

Lecture Notes in Bioinformatics

4345

Edited by S. Istrail, P. Pevzner, and M. Waterman

Editorial Board: A. Apostolico S. Brunak M. Gelfand
T. Lengauer S. Miyano G. Myers M.-F. Sagot D. Sankoff
R. Shamir T. Speed M. Vingron W. Wong

Subseries of Lecture Notes in Computer Science

Nicos Maglaveras Ioanna Chouvarda
Vassilis Koutkias Rüdiger Brause (Eds.)

Biological and Medical Data Analysis

7th International Symposium, ISBMDA 2006
Thessaloniki, Greece, December 7-8, 2006
Proceedings

Series Editors

Sorin Istrail, Brown University, Providence, RI, USA

Pavel Pevzner, University of California, San Diego, CA, USA

Michael Waterman, University of Southern California, Los Angeles, CA, USA

Volume Editors

Nicos Maglaveras

Ioanna Chouvarda

Vassilis Koutkias

Aristotle University

The Medical School

Lab. of Medical Informatics – Box 323

54124 Thessaloniki, Greece

E-mail: {nicmag,ioanna,bikout}@med.auth.gr

Rüdiger Brause

J.W. Goethe-University

Department of Computer Science and Mathematics

Institute for Informatics

Robert-Mayer Str. 11-15, 60054 Frankfurt, Germany

E-mail: rbrause@informatik.uni-frankfurt.de

Library of Congress Control Number: 2006937536

CR Subject Classification (1998): H.2.8, I.2, H.3, G.3, I.5.1, I.4, J.3, F.1

LNCS Sublibrary: SL 8 – Bioinformatics

ISSN 0302-9743

ISBN-10 3-540-68063-2 Springer Berlin Heidelberg New York

ISBN-13 978-3-540-68063-5 Springer Berlin Heidelberg New York

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

Springer is a part of Springer Science+Business Media

springer.com

© Springer-Verlag Berlin Heidelberg 2006

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India

Printed on acid-free paper SPIN: 11946465 06/3142 5 4 3 2 1 0

Preface

The area of medical and biological data analysis has undergone numerous transformations in recent years. On the one hand, the Human Genome Project has triggered an unprecedented growth in gathering new types of data from the molecular level, which has in turn raised the need for data management and processing and led to the exponential growth of the bioinformatics area. On the other hand, new bits of information coming from molecular data have started filling some long-standing gaps of knowledge, complementing the huge amount of phenotypic data and relevant medical analysis. Thus, bioinformatics, medical informatics and biomedical engineering disciplines, contributing to the vertical integration of knowledge and data, using information technology platforms and enabling technologies such as micro and nano sensors, seem to converge and explore ways to integrate the competencies residing in each field.

ISBMDA has formed a platform, enabling the presentation and integration of new research results adding huge value to this scientific endeavour. This new research information stems from molecular data and phenotypically-oriented medical data analysis such as biosignal or bioimage analysis, novel decision support systems based on new combinations of data, information and extracted knowledge and, innovative information technology solutions enabling the integration of knowledge from the molecules up to the organs and the phenotype. Thus, the 7th ISBMDA was a place for all the above mentioned competencies to come together, and for the discussion and synthesis of new approaches to our understanding of the human organism function on a multiscale level.

We would like to express our gratitude to all the authors who submitted their work to the symposium and gave the Technical Program Committee the opportunity to prepare a symposium of outstanding quality. This year we received 91 contributions and following a rigorous review procedure 44 contributions were selected for presentation at the symposium. The form of the 44 presentations selected was either oral (28 oral presentations) or poster (16 poster presentations). We would finally like to thank the Technical Program Committee and the reviewers who helped, for the preparation of an excellent program for the symposium.

December 2006

Nicos Maglaveras
Ioanna Chouvarda
Vassilis Koutkias
Rüdiger Brause

Organization

Symposium Chair

N. Maglaveras, Aristotle Univ. of Thessaloniki, Greece

Scientific Committee Coordinators

V. Maojo, Univ. Politécnica de Madrid, Spain

F. Martín-Sánchez, Institute of Health Carlos III, Spain

A. Sousa Pereira, Univ. Aveiro, Portugal

Steering Committee

R. Brause, J.W. Goethe Univ., Germany

I. Chouvarda, Aristotle Univ. of Thessaloniki, Greece

V. Koutkias, Aristotle Univ. of Thessaloniki, Greece

A. Malousi, Aristotle Univ. of Thessaloniki, Greece

A. Kokkinaki, Aristotle Univ. of Thessaloniki, Greece

Scientific Committee

A. Babic, Univ. Linköping, Sweden

R. Baud, Univ. Hospital of Geneva, Switzerland

V. Breton, Univ. Clermont Ferrand, France

J. Carazo, Univ. Autónoma de Madrid, Spain

A. Carvalho, Univ. São Paulo, Brazil

P. Cinquin, Univ. Grenoble, France

W. Dubitzky, Univ. Ulster, UK

M. Dugas, Univ. Munich, Germany

P. Ghazal, Univ. Edinburgh, UK

R. Guthke, Hans-Knoell Institut, Germany

O. Kohlbacher, Univ. Tübingen, Germany

C. Kulikowski, Rutgers Univ., USA

P. Larranaga, Univ. Basque Country, Spain

L. Ohno-Machado, Harvard Univ., USA

J. Luis Oliveira, Univ. Aveiro, Portugal

F. Pinciroli, Politecnico di Milano, Italy

D. Pisanelli, ISTC - CNR, Italy

G. Potamias, ICS - FORTH, Greece
M. Santos, Univ. Aveiro, Portugal
F. Sanz, Univ. Pompeu Fabra, Spain
W. Sauerbrei, Univ. Freiburg, Germany
S. Schulz, Univ. Freiburg, Germany
T. Solomonides, Univ. W. England, UK
C. Zamboulis, Aristotle Univ. of Thessaloniki, Greece
B. Zupan, Univ. Ljubljana, Slovenia
J. Zvarova, Univ. Charles, Czech Republic

Special Reviewers

P.D. Bamidis, Aristotle Univ. of Thessaloniki, Greece
A. Astaras, Aristotle Univ. of Thessaloniki, Greece
S. Kouidou, Aristotle Univ. of Thessaloniki, Greece
A. Malousi, Aristotle Univ. of Thessaloniki, Greece
A. Bezerianos, University of Patras, Greece
D. Perez, Univ. Politecnica de Madrid, Spain
M. Carcia-Remesal, Univ. Politecnica de Madrid, Spain
G. Calle, Univ. Politecnica de Madrid, Spain
J. Crespo, Univ. Politecnica de Madrid, Spain
L. Martin, Univ. Politecnica de Madrid, Spain
D. Manrique, Univ. Politecnica de Madrid, Spain

Table of Contents

Bioinformatics: Functional Genomics

HLA and HIV Infection Progression: Application of the Minimum Description Length Principle to Statistical Genetics	1
<i>Peter T. Hraber, Bette T. Korber, Steven Wolinsky, Henry A. Erlich, Elizabeth A. Trachtenberg, and Thomas B. Kepler</i>	
Visualization of Functional Aspects of microRNA Regulatory Networks Using the Gene Ontology	13
<i>Alkiviadis Symeonidis, Ioannis G. Tollis, and Martin Reczko</i>	
A Novel Method for Classifying Subfamilies and Sub-subfamilies of G-Protein Coupled Receptors	25
<i>Majid Beigi and Andreas Zell</i>	
Integration Analysis of Diverse Genomic Data Using Multi-clustering Results	37
<i>Hye-Sung Yoon, Sang-Ho Lee, Sung-Bum Cho, and Ju Han Kim</i>	

Bioinformatics: Sequence and Structure Analysis

Effectivity of Internal Validation Techniques for Gene Clustering	49
<i>Chunmei Yang, Baikun Wan, and Xiaofeng Gao</i>	
Intrinsic Splicing Profile of Human Genes Undergoing Simple Cassette Exon Events	60
<i>Andigoni Malousi, Vassilis Koutkias, Sofia Kouidou, and Nicos Maglaveras</i>	
Generalization Rules for Binarized Descriptors	72
<i>Jürgen Paetz</i>	
Application of Combining Classifiers Using Dynamic Weights to the Protein Secondary Structure Prediction – Comparative Analysis of Fusion Methods	83
<i>Tomasz Woloszynski and Marek Kurzynski</i>	
A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites	92
<i>George Tzanis, Christos Berberidis, and Ioannis Vlahavas</i>	
SPSO: Synthetic Protein Sequence Oversampling for Imbalanced Protein Data and Remote Homology Detection	104
<i>Majid Beigi and Andreas Zell</i>	

Biomedical Models

Markov Modeling of Conformational Kinetics of Cardiac Ion Channel Proteins	116
<i>Chong Wang, Antje Krause, Chris Nugent, and Werner Dubitzky</i>	
Insulin Sensitivity and Plasma Glucose Appearance Profile by Oral Minimal Model in Normotensive and Normoglycemic Humans	128
<i>Roberto Burattini, Fabrizio Casagrande, Francesco Di Nardo, Massimo Boemi, and Pierpaolo Morosini</i>	
Dynamic Model of Amino Acid and Carbohydrate Metabolism in Primary Human Liver Cells	137
<i>Reinhard Guthke, Wolfgang Schmidt-Heck, Gesine Pless, Rolf Gebhardt, Michael Pfaff, Joerg C. Gerlach, and Katrin Zeilinger</i>	
The Probabilities Mixture Model for Clustering Flow-Cytometric Data: An Application to Gating Lymphocytes in Peripheral Blood	150
<i>John Lakoumentas, John Drakos, Marina Karakantza, Nicolaos Zoumbos, George Nikiforidis, and George Sakellaropoulos</i>	
Integrative Mathematical Modeling for Analysis of Microcirculatory Function	161
<i>Adam Kapela, Anastasios Bezerianos, and Nikolaos Tsoukias</i>	
Searching and Visualizing Brain Networks in Schizophrenia	172
<i>Theofanis Oikonomou, Vangelis Sakkalis, Ioannis G. Tollis, and Sifis Micheloyannis</i>	

Databases and Grids

TRENCADIS – A Grid Architecture for Creating Virtual Repositories of DICOM Objects in an OGSA-Based Ontological Framework	183
<i>Ignacio Blanquer, Vicente Hernandez, and Damià Segrelles</i>	
Minimizing Data Size for Efficient Data Reuse in Grid-Enabled Medical Applications	195
<i>Fumihiko Ino, Katsunori Matsuo, Yasuharu Mizutani, and Kenichi Hagihara</i>	
Thinking Precedes Action: Using Software Engineering for the Development of a Terminology Database to Improve Access to Biomedical Documentation	207
<i>Antonio Vaquero, Fernando Sáenz, Francisco Álvarez, and Manuel de Buenaga</i>	
Grid-Based Knowledge Discovery in Clinico-Genomic Data	219
<i>Michael May, George Potamias, and Stefan Rüping</i>	

A Prospective Study on the Integration of Microarray Data in HIS/EPR	231
<i>Daniel F. Polónia, Joel Arrais, and José Luis Oliveira</i>	

Web Services Interface to Run Protein Sequence Tools on Grid, Testcase of Protein Sequence Alignment	240
<i>Christophe Blanchet, Christophe Combet, Vladimir Daric, and Gilbert Deléage</i>	

Semantics and Information Modelling

Integrating Clinical and Genomic Information Through the PrognoChip Mediator	250
<i>Anastasia Analyti, Haridimos Kondylakis, Dimitris Manakanatas, Manos Kalaitzakis, Dimitris Plexousakis, and George Potamias</i>	

OntoDataClean: Ontology-Based Integration and Preprocessing of Distributed Data	262
<i>David Perez-Rey, Alberto Anguita, and Jose Crespo</i>	

Language Modelling for the Needs of OCR of Medical Texts	273
<i>Maciej Piasecki and Grzegorz Godlewski</i>	

Biomedical Signal Processing – Time Series Analysis

The Use of Multivariate Autoregressive Modelling for Analyzing Dynamical Physiological Responses of Individual Critically Ill Patients	285
<i>Kristien Van Loon, Jean-Marie Aerts, Geert Meyfroidt, Greta Van den Berghe, and Daniel Berckmans</i>	

Time Series Feature Evaluation in Discriminating Preictal EEG States	298
<i>Dimitris Kugiumtzis, Angeliki Papana, Alkiviadis Tsimpiris, Ioannis Vlachos, and Pål G. Larsson</i>	

Symbol Extraction Method and Symbolic Distance for Analysing Medical Time Series	311
<i>Fernando Alonso, Loïc Martínez, Aurora Pérez, Agustín Santamaría, and Juan Pedro Valente</i>	

A Wavelet Tool to Discriminate Imagery Versus Actual Finger Movements Towards a Brain–Computer Interface	323
<i>Maria L. Stavrinou, Liviu Moraru, Polyxeni Pelekouda, Vasileios Kokkinos, and Anastasios Bezerianos</i>	

Biomedical Image Analysis and Visualisation Techniques

A Fully Bayesian Two-Stage Model for Detecting Brain Activity in fMRI.....	334
<i>Alicia Quirós, Raquel Montes Diez, and Juan A. Hernández</i>	
A Novel Algorithm for Segmentation of Lung Images	346
<i>Aamir Saeed Malik and Tae-Sun Choi</i>	
An Evaluation of Image Compression Algorithms for Colour Retinal Images	358
<i>Gerald Schaefer and Roman Starosolski</i>	
An Automated Model for Rapid and Reliable Segmentation of Intravascular Ultrasound Images	368
<i>Eirini Parissi, Yiannis Kompatsiaris, Yiannis S. Chatzizisis, Vassilis Koutkias, Nicos Maglaveras, M.G. Strintzis, and George D. Giannoglou</i>	

Biomedical Data Analysis and Interpretation

Supervised Neuro-fuzzy Clustering for Life Science Applications.....	378
<i>Jürgen Paetz</i>	
Study on Preprocessing and Classifying Mass Spectral Raw Data Concerning Human Normal and Disease Cases	390
<i>Xenofon E. Floros, George M. Spyrou, Konstantinos N. Vougas, George T. Tsangaris, and Konstantina S. Nikita</i>	
Non-repetitive DNA Sequence Compression Using Memoization	402
<i>K.G. Srinivasa, M. Jagadish, K.R. Venugopal, and L.M. Patnaik</i>	
Application of Rough Sets Theory to the Sequential Diagnosis	413
<i>Andrzej Zolnierak</i>	
Data Integration in Multi-dimensional Data Sets: Informational Asymmetry in the Valid Correlation of Subdivided Samples	423
<i>Qing T. Zeng, Juan Pablo Pratt, Jane Pak, Eun-Young Kim, Dino Ravnic, Harold Huss, and Steven J. Mentzer</i>	

Decision Support Systems and Diagnostic Tools

Two-Stage Classifier for Diagnosis of Hypertension Type	433
<i>Michał Wozniak</i>	
Handwriting Analysis for Diagnosis and Prognosis of Parkinson's Disease	441
<i>Atila Ünlü, Rüdiger Brause, and Karsten Krakow</i>	

A Decision Support System for the Automatic Assessment of Hip Osteoarthritis Severity by Hip Joint Space Contour Spectral Analysis . . .	451
<i>Ioannis Boniatis, Dionisis Cavouras, Lena Costaridou, Ioannis Kalatzis, Elias Panagiotopoulos, and George Panayiotakis</i>	
Modeling for Missing Tissue Compensator Fabrication Using RFID Tag in U-Health	463
<i>O-Hoon Choi, Jung-Eun Lim, Hong-Seok Na, and Doo-Kwon Baik</i>	
The Effect of User Factors on Consumer Familiarity with Health Terms: Using Gender as a Proxy for Background Knowledge About Gender-Specific Illnesses	472
<i>Alla Keselman, Lisa Massengale, Long Ngo, Allen Browne, and Qing Zeng</i>	
ICT for Patient Safety: Towards a European Research Roadmap	482
<i>Veli N. Stroetmann, Daniel Spichtinger, Karl A. Stroetmann, and Jean Pierre Thierry</i>	
Author Index	495

HLA and HIV Infection Progression: Application of the Minimum Description Length Principle to Statistical Genetics

Peter T. Hraber^{1,2}, Bette T. Korber^{1,2}, Steven Wolinsky³, Henry A. Erlich⁴,
Elizabeth A. Trachtenberg⁵, and Thomas B. Kepler⁶

¹ Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM 87501 USA

² Los Alamos National Laboratory, Los Alamos NM 87545 USA

³ Feinberg School of Medicine, Northwestern University, Chicago IL 60611 USA

⁴ Roche Molecular Systems, 1145 Atlantic Ave, Alameda CA 94501 USA

⁵ Children's Hospital Oakland Research Institute, Oakland CA 94609 USA

⁶ Department of Biostatistics and Bioinformatics, Duke University Medical Center,
Duke University, Durham NC 27708 USA

Abstract. The minimum description length (MDL) principle was developed in the context of computational complexity and coding theory. It states that the best model to account for some data minimizes the sum of the lengths, in bits, of the descriptions of the model and the data as encoded via the model. The MDL principle gives a criterion for parameter selection, by using the description length as a test statistic. Class I HLA genes play a major role in the immune response to HIV, and are known to be associated with rates of progression to AIDS. However, these genes are extremely polymorphic, making it difficult to associate alleles with disease outcome, given statistical issues of multiple testing. Application of the MDL principle to immunogenetic data from a longitudinal cohort study (Chicago MACS) enables classification of alleles associated with plasma HIV RNA abundance, an indicator of infection progression. Variation in progression is strongly associated with HLA-B. Allele associations with viral levels support and extend previous studies. In particular, individuals without *B58s* supertype alleles average viral RNA levels 3.6 times greater than individuals with them. Mechanisms for these associations include variation in epitope specificity and selection that favors rare alleles.

1 Introduction

Progression of HIV infection is characterized by three phases: acute, or early, chronic, and AIDS, the final phase of infection preceeding death [1]. The chronic phase is variable in duration, lasting ten years on average, but varying from two to twenty years. A good predictor of the duration of the chronic phase is the viral RNA level during chronic infection, with higher levels consistently associated with more rapid progression than lower levels [2]. A major challenge for treating HIV and developing effective vaccination strategies is to understand what causes variation in plasma viral RNA levels, and hence to infection progression.

The cell-mediated immune response identifies and eliminates infected cells from an individual. A central role in this response is played by the major histocompatibility complex (MHC), in humans, also known as human leukocyte antigens (HLA). Two classes of HLA genes code for codominately expressed cell-surface glycoproteins, and present processed peptide to circulating T cells, which discriminate between self (uninfected) and non-self (infected) cells [3,4].

Class I HLA molecules are expressed on all nucleated cells except germ cells. In infected cells, they bind and present antigenic peptide fragments to T-cell receptors on CD8+ T lymphocytes, which are usually cytotoxic and cause lysis of the infected cell. Class II HLA molecules are expressed on immunogenetically reactive cells, such as dendritic cells, B cells, macrophages, and activated T cells. They present antigen peptide fragments to T-cell receptors on CD4+ T-lymphocytes and the interaction results in release of cytokines that stimulate the immune response.

Human HLA loci are among the most diverse known [5,6]. This diversity provides a repertoire to recognize evolving antigens [6,7]. Previous studies of associations between HLA alleles and variation in progression of HIV-1 infection have established that within-host HLA diversity helps to inhibit viral infection, by associating degrees of heterozygosity with rates of HIV disease progression [8]. Thus, homozygous individuals, particularly at the HLA-B locus, suffer a greater rate of progression than do heterozygotes [8,9]. Identifying which alleles are associated with variation in rates of infection progression has been difficult, due in part to the compounding of error rates incurred when testing many alternative hypotheses, and published results do not always agree [10,11].

This study demonstrates the use of an information-based criterion for statistical inference. Its approach to multiple testing differs from that of standard analytic techniques, and provides the ability to resolve associations between variation in HIV RNA abundance and variation in HLA alleles.

As an application of computational complexity and optimal coding theory to statistical inference, the minimum description length (MDL) principle states that the best statistical model, or hypothesis, to account for some observed data is the model that minimizes the sum of the number of bits required to describe both the model and the data encoded via the model [12,13,14]. It is a model-selection criterion that balances the need for parsimony and fidelity, by penalizing equally for the information required to specify the model and the information required to encode the residual error.

2 Methods

The analyses detailed below apply the MDL principle to the problem of partitioning individuals into groups having similar HIV RNA levels, based on HLA alleles present in each case.

Chicago MACS HLA & HIV Data. The Chicago Multicenter AIDS Cohort Study (MACS) provided an opportunity to analyze a detailed, long-term, longitudinal set of clinical HIV/HLA data [10]. Each participant provided informed

consent in writing. Of 564 HIV-positive cases sampled in the Chicago MACS, 479 provided information about both the rate of disease progression and HLA genetic background. Progression was indicated by the quasi-stationary “set-point” viral RNA level during chronic infection. Immunogenetic background was obtained by genotyping HLA alleles from class I (HLA-A, -B, and -C) and class II (HLA-DRB1, -DQB1, and -DPB1) loci.

Viral RNA set-point levels were determined after acute infection and prior to any therapeutic intervention or the onset of AIDS, as defined by the presence of an opportunistic infection or CD4+ T-cell count below 200 per ml of plasma. Because the assay has a detection threshold of 300 copies of virus per ml [10], maximum-likelihood estimators were adjusted to avoid biased estimates of population parameters from a truncated, or censored, sample distribution [15]. Viral RNA levels were log-transformed for better approximations to normality.

High-resolution class I and II HLA genotyping [10] provided four-digit allele designations, though analyses were generally performed using two-digit allele designations because of the resulting reduction of allelic diversity and increased number of samples per allele. Because of the potential for results to be confounded by an effect associated with an individual’s ethnicity or revised sampling protocol, two separate analyses were performed, one using data from the entire cohort, and another using only data from Caucasian individuals. Sample numbers were too small to study other subgroups independently.

HLA supertypes group class I alleles by their peptide-binding anchor motifs [16]. Assignment of four-digit allele designations to functionally related groups of supertypes at HLA-A and -B loci facilitated further analysis. Where they could be determined, HLA-A and HLA-B supertypes were assigned from four-digit allele designations [10]. As with two-digit allele designations for each locus, HLA-A and -B supertypes were assessed for association with viral RNA levels. Cases having other alleles were withheld from classification and subsequent analysis.

A description length analysis determined whether HIV RNA levels were non-trivially associated with alleles at any HLA locus.

Description Lengths. The challenge of data classification is to find the best partition, such that observations within a group are well-described as independent draws from a single population, but differences in population distributions exist between groups. Whether the data are better represented as two groups, or more, than as one depends on the description lengths that result.

We use the family of Gaussian distributions to model viral RNA levels. While the MDL strategy can be applied using any probabilistic model, a log-normal distribution is a good choice for the observed plasma viral RNA values. First, the description length of the model and of the data given the model is calculated as described below, grouping all of the observations into one normal distribution, L_1 . Next, the data are broken into two partitions, L_2 , and the log-RNA values associated with HLA alleles are partitioned to minimize the description length given the constraint that two Gaussian distributions, each having their own mean and variance, are used to model the data.

For fixed $n \times n$ covariance matrix Σ , the description length is $L_\Sigma = \frac{1}{2} \log |\Sigma| + \frac{1}{2} Y' \Sigma^{-1} Y + C$, where Y is the n -component vector of observations and C is the quantity of information required to specify the partition. Logarithms are computed in base two, with fractional values rounded upwards, so that the resulting units are bits. The description length of interest results from integrating L over all covariance matrices with the appropriate structure. In practice, we use Laplace's approximation for the integral [12,17] which gives, asymptotically, $L = \frac{1}{2} \log |\hat{\Sigma}| + \frac{1}{2} Y' \hat{\Sigma}^{-1} Y + \frac{k}{2} \log n + C$, where k is the number of free parameters in the covariance model, and $\hat{\Sigma}$ is the specific covariance matrix of the appropriate structure that minimizes L_Σ .

The analog of a null hypothesis is the assumption that one group of alleles is sufficient to account for the variation in viral RNA. The description length for one group is: $L_1 = \frac{1}{2} (n + (n-1) \log s^2 + \log n \bar{x}^2 + 2 \log n)$, where n is the total number of observations, s^2 is the maximum-likelihood estimate of the population variance and $\bar{x}$ is the sample mean, computed as the Winsorized mean [15] because of truncation below the sensitivity limit of the RNA assay.

It follows that the description length for two groups can be computed as:

$$L_2 = \frac{1}{2} \sum_{i=1}^2 (n_i + (n_i - 1) \log s_i^2 + \log n_i \bar{x}_i^2 + 2 \log n_i) + C,$$

where C is an adjustment for performing multiple comparisons. Because additional information is required to specify the optimum partition, the description length is increased by a quantity related to the number of partitions evaluated, such that $C = N \log k$ bits, where N is the number of alleles observed at the partitioned locus. For $k = 2$, $C = N$.

Further partitions of alleles into more than two groups might yield a shorter description length, computed as a summation over terms in the equation for L_2 for each of the k distinct groups.

The shortest description length for any value of k indicates the best choice of model parameters, including the number of parameters, and hence, the optimum partition of N alleles into k groups. We denote this as L^* .

Algorithm. The minimum description length is found by iteratively computing the description length for each possible partition of alleles into groups and taking the minimum as optimal. Iteration consists first of determining the number of alleles, N , at a particular locus, and then incrementing through each of the $k^{(N-1)}$ possible partitions of alleles into k groups, computing the associated description length, and reporting the best results. Each iteration evaluates one possible mapping of alleles to groups. Searching through all possible partitions using the description length as an optimality criterion ensures selection of the best partition as a result of the search.

In this mapping, the ordering of groups is informative, because the ordering gives the relative dominance of alleles for diploid loci. An individual having an allele assigned to the first-order group is assigned to that group. Otherwise, the individual is assigned to the next appropriate group. Two individuals sharing one

allele might be placed in either the same group or different groups, depending on the mapping of alleles to groups in a particular iterate. For example, consider how one might group two individuals, one with alleles $A1$ and $A2$ at some locus, and another with alleles $A2$ and $A3$. Whether or not they are grouped together depends on the assignment of alleles to groups, and can be done several different ways. The algorithm enumerates each possible assignment of alleles to groups.

The extent of the search scales as k^N . In practice, the most diverse locus was HLA-B, with 30 alleles when analyzed using two-digit allele designations. For two groups, this gives $2^{30} \approx 10^8$ possible partitions. A parallel implementation requires no message passing, so computing time scales inversely with an increasing number of CPUs, or doubling available processors halves the time for iteration. The search space of 2^{30} partitions can be exhaustively evaluated in minutes on a cluster of CPUs. Unfortunately, exhaustively evaluating all three-way partitions is prohibitive, as $3^{30} \approx 2 \times 10^{14}$, over a million-fold increase in computational effort. Supertype classification reduced the diversity of possible partitions and enabled partitioning of the data into more than two groups.

3 Results

Class I & II HLA Alleles. The description length for the entire cohort as one group is $L_1 = 934$ bits; for the Caucasian subsample, it is $L_1 = 721$ bits. In general, $L_1 < L_2$ at most loci (Table 1), so the MDL criterion does not support partitioning alleles into groups that are predictive of high or low RNA levels, except at HLA-B, where $L_2 < L_1$. In the subsample, partitioning HLA-C or HLA-DQB1 alleles can also provide preferred two-way splits, though not as well as HLA-B. Further partitioning was intractable because of great allelic diversity, as previously mentioned. Partitions of HLA-B alleles provide the best groupings among all loci. Because $L_2^* < L_1$, two groups, partitioned by HLA-B alleles, provide a better description than one (Fig. 1a and 1b).

Table 1. Optimum two-way partitions at each locus, with per-locus allelic diversity (N), description lengths less the information cost to specify model parameters ($L_2 - C$), and total description lengths (L_2). (For the entire cohort, $n = 479$ and $L_1 = 934$ bits. For the Caucasian subsample, $n = 379$ and $L_1 = 721$ bits).

Class	Locus	Entire Cohort			Caucasian Subsample		
		N	$L_2 - C$	L_2	N	$L_2 - C$	L_2
I	HLA-A	19	916	935	18	703	721
	HLA-B	30	887	917*	26	681	707*
	HLA-C	14	921	935	13	706	719
II	DRB1	13	927	940	13	711	724
	DQB1	5	936	941	5	715	720
	DPB1	24	927	951	21	710	731

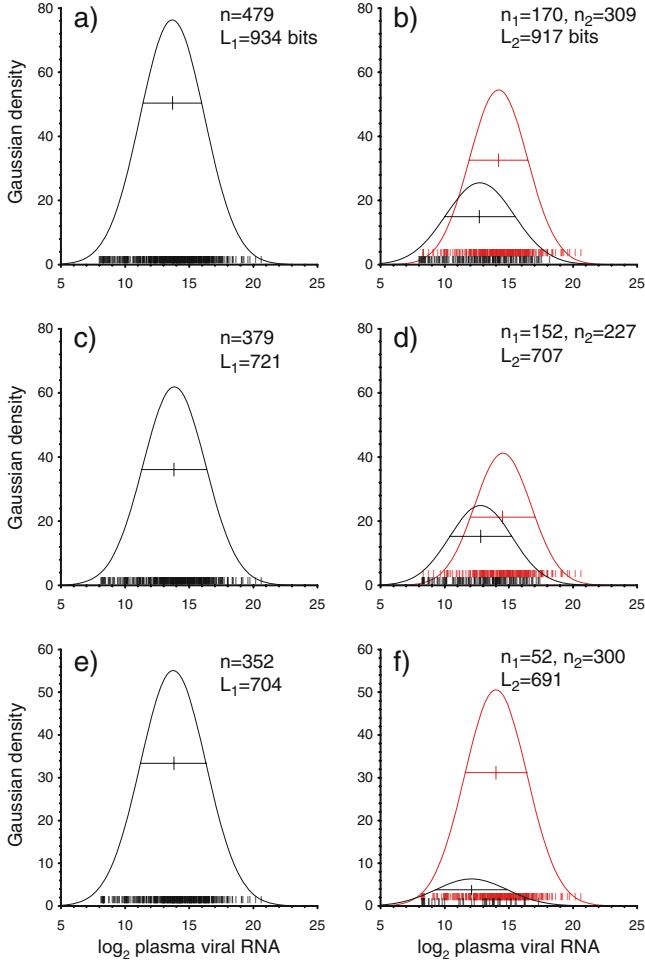


Fig. 1. Description-length comparisons of log-normal viral RNA distributions as one (L_1) or two (L_2) groups. Ordinate units are the expected number of observations between two tick marks over the abscissa, or one doubling of viral RNA. Impulses along the abscissa show individual observations, with jitter added to facilitate presentation of identical values. (a) Observations (n) from the Chicago MACS cohort lumped into one group, and (b) split into the best partition as two groups, with individuals having alleles B^*13 , B^*27 , B^*38 , B^*45 , B^*49 , B^*57 , B^*58 , or B^*81 assigned to the lower group (n_1), and remaining individuals assigned to the group with greater viral RNA (n_2). (c) Observations from the Caucasian subsample as one group, and (d) as the best split into two groups, where having alleles B^*13 , B^*27 , B^*40 , B^*45 , B^*48 , B^*49 , B^*57 , or B^*58 was the criterion for being assigned to the low viral-RNA group. Observations from individuals having two HLA-B supertype alleles, (e) in one group, and (f) partitioned into two groups, contingent on the presence of $B58s$.

Table 2. HLA-B alleles associated with low (○) or high (●) viral RNA levels. (Numbers in parentheses indicate allele abundances sampled).

Allele	Entire Cohort <i>n</i> = 479	Caucasian Subsample <i>n</i> = 379	Supertypes Only <i>n</i> = 352
<i>B7s</i>			(235)
<i>B*07</i>	(110) ●	(90) ●	●
<i>B*35</i>	(78) ●	(66) ●	●
<i>B*51</i>	(46) ●	(42) ●	●
<i>B*53</i>	(30) ●	(7) ●	●
<i>B*55</i>	(12) ●	(12) ●	●
<i>B*56</i>	(5) ●	(5) ●	●
<i>B*67</i>	(1) ○/●	—	●
<i>B27s</i>			(112)
<i>B*14</i>	(36) ●	(27) ●	●
<i>B*27</i>	(38) ○	(35) ○	●
<i>B*38</i>	(22) ○	(22) ●	●
<i>B*39</i>	(16) ●	(16) ●	●
<i>B*48</i>	(2) ●	(1) ○/●	●
<i>B44s</i>			(223)
<i>B*18</i>	(38) ●	(33) ●	●
<i>B*37</i>	(15) ●	(12) ●	●
<i>B*40</i>	(57) ●	(50) ○	●
<i>B*41</i>	(7) ●	(5) ●	●
<i>B*44</i>	(118) ●	(105) ●	●
<i>B*45</i>	(13) ○	(3) ○	●
<i>B*49</i>	(17) ○	(11) ○	●
<i>B*50</i>	(10) ●	(7) ●	●
<i>B58s</i>			(56)
<i>B*57</i>	(54) ○	(41) ○	○
<i>B*58</i>	(18) ○	(7) ○	○
<i>B62s</i>			(78)
<i>B*13</i>	(26) ○	(22) ○	●
<i>B*52</i>	(13) ●	(9) ●	●
Other			
<i>B*08</i>	(75) ●	(70) ●	—
<i>B*15</i>	(84) ●	(56) ●	—
<i>B*42</i>	(10) ●	—	—
<i>B*47</i>	(5) ●	(4) ○/●	—
<i>B*81</i>	(1) ○	—	—
<i>B*82</i>	(1) ○/●	—	—

Table 2 summarizes results of assigning HLA-B alleles to high or low viral-RNA categories. For the entire cohort, the following alleles were associated with low viral RNA levels: *B*13*, *B*27*, *B*38*, *B*45*, *B*49*, *B*57*, *B*58*, and *B*81*. The remaining alleles are associated with greater viral RNA than the first group. As already described, having any alleles associated with the first group is sufficient for an individual to be assigned to the group having lower viral RNA.

Four other groupings provide description lengths within one bit of the optimum. They do not dramatically rearrange the assignment of individuals to groups, but do provide insight as to which alleles are assigned to either group with less confidence. Among near-optimal partitions, alleles *B*67* and *B*82* were assigned to groups other than in the optimum partition.

Results were similar for the Caucasian subsample, though alleles *B*42*, *B*67*, *B*81*, and *B*82* were absent and two alleles, *B*38* and *B*40*, changed classification. Alleles not present in a sample are indicated by a dash in Table 2. Two nearly optimal partitions assigned alleles *B*47* and *B*48* to the alternative group. Figure 1 illustrates the distributions of viral RNA levels from this subsample, as one group (Fig. 1c) and as the best partition at HLA-B (Fig. 1d).

To summarize the most robust inferences from the analyses of two-digit allele designations, individuals having HLA-B alleles *B*13*, *B*27*, *B*45*, *B*49*, *B*57*, or *B*58* were associated with lower viral RNA levels than their counterparts lacking these alleles.

Comparison of groupings obtained via the MDL approach with more traditional means for statistical inference, a two-tailed, two-sample, Welch modified t-test, which does not assume equal variances, and its non-parametric variant, the Wilcoxon rank-sum test [18], was very favorable. In each case, the null hypothesis was that of no difference between the group mean log-transformed viral RNA levels, and the alternative hypothesis was that the means differ. Both tests agreed in rejecting the null hypothesis in favor of the alternative ($P < 10^{-10}$).

HLA Supertypes. Assigning the diploid, codominantly expressed HLA-A alleles to four HLA-A supertypes [16], *A1s*, *A2s*, *A3s*, and *A24s*, was possible for 399 individuals. The mapping of HLA-B alleles to five supertypes, *B7s*, *B27s*, *B44s*, *B58s*, and *B62s*, was made for 352 individuals. The resulting decrease in allelic diversity enabled analysis for $k > 2$.

Description lengths of the best k -way partitions of supertype alleles for HLA-A supertypes are: $L_1 = 793$, $L_2 = 782$, $L_3 = 789$, and $L_4 = 794$ bits. The best description length results from a two-way split, though a three-way split also yields a shorter description length than that obtained from one group. The best partition of HLA-A supertypes assigned individuals having *A1s* alleles to the low RNA group.

For HLA-B supertypes, $L_1 = 704$, $L_2 = 691$, $L_3 = 693$, and $L_4 = 697$ bits (Fig. 1e). The best model results when $k = 2$. Overall, individuals lacking *B58s* alleles averaged viral RNA levels 3.6-times greater than individuals having *B58s* supertype alleles (Fig. 1f). Thus, individuals with *B58s* alleles have significantly lower viral RNA levels than individuals without them.

Compositions of the optimum groupings of HLA-B supertypes for those individuals having two alleles that could be assigned to a supertype are summarized in Table 2. The *B*15* alleles were not analyzed as supertypes because their high-resolution genotype designations correspond to four different supertypes.

Overall, the most consistent associations with low viral RNA are among the *B58s*, and with high viral RNA, the *B7s*. Inconsistent assignments to a category occur for alleles *B*13*, *B*27*, *B*45*, and *B*49*, which are in the low viral-RNA

group when analyzed as such, but the high viral-RNA group when assigned to supertypes.

When compared with standard inferential techniques, the difference between group viral RNA levels was highly significant. This and agreement with alleles reported to be associated with variation in viral RNA levels in previously published studies indicate that using the description length as a test statistic can provide reliable inferences.

4 Discussion

MDL & Statistical Inference. The traditional statistical solution is to pose a question as follows: suppose that the simpler model (e.g., one homogeneous population) were actually true; call this the null hypothesis. How often would one, in similar experiments, get data that look as different from that expected under the null hypothesis as in the actual experiment?

This technique has limitations when the partition that represents the alternative hypothesis is not given in advance. There are then many other partitions possible and the appropriate distribution under the null hypothesis for this ensemble of tests is very difficult to estimate. Furthermore, for proper interpretation, the outcome relies upon the truth of the initial assumption: that the data are distributed as dictated by the null hypothesis.

An alternative is to choose that model that represents the data most efficiently. Here, efficiency is the amount of information, quantified as bits, required to transmit electronically both the model and the data as encoded by the model. This criterion may not seem intuitively clear on first exposure. However, it follows naturally from a profound relationship between probability and coding theory that was discovered, explored, and elaborated by Solomonoff, Kolmogorov, Chaitin, and Rissanen [19,20,21,22,23].

The idea is quite simple and elegant. It can be illustrated by analogy to the problem of designing an optimal code for the efficient transmission of natural-language messages. Consider the international Morse code. Recall that Morse code assigns letters of the Roman alphabet to codewords comprised of dots (“.”) and dashes (“—”). The codewords do not all have the same number of dots and/or dashes; it is a variable-length code.

Efficient, compact encodings result from the design of a codebook such that the shortest codewords are assigned to the most frequently encoded letters and long codewords are assigned to rare letters. Thus, *e* and *t* are encoded as “.” and “—”, respectively, while *q* and *j* are encoded as “— — . —” and “. — — —”. The theory of optimal coding provides an exact relationship between frequency and code length and thus, probability and description length.

The key departure of MDL from Morse-codelike schemes is that, while Morse code would generally be good for sending messages over an average of many texts, specific texts might be encoded even more efficiently, by encoding not only letters, but letter combinations, common words, or even phrases, perhaps as abbreviations or acronyms. However, if one is to recode for particular texts, one must first transmit the coding scheme. So perhaps one might use Morse code

to transmit the details of the new coding scheme and then transmit the text itself with the new scheme. Whether this might yield greater efficiency depends not only on how much compression is achieved in the new encoding, but also on how much overhead is incurred in having to transmit the coding scheme.

The analogy to scientific data analysis is clear. A statistical model is an encoding scheme that encapsulates the regularities in the data to yield a concise representation thereof. The best model effectively compresses regularities in the data, but is not so elaborate that its own description demands a great deal of information to be encoded. The MDL principle provides a model-selection criterion that balances the need for a model that is both appropriate and parsimonious, by penalizing with equal weights the information required to specify the model and the unexplained, or residual error.

Yet another contribution the MDL principle brings to statistical modelling is that the penalty for multiple comparisons is less restrictive than the penalty of compounded error rates incurred with canonical inferential approaches. In order to maintain a desired experiment-wide error rate, the standard adjustment is to make the per-comparison error rate considerably more stringent. With current technology, realistic sample sizes for such studies will generally be less than a thousand and stringent significance levels will be difficult to surpass. Unfortunately, fixing the false-positive error rate does not address the false-negative probability, which may leave researchers powerless to detect effects among many competing hypotheses with limited samples. Practical concerns of using MDL and related approaches are discussed in depth elsewhere [24].

Mechanisms. Of HLA supertype alleles, individuals with *B58s* have lower viral RNA levels than those who lack them, even among homozygotic individuals. Naturally, this leads one to consider mechanisms that underlie patterns found in the data. Elsewhere, we consider two hypotheses to explain the observed associations between HLA alleles and variation in viral RNA [10].

There may be allele-specific variation in antigen-binding specificity. Some alleles may have greater affinity than others for HIV-specific peptide fragments due to the peptide-binding anchor motifs they present. We were not able to identify any clear association between the frequency of anchor motifs among HIV-1 proteins and viral RNA levels in the Chicago MACS [10], though others have suggested that such a relationship might exist [25].

It may also be the case that frequency-dependent selection has favored rare alleles. Frequent alleles provide the evolving pathogen greater opportunity to explore mutant phenotypes that may escape detection by the host's immune response. By encountering rare alleles less frequently, the virus has not had the same opportunity to explore mutations that evade the host's defense response. This hypothesis is corroborated by a significant association between viral RNA and HLA allele frequency in the Chicago MACS sample [10].

Because their predictions differ, these hypotheses could be tested with data from another cohort, where a different viral subtype predominates (e.g., [26]). That is, if other alleles were associated with low viral RNA than those identified in this study, and an association between rare alleles and low viral RNA levels

were observed there, then the second hypothesis would be more viable than the first. Alternatively, if a clear association between antigen peptide-binding anchor motifs and variation in viral RNA levels were found, the first hypothesis would be more viable. Other mechanisms are also possible, and hypotheses by which to evaluate them merit consideration.

Acknowledgments

We thank Bob Funkhouser, Cristina Sollars, and Elizabeth Hayes for sharing their expertise, and researchers of the Santa Fe Institute for insight and inspiration. This research was financed by funds from the Elizabeth Glaser Pediatric AIDS Foundation, the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, National Institutes of Health, National Science Foundation award #0077503, and the US Department of Energy. We have no conflicting interests.

