extremely useful if there were a single system on a chip which could be used in the field to measure the large number of parameters of importance and make a judgment as to the safety of the water. For such, a large number of sensors are needed and a means of coordinating the readouts of the sensors into a user-friendly output from which human decisions could be made. As another example, the food processing industry needs sensors to tell if various standards of safety are met. In this case, it is important to measure the various properties of the food, for example, the viscosity and thermal conductivity of cream or olive oil [2, pp. 283 287].

In biomedical engineering, biosensors are becoming of considerable importance. General theories of different types of biosensors can be found in [3 5] while similar devices dependent upon temperature sensing are introduced in [6]. Methods for the selective determination of compounds in fluids such as blood, urine, and saliva are indeed very important in clinical analysis. Present methods often require a long reaction time and involve complicated and delicate procedures. One valuable application in the health care area is that of the use of multiple sensors for maintaining the health of astronauts where presently an array of 11 sensors is used to maintain the quality of recycled air [7], although separate control is effected by the use of an external computer. Therefore, it is desirable to develop inexpensive and miniaturized sensors that are highly selective and sensitive, and for which control and analysis is available all on the same chip. These sensors can be implemented with microelectromechanical systems (MEMS). Since they are fabricated on a semiconductor substrate, additional signal processing units can easily be integrated into the chip, thereby readily providing functions such as multiplexing and analog-to-digital conversion. In numerous other areas, one could find similar uses for a smart multisensor array from which easy measurements can be made with a small portable device. These are the types of systems on a chip (SOC) that this chapter addresses.

78.2 System on a Chip Architecture

The architecture of these systems is given in Fig. 78.1 where there are multiple inputs, sensors, and outputs. In between are smart signal processing elements including built-in self-test (BIST). In this system, there may be many classes of input signals [e.g., material (as a fluid) and user (as indicator of what to measure)]. On each of the inputs, there may be many sensors [e.g., one material may go to several sensors each of which senses a different property (as dielectric constant in one and resistivity in another)]. The sensor signals are treated as an N-vector and combined as necessary to obtain the desired outputs, of which there may be many (such as an alarm for danger and indicators for different properties). For example, a patient with kidney disease may desire a system on a chip which gives an indication of when to report to the hospital. For this, an indication of deviation of dielectric

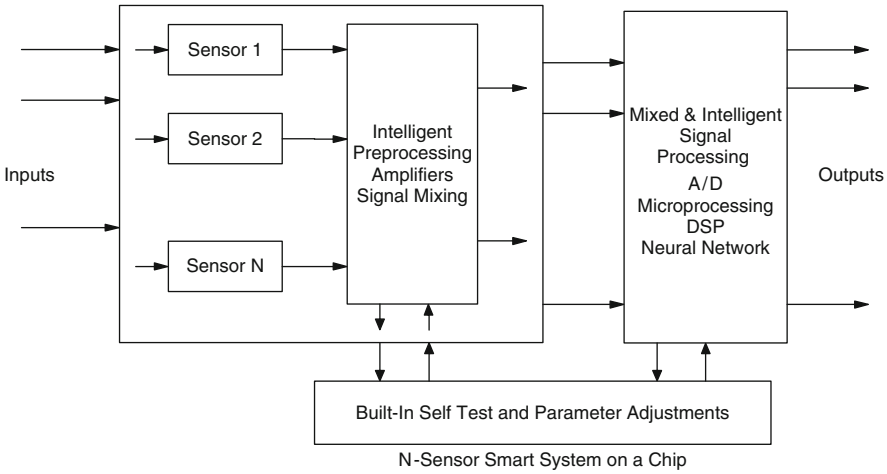


Fig. 78.1 Architecture for *N* sensor smart system on a chip

constant from normal and spectral properties of peritoneal fluid may be sensed and combined to give the presence of creatinine (a protein produced by the muscles and released in the blood) in the fluid, with the signal output being the percent of creatinine in the fluid and an alarm when at a dangerous level.

78.3 Dielectric Constant and Resistivity Sensor

The fluid sensing transistor in this sensor can be considered as a VLSI adaptation of the CHEMFET [7, p. 494] which we embed in a bridge to allow for adjustment to a null [8]. The sensor is designed for ease of fabrication in standard VLSI processing with an added glass etch step. A bridge is used such that a balance can be set up for a normal dielectric constant, with the unbalance in the presence of a body fluid being used to monitor the degree of change from the normal. The design chosen leads to a relatively sensitive system, for which on-chip or off-chip balance detection can occur. In the following, we present the basic sensor bridge circuit, its layout with a cross section to show how the chip is cut to allow measurements on the fluid, and simulation results from the Spice extraction of the layout that indicate the practicality of the concept.

Figure 78.2 shows a schematic of the sensor circuit. This is a capacitive-type bridge formed from four CMOS transistors, the two upper ones being diode-connected PMOS and the two lower NMOS, one diode connected and the other with a gate voltage control. The output is taken between the junction of the PMOS and NMOS transistors, and as such is the voltage across the midpoint with the circuit being supplied by the bias supply. As the two upper and the lower right transistors are diode connected, they operate in the saturation region while the gate

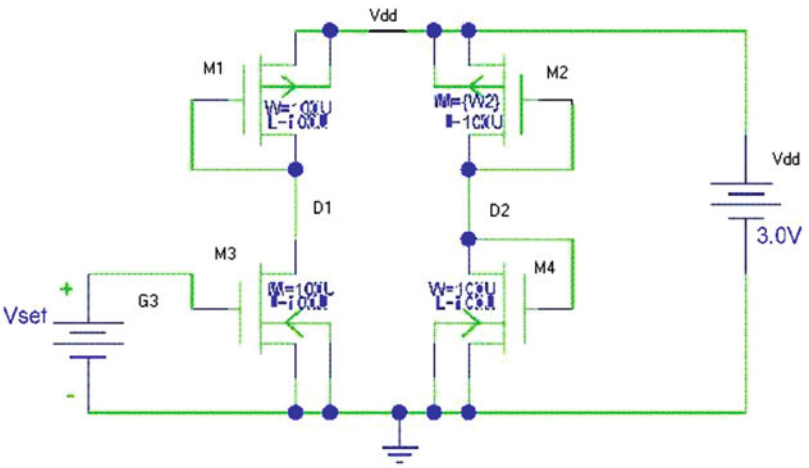


Fig. 78.2 Circuit schematic of a fluid biosensor

(the set node) of the lower left transistor, M_3, is fed by a variable DC supply allowing that transistor to be adjusted to bring the bridge into balance. The upper right transistor, M_2, has cuts in its gate to allow fluid to enter between the silicon substrate and the polysilicon gate. In so doing the fluid acts as the gate dielectric for that transistor. Because the dielectric constants of most fluids are a fraction of that of silicon dioxide, the fraction for water being about 1/4, M_2 is actually constructed out of several transistors, four in the case of water, with all of their terminals (source, gate, drain) in parallel to effectively multiply the Spice gain constant parameter *KP* which is proportional to the dielectric constant.

The sensor relies upon etching out much of the silicon dioxide gate dielectric. This can be accomplished by opening holes in protective layers by using the overglass cut available in MEMS fabrications. Since, in the MOSIS processing that is readily available, these cuts should be over an *n*-well, the transistor in which the fluid is placed is chosen as a PMOS one. And, since we desire to maintain a gate, only portions are cut open so that a silicon dioxide etch can be used to clear out portions of the gate oxide, leaving the remaining portions for mechanical support. To assist the mechanical support, we also add two layers of metal, metal-1 and metal-2, over the polysilicon gate.

A preliminary layout of the basic sensor is shown in Fig. 78.3 for M_2 constructed from four subtransistors, this layout having been obtained using the MAGIC layout program. As the latter can be used with different λ values to allow for different technology sizes, this layout can be used for different technologies and thus should be suitable for fabrications presently supported by MOSIS. Associated with Fig. 78.3 is Fig. 78.4 where a cross section is shown cut through the upper two transistors in the location seen on the upper half of the figure. The section shows that the material over the holes in the gate is completely cut away, so that an etch of

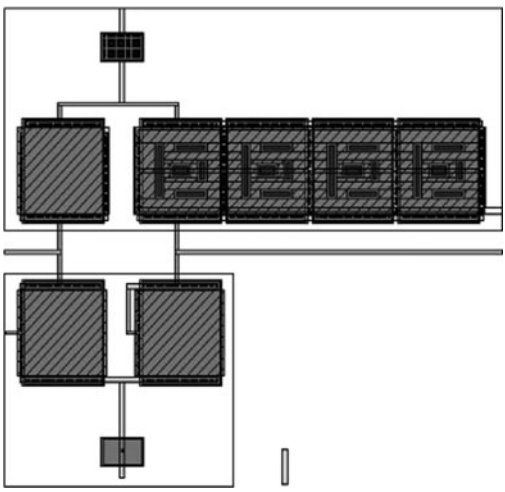


Fig. 78.3 Biosensor VLSI layout

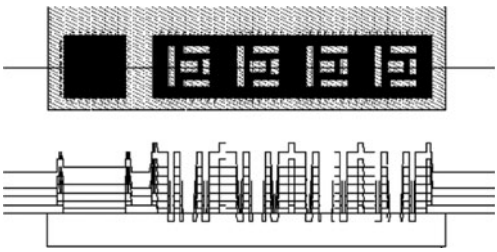


Fig. 78.4 Cross section of upper transistors

the silicon dioxide can proceed to cut horizontally under the remaining portions of the gate. The two layers of metal can also be seen as adding mechanical support to maintain the cantilevered portions of the gate remaining after the silicon dioxide etch.

To study the operation of the sensor, we turn to the describing equations. Under the valid assumption that no current is externally drawn from the sensor, the drain currents of M_1 and M_3 are equal and opposite, $I_{D3} = -I_{D1}$, and similarly for M_2 and M_4, $I_{D4} = -I_{D2}$. Assuming that all transistors are operating above threshold, since M_1, M_3, and M_4 are in saturation they follow a square law relationship while the law for M_3 we designate through a function $f(V_{set}, V_{D1})$ which is controlled by V_{set}. Thus,

$$-I_{D1} = \beta_1 (V_{dd} - V_{D1} - |V_{thp}|)^2 (1 + \lambda_p [V_{dd} - V_{D2}]), \tag{78.1a}$$

$$-I_{D1} = \beta_3 [f(V_{set}, V_{D1})(1 + \lambda_n V_{D1}) = I_{D3}, \tag{78.1b}$$

$$-I_{D2} = \varepsilon\beta_2(V_{dd} - V_{D2} - |V_{thp}|)^2(1 + \lambda_p[V_{dd} - V_{D2}]), \tag{78.2a}$$

$$-I_{D2} = \beta_4(V_{D2} - V_{thn})^2(1 + \lambda_n V_{D2}) = I_{D4}, \tag{78.2b}$$

where, for the ith transistor,

$$\beta_i = KP_i W_i/2L_i, \quad i = 1, 2, 3, 4 \tag{78.3}$$

and

$$f(x, y) = \{(x - V_{thn})^2 \text{ if } x - V_{thn} < y, 2(x - V_{thn})y - y^2 \text{ if } x - V_{thn} \geq y\}. \tag{78.4}$$

Here, V_{th}, KP, and λ are Spice parameters for silicon transistors, all constants in this case, with the n or p denoting the NMOS or PMOS case, and ε is the ratio of the dielectric constant of the fluid to that of silicon dioxide:

$$\varepsilon = \varepsilon_{fluid}/\varepsilon_{SiO_2}. \tag{78.5}$$

To keep the threshold voltages constant, we have tied the source nodes to the bulk material in the layout. In our layout, we also choose the widths and lengths of M_1, M_3, and M_4 to be all equal to 100 μm and L_2/W_2 to approximate ε. Under the reasonable assumption that the λs are negligibly small, an analytic solution for the necessary V_{set} to obtain a balance can be obtained. When M_3 is in saturation, the solution is

$$V_{D1} = V_{dd} - |V_{thp}| - \{\beta_3/\beta_1\}^{1/2}(V_{set} - V_{thn}) \tag{78.6}$$

while irrespective of the state of M_3

$$V_{D2} = \{V_{thn} + (\varepsilon\beta_2/\beta_4)^{1/2}(V_{dd} - |V_{thp}|)/[1 + \varepsilon\beta_2/\beta_4]^{1/2}\}. \tag{78.7}$$

Balance is obtained by setting $V_{D1} = V_{D2}$. Still assuming that M_3 is in saturation, the value of V_{set} needed to obtain balance is obtained from (78.6) and (78.7) as

$$V_{set} = V_{thn} + \{(\beta_1/\beta_3)\}^{1/2}(V_{dd} - |V_{thp}| - V_{thn})/[1 + (\varepsilon\beta_2/\beta_4)^{1/2}]. \tag{78.8}$$

At this point, we can check the condition for M_3 to be in saturation, this being that $V_{DS} \geq V_{GS} - V_{thn}$; since $V_{DS} = V_{D1}$ and $V_{GS} = V_{set}$, the use of (78.6) gives

$$V_{thn} < V_{set\{sat\}} \leq V_{thn} + (V_{dd} - |V_{thp}|)/[1 + (\beta_3/\beta_1)^{1/2}]. \tag{78.9}$$

Substituting the value of V_{set} at balance (78.8) shows that the condition for M_3 to be in saturation at balance is $\varepsilon\beta_2 \geq \beta_3$; this normally would be satisfied but can be guaranteed by making M_2 large enough.

Several things are added to the sensor itself per Fig. 78.1. Among these are a differential pair for direct current mode readout followed by a current mode pulse-coded neural network to do smart preprocessing to insure the integrity of the signals. Finally, a built-in test circuit is included to detect any breakdown in the sensor operation.

The sensor is sensitive to the dielectric constant of a fluid over an 11 1 range of dielectric constant most likely can be incorporated into a multisensor chip. Using standard analog VLSI MEMS processing, one can use the bridge for anomalies in a fluid by obtaining V_{set} for the normal situation and then comparing with V_{set} found for the anomalous situation. This could be particularly useful for determining progress of various diseases. For example, one way to determine kidney function and dialysis adequacy is through the clearance test of creatinine. The latter tests for the amount of blood that is cleared of creatinine per time period, which is usually expressed in ml/min. For a healthy adult, the creatinine clearance is 120 ml/min. A renal adult patient will need dialysis because symptoms of kidney failure appear at a clearance of less than 10 ml/min. Creatinine clearance is measured by urine collection, usually 12 or 24 h. Therefore, a possible use for the proposed sensor could be as a creatinine biosensor device for individual patient to monitor the creatinine level at home. An alternate to the proposed biosensor is based on biologically sensitive coatings, often enzymes, which could be used on M_2 transistor in a technology that is used for urea biosensors which are presently marketed for end stage renal disease patients [4]. The advantage of the sensor presented here is that it should be able to be used repetitively whereas enzyme-based coatings have a relatively short life. The same philosophy of a balanced bridge constructed in standard VLSI processing can be carried over to the measurement of resistivity of a fluid. In this case, the bridge will be constructed of three VLSI resistors with the fourth arm having a fluid channel in which the conductance of the fluid is measured.