References

1. McMichael, A. J., Rowland-Jones, S. L.: Cellular immune responses to HIV. *Nature* **410** (2001) 980–987
2. Mellors, J. W., Rinaldo, C. R., Jr., Gupta, P., White, R. M., Todd, J. A., Kingsley, L. A.: Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science* **272** (1996) 1167–1170
3. Germain, R. N.: Antigen processing and presentation. In: Paul, W. E. (ed.): *Fundamental Immunology*. 4th edn. Lippincott-Raven, Philadelphia (1999) 287–340
4. Williams, A., Au Peh, C., Elliott, T.: The cell biology of MHC class I antigen presentation. *Tissue Antigens* **59** (2002) 3–17
5. Bodmer, W. F. Evolutionary significance of the HL-A system. *Nature* **237** (1972) 139–145
6. Little, A. M., Parham, P.: Polymorphism and evolution of HLA class I and II genes and molecules. *Rev. Immunogenet.* **1** (1999) 105–123
7. Hill, A. V. S.: The immunogenetics of human infectious diseases. *Ann. Rev. Immunol.* **16** (1998) 593–617
8. Roger, M. Influence of host genes on HIV-1 disease progression. *FASEB J.* **12** (1998) 625–632
9. Carrington, M., Nelson, G. W., Martin, M. P., Kissner, T., Vlahov, D., Goedert, J. J., Kaslow, R., Buchbinder, S., Hoots, K., O'Brien, S. J.: HLA and HIV-1: heterozygote advantage and *B*35-Cw*04* disadvantage. *Science* **283** (1999) 1748–1752
10. Trachtenberg, E. A., Korber, B. T., Sollars, C., Kepler, T. B., Hraber, P. T., Hayes, E., Funkhouser, R., Fugate, M., Theiler, J., Hsu, M., Kunstman, K., Wu, S., Phair, J., Erlich, H. A., Wolinsky, S.: Advantage of rare HLA supertype in HIV disease progression. *Nat. Med.* **9** (2003) 928–935
11. Trachtenberg, E. A., Erlich, H. A.: A review of the role of the human leukocyte antigen (HLA) system as a host immunogenetic factor influencing HIV transmission and progression to AIDS. In: Korber, B. T., Brander, C., Haynes, B. F., Koup, R., Kuiken, C., Moore, J. P., Walker, B. D., Watkins, D. (eds.): *HIV Molecular Immunology 2001*. Theoretical Biology and Biophysics Group, LANL, Los Alamos (2001) I 43–60

12. Rissanen, J.: Stochastic Complexity in Statistical Inquiry. World Scientific, Singapore (1989)
13. Li, M., Vitányi, P.: An Introduction to Kolmogorov Complexity and its Applications. Springer-Verlag, New York (1993)
14. Hansen, M. H., Yu, B.: Model selection and minimum description length principle. *J. Am. Stat. Assoc.* **96** (2001) 746–774
15. Johnson, N. L., Kotz, S., Balakrishnan, N.: Continuous Univariate Distributions, Vol. 1. 2nd edn. Wiley Interscience, New York (1994)
16. Sette, A., Sidney, J.: Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism. *Immunogenetics* **50** (1999) 201–212
17. Lindley, D. V.: Approximate Bayesian Methods. In: Bernardo, J. M., DeGroot, M. H., Lindley, D. V., Smith, A. F. M., (eds.): Bayesian Statistics. Valencia University Press, Valencia (1980) 223–237
18. Venables, W. N., Ripley, B. D.: Modern Applied Statistics with S-PLUS. 3rd edn. Springer, New York (1999)
19. Kolmogorov, A. N.: Three approaches to the quantitative definition of information. *Prob. Inform. Transmission* **1** (1965) 4–7
20. Chaitin, G. J. On the lengths of programs for computing binary sequences. *J. Assoc. Comput. Mach.* **13** (1966) 547–569
21. Chaitin, G. J.: Algorithmic Information Theory. Cambridge University Press, Cambridge UK (1987)
22. Rissanen, J.: Stochastic complexity and modeling. *Ann. Statist.* **14** (1986) 1080–1100
23. Rissanen, J.: Hypothesis selection and testing by the MDL principle. *Comput. J.* **42** (1999) 260–269
24. Burnham, K. P., and Anderson, D. R.: Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. 2nd edn. Springer, New York (2002)
25. Nelson, G. W., Kaslow, R., Mann, D. L.: Frequency of HLA allele-specific peptide motifs in HIV-1 proteins correlates with the allele's association with relative rates of disease progression after HIV-1 infection. *Proc. Natl. Acad. Sci. (USA)* **94** (1997) 9802–9807
26. Kiepiela, P., Leslie, A. J., Honeyborne, I., Ramduth, D., Thobakgale, C., Chetty, S., Rathnavalu, P., Moore, C., Pfafferoth, K. J., Hilton, L., Zimbwa, P., Moore, S., Allen, T., Brander, C., Addo, M. M., Altfeld, M., James, I., Mallal, S., Bunce, M., Barber, L. D., Szinger, J., Day, C., Klennerman, P., Mullins, J., Korber, B., Coovadia, H. M., Walker, B. D., and Goulder, P. J. R.: Dominant influence of HLA-B in mediating the potential co-evolution of HIV and HLA. *Nature* **432** (2004) 769–774

Visualization of Functional Aspects of microRNA Regulatory Networks Using the Gene Ontology[★]

Alkiviadis Symeonidis^{1,2}, Ioannis G. Tollis^{1,2}, and Martin Reczko^{1,★★}

¹ Computer Science Department, University of Crete

² Institute for Computer Science, Foundation for Research and Technology - Hellas,
P.O. Box 1385, 711 10 Heraklion, Crete, Greece

Abstract. The post-transcriptional regulation of genes by microRNAs (miRNAs) is a recently discovered mechanism of growing importance. To uncover functional relations between genes regulated by the same miRNA or groups of miRNAs we suggest the simultaneous visualization of the miRNA regulatory network and the Gene Ontology (GO) categories of the targeted genes. The miRNA regulatory network is shown using circular drawings and the GO is visualized using treemaps. The GO categories of the genes targeted by user-selected miRNAs are highlighted in the treemap showing the complete GO hierarchy or selected branches of it. With this visualization method patterns of reoccurring categories can easily identified supporting the discovery of the functional role of miRNAs. Executables for MS-Windows are available under www.ics.forth.gr/~reczko/isbmda06

1 Introduction

MicroRNAs (miRNAs) are very small non coding DNA regions regulating the post-transcriptional activity of other genes [1]. Experimental verification of these regulations is very expensive both in time and money so prediction algorithms are being developed [2,3,4,5,6,7,8,9,10]. In the following we have used the mi-Randa program ([2,3,4]), where every prediction is also assigned a score that reflects the strength of the regulation. The genes and miRNAs together with a set of predicted regulations form a (predicted) miRNA regulatory network. Information about genes is stored in the Gene Ontology (GO) where terms related to genes' functionalities are hierarchically organized [11].

Since research is focused on predictions little attention has been paid to the produced networks, their structure and properties. Here, we present algorithms for the visualization of miRNA regulatory networks and the GO. We also propose an alternative point of view for miRNA regulatory networks through the GO.

[★] This work was supported in part by INFOBIOMED code: IST-2002-507585 and the Greek General Secretariat for Research and Technology under Program "ARIS-TEIA", Code 1308/B1/3.3.1/317/12.04.2002.

^{★★} Corresponding author.

The remainder of this paper is organized as follows: In Section 2 we discuss how the biological information is transformed into graphs and review appropriate drawing techniques. In Section 3 we discuss an interactive drawing technique that allows the user to see the portion of the miRNA regulatory network that is of interest to him. In Section 4 we explain how we adopt the use of treemaps for the visualization of the GO. In Section 5 we discuss how these two drawing techniques communicate to support our suggestion for looking at miRNA regulatory networks through the GO. Finally, in Section 6 we discuss future extensions of our work.

2 Background

2.1 From Biology to Graph Theory

Here, we discuss how miRNA regulatory networks and the GO can be mathematically modeled as graphs. A graph $G(V, E)$ is a set V of vertices, which can represent objects and a set of edges $E \subseteq \{V * V\}$, which represents relations between the vertices.

The miRNA Regulatory Network. The miRNA network consists of miRNAs and protein coding genes. We also know that miRNAs regulate certain genes. We can construct a graph that represents this information: The set of vertices V is the set of miRNAs and genes. Then, for every regulation, we insert an edge that joins the two respective vertices. These edges have a weight, the score that is associated to the respective prediction. Since no direct relations exist between two miRNAs or two genes, this is a bipartite graph where the two partitions $V1$ and $V2$ are: $V1$: the set of miRNAs, $V2$: the set of genes.

Gene Ontology. The GO stores terms that describe information related to the biological role of genes. It has three main branches with different aspects of genes functionality. The first branch, "biological process" stores information about the various broad biological goals that are served by genes. The second branch, "molecular function" stores terms that describe various tasks that are performed by individual gene products. Typically, series of molecular functions form biological processes. Finally, the third branch, the "cellular component" stores various sub-cellular locations where gene products operate. The GO is organized in a hierarchical manner, so the terms are placed in layers that go from general to specific. The first layers are quite general and are used only to create main categories.

Since the GO consortium is an active project, new terms are inserted and others are updated or removed. In order to keep information about the removed terms three additional sets are used, one for each main branch and obsolete terms are placed in them. This information can be modeled with a graph, where each term is represented by a vertex. At this point the existence of synonyms must be mentioned. Synonyms that describe the same term are used in the GO. For all the synonyms of a term only one vertex is used in order to keep the

size of the graph small. Edges are used to declare the relations. They have to be directed to show which is the general and which the specific term. The GO hierarchical construction guarantees the absence of cycles in the graph. Thus, the GO hierarchy is a Directed Acyclic Graph (DAG). Since every term appears in exactly one branch and has no edge to the other two, the GO consists of three *DAGs* merged under a common root. The three obsolete sets are also placed under this root.

2.2 Graph Drawing Techniques

Graph drawing has applications in many different fields such as computer and social networks, E-R models or any other network. In [12] a large number of graph drawing algorithms are described. Here we review appropriate drawing techniques for the visualization of miRNA regulatory networks and the GO.

Visualization of miRNA Networks. Bipartite graphs are typically drawn based on the idea of the placement of each partition in a column. The main problem is to compute the ordering of vertices in the partitions that gives the lowest number of crossings. This is an NP-complete problem [13] even if the ordering of one partition is fixed [14] so various heuristics have been proposed to determine an ordering which gives a small number of edge crossings [15,16,17,18]. The size of a miRNA regulatory network (some thousands of genes and some hundreds of miRNAs) makes the adoption of such a technique inefficient. Efforts are focused on the prediction of regulations and not the network itself, so no special techniques seem to be available for miRNA regulatory networks. Here we introduce an interactive technique using circular drawings that allows the user to see only the portion of the network that interests him and not the full network.

Visualization of the GO. DAGs (like the GO) are typically drawn using algorithms that try to display their hierarchical structure. The polyline drawing convention is the basis for such algorithms, [19,20] and the main idea is to create hierarchical layers, [21,22,23]. Another approach to the visualization of hierarchies are treemaps [24,25,26], where vertices that are low in the hierarchy are placed over those that are in the first layers in the hierarchy. This method works for trees, but can handle DAGs as well, since they can be transformed into trees. Tools like GOfish [27], Amigo[28], CGAP[29] and dagEdit [30], visualize the GO with expanding lists or trees, which can manipulate small hierarchies, but face difficulties for large hierarchies. Treemaps which are a space-efficient manner for displaying hierarchies and also support fast and easy navigation have more recently been used for the visualization of the GO [31,32]. In both of them additional information e.g. about the number of genes related to the GO terms is used and weights are assigned to the terms based on this, offering a visualization of the part of GO that is related to the available information. We also use treemaps but choose to display the GO without additional information. This way we have a general-purpose browser for the whole GO. One has still access to the relations between genes and GO terms through interactive methods.

3 Drawing the miRNA Network

For this network, the main requirement is that the bipartite structure must be clear. Also since a prediction-score is assigned to each regulation we also want to see this information. Small bipartite graphs are typically drawn using two columns, but the size of the network and the difference in the cardinalities of the two partitions (miRNAs are expected to be 1-10% of the number of genes) inhibit the adoption of such a technique. Furthermore, due to the large degree of many miRNA-vertices, there are many edge crossings and the result is very cluttered. In addition, most of our available drawing area is occupied by the GO. as we discuss later, thus the space that is available for the miRNA regulatory network is limited. While ideally one would like to have a visualization of the whole regulatory network where the bipartite structure and other properties are clear, due to the aforementioned problems this can not be achieved. But we can circumvent these difficulties with an interactive visualization technique, where the user specifies what he wants to see.

In order to draw the miRNA regulatory network interactively we use two lists, and allow the user to select genes and miRNAs. If some miRNA is selected it is drawn using a small circle and all genes that are regulated by it are placed on the periphery of a circle whose center is the miRNA. This way, we avoid drawing lines for many edges and obtain a much clearer result. An edge has to be drawn only if some gene has been previously drawn. This happens when the user selects to display some miRNA which regulates a gene that is also regulated by some other previously selected miRNA (and is drawn on its periphery). We also suggest that dots for miRNAs are a bit larger than genes, to further clarify the bipartite structure. The miRNAs are placed in two columns, in the order they are selected. The first miRNA in the left column, the second in the right, the third in the left under the first and so on. The names of the miRNAs are printed in the interior, while genes names are shown when the mouse is over them. An example is shown in figure 1(a).

The user can also select some gene from the respective list. If the gene has already been drawn on the periphery of a cycle, all miRNAs that regulate it and have not been yet drawn are placed in the lists and the respective lines are drawn to indicate the relationship. If the gene has not been already drawn, all regulating miRNAs are drawn and the gene is placed on the periphery of the first of them. The newly added miRNAs are not expanded (no genes are placed on their periphery and no lines are added) in order to keep the image as clear as possible. This is shown in figure 1. the network of fig. 1 by selecting gene ENSG00000048740 leads to the drawing in fig. 1(b).

For genes with many related miRNAs a large number of new cycles can be produced. Thus, an option to remove some miRNA from the network is available on right click. This drawing method offers many advantages. First of all, the user sees only the portion of the network that is of interest to him. Secondly, since this image is being created in a step-by-step interactive process, it is easy to remember where each term is and thus, maintain a mental map of the network. The bipartite structure of the network is also obvious: miRNAs appear in the

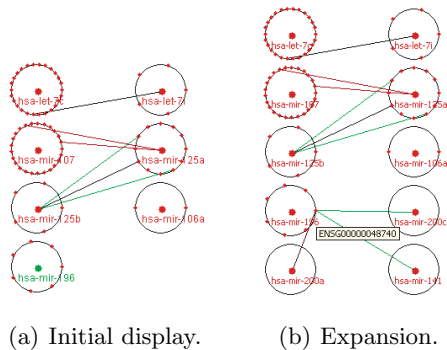


Fig. 1. The sub-network, after selecting hsa-let-7c, hsa-let-7i, hsa-mir-107, hsa-mir-125a, hsa-mir-125b, hsa-mir-106a, hsa-mir-196 and the expansion after selecting gene ENSG00000048740

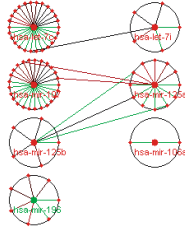
middle of the cycles and are a bit larger, while genes appear on the periphery. Due to the convention that genes on the periphery are regulated by the miRNA in the center, many edges are implied and do not have to be drawn. With some other drawing technique which displays all edges for the part of the network in Figure 1, 82 more edges would have to be drawn. Finally, this method is very fast, since every selection simply adds a few elements to the already constructed image. An equivalent gene-oriented visualization is available as an option with the roles of miRNAs and genes interchanged.

Weighted edges

Every prediction has an associated score, which represents its "strength", so it is important to show this information. To this end, the median value of the scores for the edges is computed. All edges that have score below the median are colored in red and the brightness reflects the value. The lower the score, the brighter the color. For edges that have score equal to the median black is used, while edges with higher score are colored using green where again, the brightness reflects the value. Since viewing edges is important, we have added some features. By clicking on a vertex, all the associated edges which are initially visible are hidden. So even if a complex image has been created, the user has the ability to unclutter it. By re-clicking, the edges are revealed. Similarly, for the edges that are implied by the placement of vertices on the periphery of a cycle, an option to show them is also available(Fig 2).

4 Drawing the Gene Ontology DAG

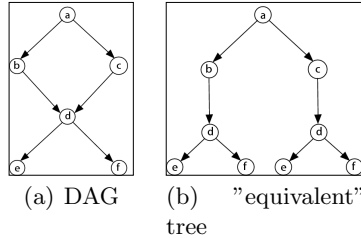
In order to display the GO DAG efficiently three aspects must be considered: a) display the whole graph, b) maintain the hierarchy, c) traverse through the graph easily. The GO DAG currently contains about 20000 terms. Any traditional graph drawing technique would fail to draw such a large graph efficiently within

**Fig. 2.** Figure 1 including in-cycle edges

the limited area of a computer monitor. The hierarchical structure though, allows for the decomposition to a hierarchical tree. This process creates a tree with even more vertices but enables us to use treemaps which is among the most space-efficient techniques to draw large trees.

4.1 Converting a DAG into a Tree

The method that decomposes a DAG to a tree is based on the placement of multiple copies for vertices with many in-coming neighbors, since these are the vertices that destroy the tree structure. Fig. 3 shows an example.

**Fig. 3.** DAG to tree conversion

First we create the root of the tree which is the (unique) source of the DAG. Then for every vertex v in the DAG, a recursive method *insertNode* is called. This method inserts the necessary copies of a vertex in the tree after ensuring that all the ancestors have been inserted. If v has been processed *insertNode* simply returns, otherwise it adds all the incoming neighbors of v in the tree one by one with recursive calls. Once a recursive call for an incoming neighbor u terminates, the necessary copies of u have been inserted in the tree. Now we can insert copies of v to be children of u as many times as the copies of u are. Repeating this for every parent of v in the DAG adds the necessary copies in the tree. For fast access to the many copies of a vertex we use a hashtable where the key is a vertex and the value is a list with all of its copies.

```

DAG2TREE( Directed acyclic graph  $G$ )
1  Hashtable vertexCopies
2  treeRoot =  $G$ .vertexWithInDegree0
3  vertexCopies.put(treeRoot, treeRoot)
4  for each vertex  $v$  in  $G$ 
5    do INSERTNODE( $v$ )

INSERTNODE( $v$ )
1  if not vertexCopies.get(v).isEmpty() //if inserted
2    then
3      return
4  for each incoming neighbor  $w$  of  $v$ 
5    do
6      INSERTNODE( $w$ )
7      for each  $w\_copy$  in vertexCopies.get(w)
8        do
9          make a new copy  $v\_copy$  of  $v$ 
10         insert  $v\_copy$  to be a child of  $w\_copy$ 
11         append  $v\_copy$  in vertexCopies.get(v)

```

The hashtable performs the search operation in constant time ($O(1)$). Hence, the cost for calling *insertNode* for one vertex in the DAG, is linear to the number of ancestors of the respective copies in the resulting tree, or equivalently to the number of incoming edges for all copies of the vertex, plus the cost of potential recursive calls. Since *insertNode* is called for every vertex creating different edges, the overall complexity is proportional to the sum of the incoming degrees of all vertices, or the number of edges in the tree. So the overall complexity is $O(E_{tree})$ where E_{tree} is the number of edges in the tree. This process leads to a tree with ~ 100.000 vertices. The careful management of synonyms which was discussed in Section 2.1 pays back here: If all synonyms were represented with different vertices, the resulting tree would have ~ 150.000 nodes.

4.2 Treemaps

Treemaps were introduced by Johnson and Shneiderman in 1991 [24,25]. In treemaps, every node of the tree is represented by some rectangular area. The main idea is to place each vertex on top of its parent's rectangle starting from the root and proceeding in a depth-first-search way. This way, layers of hierarchy are created. On the upper, visible layer only the leaves appear.

```

TREEMAP( vertex  $v$ , rectangular area area)
1   $v$ .area=area;
2  divide area to the children of  $v$ 
3  for each child  $c$  of  $v$ 
4    do TREEMAP( $c$ ,  $c$ 's computed area )

CALLTREEMAP(treeRoot,drawingArea)
1  TREEMAP(treeRoot,drawingArea)

```

If instead of the root some other vertex is used for calling the algorithm, the subtree rooted under this vertex is shown. This means that the algorithm itself can be used for zooming-in and out. For the space division in step 2 various approaches have been proposed. The original idea, "*slice and dice*" [24,25] was to divide the area to slices with size proportional to some metric defined for the vertices. This metric typically is the weight in case of weighted trees and the number of leaves in the subtree for non-weighted trees. For better visualization results the orientation of the slices alters from horizontal to vertical between subsequent layers. A drawback is that *slice and dice* tends to create long and thin rectangles. In an attempt to avoid this, Bruls et. al. proposed *squarified treemaps* [26]. They divide the area using a heuristic which aims to create rectangles as close to squares as possible. This additional constraint usually alters the placement of vertices while zooming the tree in and out. In [24] the authors also propose nesting to show internal vertices and the hierarchical structure. An additional advantage of nesting is that the unoccupied space can be used for placing labels. The thin slices of *slice & dice* create an obstacle in this concept. We support nesting in the following manner: Whenever some area is assigned to a vertex, a small border is placed on its boundary and only its interior is given as drawing area for the children. The user has the option to set the size of the border. We have chosen to use the *squarified* approach since the results of *slice & dice* for large trees are relatively poor. We also manage to keep track of an interesting vertex after zooming in and out by highlighting the path from the root of the tree to it (Fig 4). Larger figures are available at www.ics.forth.gr/~reczko/isbmda06

We also provide an option to show only a number of hierarchical layers in order to hide small, hardly visible rectangles that are produced for vertices in the deeper layers. Zoom-in allows the user to access deeper layers. Zoom-in is straightforward and is performed when some term is left-clicked. For zooming out we provide three options, which become available on right-click. The user can move one or two layers higher. Alternatively, he can move up to the root of the tree and display the whole GO.

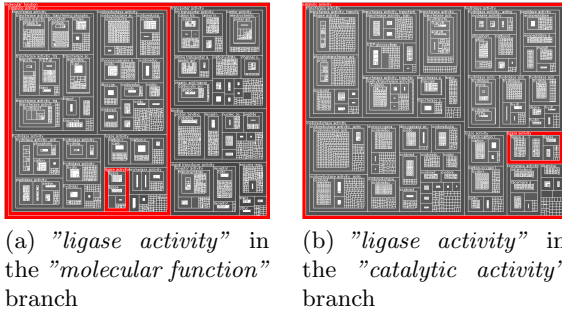


Fig. 4. Treemaps of subtrees of the GO. The term "*ligase activity*" is highlighted with a red border.

4.3 Integrated Visualization of the miRNA Regulatory Network and the Gene Ontology

We have discussed how to visualize an interesting part of the miRNA regulatory network and how to explore the GO with treemaps. Here, we discuss how one can combine them using information about GO terms that are related to genes. miRNAs are also related to GO terms through their regulating genes.

Gene Ontology Based Analysis of the miRNA Regulatory Network.

Every gene in the network is related to at least one term in the GO. So the first requirement is to see the genes that are related to some specific term. In order to ask this question, the user has to select some term and choose the "Show Genes" option. All genes that are related to the selected GO term and belong to the visible portion of the network on the left are highlighted (colored in green). An instance is illustrated in Fig. 5.

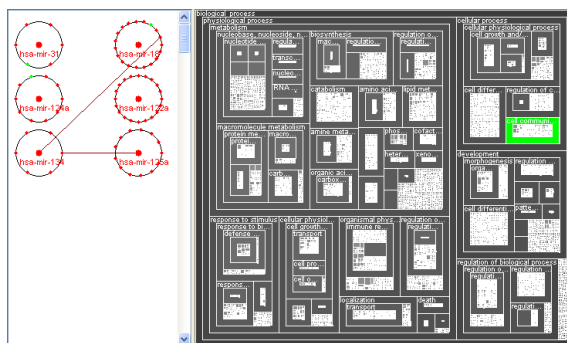


Fig. 5. Highlighted genes on the cycles on the left(87470, 68383, 101384) are related to "cell communication"

With the drawing method used for the miRNA regulatory network, only a portion of the regulatory network is drawn. This way the user sees only the part of interest and not the whole complex network, but some of the genes that are related to some GO term may not be visible. These additional genes can be shown with the option "Expand & show Genes".

As an example for the miRNAs known to be related to the category "neurogenesis" the subset targeting genes in the category "axonogenesis" is highlighted in figure 6.

Search for GO Related Properties of the miRNA Regulatory Network. Complementary, the user can search for GO properties based on genes or miRNAs. If some gene is selected, the GO terms that are related to it are colored in bright green. Because of the option to show only some layers some terms may not be visible. In this case, the closest visible ancestor is highlighted

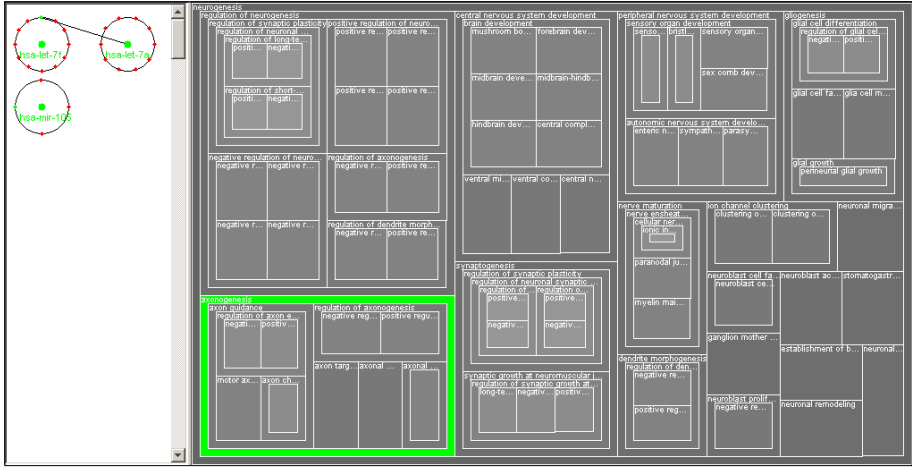


Fig. 6. Highlight visible genes and miRNAs related to "axonogenesis"

using dark green (Fig. 7(a)). If the branch is of interest, one can zoom in and identify the related term, which is in bright green (Fig. 7(b)).

An interesting relation between a miRNA and the GO is implied by the GO terms that are related to the genes it regulates.

Selecting a miRNA, the set of GO terms that are related to at least one of the genes it regulates are colored. The coloring is the same as in the previous case. For *hsa-mir-17-5p*, the miRNA with the maximum number of predicted target genes, the result appears in Figure 8.

For miRNAs with many target genes, a large portion of the GO is highlighted. The user can see the terms that are related to the genes, but does not know the number of genes that are related to these terms. So, we support an option to iterate through the genes that are regulated by some miRNA and display for every gene the related GO terms with animated highlighting. Now, the user can see how often some term is highlighted and is able to observe emerging patterns of reoccurring terms. With this feature it became obvious that 17 out of the 21 genes that are regulated by *hsa-mir-98* are related to the category "binding" or a even more specific term, emphasizing the likelihood of those predictions.

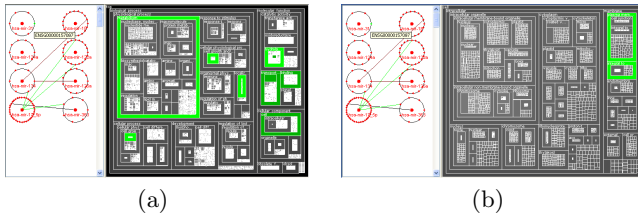


Fig. 7. Left: All terms related to gene *ENSG00000157087* (PLASMA MEMBRANE CALCIUM-TRANSPORTING ATPASE 2). Right: After zooming-in "cell".

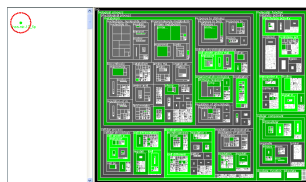


Fig. 8. All GO terms that are related to *hsa-mir-17-5p*

5 Discussion

We have introduced a novel combination of visualization methods to explore the relation of microRNAs and their target genes. The interactive construction of miRNAs and functionally related targeted genes will be of great use in discovering the functional role of many miRNAs. The treemap based browser might establish a standard 'mental map' of the GO and can independently be used to visualize any GO related information. Currently we support data in the format presented in [2] but the support for other prediction methods is planned.

References

1. R.C. Lee et. al. *C. elegans* heterochronic gene *lin-4* encodes small rnas with anti-sense complementarity to *lin-14*. *Cell*, 75:843–854, 1993.
2. B. John et. al. Human microRNA targets. *PLoS Biology*, 2:e363, 2004.
3. A. J. Enright et. al. microRNA targets in drosophila. *Genome Biol.*, 5:R1, 2003.
4. A. Stark et. al. Identification of drosophila microRNA targets. *PLoS Biology*, 1:E60, 2003.
5. B.P. Lewis et. al. Prediction of mammalian microRNA targets. *Cell*, 115:787–798, 2003.
6. S. Pfeffer et. al. Identification of virus encoded microRNAs. *Science*, 304:734–736, 2004.
7. M. Kiriakidou et. al. A combined computational-experimental approach predicts human microRNA targets. *Genes dev.*, 18:1165–1178, 2004.
8. D. Grün et. al. microRNA target predictions across seven drosophial species and comparison to mammalian targets. *PLoS Comp. Biol.*, 1:e13, 2005.
9. A. Krek et. al. Combinatorial microRNA target predictions. *Nature Genetics*, 37:495–500, 2005.
10. M. Rehmsmeier et. al. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10:1507–1517, 2004.
11. <http://www.geneontology.org>.
12. G. Di Battista et. al. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, 1999.
13. M.R. Garey and D.S. Johnson. Crossing number is np-complete. *SIAM J. Algebraic Discrete Methods*, 4:312–316, 1983.
14. P. Eades et. al. On an edge crossing problem. In *Proc. of the Ninth Australian Computer Science Conference*, pages 327–334. Australian National University, 1986.

15. P. Eades and D. Kelly. Heuristics for drawing 2-layered networks. *Ars Combin.*, 21-A:89–98, 1986.
16. P. Eades and N. Wormald. Edge crossings in drawings of bipartite graphs. *Algorithmica*, 11(4):379 – 403, 1994.
17. E. Mäkinen. Experiments on drawing 2-level hierarchical graphs. *Inter. Journ. Comput. Math.*, 36:175–181, 1990.
18. M. May and K. Szkatula. On the bipartite crossing number. *Control Cybernet*, 17:85–98, 1988.
19. G. Di Battista and R. Tamassia. Algorithms for plane representations of acyclic graphs. *Theoret. Computation Scien.*, 61:175–198, 1988.
20. G. Di Battista et. al. Constrained visibility representations of graphs. *Informat. Process. Letters*, 41:1–7, 1992.
21. K. Sugiyama et. al. Methods for visual understanding of hierarchical systems. *IEEE, Trans. Syst. Man. Cybern.*, 11(2):109–125, 1981.
22. D. J. Gschwind and T. P. Murtagh. *A Recursive Algorithm for Drawing Hierarchical Directed Graphs*. Department of Computer Science, Williams College, 1989.
23. M.J. Carpano. Automatic display of hierarchized graphs for computer aided decision analysis. *IEEE, Transact. Syst. Man. Cybern.*, 10(11):705–715, 1980.
24. B. Johnson and B. Shneiderman. Treemaps: a space-filling approach to the visualization of hierarchical information structures. In *Proc. of the 2nd Intern. IEEE Visualization Conference*, pages 284–291, Oct. 1991.
25. B. Shneiderman. Tree visualization with tree-maps: a 2d space-filling approach. *ACM Transactions on Graphics*, 11(1):73–78, Sept. 1992.
26. D. M. Bruls et. al. Squarified treemaps. In *Proceedings of the joint Eurographics and IEEE TVCG Symposium on Visualization*, pages 33–42, October 2000.
27. llama.med.harvard.edu/~berriz/GoFishWelcome.html.
28. www.godatabase.org/cgi-bin/amigo/go.cgi.
29. cgap.nci.nih.gov/Genes/GOBrowser.
30. sourceforge.net/project/showfiles.php?group_id=36855.
31. E. H. Baehrecke et. al. Visualization and analysis of microarray and gene ontology data with treemaps. *BMC Bioinformatics*, 5, 2004.
32. P McConnell et. al. Applications of tree-maps to hierarchical biological data. *Bioinformatics*, 18(9):1278–1279, 2002.

A Novel Method for Classifying Subfamilies and Sub-subfamilies of G-Protein Coupled Receptors

Majid Beigi and Andreas Zell

University of Tübingen
Center for Bioinformatics Tübingen (ZBIT)
Sand 1, D-72076 Tübingen, Germany
{majid.beigi, andreas.zell}@uni-tuebingen.de

Abstract. G-protein coupled receptors (GPCRs) are a large superfamily of integral membrane proteins that transduce signals across the cell membrane. Because of that important property and other physiological roles undertaken by the GPCR family, they have been an important target of therapeutic drugs. The function of many GPCRs is not known and accurate classification of GPCRs can help us to predict their function. In this study we suggest a kernel based method to classify them at the subfamily and sub-subfamily level. To enhance the accuracy and sensitivity of classifiers at the sub-subfamily level that we were facing with a low number of sequences (imbalanced data), we used our new synthetic protein sequence oversampling (SPSO) algorithm and could gain an overall accuracy and Matthew's correlation coefficient (MCC) of 98.4 % and 0.98 for class A, nearly 100% and 1 for class B and 96.95% and 0.91 for class C, respectively, at the subfamily level and overall accuracy and MCC of 97.93% and 0.95 at the sub-subfamily level. The results shows that Our oversampling technique can be used for other applications of protein classification with the problem of imbalanced data.

1 Introduction

G-protein coupled receptors (GPCRs) are a large superfamily of integral membrane proteins that transfer signals across the cell membrane. Through their extracellular and transmembrane domains they respond to a variety of ligands, including neurotransmitters, hormones and odorants. They are characterized by seven hydrophobic regions that pass through the cell membrane (transmembrane regions) [1], as shown in Fig. 1. Each GPCR has an amino terminal (NH₂ or N-terminal) region outside of the cell, followed by intracellular and extracellular loops, which connect the seven transmembrane regions, and also an intracellular carboxyl terminal (COOH- or C-terminal) region. GPCRs are involved in signal transmission from the outside to the interior of the cell through interaction with heterotrimeric G-proteins, or proteins that bind to guanine (G) nucleotides. The receptor is activated when a ligand that carries an environmental signal binds to a part of its cell surface component. A wide range of molecules is used as

the ligands including peptide hormones, neurotransmitters, pancrine mediators, etc., and they can be in many forms: e.g., ions, amino acids, lipid messengers and protease [2].

The function of many GPCRs are unknown and understanding the signaling pathways and their ligands in laboratory is expensive and time-consuming. But the sequence of thousands of GPCRs are known [3]. Hence, if we can develop an accurate predictor of the class (and so function) of GPCRs from their sequence it can be of great usefulness for biological and pharmacological research. According to the binding of GPCRs to different ligand types they are classified into different families. Based on GPCRDB (G protein coupled receptor data base) [3] all GPCRs have been divided into a hierarchy of 'class', 'subfamily', 'sub-sub-family' and 'type' (Fig. 2).

Because of the divergent nature of GPCRs it is difficult to predict the classification of GPCRs by means of sequence alignment approaches. The standard bioinformatics approach for function prediction of proteins is to use sequence comparison tools such as PSI-BLAST [4] that can identify homologous proteins based on the assumption of low evolutionary divergence, which is not true for GPCRs families. Here, we are facing a more difficult problem of remote homology detection, where classifiers must detect a remote relation between unknown sequence and training data.

There have been several recent developments to the classification problem specific to the GPCR superfamilies. Moriyama and Kim [5] developed a classification method based on discriminant function analysis using composition and physicochemical properties of amino acids. Elrod and Chou [6] suggested a covariant discriminant algorithm to predict GPCR's type from amino acid composition. Qian et al. [7] suggested a phylogenetic tree based profile hidden Markov model (T-HMM) for GPCR classification. Karchin et al. [8] developed a system based on support vector machines built on profile HMMs. They generated fisher score vectors [9] as features for SVM classifier from those profile HMMs. They showed that classifiers like SVMs that are trained on both positive and negative examples can increase the accuracy of GPCRs classification compared with only HMMs as generative method.

To increase the accuracy of remote homology detection by discriminative methods, researchers also focused on finding new kernels, which measure the similarity between sequences, as main part of SVM based classifiers. So after choosing an appropriate feature space, and representing each sequence as a vector in that space, one takes the inner product between these vector-space representations. Spectrum kernel [10], Mismatch kernel [11] and Local alignment kernel [12] are examples of those kernels and it has been shown that they have outperformed previous generative methods for remote homology detection.

In our study we want to classify GPCRs at the subfamily and sub-subfamily level. In this case, a problem in classification of GPCRs is the number of proteins at the sub-subfamily level. At this level in some sub-subfamilies we have only a very low number of protein sequences as positive data (minor class) compared with others (major class). In general, with imbalanced data, the SVM classifier

tends to perform best for classifying the majority class but fails to classify the minority class correctly. Because of that problem some researchers have not considered those GPCRs families, or if they have included them in their classifier they did not get as good results for them as for other families with enough data [13]. We used a new oversampling technique for protein sequences, explained in [24] to overcome that problem. Based on that method at first we make a HMM profile of those sequences and then try to increase the number of sequences in that family synthetically considering the phylogenetic tree of that family and also the distribution of other families near to that family. For classification, we use the local alignment kernel (LA kernel) that has been shown to have better performance compared with other previously suggested kernels for remote homology detection when applied to the standard SCOP test set [15]. It represents a modification of the Smith-Waterman score to incorporate sub-optimal alignments by computing the sum (instead of the maximum) over all possible alignments. Using that kernel along with our oversampling technique we could get better accuracy and Matthew's correlation coefficient for the classification of GPCRs at the subfamily and sub-subfamily level than other previously published method.

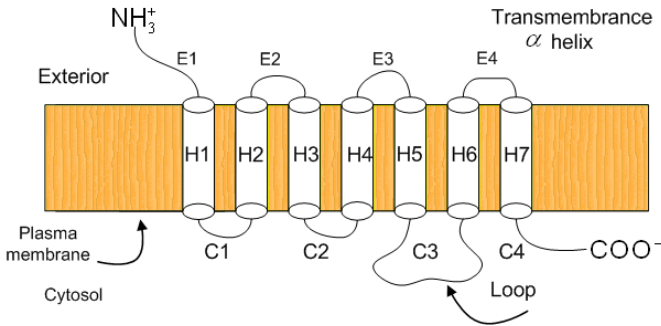


Fig. 1. Schematic representation of GPCR shown as seven transmembrane helices depicted as cylinders along with cytoplasmic and extracellular hydrophilic loops

2 Materials

The dataset of this study was collected from GPDRDB [3] and we used the latest dataset of GPCRDB (June 2005 release, <http://www.gpcr.org/7tm/>). The six main families are: Class A (Rhodopsin like), Class B (Secretin like), Class C (Metabotropic glutamate/pheromone), Class D (Fungal pheromone), Class E (cAMP receptors) and Frizzled/Smoothed family. The sequences of proteins in GPCRDB were taken from SWISS-PROT and TrEMBL data banks [14]. All six families of GPCRs (5300 protein sequences) are classified in 43 subfamilies and 99 sub-subfamilies. The three largest classes are the rhodopsin-like receptors, the secretion-like receptors and the metabotropic glutamate receptors (class A, B, and C). The rhodopsin-like family is the largest and most studied with approximately 90 percent of all receptors (4737 out of 5300).

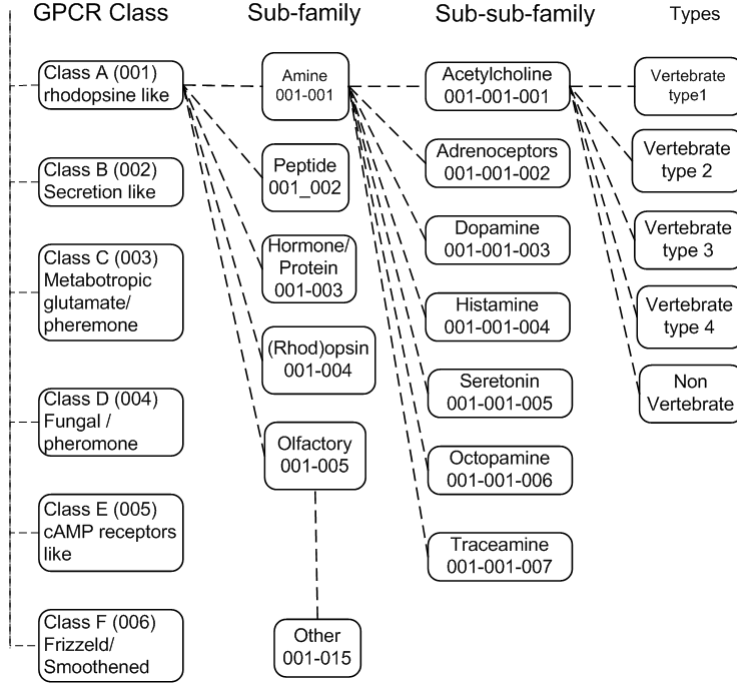


Fig. 2. GPCR family tree according to GPCRDB nomenclature

3 Algorithms

3.1 Kernel Function

In discriminative methods, a classifier learns a rule to classify unlabelled sequences into a class of proteins by using both sequences belonging to this class (positive examples) and sequences known as not belonging to that class (negative examples). Given a set of positive training sequences χ_+ and a set of negative training sequence χ_- an SVM learns a classification function $f(x)$ of the form:

$$f(x) = \sum_{i; x_i \in \chi_+} \lambda_i K(x, x_i) - \sum_{i; x_i \in \chi_-} \lambda_i K(x, x_i) \quad (1)$$

where non-negative λ_i weights are computed during training by maximizing a quadratic objective function and $K(.,.)$ is the kernel function. Given this function, a new sequence X is predicted to belong to positive dataset if the value of $f(x)$ is positive, otherwise it belongs to the negative dataset.

On the other hand, variable length protein sequences must be converted to fixed length vectors to be accepted as input to a SVM classifier. These vectors should exploit prior knowledge of proteins belonging to one family and enable us to have maximum discrimination for unrelated proteins. So the kernel function is of great importance for SVM classifiers in learning the dataset and also in

exploiting prior knowledge of proteins and mapping data from input space to feature space. The Smith Waterman (SW) alignment score between two protein sequences tries to incorporate biological knowledge about protein evolution by aligning similar parts of two sequences but it lacks the positive definiteness as a valid kernel [15]. The local alignment kernel mimics the behavior of the Smith Waterman (SW) alignment score and tries to incorporate the biological knowledge about protein evolution into a string kernel function. But unlike the SW alignment, it has been proven that it is a valid string kernel. We used this kernel for our classification task, so we give a brief introduction to that algorithm: If K_1 and K_2 are two string kernels then the convolution kernel $K_1 \star K_2$ is defined for any two strings x and y by:

$$K_1 \star K_2(x, y) = \sum_{x_1 x_2 = x, y_1 y_2 = y} K_1(x_1, y_1) K_2(x_2, y_2) \quad (2)$$

Based on work of Haussler [16] if K_1 and K_2 are valid string kernels, then $K_1 \star K_2$ is also a valid kernel. Vert et al. [12] used that point and defined a kernel to detect local alignments between strings by convolving simpler kernels. The local alignment kernel (LA) consists of three convolved string kernels. The first kernel models the null contribution of a substring before and after a local alignment in the score:

$$\forall(x, y) \in \chi^2, \quad K_0(x, y) = 1 \quad (3)$$

The second string kernel is for alignment between two residues:

$$K_\alpha^{(\beta)}(x, y) = \begin{cases} 0 & \text{if } |x| \neq 1 \text{ or } |y| \neq 1 \\ \exp[\beta s(x, y)] & \text{otherwise,} \end{cases} \quad (4)$$

where $\beta \geq 0$ controls the influence of suboptimal alignments in the kernel value and $s(x, y)$ is a symmetric similarity score or substitution matrix, e.g. BLO-SUM62.

The third string kernel models affine penalty gaps:

$$K_g^{(\beta)}(x, y) = \exp \{ \beta [g(|x|) + g(|y|)] \} \quad (5)$$

$g(n)$ is the cost of a gap of length n given by:

$$\begin{cases} g(0) = 0 & \text{if } n = 0, \\ g(n) = d + e(n - 1) & \text{if } n \geq 1, \end{cases} \quad (6)$$

where d and e are gap opening and extension costs. After that the string kernel based on local alignment of exactly n residues is defined as:

$$K_n^{(\beta)}(x, y) = K_0 * \left(K_\alpha^{(\beta)} * K_\alpha^{(\beta)} \right)^{(n-1)} * K_\alpha^{(\beta)} * K_0. \quad (7)$$

This kernel quantifies the similarity of two strings x and y based on local alignments of exactly n residues. In order to compare two sequences through all possible local alignments, it is necessary to take into account alignments with different numbers n of aligned residues:

$$K_{LA}^{(\beta)} = \sum_{i=0}^{\infty} K_{(i)}^{(\beta)}. \quad (8)$$

The implementation of the above kernel can be done via dynamic programming [12].

3.2 Synthetic Protein Sequence Oversampling (SPSO)

In classification of GPCRs at the subfamily and specially sub-subfamily level we are facing an imbalanced dataset. There have been two types of solutions to this problem. The first type, as exemplified by different forms of re-sampling techniques, tries to increase the number of minor class examples (oversampling) or decrease the number of major class examples (undersampling) in different ways. The second type adjusts the cost of error or decision thresholds in classification for imbalanced data and tries to control the sensitivity of the classifier [17, 18, 19, 20]. In protein classification problems the second type of those approaches has been applied more and a class-depending regularization parameter is added to the diagonal of the kernel matrix: $K'(x, x) = K(x, x) + \lambda n/N$, where n and N are the number of positive (or negative) instances and the whole dataset, respectively.

In GPCR classification, even with that method we could not get good results, especially at the sub-subfamily level. One important issue with imbalanced data is that making the classifier too specific may make it too sensitive to noise specially with highly imbalanced datasets, having a ratio of 100 to 1 and more, the classifier often treats positive data as noise and considers it as negative data and we also have instabilities in the classifier. It means the cost that we consider for an error can be an important issue, and sometimes choosing a value near the optimum value can give unsatisfying results. Then, in this case, only using a different error cost method (DEC) [19] is not suitable. We found out that if we can add synthetic sequences (oversampling) at the sub-subfamily level (minority class) in a way that those added sequences are related to that class and away from other classes (majority class), the accuracy of a classifier will be increased. For that, we used our newly developed algorithm named synthetic protein sequence oversampling (SPSO) technique [24] in which the minority class in the data space is oversampled by creating synthetic examples. It considers the distribution of residues of the protein sequence using a hidden Markov model profile of the minority class and also one of the majority class and then synthesizes protein sequences which can precisely increase the information of the minor class. We used this method along with the DEC method to increase the sensitivity and stability of the classifier.

4 Results

In this study we used the local alignment kernel (LA kernel) to generate vector from protein sequences. For this, we divided the data into training and test data and then build a kernel matrix K for the training data as shown in Fig. 3. Each

cell of the matrix is a local alignment kernel score between protein i and protein j . After that we normalized the kernel matrix via $K_{ij} \leftarrow K_{ij} / \sqrt{K_{ii}K_{jj}}$. We used the SPSO algorithm, explained above, for each subfamily or sub-subfamily whose number of sequences in the training set was less than 50 and more than 4, to synthetically increase the number of data up to 800 percent (depending on the number of sequences). Each subfamily or sub-subfamily is considered as positive training data and all others as negative training data. After that the SVM algorithm with RBF kernel is used for training and for highly imbalanced data (after oversampling) we also use the DEC (different error cost) method. For testing, we create feature vectors by calculating a local alignment kernel between the test sequence and all training data.

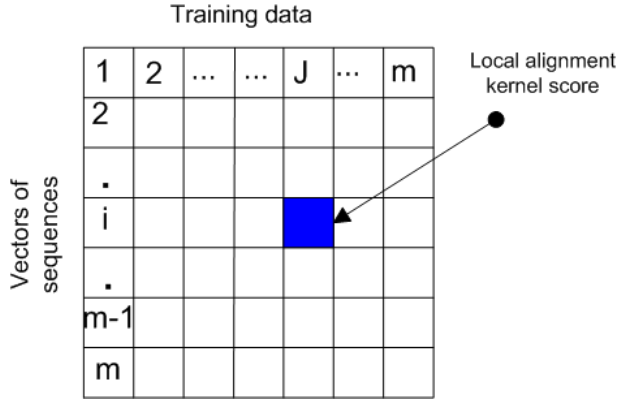


Fig. 3. Calculating the kernel matrix of the training data

In subfamily classification we randomly partitioned the data in two non-overlapping sets and used a two-fold cross validation protocol. The training and testing was carried out twice using one set for training and the other one for testing. The prediction quality was then evaluated by Accuracy (ACC), Matthew's correlation coefficient (MCC), overall Accuracy ($\overline{ACC}$) and overall MCC ($\overline{MCC}$) as follows:

$$ACC = \frac{TP + TN}{(TN + FN + TP + FP)} \quad (9)$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TN + FN)(TP + FN)(TN + FP)(TP + FP)}} \quad (10)$$

$$\overline{ACC} = \sum_{i=1}^N \frac{ACC(i)}{N} \quad (11)$$

$$\overline{MCC} = \sum_{i=1}^N \frac{MCC(i)}{N} \quad (12)$$

Table 1. The performance of our method in GPCRs subfamily classification (Class A)

Class A subfamilies	Accuracy (%)	MCC
Amine	99.9	0.99
Peptide	97.8	0.97
Hormone protein	100.0	1.00
(Rhod)opsin	99.6	0.99
Olfactory	99.9	0.99
Prostanoid	99.9	.98
Nucleotide-like	100.0	1.00
Cannabinoid	100.0	1.00
Platelet activating factor	100.0	1.00
Gonadotropin-releasing hormone	100.0	1.00
Thyrotropin-releasing hormone	100.0	1.00
Melatonin	100.0	1.00
Viral	87.0	0.8
Lysosphingolipid	100.0	1.00
Leukotriene	100.0	1.00
Overall	98.4	0.98

Table 2. The performance of our method in GPCRs subfamily classification (Class B)

Class B subfamilies	Accuracy (%)	MCC
Calcitonin	100.0	1.00
Corticotropin releasing factor	100.0	1.00
Glucagon	100.0	1.00
Growth hormone-releasing hormone	100.0	1.00
Parathyroid hormone	100.0	1.00
PACAP	100.0	1.00
Secretin	100.0	1.00
Vasoactive intestinal polypeptide	100.0	1.00
Diuretic hormone	99.1	0.91
EMR1	100.0	1.00
Latrophilin	100.0	1.00
Brain-specific angiogenesis inhibitor	100.0	1.00
Methuselah-like proteins (MTH)	100.0	1.00
Cadherin EGF LAG (CELSR)	100.0	1.00
Overall	≈ 100	0.99

(TP = true positive, TN = true negative, FP = false positive , FN = false negative, N =number of subfamily or sub-subfamily)

In our study, we used the Bioinformatics Toolbox of MATLAB to create the HMM profiles of families and the SVMlight package [23], to perform SVM training and classification.

Tables 1, 2 and 3 show the results of subfamily classification for classes A,B and C of GPCRs. We see that even when the number of sequences is low, the

Table 3. The performance of our method in GPCRs subfamily classification (Class C)

Class C subfamilies	Accuracy (%)	MCC
Metabotropic glutamate	92.1	0.84
Calcium-sensing like	94.2	0.82
Putative pheromone receptors	98.7	0.93
GABA-B	100.0	1.00
Orphan GPRC5	97.1	0.96
Orphan GPRC6	100.0	1.00
Taste receptors (T1R)	97.2	0.81
Overall	96.95	0.91

Table 4. The performance of our method in GPCRs sub-subfamily classification for Class A,B and C

Class A subfamilies	Overall Accuracy (%)	Overall MCC
Amine	97.1	0.91
Peptide	99.9	0.93
Hormone protein	100.1	1.00
(Rhod)opsin	96.6	0.95
Olfactory	98.9	0.92
Prostanoid	98.0	0.94
Gonadotropin-releasing hormone	96.1	0.93
Thyrotropin-releasing hormone	91.2	0.94
Lysosphingolipid	98.4	1.00
Class B Latrophilin	100.0	1.00
Class C Metabotropic glutamate	98.1	0.96
Calcium-sensing like	97.2	0.93
GABA-B	100.0	1.00
Overall	97.93	0.95

accuracy of our method is high. The overall accuracy for families A, B and C is 98.94%, 99.94% and 96.95%, respectively, and overall MCC for families A, B and C is 0.98, 0.99 and 0.91, respectively. The results show that almost all of the subfamilies are accurately predicted with our method.

For sub-subfamily classification we used 5-fold cross validation. Table 4 shows the results for the sub-subfamily level. We see that in this level also the accuracy is high and we could classify most of GPCRs sub-subfamilies. We could obtain an overall accuracy of 97.93% and a MCC of 0.95 for all sub-subfamilies. At this level we could increase the accuracy, especially when the number of sequences in the positive training data was less than 10, and there was no example in which with our oversampling method the accuracy decreases. Table 5 shows the result of classification in some sub-subfamilies that we used only DEC (different error cost) compared with DEC along with the SPSO method. We tried to find optimum value for both rate of oversampling and error costs. We used the numbers to show the level of family, subfamily and sub-subfamily. For example 001-001-002 means the sub-subfamily Adrenoceptors that belongs to subfamily

of Amine (001-001) and class A (001) (as shown in Fig. 2). We see that with our method the MCC in general increases and, especially when the number of sequences is low, the efficiency of our method is apparent.

5 Discussion and Conclusion

GPCR family classification enables us to find the specificity for ligand that binds to the receptor and also to predict the function of GPCRs. Our aim in this study was to develop an accurate method for classification of GPCRs at the sub-subfamily level, at which we have the problem of imbalanced data. We chose a local alignment kernel(LA kernel) as suitable kernel for our classification task. Compared with HMMs, the LA kernel takes more time during the training phase, but according to results of other researchers, the accuracy of discriminative methods with that kernel is higher than with a generative method like HMMs [8, 9, 10]. To solve the problem of imbalanced data we used the SPSO algorithm that can be used along with DEC (different error cost). It makes the classifier less sensitive to noise (here negative data) and increases its sensitivity. Based on our experiments (not showed here) in classifying sub-subfamilies of a subfamily, we get more accurate results if we select all other sub-subfamilies as negative data rather than only sequences in that subfamily, despite the fact that the learning step of the SVM classifier takes more time, because of the higher dimension of the kernel matrix. But the problem of imbalanced data in this case is severe and we tried to solve it with DEC along with the SPSO algorithm. Our study shows again that a discriminative approach for protein classification of GPCRs is more accurate than a generative approach. At the subfamily level we compared our method with that of Bhasin et al. [21]. They used an SVM-based method with dipeptide composition of protein sequences as input. The accuracy and MCC values of our method outperform theirs. For example in classification of subfamily A, the overall accuracy and MCC of their method were 97.3% and 0.97 but ours are 98.4% and .98, respectively. They did a comparison with other previously published methods like that of Karchin et al. [8] and showed that their method outperformed the others. To the best of our knowledge there is only one study which has been done for sub-subfamily classification [13]. Their

Table 5. The result of sub-subfamily classification with and without SPSO oversampling for subfamilies of Peptide(Class A)

sub-subfamily	Number of sequence	DEC		DEC+SPSO	
		Accuracy(%)	MCC	Accuracy(%)	MCC
001-002-002	17	99.7	0.81	99.9	0.97
001-002-003	19	99.9	0.94	100.0	1.00
001-002-005	12	99.9	0.91	100.0	1.00
001-002-021	20	99.8	0.66	99.9	0.91
001-002-024	4	99.7	0.38	100.0	1.00
001-002-025	5	99.9	0.79	100.0	1.00

approach is based on bagging a classification tree and they achieved 82.4% accuracy for sub-subfamily classification, which is less accurate than ours (97.93% with MCC of 0.95) despite the fact that they had excluded families with less than 10 sequences (we only excluded families with less than 4 sequences).

References

- [1] T.K Attwood, M. D. R. Croning and A. Gaulton. Deriving structural and functional insights from a ligand-based hierarchical classification of G-protein coupled receptors. *Protein Eng* 15:7-12, 2002.
- [2] T. E. Herbert and M. Bouvier. Structural and functional aspects of G protein-coupled receptor oligomerization. *Biochem Cell Biol* , 76:1-11, 1998.
- [3] F. Horn, E. Bettler, L. Oliveira L, F. Campagne, F. E. Cohhen and G. Vriend. GPCRDB information system for G protein-coupled receptors. *Nucleic Acids Res.* 31(1):294-297,2003.
- [4] S. F. Altschul, T. L. Madden, A. A. Schaffer, Z. Zhang, W. Miller W and D. J. Lipman. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402, 1997.
- [5] J. Kim, E. N. Moriyama, C. G. Warr, P. J. Clyne, and J. R. Carlson. Identification of novel multi-transmembrane proteins from genomic databases using quasi-periodic structural properties. *Bioinformatics*,16(9):767775, 2000.
- [6] D. W. Elrod and K. C. Chou. A study on the correlation of G-protein-coupled receptor types with amino acid composition. *Protein Eng.*, 15, 713715, 2002.
- [7] B. Qian, O. S. Soyer and R. R. Neubig. Depicting a protein's two faces: GPCR classification by phylogenetic tree-based HMM. *FEBS Lett.*554, 95, 2003.
- [8] R. Karchin, K. Karplus, and D. Haussler. Classifying G-protein coupled receptors with support vector machines. *Bioinformatics*, 18(1):147159, 2002.
- [9] T. Jaakkola, M. Diekhans, and D. Haussler. A discriminative framework for detecting remote protein homologies. *Journal of Computational Biology*, 7(1-2):95114, 2000.
- [10] C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: A string kernel for SVM protein classification. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing*, pages 564575, New Jersey, World Scientific, 2002.
- [11] C. Leslie, E. Eskin, A. Cohen, J. Weston, and W.S. Noble. Mismatch string kernel for SVM protein classification. *Advances in Neural Information Processing System* 15, pages 1441-1448, 2003.
- [12] J.-P. Vert, H. Saigo, and T. Akustu. Convolution and local alignment kernel. In B. Schölkopf, K. Tsuda, and J.-P. Vert (Eds.), *Kernel Methods in Computational Biology*. The MIT Press.
- [13] Y. Huang, J. Cai, Y. D. Li, Classifying G-protein coupled receptors with bagging classification tree. *Computational Biology and Chemistry* 28:275-280, 2004.
- [14] A. Bairoch, R. Apweiler. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids res.* 29, 346-349, 2001.
- [15] H. saigo, J. P. Vert, N. Ueda and T. akustu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11): 1682-1689, 2004
- [16] D. Haussler. Convolution kernels on discrete structures. Technical Report UCSC-CRL-99-10, Department of Computer Science, University of California at Santa Cruz, 1999

- [17] M. Pazzini, C. Marz, P. Murphi, K. Ali, T. Hume and C. Bruk. Reducing misclassification costs. In proceedings of the Eleventh International Conference on Machine Learning, 217-225, 1994
- [18] N. Japkowicz, C. Myers and M. Gluch. A novelty detection approach to classification. In Proceeding of the Fourteenth International Joint Conference on Artificial Intelligence, 10-15, 1995.
- [19] N. Japkowicz. Learning from imbalanced data sets: A Comparison of various strategies. In Proceedings of Learning from Imbalanced Data, 10-15, 2000.
- [20] K. Veropoulos, C. Campbell and N. Cristianini. Controlling the sensitivity of support vector machines. Proceedings of the International Joint Conference on AI, 55-60, 1999.
- [21] M. Bhasin and G. P. S. Raghava. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. *Nucleic Acids Res.* 32, 383-389, 2004.
- [22] J. D. Thompson and D. G. Higgins, and T. J. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22, 4673-4680, 1994.
- [23] T. Joachims. Making large scale svm learning practical. Technical Report LS8-24, Universitat Dortmund, 1998.
- [24] M. Beigi and A. Zell. SPSO: Synthetic Protein Sequence Oversampling for imbalanced protein data and remote homology detection. VII international symposium on Biological and Medical Data Analysis ISBMDA, 2006.

Integration Analysis of Diverse Genomic Data Using Multi-clustering Results

Hye-Sung Yoon¹, Sang-Ho Lee¹, Sung-Bum Cho², and Ju Han Kim²

¹ Ewha Womans University, Department of Computer Science and Engineering, Seoul
120-750, Korea

comet@ewhain.net, shlee@ewha.ac.kr

² Seoul National University Biomedical Informatics (SNUBI), Seoul National
University College of Medicine, Seoul 110-799, Korea

csb1749@snu.ac.kr, juhan@snu.ac.kr

Abstract. In modern data mining applications, clustering algorithms are among the most important approaches, because these algorithms group elements in a dataset according to their similarities, and they do not require any class label information. In recent years, various methods for ensemble selection and clustering result combinations have been designed to optimize clustering results. Moreover, conducting data analysis using multiple sources, given the complexity of data objects, is a much more powerful method than evaluating each source separately. Therefore, a new paradigm is required that combines the genome-wide experimental results of multi-source datasets. However, multi-source data analysis is more difficult than single source data analysis. In this paper, we propose a new clustering ensemble approach for multi-source bio-data on complex objects. In addition, we present encouraging clustering results in a real bio-dataset examined using our proposed method.

1 Introduction

Recent data mining approaches employ multiple representations to achieve more general results that are based on a variety of aspects. The extraction of meaningful feature representations yields a variety of different views on the same set of data objects using various methods. Moreover, generating high-quality results is difficult, because of the inherent noise that exists in the application of data and the inconsistency that exists among different algorithms. Therefore, recent research has shown that combining the merits of several clustering methods often yields better results than using one method alone, and that clustering ensemble techniques can be applied successfully to increase classification accuracy and stability in data mining [2][6][11].

Different clustering techniques create different errors on the same set of data objects, which means that we can arrive at an ensemble that makes more accurate decisions by combining clustering results [1]. For this purpose, diverse clustering results are grouped together into what is known as a *cluster ensemble*.

However, previous work has identified several problems in optimizing the performance of clustering ensembles. First, previous methods generally have fixed the result numbers from the applied clustering algorithms, thereby resulting in the same number of clustering results; and second, highly-overlapped clustering results often are generated, clusters that are assumed to indicate the final clustering result. These problems are fundamentally difficult, and cannot be solved to yield better results. Directly combining the same number of clustering results cannot generate a meaningful result, because of the inherent noise that exists in the data, and because of the inconsistency that exists between different clustering algorithms. It also remains difficult to say which clustering result is best, because the same algorithm can lead to different results, merely secondary to repetition and random initialization. Meanwhile, with respect to the latter ensemble combination, this method generates clustering results with the same parameters to all applied algorithms. Here too, it is difficult to say which clustering result is best; even though there are different numbers of clustering results, this is not considered a characteristic of the clustering algorithm or the applied data set.

Bioinformatics is a combined interdisciplinary subject that focuses on the use of computational techniques to assist in the understanding and organization of information associated with biological macromolecules. Bioinformatics not only deal with raw DNA sequences, but also with other various types of data, such as protein sequences, macromolecular structure data, genome data and gene expression data [9]. These various types of data provide researchers with the opportunity to predict phenomena that formerly had been considered unpredictable, and most of these data can be accessed freely on the internet. Among the features of bio-data, one is that the same variables can be used to generate different types of multi-source data through a variety of different experiments and under several different experimental conditions. These multi-source data are useful for understanding cellular function at the molecular level, and they also provide further insight into their biological relatedness by means of information from disparate types of genomic data.

This paper describes a machine learning approach to an information fusion method intended for combining and analyzing multi-source genomic data. Our proposed method involves a diversity-based clustering ensemble mechanism that identifies optimal clusters, using collaborative learning of an unsupervised clustering method, based on multi-source bio-data.

The remainder of this paper is organized as follows. The application of multi-source data and clustering ensemble methods are reviewed in Section 2. Section 3 explains the proposed diversity-based clustering ensemble method, based upon genetic algorithm (GA). Section 4 describes experimental results generated by applying the proposed method, and compares these results with those generated using three other algorithms. Finally, concluding remarks and possibilities for future research are presented in Section 5.

2 Multi-source Data Analysis and Clustering Ensemble

Although the volume of data in molecular biology is growing at an exponential rates, the key features of this biological data are not so much their volume, but their diversity, heterogeneity and dispersion. Therefore, combining and analyzing different types of data is widely acknowledged in bioinformatics and genomics.

The objective of data integration analysis is to compile information from multiple data sources, so as to generate experimental results that better fit the users' goals. Also, multi-source data analysis provides and identifies correlations more accurately, using diverse independent attributes in gene classification, clustering, and regulatory networks. The collection of bio-data sources has the property that similar data can be contained in several sources, and represented in several different ways depending upon the source. However, this multi-source data analysis is useful in understanding cellular functions at the molecular level, and in providing further insight into the cells' biological relatedness. In [10], the problem of inferring gene functional classification from a heterogeneous dataset consisting of DNA microarray expression measurements and phylogenetic profiles from whole-genome sequence comparisons is considered; [10] also demonstrates that more important information can be extracted by means of using disparate types of data.

Many genomic researchers apply clustering algorithms to gain various genetic understandings of and biological information from bio-data. Clustering algorithms comprise a technique of unsupervised learning, whereby the task is to identify interesting patterns that exist within an inadequately-labeled bio-data set [14]. However, it remains difficult to say which clustering result is best, because the same algorithm can lead to many different results, as a result of repetition and random initialization. *Clustering ensemble* is a method that combines several runs of different clustering algorithms to achieve a common partition of the original dataset, aiming for consolidation of results from a portfolio of individual clustering results. This method also combines clustering results through several clustering algorithms, to generate a specific view of the data. Each clustering algorithm outputs a clustering result or label, comprised of group labels for some or all objects.

Generating high-quality clustering results is difficult, because of the inherent noise that exists in the experimental data and the different characteristics that exist among different clustering algorithms [4][5][7][8][12]. One of the major dilemmas associated with clustering ensembles is how to combine different clustering results [3]. Previous reports describing other methods have referred to the importance of ensemble algorithms, but the methods used fixed the cluster number from the clustering algorithms and ended up with the same number of clustering results [13]. However, directly combining the same number of clustering results cannot generate a meaningful result. In addition, highly-overlapped cluster results were assumed to indicate a final clustering result, but these investigators invariably searched for the optimal cluster number as well, and reapplied that cluster number, as a parameter, to all algorithms.

Several important factors must be considered when applying clustering ensemble methods.

- (a) One must find a pertinent objective function when selecting the clustering results;
- (b) One must use pertinent clustering algorithms to apply the ensemble;
- (c) One must use an adequate fusion function to combine cluster outputs.

Diversity measures are designed to be objective functions for ensemble selection, but their performance is not convincing. Moreover, when *genetic algorithms* (GA) are used as a searching algorithm for ensemble selection, the evaluation of diversity measures may be very time consuming. To offset this problem, we now propose a method for selecting and combining cluster results. Our proposed method combines diversity measures from a multi-source dataset with the simple proposed method of GA operators, and thus allows for effective GA searching for ensemble selection.

In this paper, we assumed that our proposed method may outperform other methods in two ways. First, analysis of combined biological datasets should lead to a more understandable direction than experimental results derived from a single dataset. Second, the same variables can be used to make various types of multi-source data through different experiments and under several different experimental conditions. Therefore, we focus on optimizing the information provided by a collection of different clustering results, combining them into one final result from different data sources, using a variety of proposed methods.

3 Methods

In this section, the experimental data and experimental methods applied in this paper are explained, in detail.

3.1 Experimental Data

In this paper, the CAMDA 2006 conference dataset¹, was used as a source of multi-source data in order to test the application of the proposed method. This dataset is derived from the CDC (Center for Disease Control and Prevention) chronic fatigue syndrome (CFS) research group and contains microarray, proteomics, single nucleotide polymorphisms (SNPs), and clinical datasets. CFS is a condition that is diagnosed based upon classification criteria that are highly subjective, for the most part. The illness has no disease-specific diagnostic clinical signs or laboratory abnormalities, and it is unclear if CFS represents a single entity or a spectrum of many disorders. Prior analyses into CFS pathogenesis have not yielded further insights into the nature of this condition. One objective of the current study was to observe how our proposed method might deal with various experimental datasets on CFS, a condition for which both the clinical parameters and the pathogenesis of disease are unclear.

¹ <http://www.camda.duke.edu/camda06/datasets>

In our experiments, three data categories - microarray, proteomics and clinical - were used for application and verification. The first dataset, microarray data, is a single-channel experimental dataset that is comprised of 20,160 genes, using DNA from 177 patients. The second dataset, a proteomics dataset, was generated from three ProteinChip Array chemistries on the same samples (patients): Reversed Phase (*H50*), Metal Affinity Capture (*IMAC30*) and Weak Cation Exchange (*CM10*) to detect the maximal number of proteins. Among these several conditions of proteomics data, we applied four (2x2) different experimental conditions H50 and IMAC30 ProteinChip data under both high and low stringency conditions. Clinical data were used to validate the proposed method. We compared our method with three other clustering algorithms, using data from 64 patients who were common to both the microarray and proteomics datasets.

3.2 Diversity-Based Clustering Ensemble

Our diversity-based clustering ensemble approach is described as follows.

3.2.1 Generating Clustering Result Outputs

We first had to identify the optimal clustering algorithm for analysis of multi-source bio-data. However, we were faced with the inherent challenges due to the diverse features of multi-source data and the existence of many clustering algorithms. To counteract some of these concerns, we applied clustering algorithms with various characteristics to a given multi-source. We also constructed paired subsets with two clustering results that were composed of different numbers of clustering results from applied clustering algorithms. The next step was to select two parents as a couple, the couple with the largest number of highly-overlapped elements of the fitness function $F(t)$ to allow for crossover into the next GA operation. Continually, the previous process replaced two parents from the population to generate offspring after crossover, until an optimal subset was formed.

The following explains the order of the proposed method, by which we applied GA operators to a multi-source bio-data set.

3.2.2 Application of GA Operators

We propose new two GA operators, Selection and Crossover, in order to generate the optimal result.

■ Method for ensemble selection

Once a suitable chromosome is chosen for analysis, it is necessary to create an initial population to serve as the starting point for the GA. The following explains in the order of the proposed selection method, with examples.

1. We construct paired subsets from two clustering results, out of all the possible clustering results for the population generation. Generating the initial population for the selection operator combines different clustering results, because multi-source bio-datasets can lead to different outputs.

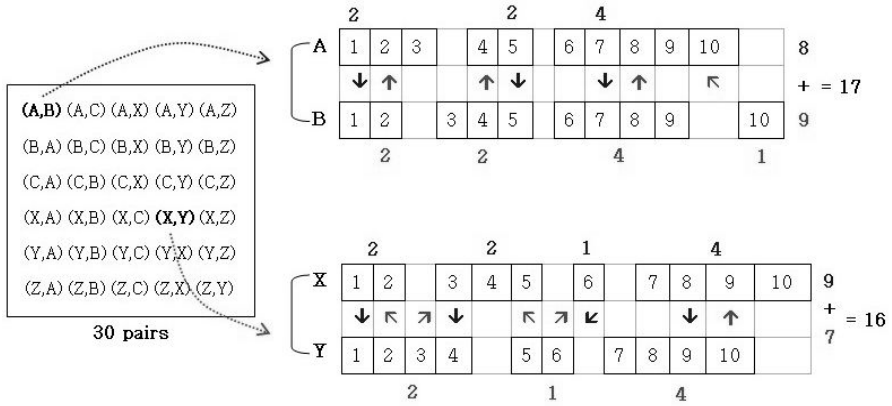


Fig. 1. Selection method for the evolutionary reproduction process

- After generating the initial population, the next step involves selecting parents for recombination. We applied the *roulette wheel selection method* as our proposed crossover operation in this paper.

Roulette wheel selection - Simple reproduction allocates offspring using a roulette wheel, with slots that are sized according to fitness value. This is one method of choosing members from a population of chromosomes with a probability that is proportional to their fitness value. Parents are selected according to their fitness value. The better the fitness of the chromosome, the greater the probability that it will be selected.

- In the initial population, we selected that pair had the higher fitness value; that is, two clustering results that form a pair with highly-overlapped elements. Suppose that bio-data containing 10 elements and a pair (A, B) with three and four clustering results are compared. The largest number of highly-overlapped elements is the representative cluster value. Specifically, the first cluster (1, 2, 3) of A is compared with the other clusters {(1, 2) (3, 4, 5) (6, 7, 8, 9) (10)} of B, as shown in Figure 1. The first cluster (1, 2, 3) from A and the first cluster (1, 2) from B have two values that are more highly-overlapped than the {(3, 4, 5) (6, 7, 8, 9) (10)} of B. Moreover, the (1, 2, 3, 4) cluster of Y has the same value as two of the highly-overlapped parents, with the other cluster being between the (1, 2) and (3, 4, 5) clusters of X. This process adds the representative values of each cluster and selects a final pair among 30 pairs population. As shown as (A, B) and (X, Y) in Figure 1, the representative values have 17 and 16, respectively. In this case, (A, B) pair has a greater probability of selection than the (X, Y) pair by having 17 value.

This is a process by which each chain is copied according to the values of the function which one wishes to optimize. It means that chains with greater fitness function values have a greater probability of contributing to the following generation, by creating offspring, than those with lesser fitness values. This operator is an artificial version of natural selection, wherein fitness is determined by the ability of individuals to survive.

The selection of a paired subset is executed whether each element in the clusters will survive or not, and this method is proposed as the crossover operator as follows.

■ Method for ensemble combination

Figure 2 shows the proposed crossover method.

1. During this phase, a pair produced by the selection phase initially is matched. For example, P_1 and P_2 are selected to two parents in the population.
2. Suppose that P_1 has three clustering results (C_{1_1} , C_{1_2} , and C_{1_3}) and P_2 has five clustering results (C_{2_1} , C_{2_2} , C_{2_3} , C_{2_4} , and C_{2_5}). First, we select the first cluster among the three clustering results from P_1 and see that it has more highly-overlapped traits than the other two clusters, when compared to clusters from P_2 .
3. This process makes progress based upon all the clustering results of P_1 . Moreover, if C_{1_1} and C_{2_3} of P_2 have the largest number of similarities, then we replace traits C_{1_1} and C_{2_3} via the following process. The C_{1_1} traits include 7, 27, 39, 58, 63, 65, 71 and 84, and C_{2_3} traits include 7, 27, 39, 58, 59, 65 and 85. In the replacement process, certain traits in C_{1_1} (63, 71, and 84) do not appear as overlapping traits in C_{2_3} . However, traits 63 and 84 in C_{1_1} do appear as traits in C_{1_2} and C_{1_3} , respectively. Consequently, traits 63 and 84 are removed, so that each trait only belongs to one cluster. The remaining trait in C_{1_1} (trait 71) is taken from C_{2_3} , so that it does not appear in any other cluster.
4. Finally, the new clustering solution is represented by the first offspring's possessing traits (C_{2_1} , C_{2_2} , revised C_{1_1} , C_{2_4} , and C_{2_5}).
5. This crossover operation is repeated once more, by selecting a cluster from P_2 to generate the second offspring. Two parents, P_1 and P_2 , are replaced by new offspring in the final population.
6. After replacement, we again compute the fitness function in the new paired non-empty subsets to generate two clustering results; then we determine another pair of new candidates for the subsequent parent selection; and repeat the stages above.

Our proposed crossover operation exchanges the clustering traits from different clustering results and traits with highly-overlapped and meaningful information inherited by the offspring, until we ultimately achieve an optimal clustering result.

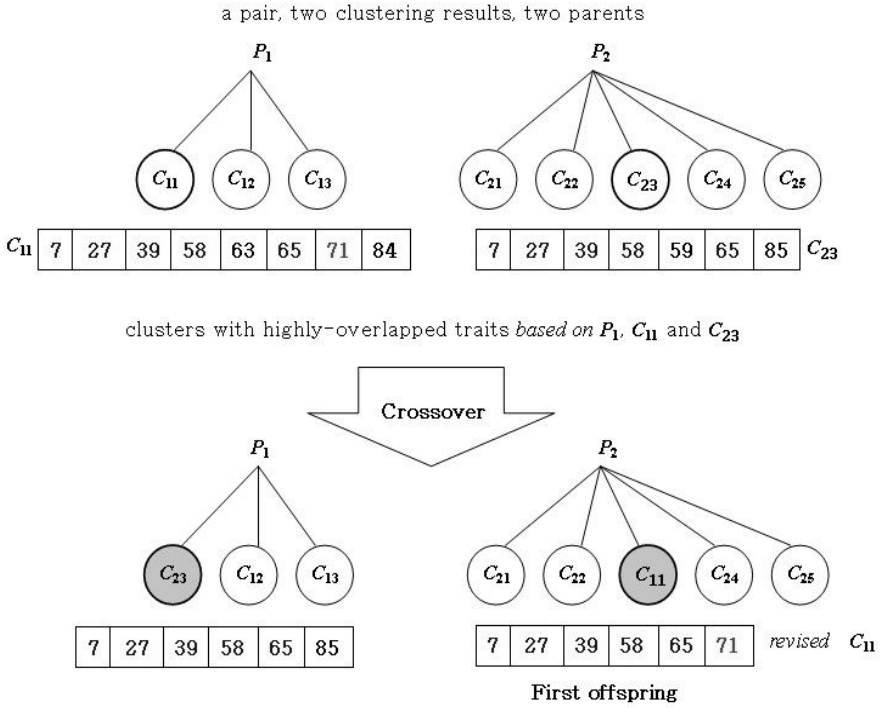


Fig. 2. Crossover operation for generating first offspring, based upon the parent P_1

4 Experimental Results

In this paper, the CLUSTER analysis tool² was used to generate clustering outputs from various clustering algorithms. Our experimental environment was conducted on Pentium 4 PC with 2.8G Hz CPU and 1GB. The proposed method was implemented using JAVA 1.4.2 language. The CLUSTER analysis tool performs a variety of types of clustering algorithms: hierarchical clustering, self-organizing maps (SOMs), k -means clustering, and principal component analysis. Of these, we applied hierarchical clustering, self-organizing maps, and k -means clustering algorithms, and compared the results generated using CLUSTER to those of our proposed method.

To generate clustering results using three applied algorithms, we set parameters as in Table 1.

For data analysis and validity testing, we selected 44 patients in common between the clinical, microarray and proteomics datasets.

Table 2 lists the comparisons between our method and the other clustering algorithms created by the parameter change using the H50 low and IMAC30 high-proteomics dataset. This demonstrates that the results generated using a

² <http://rana.lbl.gov/EisenSoftware.htm>

Table 1. Parameters applied to the clustering algorithms of the CLUSTER tool

Algorithms	Parameters
Hierarchical	All linkage clustering, based on arrays
SOMs	Ydim: 5,7,9 and 200–2,000 iterations, based on arrays
<i>k</i> -means	max cycles: 100 and <i>k</i> =3,4,5, based on arrays

Table 2. A comparison of the clustering algorithms

H50.Low				
Cluster <i>k</i> -means	Hierarchical	SOMs	Our method	Actual value
#				
3	(W,L,L)	(L,L,L)	(W,W,L/W)	(W,W,W)
4	(L,L,L/W,W)	(L,W,L,L)	(L,W,L/W,L/W)	(L,W,L,L)
5	(L,L,L,L,W)	(L,W,L,L/W,L)	(L,L,L/W,L/W,W)	(L,L,L/W,L,W)

IMAC30.High				
Cluster <i>k</i> -means	Hierarchical	SOMs	Our method	Actual value
#				
3	(L,L/W,L)	(L,W,L)	(L,L,L/W)	(L,L,L)
4	(L,W,L,W)	(L,L,W,L)	(W,L/W,L,L)	(L/W,L/W,L,L/W)
5	(L,W,L,W,L)	(W,L,L,W,L)	(L,L,L,L/W,L)	(L,W,L,L,L)

clustering algorithm when we have no previously defined clusters are no more consistent with the three clinical datasets than the proposed method. Specifically, the clinical dataset from CAMDA was classified into three cluster groups, based upon the overall severity of CFS symptoms- least symptoms (L), mid-level symptoms (M), and worst symptoms (W).

For validity testing, we chose to use those representative symptoms with the largest number of similarities. The representative values that are similar between the proposed method and the three different algorithms are written in bold characters. In Table 2, we discover that the results generated by our diversity-based clustering ensemble method more closely agree with the clusters classified using clinical data than the results produced by any of the other clustering algorithms. Here, L/M and M/W are found to cluster in the same ratio as the number of patients classified as least/middle and middle/worst.

Table 3 compares the cluster results for single source datasets (individual microarray and proteomics data) with the true classified clusters of the clinical dataset, using roulette wheel selection.

As shown in table 3, the proposed method demonstrates that five cluster results generate the best fitness value in paired clustering of various data sources. That is, this final cluster result number produced the most representative selected pair in the paired population. We chose the symptomatic class with the most representation and the largest number of similarities for validity testing.

Table 3. A comparison of the microarray and proteomics datasets

		Diversity-based Clustering Ensemble			Actual classification
Data set	Cluster results#	Least (L)	Moderate (M)	Worst (W)	Representative value
Microarray	1	2	3	0	L
	2	5	3	3	L/W
	3	6	1	6	W
	4	4	3	2	L
	5	2	1	3	M
Total		44 patients			44 patients

		Diversity-based Clustering Ensemble			Actual classification
Data set	Cluster results#	Least (L)	Moderate (M)	Worst (W)	Representative value
Proteomics	1	1	1	0	L/M
	2	2	1	3	W
	3	5	2	5	W
	4	6	5	4	L
	5	5	2	2	L
Total		44 patients			44 patients

Table 4. A comparison of the microarray and proteomics datasets

		Diversity-based Clustering Ensemble			Actual classification
Data set	Cluster results#	Least (L)	Moderate (M)	Worst (W)	Representative value
Multi-source data sets	1	3	2	3	L/W
	2	5	2	5	L/W
	3	6	5	4	L
	4	5	2	2	L
Total		44 patients			44 patients

From these result tables, we found that the proteomics data yielded better experimental results than the microarray data, because the proteomics data more closely agrees with the clusters classified using the clinical data (comparison in bold typeface).

We also explain the experimental results of multi-source datasets. As more data sources are added to the experiment (combined microarray and proteomics data), the experimental results lead to better cluster solutions. As shown in Table 4, using the proposed method on four clusters produced the best fitness value among the generated paired subsets, and the four-cluster results were most comparable to the actual clinical data.

Here, multi-source datasets using our proposed method mostly agree with the clusters classified by clinical data. In addition, the cluster results using a data

source are no more consistent with the three symptomatic classes (L, M, and W) of the clinical dataset than the multi-source dataset generated by our proposed method. Therefore, we can say that our proposed method yields better cluster results than applying clustering algorithms to a single data source.

5 Conclusion and Discussion

We proposed a diversity-based clustering ensemble approach to generate optimal clusters on multi-source bio-data, by designing and applying new operators of the GA. We initially considered the problems inherent in combining different clustering results, by considering multi-source bio-data characteristics and the analysis of different clustering results. We also considered characteristics that present optimal cluster results from different clusters and different clustering algorithms. The experimental results show that a combined clustering approach using multi-source bio-data can generate better cluster results than those obtained using just one data source. In addition, combining clustering results from different clustering algorithms can produce better end-result clusters than the single clustering results generated using a single clustering algorithm. We need not remove elements for preprocessing, nor fix the same number of clusters during the first application step, because the GA approach is a stochastic search method that has been successfully applied in many search, optimization, and machine learning problems.

The experimental methods introduced in this paper suggest several avenues that can be taken for future research. One direction would be to identify other bio-information based on genes, as opposed to patients, in multi-source datasets. Our experimental datasets were consistent in that the rows of genes and columns of patients reflected the same level of CFS disease. We applied the columns data based on patients. Another direction, since three biological data types were used for multi-source analysis, would be to include multiple biological data types in order to discover optimal cluster results and then to again apply our proposed method. Another important task would be to develop a more theoretically and experimentally-justified verification system of multi-source data than we currently have.