78.4 Built-In Self-Test (BIST)

The BIST can interface with the sensors and other circuits under consideration. It can be built upon modifications of circuits and ideas available in the literature, such as the use of oscillations for mixed signal testing including the production line technique of using standard ring oscillator properties. The BIST is needed due to the fact that there are many interacting subsystems, and an error in one can perhaps drastically affect the operation of others.

BIST circuitry consists of a controller, a pattern generator, and a multiple input signature analyzer. The BIST method allows core testing to be realized by commanding the core BIST controller to initiate self-test and by knowing what the correct result should be. On-chip testing of embedded memories can be realized either by

multiplexing their address and data lines to external SOC I/O pads or by using the core processor to apply enough read/write patterns of various types to ensure the integrity of the memory. This technique works best for small embedded memories. Some recommend providing embedded memories with their own BIST circuitry.

For BIST to be effective, there must be a means for on-chip test response measurement, on-chip test control for digital and analog test, and I/O isolation. There are three categories of measurements that can be distinguished: DC static measurements, AC dynamic measurements, and time domain measurements. The first of these, DC static measurements, includes the determination of the DC operating points, bias and DC offset voltages, and DC gain. DC faults can be detected by a single set of steady-state inputs. AC dynamic measurements measure the frequency response of the system under test. The input stimulus is usually a sine-wave form with variable frequency. Digital signal processing (DSP) techniques can be employed to perform harmonic spectral analysis. Time domain measurements derive slew rate, rise and delay times using pulse signals, ramps or triangular waveforms as the input stimuli of the circuit.

78.5 Summary

In this chapter, we developed a general framework for the design and fabrication of a multisensor system on a chip, which includes intelligent signal processing, as well as a built-in self-test and parameter adjustment units. Further, we outlined its architecture, and developed a transistorized bridge fluid biosensor for monitoring changes in the dielectric constant of a fluid, which could be of use for in-home monitoring of kidney function of patients with renal failure.

Acknowledgments This research was sponsored in part by the 2007 Wertheim Fellowship, US Naval Academy.

References

1. De Zuane J (1990), Handbook of drinking water quality: Standards and control. Van Nostrand Reinhold, New York.
2. Singth RP, Heldman DR (1984), Introduction to food engineering. Academic, New York.
3. Van der Schoot BH, Berveld P (1988), Use of immobilized enzymes in FET detectors. In: Guilbault GG, Mascini M, Riedel Publishing Co. (ed) Analytical uses of immobilized biological compounds for detection medical and industrial uses. Riedel, Dordrecht.
4. Eggins BR (1996), Biosensors: An introduction. Wiley Teubner, New York.
5. Scheller F, Schubert F (1992), Biosensors. Elsevier, Amsterdam.
6. Van Herwaarden AW, Sarro PM, Gardner JW, Bataillard P (1994), Liquid and gas micro calorimeters for biochemical measurements. Sensors and Actuators 43:24 30.
7. Turner AFP, Karube I, Wilson GS (1987), Biosensors: Fundamentals and applications. Oxford University Press, Oxford.
8. Sellami L, Newcomb RW (1999) A mosfer bridge fluid biosensor. IEEE International Sympo sium on Circuits and Systems, May 30 June 2, Vol V, pp 140 143.

Chapter 79
Visual Presentation as a Welcome Alternative to Textual Presentation of Gene Annotation Information

Jairav Desai, Jared M. Flatow, Jie Song, Lihua J. Zhu, Pan Du, Chiang-Ching Huang, Hui Lu, Simon M. Lin, and Warren A. Kibbe

Abstract The functions of a gene are traditionally annotated textually using either free text (Gene Reference Into Function or GeneRIF) or controlled vocabularies (e.g., Gene Ontology or Disease Ontology). Inspired by the latest word cloud tools developed by the Information Visualization Group at IBM Research, we have prototyped a visual system for capturing gene annotations, which we named Gene Graph Into Function or GeneGIF. Fully developing the GeneGIF system would be a significant effort. To justify the necessity and to specify the design requirements of GeneGIF, we first surveyed the end-user preferences. From 53 responses, we found that a majority (64%, $p < 0.05$) of the users were either positive or neutral toward using GeneGIF in their daily work (acceptance); in terms of preference, a slight majority (51%, $p > 0.05$) of the users favored visual presentation of information (GeneGIF) compared to textual (GeneRIF) information. The results of this study indicate that a visual presentation tool, such as GeneGIF, can complement standard textual presentation of gene annotations. Moreover, the survey participants provided many constructive comments that will specify the development of a phase-two project (http://128.248.174.241/) to visually annotate each gene in the human genome.

Keywords Gene function · Social networking · Visualization · Word cloud

79.1 Introduction

Genes in the human genome have been predominantly annotated using unstructured text. For example, the Gene Reference Into Function (GeneRIF) provides a tool to include one or more 255-character-long "gene function" statements that couple a

W.A. Kibbe (✉)
The Biomedical Informatics Center and The Robert H. Lurie Comprehensive Cancer Center, Northwestern University Feinberg School of Medicine, Chicago, IL 60611, USA
e mail: wakibbe@northwestern.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 79,

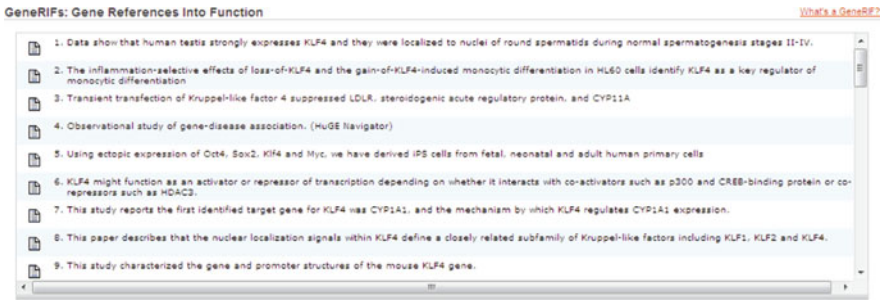


Fig. 79.1 GeneRIF annotation of gene KLF4 (human). Each GeneRIF is a statement up to 255 character long. Note that only 9 out of the 49 GeneRIFs are presented

specific publication with a gene [4, 6]. An example GeneRIF annotation of the human Kruppel-like factor 4 (KLF4, GeneID:9314) gene is shown in Fig. 79.1. For genes with more than about ten GeneRIFs, it is time-consuming to review the knowledge present in GeneRIFs.

Gene Ontology annotations [3] and Disease Ontology annotations [5] of a gene are more compact and the ontological structure makes these annotations much easier for a human reader to parse, in addition to the advantages of these ontologies for semantic reasoning and inference. However, these ontological systems require training to use consistently and accurately, and require a significant investment in curatorial time to build the ontological structure.

We investigated a different approach to present the genome annotation data. Research in human cognition has suggested that visual presentation can facilitate human learning and knowledge acquisition [2, 7, 10]. New semantic web tools, such as word clouds, appear to be ideally suited for helping people rapidly parse large amounts of textual data. Thus, we explored the impact of using a word cloud visual presentation of gene annotation information using the latest visualization tools developed by the Information Visualization Group at IBM Research (http://manyeyes.alphaworks.ibm.com/manyeyes). We call this visual annotation of a gene a "Gene Graph Into Function (GeneGIF)." Results from the user survey suggest that the visual presentation (GeneGIF) is complementary to the raw text presentation (GeneRIF) in current use.

79.2 Results

79.2.1 Word Clouds: A Direct Application of Existing Visualization Tools

A word cloud is a visual display of a set of words, where the font, size, color, or even movement can represent some underlying information. When a reader is in the

Fig. 79.2 Word clouds for the gene annotation of KLF4. (**a**) Using the tag cloud algorithm (**b**) using the Wordle algorithm

process of acquiring new information by examining evidences in a cursory manner, a word cloud can be very effective. Our first attempt was to apply the existing "word clould" tools of "tag-cloud" and "Wordle" from the "Many Eyes" project of IBM Research to display the words with their fonts proportional to their frequencies in the GeneRIFs of a gene, and the colors were added to enhance readability. Compared with the tag cloud tool that presents the words alphabetically (Fig. 79.2a) the Wordle tool utilizes the screen space more effectively (Fig. 79.2b).

79.2.2 Incorporating Domain-Specific Knowledge to Improve Wordle

Although Fig. 79.2b effectively summarizes the functions of KLF4, such as its role in DNA damage, cell cycle, promoter activity, and cancer, a careful examination of Fig. 79.2b suggests the following problems:

- Self-referencing: The gene symbol, "KLF4," is a self-reference. Although it appears as the most frequent word, it carries little extra information. Thus, the self-referencing words should be removed.
- Inflected words: The plural form of "cells" carries redundant information with the single form of "cell." Other examples include "KLF4" and "klf4" (upper vs. lower case). Thus, word stemming should be included in the algorithm.
- Stopwords: Although the common English stopwords are removed, some domain-specific stopwords, such as "gene," "paper," or "study," should also be removed.
- Phrases: Currently, the Wordle algorithm cannot identify phrases such as "cell cycle."

To count the word frequency more accurately, we applied the Porter Stemming Algorithm [9] to reduce words to word stems: for example, the inflected words "stimulates," "stimulated," and "stimulating" are all reduced to the stem form of "stimulate." In addition, we use the vocabulary list from the Gene Ontology [1] to identify phrases such as "cell cycle" and count them as a unit.

As noted, most word cloud generation engines remove stopwords from the input set before generating the cloud. Stopwords are overrepresented words or phrases

Table 79.1 Stopwords. The 20 most frequently occurring words (after stemming) in the entire GeneRIF dataset (*left*) and the Brown Corpus (*center*) are shown, and the overlaps between the two lists are highlighted. *On the right* is a list of manually identified stopwords, and its overlaps with the GeneRIF corpus list are *highlighted*

GeneRIF Corpus

Word(stem)	Frequency
Of	349654
The	271839
In	237674
And	235358
A	131960
Gene	105136
To	91638
Is	78351
Associ	75561
Cell	75300
Studi	66158
That	65412
With	61829
Diseas	60138
Observ	59924
Huge	57855
Navig	57717
By	56173
Express	54920
Active	54089

Brown Corpus

Word	Frequency
The	69970
of	36410
and	28854
to	26154
a	23363
in	21345
that	10594
is	10102
was	9815
He	9542
for	9489
it	8760
with	7290
as	7251
his	6996
on	6742
be	6376
at	5377
by	5307
I	5180

Expert Curation

Word
Paper
Summar
Review
Survey
Gene
Protein
Express
Show
Exhibit

which constitute parts of speech which occur frequently but convey nonspecific information. In text of a general nature, it suffices to remove definite and indefinite articles, prepositions, pronouns, and so on. However, in the application specific sense, there can be a large set of words which are redundant. In examining GeneRIFs, for example, common biological terms such as "gene" or "protein" will occur frequently and were added to a list of GeneGIF stopwords. To do this more formally, we used the entire GeneRIF as a corpus to identify the 50 most frequently occurring words (top 20 shown in Table 79.1). In contrast with the common English stopwords identified from the Brown corpus [8], we call the domain-specific stopwords "bio-stopwords." We combined the three lists from (Table 79.1) to remove the stopwords in GeneRIF.

The final visual presentation of KLF4 is shown in Fig. 79.3. We call this improved visual annotation of a gene "Gene Graph Into Function" (GeneGIF). The GeneGIF of KLF4 quickly summarizes the major functions of KLF4 from 49 entries of GeneRIF by displaying the more frequent keywords in bigger font: KLF4 is a *cell-cycle checkpoint* protein that prevents *mitosis* after *DNA damage*, and is thought to function as a *tumor suppressor*. KLF4 plays an important role in the *tumorigenesis* of *intestinal cancers*, especially *colorectal* adenocarcinomas. Decreased expression of KLF4 has been demonstrated in surgically resected colorectal cancers. The normal function of KLF4 seems to require the wild type *p53* protein. (The underscored words above are the keywords identified by GeneGIF.)

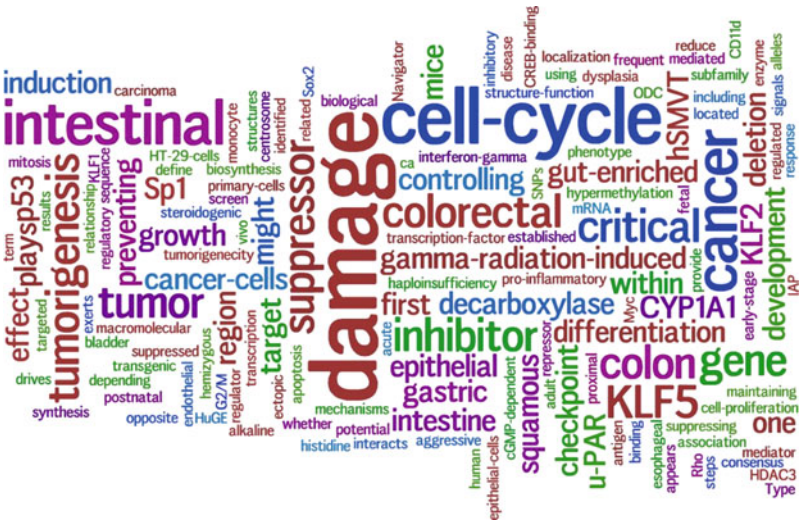


Fig. 79.3 Improvements over the Wordle algorithm. We removed all of the self references, stopwords (English stopwords and bio stopwords) and hyphenate the phrases (for example, cell cycle, interferon gamma, and cancer cells are identified)

79.3 User Survey

To test the utility of GeneGIF, we surveyed the users of gene annotations. We sent a survey to participants of the MAQC project who are experts on genomic data analysis using microarrays. Prospective participants were informed with the purpose, procedure, and handling of the survey. Since it was an anonymous survey, signatures for the consent were waived. 53 responses were collected. Statistical tests of the survey results suggest that in terms of acceptance, a majority (point estimate: 64%; 95% confidence interval: 50 77%) of the users were either positive or neutral toward using GeneGIF in their daily work; in terms of preference, a slight majority (51%, not statistically significant) of the users favored visual (GeneGIF) information compared to textual (GeneRIF) information. An Analysis of Variance (ANOVA) suggests no significant association of the outcome (GeneGIF vs. GeneRIF) with either gender, age, field of study, English as first language, or education level.