References

1. Alexander, P.T., Behrouz, M-B., Anil, K.J., William, F.P.: Adaptive clustering ensembles. *Proceedings of the International Conference on Pattern Recognition*, **1** (2004) 272–275
2. Alexander, S., Joydeep, G.: Cluster ensembles-A knowledge reuse framework for combining partitionings. *Journal of Machine Learning*, **3** (2002) 583–617
3. Ana, L.N. Fred., Anil, K.J.: Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **27** (2005) 835–850
4. Duda, R.O., Hart., P.E., Stork, D.G.: "Pattern classification", seconded. Wiley, (2001)

5. Everitt, B.: "Cluster analysis. John Wiley and Sons", Inc., (1993)
6. Greene, D., Tsybal, A., Bolshakova, N., Cunningham, P.: Ensemble clustering in medical diagnostics. Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems, (2004) 576–581
7. Jain, A.K., Murty, M.N., Flynn, P.J.: Data clustering: A review. ACM Computing Surveys, **31** (1999)
8. Kaufman, L., Rosseeuw, P.J.: "Finding groups in data: An introduction to cluster analysis", John Wiley and Sons, Inc., (1990)
9. Larray, T.H.Yu., Fu-lai, C., Stephen, C.F.: Using emerging pattern based projected clustering and gene expression data for cancer detection. Proceedings of the Asia-Pacific Bioinformatics Conference, **29** (2004) 75–87
10. Pavlidis, P., Weston, J., Cai, J., Grundy, W.N.: Learning gene functional classifications from multiple data types. Journal of Computational Biology, **9** (2002) 401–411
11. Qiu, P., Wang, Z. J., Liu, K.J.: Ensemble dependence model for classification and prediction of cancer and normal gene expression data. Bioinformatics and Bioengineering, (2004) 251–258
12. Theodoridis, S., Koutroumbas, K.: "Pattern recognition", Academic Press (1999)
13. Xiaohua, H., Illhoi, Y.: Cluster ensemble and its applications in gene expression. Proceedings of the Asia-Pacific Bioinformatics Conference, **29** (2004) 297–302
14. Zhou, Z.-H., Tang, W.: Clustering ensemble. Knowledge-Based Systems, (2006)

Effectivity of Internal Validation Techniques for Gene Clustering

Chunmei Yang¹, Baikun Wan¹, and Xiaofeng Gao²

¹ Department of Biomedical Engineering and Scientific Instrumentations, Tianjin University,
Tianjin 300072, China

² Motorola (China) Electronics Ltd. Tianjin 300457, China
yangcm@tju.edu.cn

Abstract. Clustering is a major exploratory technique for gene expression data in post-genomic era. As essential tools within cluster analysis, cluster validation techniques have the potential to assess the quality of clustering results and performance of clustering algorithms, helpful to the interpretation of clustering results. In this work, the validation ability of Silhouette index, Dunn's index, Davies-Bouldin index and FOM in gene clustering was investigated with public gene expression datasets clustered by hierarchical single-linkage and average-linkage clustering, K-means and SOMs. It was made clear that Silhouette index and FOM can preferably validate the performance of clustering algorithms and the quality of clustering results, Dunn's index should not be used directly in gene clustering validation for its high susceptibility to outliers, while Davies-Bouldin index can afford better validation than Dunn's index, exception for its preference to hierarchical single-linkage clustering.

Keywords: gene expression data, gene clustering, cluster validation, internal validation measure.

1 Introduction

Clustering is a major exploratory technique for gene expression data analysis in post-genomic era [1-4]. Genes with related functions can be clustered into groups of co-expressed genes according to the similarities in their expression profiles, which is helpful to understand gene function, gene regulation, cellular processes, and subtypes of cells. Many clustering algorithms have been proposed for gene expression data, among which the classical Hierarchical clustering, K-means clustering and Self-organizing maps (SOMs) are widely used for their convenience [5,6]. However, the clusters obtained by different clustering algorithms can be remarkably different [7]. And, most current clustering algorithms do not provide estimates of the significance of the results returned. The produced partitions are commonly assessed by subjective visual inspection using prior biological knowledge [8]. Whether the clusters actually correspond to the real structure in the data is somewhat less frequently considered. Furthermore, most clustering algorithms return clusters even in the absence of actual

structure, generating results without signification. Hence, assessing the clustering results and interpreting the clusters found are as important as generating the clusters. Cluster validation techniques are clearly essential tools within cluster analysis, and their frequent neglect in the post-genomic literature hampers progress in the field. For example, many novel clustering algorithms are insufficiently evaluated, such that users remain unaware of their relative strengths and weaknesses. Cluster validation techniques have the potential to provide an analytical assessment of the amount and type of structure captured by a partitioning, and should therefore be a key tool in the interpretation of clustering results.

Validation techniques can be divided into two main categories: external and internal validation measures [9]. External validation measures evaluate a clustering result by comparing it to a given “gold” standard which is another partition of the objects, for example, F-measure and Rand index. Evidently, this is useful to permit an entirely objective evaluation and comparison of clustering algorithms and clusters on benchmark data. In cases where no “gold” standard is available, an evaluation based on internal validation measures becomes appropriate. Internal validation techniques do not use additional knowledge in the form of class labels, but base their quality estimate on the information intrinsic to the data. Specifically, they attempt to measure how well a given partitioning corresponds to the natural cluster structure of the data. Without the demand for external “gold” standard, internal validation techniques can be used more extensively. Handl et al [10] reviewed the fundamental concepts behind different types of validation techniques and discussed some of their biases and problems.

In the current research work on clustering of gene expression data, Silhouette index and Dunn’s index are mostly used to evaluate the clustering quality and predict the number of clusters [7,10-12], while Davies-Bouldin index is less used. By overview of the literates, Bolshakova et al. studied the validity of Silhouette index, Dunn’s index and Davies-Bouldin index in predicting the number of clusters for sample clustering in their two papers [11,13]. Considering the ultimate differences between gene clustering and sample clustering [1], can these validation measures be used in gene clustering? Moreover, Yeung et al. [7] proposed Figure of merit (FOM) for the cluster validation of gene expression data. What are the relative strengths of these measures in the validation of gene clustering? What’s the performance of Davies-Bouldin index compared to Dunn’s index? Aiming at these questions, we examined the performances of Silhouette index, Dunn’s index, Davies-Bouldin index and FOM in the validation of gene clustering.

2 Materials and Methods

2.1 Gene Expression Datasets

Yeast Sporulation Data (Spor): The sporulation data was collected and analyzed by Chu et al. [14] to explore the temporal program of gene expression during sporulation

of yeast *Saccharomyces cerevisiae*. We used a subset named Spor containing 161 genes from six distinct temporal patterns they observed.

Yeast Cell Cycle Data (Cellcycle): The yeast cell cycle data [15] showed the fluctuation of expression levels of approximately 6000 genes over two cell cycle (17 time points). We used two different subsets of this data from Yeung [7]. The first subset containing 384 genes whose expression levels peak at different time points corresponding to the five phase of cell cycle. The second subset consists of 237 genes corresponding to four categories in the MIPS database [16]. Conveniently, the subset with the 5-phase criterion is named Cellcycle_384, and that with MIPS criterion is named Cellcycle_237.

Rat CNS Data (CNS): The rat CNS dataset was obtained by Wen et al. [17] using reverse transcription-coupled PCR to study the expression levels of 112 genes during rat central nervous system development over nine time points. The 112 genes are classified into four functional categories based on prior biological knowledge.

Yeast Galactose Data (GAL): The yeast galactose data was collected by Ideker et al. [18] using cDNA microarrays to examine genes expression during galactose utilization. Yeung et al. [19] extracted a subset of 205 genes from this data as GAL dataset, whose expression patterns reflecting 4 functional categories in the Gene Ontology (GO) listings.

Human Serum Data (Serum): The human serum data obtained by Iyer et al. [20] reflects the temporal program of transcription during the response of human fibroblasts to serum. Xu et al. [21] analyzed this data and partitioned the profiles of 517 genes into five meaningful clusters. The 517 genes' expression profiles construct our dataset Serum.

2.2 Research Methods

We implemented hierarchical single-linkage clustering (HSL), hierarchical average-linkage clustering (HAL), K-means and SOMs on these expression datasets, and calculated Silhouette index, Dunn's index, Davies-Bouldin index as well as FOM of the resulting gene clusters to investigate their effectiveness in cluster validation. According to Yeung et al. [7], Datta et al. [6] and our previous research [22], HAL tends to produce better clusters than HSL, while K-means and SOMs have good predictive power. Moreover, we randomly partition the genes into random clusters for strengthened control [7]. Obviously, the random clusters are unacceptable. Therefore, validation indices of good clusters should distinguish from those of random clusters. In the experiment, the random partition was repeated 1000 times and the average was taken as the final index. Here, the number of clusters was set as varying from 2 to 30, standard Euclidean distance was used as similarity metric. Please refer to Ref. [23] for detailed description of the clustering algorithms.

For the definition of Silhouette index, Dunn's index and Davies-Bouldin index, the readers are recommended to Ref. [24, 25]. According to the studies of Bolshavoka et al.[11] and Azuaje et al.[26], we selected the average-linkage distance and the average diameter in the calculation of Dunn's index and Davies-Bouldin index. What's more, the definition of the adjusted 2-norm FOM in Ref. [7] was used in the calculation.

3 Results and Discussion

3.1 Silhouette Index

The Silhouette indices from the four clustering algorithms and random partition on the six expression datasets are illustrated in Fig.1. On each dataset, Silhouette curves from the three accepted clustering algorithms, HAL, K-means and SOMs, are distinctly above that from random partition, indicating remarkably better results. What's more, Silhouette index from HAL, K-means and SOMs outperforms that from HSL in most cases. It illustrates that Silhouette index can properly reflect the quality of gene clusters and the performance of clustering algorithms.

At the same time, Silhouette index from random partition varies sharply along different number of clusters, K , implying the existence of bias with respect to K . The bias, unavoidably, would affect the ability of Silhouette index for estimating the number of clusters. As an example, the Silhouette indices from HAL, K-means and SOMs, as K varying from 2 to 8 clusters, are listed in table 1. According to the definition of Silhouette index, the partition maximizing the index is taken as the optimal partition [24]. In table 1, the values in italics represent the predicted number of clusters in the small range of K , while the number of classes is bracketed below each dataset. As values with asterisk indicating cases of correct estimate, there are only 3 cases where the number of clusters is correctly predicted over such a small range of K . It informs us of the fact that Silhouette index can predict hardly the number of clusters for the bias in its definition.

3.2 Dunn's Index and Davies-Bouldin Index

The main goal of Dunn's index is to maximize inter-cluster distance while minimizing intra-cluster distance, and large values of Dunn's index correspond to good clusters. Dunn's index from clustering algorithms and random partition on the 6 expression datasets are shown in Fig. 2. Here, Dunn's curves generated by good clustering algorithms, HAL, K-means or SOMs, can be hardly distinguished from that of random partition, even at the number of classes. It's likely that only the minimum inter-cluster and maximum intra-cluster distances are considered in the definition of Dunn's index, resulting in the significant impact of outliers on the final performance [10]. Therefore, the nowadays Dunn's index should not be directly used in the validation of gene clustering.

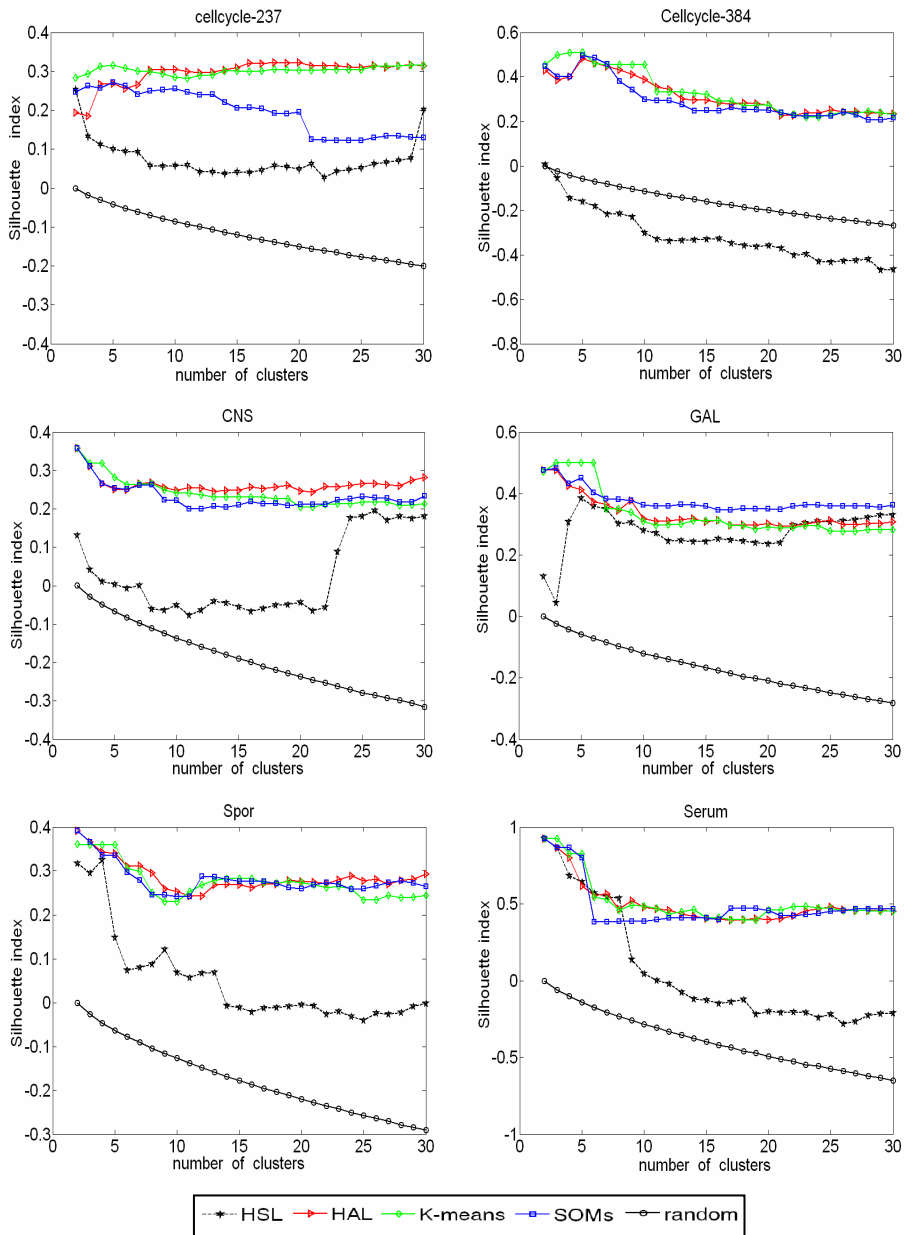


Fig. 1. Silhouette index of the clusters from *HSL*, *HAL*, *K-means* and *SOMs*, in contrast to that from random partition (*random*) on the six gene expression datasets

Table 1. Silhouette index of the clusters from HAL, K-means and SOMs on the 6 datasets. Italic entries represent the predicted number of clusters by the index, and the entries with asterisk indicate cases of correct estimate.

Datasets	Algorithms	K=2	K=3	K=4	K=5	K=6	K=7	K=8
Cellcycle_237 (4)	HAL	0.1944	0.1853	0.2667	<i>0.2688</i>	0.2564	0.2655	0.3047
	K-means	0.2837	0.2941	0.3122	<i>0.3156</i>	0.3084	0.3011	0.2998
	SOMs	0.2475	0.2629	0.2566	<i>0.2713</i>	0.2624	0.2417	0.2500
Cellcycle_384 (5)	HAL	0.2817	0.2542	0.2592	<i>0.3085</i> *	0.2969	0.2822	0.2745
	K-means	0.4554	0.4978	0.5083	<i>0.5125</i> *	0.4611	0.4546	0.4546
	SOMs	0.4479	0.4025	0.4009	<i>0.4961</i> *	0.4847	0.4582	0.3798
CNS (4)	HAL	<i>0.3583</i>	0.3111	0.2664	0.2510	0.2500	0.2666	0.2686
	K-means	<i>0.3581</i>	0.3181	0.3181	0.2822	0.2627	0.2646	0.2652
	SOMs	<i>0.3583</i>	0.3111	0.2664	0.2541	0.2512	0.2613	0.2633
GAL (4)	HAL	<i>0.4769</i>	0.4768	0.4251	0.4127	0.3736	0.3646	0.3445
	K-means	0.4717	<i>0.5005</i>	0.5004	0.5004	0.5004	0.3508	0.3508
	SOMs	0.4769	<i>0.4815</i>	0.4317	0.4505	0.4030	0.3832	0.3815
Serum (5)	HAL	<i>0.9233</i>	0.8679	0.8021	0.6162	0.5552	0.5660	0.4687
	K-means	<i>0.9233</i>	<i>0.9233</i>	0.8277	0.8277	0.5427	0.5309	0.4602
	SOMs	<i>0.9233</i>	0.8679	0.8679	0.8021	0.3841	0.3841	0.3862
Spor (6)	HAL	<i>0.3919</i>	0.3664	0.3429	0.3408	0.3104	0.3124	0.2952
	K-means	<i>0.3607</i>	0.3599	0.3590	0.3590	0.3062	0.2998	0.2500
	SOMs	<i>0.3919</i>	0.3664	0.3372	0.3350	0.2974	0.2794	0.2455

Davies-Bouldin index is based on geometrical considerations with the same basic rationale as Dunn's index, but defined as the ratio of the sum of within-cluster scatter to between-cluster separation [25]. Smaller values of Davies-Bouldin index indicate better clusters. Figure 3 illustrates the Davies-Bouldin index curves from the four clustering algorithms and random partition. Curves corresponding to the 4 clustering algorithms are remarkably distinguished from that to random partition, indicating the clear difference between their performances. At the same time, Curve from HSL is either comparative to or better than those from K-means, HAL and SOMs. Therefore, Davies-Bouldin index can give a much better validation for gene clustering than Dunn's index, but prefer to HSL. In the definition of Davies-Bouldin index, more information of the data was used, resulting in distinctly improved result. However, the "max" operation in its definition leads it susceptible to the partition of outliers, thereby preferring to elongate clusters. Similarly, the variation of Davies-Bouldin index from random partition indicates statistical bias from number of clusters, which would affect its ability for estimating the number of clusters too.

3.3 FOM

FOM measures the predicting strength of algorithms when clustering datasets into K clusters. At the same K, smaller FOM indicates better clustering result. FOM of the four clustering algorithms and random partition are illustrated in Fig. 4. In the figure, the difference between good clusters and random clusters, as well as that between HAL and HSL, are properly revealed. Therefore, FOM is effective to validate gene

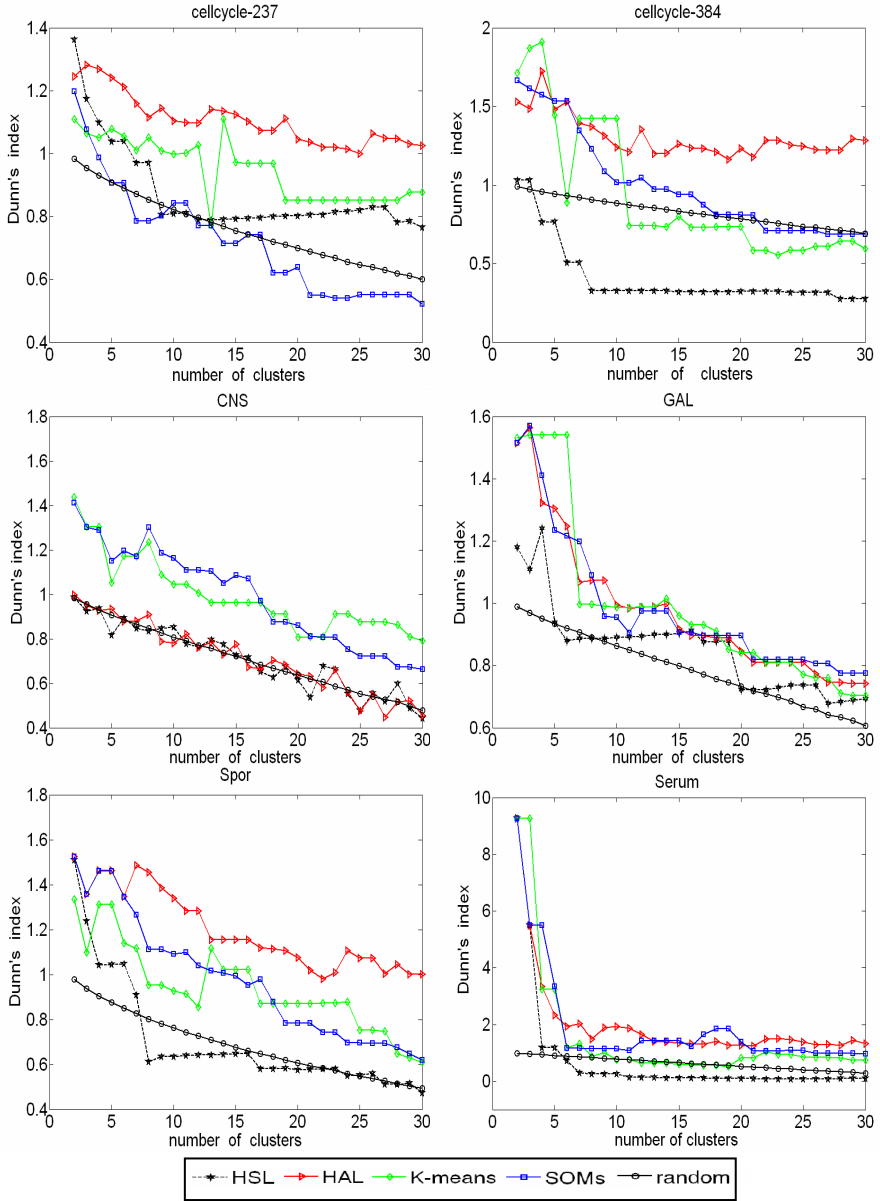


Fig. 2. Dunn's index of the clusters from *HSL*, *HAL*, *K-means* and *SOMs*, in contrast to that from random partition (*random*) on the six gene expression datasets

clustering results and able to reflect the different performance of clustering algorithms. As FOM estimates optimal number of clusters by steep decline in the curve, it's ambiguous in some degree.

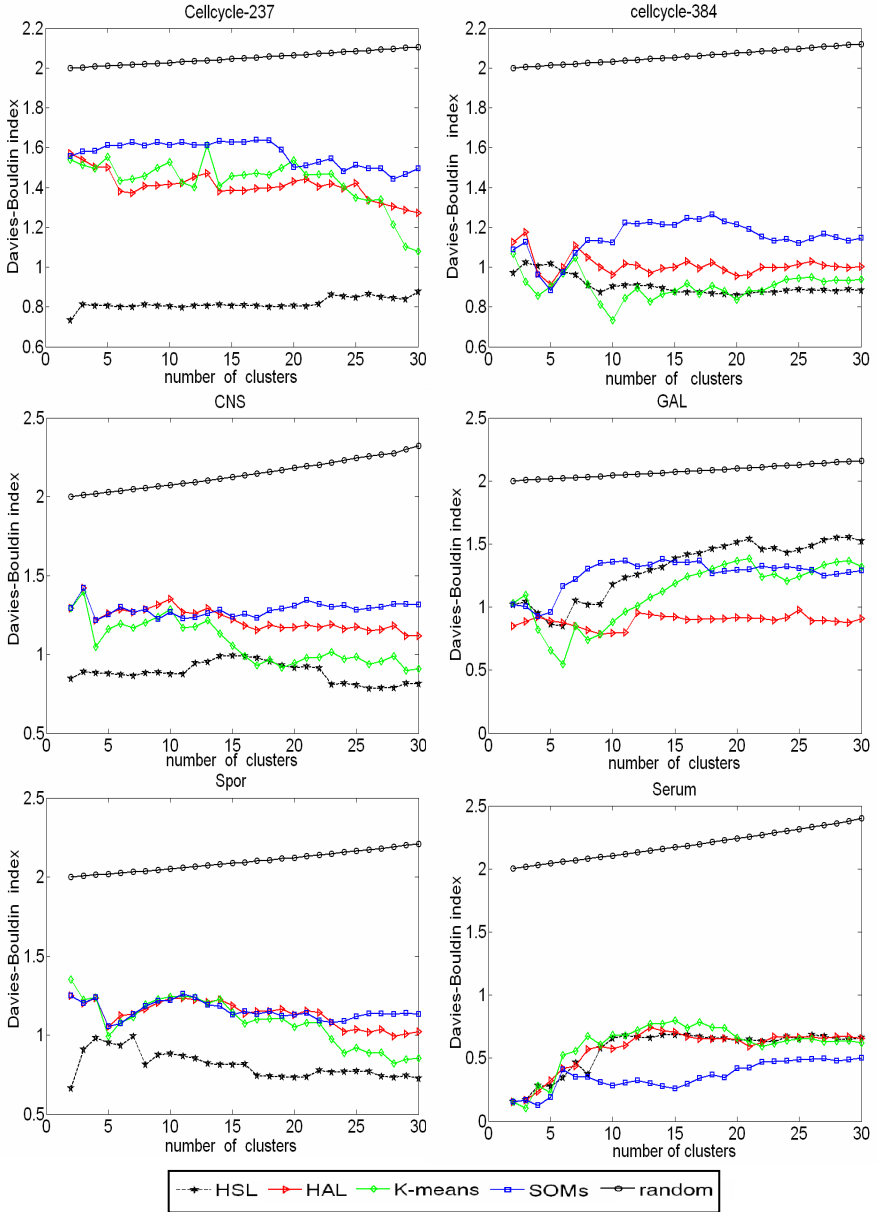


Fig. 3. Davies-Bouldin index of the clusters from *HSL*, *HAL*, *K-means* and *SOMs*, in contrast to that from random partition (*random*) on the six gene expression datasets

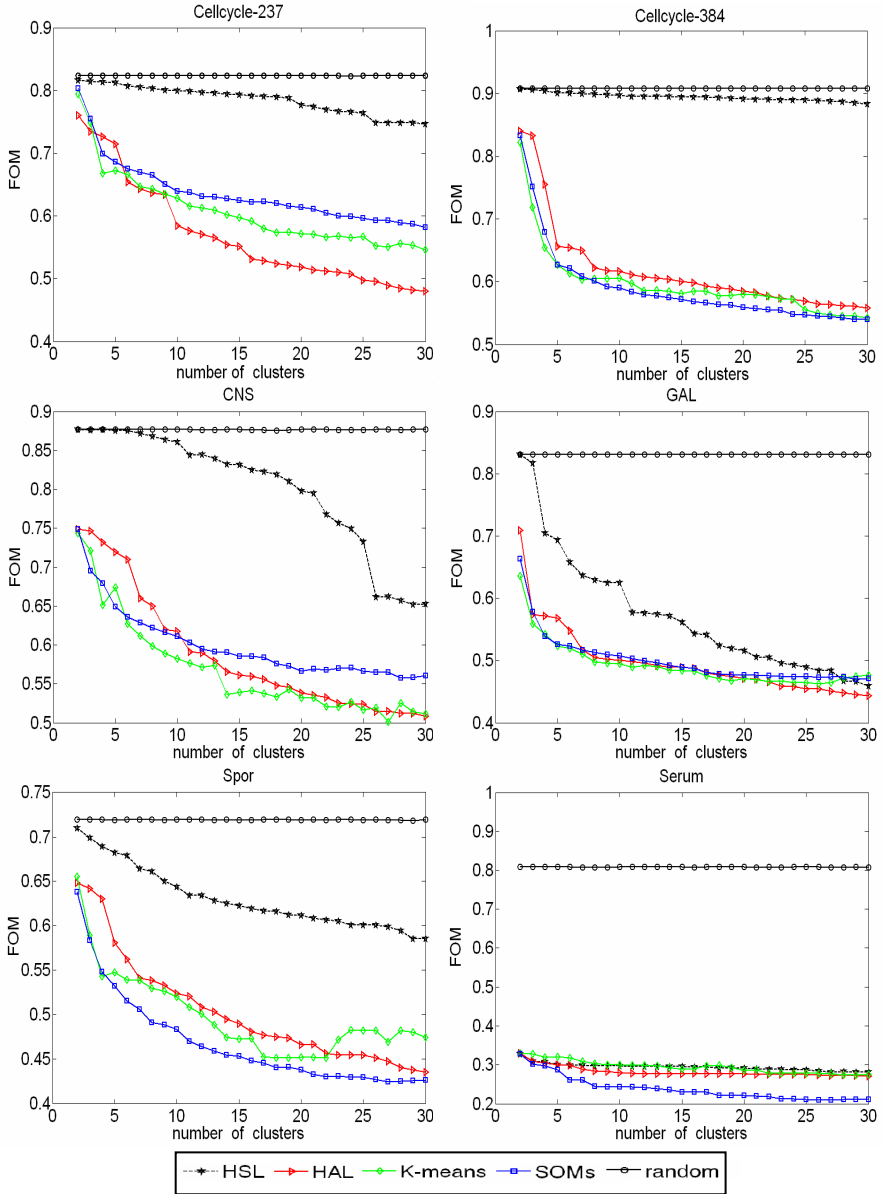


Fig. 4. FOM of the clusters from *HSL*, *HAL*, *K-means* and *SOMs*, in contrast to that from random partition (*random*) on six gene expression datasets

4 Conclusions

In this work, the ability of Silhouette index, Dunn's index, Davies-Bouldin index and FOM for the validation of gene clustering was investigated with public gene

expression datasets clustered by hierarchical single-linkage clustering, hierarchical average-linkage clustering, K-means and SOMs. It was made clear that Silhouette index and FOM can preferably reflect the performance of clustering algorithms and the quality of clustering results, Dunn's index should not be used directly in gene clustering validation for its high susceptibility to outliers, while Davies-Bouldin index can afford better validation than Dunn's index, exception for its preference to HSL. With regard to the application of these internal validation techniques in predicting the optimal number of clusters for gene clustering, a lot of further work should be done.

References

1. Jiang, D., Tang, C., Zhang, A.: Cluster analysis for gene expression data: a survey. *IEEE Transactions on Knowledge and Data Engineering*, 16 (2004) 1370-1386
2. Amir, B., Friedman, N., Yakhini, Z.: Class discovery in gene expression data. *RECOMB* (2001) 31-38
3. Quackenbush, J.: Computational analysis of microarray data. *Nat. Rev. Genet.* 2 (2001) 418-427.
4. Slonim, D.K.: From patterns to pathways: gene expression data analysis comes of age. *Nature Genetics*, 32 (2002) 502-508
5. Sherlock, G.: Analysis of large-scale gene expression data. *Current Opinion in Immunology*, 12 (2000) 201-205
6. Datta, S., Datta, S.: Comparisons and validation of statistical clustering techniques for microarray gene expression data. *Bioinformatics*, 19 (2003) 459-466
7. Yeung, K.Y., Haynor, D.R., Ruzzo, W.L.: Validating clustering for gene expression data. *Bioinformatics*, 17 (2001) 309-318.
8. Eisen, M.B., Spellman, P.T., Brown, P.O., et al.: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95 (1998) 14863-14868
9. Halkidi, M.: On clustering validation techniques. *J. Intell. Inform. Syst.* 17 (2001) 107-145.
10. Handl, J., Knowles, J., Kell, D.B.: Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21 (2005) 3201-3212
11. Bolshakova, N., Azuaje F.: Cluster validation techniques for genome expression data. *Signal Processing*, 83 (2003) 825-833
12. Ji, X.L., Li, L.J., Sun, Z.R.: Mining gene expression data using a novel approach based on hidden Markov models. *FEBS Letters*, 542 (2003) 125-131
13. Bolshakova, N., Azuaje, F.: Improving expression data mining through cluster validation. *Proc. of the 4th Annual IEEE conf. on Information Technology Application in Biomedicine* (2003) 19-22.
14. Chu, S., DeRisi, J., Eisen, M., et al.: The transcriptional program of sporulation in budding yeast. *Science*, 282 (1998) 699-705
15. Cho, R.J., Campbell, M.J., Winzeler, E.A., et al.: A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2 (1998) 65-73
16. Tavazoie, S., Huges, J.D., Campbell M.J., et al.: Systematic determination of genetic network architecture. *Nature Genetics*, 22 (1999) 281-285
17. Wen, X.L., Fuhrman, S., Michaels, G.S., et al.: Large-scale temporal gene expression mapping of central nervous system development. *Proc Natl Acad Sci USA*, 95 (1998) 334-339

18. Ideker, T., Thorsson, V., Ranish, J.A., et al.: Integrated genomic and proteomic analyses of a systemically perturbed metabolic network. *Science*, 292 (2001) 929-934
19. Yeung, K.Y., Medvedovic, M., Bumgarner, R.E.: Clustering gene expression data with repeated measurements. *Genome Biology*, 4 (2003) R34
20. Iyer, V.R., Eisen, M.B., Ross, D.T., et al.: The transcriptional program in the response of human fibroblasts to serum. *Science*, 283 (1999) 83-87
21. Xu, Y., Olman, V., Xu, D.: Clustering gene expression data using a graph-theoretic approach: an application of minimum spanning trees. *Bioinformatics*, 18 (2002) 536-545
22. Yang, C.M., Wan, B.K., Gao, X.F.: Selections of data preprocessing methods and similarity metrics for gene cluster analysis. *Progress in Nature Science*, 16 (2006) 607-713
23. Yang, C.M., Wan, B.K., Gao, X.F.: Data preprocessing in cluster analysis of gene expression. *Chin Phys Lett*, 20 (2003) 774-777
24. Rousseuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 (1987) 53-65
25. Bezdek, J.C., Nikhil, R.P.: Some new indexes of cluster validity. *IEEE Transactions on systems, man, and cybernetics*, 28 (1998) 301-315
26. Azuaje, F.: A cluster validity framework for genome expression data. *Bioinformatics*, 18 (2002) 319-320.

Intrinsic Splicing Profile of Human Genes Undergoing Simple Cassette Exon Events

Andigoni Malousi¹, Vassilis Koutkias¹, Sofia Kouidou², and Nicos Maglaveras¹

¹ Lab. of Medical Informatics, Faculty of Medicine, Aristotle University of Thessaloniki,
54124, P.O. Box 323, Thessaloniki, Greece
{andigoni,bikout,nicmag}@med.auth.gr

² Dept. of Biological Chemistry, Faculty of Medicine, Aristotle University of Thessaloniki,
54124, Thessaloniki, Greece
kouidou@auth.gr

Abstract. Alternative pre-mRNA splicing presides over protein diversity and organism complexity. Alternative splicing isoforms in human have been associated with specific developmental stages, tissue-specific expressions and disease-causing factors. In this study, we identified and analysed intrinsic features that discriminate non-conserved human genes that undergo a single internal cassette exon event from constitutively spliced exons. Context-based analysis revealed a guanine-rich track at the donor of the cassette's upstream intronic region that is absent in the constitutive dataset, as well as significant differences in the distribution of CpG and A3/G3 sequences between the alternative and the constitutive intronic regions. Interestingly, introns flanking cassette exons are larger than the constitutive ones, while exon lengths do not vary significantly. Splice sites flanking cassette exons are less identifiable, while splice sites at the outer ends are 'stronger' than constitutive introns. The results indicate that specific intrinsic features are linked with the inclusion/excision of internal exons which are indicative of the underlying selection rules.

Keywords: Alternative splicing, cassette exons, splice sites, intrinsic features.

1 Introduction

Pre-mRNA splicing in complex organisms is not merely a deterministic process. Alternative splicing (AS) is one of the most significant differentiation mechanisms of gene expression and protein synthesis that is estimated to occur in as many as 74% of the roughly 26.000 human genes, justifying the discrepancy between the unexpectedly low number of human genes and the abundance of protein domains (~90.000) [1], [2]. Recent studies have shown that tissue-specific protein functions and species segregation are predominantly caused by alternative gene expressions that are associated with either the binding of specific trans-factors or context-based features [3]. Experimental studies have also revealed a strong association of AS with certain developmental stages and the implication of genetic diseases [4], such as cancer [5].

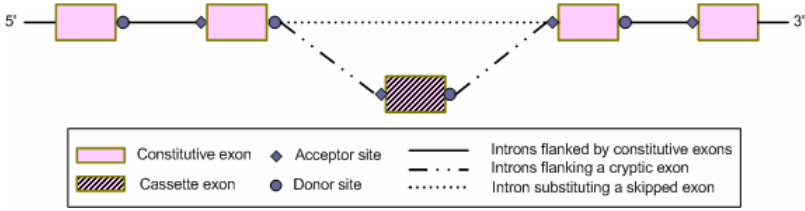


Fig. 1. Internal cassette exon event. A cassette exon is either expressed or spliced out from the AS isoforms.

Although AS is a common transcriptional event, still our knowledge does not suffice to assemble the puzzle of its underlying expression mechanisms [1]. What is presently known on AS is mainly a result of aligning sequences against Expressed Sequence Tags (ESTs) that is still subjected to the species EST coverage as well as to ambiguities introduced by the computational delineation of coding regions [6], [7].

An alternatively spliced gene can be edited in different coding forms generating multiple human protein domains with partially or completely different functionalities in excess (70%-88% [1]). Depending on the structural context, putative AS events are classified into cassette exon, intron retention, mutual exclusive exon, alternative 5', 3' splice sites (ss), alternative promoter, and initial/terminal exon events. Simple AS events involve a single type of alternative variants, while complex events result from combination of more than one AS event types. In human, 53% of all alternative exon events are expressed in form of cassette exons [8], which are either cryptic or skipped, i.e., are either expressed or spliced out from the alternative variant respectively (Fig. 1). There are two dominant evolution models of AS [3]. The first model suggests that AS events are in virtue of changes in the splice site composition, which make splice junctions more vulnerable in skipping events. The second, more elaborate model, argues that the splicing machinery is mediated by specific trans-acting factors, such as *SR* and *hnRNP* proteins, that bind specific sequence motifs and activate or inhibit AS events. In both models the role of compositional features of the DNA sequences is important.

The present study performs a thorough investigation of the context-based features that discriminate cassette from constitutively spliced exons and estimates the extent to which context bias influences the expression of 'weaker' exon candidates or suppresses the activation of 'stronger' exons. Since it is fairly more difficult to associate specific intrinsic features with the expression of cassette exon events when complex alternative events are observed, we compiled a basically stable dataset consisting of human genes in which only one exon is either skipped or cryptic with no modifications observed in the flanking regions. Given the genomic sequences describing genes matching these criteria, we evaluated the performance of intrinsic methods over the alternatively spliced exons and estimated the predictive power over splice sites flanking cassette exons. Furthermore, we classified genes into 5 subsets corresponding to the exonic and intronic sequences flanking constitutive and alternatively spliced exons. For each one of the resulting datasets we identified and

analysed specific compositional features at the proximity of the splicing tracks that differentiate alternatively from constitutively spliced genes.

2 Materials and Methods

The source datasets are built over the AltSplice database, a high quality dataset of transcript-confirmed splice patterns and AS events that is maintained by the ASD (Alternative Splicing Database) consortium. In principal, AltSplice data contain computationally delineated AS events that are basically generated by comparing EST/mRNA alignments with genomic sequences [9]. In the context of AltSplice, the human R2 release (April 2005, Ensembl 27.35a.1) is used which is a compilation of 16293 EST/mRNA confirmed genes that undergo 33338 exon events. 13799 out of the 18815 cassette exon events involve the expression of a simple cryptic/skipped exon with no other modifications observed in the flanking exonic regions, while 3147 genes are confirmed with a single alternative splice pattern.

2.1 Derivation of Constitutive and Alternative Intron/Exon Datasets

Since it is fairly more difficult to associate alternative expressions with specific factors when complex events are observed, we generated a dataset of human genes that appear with a single alternative isoform in which all exons are part of the pre-mRNA transcript except for an alternatively spliced exon that is either cryptic or skipped (cassette exon)¹. The final dataset is a compilation of human genes that match the following criteria: (1) Intron retention, mutual exclusive and complex splicing events were disqualified, since evaluation of the prediction accuracy is problematic in cases where overlapping structures are conditionally expressed. (2) Genes that are conserved between human and mouse are excluded so as to compare specific intrinsic features characterising non-conserved genes with those extracted from conserved genesets. (3) All initial and terminal exons are disqualified, since they are poorly modelled and difficult to identify [10]. (4) EST evidence of the AS events is used to distinguish dominant and minor splice forms and, in cases where an EST alignment results in internal ambiguities, the corresponding gene is excluded. (5) The compiled dataset is comprised of human genes that contain 20 exons at most.

The final data are classified into three datasets of 373 sequences each, corresponding to cassette exons, namely, CAED (CAssette Exon Dataset) and their flanking intronic regions, namely, UCID (Upstream Cassette Intron Dataset) and DCID (Downstream Cassette Intron Dataset). As control we built a dataset of 1466 constitutively spliced exons (COED, CONstitutive Exon Dataset) corresponding to internal exons of the same 373 genes that are expressed in all splicing variants. For the same purpose, we generated a dataset of 1039 sequences matching the intronic regions that are flanked by constitutively spliced exons (COID, CONstitutive Intron Dataset). The resulting 3 intronic and 2 exonic datasets were used for the analyses described below.

¹ Datasets are available upon request.

2.2 Analysis

To quantify the strength of the splice sites for each dataset we used two models introduced by Yeo et al. [11] and Shapiro and Senapathy (S&S) [12], respectively. The first model accounts for adjacent and non-adjacent dependencies between nucleotides of short motifs, corresponding to the donor's and acceptor's flanking tracks, and quantifies the likelihood of a candidate human splice site to be an actual one using a maximum entropy score model (MaxENT). Apart from the maximum entropy modelling, an inhomogeneous 1st-order Markov model (I1MM) and a position-specific weight matrix model (PWMM) were used to estimate the strength of the splice sites of the UCID, DCID and COID datasets implemented by *MaxEntScan*². The S&S score model was implemented for the same purposes. S&S splice model applies a scoring and ranking scheme over predefined species-specific nucleotide weight tables. S&S uses different scoring formulas for the donor and acceptor sites. In both cases the higher the score, the higher the likelihood of being an actual intron excision site. Finally, non-parametric chi-square tests were used to determine the statistical significance for various bivariate tabular analyses (significance level $p < 0.01$).

The construction and cleaning up of the datasets were performed based on appropriate Perl scripts. The performance of the intrinsic prediction methods was evaluated through appropriate Java and Perl wrappers. Perl was also used to parse data and to build regular expressions addressing various pattern matching problems. Gene prediction accuracy was evaluated at the exon level. Any predicted exon matching both reference exon ends is characterised as complete match. Likewise, exons matching either 5' or 3' ends are termed partial matches. ClustalW ungapped alignments [13] of the donor and acceptors 20-mers and position-based weight matrices gave closer insights of the base biases that may be associated with the splicing selection mechanism in both the constitutive and AS datasets. The resulting multiple alignments were visualised via the Jalview editor [14].

3 Results

3.1 Evaluation of Intrinsic Prediction Methods

A principal question when searching for discriminative patterns of alternatively spliced exons is whether these exons are equally predictable compared with the constitutive dataset. A significant step is thus to evaluate the performance of the currently available prediction techniques over the alternatively spliced dataset. Potential differences may be indicative of unknown selection mechanisms that conditionally act over splicing variants. Currently, no exhaustive evaluation on alternative exon datasets is available and the prediction of splicing variants is poorly addressed. Even so, intrinsic computational techniques can provide closer insights for the synthesis of high likely alternative variants [10], [15].

² <http://genes.mit.edu/burgelab/maxent/>

Evaluation of gene finders over constitutive datasets has shown highly accurate predictions at nucleotide and exon level; however, the overall prediction accuracy is still arguable at the gene level, especially when alternative exon expressions are considered [16]. This is justified by the fact that, even if all (constitutive and alternative) exons are correctly predicted, it is difficult to compile the constitutive and alternative gene assemblies from a pool of delineated exons. Although still unfeasible to build multiple splicing variants from a single DNA sequence, high likely suboptimal exons may indicate potential alternatively spliced exons [10]. The present study considers exclusively simple cassette exon events which involve a unique exon inclusion/excision, thus, an optimal gene parse extracted among candidate exons is not radically affected compared with genes undergoing complex splicing events. As a result, the correct delineation of an alternative exon in the optimal gene parse implies a unique splicing mechanism which, at least when the intrinsic characteristics are considered, is not differentiated between constitutive and alternative exons. In this case, the conditional activation of specific splice sites may be subjected to trans-acting elements, such as the binding of specific intronic/exonic splicing factors that regulate splicing by either suppressing (silencers) or increasing (enhancers) the transcription levels [3].