79.4 Discussion

The state of knowledge about a given gene changes, and these changes are reflected in the literature. Although the GeneRIF provides a mechanism to keep the functional annotation of a gene up-to-date, reading through GeneRIF entries to identify

significant and recurring points is not easy when there are dozens or even hundreds of GeneRIFs.

For the first time, we have prototyped a visual system of gene annotation (GeneGIF) by summarizing the phrases used in a collection of GeneRIFs. As the user comments indicate, GeneGIF is much more effective in getting a rough overview of the gene's major functions while GeneRIF can provide more detailed and precise information. Therefore, GeneGIFs are complementary to the raw textual display of GeneRIFs. The MAQC respondents also pointed out that the current prototype of GeneGIF is very primitive. For instance, we can make the GeneGIF clickable and directly linked it to individual GeneRIF items with the keyword highlighted. Based on these positive feedbacks, we have begun a phase II project to use GeneGIF to annotate each gene in the human genome (http://128.248.174.241/).

We have also found that the same visual representation can be used for more than just single genes. We have used gene lists from gene expression experiments to build word clouds that are based on a collection of GeneRIF collections. This is a rapid way to identify functional pathways that are affected in the collection. Another application is to directly graph gene expression data into the cloud structure. For example, position can be used to define whether the expression was negative or positive (right to left, respectively), the size of the term for expression magnitude, colored grouping for related biomarkers (e.g., common pathway), and even some degree of movement (vibration) to express the noise/discrepancy. These various types of data are all amenable to word cloud visualization.

79.5 Materials and Methods

Gene annotations were downloaded from the NCBI Entrez database in January 2009. The word cloud was created using the Wordle algorithm from the "Many Eyes" project of IBM Research (http://manyeyes.alphaworks.ibm.com/manyeyes/). The survey was designed using the django-survey application (http://code.google.com/p/django-survey/). Other programs were written in Python. Normal approximation was used to estimate the 95% confidence interval for the proportion of users who were positive or neutral toward using GeneGIF in their daily work. A one sample z-test was used to test users' preference of using GeneGIF to GeneRIF, i.e., the proportion of users who prefer using GeneGIF is greater than 50%.

Acknowledgments The authors would like to thank Martin Wattenberg, Matthew M Mckeon, and Jonathan Feinberg at IBM Research for helpful discussion of Wordle and comments on this manuscript. The authors would also like to thank Rhett Sutphin at NUCATS for exploring the application programing interface to Wordle.

References

1. Ashburner, M., C. A. Ball, et al. (2000). “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.” Nature Genetics **25**(1): 25 9.
2. Childers, T. L., M. J. Houston, et al. (1985). “Measurement of individual differences in visual versus verbal information processing.” *Journal of Consumer Research* **12**(2): 125 134.
3. Harris, M. A., J. Clark, et al. (2004). “The Gene Ontology (GO) database and informatics resource.” *Nucleic Acids Research* **32**(Database issue): D258 61.
4. Maglott, D., J. Ostell, et al. (2007). “Entrez Gene: gene centered information at NCBI.” *Nucleic Acids Research* **35**(Database issue): D26 31.
5. Osborne, J. D., J. Flatow, et al. (2009). “Annotating the human genome with disease ontol ogy.” *BMC Genomics* **10**(Suppl 1):S6.
6. Osborne, J. D., S. Lin, et al. (2007). “Other riffs on cooperation are already showing how well a wiki could work.” *Nature* **446**(7138): 856.
7. Plass, J. L., D. M. Chun, et al. (1998). “Supporting visual and verbal learning preferences in a second language multimedia learning environment.” *Journal of Educational Psychology* **90**(1): 25 36.
8. Weiss, S. M. (2005). *Text mining : predictive methods for analyzing unstructured information*. New York, Springer.
9. Willett, P. (2006). “The Porter stemming algorithm: then and now.” *Program Electronic Library and Information Systems* **40**(3): 219 223.
10. Wyer, R. S., Y. W. Jiang, et al. (2008). “Visual and verbal information processing in a consumer context: Further considerations.” *Journal of Consumer Psychology* **18**(4): 276 280.

Chapter 80
Automatic FRAP Analysis with Inhomogeneous Fluorescence Distribution and Movement Compensation

Harri Polonen, Maurice Jansen, Elina Ikonen, and Ulla Ruotsalainen

Abstract The analysis of fluorescence recovery after photobleaching (FRAP) data is complicated by the measurement noise, inhomogeneous fluorescence distribution, and image movement during experiment. Conventionally, these issues are tackled by data preprocessing and averaging, which causes loss of quantitative properties. In this study, we present a method which automatically estimates and compensates both the movement and inhomogeneous fluorescence distribution within the data analysis. The method is based on modeling the raw FRAP data with a parametric matrix and searching for maximum likelihood parameters between the model and the data. The developed method also automatically estimates also the bleach profile, immobile fraction, and noise variance. Suitable numerical computational method was developed and implemented in a computer grid. Simulated and experimental FRAP data was created and analyzed to evaluate the method.

Keywords Data analysis · Data preprocessing · Fluorescence distribution · FRAP analysis · Numerical computations

80.1 Introduction

The idea of Fluorescence Recovery After Photobleaching (FRAP) experiments is several decades old [1] and used with various applications [2, 3] examining the dynamic behavior of a cell. For long, the FRAP data analysis was based on calculating the average bleach area intensities at each acquisition time point and fitting them to a formal solution of diffusion equation. Although the mathematical

H. Pölönen (✉)
Department of Signal Processing, Tampere University of Technology, Korkeakoulunkatu 10, 33720 Tampere, Finland
e mail: harri.polonen@tut.fi

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 80,

justification [4, 5] of such methods is indisputable, they are highly dependent on the correctness of initial assumptions and are very sensitive to data defectiveness such as cell movement, high noise, and low number of images. Recently, spatial and likelihood based methods[6 8] have been developed as an attempt to take into account the whole FRAP data in the analysis and improved accuracy was reported in comparison to conventional methods. However, image movement was not considered and the cell inhomogeneity was dealt with either averaging the pre-bleach images or assuming rotational symmetry.

Our aim was to create an automatic FRAP analysis method with built-in inhomogeneous fluorescence distribution estimation and image movement compensation. The method is based on modeling the raw microscopy data with a parametric matrix model. The model is built from several components, including diffusion simulation according to the diffusion theory. We constructed a likelihood framework for the raw data and use maximum likelihood estimation to determine the best model parameters. A suitable parameter optimization method was developed and implemented into a computer grid. We created simulated and experimental FRAP data to test the developed method under various circumstances.

80.2 Methods

Fluorescence concentration model $\mathbf{M}$ is built from parametric components as

$$\mathbf{M} = m(\gamma \mathbf{H_B} + (1-\gamma)(\mathbf{H} + f(\mathbf{H_B} - \mathbf{H}))) \tag{80.1}$$

where m is a function modeling image movement, γ determines the immobile fraction, $\mathbf{H}$ describes the inhomogeneous equilibrium state of fluorescence, $\mathbf{H_B}$ denotes $\mathbf{H}$ with bleaching applied, and f is a function modeling diffusion. The various components and additional parameters of the model are described more in detail in the following sections.

The inhomogeneous equilibrium state of the fluorescence is modeled with a matrix $\mathbf{H}$, whose elements $\mathbf{H}(x)$ describe the relative amount of fluorescence within a corresponding pixel x. Each element of the matrix $\mathbf{H}$ is estimated as a separate parameter in the likelihood maximization process. This, the inhomogeneous equilibrium state, is automatically determined and no pre-estimation or correction for the inhomogeneity is needed. Often the resolution of $\mathbf{H}$ can be lower than the image resolution because usually there are no sharp edges, i.e., high frequency components, in a microscope image of a cell that need to be modeled. By excluding the high frequencies, overfitting to random pixel-to-pixel differences can be avoided and less parameters are needed to describe the inhomogeneity. The best resolution for $\mathbf{H}$ naturally depends on the experiment.

The bleach profile defines the relative effect of laser bleaching to the fluorescence concentration in the model. The bleach profile is modeled here with scaled and bounded Gaussian distribution constructed from three parameters: bleach

power $\alpha > 0$, bleach shape $\mathbf{\Sigma} > 0$, and bleach level $0 < \beta < 1$. Formally, the relative bleach effect $\mathbf{B}(x \mid \alpha,\beta,\mathbf{\Sigma})$ in pixel x is defined as

$$\mathbf{B} = 1 - min\left\{\beta, \frac{\alpha}{(2\pi)^{n/2}|\Sigma|^{1/2}} e^{\frac{1}{2}(x \quad \mu)^T \Sigma^{-1}(x \quad \mu)}\right\}, \tag{80.2}$$

where μ is the (known) bleach location. The bleaching is then modeled by multiplying the estimated bleach profile and the estimated inhomogeneous fluorescence equilibrium state pixel by pixel as $\mathbf{H}_B(x) = \mathbf{H}(x)\mathbf{B}(x)$. Note that although the above model formally contains a Gaussian distribution, it allows a wide variety of bleach profiles from pure Gaussian (with small α) to practically uniform disk (high α). Bleaching is applied both to mobile and immobile fractions of the model matrix.

The diffusion after photobleaching is here modeled similarly as in [8] through n-dimensional convolution between the heat kernel and initial fluorescence distribution. Heat kernel, i.e., the fundamental solution of Fick's second law describes the distribution of a point source initially located at zero

$$\Phi(x \mid D, t) = \frac{1}{(4\pi Dt)^{n/2}} exp\left(\frac{|x|^2}{4Dt}\right) \tag{80.3}$$

where D denotes the diffusion coefficient. The heat kernel can be used to obtain the evolution of an initial distribution over a time period t as n-dimensional convolution between the kernel and the initial matrix as

$$f(\mathbf{X} \mid D, t) = \mathbf{X} \otimes \Phi(x \mid D, t). \tag{80.4}$$

Diffusion rarely occurs uniformly due to inhomogeneous cell structure, and thereby it is common to apply the diffusion modeling to the deviation from the equilibrium state [9]. Here, we apply the diffusion dynamics to the division ($\mathbf{H}_B - \mathbf{H}$) between the inhomogeneous equilibrium state and the inhomogeneous post-bleach state [see (80.1)]. Note that although Brownian motion is a widely used model for diffusion of fluorescent molecules; it may not be a realistic model for all FRAP experiments and the correct choice of model and equations is dependent, for example, on the size of the fluorescent molecule and cellular environment. For a review of applications and corresponding models see, e.g., [2, 10]. Diffusion dynamics is applied only to mobile fraction of the model.

By *image movement* we mean here that the whole visible part of the cell in the image moves linearly at constant velocity and direction during the experiment. This can be caused, for example, by cell motility or mechanical drift either of the sample or within the microscope caused by temperature equilibrations. We model the movement here as moving the camera in the opposite direction analogously. With the lateral movement direction τ (in radians), lateral velocity r_l, and z-direction velocity r_z, the location of the camera initially at $y = (y_1,y_2,y_3)$ is shifted in time t to

$x = (y_1 + tr_l\cos(\tau), y_2 + tr_l\sin(\tau), y_3 + r_z)$. To obtain the image after movement, an integral has to be then calculated over each pixel area/volume in the new camera location as

$$m(\mathbf{X} \mid t, \tau, r_l, r_z) = \int_{x\ \ 0.5}^{x+0.5} \mathbf{X}(z) d^n z, \tag{80.5}$$

where the new pixel location x is defined through parameters t, τ, r_l, and r_z. In practice, the integration can be performed with a weighted sum of pixel values.

Maximum likelihood estimation is used to determine the parameters that produce the best model for the acquired raw FRAP data. We assume that the shot noise is dominant to other noise sources, and the image noise follows thereby Poisson type statistics with variance dependent on the square root of the mean [11]. The raw (noisy) image $\mathbf{N}$ is thus modeled through modified Poisson probability density function

$$N(x,t) \sim Poiss\left(\sqrt{C(x,t)}\right) - \sqrt{\mathrm{C}(x,t)}\left(\rho - \sqrt{\mathrm{C}(x,t)}\right), \tag{80.6}$$

where $\mathbf{C}(x,t)$ denotes the (unknown) noiseless pixel value and ρ controls the noise variance. To find the best model for the observed raw FRAP data, we substitute the unknown noiseless pixel value $\mathbf{C}(x,t)$ with the model matrix value $\mathbf{M}(x,t/\theta)$ defined in (80.1), where θ denotes the set of all estimated parameters. Namely, θ consists of heterogeneity matrix $\mathbf{H}$, bleach profile parameters α, β, and $\mathbf{\Sigma}$, Immobile fraction γ, diffusion coefficient D, movement direction τ and velocities r_l, r_z, and noise variance multiplier ρ.

Based on (80.6), *the pixel-wise likelihood* value of the model is

$$p(x,t \mid \theta) = \frac{\left(\rho\sqrt{M(x,t|\theta)}\right)^{N(x,t)+\sqrt{M(x,t|\theta)}\left(\rho\ \ \sqrt{M(x,t|\theta)}\right)} exp\left(-\rho\sqrt{M(x,t|\theta)}\right)}{\left(N(x,t) + \sqrt{M(x,t|\theta)}\left(\rho - \sqrt{M(x,t|\theta)}\right)\right)!}. \tag{80.7}$$

The joint likelihood is then given by the product of all pixel-wise likelihoods in every pixel and time point. The parameters that produce the maximum value for the joint likelihood of all pixels x at all acquisition time points t is searched for.

The following *hybrid algorithm* was developed and implemented for parameter optimization:

1. Choose initial guess for parameters θ
2. Perform random search for 120 s
3. Perform differential evolution [12, 13] optimization for 300 s
4. Run Nelder Mead [14] algorithm for 300 s
5. If {"improvement in likelihood in last three iterations" $> 0.1\%$}, repeat from 2.

We compared this hybrid with other algorithms, including Nelder Mead, differential evolution and a genetic algorithm, and the hybrid was found to perform best

in parameter estimation. To create and estimate large simulated datasets, we used a computer grid with more than 800 computer cores to perform the optimizations.

As a *reference method*, we used the approach presented by Sprague [4] based on closed-form solution by Soumpasis [5]. This method is based on calculating the average value within the bleach area from the image at each acquisition time point and fitting them to the theoretical solution of diffusion equation.

80.3 Experiments

We created two-dimensional *simulated FRAP data* to evaluate the developed method. Simulated data was built by setting random values for the parameters θ and time vector t, calculating the corresponding "true" (noiseless) images and contaminating the images with Poisson type noise according to the set ρ value. There were 200 simulated FRAP experiments in total with the following parameters: time vector $t = (0,2,...,30)$, image resolution 32×32, noise variance $\rho = 5$, bleach parameters $\alpha = 30$, $\beta = 1$, diagonal $\mathbf{\Sigma}$ with both elements equal to 0.1, and immobile fraction $\gamma = 0$. Matrix $\mathbf{H}$ was sized 5×5 pixels with each value randomly chosen from uniform integer distribution on interval (34,. . .,64). The movement parameter r_l was increased linearly in the experiments from 0 to 0.2 (pixels/time) resulting as maximum final movement of six pixels. Two simulated test groups were created in the data by setting the diffusion coefficient value either $D = 0.15$ or $D = 0.30$ (pixels^2/time). The methods' robustness and capability to separate the treatment groups could thereby be tested.

Experimental data was created to test the inhomogeneity estimation. The FRAP experiments were performed on a Zeiss LSM 5 Duo confocal microscope equipped with a Plan-Apochromat 10×/0.45 M27 objective. FITC-dextran was dissolved at 1.5 mg/ml in PBS buffer (pH 7.4) containing 30% (w/w) glucose. Up to 200 8-bit images with resolution 128×128 pixels were acquired with an interval of 447 ms. The solution was filled into microslides with ground cavity, overlaid with a cover slip and sealed with silicon grease. Movement of the bleached area was achieved by slightly tilting the sample during imaging causing a local temperature rise within the sample and resulting in a directional flow inside the sample solution.

The simulated FRAP dataset was first quantified with the developed model with homogeneous and 5×5 sized inhomogeneous fluorescence distribution matrix $\mathbf{H}$. In Fig. 80.1, it can be seen that compensation of inhomogeneous fluorescence distribution is crucial in diffusion coefficient estimation. When the inhomogeneity is compensated the two treatment groups are well separable, while with homogeneous $\mathbf{H}$ it is difficult to separate the groups.

The effect of increasing image movement was then tested, and the results are shown in Fig. 80.2. The reference method fails systematically to separate the treatments. Without movement compensation, the diffusion coefficient estimates from the developed method get clearly overestimated with increased movement and

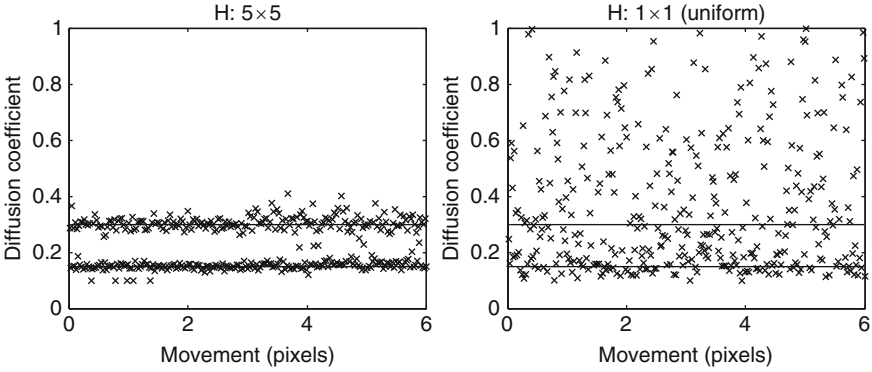


Fig. 80.1 Diffusion coefficient estimates from simulated data using heterogeneous and homoge neous fluorescence distribution

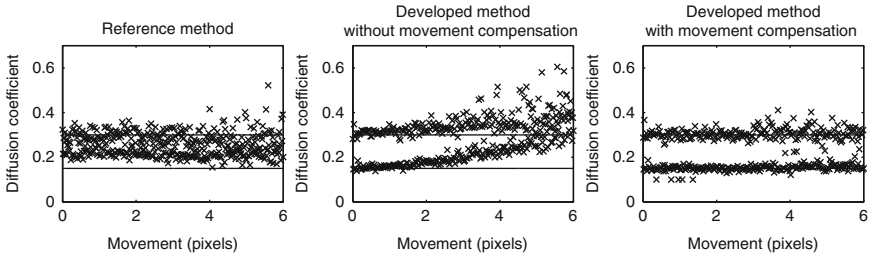


Fig. 80.2 Diffusion coefficient estimates from simulated data with increasing movement

it is finally hard to separate the treatment groups. With movement compensation, the increased movement does not affect the diffusion coefficient estimates remarkably.

The microscopy data was estimated with the developed full model, including both inhomogeneity and movement compensation. The microscopy data was down-sampled to 20 time frames by including only every tenth frame to make the data analysis more challenging. The fluorescence distribution estimates with various **H** resolutions can be seen in Fig. 80.3 with the region of interest (32 × 32 pixels) from the microscope data. It seems that the inhomogeneity can be estimated quite well with the 5 × 5 matrix, and there may be some noise-related artifacts in 20 × 20 sized **H**.

80.4 Conclusion

We have introduced an automatic FRAP analysis method to offer robust and accurate results in the case of image movement and inhomogeneous fluorescence distribution. The method requires no pre-correction for the movement or

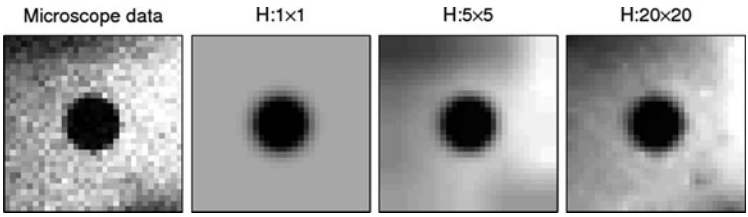


Fig. 80.3 Microscope data and inhomogeneous fluorescence distribution estimates with three different resolutions

pre-estimation of the inhomogeneous fluorescence distribution, but compensates these issues automatically. The simulation results showed that both the image movement and fluorescence inhomogeneity can cause large errors in estimates if not taken into account. The developed method was able to compensate these issues very well and the results were substantially better. In comparison to the conventional method, treatment groups were more separated in the results of developed method. From the microscopy data, the inhomogeneous fluorescence distribution could be estimated well even with lowered resolution.

Acknowledgments The work was financially supported by the Academy of Finland under the grant 213462 (Finnish Centre of Excellence Program (2006 2011)). HP received additional support from Jenny and Antti Wihuri Foundation. MJ and EI thank Sigrid Juselius Foundation, Helsinki Biomedical Graduate School, and Biomedicum Helsinki Foundation and Molecular Imaging Unit, Helsinki.

References

1. Lippincott Schwartz J, Snapp E, Kenworthy A (2001) Studying protein dynamics in living cells. Nat Rev Mol Cell Biol, 2(6):444 446
2. Reits E A, Neefjes J J (2001) From fixed to FRAP: measuring protein mobility and activity in living cells. Nat Cell Biol, 3(6):E145 147
3. Axelrod D, Koppel D E, Schlessinger J et al. (1976) Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. Biophys J, 16(9):1055 1059
4. Sprague B L, Pego R L, Stavreva D A et al. (2004) Analysis of binding reactions by fluorescence recovery after photobleaching. Biophys J, 86(6):3473 3475
5. Soumpasis D M (1983) Theoretical analysis of fluorescence photobleaching recovery experi ments. Biophys J, 41(1):95 97
6. Siggia E D, Lippincott Schwartz J, Bekiranov S (2000) Diffusion in inhomogeneous media: theory and simulations applied to whole cell photobleach recovery. Biophys J, 79(4):1761 1770
7. Irrechukwu O N (2009) Improved estimation of solute diffusivity through numerical analysis of frap experiments. J Cell Mol Biol, 2(1):104 117
8. Jonasson J K, Loren N, Olofsson P et al. (2008) A pixel based likelihood framework for analysis of fluorescence recovery after photobleaching data. J Microsc, 232(2):260 269
9. Tannert A, Tannert S, Burgold S et al. (2009) Convolution based one and two component FRAP analysis: theory and application. Eur Biophys J, 38(5):649 651

10. Chen Y, Lagerholm B C, Yang B et al. (2006) Methods to measure the lateral diffusion of membrane lipids and proteins. Methods, 39(2):147 153
11. Goldman R D (2004) Live cell imaging. CSHL Press, New York
12. Lampinen J A, Price K V, Storn R M (2005) Differential evolution A practical approach to global optimization (Natural computing series). Springer, New York
13. Polonen H, Tohka J, Ruotsalainen U (2009) Automatic quantification of fluorescence from clustered targets in microscope images. Lecture Notes in Computer Science, 5575:667 675
14. Nelder J A, Mead R (1965) A simplex method for function minimization. Comput J, 7:308 313

Chapter 81
Sorting Circular Permutations by Bounded Transpositions

Xuerong Feng, Bhadrachalam Chitturi, and Hal Sudborough

Abstract A k-bounded ($k \geq 2$) transposition is an operation that switches two elements that have at most $k - 2$ elements in between. We study the problem of sorting a circular permutation π of length n for $k = 2$, i.e., adjacent swaps and $k = 3$, i.e., short swaps. These transpositions mimic microrearrangements of gene order in viruses and bacteria. We prove a $(1/4)n^2$ lower bound for sorting by adjacent swaps. We show upper bounds of $(5/32)n^2 + O(n \log n)$ and $(7/8)n + O(\log n)$ for sequential and parallel sorting, respectively, by short swaps.

Keywords Bounds · Genomic mutations · Sorting · Transpositions

81.1 Introduction

In nature, some species have similar genetic make up and differ only in the order of their genes. The rearrangements, such as inversions, transpositions, and translocations, that span multiple genes alter the gene order in genome. Studies show that numerous rearrangements with a limited span happen during the draft of two species. For example, many rearrangements with a span less than 1 Mb happen between human and mouse genome [1]. Also, the genome of most bacteria and viruses, for example, *Escherichia coli*, is circular. Finding the shortest rearrangement path between two related bacteria or viruses is useful in drug discovery and vaccine development. Thus, certain limited span rearrangement events that mimic mutations of bacteria (virus) can be modeled as transformation of one circular permutation, i.e., *cperm*, into another by length bounded transpositions. We study the length bounds of two (*adjacent swap* [2]) and three (*short swap* [3, 4]). Sorting a cperm in parallel by short swaps models information exchange in hybrid mesh ring topology.

B. Chitturi (✉)
Department of Biochemistry, University of Texas SW Medical Center, Dallas, TX 75390, USA
e mail: chalam@utdallas.edu

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 81,

Jerrum [5] showed that finding the minimum number of adjacent swaps to sort a cperm is in P. In [6], Pevzner gave a $2\lfloor n/2 \rfloor \lceil n/2 \rceil$ upper bound for the same. Heath and Vergara [3] gave a (4/3)-approximation algorithm for sorting a linear permutation (or simply *lperm*) by short swaps. They also show a $(1/4)n^2 + O(n)$ upper bound and an $(1/6)n^2 + \Omega(n)$ lower bound. Feng et al. [4] improved the upper bound to $(3/16)n^2 + O(n \log n)$. Mahajan [7] explored the combinatorial aspects of the short swaps on lperms. Related problems appear in [8 11].

In Sect. 81.2, by applying the results in [5], we prove a lower bound of $(1/4)n^2$ for sorting a cperm by adjacent swaps, and we conjecture an upper bound of $(1/4)n^2$. In Sect. 81.3, for sorting a cperm by short swaps, we show a lower bound of $(1/8)n^2$ based on results in [3], and we show $(5/32)n^2 + O(n \log n)$ upper bound. In Sect. 81.4, we sort a cperm by short swaps in $(7/8)n + O(\log n)$ parallel steps. Section 81.5 states conclusions and open questions.

81.2 Bounds for Sorting by Adjacent Swaps

For a cperm π of length n, we mark the position of its elements by 1, 2, 3,. . ., n as shown in Fig. 81.1. Let π_i denote the element at position i. For simplicity, we write a cperm by listing its elements at position 1, 2, 3,. . ., n in sequence. For example, the cperm in Fig. 81.1 is π: (4 7 6 1 3 5 2 8) and the circular identity permutation I of length n is 1, 2, 3,. . ., n. The permutation π in Fig. 81.1 has 14 inversions; hence, if it is linear, 14 adjacent swaps are required. For a cperm π, we seek a shortest sequence of permutations $\gamma_1, \gamma_2, \ldots, \gamma_k$ such that $\pi \cdot \gamma_1 \cdot \gamma_2 \cdot \cdots \cdot \gamma_k = I$.

First, we introduce relevant notations from Jerrum [5]. For a cperm π, we define its *displacement vector* (or just *dvec*) x, where $x = (x_1, x_2, x_3, \ldots, x_n)$ such that $x_i = k - i$ and $\pi_k = i$ $(i, k \in [1, n])$. For the permutation π in Fig. 81.1, its dvec x is (3 5 2 −3 1 −3 −5 0). For any dvec x, we have: $\sum_{i=1}^{n} x_i = 0$.

Based on a dvec x, the *crossing number* (*cnum*) is defined. It is similar to the inversion number defined for a lperm. Let the dvec be $x = (x_1, x_2, x_3, \ldots, x_n)$; for any pair of i, j in [1, n], let $r = i - j$ and $s = (i + x_i) - (j + x_j)$. We define the cnum $C_{ij}(x)$ of i and j with respect to x by the following equation (81.1):

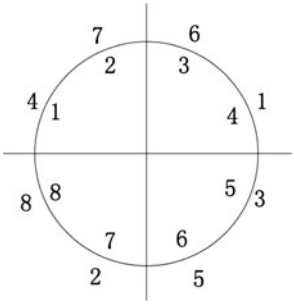


Fig. 81.1 Circular permutation π: (4 7 6 1 3 5 2 8)

$$C_{ij}(x) = \begin{cases} |\{k \in [r,s] | k := 0 (\mathrm{mod} n)\}|, & \text{if } r \leqslant s, \\ -|\{k \in [s,r] | k := 0 (\mathrm{mod} n)\}|, & \text{if } s < r. \end{cases} \tag{81.1}$$

Intuitively, if $\pi \cdot \gamma_1 \cdot \gamma_2 \cdot \cdots \cdot \gamma_k$ is a sequence that sorts π into I, $C_{ij}(x)$ measures the number of times that elements i and j cross. If i crosses j in a clockwise direction, then $C_{ij}(x) > 0$, otherwise $C_{ij}(x) < 0$; $C_{ij}(x) = -C_{ji}(x)$. If there is no number k in the specified range such that $k = 0 (\mathrm{mod}\ n)$, then $C_{ij}(x) = 0$. The relationship between a dvec x and the corresponding cnum is given by

$$x_i = \sum_{k=1}^{n} C_{ik}(x). \tag{81.2}$$

In a lperm, let π_i and π_j be two numbers, where $i < j$. If π_i needs to move to the right by p_1 positions and π_j needs to move to the left by p_2 positions, where $p_1 + p_2 > j - i$, then $p_1 + p_2 - 1$ adjacent swaps are necessary. However, cperms have an advantage when $p_1 + p_2 > n$, it takes fewer adjacent swaps if π_i is moved to the left and π_j is moved to the right. This concept motivates the following notations and Lemmas 81.1 and 81.2.