Table 1. Evaluation of the intrinsic methods: Sn: Internal exon sensitivity, i.e., proportion of the reference exons that are correctly predicted, cm: complete matches, pm: partial matches, TP: fraction of the true positive predictions

(a) Exon level accuracy of <i>GENSCAN</i>					
Dataset	Sn (%)	Optimal (%)	Suboptimal (%)	5' ss (%)	3' ss (%)
COED/cm	77.90	87.10	12.87	-	-
COED/pm	35.27	35.01	64.99	39.85	60.15
CAED/cm	45.58	90.59	9.41	-	-
CAED/pm	9.12	67.65	32.35	38.24	61.76

(b) Prediction accuracy of <i>GENESPLICER</i>					
Dataset	Correct predictions	TP av. score	Total predictions	Total av. score	TP(%)
DCID/donor	235	7.41	16374	4.63	63
UCID/acceptor	220	5.39	6885	3.35	59

In this study, we selected two highly accurate intrinsic methods, namely, *GENSCAN* [17] and *GENESPLICER* [18]. *GENSCAN* is a gene structure finder that outperforms other computational techniques and unlike most intrinsic methods it also identifies suboptimal exons that may be indicative of AS variants [16]. *GENESPLICER* is a standalone splice site predictor that combines several splicing prediction techniques and exhibits satisfactory results in terms of the overall predictive power and computational efficiency. Both tools implement probabilistic methods on species-specific gene and splice site models and extract high quality gene assemblies and splice sites respectively based on predefined thresholds. Table 1 summarises

the results obtained over both the constitutive and the cassette exons. The overall accuracy levels were estimated over internal exons. Initial and terminal exons are excluded due to ambiguities introduced by the EST evidence. It is evident that constitutively spliced exons are identified with significant higher sensitivity than cassette exons, when both complete (77.90%/45.58%) and partial (35.27%/9.12%) matches are encountered. However, while partially predicted constitutive exons are most frequently associated with lower a posteriori likelihood ($p > 0.1$), partially predicted cassette exons are mostly characterised as optimal (67.65% vs. 35.01%). In both constitutively and alternatively spliced exons *GENSCAN* exhibits significantly higher predictive power at the 3' end of the exon. The internal specificity of both constitutive and cassette exons over those predicted by *GENSCAN* is 76.60% (not shown).

GENESPLICER identifies short consensus sequences flanking donor/acceptor dinucleotides and assesses the probability of being an actual splice site using Maximal Dependence Decomposition (MDD) with Markov Models (MM). Predictions are made using the default sensitivity threshold against human splice models. *GENESPLICER* correctly identified 63% and 59% of the actual donor and acceptor sites respectively over the intronic datasets. The major weakness of the splice site predictors is the increased number of false positives caused by the degeneracy of the motifs describing splice junctions. For the set of potential splice sites identified by *GENESPLICER*, false positives are more frequent at the donor sites rather than the acceptors (98.6%, 68% respectively). The score is computed by the difference of the log-odds ratios between the score returned by the actual MM and the false MM.

3.2 Intron, Exon Length Distributions and Symmetry Estimation

Magen et al. [19] have reported an elevated frequency (68.6%) of conserved alternative exons that are symmetrical (divisible by 3) compared with the species-specific alternative exons that exhibit equal symmetry levels with the constitutive exons (~40%). Analysis of the species-specific cassette and constitutive exon lengths of the CAED and COED substantiates these findings, since in the present study symmetrical exons are observed with similar frequencies (41.82%, 41.06% respectively).

Several studies have also concluded in that conserved AS exons are shorter [20], [21]. In this non-conserved exon dataset the average size is slightly lower (149.68/153.26 nt). Interestingly, for symmetrical cassette exons the average length is 135.58 nt compared with the symmetrical constitutive exons (149.14 nt), the difference though is not statistically significant. A more significant difference is observed on the average length of the introns that contain cassette exons. The average intron length that contains an alternatively spliced exon is 11.62 kb, while the corresponding intronic sequences are significantly smaller in size (3.98 kb, $p=0$).

Constitutive intron lengths are clustered into smaller sizes compared to the intronic tracks flanking the cassette exons, while a considerable number of introns with cryptic/skipped exons are dispersed into longer sequences when cassette exons are not encountered (>30 kb). Introns of the UCID are slightly longer than introns of the DCID and, even when a cryptic exon is expressed, the average flanking intronic sequence length is longer than the average length of the constitutive intronic sequences.

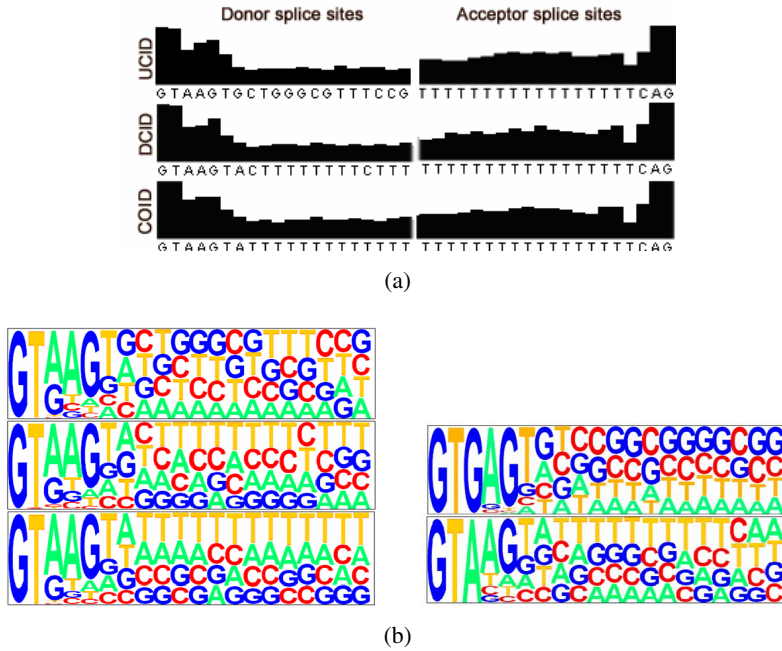


Fig. 2. (a) Consensus 20-mers extracted from ungapped multiple alignments of the donor/acceptor splice tracks. (b) Position-specific base distribution of the donor sites (*left-hand side*) contained in the UCID (*uppermost*), DCID and COID (*undermost*) and the UCID/G3, UCID/A3 sequences (*right-hand side*).

3.3 Context-Based Analysis

Intronic sequences adjacent to cassette exons are analysed with respect to the splicing rule followed. In both UCID and DCID, acceptor splice junctions are unexceptionally recognised by the AG dinucleotide. Donor sites though exhibit low frequency non-canonical GC dinucleotides which are more apparent at the downstream intronic sequences (2%) than the corresponding upstream donor sites (0.5%). All introns flanked by constitutive exons are consistent with the canonical GT-AG splicing form.

Fig. 2(a) illustrates the ungapped alignments of the donor and acceptors sites. Cassette consensus at both acceptor sites is consistent with the constitutive acceptor. Similarly, donor consensus of COID and DCID exhibits common positional characteristics. Interestingly, donor consensus of the UCID is differentiated from the corresponding constitutive consensus at the + 11 to +16 tag, with guanine being over-represented in this track. G-selection is also observed at the conserved +7 site. Position-specific weight matrices of the 20-mers at the 5' and 3' ends of the constitutive and alternative intronic tracks (UCID, DCID) are visualised through pictograms in Fig. 2(b). Furthermore, an activation of guanine residues at positions

+7..+16 relative to the beginning of the UCID donor site is observed, while a stable low guanine frequency characterises the constitutive donor splice junctions. A similar consensus to those matching constitutive splice sites is observed at the donor site downstream the cassette exons. These findings are consistent with the observations of McCullouch et al. [22], who have reported that in small introns of lower eukaryotes guanine triplets proximal to 5' splice sites cause preferential utilisation of the 5' splice site upstream of the triplets or the 3' splice sites downstream the guanine triplets. This observation may indicate potential correlation with splice site selection at both intron ends.

GT-AG splicing sites were further classified into A3 and G3 sequences depending on the nucleotide at the third position. This classification has been proposed by Clark et al. [23], as of being informative in the identification of AS events. A3 sequences appear with twofold frequency compared to the G3 sequences, totally corresponding to 95.23% in all datasets. Analysis of the intronic sequence lengths for each intronic dataset shows a statistical significant association with nucleotides at the third intronic position. The slightly longer A3 and G3 sequences of the UCID compared with the corresponding DCID introns are apparently linked with the increased average length of the UCID over the DCID introns.

We further analysed the succeeding 20-mers of the UCID/G3 and UCID/A3 sequences. The G/T ratio of the G3 and A3 sequences is 1.44 and 0.75 respectively. Specifically, analysis of the dimers succeeding the A3/G3 of the UCID confirms a strong bias between G3 sequences with AG dinucleotides at the +4,+5 positions (83.46%), which is less evident in the A3 sequences (46.70%). Similar findings are observed for the DCID and COID sequences (64.96%/43.16% and 82.27%/47.50% respectively). When the G3 sequences are encountered, the average number of guanine residues within the subsequent 9-mers is 2.29/2.88/2.20 in the COID, UCID and DCID respectively, while guanine is less evident in the A3 sequences (1.73/2.16/1.89). G3 intronic sequences are also more frequent in the UCID and this difference is statistically significant ($p=0.003$), contrary to the corresponding DCID where, despite the increased G3 sequences, the difference with the constitutive dataset does not constitute a statistically significant observation.

Acceptor sites are less divergent. A polypyrimidine-rich region (high frequency of Ts and Cs) at the proximal acceptor site is an indication of a polypyrimidine track that interacts with guanine at the 5' end of the intron during pre-RNA splicing. Position-specific analysis of the acceptor sites (Fig. 2(a)) gave no indication of potential alteration in the expression polypyrimidine track. However, polypyrimidine tracks may be sited at different positions and vary in length. PWMs calculate the base frequencies at each position and therefore do not suffice to capture this type of information. Therefore, in order to locate polypyrimidine tracks we identified potential branch points using the (C/T)T(A/G)A(C/T) consensus in an extended region of 100 nt upstream the acceptor sites. Branch point signals are identified in the 63.57% of the constitutive introns, 65.57% of the alternative introns proximal to a cassette exon at the 3' end and 66.30% of the alternative introns flanked by a cassette exon at the 5' end. These almost identical frequencies indicate that the branch point consensus is not differentiated between constitutive and cassette exons [24].

3.4 C+G Isoform, CpG Content of Cassette and Constitutive Exons

To investigate the degree of compositional similarities between the alternatively as well constitutively spliced exons and their flanking intronic regions, we first examined the C+G and CpG content (i.e., the frequency of either Cs or Gs monomers and the frequency of the CG dinucleotides respectively). Secondly, we classified intronic regions according to the A3/G3 type and re-estimated C+G content at the upstream, downstream and constitutive intronic regions. Table 2 summarises the average CpG and C+G content of the sequences of all datasets.

Constitutive exons are slightly more CpG-rich compared with the corresponding cassette exons. Although the C+G content between exon and introns does not vary significantly, CpGs are more frequent in exonic regions compared with the constitutive and alternative intronic regions. Interestingly, intronic regions flanking 5' end of cassette exons are more CpG-rich than the corresponding downstream intronic region that is consistent with the elevated G-rich tags at the donor site of the 5' ends intronic region shown in Fig. 2(a). G3 sequences of all datasets are more CpG-rich compared with the A3 datasets, however, the average C+G content is not radically changed due to the approximately twofold frequency of the A3 sequences.

Table 2. Frequency of CpG and C+G content. C+G content for A3/G3 sequences is estimated within oligomers downstream the donor sites (+7..+16).

Dataset	CpG(%)	C+G(%)	C+G/A3(%)	C+G/G3(%)
CAED	2.14	49.77	-	-
UCID	1.57	45.06	46.67	60.57
DCID	1.17	44.24	42.08	52.97
COED	2.20	48.30	-	-
COID	1.02	43.15	39.36	52.32

3.5 Splice Site Score Models

The strength of the splice sites is assessed using the splice site models introduced by Yeo et al. [11] (MaxENT, I1MM, PWMM) and Shapiro et al. [12] (S&S). The MaxENT, I1MM and PWMM use 9-mer sequences at positions -1..+6 for donor splice sites (GT at +1,+2 positions) and 23-mer sequences at positions -20..+3 for acceptor sites (AG at -2,-1 positions). Accordingly, the S&S splice models account for positional dependencies of the 9-mers at the same positions for the donor splice sites and the 15-mers at positions -14..+1 for acceptors sites (AG at -2,-1 positions). Table 3 summarises the average splice scores obtained for each dataset.

The average strength of the consensus sequences describing splice sites for each dataset is uniformly characterised by the incorporated models. Donor and acceptor consensus sequences that flank cassette exons (DCID donor, UCID acceptor) are 'weaker' than the splice sites of the constitutive intronic sequences and, consequently, less identifiable by the splicing machinery. This observation indicates a significant contribution of the intrinsic profile in the expression of the splice sites; nevertheless, the relative scores do not vary significantly. Interestingly, donor and acceptor sites at

the outer ends of the introns flanking cassette exons (UCID donor, DCID acceptor) are apparently ‘stronger’ than the corresponding consensus of the constitutive dataset. Furthermore, these highest scoring splice sites exhibit minor deviation levels. This observation is verified by all splice models and implies a multi-factorial association of the intrinsic features with the splice site selection mechanism that accounts dependencies among both neighbouring and distant nucleotides.

Table 3. Average splice site scores and standard deviations for each dataset

Dataset	MaxENT	IIMM	PWMM	S&S
UCID/donor	8.74±1.85	8.42±2.06	8.32±2.26	84.03±7.92
UCID/acceptor	7.72±2.89	8.17±2.98	8.70±3.87	86.32±7.27
DCID/donor	7.85±2.89	7.56±2.30	7.63±2.56	81.69±8.42
DCID/acceptor	9.01±2.35	9.42±2.51	10.13±3.35	88.24±6.26
COID/donor	8.47±2.45	8.04±2.38	8.15±2.45	82.83±8.64
COID/acceptor	7.99±3.18	8.50±3.31	8.91±3.98	86.63±7.49

4 Discussion

The molecular mechanisms that preside over the switching on/off of specific exons involve the combination of specific intrinsic and trans-acting factors that is still poorly characterised. In the present study, we identified intrinsic features that differentiate constitutive and cassette exons by performing two types of analysis. First, we evaluated the performance of computational techniques on alternative exon datasets and then we analysed specific compositional and structural features that are likely associated with the expression of the underlying splicing machinery.

What the assessment of the intrinsic prediction methods revealed is that cassette exons are not fully missed but are significantly less likely to be part of the optimal transcript. This observation indicates that the expression mechanisms are not differentiated; however, other factors either intrinsic or extrinsic are likely to enhance or suppress the activation of these exons. As a result, multiple weaker interactions can result in ‘stronger’ splice sites, while a strong splice site may be under-expressed when certain co-factors are present/absent. In addition, introns with high scoring splice sites may contain exons that are flanked by ‘weaker’ donor and acceptor consensus. In such case, cryptic exons are less identifiable by the splicing machinery and therefore more likely to be spliced out from the mature RNA.

A strong preservation of the reading frame in genes with confirmed exon inclusion/excision has been reported for genes that are conserved between human and mouse [19]. This observation suggests a propensity to preserve the structural features of the protein domains. In the present study, the analysis of non-conserved human genes indicates that species-specific alternatively spliced genes less frequently preserve the reading frame and most commonly induce a frameshift. This is consistent with the Magen et al. observations [19], who estimated an excess of the frameshift to 55-77% of the non-conserved genes, compared with the human-mouse orthologs which are more frequently symmetrical. Analysis of the splice site strength revealed that cassette exons border with ‘weaker’ splice sites compared to the constitutive.

This has also been reported by Itoh et al. [25], however, the insertion/excision of these sites is likely to be subjected to specific transactors that act as enhancers or silencers in the regulation of the mRNA splicing mechanism [3].

The presence of GT dinucleotides at the donor and the complimentary CA sequences at the acceptor site are probably associated with the splicing mechanism. GT sequences are known to assume flexible phosphate backbone conformation, which could facilitate the hinging of the DNA helix. CA together with the GT sequences could participate in a double-stranded sequence which determines the endpoints of the spliced area. It is also conceivable that multiple GTs at the UCID could contribute to the multiplicity of the spliced product. Finally, the elevated frequency of the C+G content succeeding G3 sequences of the UCID donor sites compared to the A3 sequences is probably an indication of the binding of specific CG-rich factors such as *SP1* proteins.

The results of the present study were evaluated over a stable dataset of non-conserved human genes that undergo a simple, internal cassette exon event. Complex cassette exons and other types of AS events were disqualified in order to extract certain intrinsic features that are associated with cassette exon events. Assuming a differentiated AS mechanism, these features are less likely to be observed in genes underlying other than cassette exons events. Still other sources of ambiguities, such as the presence of pseudogenes and paralogous genes are not efficiently addressed and remain difficult to be modelled via computational means. Finally, other factors associated with AS, such as specific tissue expression, pathogenic and developmental stage, are still ill-defined due to the lack of appropriate EST databases.

Conclusively, AS has been recently acknowledged as the principal mechanism underlying protein diversification [26]. Although of great importance, the prevalence of AS in complex organisms is yet under investigation and the implication of genetic diseases remains unclear. Computational delineation of alternatively spliced protein-coding exons and combined association with functional characteristics is expected to give closer insights on the underlying molecular mechanisms. In this context, the present study delineates specific intrinsic features that distinguish cassette from constitutive exons towards the construction of a highly accurate AS prediction schema.

References

1. Modrek, B., Lee, C.: A genomic view of alternative splicing. *Nat. Genet.*, 30 (2002) 13-19
2. Johnson, J.M., Castle, J., Garrett-Engele, P., Kan, Z., Loerch, P.M., Armour, C.D., Santos, R., Schadt, E.E., Stoughton, R., Shoemaker, D.D.: Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science*, 302 (2003) 2141-2144
3. Ast, G.: How did alternative splicing evolve? *Nat. Rev. Genet.*, 5 (2004) 773-782
4. Faustino, N.A., Cooper, T.A.: Pre-mRNA splicing and human disease. *Genes Dev.*, 17 (2003) 419-437
5. Venables, J.P.: Aberrant and alternative splicing in cancer. *Cancer Res.*, 64 (2004) 7647-7654
6. Modrek, B., Resch, A., Grasso, C., Lee, C.: Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, 29 (2001) 2850-2859

7. Graveley, B.R.: Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, 17 (2001) 100-107
8. Stamm, S., Zhu, J., Nakai, K., Stoilov, P., Stoss, O., Zhang, M.Q.: An alternative-exon database and its statistical analysis. *DNA Cell Biol.*, 19 (2000) 739-756
9. Stamm, S., Riethoven, J.J., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Tang, Y., Barbosa-Morais, N.L., Thanaraj, T.A.: ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Res.*, 34 (2006) D46-D55
10. Mathe, C., Sagot, M.F., Schiex, T., Rouze, P.: Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, 30 (2002) 4103-4117
11. Yeo, G., Burge, C.B.: Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, 11 (2004) 377-394
12. Shapiro, M.B., Senapathy, P.: RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Res.*, 15 (1987) 7155-7174
13. Thompson, J.D., Higgins, D.G., Gibson, T.J.: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22 (1994) 4673-4680
14. Clamp, M., Cuff, J., Searle, S.M., Barton, G.J.: The Jalview Java alignment editor. *Bioinformatics*, 20 (2004) 426-427
15. Foissac, S., Schiex, T.: Integrating alternative splicing detection into gene prediction. *BMC Bioinformatics*, 6 (2005) 25-34
16. Rogic, S., Mackworth, A.K., Ouellette, F.B.: Evaluation of gene-finding programs on mammalian sequences. *Genome Res.*, 11 (2001) 817-832
17. Burge, C., Karlin, S.: Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268 (1997) 78-94
18. Pertea, M., Lin, X., Salzberg, S.L.: GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.*, 29 (2001) 1185-1190
19. Magen, A., Ast, G.: The importance of being divisible by three in alternative splicing. *Nucleic Acids Res.*, 33 (2005) 5574-5582
20. Ladd, A.N., Cooper, T.A.: Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.*, 3 (2002) reviews0008.1-16
21. Sorek, R., Shamir, R., Ast, G.: How prevalent is functional alternative splicing in the human genome? *Trends Genet.*, 20 (2004) 68-71
22. McCullough, A.J., Berget, S.M.: G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. *Mol. Cell Biol.*, 17 (1997) 4562-4571
23. Clark, F., Thanaraj, T.A.: Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, 11 (2002) 451-464
24. Thanaraj, T.A., Stamm, S.: Prediction and statistical analysis of alternatively spliced exons. *Prog. Mol. Subcell. Biol.*, 31 (2003) 1-31
25. Itoh, H., Washio, T., Tomita, M.: Computational comparative analyses of alternative splicing regulation using full-length cDNA of various eukaryotes. *RNA*, 10 (2004) 1005-1018
26. Maniatis, T., Tasic, B.: Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, 418 (2002) 236-243

Generalization Rules for Binarized Descriptors

Jürgen Paetz

J.W. Goethe-Universität Frankfurt am Main,
60439 Frankfurt am Main, Germany

Abstract. Virtual screening of molecules is one of the hot topics in life science. Often, molecules are encoded by descriptors with numerical values as a basis for finding regions with a high enrichment of active molecules compared to non-active ones. In this contribution we demonstrate that a simpler binary version of a descriptor can be used for this task as well with similar classification performance, saving computational and memory resources. To generate binary valued rules for virtual screening, we used the GenIntersect algorithm that heuristically determines common properties of the binary descriptor vectors. The results are compared to the ones achieved with numerical rules of a neuro-fuzzy system.

1 Introduction

The process of virtual screening is defined as the selection of molecules with certain properties [1,2,3]. Usually, bioactivity is the main search criterion. Different techniques based on statistical, physical, or topological information of the molecules have been published [4,5,6,7]. The information is encoded in descriptor vectors where each descriptor encodes one property. Topological information of the descriptors can be two-, three-, or four-dimensional. 2D descriptors encode properties of the molecular graph. For 3D descriptors 3D coordinates of a molecule are utilized. Since a molecule might appear in different conformations, 4D descriptors consider these conformations. The higher the dimensions of the descriptor vectors are, the higher are the computational costs and the memory use.

At an early stage of virtual screening, where millions of molecules need to be screened for example, the 2D descriptors can be applied due to their lower computational costs. In our virtual screening example we used the CATS2D (Chemically Advanced Template Search for two-dimensional structures) descriptor [8]. The CATS2D descriptor vector is 150-dimensional, but it can be shortened by dimension reduction methods. To test different algorithms we used a library of 4705 ligands, that are active for different, actual drug targets [9]. Twelve drug targets are considered, such as ACE, COX2, and HIV. The ligands are called “active” when they are strong binders to the larger drug targets. A binary number uses only one bit while a numerical value can be a 32bit number for example, so that a binary representation reduces clearly the costs, and much more molecules can be screened.

Previous work was concerned with generating interval rules with high enrichment factors by using a neuro-fuzzy approach [10,11,12] that is specialized on handling numerical values, although there are heuristic extensions for symbolic data [13]. As we have found out recently, the CATS2D descriptor can be used in a binarized variant without significant performance loss in most cases when virtual screening is done by utilizing the Manhattan distance measure [14].

In this contribution we tried out the generation of binary generalization rules with GenIntersect [15] based on the binary version of the CATS2D descriptor. Our method GenIntersect considers intersections in a levelwise heuristic manner. We give a comparison of the best virtual screening rules for every drug target found by the neuro-fuzzy system with numerical values and by GenIntersect with binary values.

In the next section we repeat the ideas of the CATS2D descriptor and its binary variant and explain how we applied them to the benchmark database. In Section 3 we discuss the GenIntersect methodology for our purposes, followed by the results in Section 4. Additional pseudocodes are given in the appendix.

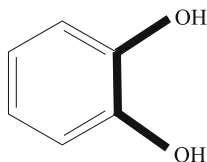


Fig. 1. A dd3 pair. OH is of type “d”.

2 The Binary Descriptor

The CATS2D descriptor encodes the 2D topological information of a molecular graph [8]. Therefore, types of atoms or atom groups are considered. The following types are available: hydrogen-bond donors (*d*), hydrogen-bond acceptors (*a*), positively charged atoms/groups (*p*), negatively charged atoms/groups (*n*), and lipophilic atoms/groups (*l*), respectively. The appearance of type pairs in the graph is counted and then normalized, so that numerical values result. Algorithm 1 in the appendix presents the encoding process in pseudocode. In Fig. 1 an example of a dd3 pair is depicted. The resulting 150-dimensional vector, using 15 type pairs and bond length $0, \dots, 9$, is noted in Formula (1):

$$\begin{aligned}
 &(dd0, da0, dp0, dn0, dl0, aa0, ap0, an0, al0, \\
 &\quad pp0, pn0, pl0, nn0, nl0, ll0; \dots; \\
 &dd9, da9, dp9, dn9, dl9, aa9, ap9, an9, al9, \\
 &\quad pp9, pn9, pl9, nn9, nl9, ll9)
 \end{aligned} \tag{1}$$

For calculation of the binary variant of descriptor vectors there are two possibilities: a) a posteriori set every numerical descriptor value > 0 to 1, and b) count only the first appearance of an atom type. Then, the counters in the descriptors need not to be incremented again and need no normalization. Only by using b) the computational costs and the allocated memory are reduced.

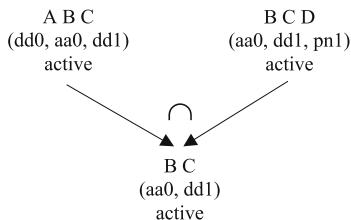


Fig. 2. Genralization of two itemsets to one itemset

3 Generalization Rules

In this section we present the main ideas behind generalization rule learning by intersections [15]. We adapt an example to the chemical application language. The items A, B, C , and D are given for example by $A = dd0$, $B = aa0$, $C = dd1$ and $D = pn1$. A set of items $\{A, B, C, \dots\}$, a so called itemset, is shortly noted as $ABC \dots$. Let $I_1 = ABC$ and $I_2 = BCD$. The common features of both itemsets are B and C , an $aa0$ and a $dd1$ group. The example is depicted in Fig. 2. The generation of intersections is a natural approach to generalization. Of course, only B and C alone would describe common properties of I_1 and I_2 , but the combination describes the common properties more exactly with regard to unknown itemsets (i.e. new molecules). Class labels can be added to the itemsets, e.g. $ABC \rightarrow$ “active”. Then, itemsets with the same class label are intersected only. The combinatorial explosion of such combinations might be a problem, but as we will see, heuristic approaches suffice for data analysis. Due to randomness, many intersections are composed of only a clearly smaller number of items than are present in the original itemsets. This is due to the fact that the intersection operator can be seen as a recombination operator, known from evolutionary algorithms [16]. The intersection process can be performed iteratively with the new intersected itemsets, e.g. with $I_3 = ABD$ we get BC , AB , and BD in the first iteration (“level 1”), and then B in the next iteration (“level 2”).

In the following we define necessary terms in the context of generalization with molecules. Since we are interested mainly in active molecules we give the definitions of the measures only for class “active”.

Definition 3.1

- a) An **item** is an elementary descriptor property of a molecule (one atom pair).
- b) An **itemset** is a set of items.
- c) The finite set of itemsets is noted as $\mathcal{IS}$.
- d) An intersection K of two sets I, J is called **nontrivial** if $K \neq I$, $K \neq J$ and $K \neq \emptyset$.
- e) A rule " $I \Rightarrow \text{active}$ " for active compounds is shortly called a **rule**.
- f) Here, the **frequency** $freq(I)$ of the itemset $I \in \mathcal{IS}$ or $freq(I \Rightarrow \text{active})$ of a rule $I \Rightarrow \text{active}$ is the proportion of the number of itemsets that contain the itemset I to the total number of itemsets of class "active", i.e.

$$freq(I \Rightarrow \text{active}) := \frac{\#\mathcal{IS}(I)}{\#\mathcal{IS}_{\text{active}}} . \quad (2)$$

- g) The **confidence** of a rule $I \Rightarrow \text{active}$ is defined as the number of rules of class "active" that contain I divided by the number of all itemsets that contain I :

$$conf(I \Rightarrow \text{active}) := \frac{\#\mathcal{IS}_{\text{active}}(I)}{\#\mathcal{IS}(I)} . \quad (3)$$

- h) The **enrichment factor** of a rule $I \Rightarrow \text{active}$ is defined as its confidence divided by the a priori probability p_{active} of the class "active" compounds:

$$ef(I \Rightarrow \text{active}) := \frac{conf(I \Rightarrow \text{active})}{p_{\text{active}}} . \quad (4)$$

The performance measures can be extended to more classes, but this is not required here. We are interested in finding rules in a heuristic manner with high enrichment factors on a test set. For analysis the data is divided in 50% training data and 50% test data. It uses the set $\mathcal{IS}$ of all training itemsets as a starting point for levelwise iteration of class "active" itemsets.

The generalization process is presented as pseudocode in Algorithm 2 in the appendix. Additionally to Algorithm 2 we use the heuristic of allowing only a maximum number of intersections in every level for every item when considering the full high dimensional descriptor space. A maximum number of levels is set, too. The complete and not heuristic generalization of a data set is only reliable if the data set is not too complex or if calculation time does not matter. We introduced measures, the so called generalization indices [17], that are based on the number of newly generated intersections within a sliding window. They are not used here, since we want to demonstrate the general ability of the algorithm to find rules for active molecules with high ef, but see the "Conclusions" section. In the following list, the main pro (P) and cons (C) are summarized [17]:

- (C) combinatorial explosion of generated rules is possible.
- (C) checking of already generated intersections in a previous level is costly.
- (P) heuristics could be used.
- (P) in one level more than one item of the rules could be removed.

- (P) robust rules with regard to unknown itemsets are generated.
- (P) no intersected itemsets with zero frequency are generated (what is the case with candidate sets for A-priori [18]).
- (P) missing values cause no technical problems in the algorithm.
- (P) Algorithm 2 can handle uncertain class labels, i.e. the same itemsets with different class labels, compared to candidate elimination [19].
- (P) optimality: without using heuristics there are no shorter rules that are more performant.

After the generation of rules it is useful to calculate the importance of the items. Items that are placed in rules with higher frequency and higher confidence are rated as being more important. Items in shorter rules are weighted with a higher factor. A formal definition is given in Def. 3.2.

Definition 3.2

Let $\{I_1 \Rightarrow \text{active}, \dots, I_r \Rightarrow \text{active}\}$ the set $\mathcal{R}$ of generated rules. Then, the **importance** (for class “active”) of an item A is defined as:

$$\text{imp}(A) := \sum_{i=1, A \in I_i}^r \text{freq}(I_i \Rightarrow \text{active}) \cdot \text{conf}(I_i \Rightarrow \text{active}) \cdot \frac{1}{|I_i|} . \quad (5)$$

4 Results

In the following the previous results are explained shortly for our purposes. Then, the new results are given together with a comparison.

4.1 Previous Results with a Neuro-fuzzy System

The Algorithm 2 is based on the generalization of symbolic properties. When considering the CATS2D descriptor the binarized values were used with this algorithm. Before, we had already performed analysis directly with the numerical values. Therefore, we used the neuro-fuzzy system [10]. Details about the system and the experiments can be found in [20]. The basic idea was to use hyper rectangular cuts of the hyper trapezoid activation functions of the neuro-fuzzy network. Only 0-cuts were considered, that are the hyper rectangles at the bottom of the hyper trapezoid. For such hyper rectangles R enrichment factors $ef(R)$ can be calculated analog to the definition given here for the generalization rules. The rules R have the format: “... **and if** var_j **in** (a_j, b_j) **and ... then** class ...”. Irrelevant attributes in R are those with infinite interval borders assigned “... **and if** var_j **in** $(-\infty, +\infty)$ **and ... then** class ...”. This property is used for calculating the importance of attributes of the dataset.

Definition 4.1 (cf. Def. 3.2)

Let $\mathbf{R}$ be the set of all generated rules. An attribute B is called *irrelevant* (irr.) for a rule $R \in \mathbf{R}$, if the corresponding interval borders are infinite. For

a given attribute A , let $\{R_1^A, \dots, R_r^A\} \subset \mathbf{R}$ be the subset of the r generated rule prototypes of class “active” where attribute A is not irrelevant. Let p_{active} be the a priori probability of the data of class “active”. Let conf be the confidence of a rule for class “active”. The **importance** (for class “active”) of an attribute A is defined as

$$\text{imp}_{\text{active}}(A) := \frac{1}{p_{\text{active}}} \cdot \sum_{i=1}^r \text{freq}(R_i^A) \cdot \text{conf}(R_i^A) \cdot f_{\text{irr},i,A} \quad (6)$$

with $f_{\text{irr},i,A} = \frac{1}{|\{B \mid B \text{ not irr. in } R_i^A\}|}$.

This definition could be adapted for class “inactive” as well. Due to the learning procedure the importance for both classes is significantly correlated and thus can be added to an overall importance value for every attribute. Note, that this is not the case in the analog Definition 3.2. The following abbreviations are used: ACE = angiotensin converting enzyme, COX2 = cyclooxygenase 2, CRF = corticotropin releasing factor (antagonists), DPP = dipeptidylpeptidase, GPCR = G-protein coupled receptor, HIV human immunodeficiency virus protease, HOR = hormone receptor, MMP = matrix metalloproteinase, NK = neurokinin receptor, PPAR = peroxisome proliferator-activated receptor, SEC = secretase, THROM = thrombin. The selected numbers of dimensions out of 150 were [20]: ACE (23), COX2 (20), CRF (18), DPP IV (24), GPCR (21), HIV (23), HOR (24), MMP (21), NK (19), PPAR (13), SEC (20), THROM (17). The numbers were chosen as about the half of the attributes with an importance greater than zero to perform a nontrivial dimension reduction. All measures were then calculated on the 50% test data. The best ef values are noted in Table 1. To avoid statistical outliers we considered only rules with a minimum of 0.8% frequency on the test data.

In mean the rules in Table 1 are composed of 6.8 (S.D. 2.5) attributes. The average frequency is 1.4% (S.D. 0.5%). The ef values are individual with respect to the drug target. Without dimension reduction the enrichment factor is only higher in the ACE case [20], cf. Table 1.

4.2 New Results with GenIntersect

The first experiment series with the GenIntersect was performed using all positive items (“property present in the molecule”) in the itemset for every molecule, so that a maximum of 150 items could be present in every itemset, but usually clearly less than the half were actually present. Here, the absence of properties was not coded, since it would led to itemsets of fixed length 150. For the initial experiment with all dimensions we used a maximum of two levels, and each itemset was allowed to generate three new itemsets. The results are presented in Table 2.

With Definition 3.2 the attributes with the highest importance were selected. Then, a new experiment series was settled for every drug target with a maximum of two levels, and each itemset was allowed to generate five new itemsets. We used the same reduced numbers of dimensions, but selected with the binary

Table 1. Enrichment factors (ef), frequency (freq), and number of numerical attributes of the best rule using the neuro-fuzzy system with 2352 test data entries. For GPCR data with one decimal precision. *: Without dimension reduction the ef, freq, attr. were 37, 1.0, 3 in the ACE case, respectively.

Drug target	ef (dim. red.)	freq [%]	attributes
ACE*	28	2.5	7
COX2	19	1.3	4
CRF	33	0.9	4
DPP IV	27	1.2	5
GPCR	2.6	1.6	8
HIV	31	1.1	8
HOR	18	2.2	5
MMP	52	0.9	4
NK	15	1.1	12
PPAR	14	1.6	9
SEC	34	0.9	9
THROM	19	0.9	7

Table 2. Enrichment factors (ef), frequency (freq), and number of binary attributes of the best rule using GenIntersect (Algorithm 2) with 2352 test data entries. For GPCR data with one decimal precision.

Drug target	ef	freq [%]	attributes
ACE	34	2.1	49
COX2	11	2.4	16
CRF	4	4.4	19
DPP IV	24	1.4	32
GPCR	2.9	1.1	36
HIV	26	0.8	38
HOR	9	3.5	41
MMP	58	1.0	31
NK	4	2.4	23
PPAR	11	1.6	42
SEC	10	1.4	27
THROM	20	1.4	44

importance measure. We omit the results here, since the number of attributes was too small in the binary case to achieve good results, compared to the number of attributes in Table 2. A clearly higher number of binary items is needed in the best rules, compared to the numerical features of the neuro-fuzzy interval rules. In mean the number of attributes is 33.2 (S.D. 10.4), what is about five times more attributes than before, but with a single bit representation only for each attribute. The frequency is 2.0% (S.D. 1.1%) on average. The bit representation even with more attributes reduces the memory requirements clearly compared

to a 16- or a 32-bit representation of numerical attributes ($33.2 \cdot 1 = 33.2$ and $6.8 \cdot 32 = 217.6$).

The efs were in most cases better with the neuro-fuzzy system, but in three cases the ef values were better in the binary case (GPCR, MMP, THROM). The required time for analysis was in mean clearly above 1h in the neuro-fuzzy case and clearly below 1h in the binary case (about one third of the time amount) when using strong heuristic restrictions as in our case.

The important attributes of the neuro-fuzzy analysis are not necessarily the same as in the binary case. In fact, the set of important numerical and binary attributes may differ clearly. A similar observation was made in [14]. As an example we give the rules in the MMP case and consider the most important attributes. – The best neuro-fuzzy interval rule ($ef = 52$, $freq = 0.9\%$) is:

if var 1 (dd0) ≥ 0.00 **and** var 10 (pp0) ≤ 0.50 **and** var 16 (dd1) ≥ 0.00 **and** var 96 (aa6) ≥ 0.06 **then** class “active”.

The best binary rule ($ef = 58$, $freq = 1.0\%$) as an itemset of the present CATS2D dimensions is:

{1, 2, 6, 15, 16, 17, 30, 32, 39, 45, 47, 51, 54, 60, 69, 77, 80, 81, 84, 90, 92, 95, 96, 99, 105, 110, 114, 120, 125, 129, 144} $\rightarrow$ “active”

Note, that in the latter rule the items 1, 16, and 96 are present, the item 10 is not. Hence, the rule goes together with the interpretation of the neuro-fuzzy rule, although the binary rule is more a kind of a “top-down” rule with more necessary items. The six most important numerical features for all neuro-fuzzy interval rules are in descending order: 10, 16, 13, 1, 46, and 96 (three of them present in the first rule). The six most important rules for all binary rules are (again in descending order): 15, 6, 69, 54, 30, 45 (all six present in the second rule). It is remarkable that the six most important dimensions differ, so that the MMP binary rules are composed of different features than the MMP numerical rules. When using both representations, binary and numerical, in the virtual screening the diversity of found active molecules could be increased, although this cannot be guaranteed for all drug targets. For example it is clear that the binary rule with $ef = 58$ and $freq = 1.0\%$ describe more active molecules compared to the numerical rule with $ef = 52$ and $freq = 0.9\%$. The same argument holds for the THROM case. Even the finding of a small number of additional active molecules is important because potentially powerful candidates for druglike compounds could be among them.

5 Conclusion

We introduced binary generalization rules for virtual screening and compared the results with numerical neuro-fuzzy interval rules. Both paradigms allow for virtual screening of molecules, but with different resulting rules. The numerical neuro-fuzzy rules are more precise, and they perform in most cases better, i.e. rules with higher ef values were obtained. The heuristic binary approach is faster and the bit representation requires less memory what is of importance when mil-

lions of molecules should be screened with the generated rules. Due to different search spaces with different important attributes, binary rule generation performed better in three out of twelve cases. It is assumed that both approaches will find different molecules, enhancing the overall diversity of active molecules, what is clear in the MMP and THROM cases. In other cases this follows from the different importance of the attributes. Since we used the binary generalization approach with strong heuristic restrictions, future work could be the usage of finer heuristics and the tuning of the generalization approach. Another task for the future will be the deeper exploration of common and different properties in binary and numerical search spaces.

Acknowledgement. The author thanks the Chair for Bio- and Cheminformatics at J.W. Goethe-Universität for let me have the chemical data for research.

References

1. Ajay, Predicting Drug-Likeness: Why and How?. *Current Topics in Medicinal Chemistry* **2**(12) 1273–1286, 2002.
2. Xu, H., Retrospect and Prospect of Virtual Screening in Drug Discovery. *Current Topics in Medicinal Chemistry* **2**(12) 1305–1320, 2002.
3. H.-J. Böhm and G. Schneider, *Virtual Screening for Bioactive Molecules*, Weinheim: Wiley VCH, 2000.
4. Lyne, P.D. Structure-Based Virtual Screening: An Overview. *Drug Discovery Today* **7**(20) 1047–1055, 2002.
5. Schneider, G., Böhm, H.-J. Virtual Screening and Fast Automated Docking Methods. *Drug Discovery Today* **7**(1) 64–70, 2002.
6. Borgelt, C., Berthold, M.R., Mining Molecular Fragments: Finding Relevant Substructures of Molecules. *Proc. of the 2nd IEEE Int. Conf. on Data Mining (ICDM)*, Maebashi City, Japan, 51–58, 2002.
7. Todeschini, T., Consonni, V., *Handbook of Molecular Descriptors*, Weinheim: Wiley-VCH, 2000.
8. Schneider, G., Neidhart, W., Giller, T., Schmid, G., Scaffold Hopping by Topological Pharmacophore Search: A Contribution to Virtual Screening, *Angewandte Chemie, International Edition* **38**(19) 2894–2895, 1999.
9. Schneider, P., Schneider, G., Collection of Bioactive Reference Compounds for Focused Library Design, *QSAR & Combinatorial Science* **22**, 713–718, 2003.
10. Huber, K.-P., Berthold, M.R., Building Precise Classifiers with Automatic Rule Extraction, *Proc. of the IEEE Int. Conf. on Neural Networks (ICNN)*, Perth, Western Australia, 1263–1268, Univ. of Western Australia, 1995.
11. Paetz, J., Metric Rule Generation with Septic Shock Patient Data, *Proc. of the 1st Int. Conf. on Data Mining (ICDM)*, San Jose, CA, USA, 637–638, 2001.
12. Paetz, J., Knowledge Based Approach to Septic Shock Patient Data Using a Neural Network with Trapezoidal Activation Functions, *Artificial Intelligence in Medicine* **28**(2) (Special Issue on Knowledge-Based Neurocomputing in Medicine), 207–230, 2003.
13. Berthold, M.R., Mixed Fuzzy Rule Formation, *International Journal of Approximate Reasoning* **32**, 67–84, 2003.

14. Fechner, U., Paetz, J., Schneider, G., Comparison of Three Holographic Fingerprint Descriptors and Their Binary Counterparts, *QSAR & Combinatorial Science* **24**, 961–967, 2005.
15. Paetz, J., Intersection Based Generalization Rules for the Analysis of Symbolic Septic Shock Patient Data, *Proc. of the 2nd IEEE Int. Conf. on Data Mining (ICDM)*, Maebashi City, Japan, 673–676, 2002.
16. Beyer, H.-G., An Alternative Explanation for the Manner in Which Genetic Algorithms Operate, *BioSystems* **41** 1–15, 1997.
17. Paetz, J., Durchschnittsbasierte Generalisierungsregeln Teil I: Grundlagen. Frankfurter Informatik-Berichte Nr. 1/02, Institut für Informatik, Fachbereich Biologie und Informatik, J.W. Goethe-Univ. Frankfurt am Main, Germany, ISSN 1616–9107, 2002.
18. Agrawal, R., Skrikant, R.: Fast Algorithms for Mining Association Rules, *Proc. of the 20th Int. Conf. on Very Large Databases (VLDB)*, Santiago de Chile, Chile, 487–499, 1994.
19. Mitchell, T.M., *Machine Learning*, New York: McGraw-Hill, 1997.
20. Paetz, J., Schneider, G., Virtual Screening Using Local Neuro-Fuzzy Rules, *Proc. of the 13th IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE)*, Budapest, Hungary, 861–866, 2004.

Appendix

Algorithm 1 (CATS2D)

```

for every atom  $A$  in a molecule  $M$  do
    assign no, one, or two group labels  $\in L := \{d, a, p, n, l\}$  to  $A$ ;
end
for every pair of atoms  $(A, B)$  do
    calculate distance  $d$  as shortest path of  $A$  to  $B$ 
    by bond counting  $\in K := \{0, 1, 2, \dots, 8, 9\}$ ;
    increment a counter for the appropriate pairs  $(x, y, z) \in L \times L \times K$ ;
end
for all  $15 \cdot 10 = 150$  counters do
    divide counter by the sum  $|x| + |y|$  of the
    occurrences of the atom types  $x, y$ ;
end

```

Algorithm 2 (GenIntersect)

Input parameters:

Set of itemsets $\mathcal{IS}$,

initial frequency list F (counts number of identical itemsets),

maxlevel (indicates, how many levels will be used for heuristic generalization),

and

thresholds γ_i (minimum thresholds for the performance measures of Def. 3.1)

Output parameters:

$\mathcal{IS}^F$ (set of generalized rules, including the initial rule itemset $\mathcal{IS}$),

level (indicates, how many levels effectively were needed, i.e. level $\leq$ maxlevel),

startindexlevel (itemsets with index startindexlevel(n) to startindexlevel($n + 1$) $- 1$ are generated in level n for $n \geq 1$), and the final (filtered) frequency list $F_{\text{final}}^{\geq}$.

1. initialization

$\mathcal{IS}_{\text{new}} := \mathcal{IS}$;

level := 1;

startlevel(level) := 1;

endofalg := false;

2. generate intersections in the levels

while endofalg = false **do**

startlevel(level+1) := $\#(\mathcal{IS}_{\text{new}}) + 1$;

oldIS := $\# \mathcal{IS}_{\text{new}}$;

% pass through actual level

for i = startlevel(level) **to** startlevel(level+1)-2

% pass through itemsets *without* considering itemsets

% of the preceding levels

for j = i+1 **to** startlevel(level+1)-1

Inter := $\mathcal{IS}_{\text{new}}(i) \cap \mathcal{IS}_{\text{new}}(j)$;

if Inter is a nontrivial intersection

and (Inter $\notin \mathcal{IS}_{\text{new}}$) **then**

$\mathcal{IS}_{\text{new}}(\# \mathcal{IS}_{\text{new}} + 1) := \text{Inter}$;

end

end

end

3. check termination

if ($\# \mathcal{IS}_{\text{new}} = \text{oldIS}$) **or** (level $\geq \text{maxlevel}$) **then**

endofalg := true;

$\mathcal{IS}^F := \mathcal{IS}_{\text{new}}$;

else

level := level + 1;

end

end while

4. determine the final frequency list F_{final} by calculating the frequency for all new rules and expand F . Calculate the confidence, too.

5. filter all rules that have performance measures higher then all γ_i .

Application of Combining Classifiers Using Dynamic Weights to the Protein Secondary Structure Prediction - Comparative Analysis of Fusion Methods

Tomasz Wołoszynski and Marek Kurzynski

Wroclaw University of Technology, Faculty of Electronics, Chair of Systems and Computer Networks, Wyb. Wyspińskiego 27, 50-370 Wroclaw, Poland

tomasz.wołoszynski@pwr.wroc.pl,

marek.kurzynski@pwr.wroc.pl

Abstract. We introduce common framework for classifiers fusion methods using dynamic weights in decision making process. Both weighted average combiners with dynamic weights and combiners which dynamically estimate local competence are considered. Few algorithms presented in the literature are shown in accordance with our model. In addition we propose two new methods for combining classifiers. The problem of protein secondary structure prediction was selected as a benchmark test. Experiments were carried out on previously prepared dataset of non-homologous proteins for fusion algorithms comparison. The results have proved that developed framework generalizes dynamic weighting approaches and should be further investigated.

1 Introduction

Information fusion has been investigated with much attention in recent years. The idea of using ensemble of classifiers instead of single one proved to be useful, assuring higher classification accuracies in many pattern recognition problems. In general, combining methods may be divided into two groups: classifier fusion and classifier selection. The first one assumes that the final decision should be made using all classifiers outputs. The latter chooses single classifier with the highest local competence and relies only on its supports. In Sect. 2 we present common framework for both weighted average combiners with dynamic weights (WAD) and combiners which estimate local competence dynamically (LCE) [17]. Described approaches make sense in problems where similarity between objects can be measured. Although continuous character of input features seems to be a good criterion for selecting a specific task, it may be very interesting to examine performance of introduced fusion methods elsewhere. The protein secondary structure prediction, being one of the most important challenges in computational biology provides us with such testing data. Two main differences between classical pattern recognition problem and predicting three-dimensional conformation of a protein making the task more demanding are: variable length of

input object and computation of distance between two proteins using evolutionary matrix. The idea of using classifier fusion in protein secondary structure prediction has already been put into practice by researchers [10] giving slightly better scores than the individual methods. The reason why combining approach becomes more popular is mainly due to existence of many basic prediction algorithms which can be treated as base classifiers and incessantly growing size of protein databases. In Sect. 3 we describe previously prepared protein dataset and discuss the results of benchmark tests performed for proposed combining algorithms and few other fusion methods for comparison. Conclusions for presented combiners are given afterwards in Sect. 4.

2 Combining Methods

2.1 WAD and LCE Combiners

We are given the ensemble of N base classifiers, each of them producing a row vector with supports for M classes. All of the support vectors form a decision profile matrix $DP(x)$ [16,17] for any input object x :

$$DP(x) = \begin{bmatrix} d_{1,1}(x) & \cdots & d_{1,M}(x) \\ \vdots & & \vdots \\ d_{N,1}(x) & \cdots & d_{N,M}(x) \end{bmatrix}, \quad (1)$$

where $d_{n,m}(x)$ denotes support of n -th classifier for m -th class for object x . Without loss of generality we can restrict x within the interval $[0, 1]$ and additionally $\sum_{m=1}^M d_{n,m}(x) = 1$. We assume that weights $w_{n,m}(x)$ ($n = 1, \dots, N$, $m = 1, \dots, M$) used both by WAD and LCE combiners in fusion procedure depend on the input object x and form matrix $W(x)$. For a WAD combiner final support for class m is given by weighted sum of supports of base classifiers, viz.

$$\mu_m(x) = \sum_{n=1}^N w_{n,m}(x) d_{n,m}(x). \quad (2)$$

In the case of LCE combiner this support is equal to the support of base classifier with the greatest local (at point x) competence. As a competence measure we adopt the sum of classifier weights, which leads to the following final support formula:

$$\mu_m(x) = d_{n,m}(x), \text{ where } \sum_{m=1}^M w_{n,m}(x) = \max_k \sum_{m=1}^M w_{k,m}(x). \quad (3)$$

The class with the highest final support is assigned to the input object x .

2.2 Framework for Combiners Using Dynamic Weights

Let us assume that the feature space is divided into K disjoint regions R_k . Suppose that E^* and E_k^* are fusion methods with the best possible static

(independent of x) weight matrix for the whole feature space and for region k , respectively. The following inequality holds:

$$\forall_{k=1,\dots,K} P_C(E_k^*|R_k) \geq P_C(E^*|R_k) , \quad (4)$$

where $P_C(E|R_k)$ denotes probability of correct classification of the ensemble E under condition that object x lies in region R_k . It is clear that the feature space division provides us with better classification accuracy:

$$P_C(fusion) = \sum_{k=1}^K P_C(E_k^*|R_k) P(R_k) \geq P_C(E^*) , \quad (5)$$

where $P(R_k)$ denotes probability that input object x lies in region R_k . If we split feature space into infinite number of regions so that each of them shrinks to a single point we get:

$$P_C(fusion) = \int P_C(E_x^*|x) f(x) dx , \quad (6)$$

where $f(x)$ denotes probability density function of features and $P_C(E_x^*|x)$ is probability of correct classification of the best possible ensemble at point x . The latter takes value within the range $[0, 1]$. Therefore, in order to maximize the probability (6) it is sufficient to maximize just conditional probability $P_C(E_x^*|x)$ for any given x :

$$\max_E P_C(fusion) \equiv \max_E P_C(E_x^*|x) . \quad (7)$$

This approach can be used only with objects x_l for which class memberships i_l are known. We denote such learning set by $S_L = \{(x_l, i_l), l = 1, \dots, L\}$ and its cardinality by L . The classification algorithm based on proposed framework becomes completely determined by the way of creating the weight matrix $W(x_l)$ for any object x_l from the set S_L . For any other object x we suggest finding the weight matrix $W(x)$ by following equation:

$$W(x) = \sum_{l=1}^L g(x, x_l) W(x_l) , \quad (8)$$

where $g(x, x_l)$ is a function dependent on the distance $d(x, x_l)$ between objects x and x_l . The way of creating matrices $W(x_l)$ as well as defining function $g(x, x_l)$ are parameters of introduced algorithm. We have adapted two LCE methods proposed in the literature to framework described above: distance-based k -nn [8] and potential functions [17]. Additionally we have created two new algorithms using presented model. All of them were tested using two types of distance dependent function: $g_1(x, x_l) = \frac{1}{d(x, x_l)}$ and $g_2(x, x_l) = \frac{1}{1+(d(x, x_l))^2}$. Description of mentioned methods is shown in Table 1.

Table 1. Tested algorithms presented in accordance with proposed framework

Tested combiners	The way of creating weight matrix $W(x)$, $r = i_l$	Distance dependent function
CC1 (Distance- based k -nn)	$w_{n,r}(x_l) = d_{n,r}(x_l)$	$g_1(x, x_l)$
CC2 (Potential functions)	$w_{n,r}(x_l) = \begin{cases} 1 & \text{if } d_{n,r}(x_l) = \max_j d_{n,j}(x_l) \\ -1 & \text{otherwise} \end{cases}$	$g_2(x, x_l)$
CC3	$w_{n,r}(x_l) = \begin{cases} d_{n,r}(x_l) & \text{if } d_{n,r}(x_l) = \max_j d_{n,j}(x_l) \\ -\max_j d_{n,j}(x_l) & \text{otherwise} \end{cases}$	$g_1(x, x_l)$
CC4	$w_{n,r}(x_l) = \begin{cases} d_{n,r}(x_l) & \text{if } d_{n,r}(x_l) = \max_j d_{n,j}(x_l) \\ -\max_j d_{n,j}(x_l) & \text{otherwise} \end{cases}$	$g_2(x, x_l)$
CC5	$w_{n,r}(x_l) = \begin{cases} d_{n,r}(x_l) & \text{if } d_{n,r}(x_l) = \max_j d_{n,j}(x_l) \\ 0 & \text{otherwise} \end{cases}$	$g_1(x, x_l)$
CC6	$w_{n,r}(x_l) = \begin{cases} d_{n,r}(x_l) & \text{if } d_{n,r}(x_l) = \max_j d_{n,j}(x_l) \\ 0 & \text{otherwise} \end{cases}$	$g_2(x, x_l)$

3 Application to Protein Secondary Structure Prediction

3.1 Introduction to Protein Prediction

The problem of secondary structure prediction for a given protein is of great importance in the field of drug designing. Current measuring methods providing three-dimensional protein structures i.e. X-ray crystallography using diffraction images or NMR are based on expensive and long processes, therefore computational techniques are used to overcome these disadvantages. Because most of the properties (functions) of proteins are closely related to their 3D shapes it is essential to determine the secondary structures of amino acid sequence for its further study.

A protein is a biomolecule constructed from amino acid units. A protein chain represented by a string of amino acid units called protein primary structure is encoded as a sequence of characters over the alphabet $\{C, S, T, P, A, G, N, D, E, Q, H, R, K, M, I, L, V, P, Y, W, -, X\}$. The symbols C to W represent 20 amino acids, the symbol $-$ refers to a gap in the alignment as the result of an insertion or deletion of an amino acid residue over evolutionary history. The symbol X indicates an undefined amino acid and can occasionally be encountered in protein sequence data. The sequence length (number of residues in the chain) depends on the protein. Amino acid sequence forms local conformation called protein secondary structure. Eight forms of secondary structure assignment were developed in 1983 [3] and are widely used today, but secondary structure prediction is often limited to three of those: helical or α -helix (encoded by letter H), extended or β -sheets (E) and loops also called reverse turn or coil (C). α -helices are strengthened by hydrogen bonds between every fourth amino acid so that the protein backbone adopts a helical configuration. In β -sheets the hydrogen bonding is non-local. They adopt a parallel or anti-parallel sheet

configuration. Other structural elements such as bends and turns are classified as loops. In the problem of protein secondary structure prediction, which can be considered as an classification task, the inputs are the amino acid sequences while the output is the predicted structure also called conformation which is the combination of α -helix, β -sheets and loops states. An example of typical protein sequence and its conformation class is shown in Table 2. Since experimental

Table 2. The example of primary and secondary structure of protein 16PK

Protein codename	16PK														
# of amino acid	1	2	3	4	5	6	7	...	409	410	411	412	413	414	415
Primary structure	E	K	K	S	I	N	E	...	V	T	V	L	D	D	K
Secondary structure (conformation class)	C	E	C	E	H	H	H	...	H	H	C	C	C	E	C

evidence has shown that the conformation of proteins is determined predominantly by their amino acid sequence [3,23], many attempts have been made to predict protein structures from their primary sequence. In the past thirty years, scientists have made great efforts to solve this problem. A huge number of algorithms had been designed to predict protein secondary structure. These algorithms were based on different approaches and achieved different accuracies. In particular machine learning techniques and pattern recognition methods have been actively investigated including neural networks [1,24], decision trees [18], Hidden Markov Models [2], Support Vector Machines [20], data mining [21], to name only a few. Although over the years the quality of protein secondary structure prediction has been incessantly improved, it is still not satisfactory. The current algorithmic methods give the predictive accuracies of protein secondary structure in the range of 65% to 80% [2,24]. In the next subsection we present results of applications of proposed combining classifiers methods, which give the top performance of protein prediction in contrast with those published in literature. The new methods reached an accuracy above 78% which is very close to the best accuracy achieved by single methods.

3.2 Experiments and Results

We have derived non-homologous protein dataset from PDBSELECT [12] with 25% similarity threshold. The total number of 583 proteins with 49322 residues were selected for the experiment. Only proteins with at most 150 amino acids in the sequence and fully described by 20 letter alphabet (no X and – symbols) were taken into account. PDBFINDER2 [15] was used for finding DSSP [14] predictions (class memberships). We have reduced number of classes on DSSP output to previously described three (H, E, C). The ensemble of base classifiers is built of 3 different methods: GOR IV [6] (based on information theory and Bayes method), HNN [9] (hierarchical neural network) and SOPMA [7] (based on multiple alignments). The results were gathered using NPS@ server [5]. Each

of classifiers gives three degrees of support for every amino acid in the chain. The outputs were processed using modified *softmax* method [4] ensuring that they always sum to one. Hence, the outputs can be interpreted as the conditional probabilities that a given input belongs to each of the three classes.

In protein sequence analysis we need to know functional or structural relationship between two sequences [13]. We are mainly interested in how similar they are. The distance between a pair of amino acids from different proteins was computed using BLOSUM 30 [11] scoring matrix A with window size of 11. Each entry a_{ij} of matrix A represents the „normalized probability” (score) that amino acid i can mutate into amino j during evolution. Calculations of distances were done in following manner. For every possible pair of amino acids (put in the middle of 11 sized window) from different proteins an evolutionary scores were checked in the BLOSUM 30 matrix. Next the best possible match (therefore a match with itself) for analysed amino acid was computed using the same scoring matrix. The quotient of two described values was used as a normalized distance measure. Finally ten most similar amino acids were taken into account during computation process.

To measure the accuracy of a prediction some parameters have been defined in the literature [19]. The most widely used accuracy index for secondary structure prediction is the three-state per-residue accuracy (Q_3) for the whole class set and Q_H , Q_E and Q_C for α -helix, β -sheets and coil, namely:

$$Q_3 = \frac{P_H + P_E + P_C}{T}, \quad Q_H = \frac{P_H}{T_H}, \quad Q_E = \frac{P_E}{T_E}, \quad Q_C = \frac{P_C}{T_C} . \quad (9)$$

$P_H(T_H)$, $P_E(T_E)$ and $P_C(T_C)$ are the number of residues predicted correctly (number of residues in total) in state α -helix, β -sheets and coil, respectively, while T is the total number of all residues. Although this kind of measure is very common in pattern recognition problems it may be misleading when dealing with protein secondary structure prediction. Segment overlap rate SOV [23] was developed specially for this task and is much more competent. The per-segment accuracy SOV measures the percentage of segments of secondary structure correctly predicted, where a segment is a continuous set of amino acids. Therefore it gives a prediction score with respect to the structural level instead of taking into account only single residues. The approach ensures that the general secondary structure shape of a protein is the most important factor.

Prediction accuracies for base classifiers and simple combining methods such as max, mean and majority voting [16] for the whole dataset are presented in Table 3. First of all it should be stated that all base classifiers give quite distinct predictions. Still SOPMA method seems to be the best one among others. Combiners such as max and mean are always less accurate than the best single classifier but in overall they are superior to both GOR IV and HNN. Majority voting combiner gets the highest score for predicting class C and is almost as good as SOPMA algorithm for total conformational state prediction. The results given for oracle classifier are very interesting and meaningful. It is on average at least 10 percentage points better than any of other methods. This proves that there is much space for improvement for combining algorithms. Tests of methods

described in Sect. 2 were carried out with ten-fold cross validation. Results are shown in Table 4. The CC1 fusion approach is inferior to all other combiners. This fact may be caused by the way of computing weight matrix where no penalty policy was applied. Similar situation can be seen for CC5 and CC6 algorithms despite the best accuracies for class C. The latter two were tested using WAD final supports (2). The best three combiners are CC2, CC3 and CC4. Each of them is better than majority voting method in all classes, but surprisingly the overall scores are lower. Nonetheless they assure good performance and all were examined using LCE approach (3), which is worth mentioning.

Table 3. Prediction accuracies for base classifiers and selected combiners (in per cent)

	Q _H	Q _E	Q _C	Q ₃	SOV _H	SOV _E	SOV _C	SOV ₃
GOR IV	59.26	60.78	65.83	60.43	67.26	67.37	71.62	72.02
HNN	68.09	56.45	74.48	66.65	73.85	60.66	79.23	75.77
SOPMA	72.85	64.35	68.04	68.73	78.23	71.12	76.27	79.07
MAX	65.34	59.87	72.32	65.53	72.15	65.73	78.50	76.10
MEAN	65.53	59.78	74.39	66.65	72.54	65.71	78.86	76.72
VOTE	68.65	60.41	75.15	68.60	75.06	65.84	80.70	78.36
ORACLE	73.06	68.10	77.95	82.30	86.63	80.75	91.24	89.40

Table 4. Prediction accuracies for proposed fusion methods and two combiners adapted to presented framework (in per cent)

	Q _H	Q _E	Q _C	Q ₃	SOV _H	SOV _E	SOV _C	SOV ₃
CC1 (Distance-based k -nn)	65.84	58.12	74.60	66.06	73.33	64.96	80.64	76.44
CC2 (Potential functions)	68.18	60.55	74.31	67.79	76.31	68.18	82.09	78.45
CC3	67.84	60.30	74.50	67.67	76.09	67.80	82.01	78.16
CC4	67.83	60.29	74.50	67.67	76.07	67.82	82.02	78.15
CC5	65.17	56.17	78.10	66.81	73.40	64.99	82.39	76.82
CC6	65.18	56.19	78.10	66.80	73.37	65.04	82.39	76.79

4 Conclusions

We have introduced a framework for combining classifiers based on dynamic weights. Two parameters in our model: distance dependent function and weight matrix allow us to modify the fusion process in many ways. The generalization ability of developed algorithm was proven by adapting existing LCE combiners in accordance with our approach. Future investigation should be focused on selecting the most proper parameters for a given problem. Improvement of method proposed for computing the weight matrix for particular input object x

would also be desirable. The accuracies gained during experiments for protein secondary structure prediction are satisfactory in comparison to other types of combiners. However it should be stated that the process of adapting protein dataset to the pattern recognition model could be done in different manners providing even better performance of introduced fusion methods.

References

1. Akkaladevi S., Balkasim M., Pan Y., Protein Secondary Structure Prediction Using Neural Network and Simulated Annealing Algorithm, Proc. 26th Int. Conf. of the IEEE EMBS, San Francisco (2004) 2987–2990
2. Aydin Z., Altunbasak Y., Borodovsky M., Protein Secondary Structure Prediction with Semi Markov HMMs, Proc. 26th Int. Conf. of the IEEE EMBS, San Francisco (2004) 2964–2967
3. Bermann H., Westbrook Z., Feng G. et al., The Protein Data Bank, Nucleic Acid Res. 2 (2000) 235–242
4. Bridle J., Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters, [in] Neural Information Processing Systems **2**, D. Touretzky (ed.), (San Mateo, CA) Morgan Kaufmann (1990) 211–217
5. Combet C., Blanchet C., Geourjon C., Deleage G., NPS@: Network Protein Sequence Analysis, TIBS Vol. 25, No. 3 (2000) 147–150
6. Garnier J., Gibrat J-F., Robson B., GOR secondary structure prediction method version IV, Methods in Enzymology R.F. Doolittle Ed., Vol. 266 (1996) 540–553
7. Geourjon C., Deleage G., SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments, Comput Appl Biosci (1995) 681–684
8. Giacinto G., Roli F., Design of effective neural network ensembles for image classification processes, Image Vision and Computing Journal (2001) 699–707
9. Guermeur Y., Combinaison de classifieurs statistiques, Application a la prediction de structure secondaire des proteines, PhD Thesis
10. Guermeur Y., Combining discriminant models with new multi-class SVMs, Pattern Analysis & Applications (2002) 5: 168–179
11. Henikoff S., Henikoff JG., Amino acid substitution matrices from protein blocks, PNAS USA (1992) 10915–10919
12. Hobohm U., Sanders C., Enlarged representative set of protein structures, Protein Science (1994) 522
13. Jones D., Protein Secondary Structure Prediction Based on Position-Specific Scoring Matrices, J. Mol. Biol. **292** (1999) 397–407
14. Kabsch W., Sanders C., Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers (1983) 2577–2637
15. Krieger E., Hooft R., Nabuurs S., Vriend G., PDBFinderII - a database for protein structure analysis and prediction (2004)
16. Kuncheva L., A theoretical study on six classifier fusion strategies, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 2 (2002) 281–286
17. Kuncheva L., Combining pattern classifiers: methods and algorithms, John Wiley & Sons, New Jersey (2004)
18. Selbig J., Mevissen T., Lengauer T., Decision-tree Based Formation on Consensus Secondary Structure Prediction, Bioinformatics **15** (1999) (1039–1046)

19. Wang L., Liu J., Zhou H., A Comparison of Two Machine Learning Methods for Protein Secondary Structure Prediction, Proc. 3rd Int. Conf. on Machine Learning and Cybernetics, Shanghai (2004) 2730–2735
20. Xiaochung Y., Wang B., A Protein Secondary Structure Prediction Framework Based on the Support Vector Machine, Proc. 4th Int. Conf. on Information Management WAIM03, LNCS 2762, Springer Verlag (2003) 266–277
21. Xu H., Lau K., Lu L., Protein Secondary Structure Prediction Using Data Mining Tool C5, Proc. 11th IEEE Int. Conf. on Tools in AI, Chicago (1999) 107–110
22. Zaki M., Shan J., Bystroff C., Mining Residue Contacts in Protein Using Local Structure Prediction, IEEE Trans. on SMC **33** (2003) 258–264
23. Zemla A., Venclovas C., Fidelis K., Rost B., Some measures of comparative performance in the tree casps. PROTEINS: Structure, Function, and Genetics (1999) 220–223
24. Zhang B., Chen Z., Murphey Y., Protein Secondary Structure Prediction Using Machine Learning, Proc. IEEE Int. Conf. on Neural Networks, Montreal (2004) 532–537

A Novel Data Mining Approach for the Accurate Prediction of Translation Initiation Sites

George Tzanis, Christos Berberidis, and Ioannis Vlahavas*

Department of Informatics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece
{gtzanis,berber,vlahavas}@csd.auth.gr
<http://mlkd.csd.auth.gr>

Abstract. In an mRNA sequence, the prediction of the exact codon where the process of translation starts (Translation Initiation Site – TIS) is a particularly important problem. So far it has been tackled by several researchers that apply various statistical and machine learning techniques, achieving high accuracy levels, often over 90%. In this paper we propose a machine learning approach that can further improve the prediction accuracy. First, we provide a concise review of the literature in this field. Then we propose a novel feature set. We perform extensive experiments on a publicly available, real world dataset for various vertebrate organisms using a variety of novel features and classification setups. We evaluate our results and compare them with a reference study and show that our approach that involves new features and a combination of the Ribosome Scanning Model with a meta-classifier shows higher accuracy in most cases.

1 Introduction

The last decades has seen a rapid progress in two major scientific areas, biology and computer science. Lately, the field of data mining and machine learning has provided biologists, as well as experts from other areas, a powerful set of tools to analyze new data types in order to extract various types of knowledge fast, accurately and reliably. These tools combine powerful techniques from different areas such as statistics, mathematics, artificial intelligence, algorithmics and database technology. This fusion of technologies aims to overcome the obstacles and constraints posed by the traditional statistical methods.

Translation is one of the basic biological operations that attract biologists' attention. Translation along with replication and transcription make possible the transmission and expression of an organism's genetic information. The initiation of translation plays an important role in understanding which part of a sequence is translated and consequently what is the final product of the process. A sequence contains a number of sites where the translation might initiate. However, only one of them is the true translation initiation site (TIS). The recognition of the true TIS among the candidate TISs is not a trivial task as it requires the highest possible accuracy.

* This work was partially supported by the Greek R&D General Secretariat through a PENED program (EPAN M.8.3.1, No. 03E D 73).

Classification and meta-classification methods have been used in order to deal with this problem.

In this paper, we propose the use of a new feature set along with a combination of meta-classifiers with the *Ribosome Scanning Model* (RSM). We test the feature set with 2 different statistics and then use the extracted features to train 7 different classifiers and a meta-classifier. Then we estimate the prediction accuracy of our approach using a state of the art evaluation method, namely 10 times 10-fold cross validation. We also train the same classifiers and perform the same evaluation using the feature sets from a reference study [12] and compare the results. In most cases the proposed approach has a clear advantage against the reference study, showing that both our feature set and our classification setup are more effective in terms of accuracy.

This paper is organized as follows: In the next section, we provide a concise review of the literature on TIS prediction. Section 3 contains a brief introduction on the biological problem attacked in our study. In Section 4 we describe the mining approach we propose and in Section 5 we explain the experimental methodology we followed, show the results of the extensive experiments we performed and finally the evaluation and comparison of our work with a reference study. In the last sections we summarize our paper with our conclusions and directions for future research.

2 Related Work

Since 1982 the prediction of TISs has been extensively studied using biological approaches, data mining techniques and statistical models. Stormo et al. [20] used the perceptron algorithm to distinguish the TISs. In 1987 Kozak developed the first weight matrix for the identification of TISs in cDNA sequences [8]. The consensus pattern derived from this matrix is GCC[AG]CCatgG (the bold residues are the highly conserved positions). Meanwhile, Kozak and Shatkin [10] had proposed the scanning model of translation initiation, which was later updated by Kozak [9]. According to this model translation initiates at the first start codon that is in an appropriate context.

Pedersen and Nielsen [16] make use of artificial neural networks to predict the TISs achieving an overall accuracy of 88% in *Arabidopsis thaliana* dataset and 85% in vertebrate dataset. Zien et al. [25] studied the same vertebrate dataset, but instead of neural networks employed support vector machines using various kernel functions. Hatzigeorgiou [4] proposed an ANN system named “DIANA-TIS” consisting of two modules: the consensus ANN, sensitive to the conserved motif and the coding ANN, sensitive to the coding or non-coding context around the start codon. The method applied in human cDNA data and 94% of the TIS were correctly predicted. Salamov et al. [19] developed the program ATGpr, using a linear discriminant approach for the recognition of TISs by estimating the probability of each ATG codon being the TIS. Nishikawa et al. [15] presented an improved program, ATGpr_sim, which employs a new prediction algorithm based on both statistical and similarity information. In [11] Gaussian Mixture Models were used for the prediction of TISs improving classification accuracy.

Feature generation and correlation based feature selection along with machine learning algorithms has also been employed [13, 24]. In these studies a large number of k -gram nucleotide patterns were utilized. By using a scanning model an overall accuracy of 94% was attained on the vertebrate dataset of Pedersen and Nielsen. Later, in [12] the same three-step method was used, but k -gram amino acid patterns were considered, instead of nucleotide patterns.

Nadershahi et al. [14] compared five methods -firstATG, ESTScan, Diogenes, Netstart [16] and ATGPr [19]- for the prediction of the TIS. For the comparison a dataset of 100 Expressed Sequence Tag (EST) sequences, 50 with and 50 without a TIS, was created. ATGPr appeared to outperform the other methods over this dataset.

3 Background Knowledge

Translation is the second process of protein synthesis. In particular, after a DNA molecule has been transcribed into a messenger RNA (mRNA) molecule, an organelle called ribosome scans the mRNA sequence. The ribosome reads triplets, or *codons*, of nucleotides and “translates” them into amino acids. An mRNA sequence can be read in three different ways in a given direction. Each of these ways of reading is referred to as *reading frame*.

Translation, usually, initiates at the AUG codon nearest to the 5′ end of the mRNA sequence. However, this is not always the case, since there are some escape mechanisms that allow the initiation of translation at following, but still near the 5′ end AUG codons. Due to these mechanisms the recognition of the TIS on a given sequence becomes more difficult.

After the initiation of translation, the ribosome moves along the mRNA molecule, towards the 3′ end (the direction of translation is 5′ → 3′) and reads the next codon. This process is repeated until the ribosome reaches a stop codon. For each codon read the proper amino acid is brought to the protein synthesis site by a transfer RNA (tRNA) molecule. The amino acid is joined to the protein chain, which by this way is elongated.

A codon that is contained in the same reading frame with respect to another codon is referred to as *in-frame codon*. We call *upstream* the region of a nucleotide sequence from a reference point towards the 5′ end. Respectively, the region of a nucleotide sequence from a reference point towards the 3′ end is referred to as *downstream*. In TIS prediction problem the reference point is an AUG codon. The above are illustrated in Fig. 1.

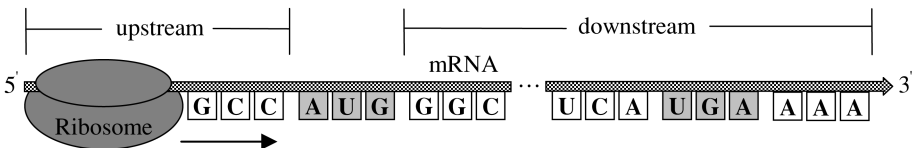


Fig. 1. Translation initiation – The ribosome scans the mRNA sequence from the 5′ end to the 3′ end until it reads an AUG codon. If the AUG codon has appropriate context, the translation initiates at that site and terminates when a stop codon (i.e. UGA) is read. An in-frame codon is represented by three consecutive nucleotides that are grouped together.

4 The Proposed TIS Prediction Approach

We propose a machine learning approach, focusing on all stages of the prediction, namely the data selection, the feature extraction, the training of the classifier and the evaluation of the effectiveness.

4.1 Datasets

The dataset we used in our study consists of 3312 genomic sequences collected from various vertebrate organisms, acquired from the Kent Ridge Biomedical Data Set Repository (<http://sdmc.i2r.a-star.edu.sg/rp>). Being DNA sequences, they contain only the letters A, C, G and T. Therefore, a candidate TIS is referred to as ATG codon instead of AUG codon.

The sequences of the dataset were extracted from GenBank, release 95 [2]. Only nuclear genes with an annotated start codon were selected. The DNA sequences have been processed and the introns have been removed. From the resulting dataset, the selected sequences contain at least 10 nucleotides upstream of the initiation point and at least 150 nucleotides downstream (with reference to A in the ATG codon). All sequences containing non-nucleotide symbols in the interval mentioned above (typically due to incomplete sequencing) were excluded. Moreover, the dataset has been gone through very thorough reduction of redundancy [16].

4.2 Features

One of the most important tasks in prediction is the extraction of the right features that describe the data. This is also a particularly crucial point in our approach in terms of novelty and performance. The basic features used in our approach are summarized in Table 1. Some of them (features 1, 2, 12-15) have been already studied in previous research works [12, 13, 24]. However, there is a number of new features that we propose and study in this paper. One set of features (features 3), that have been proposed in previous works of ours [21, 22], concern the periodic occurrence of particular nucleotides at a specific position inside an in-frame codon (Figure 2). Another set of features that has not been studied yet (features 4) includes features that count the amino acids that appear at each in-frame position. The numbering of each position of a sequence is presented in Figure 3. For numbering we used the same conventions as in other studies [12, 13, 21, 22, 24]. What's more important, we propose a number of extra features based on the chemical properties of amino acids that haven't been considered before. These features are counts of hydrophobic, hydrophilic, acidic, or basic amino acids, as well as counts of aromatic, aliphatic, or neither aromatic, nor aliphatic amino acids (features 5-11). We used a window size of 99 nucleotides upstream and 99 nucleotides downstream the ATG for calculating the values of the features.

position:	1	2	3	1	2	3		1	2	3	1	2	3
	5'	G	C	C	A	C	C	A	T	G	G	C	A
								T	C	G	3'		

Fig. 2. The positions of nucleotides inside the in-frame codons

Table 1. The features used in our approach

Features	Description
1 up_ x down_ x	Count the number of amino acid x in the upstream and downstream region respectively.
2 up-down_ x	Counts the difference between the number of occurrences of amino acid (or set of amino acids) x in the upstream region and the number of occurrences of x in the downstream region.
3 up_pos_ k_x down_pos_ k_x	Count the number of occurrences of nucleotide x in the k^{th} position of the in-frame codons ($k \in \{1, 2, 3\}$) in the upstream and downstream region respectively.
4 pos_ $-3k$ pos_ $3(k+1)$	Concern the presence of amino acids at in-frame positions in the upstream and downstream region respectively ($k \geq 1$).
5 up_hydrophobic down_hydrophobic	Count the number of hydrophobic amino acids in the upstream and downstream region respectively.
6 up_hydrophilic down_hydrophilic	Count the number of hydrophilic amino acids in the upstream and downstream region respectively.
7 up_acidic down_acidic	Count the number of acidic amino acids in the upstream and downstream region respectively.
8 up_basic down_basic	Count the number of basic amino acids in the upstream and downstream region respectively.
9 up_aromatic down_aromatic	Count the number of aromatic amino acids in the upstream and downstream region respectively.
10 up_aliphatic down_aliphatic	Count the number of aliphatic amino acids in the upstream and downstream region respectively.
11 up_non_aromatic/ aliphatic down_non_aromatic/ aliphatic	Count the number of amino acids that are not aromatic nor aliphatic in the upstream and downstream region respectively.
12 up_ -3 _[AG]	A Boolean feature that is true if there is an A or a G nucleotide three positions before the ATG codon, according to Kozak's pattern (GCC[AG]CCatgG).
13 down_ $+1$ _G	A Boolean feature that is true if there is a G nucleotide in the first position after the ATG codon, according to Kozak's pattern (GCC[AG]CCatgG).
14 up_ATG	A Boolean feature that is true if there is an in-frame upstream ATG codon.
15 down_stop	A Boolean feature that is true if there is an in-frame downstream stop codon (TAA, TAG, TGA).

position: -6 -5 -4 -3 -2 -1 +1 +2 +3 +4 +5 +6 +7 +8 +9
 5' G C C A C C A T G G C A T C G 3'

Fig. 3. The positions of nucleotides relative to an ATG codon

4.3 Feature Selection Algorithms

For the conduction of our experiments we have utilized the Weka library of machine learning algorithms [23]. We have used the following feature selection methods:

- *Chi-Squared*. Evaluates the worth of an attribute by computing the value of the X^2 statistic with respect to the class.
- *Gain Ratio*. Evaluates the worth of an attribute by measuring the gain ratio with respect to the class.

4.4 Classification Algorithms

For classification we have used the following classification algorithms:

- *C4.5*. Algorithm for generating a decision tree [18].
- *RIPPER*. This is a propositional rule learner called Repeated Incremental Pruning to Produce Error Reduction. [3].
- *Decision Table*. This algorithm implements a simple decision table majority classifier [7].
- *Naïve Bayes*. A Naive Bayes classifier [5].
- *SVM*. This is the John Platt's [17] sequential minimal optimization algorithm for training a support vector classifier. The Weka implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default.
- *Multilayer Perceptron*. This algorithm implements a neural network that uses back-propagation to train.
- *k-Nearest Neighbors classifier*. The algorithm normalizes attributes by default and can do distance weighting. We have used this algorithm with 1-nearest neighbor. More information about this algorithm can be found in [1].

The idea of meta-classifier systems is an attempt to construct more accurate classification models by combining a number of classifiers. Classifier combination includes two main paradigms: classifier selection and classifier fusion. In the first case a new instance is classified by selecting the appropriate classifier, while in the second case a new instance is classified according to the decisions of all the classifiers. We implemented and used two meta-classification algorithms of the second paradigm:

- *Simple Voting*. This algorithm combines the decisions of multiple classifiers and makes the final decision by considering the majority of the votes. Each classifier participates equally to the voting.
- *Weighted Voting*. This algorithm also applies voting but each classifier participates with a different weight to the voting procedure. The weight for each classifier is its classification accuracy. In other words, the more accurate classifiers contribute more to the final result than the less accurate ones.

4.5 Evaluation Method

In order to evaluate the results of our experiments we have used stratified 10-fold cross-validation (CV). In particular, the performance of a classifier on a given dataset using 10-fold CV is evaluated as follows. The dataset is divided into 10 non-overlapping almost equal size parts (folds). In stratified CV each class is represented in each fold at the same percentage as in the entire dataset. After the dataset has been divided, a model is built using 9 of the folds as a training set and the remaining fold as a test set. This procedure is repeated 10 times with a different test set each time. The use of the 10-fold CV was based on the widely accepted study of R. Kohavi [6]. The results of this work indicate that for many real-word datasets, the best method to use for model selection is stratified 10-fold CV, even if computation power allows using more folds.

Furthermore, in order to increase the reliability of the evaluation, we have repeated each experiment 10 times, each time generating randomly different folds and we finally took into account the average of the 10 independent repeats.

5 Experimental Results

In this section we present the results of the experiments we conducted, compared to a reference study [12] on the same dataset. The first subsection describes the results of feature selection and the second subsection deals with the results of classification. Finally, a discussion about the results is given in the third subsection.

5.1 Feature Selection Results

The best features selected by the two feature selection methods are presented in Table 2. When using the X^2 statistic for feature selection the 8 out of the 10 top features are the ones we propose (5 are proposed in this paper and the other 3 have been proposed in [21]). When using the gain ratio measure, the 6 out of the 10 top features are the ones we propose (4 are introduced in this paper and the other 2 have been proposed in [21]). It should be noted here that the feature `down_stop` of the reference study is not Boolean (see Table 1). Instead, it counts the number of the in-frame downstream stop codons. Moreover, in this study, features of the form `down_xy` and `up_xy` are also considered, where x and y are amino acids, or a stop codon.

5.2 Classification Results

After extensive experiments, we present the results produced by three different setups:

- *Setup 1.* The use of a meta-classifier (simple voting or weighted voting) for predicting the TIS.
- *Setup 2.* The use of a meta-classifier (simple voting or weighted voting) incorporated with the *ribosome scanning model* (RSM). The first candidate TIS that appears inside a sequence and has received more than 50% of votes for being a TIS is selected as the true TIS. The remaining candidates of the same sequence are

considered not to be TISs, even if they have received more than 50% of votes for being a TIS.

- *Setup 3.* We propose the use of the RSM based on the results of a meta-classifier (simple voting or weighted voting). Among all candidate TISs of a sequence, the one that has received the larger number of (positive) votes is considered as the true TIS. The remaining candidates of the same sequence are considered as non-TISs.

Table 2. The top features that were selected by the feature selection methods from the set of features we propose (middle column) and from the set of features used in the reference study [12] (last column). The features are ordered according to their ranking.

FS Method	Our Features	Reference Features
X^2	up_ATG down_1_G down_hydrophobic down_non_aromatic/aliphatic down_3_C down_stop down_aliphatic up-down_non_aromatic/aliphatic up-down_hydrophobic down_2_T	up_ATG down_stop up_M down_A down_L down_V down_E down_D up_-3_[AG] down_G
Gain Ratio	up_ATG down_stop up_M up_-3_[AG] down_non_aromatic/aliphatic down_1_G up-down_non_aromatic/aliphatic down_hydrophobic up-down_hydrophobic down_2_C	down_stop up_RR up_NH down_MY up_ATG up_M down_Lstop down_stopR down_stopS down_Pstop

The results of the three setups using either simple voting, or weighted voting are presented in Tables 3 and 4. The first table contains the results produced using the set of features we propose, while the second one contains the results produced using set of features utilized in the reference study.

Figure 4 depicts the results for the 3rd Setup, which is introduced in this paper. In particular, the results obtained using the features proposed here are compared to the results obtain using the features of the reference study [12]. The results of each of the seven classifiers are not presented here for brevity since only their output is considered in the meta-classification step.

Table 3. Classification results using the set of features we propose. The grayed cells indicate the setup that achieves the highest accuracy.