Let $\max(x)$ and $\min(x)$ denote the maximum and the minimum component value of a dvec x. Let $x_i = \max(x)$ and $x_j = \min(x)$. If $\pi \neq \mathrm{I}$, then $x_i > 0$ and $x_j < 0$. For the dvec x shown in Fig. 81.1, we have $\max(x) = x_2 = 5 > 0$ and $\min(x) = x_7 = -5 < 0$. The transformation T_{ij} $(i \neq j)$ on x is defined as follows: if T_{ij} $(i \neq j)$ is applied to x, then the resulting $x' = T_{ij}(x)$, where x' is given by

$$x'_k = x_k, \text{ if } k \neq i \text{ or } j, \quad x'_i = x_i - n, \quad x'_j = x_j + n. \tag{81.3}$$

If there exists at least one pair of indices i and j such that $x_i - x_j > n$, then we say that the transformation $T_{ij}(x)$ *strictly contracts* x. Otherwise, if for all values of i and j, $x_i - x_j \leq n$, then we say that x admits no strictly contracting transformation.

Lemma 81.1 [5]. *For a circular permutation π, if its displacement vector x admits no strictly contracting transformations, then for all $i, j \in [1, n]$, $\mathrm{C_{ij}(x)} = \{-1, 0, +1\}$.*

Lemma 81.1 shows that if $\pi \cdot \gamma_1 \cdot \gamma_2 \cdot \cdots \cdot \gamma_k$ is a shortest sequence that sorts π into I, then for any two elements $i, j \in [1, n]$ and $i \neq j$, the element i crosses the element j at most one time. We build an $n \times n$ *cnum matrix* $M(x)$, where $M_{ij}(x) = C_{ij}(x)$ for all $i, j \in [1, n]$. In Fig. 81.1, since $\max(x) = x_2 = 5$ and $\min(x) = x_7 = -5$, $\max(x) - \min(x) = 10 > n\ (=8)$, and so T_{27} strictly contracts x. Applying T_{27} to x, we obtain the new dvec $x' = (3\ -3\ 2\ -3\ -1\ -3\ 3\ 0)$. Figure 81.2 shows the cnum matrix for π with x'.

For an arbitrary cperm π length n, let $I(x)$ be defined as: $I(x) = \frac{1}{2}\sum_{i,j=1}^{n} |M_{ij}(x)|$.

Lemma 81.2 [5]. *For a circular permutation π, let x be its displacement vector that admits no strictly contracting transformations. Then, the minimum number of steps needed to sort π into identity permutation I by adjacent transposition equals I(x).*

Fig. 81.2 Crossing number matrix for π: (4 7 6 1 3 5 2 8)

$$\begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ -1 & 0 & -1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & 0 & -1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & -1 & 0 \end{pmatrix}$$

Fig. 81.3 A hard permutation to sort by adjacent swaps

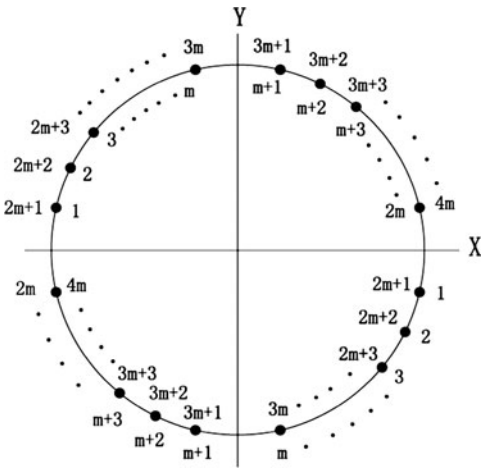


Figure 81.2 has a cnum matrix for Fig. 81.1 and contains 20 nonzero values. Thus, at least 10 adjacent swaps are needed to sort it.

Lemma 81.3. *A lower bound for sorting circular permutation π of length n by adjacent swaps is* $(1/4)n^2$.

Proof. Without loss of generality, we assume n to be a multiple of 4, i.e., $n = 4m$.

Consider a special cperm σ as shown in Fig. 81.3. The elements of σ are

$$\sigma_i = \begin{cases} (n/2) + i, & \text{if } i \in [1, (n/2)], \\ i - (n/2), & \text{if } i \in [(n/2) + 1, n]. \end{cases}$$

The dvec x for σ is (n/2,. . .,n/2, $-n$/2,. . .,$-n$/2). So, $\sum_{i=1}^{n} |x_i| = (1/2)n^2$ (say S). Thus, by Lemma 81.2 the minimum number of steps to sort σ is $I(x) = S/2 = (1/4)n^2$.

We observe that a natural algorithm to sort a cperm by adjacent swaps is to exchange large numbers at the front of the permutation, with small numbers at the end using fewer steps than the number of inversions. (Here, inversions of a cperm are the inversions of the corresponding lperm that starts at position 1 and ends at position n.) That is, the total number of inversions gives the exact number of adjacent swaps to sort a lperm, i.e., one does not have an adjacent swap involving

π_1 and π_n (i.e., no *wrap around*). Also, we note that in a cperm, an element that is more than $n/2$ positions away from its destination can reach there in less than $n/2$ adjacent swaps by using wrap around. Lemma 81.4 specifies when the number of adjacent swaps can be reduced using wrap around. It describes when there is a sequence, say of k swaps that decreases the number of inversions by more than k.

Lemma 81.4. *If a permutation π of length n has a number x in position d_x and a number y in position $n - d_y$ and $x - y > d_x + d_y$, then π can be transformed by a sequence t of adjacent transpositions into a permutation π' such that $|t| <$ inversions(π) $-$ inversions(π').*

Proof. Let x and y be numbers in π at positions d_x and $n - d_y$, respectively. Furthermore, assume $x - y > d_x + d_y$ and that $x - y - (d_x + d_y)$ is the maximum for any pair x, y. In $d_x + d_y$ adjacent swaps, x can be moved to position 1, y can be moved to position n, and they can then be transposed. There are $x - 1$ numbers smaller than x in π, and there are $n - y$ numbers larger than y in π. Consequently, by moving x to position n (through numbers in positions $1,\ldots, d_{x\ 1}$ that are smaller than x), we eliminate at least $x - d_x$ inversions (i.e., inversions involving x and numbers smaller than x) and create at most $n - x$ new inversions (i.e., inversions involving numbers larger than x which now occur before x).

Note that, if any one of the numbers, say x', in positions $1,\ldots,d_{x\ 1}$ were larger than x, then $x' - y - (d_x + d_y)$ would be larger than $x - y - (d_x + d_y)$, which contradicts our assumption that $x - y - (d_x + d_y)$ is maximum.

Similarly, by moving y to position 1 (through numbers in positions $n - d_y,\ldots,n$ that are larger than y), we eliminate at least $n - y - d_y$ (i.e., inversions involving numbers larger than y and y itself) and create at most y new inversions (i.e., inversions involving numbers smaller than y which now occur after y). No other inversions are created or destroyed through this sequence of swaps. So the number of inversions decreases by $(2x - d_x - n) + (n - 2y - d_y) = 2(x - y) - (d_x + d_y)$.

As the number of adjacent swaps used is $d_x + d_y$, it follows that the new permutation π' is obtained by a sequence t of adjacent swaps whose length is less than the change in the inversion count, namely inversions(π) $-$ inversions(π').

The proof shows that the decrease in the number of inversions, namely, $2(x - y) - (d_x + d_y)$ must be greater than the number of steps $(d_x + d_y)$. This is so, when $2(x - y) > 2(d_x + d_y)$, i.e., $x - y > d_x + d_y$.

Conjecture 81.1. *The upper bound on the number of inversions of circular permutation π of length* n *with wrap around is (1/4)*n^2.

Justification. *We note that the reverse order* R = (n, n − 1,. . ., 1) *which has the maximum number of* (1/2)(n^2 − n) *inversions for a lperm can be decomposed into two independent lperms* A = (3n/4, 3n/4 − 1,. . ., n/4 + 1) *and* B = (n/4,. . ., 1, n, n − 1,. . ., 3n/4 + 1) *due to wrap around.* A *is a reverse order of size* n/2 *(where* n/4 *is added to each element) with* (1/8)(n^2 − n) *inversions.* B *needs to be transformed to* (3n/4 + 1,. . ., n, 1,. . ., n/4 − 1, n/4)*, which is reverse of* B*; thus,* B *also has* (1/8) (n^2 − n) *inversions. So* R *can be sorted in* (1/4)(n^2 − n) *adjacent swaps.*

The maximum number of inversions (that we found) is $(1/4)n^2$ *for* $(n/2 + 1,\ldots,1, n, 1, 2,\ldots, n/2)$.

Conjecture 81.2. *An upper bound for sorting an arbitrary circular permutation* π *of length* n *by adjacent swaps is* $(1/4)n^2$.

Justification. *Let* π *be a permutation such that the minimum number of adjacent swaps (with wrap around) to sort it is exactly inversions* (π). *In other words, by Lemma 81.4, the numbers* x *and* y, *respectively, in positions* d_x *and* $n - d_y$, *respectively, must satisfy* $x - y = d_x + d_y$. *Let* i *be the number in position* n. *It follows that the number in position* $j\ (1 \le j \le n/2)$ *must be no larger than* $i + j$. *Let the number in position* 1 *be* k, *so* $k \le i + 1$, *it follows that the number in position* $n - j\ (1 \le j \le n/2)$ *must be at least* $k - j - 1$. *It follows that the permutation* π *which requires inversions* (π) *adjacent swaps is* $(n/2 + 1,\ldots, 1, n, 1, 1,\ldots, n/1)$ *which has* $(1/4)n^2$ *inversions (which is also the maximum number that we found).*

81.3 Sorting by Short Swaps

A short swap exchanges two elements with at most one element in between. We study upper and lower bounds for sorting cperms by short swaps. We introduce necessary notations and results from [3]. For a lperm π of length n, for each element $a = \pi_i\ (1 \le i \le n)$, we draw its *vector* as a directed line that begins at position i and ends at position a. We call the resulting diagram a *vector diagram* of π, denoted as V_π. The total length of the vectors in V_π, called *vector sum* is denoted as $|V_\pi|$, can be computed by $\sum_{i=1}^{n} |\pi_i - i|$. Figure 81.4 shows V_π (=14) for $\pi = (4\ 3\ 5\ 1\ 6\ 2)$. Two elements π_i and π_j are called *m vector-opposite* if their vectors have opposite directions and satisfy the relation $\pi_j \le i \le j \le \pi_i$ where $m = j - i$. In Fig. 81.4, π_3 and π_4 are 1-vector opposite π_1 and π_4 are 3-vector opposite. The following properties shown for lperms in [3] are also applicable to cperms (by not using wrap around):

$$\text{For any lperm } \pi,\ |V_\pi| \leqslant \lfloor n^2/2 \rfloor. \tag{81.4}$$

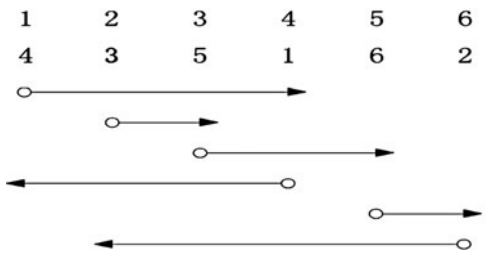


Fig. 81.4 Vector diagram

For unsorted lperm π, there is at least one pair of m vector - opposite elements. (81.5)

Lemma 81.5. *A lower bound for sorting circular permutation of length n by short swaps is* $(1/8)n^2$.

Proof. According to (81.4), for any cperm π, $V_\pi \leqslant \lfloor n^2/2 \rfloor$. The permutation $(n, n-1, \ldots, 1)$ has $|V_\pi| = \lfloor n^2/2 \rfloor$. By a short swap, one can reduce $|V_\pi|$ by at most 4. Thus, one needs at least $(1/8)n^2$ short swaps.

According to (81.5), for any unsorted cperm π, there exists at least one pair of m-vector ($m \geq 1$) opposite elements. By transposing two m-vector opposite elements, one reduces $|V_\pi|$ by at least 2 in each step [3]. Thus, transposing vector opposite elements yields a 2-approximation algorithm for sorting a cperm by short swaps.

Theorem 81.1. *An upper bound for sorting a circular permutation* π *of length n by short swaps is* $(5/32)n^2 + O(n \log n)$.

Proof. For a cyclic permutation π of length n, without loss of generality, assume $n = 4k$ and assume that the identity permutation is as shown in Fig. 81.3. Since $n = 4k$, the X-axis and Y-axis, as shown, divide the permutation into four equal size parts. Each part has k numbers. Part 1 contains the positions $(1, \ldots, k)$, etc. We call the partial permutation above (below) X-axis the *upper* (*lower*) *half* and the partial permutation to the left (right) of Y-axis the *left* (*right*) *half*. We call the value range $[1, 2k]$ as *range_low* ($[2k+1, 4k]$ as *range_high*). We call the numbers of *range_low* positioned in the lower half as *misplaced_low*. Likewise, we call the numbers of *range_high* positioned in the upper half as *misplaced_high*.

Assume that *misplaced_low* has m numbers (*misplaced_high* will also have m numbers). First, with short swaps we exchange *misplaced_low* with *misplaced_high*. Then all the numbers of *range_low* are in the upper half, and all the numbers in *range_high* are in the lower half yielding two independent lperms, one in each half. For lperms, the upper bound shown in [4] by Feng et al. holds.

Among the *misplaced_high*, assume that b_1 of them are in Part 1, b_2 of them are in Part 2, and in *misplaced_low* assume that b_3 of them are in Part 3, and b_4 of them are in Part 4. Clearly, $b_1 + b_2 = b_3 + b_4 = m$. There are 16 inequalities among b_1, b_2, b_3 and b_4. We consider Case 1: $b_1 \leq b_3 \leq b_4 \leq b_2$. Similar analysis applies to other 15 cases. We exchange *misplaced_low* with *misplaced_high* by moving the $2m$ numbers to positions close to X-axis and exchanging them across the X-axis. Figure 81.5 shows the worst case where the numbers to be gathered near the X-axis are far away from it.