FS Method	# Top Features	Simple Voting			Weighted Voting		
		Setup 1	Setup 2	Setup 3	Setup 1	Setup 2	Setup 3
X^2	5	86.20%	86.34%	86.32%	86.20%	86.34%	86.38%
	10	90.36%	91.58%	93.30%	90.36%	91.58%	93.65%
	15	94.60%	95.19%	95.53%	94.60%	95.19%	96.01%
	20	94.71%	95.34%	95.89%	94.71%	95.34%	96.25%
	25	94.78%	95.41%	95.85%	94.79%	95.41%	96.19%
	30	94.79%	95.37%	95.79%	94.78%	95.37%	96.21%
Gain Ratio	5	93.09%	94.16%	94.87%	93.09%	94.16%	94.77%
	10	94.10%	94.99%	95.12%	94.10%	94.99%	95.46%
	15	94.58%	95.14%	95.27%	94.58%	95.14%	95.75%
	20	94.60%	95.16%	95.34%	94.60%	95.16%	95.79%
	25	94.66%	95.23%	95.53%	94.66%	95.23%	95.98%
	30	94.70%	95.28%	95.60%	94.70%	95.28%	95.99%

Table 4. Classification results using the set of features utilized in the reference study. The grayed cells indicate the setup that achieves the highest accuracy.

FS Method	# Top Features	Simple Voting			Weighted Voting		
		Setup 1	Setup 2	Setup 3	Setup 1	Setup 2	Setup 3
X^2	5	87.98%	90.71%	91.64%	87.98%	90.71%	90.90%
	10	91.75%	92.57%	93.68%	91.75%	92.57%	93.70%
	15	91.93%	92.61%	93.76%	91.93%	92.61%	93.84%
	20	92.40%	92.90%	94.07%	92.40%	92.90%	94.25%
	25	92.40%	92.85%	94.02%	92.40%	92.85%	94.19%
	30	92.53%	92.91%	94.18%	92.53%	92.91%	94.32%
Gain Ratio	5	82.62%	86.56%	87.42%	82.62%	86.56%	84.79%
	10	85.33%	91.80%	91.67%	85.33%	91.80%	87.59%
	15	87.23%	91.00%	92.01%	87.23%	91.00%	89.45%
	20	87.97%	90.98%	92.27%	87.97%	90.98%	89.82%
	25	88.00%	90.98%	92.33%	88.00%	90.98%	89.90%
	30	88.02%	91.00%	92.35%	88.02%	91.00%	89.96%

5.3 Discussion

The classification accuracy achieved by the setup proposed in this study (Setup 3) is higher in almost all cases when using the set of features we propose. When using the

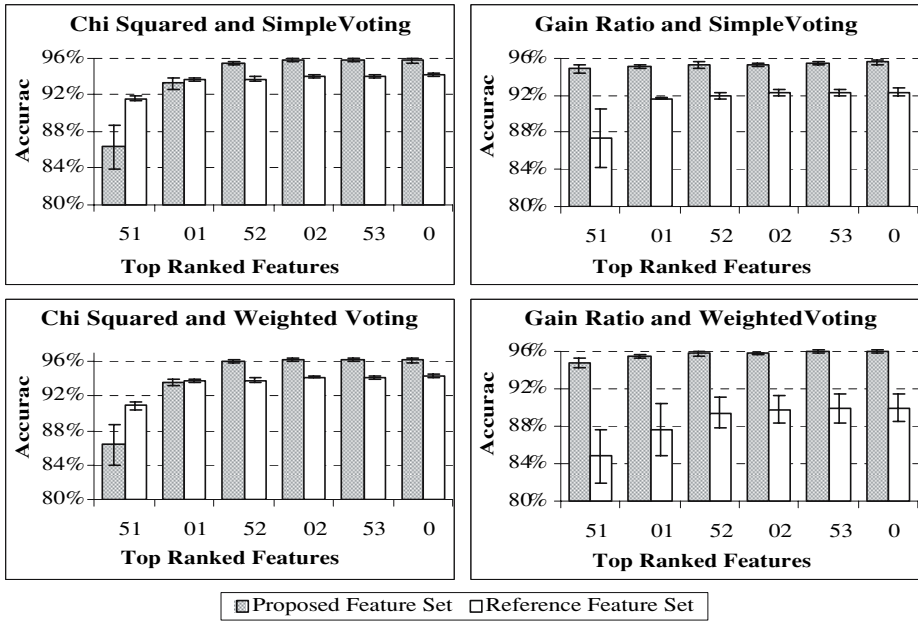


Fig. 4. Comparison of the results obtained using the proposed feature set against the reference feature set [17], for Setup 3. The error bars are calculated as ± 2 standard deviations.

features of the reference study, Setup 3 outperforms Setups 1 and 2 with the χ^2 statistic for feature selection. However, this is not the case when the Gain Ratio measure is used. Moreover, when the features we propose are used, the best accuracy is achieved with χ^2 statistic for feature selection and the top 20 ranked features along with the weighted voting meta-classification algorithm and Setup 3 (96.25%). At the other hand, when the features of the reference study are used, the best accuracy is achieved with χ^2 and the top 30 ranked features along with the weighted voting meta-classification algorithm and Setup 3 (94.32%).

The results of Setups 1 and 2 for both simple and weighted voting are identical. This happens because there are zero or very few decisions of the simple voting schema that are based on a majority near 50%. The percentage of majority is high enough, so that the weighting of the votes do not affect the results. However, in Setup 3 the results are better when using the weighted voting schema.

The incorporation of the ribosome scanning model (RSM) (Setup 2) provides better classification accuracy in almost all cases. Moreover, the improvement is much higher when RSM is used with the features of the reference study. Specifically, the improvement achieved is 6.48 percentage units. The improvement when the set of our proposed features are used is only 1.22 percentage units. However, when using the set of features we propose in Setup 1, the classification accuracy is in most cases higher than the results achieved using the reference study features enhanced with RSM (Setup 2). This implies that a large portion of the improvement provided by the use of RSM is incorporated in the features we use. In other words, although the use of RSM greatly enhances the effectiveness of the reference feature set, its effect on our feature

set (used in Setup 3) is not that significant because the accuracy attained by the latter alone is already the highest so far.

Although previous works [13, 21, 24] have shown that the feature that counts the distance of a candidate ATG from the start of the sequence is very good for discriminating a TISs from non-TISs we excluded it from our study. The reason is that this feature is highly affected by the intrinsic characteristics of the sequences contained in the dataset we used. For example, for each sequence there are always 150 nucleotides downstream of the TIS and there is up to a maximum of 150 nucleotides upstream.

6 Conclusions and Future Work

In this paper we tackle the problem of the prediction of Translation Initiation Sites in genome sequences. We implement a machine learning approach that shows higher accuracy than previous approaches on a public vertebrate dataset. First, we provide a review of the literature on this task and a short section on biological background knowledge. By extensive experiments using two different statistics (X^2 and Gain Ratio) we propose the use of a novel feature set that leads to higher accuracy when compared to the feature sets of the reference study. Then, we introduce a new prediction setup that utilizes meta-classifiers and the ribosome scanning model in order to achieve higher accuracy. We support our claims by extensive experiments using 7 different classifiers along with 2 meta-classifiers. Then, we evaluated our results by performing 10 times 10-fold cross validation, in order to prove the reliability of our approach.

In the near future we are going to apply our approach on more datasets, run more experiments with more classifiers and new feature sets. We also plan to investigate the application of our approach on other functional site prediction problems, such as splice sites.

References

1. Aha, D., Kibler, D. Instance-based learning algorithms, *Machine Learning* (1991) 6, 37-66
2. Benson, D., Boguski, M., Lipman, D., Ostell, J. Genbank. *Nucleic Acids Research* 25, (1997) 1-6
3. Cohen, W.: Fast Effective Rule Induction. In *Proceedings of the 12th International Conference on Machine Learning*. Morgan Kaufmann, Lake Tahoe, USA (1995) 80-89
4. Hatzigeorgiou, A.: Translation Initiation Start Prediction in Human cDNAs with High Accuracy. *Bioinformatics* (2002) 18(2) 343-350
5. John, G.H., Langley, P.: Estimating Continuous Distributions in Bayesian Classifiers. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann, San Mateo, USA (1995) 338-345
6. Kohavi, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection, In *Proceedings of 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995
7. Kohavi, R. The Power of Decision Tables. In *Proceedings of the 8th European Conference on Machine Learning (ECML'95)*, LNAI 914, 174-189. Springer Verlag, 1995.

8. Kozak, M.: An Analysis of 5'-Noncoding Sequences from 699 Vertebrate Messenger RNAs. *Nucleic Acids Research* (1987) 15(20) 8125-8148
9. Kozak, M.: The Scanning Model for Translation: An Update. *The Journal of Cell Biology* (1989) 108(2) 229-241
10. Kozak, M., Shatkin, A.J.: Migration of 40 S Ribosomal Subunits on Messenger RNA in the Presence of Edeine. *Journal of Biological Chemistry* (1978) 253(18) 6568-6577
11. Li, G., Leong, T-Y, Zhang, L: Translation Initiation Sites Prediction with Mixture Gaussian Models in Human cDNA Sequences. *IEEE Transactions on Knowledge and Data Engineering* (2005) 8(17) 1152-1160
12. Liu, H., Han, H., Li, J., Wong, L.: Using Amino Acid Patterns to Accurately Predict Translation Initiation Sites. *In Silico Biology* (2004) 4(3) 255-269
13. Liu, H., Wong, L.: Data Mining Tools for Biological Sequences. *Journal of Bioinformatics and Computational Biology*, (2003) 1(1) 139-168
14. Nadershahi, A., Fahrenkrug, S.C., Ellis, L.B.M.: Comparison of computational methods for identifying translation initiation sites in EST data. *BMC Bioinformatics* (2004) 5(14)
15. Nishikawa, T., Ota, T., Isogai, T.: Prediction whether a Human cDNA Sequence Contains Initiation Codon by Combining Statistical Information and Similarity with Protein Sequences. *Bioinformatics* (2000) 16(11) 960-967
16. Pedersen, A.G., Nielsen, H.: Neural Network Prediction of Translation Initiation Sites in Eukaryotes: Perspectives for EST and Genome analysis. In *Proceedings of the 5th International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, USA (1997) 226-233
17. Platt, J. Fast Training of Support Vector Machines using Sequential Minimal Optimization. In *Advances in Kernel Methods - Support Vector Learning*, B. Schoelkopf, C. Burges, A. Smola (Eds.), MIT Press, (1998)
18. Quinlan, J.R.: C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, USA (1993)
19. Salamov, A.A., Nishikawa, T., Swindells, M.B.: Assessing Protein Coding Region Integrity in cDNA Sequencing Projects. *Bioinformatics* (1998) 14(5) 384-390
20. Stormo, G.D., Schneider, T.D., Gold, L., Ehrenfeucht, A.: Use of the 'Perceptron' Algorithm to Distinguish Translational Initiation Sites in *E. coli*. *Nucleic Acids Research* (1982) 10 (9) 2997-3011
21. Tzani, G., Berberidis, C., Alexandridou, A., Vlahavas, I.: Improving the Accuracy of Classifiers for the Prediction of Translation Initiation Sites in Genomic Sequences. In *Proceedings of the 10th Panhellenic Conference on Informatics (PCI'2005)*, Volos, Greece, (2005) 426 – 436
22. Tzani, G., Vlahavas, I.: Prediction of Translation Initiation Sites Using Classifier Selection. In *Proceedings of the 4th Hellenic Conference on Artificial Intelligence (SETN'06)*, G. Antoniou, G. Potamias, D. Plexousakis, C. Spyropoulos (Eds.), Springer-Verlag, LNAI 3955, Heraklion, Greece (2006) 367 - 377
23. Witten, I.H., Frank, E.: *Data Mining: Practical Machine Learning Tools with Java Implementations*. Morgan Kaufmann, San Francisco (2000)
24. Zeng F., Yap H., Wong, L.: Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites. In *Proceedings of the 13th International Conference on Genome Informatics*, Tokyo, Japan (2002) 192-200
25. Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., Müller, K.R.: Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites. *Bioinformatics* (2000) 16(9) 799-807

SPSO: Synthetic Protein Sequence Oversampling for Imbalanced Protein Data and Remote Homology Detection

Majid Beigi and Andreas Zell

University of Tübingen
Center for Bioinformatics Tübingen (ZBIT)
Sand 1, D-72076 Tübingen, Germany
{majid.beigi, andreas.zell}@uni-tuebingen.de

Abstract. Many classifiers are designed with the assumption of well-balanced datasets. But in real problems, like protein classification and remote homology detection, when using binary classifiers like support vector machine (SVM) and kernel methods, we are facing imbalanced data in which we have a low number of protein sequences as positive data (minor class) compared with negative data (major class). A widely used solution to that issue in protein classification is using a different error cost or decision threshold for positive and negative data to control the sensitivity of the classifiers. Our experiments show that when the datasets are highly imbalanced, and especially with overlapped datasets, the efficiency and stability of that method decreases. This paper shows that a combination of the above method and our suggested oversampling method for protein sequences can increase the sensitivity and also stability of the classifier. Our method of oversampling involves creating synthetic protein sequences of the minor class, considering the distribution of that class and also of the major class, and it operates in data space instead of feature space. This method is very useful in remote homology detection, and we used real and artificial data with different distributions and overlappings of minor and major classes to measure the efficiency of our method. The method was evaluated by the area under the Receiver Operating Curve (ROC).

A dataset is imbalanced if the classes are not equally represented and the number of examples in one class (major class) greatly outnumbers the other class (minor class). With imbalanced data, the classifiers tend to classify almost all instances as negative. This problem is of great importance, since it appears in a large number of real domains, such as fraud detection, text classification, medical diagnosis and protein classification [1,2]. There have been two types of solutions for coping with imbalanced datasets. The first type, as exemplified by different forms of re-sampling techniques, tries to increase the number of minor class examples (oversampling) or decrease the number of major class examples (undersampling) in different ways. The second type adjusts the cost of error or decision thresholds in classification for imbalanced data and tries to control the sensitivity of the classifier [3,4].

Undersampling techniques involve loss of information but decrease the time of training. With oversampling we do not lose the information but instead it increases the size of the training set and so the training time for classifiers. Furthermore, inserting inappropriate data can lead to overfitting. Some researchers [2] concluded that undersampling can better solve the problem of imbalanced datasets. On the other hand, some other researchers are in favor of oversampling techniques. Wu and Chang [5] showed that with imbalanced datasets, the SVM classifiers learn a boundary that is too close to positive examples. Then if we add positive instances (oversampling), they can push the boundary towards the negative data, and we have increased the accuracy of classifier.

To decide the question of oversampling vs. undersampling, two parameters should be taken into consideration: the *imbalance ratio* and the distribution of data in imbalanced datasets. The *imbalance ratio* ($\frac{\text{NumberOfMinorityData}}{\text{NumberOfMajorityData}}$) is an important parameter that shows the degree of imbalance. In undersampling we should be sure of the existence of enough information in the minor class and also of not losing the valuable information in the major class. We found out that the oversampling technique can balance the class distribution and improve that situation. But the distribution of inserted positive instances is of great importance. Chawla et al. [6] developed a method for oversampling named Synthetic Minority Oversampling Technique (SMOTE). In their technique, between each positive instance and its nearest neighbors new synthetic positive instances were created and placed randomly between them. Their approach proved to be successful in different datasets.

On the other hand Veropoulos et al. [4] suggested using different error costs (DEC) for positive and negative classes. So the classifier is more sensitive to the positive instances and gets more feedback about the orientation of the class-separating hyperplane from positive instances than from negative instances.

In protein classification problems the efficiency of that approach (Veropoulos et al. [4]) has been accepted. In kernel based protein classification methods [7,8,1] a class-depending regularization parameter is added to the diagonal of the kernel matrix. But, based on our experiments, if the dataset is highly imbalanced and has overlapping data, choosing a suitable ratio of error costs for positive and negative examples is not always simple and sometimes the values near the optimum value of the error cost ratio give unsatisfying results.

We propose an oversampling technique for protein sequences in which the minority class in the data space is oversampled by creating synthetic examples. Working with protein data in data space instead of feature space allows us to consider the probability distribution of residues of the sequence using a HMM-profile of the minority class and also one of the majority class and then synthesize protein sequences which can push precisely the boundary towards the negative examples. So we increase the information of the minor class. Our method of oversampling can cause the classifier to build larger decision regions for the minor class without overlapping with the major class. In this work we used real and artificial data with different degrees of overlapping and imbalance ratio to show the efficiency of our methods and we also suggest that our algorithm can

be used along with DEC methods to increase the sensitivity and stability of the classifier. As SVM classifiers and kernel methods outperformed other methods in protein classification [7,1,8], we discuss the efficiency of our oversampling technique when used with kernel-based classifiers.

1 Algorithm

Given a set of positive training sequences (minor class) S_+ and a set of negative training sequences (major class) S_- we want to create synthetic protein sequences $S_{synthetic}$ as mutated replicas of each sequence of the minor class, provided that those synthetic sequences are created by an HMM profile (Hidden Markov Model profile) of the minor class and are phylogenetically related to that class and far away from the major class. For this, at first we build a multiple alignment of the sequences of the minor class using ClustalW [9] and then we train a hidden Markov model profile with length of the created multiple alignment sequences for each class (positive data and every family belonging to the negative data).

For every sequence in the minor class we create another mutated sequence synthetically. For that, we consider an arbitrary N_m as number of start points for mutation in that sequence. We suppose the $HMMp_+$ (hidden Markov model profile of positive instances) has emitted another sequence identical to the main sequence until the first point of mutation. From that point afterward we assume that $HMMp_+$ emits new residues until the emitted residue is equal to a residue in the same position in the main sequence. From this residue, all residues are the same as residues in the original sequence until the next point of mutation (Fig. 1).

Algorithm. SPSO(S_+, S_-)

Input : S_+ , set of sequences of minority class; S_- , set of sequences of majority class

Output: $S_{synthetic}$, set of synthetic protein sequences from the minority class

```

1 Create HMM profile of set  $S_+$ , call it  $HMMp_+$  ;
2 Create array of HMM profiles consisting of all families belonging to  $S_-$ , call it
   $HMMp_-[]$ ;
3 Choose an arbitrary number as number of start points for mutation, call it  $N_m$ ;
4 for  $i \leftarrow 1$  to  $|S_+|$  do
5    $s = S_+[i]$  ;
6   repeat
7     Create an array of sorted non-repeating random numbers with size of  $N_m$ 
      as array of start points for mutation, call it  $P_m$  ;
8      $S_{synthetic}[i] = \text{newSeq}(s, HMMp_+, P_m)$ ;
9      $p_+ = P_e(S_{synthetic}[i], HMMp_+)$  ;
10     $p_-[] = P_e(S_{synthetic}[i], HMMp_-[])$  ;
11  until  $p_+ < \min p_-[]$ ;
12 end
13 return  $S_{synthetic}$ 
```

Function. `newSeq( $s, HMMp_+, P_m$ )`

Input : s , original sequence; $HMMp_+$, HMM profile of set S_+ to which s belongs; P_m , array of start points for mutation

Output: $s_{synthetic}$, synthetic sequence from s

```

1  $s_{synthetic} = s$  ;
2 for  $i \leftarrow 1$  to  $|P_m|$  do
3    $p = P_m[i]$  ; (* assume that  $HMMp_+$  in position  $p$  has emitted  $s[p]$  *)
4   repeat
5      $s_{synthetic}[p + 1] =$  emitted residue in position  $p + 1$  by  $HMMp_+$  ;
6      $p = p + 1$  ;
7   until ( $newres \neq s[p]$ ) && ( $p < |HMMp_+|$ );
8 end
9 return  $s_{synthetic}$ 

```

In this way, if the point of mutation belongs to a low entropy area of the HMM profile the emitted residue will be very similar to the main sequence (will have few mutations). We expect the emittance probability of the synthesized sequence with $HMMp_+$ to be higher than with $HMMp_-$, if not (very rarely), we synthesize another one or we decrease the value of N_m . The N_m parameter can adjust the radius of the neighborhood of the original sequences and the synthesized sequences. With larger values of N_m , the algorithm creates sequences that are phylogenetically farer away from main sequences and vice versa. We used another routine to find a suitable value of N_m . At first, in the minor class, we find the protein sequence which has the highest emission probability with the HMM profile of the minor class and consider it as root node. Then, we suppose the root node has been mutated to synthesize all other sequences in the minor class through the *newSequence* procedure of our algorithm. It means each sequence is a mutated replica of the root node sequence which is emitted by the HMM profile of the minor class. We gain the value of N_m for each sequence. Then, we get the average of all those values as N_m entry for the SPSO algorithm.

With each call of the SPSO algorithm, we double the minor class. As an example of random synthesizing of sequences, Fig. 1(upper) shows the phylogenetic tree of the original sequences and the synthesized sequences for the vasoactive intestinal polypeptide family of class B (9 out of 18 sequences were randomly selected). It is shown that the synthesized sequences of most original sequences have less distance to them than to other sequences. In that figure (lower) we see two synthetic sequences of $s1$ with different values of N_m . In the low entropy area of the HMM profile of that family we have less mutations.

2 Datasets

To evaluate the performance of our algorithm, we ran our experiments on a series of both real and artificial datasets, whose specification covers different complexity and allows us to fully interpret the results. We want to check its efficiency with different ratio of imbalance and complexity. Fig. 2 shows the

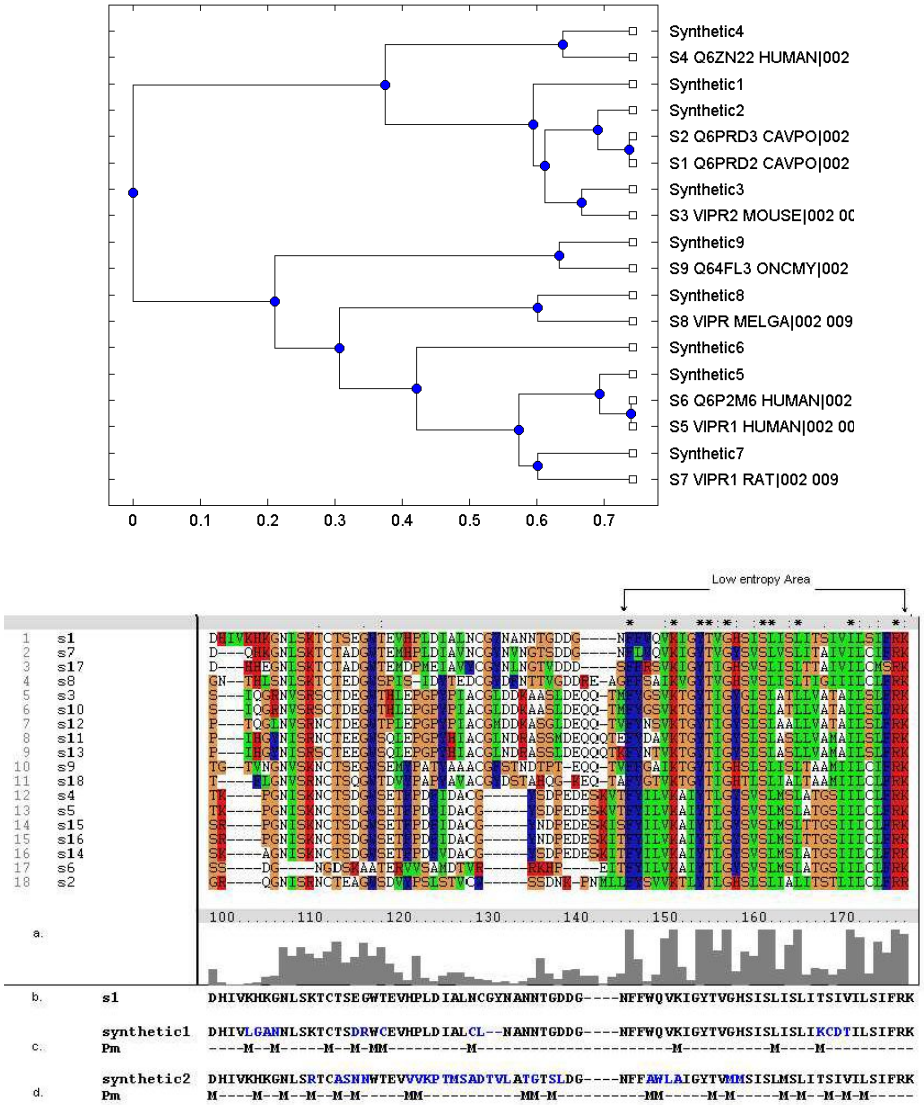


Fig. 1. The phylogenetic tree of the original and the synthesized sequences from the "vasoactive intestinal polypeptide" family of GPCRs (**upper**) and an example of the SPSO algorithm for sequences from the above family (**lower**). **a.** Multiple sequence alignment and low entropy area of that family **b.** A part of sequence s1. **c.** Synthetic sequence of s1 with $N_m=50$. **d.** Synthetic sequence of s1 with $N_m=100$.

pictorial representation of our datasets. In the first one, the distribution of the positive and negative data are completely different and they are separate from each other. With that distribution, we want to see, how the imbalance ratio affects the performance of the classifier by itself. The second one shows datasets

in which positive data are closer to negative data and there is an overlap between the minor and major classes. With this distribution, we can consider both the ratio of imbalance and overlap of the datasets in our study. The third one is a case where the minor class completely overlaps with the major class and we have fully overlapping data.

We used the G-protein coupled receptors (GPCRs) family as real data and then created artificial data based on it. G-protein coupled receptors (GPCRs) are a large superfamily of integral membrane proteins that transduce signals across the cell membrane [10]. According to the binding of GPCRs to different ligand types they are classified into different families. Based on GPCRDB (G protein coupled receptor database) [11] all GPCRs have been divided into a hierarchy of ‘class’, ‘subfamily’, ‘sub-sub-family’ and ‘type’. The dataset of this study was collected from GPCRDB and we used the latest dataset (June 2005 release, <http://www.gpcr.org/7tm/>). The six main families are: Class A (Rhodopsin like), Class B (Secretin like), Class C (Metabotropic glutamate/pheromone), Class D (Fungal pheromone), Class E (cAMP receptors) and Frizzled/Smoothed family. The sequences of proteins in GPCRDB were taken from SWISS-PROT and TrEMBL [12]. All six families of GPCRs (5300 protein sequences) are classified in 43 subfamilies and 99 sub-subfamilies.

If we want to classify GPCRs at the sub-subfamily level, mostly we have only a very low number of protein sequences as positive data (minor class) compared with others (major class). We chose different protein families from that level to cover all states of complexity and imbalance ratio discussed above (Fig. 2). In some experiments we made artificial data using those families and synthesized sequences from them (discussed later). We used numbers to show the level of family, subfamily and sub-subfamily. For example 001-001-002 means the sub-subfamily Adrenoceptors that belongs to subfamily of Amine (001-001) and class A (001).

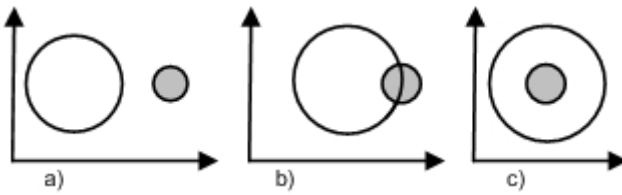


Fig. 2. Pictorial representation of the minor (shaded circle) and minor classes of our datasets

3 Experiments

We selected the peptide subfamily (001-002) of Class A (Rhodopsin-like) to classify its 32 families (or sub-subfamily level of class A). We built HMM profiles of all families and measured the probability of emission of sequences belonging

to each one by all HMM profiles. We saw that the emission probability of each sequence generated by the HMM profile of its own family is higher than that of almost all other families. So we can conclude that the distribution of the peptide subfamily in a suitable feature map can be considered as in Fig. 2.a. In this study, we used the local alignment kernel (LA kernel) [13] to generate vectors from protein sequences. It has been shown that the local alignment kernel has better performance than other previously suggested kernels for remote homology detection when applied to the standard SCOP test set [8]. It represents a modification of the Smith-Waterman score to incorporate sub-optimal alignments by computing the sum (instead of the maximum) over all possible alignments. We build a kernel matrix K for the training data. Each cell of the matrix is a local alignment kernel score between protein i and protein j . Then we normalize the kernel matrix via $K_{ij} \leftarrow K_{ij}/\sqrt{K_{ii}K_{jj}}$. Each family is considered as positive training data and all others as negative training data. After that the SVM algorithm with RBF kernel is used for training. For testing, we created feature vectors by calculating a local alignment kernel between the test sequence and all training data. The number of sequences in the peptide subfamily is in the range of 4 to 251 belonging to (001-002-024) and (001-002-008), respectively. Thus the *imbalance ratio* varies from $\frac{4}{4737}$ to $\frac{251}{4737}$. Fig. 3.a shows the result of SPSO oversampling for classification of some of those families. We see that this method can increase the accuracy and sensitivity of the classifier faced with highly imbalanced data without decreasing its specificity. The minority class was oversampled at 100%, 200%, 300%,..., 800% of its original size. We see that the more we increase the synthetic data (oversample) the better result we get, until we get the optimum value of 100%. It should be noted that after oversampling, the accuracy of classifiers for the major class didn't decrease.

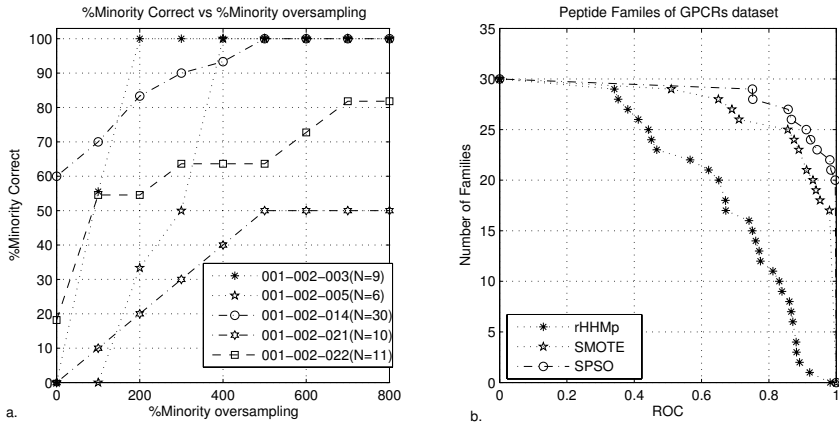


Fig. 3. a. %Minority correct for SPSO oversampling for some families of peptide sub-family (N number of sequences). b. Comparison of several methods for oversampling. The graph plots the total number of families for which a given method exceeds an ROC score threshold.

We compared our method with two other methods. The first one was SMOTE (Synthetic Minority Oversampling Techniques) [6] that operates in the feature space rather than in data space, so it works with all kind of data. The second comparison was done with randomly oversampling, in which we create random sequences by the HMM profile of each family. For this, like our method, we build a multiple alignment of the minor class sequences using ClustalW and then train a hidden Markov model profile with length of the created multiple alignment sequence. Then, we create random sequences by the HMM profile of each family. In this method we don't have enough control on the distribution of created random sequences. We call this method rHMMp in this paper.

In our study, we used the Bioinformatics Toolbox of MATLAB to create the HMM profiles of families and the SVMlight package [14], to perform SVM training and classification.

We used the Receiver Operating Characteristic (ROC) graphs [15] to show the quality of the SPSO oversampling technique. An ROC graph characterizes the performance of a binary classifier across all possible trade-off between the classifier sensitivity (TP_{rate}) and false positive error rates (FP_{rate}) [16]. The closer the ROC score is to 1, the better performance the classifier has. We oversampled each minority class with the three different methods noted above, until we got the optimum performance for each of them. At that point, we calculated the ROC score of all methods.

Fig. 3.b shows the quality of classifiers when using different oversampling methods. This graph plots the total number of families for which a given method exceeds an ROC score threshold. The curve of our method is above the curve of other methods and shows better performance. In our method and in SMOTE, the inserted positive examples have been created more accurately than random oversampling (rHMMp). Our method (SPSO) outperforms the other two methods especially for families in which we have a low number of sequences, although the quality of the SMOTE is comparable to the SPSO method.

To study the second and third representation of the dataset shown in Fig. 2 we had to create some sequences synthetically. At first we built the HMM profile of each family of the peptide families and then computed the probability score of each sequence when emitted not only by the HMM profile of its own family but also from all other families. The average of those scores for sequences of each family when emitted by each HMM profile can be used as a criterion for the closeness of the distribution of that family to other families and how much it can be represented by their HMM profiles. In this way we can find the nearest families to each peptide family. After that we synthesized sequences for each family through the *newSequ* procedure of the SPSO algorithm, provided that it is emitted by the HMM profile of another near family and not by its own HMM profile. So after each start position for mutation (Fig.1 (lower)) we have residues that are emitted by another HMM profile instead of its own HMM profile and there is an overlap for the distribution of synthesized sequences between those two families. The degree of overlapping can be tuned by the value of N_m (number of mutations). This dataset (original and new synthesized sequences) can be

Table 1. ROC scores obtained on the partially overlapping classes created from peptide families of GPCR dataset, by various methods. DEC = different error cost.

Partially overlapping classes- ROC scores				
minority class	# of sequences	SMOTE	DEC	SPSO
001 – 002 – 015'	16	0.863	0.943	0.951
001 – 002 – 016'	122	0.821	0.912	0.929
001 – 002 – 017'	68	0.854	0.892	0.884
001 – 002 – 018'	74	0.912	0.871	0.891
001 – 002 – 020'	86	0.972	0.975	0.984
001 – 002 – 021'	40	0.695	0.739	0.723
001 – 002 – 022'	44	0.725	0.762	0.751
001 – 002 – 023'	48	0.965	0.982	0.996
001 – 002 – 024'	8	0.845	0.834	0.865
001 – 002 – 025'	10	0.945	0.972	0.987
overall ROC-score		0.859	0.882	0.896

considered as partially overlapping dataset (Fig. 2.b). If we create more sequences using other HMM profiles the distribution of the dataset is fully overlapping (Fig. 2.c). To study the partially overlapping datasets, we selected 10 families of peptide families and built the synthesized sequences as noted above. To create the fully overlapping dataset, we performed that routine for each family using the HMM profile of three families near to the original family, separately.

We compared our oversampling technique with the SMOTE oversampling technique and the different error cost (DEC) method [4]. Tables 1 and 2 show the results. We see that in general SPSO outperforms the SMOTE and DEC methods, and the performance of the classifier with the SPSO oversampling technique in fully overlapped datasets is more apparent. When there is more overlapping between the minor and major classes, the problem of imbalanced

Table 2. ROC scores obtained on the Fully overlapping classes created from peptide families of GPCR dataset by various methods.

Fully overlapping classes- ROC scores				
minority class	# of sequences	SMOTE	DEC	SPSO
001 – 002 – 015"	32	0.673	0.680	0.724
001 – 002 – 016"	244	0.753	0.775	0.821
001 – 002 – 017"	136	0.672	0.652	0.643
001 – 002 – 018"	148	0.591	0.624	0.672
001 – 002 – 020"	172	0.763	0.821	0.858
001 – 002 – 021"	80	0.632	0.689	0.681
001 – 002 – 022"	88	0.615	0.812	0.854
001 – 002 – 023"	96	0.912	0.942	0.968
001 – 002 – 024"	16	0.716	0.768	0.819
001 – 002 – 025"	20	0.908	0.902	0.921
overall ROC-score		0.723	0.766	0.796

data is more acute. So the position of the inserted data in the minor class is more important and in our algorithm it has been done more accurately than in SMOTE method. With regard to the time needed for each algorithm, DEC has an advantage compared to our method, because in the oversampling technique the minor class, depending on the number of its instances, is oversampled up to 10 times (in our experiments) which increases the dimension of the kernel matrix. In contrast, in the DEC method choosing the correct cost of error for minority and majority classes is an important issue. One suggested method is to set the error cost ratio equal to the inverse of the imbalance ratio. But, based on our experiments (not shown here) that value is not always the optimum, and especially in partially and fully overlapped datasets we had instability of performance even with values near the optimal value. Based on our experiments in the well-separated imbalanced data the quality of DEC is very near to the SPSO method and for some experiments, even better, and we could find optimum value for error cost ratio simply. So perhaps with this kind of datasets one should prefer the DEC method. But with partially and fully overlapping data, we found that our oversampling method in general has better performance, and if it is used along with the DEC method, it not only increases the performance of the classifier but it also makes finding the value for the error cost ratio simpler. We also have more stability with values close to the optimum value of the error cost ratio. The graphs in Fig. 4.a and Fig. 4.b show the value of the ROC score of classifier for partially overlapped artificial sequences from the family of 001-002-024 (001 – 002 – 024') when the DEC method and DEC along with SPSO (400% oversampling) were applied. We see that when SPSO oversampling is used we have stability in ROC score values and after the optimum value, the ROC score does not change. The drawback is, that we again have to find the

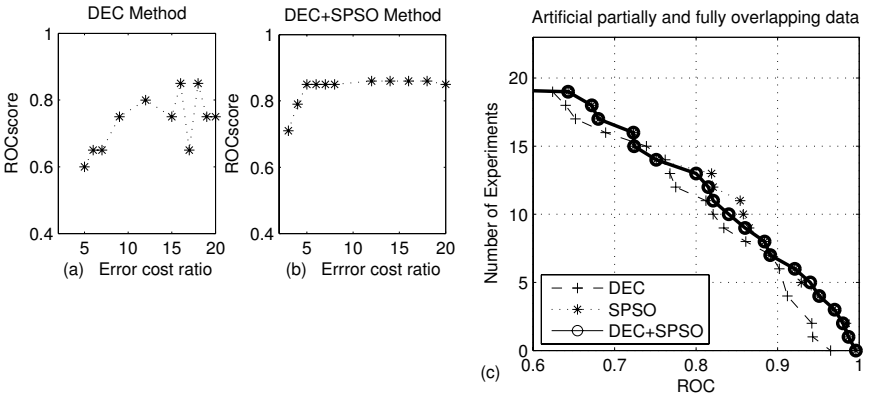


Fig. 4. (a) Comparison of the ROC score at different error cost ratios for artificial sequences of 001 – 002 – 024' in classifier with DEC method. (b) classifier with DEC + SPSO methods (400% oversampled). (c). Comparison of DEC, SPSO and DEC+SPSO methods for imbalanced data. The graph plots the total number of experiments of partially and fully overlapped imbalanced artificial data for which a given method exceeds an ROC score threshold.

best value for the error cost ratio and the rate of oversampling through the experiment by checking different values, but in less time compared to only the DEC method because of the stability that was shown in Fig. 4.b. We used that method for all partially and fully overlapping artificial data (Table 1 and 2). For each experiment we oversampled data in different rates and selected different values of error cost ratio until we got the best result. The results in Fig. 4.c show that for those kind of data the ROC scores of SPSO and DEC + SPSO are nearly the same. But in the second method (DEC + SPSO), we needed to oversample data less than in SPSO only method and we could find the best value of the error cost ratio sooner than in DEC only. With less rate of oversampling in SPSO we get less accurate results but we can compensate that with DEC.

4 Conclusion

In this work, we suggested a new approach of oversampling for the imbalanced protein data in which the minority class in the data space is oversampled by creating synthetic protein sequences, considering the distribution of the minor and major classes. This method can be used for protein classification problems and remote homology detection, where classifiers must detect a remote relation between unknown sequence and training data with an imbalance problem. We think that this kind of oversampling in kernel-based classifiers not only pushes the class separating hyperplane away from the positive data to negative data but also changes the orientation of the hyperplane in a way that increases the accuracy of classifier. We developed a systematic study using a set of real and artificially generated datasets to show the efficiency of our method and how the degree of class overlapping can affect class imbalance. The results show that our SPSO algorithm outperforms other oversampling techniques. In this paper, we also presented evidences suggesting that our oversampling technique can be used along with DEC to increase its sensitivity and stability. For further work, we hope to find an algorithm for finding the suitable rate of oversampling and error cost ratio when DEC and SPSO methods are used together.

References

1. C.Leslie, E. Eskin, A. Cohen, J. Weston, and W.S. Noble. Mismatch string kernel for svm protein classification. *Advances in Neural Information Processing System*, pages 1441–1448, 2003.
2. A. Al-Shahib, R. Breitling, and D. Gilbert D. Feature selection and the class imbalance problem in predicting protein function from sequence. *Appl Bioinformatics*, 4(3):195–203, 2005.
3. N. Japkowicz. Learning from imbalanced data sets: A comparison of various strategies. *In Proceedings of Learning from Imbalanced Data*, pages 10–15, 2000.
4. K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. *Proceedings of the International Joint Conference on AI*, pages 55–60, 1999.

5. G. Wu and E. Chang. Class-boundary alignment for imbalanced dataset learning. *In ICML 2003 Workshop on Learning from Imbalanced Data Sets II, Washington, DC*, 2003.
6. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence and Research*, 16:321–357, 2002.
7. C. Leslie, E. Eskin, and W. S. Noble. The spectrum kernel: A string kernel for svm protein classification. *Proceedings of the Pacific Symposium on Biocomputing*, page 564575, 2002.
8. H. saigo, J. P. Vert, N. Ueda, and T. akustu. Protein homology detection using string alignment kernels. *Bioinformatics*, 20(11):1682–1689, 2004.
9. J. D. Thompson, D. G. Higgins, and T. J. Gibson. Clustalw: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.
10. T. K Attwood, M. D. R. Croning, and A. Gaulton. Deriving structural and functional insights from a ligand-based hierarchical classification of g-protein coupled receptors. *Protein Eng.*, 15:7–12, 2002.
11. F. Horn, E. Bettler, L. Oliveira, F. Campagne, F. E. Cohhen, and G. Vriend. Gpcrdb information system for g protein-coupled receptors. *Nucleic Acids Res.*, 31(1):294–297, 2003.
12. A. Bairoch and R. Apweiler. The swiss-prot protein sequence data bank and its supplement trembl. *Nucleic Acids Res.*, 29:346–349, 2001.
13. J.-P.Vert, H. Saigo, and T.Akustu. *Convolution and local alignment kernel In B. Schoelkopf, K. Tsuda, and J.-P.Vert (Eds.), Kernel Methods in Computational Biology*. The MIT Press.
14. T. Joachims. Macking large scale svm learning practical. *Technical Report LS8-24*, Universitat Dortmund, 1998.
15. F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 423:203–231, 2001.
16. J. Swet. Measuring the accuracy of diagnostic systems. *Science.*, 240:1285–1293, 1988.