As shown in Fig. 81.5, we move b_4 numbers of *range_high*, where b_1 numbers are in Part 1 and $b_2 - b_3 = b_4 - b_1$ numbers are in adjacent segment of Part 2, to positions $1, 2, \ldots, b_4$. Similarly, we move b_4 numbers of *range_low* in Part 4 to

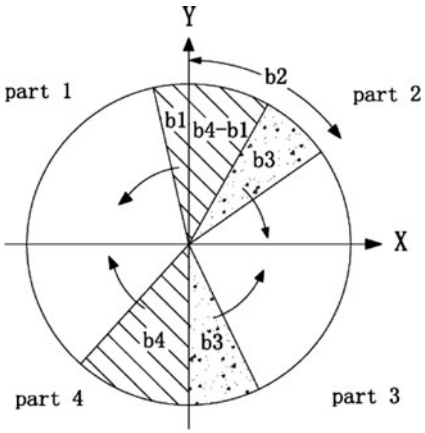


Fig. 81.5 A worst case scenario

positions $4k - b_1 + 1$, $4k - b_1 + 2$,. . .$4k$. Then, the b_4 numbers above X-axis in Part 1 and b_4 numbers below X-axis in Part 4 are exchanged.

We move b_3 numbers of *range_high* in Part 2 close to the X-axis in a clockwise direction to positions $2k - b_3 + 1$, $2k - b_3 + 2$,. . ., $2k$. Likewise, we move b_3 numbers of *range_low* in Part 3 in counter-clockwise direction to positions $2k + 1$, $2k + 2$,. . ., $2k + b_3$. Then, the b_3 numbers above X-axis in Part 2 and b_3 numbers below X-axis in Part 3 are exchanged. During the procedure, we use short swaps. $O(1)$ adjacent swaps are used at the beginning or at the end when necessary.

The tasks that accomplish the above procedure are:

(1) Move b_4 numbers in Part 1 and Part 2 in at most $b_4 \times (n/4 - b_1)/2$ steps.
(2) Move b_4 numbers in Part 4 in at most $b_4 \times (n/4 - b_4)/2$ steps.
(3) Exchange the upper b_4 numbers with the lower b_4 numbers by short swaps in $b_4 \times (b_4/2)$ steps.
(4) Move b_3 numbers in Part 2 in at most $b_3 \times (n/4 - b_2)/2$ steps.
(5) Move b_3 numbers in Part 3 in at most $b_3 \times (n/4 - b_3)/2$ steps.
(6) Exchange the upper b_3 numbers with the lower b_3 numbers by short swaps in $b_3 \times (b_3/2)$ steps.

The summation of all the steps yields $(1/4)nm - (1/2)(b_2b_3 + b_1b_4)$. Since $b_1 + b_2 = b_3 + b_4 = m$, $b_1 = m - b_2$ and $b_3 = m - b_4$. By substituting b_1, b_3 into the formula above, we have a function $f(m, b_2, b_4)$ with three variables: $f(m, b_2, b_4) = (1/4)nm - (1/2)m(b_2 + b_4) + b_2b_4$.

By definition, $0 \leq m \leq n/2$. For Case 1, we have $0 \leq b_4 \leq b_2 \leq n/4$. We can prove that the function f has its maximum value of $(1/16)n^2$.

The steps shown above yield two independent lperms each of size $n/2$ (upper and lower half). We use the algorithm described in [4] to sort them separately in $2 \times (3/16)(n/2)^2 + O(n \log n)$ steps. The function f shown above has the maximum value of $(1/16)n^2$. So, in total the algorithm takes $(5/32)n^2 + O(n \log n)$ steps.

81.4 Sorting in Parallel by Short Swaps

A lower bound for sorting a cperm by short swaps in parallel is directly applicable to hybrid mesh ring network. We show a lower bound with the help of *Short-SwapParallelSort* algorithm. We use the terminology from Theorem 81.1 of Sect. 81.3. We solve the case depicted in Fig. 81.5; the rest of the cases are equivalent.

ShortSwapParallelSort (π, n)

1. Exchange the m numbers in the *misplaced_low* with m numbers in *misplaced_high*.
2. ParallelQuickSort$(\pi_1, \pi_2, \ldots, \pi_{n/2})$.
3. ParallelQuickSort$(\pi_{(n/2)+1}, \pi_{(n/2)+2}, \ldots, \pi_n)$.

Algorithm *ParallelQuickSort* is a parallel quicksort algorithm with short swaps. Each time, for a partial permutation π_i, π_{i+1},. . ., π_j, the algorithm picks $\lfloor (i+j)/2 \rfloor$ as the pivot and divides the permutation into two parts of equal size.

Let us consider shifting the elements (e_1, e_2,. . ., e_j) at positions (1, 2,. . ., j) counter clockwise (cc) by $k > 0$ positions in parallel. In the first timeslot (*slot*), e_1 and e_2 can be moved in parallel by short swaps to the positions $4m - 1$ and $4m$ ($p(4m - 1)$ and $p(4m)$) respectively. In the subsequent slots, these elements move along odd and even positions respectively, with short swaps of type (i, $i + 2$) (moving 2 positions cc per move). If e_3 makes a short swap in cc direction, i.e., (x, 3), then $x = 1$ or 2. In slot 1, (1, 3) and (2, 3) are detrimental as they prevent e_1 and e_2 from moving cc. Similarly, in slot 1, (2, 4) is detrimental and (3, 4) is not useful. Thus, e_3 and e_4 execute their first short swaps ((1, 3) and (2, 4)) in slot 2. It follows that e_j executes its first short swap in $j/2$ slot. It needs $k/2$ more slots to reach its destination (the last element) in slot $(j + k)/2$. Note that all fractional values of slots are rounded up.

ParallelQuickSort (π_i, π_{i+1},. . ., π_j)

1. pick pivot $p = \lfloor (i+j)/2 \rfloor$
2. $k_1 = i,\ k_2 = p,\ m = 0$ // $k_1 = p + 1,\ k_2 = j,\ m = 0$
3. **L1**) **while**$(\pi_{k1} \leq p)\ k_1 = k_1 + 1$ // **L1**) **while**$(\pi_{k1} \leq p)\ k_1 = k_1 + 1, m = m + 1$
4. **while**$(\pi_{k2} > p)\ k_2 = k_2 - 1, m = m + 1$ //**while** $(\pi_{k2} > p)\ k_2 = k_2 - 1$
5. if $(k_1 < k_2)$ start moving π_{k1} to k_2 and go to $L1$
   // if $(k_1 < k_2)$ start moving π_{k2} to k_1 and go to $L1$
6. In parallel, exchange m numbers to the left of position p with m numbers to its right in $O(m/2)$ time
7. ParallelQuickSort$(\pi_i, \pi_{i+1}, \ldots, \pi_p)$
8. ParallelQuickSort$(\pi_{p+1}, \pi_{p+2}, \ldots, \pi_j)$

The parallel short swaps required to sort the cperm depicted in Fig. 81.5 are described below. The remaining cases are sorted similarly.

First, Step 1 of *ShortSwapParallelSort* is executed followed by two copies of *ParallelQuickSort* that are executed in parallel. Step 1 of *ShortSwapParallelSort* has two independent threads T_1 and T_2 described hereunder. T_1: Step (a) Move b_4 numbers in Part 1 and Part 2 in counter clockwise direction close to X-axis. Step (b) Move b_4 numbers in Part 4 in clockwise direction close to X-axis. Step (c) Exchange the upper b_4 numbers with the lower b_4 numbers. T_2: Step (a) Move b_3 numbers in Part 2 in clockwise direction close to X-axis. Step (b) Move b_3 numbers in Part 3 in counter clockwise direction close to X-axis. Step (c) Exchange the upper b_3 numbers with the lower b_3 numbers.

Within each thread: Step (c) is dependent upon Steps (a) and (b) which are independent but partial execution of Steps (a) and (b) is sufficient for Step (c) to begin. If at least two elements a and b that need to be exchanged exist on either side of X-axis ($p(1)$ and $p(4m)$), then Step (c) can be executed. Otherwise, Step (c) will wait for Steps (a) and/or (b) to provide them (further explained below).

During Step (a) of T_1, in the slot $(n/4 - b_1)/2$, the first two elements arrive at $p(1)$ and $p(2)$, and then the rest follow. During Step (b) of T_1, in slot $(n/4 - b_4)/2$, the first two elements arrive at $p(4m)$ and $p(4m - 1)$. Because $b_4 \geq b_1$ the first two elements from Step (b) (the elements of Part 4) arrive no later than those from Parts 1 and 2. So in the slot $(n/4 - b_1)/2 + 1$, the short swaps $(2, 4m)$ and $(1, 4m - 1)$ (the first short swaps of Step (c)) can be executed exchanging two elements above the X-axis with two elements below the X-axis. In the subsequent slots, (an) two elements arrive(s) at X-axis from each (one) direction in slot t and are swapped across X-axis in slot $t + 1$ until all elements reach their respective halves. Note that elements might be arriving only from one side because the elements on the other side are already present. Thus, in every alternate slot two pairs of elements are exchanged across X-axis consuming b_4 slots. Thus, in $(n/4 - b_1)/2 + b_4$ slots all the elements are moved to and exchanged across X-axis. For illustration, consider the portion of permutation on either sides of X-axis (denoted by "|") on successive slots, where (1,. . ., 6) in Part 4 and (a,. . ., f) in Part 1 need to be exchanged across X-axis, * represents an element that is already in the correct part: (123456|**ab**cd**ef) → (123456|ab**cd**ef**) → (1234ab|56cd**ef****) → (12ab34|cd56ef******) → (ab12cd|34ef56******) → (abcd12|ef3456******) → (abcdef|123456******).

The cases where the elements to be exchanged across X-axis that are nearer to X-axis are easier. Also, the elements in Parts 1 and 4 can be exchanged in $n/4$ slots. Similar analysis holds for T_2. In Fig. 81.5, $b_4 \geq b_3$ and the elements in Parts 1 and 2 are farther away from X-axis than b_3 elements (of Parts 2 or 3). Thus, T_1 determines the parallel time given by $(n/4 - b_1)/2 + b_4$ with a maximum value of $3n/8$. Thus, Step 1 of *ShortSwapParallelSort* in $3n/8$ slots yields two lperms of size $n/2$, (in upper and lower halves) which can be sorted in parallel by quicksort. All calls to quicksort are made on disjoint positions. So they can execute in parallel.

The parallel time of recursive quicksort is bounded by its longest thread. The first quicksort call is on the array of length $s(=n/2)$. Let $k \leq s/2$ be the number of elements that need to be exchanged across pivot position. If $k = s/2$, then as described earlier, this can be executed in $s/2$ slots. Otherwise, in at most $(s/2 - k)/2$ slots pairs of

elements can be exchanged as described in Step 1 of *ShortSwapParallelSort* and finish in $s/4 + k/2$ slots (maximum value of $s/2$). So in $s/2$ moves, we obtain a problem of half the size yielding a recurrence $T(s) = T(s/2) + s/2 + O(1)$ with solution $T(s) = s + O(\log s)$ ($s = n/2$). Thus, the total parallel time of *ShortSwapParallelSort* is $3n/8 + n/2 + O(\log n) = 7n/8 + O(\log n)$. Theorem 81.2 directly follows from this result.

Theorem 81.2. *A circular permutation can be sorted in parallel by short swaps in* $7n/8 + O(\log n)$ *time*.

81.5 Conclusions and Open Questions

We prove new bounds for sorting circular permutations by adjacent and short swaps. We give a lower bound of $(1/4)n^2$ for adjacent swaps. We identify a class of circular permutations that can be sorted by fewer adjacent swaps than their inversion number, based on this, we conjecture an upper bound of $(1/4)n^2$. We design an algorithm with $(5/32)n^2 + O(n \log n)$ upper bound for short swaps which can be extended for sorting by k-bounded ($k > 3$) transpositions. We also give an upper bound of $7n/8 + O(\log n)$ for sorting by short swaps in parallel. It is not known whether our bounds for short swaps are optimal.

References

1. Pevzner, P.A. and Tesler, G. Genome rearrangements in mammalian evolution: Lessons from human and mouse genomes. Genomes (2003) 13(1):37 45
2. Chitturi, B., Sudborough, I.H., Voit, W., and Feng, X. Adjacent Swaps on Strings, LNCS 5092. Springer, Berlin, pp. 299 308, 2008
3. Heath, L.S. and Vergara, J.C. Sorting by short swaps. Journal of Computational Biology (2003) 10(4):75 89
4. Feng, X., Sudborough, I.H., and Lu, E. A fast algorithm for sorting by short swap. Proceeding of the 10th IASTED International Conference on Computational and Systems Biology, pp. 62 67, 2006
5. Jerrum, M. The complexity of finding minimum length generator sequences. Theoretical Computer Science (1985) 36:265 289
6. Pevzner, P.A. Computational Molecular Biology: An Algorithmic Approach. MIT, Cambridge, 2000
7. Mahajan, M., Rama, R., and Vijayakumar, S. On sorting by 3 bounded transpositions. Discrete Mathematics (2006) 306(14):1569 1585
8. Bruijn, N.G. Sorting by means of swapping. Discrete Mathematics (1974) 9:333 339
9. Hartman, T. and Shamir, R. A simpler 1.5 approximation algorithm for sorting by transposi tions. Proceedings of the 14th Symposium on Combinatorial Pattern Matching (CPM), LNCS 2089. Springer, Berlin, pp. 156 169, 2003

10. Walter, M.E., Dias, Z., and Meidanis, J. Reversal and transposition distance of linear chromo somes. Proceedings of SPIRE'98 String Processing and Information Retrieval: A South American Symposium, pp. 96 102, 1998
11. Solomon, A., Sutcliffe, P., and Lister, R. Sorting circular permutations by reversal. Proceed ings of the Workshop on Algorithms and Data Structures, Carleton University, Ottawa, Canada, LNCS 2748, Springer, Berlin, pp. 319 328, 2003

Chapter 82
Annotating Patents with Medline MeSH Codes via Citation Mapping

Thomas D. Griffin, Stephen K. Boyer, and Isaac G. Councill

Abstract Both patents and Medline are important document collections for discovering new relationships between chemicals and biology, searching for prior art for patent applications and retrieving background knowledge for current research activities. Finding relevance to a topic within patents is often made difficult by poor categorization, badly written descriptions, and even intentional obfuscation. Unlike patents, the Medline corpus has Medical Subject Heading (MeSH) keywords manually added to their articles, giving a medically relevant taxonomy to the 18 million article abstracts. Our work attempts to accurately recognize the citations made in patents to Medline-indexed articles, linking them to their corresponding PubMed ID and exploiting the associated MeSH to enhance patent search by annotating the referencing patents with their Medline citations' MeSH codes. The techniques, system features, and benefits are explained.

Keywords MeSH · Pubmed · Patent · Medline · Citation

82.1 Introduction

To be successful today, biotechnology companies need a full understanding of the patent landscape for a particular biotechnology field. Patent documents contain a great deal of information that can be useful for making decisions regarding investments of resources, avoiding potential infringement and understanding the state-of-the-art. However, they are usually not written with the intent of serving as a resource for information and research [1], but rather to fulfill a mix of legal

T.D. Griffin (✉)
IBM Almaden Research Center, San Jose, CA 95120, USA
e mail: tdg@us.ibm.com

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 82,

requirements for protection. No common language terms are used within industry subfields. Keywords are usually not standardized or possibly not present at all. The text itself can be written to intentionally obfuscate the meaning or may simply be difficult to understand.

The PubMed database [2] is another document set that is highly important to biotechnology, containing over 18 million bibliographic abstracts for articles relevant to the NIH. The Medline portion of PubMed is annotated by the US National Library of Medicine (NLM) with Medical Subject Heading (MeSH) Codes [3]. Staff at the NLM index the Medline-covered articles using the MeSH codes, applying this hierarchical, controlled language across the entire set of articles. Having the MeSH index over the PubMed data assists users to find articles relevant to their field or particular search topic. Efforts have been made to try to automate the application of MeSH codes to the articles [4], but the articles are generally written to communicate their ideas and often peer reviewed for such clarity. With patents, this is not the case.

Creating, tracking, and mining patent information is a critical aspect of business for IBM and their customers. To support IP-related activities, IBM has created the Strategic Information Mining Platform for IP Excellence (SIMPLE) [5]. It includes a full-document patent database with over 11 million US, EP, and WO patents and applications, as well as a full PubMed database, containing over 18 million PubMed abstracts. Both databases have been indexed with a solr/lucene full-text index [6] that supports fielded search, wildcards, and other features. On top of these components sits analytic services [7] and an end user accessible website. The patent and PubMed parts of the service have been independent and only loosely coupled at the interface level. Many patent documents contain an "Other Publications" section that lists citations from inventors and examiners that are made to nonpatent prior art. Often these are references to articles in the scientific literature, and often these citations are for articles that are present in the PubMed index. Knowing when that is the case is not obvious, but if it were available, it would permit deeper integration of the patent and PubMed data sets. The goal is to create a mapper from free-text citations in patents to the specific PubMed articles they refer, and then exploit the MeSH headings of the linked-to articles [8].

82.2 Recognizing Patent References to PubMed Articles

The mapper's technique to match free text citations in patents to entries in the PubMed article database relies upon three steps. First, parse the free text citation in the patent into fields using the most consistent field rules available, being careful to have high confidence in all rules that have matched. Second, use the labeled fields to build queries against the standardized and accurately fielded PubMed search index and database, looking to end up with a single hit from the index that matches the available field information. Finally, to reduce false positive matches, eliminate

WO	WO 92/08777	5/1992
WO	WO 92/09291	6/1992
WO	WO-93/10776	6/1993
WO	WO 93/25211	12/1993
WO	WO-94/03147	2/1994
WO	WO-94-06440	3/1994
WO	WO 94/09798	5/1994
WO	WO 94/13305	6/1994
WO	WO 95/26198	10/1995
WO	WO-96/19184	6/1996

OTHER PUBLICATIONS

Cameo Chemical Data Sheet, "Strontium Sulfide".*
Bilotto, Gerardo, et al., "Effects of Ionic and Non-Ionic Solutions on Intradental Nerve Activity in the Cat", *Pain*, 32:231-38 (1988).
Celerier, et al., "Modulatory Effects of Selenium and Strontium Salts on Keratinocyte-Derived Inflammatory Cytokines", *Arch. Dermatol. Res.*, 287:680-82 (1985).
Brooks, et al., "Cutaneous Allergy to Insulin Preparations", *Pract. Diabetes*, 11(6):236-238, 1994.
Budavari, Susan, et al. (Eds.), *The Merck Index*, (Merck & Co., Inc., 1989), pp. 708, 869-872, 1211-1219, 1298-1299, 1395, 1576.
Cohen, et al., "Safety and Efficacy of Human IgE Pentapeptide", *Ann Allergy*, 52(2):83-86 (1984).
Degroot, *Unwanted Effects of Cosmetics and Drugs Used in Dermatology*, 1995, p. 229.
Dolynchuk, et al., "Tropical Purtescine (Fibrostat) in Treatment of Hypertrophic Scars: Phase II Study", *Plast. Reconstr. Surg.*, 97(1):117-123, Jan. 1996.
Drozdiak, W., "Keeping Svelte and Healthy With Vitamin Sea", *The Washington Post*, Issued Mar. 25, 1996, Section A, p. 11.
Drug Facts and Comparisons, 1995 Edition (Facts and Comparisons, St. Louis, MO), pp. 37-38.
Fitzpatrick, Thomas B., et al., (Eds.), *Dermatology in General Medicine*, (McGraw-Hill, Inc. 1993) 4th ed., vol. 1, pp. 501, 508-510, 1393-1396.

Fig. 82.1 "OTHER PUBLICATIONS" Section from the US Patent 7404967

from consideration citations made in patent documents that can be obviously labeled as specific non-PubMed document types, such as GenBank accessions, Derwent World Patent Index (DWPI) references, or PCT patent application references.

The US Patent "Other Publications" section is a catch-all of all nonpatent references made by the inventor or the examiner. Figure 82.1 shows an example, where the first reference is simply the title of a chemical company's data sheet, the second an article reference to the Medline-indexed journal *Pain*. The goal is to recognize references to items indexed in PubMed whenever possible and create a direct reference to the PubMed ID number (PMID) for that article.

82.2.1 Start by Labeling Easily Found Fields

Inspection of numerous examples of citations led to a few observations from which reliable mapping rules were written. First and most directly, some patents make direct use of PMID numbers in their citations, usually preceded by the string "PMID" or "PubMed". These are the easiest references to get right, as they do not even require a PubMed database or index to confirm assumptions. The patent of Fig. 82.1 includes this later reference:

Grynpas et al., "Strontium Increases Vertebral Bone Volume in Rats at a Low Dose That Does Not Induce Detectable Mineralization Defect," *Bone*, 1996, 253 9, 18(3) PMID 8703581.

The "PMID 8703581" is a direct inclusion of the PMID in the citation. This was the basis of the first rule for matching citations to PMID if the citation contains the string "pmid" (or "pubmed"), then any five digit or larger integer following is assumed to be the PMID directly. The five digit rule excludes PMIDs in the 1 9,999

range, but that is a tradeoff to eliminate false matches on non-IDs, especially page counts. This rule maps references to PMIDs in 947 US patents.

For the remaining citations, the next step is to label parts of the citation as specific fields which can be used to query the PubMed index to try and find a unique match. Instead of attempting to label all fields, such as publish date, authors, and journal name and title, the focus is on particular elements that are most easily correctly labeled due to regular syntax. The easiest field to label is the year of publication, labeled as pubyear. The pubyear rule is to match any integer in the range of 1900 2009. The java code is included below and indicative of the style of other regular-expression-based rules:

```
String tagPubyear(String full_citation_in) {
   String ret_string = full_citation_in.replaceAll("([^-A-Za-z0-
9])(19[0-9][0-9])([^-A-Za-z0-9])",    "$1<pubyear>$2</pubyear>$3");
   ret_string = ret_string.replaceAll("([^-A-Za-z0-9])(200[0-
9])([^-A-Za-z0-9])", "$1<pubyear>$2</pubyear>$3");
   return(ret_string);
}
```

Pubmonth is the published month. A rule looks at the text immediately before any pubyear for an appropriate integer or three character month abbreviations. Volume and pages fields are labeled with rules looking for certain punctuation, abbreviations, and integer patterns. The rules are designed to match as many cases as possible without being too broad. When in doubt, rules are made more specific so that there is higher confidence in each field label, even if only a few fields are labeled.

Author names are another key field for which syntax is fairly consistent. For papers with more than one author, only the first author is listed with their last name and possibly first initials, followed by some variant of “et al.”. The “et al.” is the critical key from which the author can be positioned within the full citation and the end of the authors area known for sure. The mapper has rules to find a word and possibly first initials before any “et al.” variant in the citation, label it as the author and label the first word of that section as “firstauthor lastname”. When “et al.” is not found, say for single-author papers or ones where multiple authors have been listed explicitly, a secondary author rule is applied which looks for a word followed by initials at the very start of the citation, as convention arranges most citations with the authors field first.

Once fields have been labeled, certain found fields and their field names are combined via “AND” to form a fielded query against the PubMed index, looking to find a single result returned. Here is an example of a raw citation and the query created from the initial labeling of the fields:

Bilotto, Gerardo, et al., “Effects of Ionic and Non-Ionic Solutions on Intradental Nerve Activity in the Cat”, *Pain*, 32:231 38 (1988).

Query = author main1:“Bilotto, Gerardo” AND pubyear:1988 AND journal volume:32

That query returns a single hit: PMID = 3362559. Note the use of the author main1 field that searches the first author listed for a PubMed article. By searching only the first author field, the search space for authors is greatly narrowed, which improves disambiguation. As a minimum threshold of recognition, the initial query is only attempted when at least an author and pubyear have been labeled. When this initial query results in one hit, the mapper is done and the whole citation is linked with the returned PMID.

82.2.2 Refine When Necessary

If the initial query has zero or more than one result, the mapper tries using more broad terms to widen the search, or adds terms to narrow the search. In the case of no initial result, the author value is changed to only include the last name without any initials or positional first name, assuming a last name and first name order. The query is repeated with just last name for author main1. If exactly one hit is returned, the mapper succeeds. If no hits are returned, the entry is abandoned. If there is more than one hit, for this stage or after the initial search stage, terms are added to narrow the search results.

The main way for adding a term is to assume that the first word except "the" after the labeled authors is the start of the article title. This is a crude title labeler, but the method is only used when search is still finding more than one possible article and the eligible article set needs to be narrowed. In the example given above, if the initial query had returned more than one hit instead of exactly one hit, the query would be retried as this:

Query = author main1:"Bilotto, Gerardo" AND pubyear:1988 AND journal volume:32 AND title:Effects

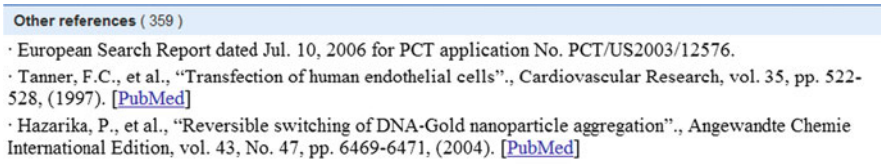
Once there is a single-hit result, the mapper performs one final validation check. If there is a labeled "pages" section in the citation and the pages value for the tentative PMID article match is known, the mapper compares the first integer found in both. If this starting page number matches, the result is approved, but otherwise it is rejected. This final check helps to eliminate false positive cases, where either the initial query of author and pubyear may have coincidentally matched, or where a refined query, including title, may not have led to a true match but rather a similar year/author/first title word from some other journal.

82.2.3 Eliminate Obvious Non-PubMed References

Another mechanism for avoiding false positives is to spot frequently occurring citation types that are not PubMed and simply not try to match them at all. Several categories of these references include Derwent World Patent Index references

Table 82.1 Document counts for some non PubMed reference types

Reference type	Document count
PCT	60,122
DWPI	27,313
GenBank	6,065
All documents	4,195,929


Other references (359)

· European Search Report dated Jul. 10, 2006 for PCT application No. PCT/US2003/12576.

· Tanner, F.C., et al., "Transfection of human endothelial cells"., Cardiovascular Research, vol. 35, pp. 522-528, (1997). [PubMed]

· Hazarika, P., et al., "Reversible switching of DNA-Gold nanoparticle aggregation"., Angewandte Chemie International Edition, vol. 43, No. 47, pp. 6469-6471, (2004). [PubMed]

Fig. 82.2 Screen view of the "Other references" section from patent US7332283

(DWPI), GenBank accession numbers and WO PCT application numbers. Table 82.1 lists counts for some of these easy to spot non-PubMed reference types.

The techniques used failed most often in cases where the reference information is very limited, possibly not including title. The base assumption about a pubyear and first author name being a good start toward a PubMed match breaks down when similar information appears in other types of references, like PCT references from the same author in the same year as other journal articles. Recognizing these frequently occurring non-article references helps reduce matching of incorrect article types. Note that in some cases, false positives may still yield useful information as they may well be from the same author, in the same year, and on the same topic as the target reference. Their MeSH headings could be quite similar to the actual cited article. The mapper still tries to do its best to avoid false positives.

82.3 Exploit the Linkages

After a successful match, the mapper creates an entry in a three column table, including patent number, citation number, and pubmed id. As a test set, the citation mapper was run against the 156,177 distinct patents issued from Jan 1, 2008 Sept 2009 that had "Other Publications" citations. There are 1,904,169 individual references in this set. The mapper matched 443,558 patent number/citation number/PMID triplets which were loaded into a new table. This table formed a bridge between the US patent and PubMed databases which enabled cross-linking and pulling of various fields from one database to records in the other. The application features created so far to exploit this linkage are a link on PubMed articles to show all referencing patents, a link on each citation a patent makes to the full record of

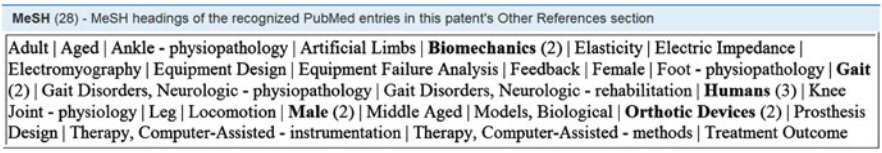


Fig. 82.3 Screen view of the "MeSH" section from patent US7431737

recognized PubMed articles, and a section on the patent record that summarizes all the MeSH headings seen for the recognized citations.

In the "Other references" section on our patent web application record view, a [PubMed] link was added to the end of all recognized citations, see Fig. 82.2. The links jump directly to each PubMed article's record view where fields, including Abstract and MeSH descriptor, are shown. The field labels found for each citation are not shown and only the end result of a linkage is made. This is because, for most matched citations, actually very few field sections are labeled, including never labeling a full title or journal name.

A "MeSH" section is added for patents with matched cited PubMed references, showing a summary of the MeSH descriptors found for the cited MeSH-containing PubMed articles, arranged alphabetically. See Fig. 82.3. The count is shown in parenthesis and the descriptor is bolded to provide some visual emphasis when there is more than one citation with the same MeSH value.

82.4 Conclusions and Future Work

Simple rules for labeling a few specific field sections in patent citations can create a citation mapper that matches raw citations in patents to well-fielded data from the PubMed article database. Once linked, the cited articles' MeSH terms are brought into the patent data space to amend PubMed and patent record views. Future plans include improving the mapper's field labeling rules by exploiting more frequently seen citation features like italics present around journal names and quotes around titles. We may add a separate engine such as ParsCite [9, 10] as a secondary field labeler.

References

1. Weaver, D. Don't overlook the rigorously reviewed novel work in patents. *Nature*, Vol 461, No 17. Sept 2009.
2. PubMed.gov. http://www.ncbi.nlm.nih.gov/pubmed/.
3. Introduction to MeSH 2010. http://www.nlm.nih.gov/mesh/introduction.html.

4. Rak, R., Kurgan, L., and Reformat, M. Multilabel Associative Classification Categorization of MEDLINE Articles into MeSH Keywords. *IEEE Engineering in Medicine and Biology Magazine*. pp. 47 55. Mar/Apr 2007.
5. Chen, Y., Spangler, W. S., He, B., Behal, A., Kato, L., Griffin, T., Alba, A., Kreulen, J., Boyer, S., Zhang, L., Wu, X., and Kieliszewski, C. SIMPLE: A Strategic Information Mining Platform for IP Excellence. *1st Workshop on Large scale Data Mining: Theory and Applications in Conjunction with ICDM09*. Miami, Florida, USA. Dec. 2009.
6. Apache Solr Project. http://lucene.apache.org/solr/.
7. Hasan, M., Spangler, W. S., Griffin, T. D., and Alba, A. "COA: Finding Novel Patents through Text Analysis", *Proceedings of the 15th ACM SIGKDD International Conference on Know ledge Discovery and Data Mining*, Paris, France 2009, pp. 1175 1184, 2009.
8. Angell, R., Boyer, S., Cooper, J., Hennessy, R., Kanungo, T., Kreulen, J., Martin, D., Rhodes, J., Spangler, W. S., and Weintraub, H. System and method for annotating patents with MeSH data, US Patent Application, Publication number US20070112833. May 17, 2007.
9. Councill, I. G., Lee Giles, C., and Kan, M. Y. "ParsCit: An open source CRF reference string parsing package", In *Proceedings of the Language Resources and Evaluation Conference* (LREC 08), Marrakesh, Morrocco, May 2008.
10. ParsCite: An open source CDF Reference String Parsing Package, http://wing.comp.nus.edu.sg/parsCit/.

Chapter 83
Some New Measures of Entropy, Useful Tools in Biocomputing

Angel Garrido

Abstract The basic problem rooted in Information Theory (IT) foundations (Shannon, Bell Syst Tech J 27:379 423 and 623 656, 1948; Volkenstein, Entropy and Information. Series: Progress in Mathematical Physics, 2009) is to reconstruct, as closely as possible, the input signal after observing the received output signal.

The Shannon information measure is the only possible one in this context, but it must be clear that it is only valid within the more restricted scope of coding problems that C. E. Shannon himself had seen in his lifetime (Shannon, Bell Syst Tech J 27:379 423 and 623 656, 1948). As pointed out by Alfred Rényi (1961), in his essential paper (Rényi, Proc. of the 4th Berkeley Symposium on Mathematics, Statistics and Probability, 547 561, 1961) on generalized information measures, for other sorts of problems other quantities may serve just as well as measures of information, or even better. This would be supported either by their operational significance or by a set of natural postulates characterizing them, or preferably by both. Thus, the idea of generalized entropies arises in scientific literature.

We analyze here some new measures of Entropy, very useful to be applied on Biocomputing (Ulanowicz and Hannon, Proc R Soc Lond B 232:181 192, 1987; Volkenstein, Entropy and Information. Series: Progress in Mathematical Physics, 2009).

Keywords Computational Biology · AI · Fuzzy Measure Theory

83.1 The Maximum Entropy Principle

The maximum entropy principle (MEP) is a postulate [1, 2] about a universal feature of any probability assignment on a given set of propositions: events, hypotheses, etc.

H.R. Arabnia (ed.), *Advances in Computational Biology*,
Advances in Experimental Medicine and Biology 680,
DOI 10.1007/978 1 4419 5913 3 83,

Let some testable information about a probability distribution function be given. Consider the set of all trial probability distributions that encode this information. Then, the probability distribution that maximizes the information entropy is the true probability distribution with respect to the testable information prescribed. This principle was first stated by Jaynes in his paper [1], where he emphasized some natural correspondence between Statistical Mechanics and Information Theory. He suggested that the Entropy denoted by S in Statistical Mechanics and H in Information Theory are basically the same thing. Consequently, Statistical Mechanics should be seen just as a particular application of a general tool of Logical Inference and Information Theory.

Given testable information, the maximum entropy procedure consists of seeking the probability distribution which maximizes the information entropy, subject to the constraints of the information. This constrained optimization problem will be typically solved using the analytical method known as Lagrange Multipliers.

Entropy maximization (max H), with no testable information, takes place under a single constraint: the sum of the probabilities must be one. Under this constraint, the maximum entropy probability distribution is the well-known U (uniform distribution). The MEP can thus be considered as a generalization of the classical Principle of Indifference (due to Laplace), also known as the Principle of Insufficient Reason.

The development of the idea of entropy of random variables and processes by Shannon gives support to the Ergodic Theory. Entropy and related information measures provide descriptions of the long-term behavior of random processes, and that this behavior is a key factor in developing aspects as the Coding Theorems of IT (proofs on the lower bounds), very interesting questions on Computational Biology, or in Ecology, modeling the competence between species.

83.2 Metric Entropy

We may also consider [3, 4] the so-called Metric Entropy, also called Kolmogorov Entropy, or Kolmogorov Sinai (K-S) Entropy. In a dynamical system, the metric entropy is equal to zero for nonchaotic motion, and is strictly greater than zero for chaotic motion.

In Thermodynamics, Prigogine entropy is a very frequently used term to refer to the splitting of Entropy into two variables, one being that which is "exchanged" with the environment, and the other being a result of "internal" processes. It holds

$$\mathrm{d}S = \mathrm{d}_{\mathrm{e}}S + \mathrm{d}_{\mathrm{i}}S \tag{83.1}$$

This expression is sometimes referred to as the Prigogine entropy equation. Such new function results, according to Prigogine, because the Entropy of a System is an Extensive Property, i.e., if the system consists of several parts, the total entropy is equal to the sum of the entropies of each part, and the change in entropy can be split

into two parts, being these d_eS and d_iS, denoting as d_eS the flow of entropy, due to interactions with the exterior, and denoting as d_iS the contributions due to changes inside the system.

83.3 Topological Entropy

Let (X, d) be a compact metric space, and let $f: X \to X$ be a continuous map.

For each $n > 0$, we define a new metric, d_n, by

$$d_n(\{x, y\}) = \max\left\{d\left(f^i(x), f^i(y)\right) : \ 0 < i < n\right\}. \tag{83.2}$$

Two points, x and y, are close with respect to this metric, if their first n iterates (given by f_i, $i = 1, 2, \ldots$) are close.

For $\varepsilon > 0$, and $n \in N^*$, we say that $S \subset X$ is a (n, ε)-separated set, if for each pair, x, y, of points of S, we have $d_n(x, y) > \varepsilon$.

Denote by $N(n, \varepsilon)$ the maximum cardinality of a (n, ε)-separated set. It must be finite because X is compact. In general, this limit may exist, but it could be infinite. A possible interpretation of this number is measure of the average exponential growth of the number of distinguishable orbit segments. So we could say that the higher the topological entropy is, the more essentially different orbits we have [3, 5, 6]. Hence, $N(n, \varepsilon)$ shows the number of "distinguishable" orbit segments of length n, assuming that we cannot distinguish points that are less than ε apart.

The topological entropy of f is then defined by

$$H_{\text{top}} = \lim_{r \to 0} \left\{ \limsup_{n \to \infty} (1/n) \log N(n, r) \right\}. \tag{83.3}$$

Topological entropy was introduced, in 1965, by Adler, Konheim, and McAndrew.

83.4 Entropy on Intuitionistic Fuzzy Sets

The notion of Intuitionistic Fuzzy Set (IFS) was introduced by Atanassov (1983), and then developed by authors as Hung and Yang, among others. Recall that an Intuitionistic Set is an incompletely known set. An IFS must represent [3, 5] the degrees of membership and nonmembership, with a certain degree of doubt. For this reason, they have been widely used in applied fields. Therefore, the apparition here of IFS, instead of FS, permits the introduction of another degree of freedom, in set descriptions. Such a generalization of FS gives us a new possibility to represent imperfect knowledge. Thus, we are able to describe many real problems in a more adequate way.

A very frequent measure of fuzziness is the Entropy of FSs, which was first mentioned by Zadeh (1965). But recently, two new definitions have been proposed,

by Szmidt and Kacprzyk (2001), and by Burillo and Bustince (1996). The first one is a nonprobabilistic entropy measure, which departs on a geometric interpretation of an IFS. And by the second, it would be possible to measure the degree of intuitionism of an IFS.

We can also generalize from IFS to a Neutrosophic Set, abridgedly *N-Set* [7], a concept due to Smarandache (1995). An IFS is a set that generalizes many previously existing classes of sets. In particular, FS and its first generalization, IFS.

Let U be a universe of discourse, and let M be a set included in U. An element x from U is denoted, with respect to the set M, as $x\ (T, I, F)$. And it belongs to M in the following way: it is $T\%$ in the set (membership degree); $I\%$ indeterminate (unknown if it is in the set); and $F\%$ not in the set (nonmembership degree). Here T, I, and F components are real standard/nonstandard subsets, included in the nonstandard unit interval, representing truth, indeterminacy, and falsity percentages, respectively. It is possible to define a measure of the Neutrosophic Entropy, abridgedly called *N-Entropy*, as the summation of the respective entropies of three subsets, T, I, and F.

83.5 Negentropy

Thermodynamics is usually considered as the keystone of modern science, and there are few principles that have generated more controversies [2, 4, 8, 9] than the Second Law of Thermodynamics (2nd LT).

Until recently, it was thought that the process of life was in unavoidable contradiction with this law. But such "contradiction" is only apparent. The 2nd LT makes no distinction between living and nonliving things. But living things are characterized by a very high degree of assembly and structure. Every isolated system (as may be the Universe) moves toward a state of maximum entropy. Its entropy can never decrease. Hence, the decrease in entropy that accompanies the growth of living structures must be accompanied by an increase in entropy in the physical environment. In his famous book *What is life?* Erwin Schrödinger [8] analyzed the life as a state of very low probability because of the necessary energy to create and sustain it. The "vital force" that maintains life is energy.

Living things preserve their low levels of entropy throughout time because they receive energy from the surroundings in the form of food [9]. They gain its order at the expense of disordering the nutrients they consume.

The entropy of a system represents the amount of uncertainty one observer has about the state of the system. The simplest example of a system will be a random variable. Information measures the amount of correlation between two systems, and it reduces to a mere difference in entropies.

As islands of order in a sea of chaos, organisms are far superior to human-built machines [2, 9]. And the body concentrates order. It continuously self-repairs, being the metabolism a sure sign of life. Brillouin says that a living system imports negentropy and stores it. So the denoted by J (negentropy) function is the entropy

that it exports to keep its own entropy low. The negentropy is also useful as a measure of distance to normality. It is always nonnegative, $J > 0$. And it will be also linear by any linear invertible change of coordinates. It vanishes if and only if the signal is Gaussian.

Some biologists speak in terms of the entropy of an organism [9], or about its antonym, negentropy, as a measure of the structural order within such organism. Being entropy defined as a measure of how close a system is to equilibrium, i.e., to perfect internal disorder. Living things and ordered structures appear because entropy can be lowered locally by a external action. It may be reduced in different systems, as the cold chamber of a refrigerator, being such a case possible by an increase of S in their surroundings.

83.6 Graph Entropy

Graph Entropy is a functional on a graph, $G = (V, E)$, with P a probability distribution on its node (or vertex) set, V. It is denoted by GE. Such concept [10] was introduced as the solution of a coding problem formulated on Information Theory. Because of its sub-additivity, it has become a useful tool in proving some lower bounds that result in Computational Complexity Theory. The search for exact additivity has produced certain interesting combinatorial structures. One of such results is the characterization of perfect graphs by the additivity of GE. It is defined as

$$H(\{G,P\}) = \min\left\{\sum p_i \log p_i\right\}. \tag{83.4}$$

Note that such function is convex. It tends to $+\infty$ on the boundary of the nonnegative orthant of $\boldsymbol{R}n$. And also tends monotonically to $-\infty$ along the rays from the origin. So such minimum is always achieved and it will be finite.

Statistical entropy is a probabilistic measure of uncertainty, or ignorance about data, whereas *Information* is a measure of a reduction in that uncertainty [4, 11, 12]. The Entropy of a probability distribution is just the expected value of the information of such distribution. All these ideas provide us with improved tools to advance not only in Computing Biology [9], but also in Modeling and Optimization, Economics, Ecology, and so on.

References

1. Jaynes TE (1948). Probability Theory: The Logic of Science. Cambridge University Press, Cambridge, current edition 2003.
2. Prigogine I (1947). Etude Thermodynamics des Phenomenes Irreversibles. Dunod editeur, Paris.

3. Garmendia L (2005). The Evolution of the Concept of Fuzzy Measure. Studies in Computa tional Intelligence. Springer Verlag, Berlin, 5: 186 200.
4. Volkenstein F (2009). Entropy and Information Series: Progress in Mathematical Physics, 57, Birkhäuser Verlag, Basel.
5. Burillo P, Bustince H (1996). Entropy on intuitionistic fuzzy sets, and interval valued fuzzy sets. Fuzzy Sets and Systems 78: 305 316.
6. Garrido A (2006). Additivity and Monotonicity in Fuzzy Measures. Plenary Talk in ICMI45, at Bacau University, then published at SCSSM journal, 16: 445 458.
7. Smarandache F (1999). A Unifying Field in Logics. Neutrosophy: Neutrosophic Probability, Set, and Logic. American Research Press, Rehoboth, USA.
8. Schrödinger E (1992). What is Life? Cambridge University Press, Cambridge.
9. Ulanowicz RE, Hannon BM (1987). Life and the production of entropy. Proc. of the Royal Society of London. Series B 232: 181 192.
10. Simonyi C (2002). Graph Entropy: A Survey. DIMACS Series in Discrete Mathematics and Theoretical Computer Science.
11. Rényi A (1961). On measures of information and entropy. Proc. of the 4th Berkeley Sympo sium on Mathematics, Statistics and Probability: 547 561.
12. Shannon CE (1948). A Mathematical Theory of Communication. Bell System Technical Journal 27: 379 423 and 623 656. Later, it appears as book, at Illinois Press, 1963. Also as E print.

Index