Markov Modeling of Conformational Kinetics of Cardiac Ion Channel Proteins

Chong Wang^{1,2}, Antje Krause¹, Chris Nugent², and Werner Dubitzky²

¹ University of Applied Sciences Wildau, Bahnhof str. 1,
15745 Wildau, Germany

{[cwang](mailto:cwang@igw.tfh-wildau.de),[akrause](mailto:akrause@igw.tfh-wildau.de)}@igw.tfh-wildau.de

² University of Ulster, Coleraine, Northern Ireland, UK

{[cd.nugent](mailto:cd.nugent@ulster.ac.uk),[w.dubitzky](mailto:w.dubitzky@ulster.ac.uk)}@ulster.ac.uk

Abstract. Markov modeling of conformational kinetics of cardiac ion channels is a prospective means to correlate the molecular defects of channel proteins to their electrophysiological dysfunction. However, both the identifiability of the microscopic conformations and the estimation of the transition rates are challenging. In this paper, we present a new method in which the distribution space of the time constants of exponential components of mathematical models are searched as an alternative to the consideration of transition rates. Transition rate patterns were defined and quasi random seed sequences for each pattern were generated by using a multiple recursive generator algorithm. Cluster-wide Monte Carlo simulation was performed to investigate various schemes of Markov models. It was found that by increasing the number of closed conformations the time constants were shifted to larger magnitudes. With the inclusion of inactivation conformation the time distribution was altered depending on the topology of the schemes. Further results demonstrated the stability of the morphology of time distributions. Our study provides the statistical evaluation of the time constant space of Markov schemes. The method facilitates the identification of the underlying models and the estimation of parameters, hence is proposed for use in investigating the functional consequences of defective genes responsible for ion channel diseases.

Keywords: Markov model, mutation, ion channel, conformation, time constant.

1 Introduction

One of the predominant characteristics of cardiac ion channels is their underlying conformational changes in response to the voltage difference across the cell membrane. Patch clamping recordings of the ion flux through single ion channels demonstrate the fluctuation from closed to open conformation states and vice versa with the dwell time distribution of each state manifesting a stochastic property [1]. However, macroscopic kinetic measurements of such channels indicate that ion channel protein molecules may undergo conformational changes

other than a single closed or an open conformation. For example, the rectification phenomenon observed in the cardiac delayed rectifier potassium channels [2] implies the existence of an inactivation conformation state.

In the heart, the rhythmic transition among the conformations in the cardiac ion channel proteins is responsible for the normal electrophysiological function of cardiac cells, hence the rhythmic heart beat. Any aberrant function in the channel protein molecules may lead to the alteration of the kinetic properties and the permeability of ion flux, thus resulting in abnormal electrophysiological function. The advances in linkage study have unraveled the heterogeneous molecular causes of the ion channel diseases like long QT syndrome (LQTS) [3], which can lead to lethal consequences. For example, the mutants in the gene *SCN5A*, encoding alpha subunits of sodium channel proteins, causes the channel either to exhibit a gain of function or a loss of function. The gain of function in a mutated sodium channel has been identified to cause LQTS subtype 3 [4]. Conversely, the loss of function in a defected sodium channel protein has been attributed to the Brugada syndrome [2], [3]. Over the last decade, the cellular functions of a large number of mutations found in cardiac ion channels have been experimentally characterized. More and more data has been produced through the co-expression of mutant and wide-type subunits in *Xenopus* oocyte expression system [2] or mammalian expression systems [5]. Integration of this information into the physiological models of cells and tissues and even organs becomes attractive in explaining and providing a means of understanding for the molecular mechanism of cardiac arrhythmia.

Markov modeling was first suggested to characterize kinetics of a sodium ion channel by French and Horn [6]. This approach received further attention due to Clancy and Rudy's pioneering work [4] in their attempts to model the mutant sodium channel and to investigate the effect of mutated proteins on the electrophysiological function at the cellular level. Since then, Markov modeling of conformational kinetics of ion channels has emerged as a prospective approach to correlating the molecular property of channel proteins to their electrophysiological function. Nevertheless, the characterization of various mutations residing in the different functional domains of a channel protein is a challenging concept. The big barrier lies in the identifiability of the underlying conformations in the specific mutant or variant, and the estimation of the transition rates among possible conformational states. The conformation changes in channel proteins are not directly observable, and only the ion flux or current conducted by a certain conformational state may be recorded instead [7]. Assume that an ion channel undergoes three conformational changes, two closed (C1, C2) and one open (O). Several Markov schemes may be postulated, e.g., two linear forms C1-C2-O and C1-O-C2 or a triangular form in which each conformation state is adjacent to the other two states. The question then arises as to which of them represents the underlying conformational kinetics in the optimal manner. Another problem is to estimate the transition rate function, especially the estimation of the initial parameters. This is considered complex due to the strong sensitivity of the common method used to fit the experimental data. An initially incorrect estimation

may either lead to non-convergence in the search for transition rate function or cause a very large computational load if convergence were to be achieved.

In this paper, we present a new method in an attempt to improve the quality of Markov modeling of conformational kinetics of ion channels. As an alternative to determining the transition rate directly, the time constants of the exponential components of the Markov schemes are searched. The transition rate patterns have been defined and quasi random seed sequences were generated to construct data spaces of transition rates for each scheme. The time constants of the exponential components were obtained through solving ordinary differential equations and performing a nonlinear square fit. The whole distribution of the time constants of a specific model was obtained through cluster-wide Monte Carlo simulation, whereby the synchronic communication functions of the message passing interface (MPI) [8] were used. Our data shows that the time constant distribution of a certain scheme is determined by the number of conformation states in a scheme, the topology of the scheme and more importantly the transition rate pattern. Our study sheds some light on identifying the conformation states with a Markov property and on explaining the underlying conformational kinetics of cardiac ion channels. In addition this approach provides the opportunity of accelerating the model choice and the interpretation of genetic mutations effecting cardiac physiological function.

2 Methods

2.1 Mathematical Formulism of a General Markov Scheme

The Markov property of the kinetics of an ion channel refers to a discrete-state stochastic process in which the conformational transition between discrete states is determined solely by the current state, independent of the history of transition and how long it has been in that state [7]. Since each state represents an energetically stable conformation of the channel, the time evolution of a channels kinetics can be formulated mathematically as a time-homogeneous Kolmogorov system as presented in Equation (1):

$$\frac{dp(t)}{dt} = Qp(t) , \quad (1)$$

where the element of $p(t)$ defines the probability of being in a state at any given time. Q is a structure matrix whose elements are given in Equation (2) and Equation (3):

$$q_{ij} = k_{ij}, \quad i \neq j , \quad (2)$$

$$q_{ii} = - \sum_{j \neq i} k_{ij} , \quad (3)$$

where k_{ij} is the transition rate (ms^{-1}) from the conformation state i to the conformation state j at a given time t . Equation (1) was solved by the Embedded Runge-Kutta Prince-Dormand method.

2.2 Exponential Components

From the numerical solution of Equation (1), we can derive the mean current flowing through an ensemble of channels. Assume that $P_o(t)$ is the channel open probability, hence the current across the membrane at time t can be expressed as in Equation (4):

$$I(t) = G_{max}P_o(t)(V - V_{rev}) , \quad (4)$$

where G_{max} is the maximum conductance and V_{rev} is the reversal potential.

To investigate the time constants of a certain Markov scheme, we used the general formulism of the exponential function as depicted in Equation (5):

$$I(t) = a_0 + \sum_i a_i e^{-t/\tau_i} , \quad (5)$$

where a_0 determines the start of the exponential or the value to which the exponential decays or rises, depending on the type of exponential data respectively. τ_i is the time constant to be searched. The reason for this to be considered is that the activation time course of a channel is measured by fitting the rise phase of a current trace as a single exponential function and the deactivation time course is obtained by fitting the decay of the tail current as a double exponential function. In the present work the Levenberg-Marquardt algorithm [9] was used to fit the representation of Equation (5) to the solution of a Markov scheme.

2.3 Transition Rate Pattern

Given a set of rate constant data, the transition matrix can be determined from Equation (2) and Equation (3). Accordingly, the time constant of Equation (5) can be obtained by fitting the function to the numerical solution of Equation (1). However, in order to obtain the whole space of time constants that a theoretical Markov scheme has the data space of rate constants must be first of all determined. In this study, each of the transition rates was defined as having a value belonging to either the following type as represented in Equation (6) and Equation (7):

$$k_{ij} \in (0, 1] , \quad (6)$$

or

$$k_{ij} \in (1, m], \quad m = 10^n, \quad n = 1, 2, 3, \dots \quad (7)$$

Thus, the data space of a scheme with N transition rates consists of 2^N patterns of data.

2.4 Monte Carlo Simulation on a Cluster

Based on the transition rate pattern defined above, we generated n -tuple of transition rate data as the input parameters to Equation (1). To achieve a good

property of randomness, the algorithm of multiple recursive generator (MRG) [9] was used to generate quasi random sequences of seeds using the entropy generator. In other words, the quasi random generators for each pattern was seeded with a distinct sequence of seeds. Because the generation was very computationally demanding, the whole computation was partitioned into small tasks, which were then submitted on SGE-managed cluster systems. The statistical evaluation of the time constant distribution was implemented using the MPI function: MPI_Bcast, MPI_Allreduce and MPI_Barrier [8]. The constraints as described in Equation (8) were introduced to accommodate for the fact that a Markov scheme will not have any physical meaning if any solution of the equation has a negative value.

$$\sum p_i(t) = 1, \quad p_i(t) > 0, \quad (8)$$

where $p_i(t)$ is the probability of ion channel proteins in the i th conformation at any given time as mentioned in Section 2.1.3.

3 Results

Through solving ordinary differential equations and performing nonlinear square fitting, the distribution of the time constants of seven Markov schemes was investigated. In contrast to the data space of transition rates which was generated by the MRG algorithm in a SGE-managed distributed system, the time constant space was obtained by Monte Carlo simulation in a parallel environment with an implementation of MPI.

3.1 Partition of Large Computational Tasks into Multiple Tasks

Quasi random seed sequences for seven schemes with a different number of transition rates were computed on a Linux cluster. Figure 1 shows the computational

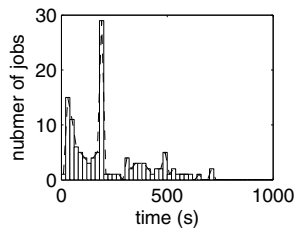


Fig. 1. Two high peaks indicated that most of the tasks were distributed to two types of nodes in the cluster, one with mean CPU time 178.43 ± 3.785 (s) (mean $\pm$ s.e.; $n=42$), another having mean CPU time 48.13 ± 3.662 (s) (mean $\pm$ s.e.; $n=38$)

time distribution for generating 1024 seed sequences, each consisting of 10 seeds of the type unsigned long integer. The total task was partitioned into 128 small tasks. The SGE scheduler distributed the tasks (jobs) to the nodes.

3.2 Closed Conformation and Time Constant Distribution

The effect of closed conformation was investigated by comparing the time constant distribution of schemes with different numbers of closed states. Through all the schemes, the results indicated that the increase in the number of closed states shifted the distribution to large magnitudes. Three schemes have been illustrated in Figure 2. As shown on the right side of the Figure, a scheme with a pathway of three closed conformations extended the time constant to the larger magnitude area in comparison with a scheme with less closed conformations. The elevation in the curve morphology suggests that the probability of obtaining a specific time constant from a scheme with more closed conformations can be enhanced when the time interval is fixed through all the schemes.

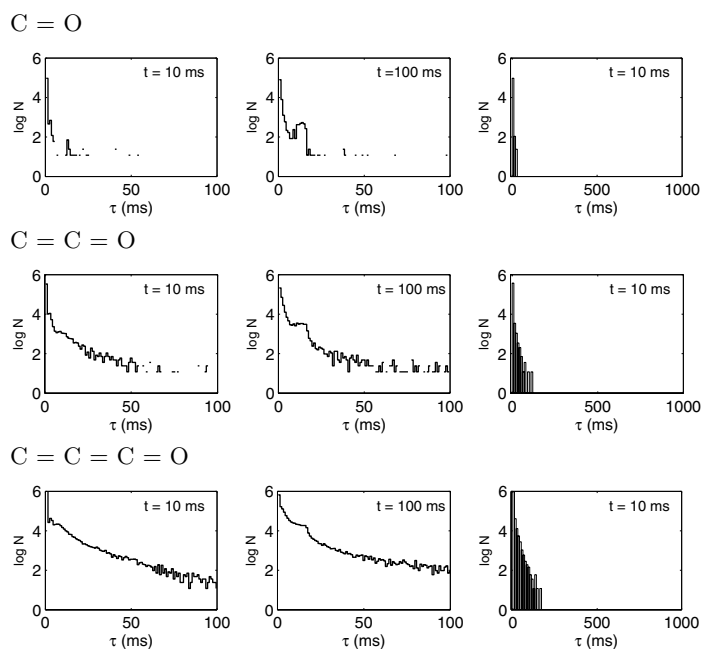


Fig. 2. Effect of closed conformation. The curves on the left hand side of the Figure were obtained with 10 (ms) elapsed time duration; the curves in the center were computed by 100 (ms) elapsed time; the curves on the right hand side were obtained with 10 (ms) elapsed time duration, but an evaluation interval of 1000 (ms).

3.3 Inactivation Conformation and Topology of Model Scheme

In contrast to closed conformation, the inactivation conformation affected the time constant distribution depending on the topology of the specific scheme. Figure 3 summarizes the results obtained from six schemes. Compared to the upper set of histograms, the middle set of histograms indicates that the inclusion of the inactivation conformation in a linear form extended the distribution to encompass

time constants with large magnitudes. Conversely, a comparison of the lower set of histograms with the upper ones demonstrates that the transition from closed conformations and open conformations to the inactivation conformation led to a reduction of time constants in the magnitude as well as in the probability.

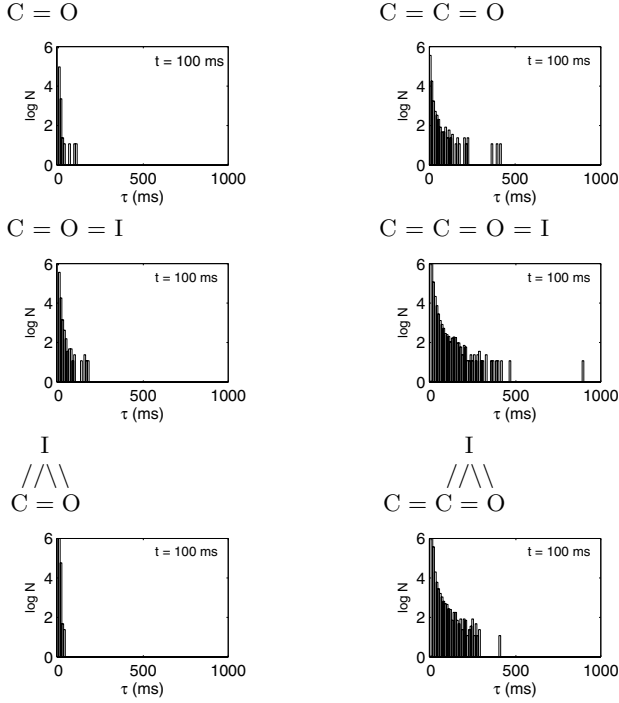


Fig. 3. Role of inactivation conformation in the determination of time constants. The upper set of histograms was obtained with schemes without inactivation states; the middle set of histograms was with inclusion of inactivation states in a linear form and the lower set of histograms with an inactivation state but in triangular form.

3.4 Role of Transition Rate Pattern

Following on the definition in the previous Methods Section, we know that a Markov scheme with N transition rates can have 2^N patterns of data. For example, the transition rate pattern for the scheme $C = C = O$ is 16. The time constant distribution of this scheme was statistically evaluated and represented in the form of histograms as shown in Figure 4. A and B in the figures denote a transition rate in either of the transition rate types defined in Equation (6) and Equation (7). As can be seen the distributions are significantly different. Given a time constant of 50 (ms) obtained from the kinetic experiment with a specific cardiac ion channel, the underlying transition rates with respect to the pattern AABA would be more probably identified than with a pattern like BABA or BBBA if the three-state scheme was used. The distributions with respect to the

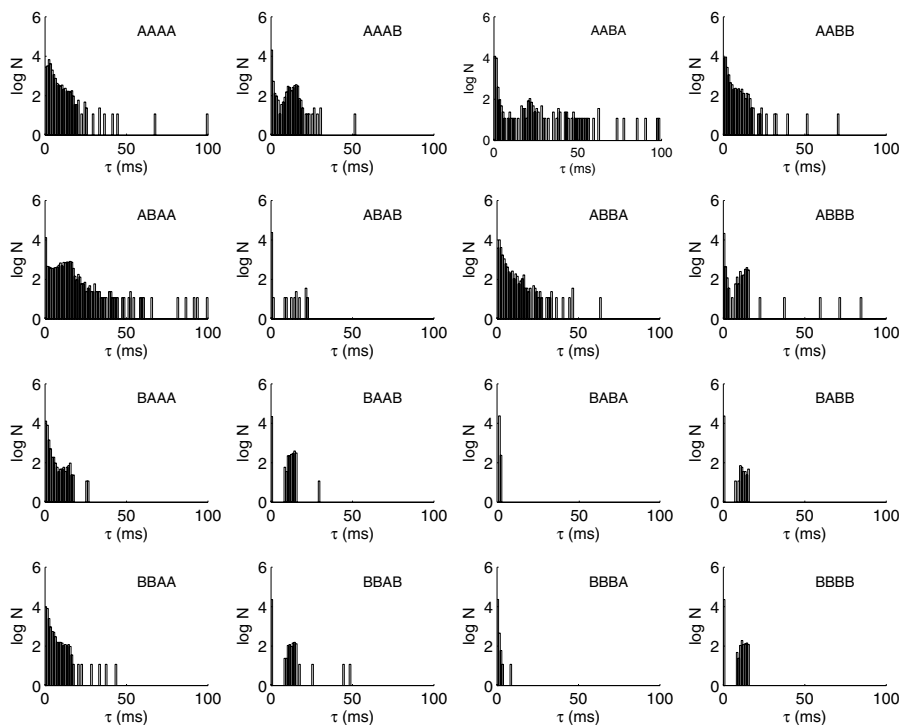


Fig. 4. The 16 patterns of the scheme $C = C = O$. A denotes a transition rate in the range (0, 1) and B denotes a transition rate in (1, m), $m=100$.

latter two patterns in Figure 4 show that their maximal time constants were much less than 50 (ms). In a transition rate search using the above three-state scheme, transition rates located in the latter two patterns led to either non-convergence or the violation of the constraint as defined in Equation (8). Similar results were also obtained from other Markov schemes investigated in this study.

3.5 Other Characteristics of the Time Constant Distribution

Beside the conformational states, we have also investigated the other factors which may affect the distribution of time constants of a specific scheme. In Figure 5, the upper two sets of curves show the time constant distributions of the scheme $C = C = O$ obtained in different conditions. Either in large evaluation intervals or in small evaluation intervals the morphology of distribution remained stable. The lower two sets of curves illustrate the time constant distributions of the scheme $C = C = O = I$. Compared to the set of curves obtained with an elapsed time interval 10 (ms), the set of curves with 100 (ms) time interval was elevated in an ordinate direction; however, there were no clear changes in the magnitude of time constants (abscissa direction). All these results suggested that the time constant distribution of a certain scheme has its inherent property,

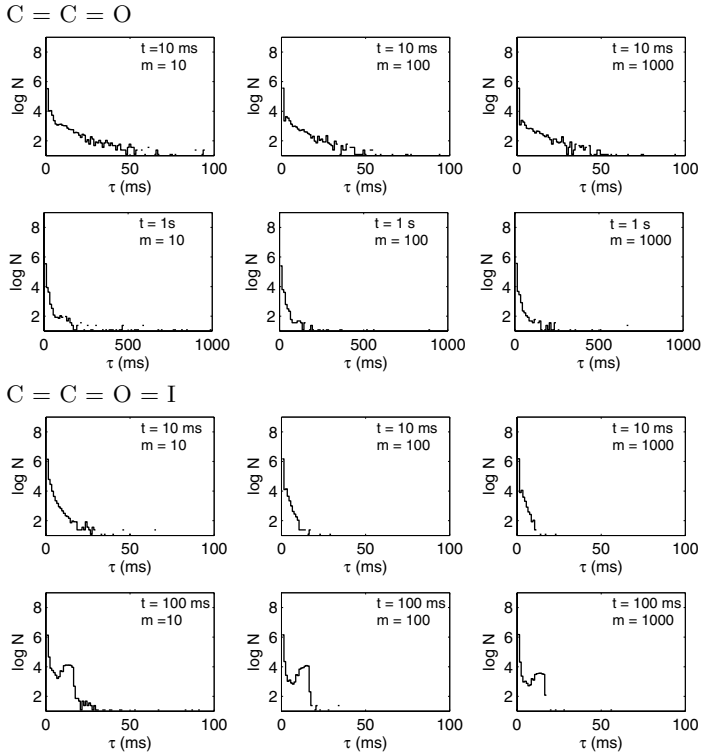


Fig. 5. Stable morphology of the time constant distribution. The upper two sets of curves were obtained with 10 (ms) and 1 (s) elapsed times respectively; the lower two sets of curves were the distributions with 10 (ms) and 100 (ms) elapsed times respectively. In both schemes, the maximum rate values ranged from 10 to 1000.

and increasing the elapsed time duration can lead to obtaining a more precise probability density function.

An example of the model proposal by means of the time constant distribution is given in Table 1. To cover a range of time constants to 500 (ms) a scheme with at least three closed (C) conformations must be used. For the rapid delayed rectifier potassium channel in the canine ventricle and the channel formed by coexpression of HERG and KCNE2, an inclusion of inactivation (I) state has more close time constants to cover a wider range from 50 (ms) to 600 (ms).

Table 1. Model proposal based on time constants

		Activation τ (ms)		Model proposal	Reference
		0 mV	+20 mV		
Native I_{kr}	Rabbit ventricle	400	200	3 C	11
	Canine ventricle	600	50	3 C + 1 I	
HERG-hKCNH2	In HEK 293 cells	583	288	3 C + 1 I	5

4 Discussion

The performance of the Monte Carlo simulation is dependent on the quality of the data space of transition rates. To assure the quality, a random number generator with a long period and a low serial correlation is preferred. In the current study, we used a fifth-order multiple recursive generator by L'Ecuyer et al. [9], [10] which has a period of about 10^{46} . Since the common generators show a significant correlation from sequence to sequence, if started with nearby seeds, n -tuple of transition rate data must be produced by generators which are maximally decorrelated at the seed level. To meet this requirement, we used the kernel-based entropy generator to calculate the seeds on Linux nodes. Nevertheless, we noted that reading the entropy generator is very slow, especially when a large number of random numbers must be computed on a single node. The process may hang if there are not sufficient entropies. Thus, we partitioned the overall computational task into many smaller tasks, and submitted to the SGE master, which managed the distribution and running of the jobs. Our SGE distribution system consists of four types of nodes, and among them two types, 3555.32 (MIPS) and 4734.97 (MIPS), accounting for 58% and 32% respectively. Thus, in this heterogeneous system, the jobs submitted should be distributed to and performed on the two dominant groups of nodes. Our results show two high peaks in the distribution of running times (Figure 1), which are in good agreement with our expectations.

The common approach used in modeling kinetics of ion channels is to extract rate constants for some Markov schemes by either a maximum likelihood method [1], [6] or histogram analysis of the life time duration of closed and open conformations [12]. The choice of a model scheme is somewhat empirical. Whether a scheme can characterize the underlying kinetics more closely remains unsolved. The estimates of parameters are also biased if the number of sweeps does not suffice. Importantly, the function characterization of mutant channel proteins are widely carried out in the heterologous expression systems such as *Xenopus* oocytes or transected HEK293 cell lines [2], [4], [5]. In such experimental studies, the sweeps of in flux through a single channel are not recorded; instead, current traces or tail currents in response to discrete conditional voltages are measured. In spite of the difficulty in identifiability due to the aggregation of conformational conductance, we noted that the time constants characterizing such dynamic processes as activation, inactivation of ion channels are more distinct from one another. Each mutated channel can have a discrete number of time constants in a physiological condition. Given that each scheme of Markov model has an inherent distribution space of the time constants and its exponential components, it is more practical to investigate the probability of the experimental rate values in the distribution space of a model and predict how the model scheme can represent the underlying conformational kinetics. The current study demonstrated that the distribution space of time constants of various Markov schemes had their inherent property and could be used to predict the underlying conformations by searching the probability of experimental kinetics. Our model proposal for the HERG-KCNE2 channel is consistent with that used by Mazhari

et. al. [5], and subsequently provided statistical evidence for the choice of a five state Markov model. In addition, for each scheme, pattern-specific distributions in our method were also provided, which can help to make a quick decision of initial parameters as shown in Section 3.4.

5 Conclusions

In summary, an important aspect of this study has been the investigation of the time constant distribution of Markov schemes, and to explore the possibility of this approach in support of the future functional interpretation of the mutant cardiac ion channel. Our data shows that increasing the number of closed conformations shifted the time constants to larger magnitudes; the inclusion of inactivation conformations altered the time distribution, depending on the topology of the schemes. The stability of the morphology of the distribution over a range of rate patterns suggests a specific Markov scheme has its inherent distribution of time constants. Our study provides the statistical evaluation of the time constant space of Markov schemes and the method facilitates the fast identification of the underlying models of cardiac channels and the choice of initial parameters. These results provide an indication that the approach is suitable for investigating functional consequences of defective genes responsible for cardiac ion channel diseases.

Acknowledgments. We thank Dr. Juergen Rose, Department of Bioinformatics, University of Applied Sciences Wildau, for his kind support in testing different packages of programs.

References

1. Horn, R., Vandenberg C.: Statistical Properties of Single Sodium Channels. *J. Gen. Physiol.*, 1984. 84(4): p. 505-534.
2. Sanguinetti, M.C., et al.: Spectrum of HERG K⁺-Channel Dysfunction in an Inherited Cardiac Arrhythmia. *PNAS*, 1996. 93(5): p. 2208-2212
3. Antzelevitch, C. and Shimizu, W.: Cellular Mechanisms Underlying the Long QT Syndrome. *Curr Opin Cardiol*, 2002. 17(1): p. 43-51.
4. Clancy, C.E. and Rudy Y.: Na⁺ Channel Mutation That Causes Both Brugada and Long-QT Syndrome Phenotypes: A Simulation Study of Mechanism. *Circulation*, 2002. 105(10): p. 1208-1213.
5. Mazhari, R., et al.: Molecular Interactions Between Two Long-QT Syndrome Gene Products, HERG and KCNE2, Rationalized by In Vitro and In Silico Analysis. *Circ Res*, 2001. 89(1) p. 33-38.
6. French, R.J., and Horn R.: Sodium Channel Gating Models, Mimics and Modifiers. *Annu Rev Biophys Bioeng.*, 1983. 12: p. 319-56.
7. Qin, F., Auerbach, A. and Sachs, F.: A Direct Optimization Approach to Hidden Markov Modeling for Single Channel Kinetics. *Biophys. J.*, 2000. 79(4): p. 1915-1927.
8. Gropp, W., Lusk, E., et al.: A High-Performance, Portable Implementation of the MPI Message-Passing Interface Standard. *Parallel Computing*. 1996. 22(6): p. 789-828.

9. Glassy M. et al: GNU Scientific Library Reference Manual (2nd Ed.). ISBN 0954161734.
10. L'Ecuyer P., Blouin F., and Coutre, R.: A Search for Good Multiple Recursive Random Number Generators. ACM Transactions on Modeling and Computer Simulation 1993. 3: p. 87-98.
11. Tseng, G.-N.: IKr: The hERG Channel. Journal of Molecular and Cellular Cardiology, 2001. 33(5): p. 835-849.
12. Scanley, B., et al., Kinetic Analysis of Single Sodium Channels from Canine Cardiac Purkinje Cells. J. Gen. Physiol., 1990. 95(3): p. 411-437.

Insulin Sensitivity and Plasma Glucose Appearance Profile by Oral Minimal Model in Normotensive and Normoglycemic Humans

Roberto Burattini¹, Fabrizio Casagrande¹, Francesco Di Nardo¹, Massimo Boemi², and Pierpaolo Morosini³

¹ Department of Electromagnetics and Bioengineering, Polytechnic University of Marche, 60131 Ancona, Italy

² Metabolic Disease and Diabetes Unit, Italian National Institute of Research and Care on Aging (INRCA-IRCCS), 60131 Ancona, Italy

³ Unit of Internal Medicine, "C. G. Mazzoni" General Hospital, 63100 Ascoli Piceno, Italy

Abstract. To evaluate the whole body insulin sensitivity, S_{IW} , and the rate of appearance, $R_a(t)$, of ingested glucose into plasma, the oral minimal model of glucose kinetics (OMM) was applied to insulinemia and glycemia data from six volunteer, normotensive and normoglycemic subjects, who underwent a 300 min oral glucose tolerance test (OGTT). Results from a full 22-sampling schedule (OGTT_{300/22}), were compared with two reduced schedules consisting of 12 samples (OGTT_{300/12}) and 11 samples (OGTT_{300/11}), respectively. The three protocols yielded virtually the same values of insulin sensitivity (denoted as S_{IW}^{22} , S_{IW}^{12} and S_{IW}^{11} , respectively) with intraclass correlation coefficients being not lower than 0.74. The $R_a(t)$ profiles reproduced by the OMM after application to the OGTT_{300/22} and the OGTT_{300/12} data were practically indistinguishable, whereas the profile obtained from the OGTT_{300/11} was characterized by a 33% overshoot at the 30th minute, followed by a 22% undershoot at the 60th minute. Our results suggest that the reduced OGTT_{300/12} is suitable to facilitate the assessment of insulin sensitivity and plasma glucose appearance profile in pathophysiological studies by the OMM.

Keywords: Glucose kinetics, insulin action, metabolic syndrome, OGTT.

1 Introduction

The measure of an insulin sensitivity index, defined as the ability of insulin to control glucose production and utilization, is of primary importance in the assessment of glucose regulatory system efficiency. In the effort to quantify insulin sensitivity in subjects presenting varying degrees of glucose tolerance, it is desirable to have a method which works during normal life condition, like a meal glucose tolerance test (MGTT), or under standardized oral glucose administration, like an oral glucose tolerance test (OGTT) [1]-[5]. Models for interpretation of data from oral tests, however, enhance the difficult problem of estimating the rate of appearance into plasma, $R_a(t)$, of glucose taken by mouth and absorbed from the gastrointestinal tract. Recently, this problem has been addressed by an oral minimal model (OMM) derived

from coupling the classic minimal model of glucose kinetics with a parametric description of $R_a(t)$ by a piecewise linear function [2]-[4].

The purpose of the present study was to test the applicability of the OMM in our clinical setting, in order to evaluate the whole body insulin sensitivity, S_{IW} , and the $R_a(t)$ from a 300 min OGTT applied to a group of six normotensive and normoglycemic patients, not affected by the metabolic syndrome, MS [6]. The perspective was to build control values useful to pursue future pathophysiological studies. Because easier and less costly applications of oral glucose tests, interpreted with the OMM, can be accomplished by optimization of the sampling schedule, we compared S_{IW} estimates and $R_a(t)$ predictions obtained from 22 blood samples with those obtained from reduced 12-sample and 11-sample protocols.

2 Methodology

2.1 Subjects and Protocol

This study included six subjects (3 men and 3 women with mean age of 46.2 (SD=13.2) yr and body mass index, BMI, of 24.2 (SD=3.4) kg/m². All they gave informed consent to the procedures approved by the Ethics Committee.

Special care was exercised in recruiting these subjects, in order to avoid hypertension and MS as confounding factors in evaluating insulin sensitivity. The MS was defined according to the Adult Treatment Panel III criteria [6], that is, presence of 3 or more of the following criteria: fasting glucose ≥ 6.1 mmol/l (110 mg/dl), waist circumference >102 cm in men and >88 cm in women; triglycerides ≥ 1.695 mmol/l (150 mg/dl); HDL cholesterol <1.036 mmol/l (40 mg/dl) in men and <1.295 mmol/l (50 mg/dl) in women; blood pressure $\geq 130/85$ mmHg. Subjects were excluded from participation if they had a past history of diabetes mellitus or had a fasting glycemia ≥ 110 mg/dl. Thus, our recruited subjects were normoglycemic, normotensive and showed no more than two of the remaining three ATP III criteria.

Each subject underwent an OGTT for measurement of insulinemia and glycemia data, starting at 8:30 a.m., after overnight fast. One fasting blood sample was taken immediately before a 75 g glucose load administration ($t=0$), and 21 more blood samples were taken at minutes 10, 20, 30, 45, 60, 75, 90, 105, 120, 135, 150, 165, 180, 195, 210, 225, 240, 255, 270, 285 and 300, after glucose load (OGTT_{300/22}) [1]. Results from this protocol were compared with those obtained from a reduced OGTT_{300/12} protocol, consisting of 12 blood samples drawn at minutes 0, 10, 20, 30, 45, 60, 90, 120, 150, 180, 240 and 300. A further 11-sample protocol (OGTT_{300/11}), that differs from the OGTT_{300/12} for the omission of the 45th-minute sample, was also considered [4].

2.2 Model Identification

The oral minimal model (OMM) [2]-[4] uses the changes in plasma glucose and insulin concentrations observed in basal state and after the administration of 75 g glucose load, to derive insulin sensitivity, S_{IW} , and the rate of appearance into plasma of ingested glucose, $R_a(t)$. Model equations are [2]:

$$\frac{dG(t)}{dt} = -[S_G + X(t)] \cdot G(t) + S_G \cdot G_b + \frac{R_a(\alpha, t)}{V} \quad (1)$$

$$\frac{dX(t)}{dt} = -p_2 \cdot X(t) + p_2 \cdot S_I \cdot [I(t) - I_b] \quad (2)$$

with initial conditions:

$$G(0) = G_b \quad (3)$$

$$X(0) = 0 \quad (4)$$

In equations 1 and 2, $G(t)$ is glycemia (mg/dl), $I(t)$ is insulinemia ($\mu\text{U/ml}$), the suffix "b" denotes basal values, $X(t)$ is insulin action on glucose production and disposal (min^{-1}), V is distribution volume (dl/kg), S_G is the fractional (i. e. per unit distribution volume) glucose effectiveness (min^{-1}), S_I and p_2 are free model parameters. Namely, S_I is fractional insulin sensitivity [$\text{min}^{-1}/(\mu\text{U} \cdot \text{ml}^{-1})$], while p_2 is the rate constant that accounts for the dynamics of insulin action (min^{-1}). Whole body insulin sensitivity, S_{IW} , is given by the product $S_I \cdot V$ [$\text{min}^{-1} \cdot \text{dl} \cdot \text{kg}^{-1}/(\mu\text{U} \cdot \text{ml}^{-1})$].

The time course of the rate of appearance of glucose into plasma, $R_a(\alpha, t)$, was described by a piecewise-linear function with break-points, t_i , at 0, 15, 30, 60, 90, 120, 180 and 300 min, and seven coefficients, α_i ($\text{mg} \cdot \text{min}^{-1}/\text{kg}$) [2]-[4]:

$$R_a(\alpha, t) = \alpha_{i-1} + \frac{\alpha_i - \alpha_{i-1}}{t_i - t_{i-1}} (t - t_{i-1}); \quad (5)$$

$$t_{i-1} \leq t \leq t_i; \quad i = 1, \dots, 7$$

$$R_a(\alpha, t) = 0; \quad \text{otherwise} \quad (6)$$

In equations 1, 5 and 6, α denotes the parameter vector $[\alpha_1, \alpha_2, \dots, \alpha_7]^T$.

2.3 Parameter Estimation

The OMM was identified by fitting to $G(t)$ data. Because V is not identifiable and S_G not uniquely identifiable [2], V and S_G were given the mean reference values previously determined from model independent, dual tracer OGTT studies [3], i.e., $V = 1.34$ dl/kg, $S_G = 0.028$ min^{-1} , while model parameters p_2 , S_I , and $\alpha_1 \dots \alpha_6$ were estimated in each individual. Under the assumption that the rate of appearance wears out within the 300th minute, α_7 was posed equal to 0. Maximum a posteriori Bayesian estimation was employed for p_2 , assumed normally distributed with mean 0.012 min^{-1} and $\text{SD}\% = 10\%$. As in [2] the area under the $R_a(t)$ profile was assumed to be equal to the total amount of ingested glucose, D , multiplied by the fraction, f , of it that is actually absorbed. In accordance with Dalla Man et al. [3], f was given the value of 0.87.

Parameter estimation was accomplished making use of a weighted non-linear least-squares estimation technique implemented by the SAAM II software (SAAM Institute, University of Washington, Seattle, WA) [7], [8]. Weights were optimally chosen, i.e., equal to the inverse of the variance of the glucose measurement errors. The errors associated with total glucose measurements were assumed to be normally distributed random variables with zero mean and a constant percent coefficient of

variation (CV%) equal to 2%. To assess the goodness of glucose data fit we analyzed the weighted residuals, i.e. the differences between the data and model-predicted values, multiplied by the square root of a weight proportional to the reciprocal of datum SD [7]-[9].

2.4 Reproducibility Analysis

Intraclass correlation coefficient [10] (irrespective of CV%) was used to evaluate the agreement of S_{IW} estimates from the full and reduced sample OGTTs.

3 Results

Mean fasting glycemia and insulinemia were 79.3 (SD=10.9) mg/dl and 5.03 (SD=1.79) μ U/ml, respectively. Waist circumference was 86.2 (SD=12.8) cm. Serum triglycerides and HDL cholesterol were 77.0 (SD=55.5) mg/dl and 55.7 (SD=15.4) mg/dl, respectively.

Mean experimental glycemia, $G(t)$, and insulinemia, $I(t)$, levels as a function of time are shown in Fig. 1.

Mean weighted residuals of glucose data fit, as obtained from the full and the reduced sample protocols, are shown in Fig. 2.

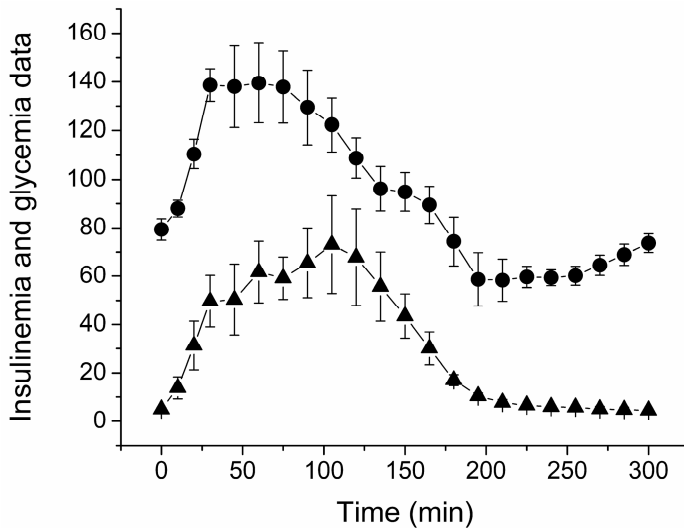


Fig. 1. Glycemia (full circles, mg/dl) and insulinemia (full triangles, μ U/ml) as a function of time, measured during our OGTT_{300/22} protocol. Values are expressed as means $\pm$ SE over six subjects.

Mean estimates of whole body insulin sensitivity (S_{IW}^{22} from the OGTT_{300/22}, S_{IW}^{12} from the OGTT_{300/12} and S_{IW}^{11} from the OGTT_{300/11}) are compared in Fig. 3. Means

and 95% confidence intervals (CI), expressed as $10^{-4} \cdot \text{min}^{-1} \cdot \text{dl} \cdot \text{kg}^{-1} / (\mu\text{U} \cdot \text{ml}^{-1})$, were $S_{IW}^{22} = 13.6$ (9.3-17.8), $S_{IW}^{12} = 13.5$ (8.8-18.3) and $S_{IW}^{11} = 13.4$ (8.3-18.4). Mean CV% of the estimates did not exceed 3%. Intraclass correlation coefficient was 0.80 between S_{IW}^{22} and S_{IW}^{12} , and 0.74 between S_{IW}^{22} and S_{IW}^{11} .

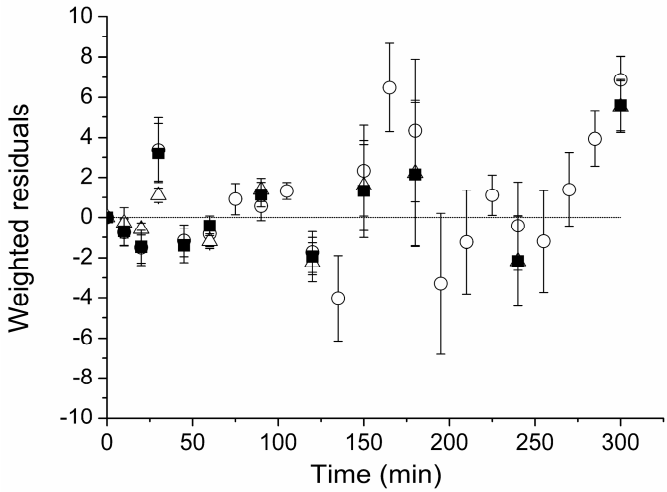


Fig. 2. Mean ($\pm$ SE) weighted residuals computed after fitting the OMM to glycemia data from the full OGTT_{300/22} (open circles), OGTT_{300/12} (full squares) and OGTT_{300/11} (open triangles).

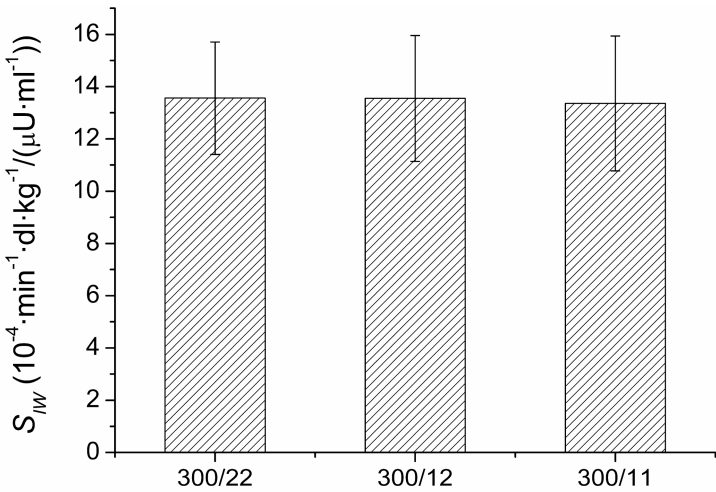


Fig. 3. Mean ($\pm$ SE) estimates of insulin sensitivity, S_{IW} , obtained from fitting the OMM to glycemia data from OGTT_{300/22}, OGTT_{300/12} and OGTT_{300/11} protocols

Means and 95% confidence intervals (CI) of p_2 estimates, expressed as $10^{-3} \cdot \text{min}^{-1}$, were $p_2^{22}=18.8$ (13.6-24.1), $p_2^{12}=16.0$ (12.9-19.0) and $p_2^{11}=16.0$ (12.9-19.1). Mean CV% did not exceed 6%.

Table 1. Mean estimates of α_i coefficients. Values are expressed as means $\pm$ SE over six subjects. Measure units of α_i are $\text{mg} \cdot \text{min}^{-1} / \text{kg}$. Precision of parameter estimates (CV% $\pm$ SE), computed as SD of parameter estimate divided by the parameter estimate and multiplied by 100, is given in round brackets.

α_i	OGTT _{300/22}	OGTT _{300/12}	OGTT _{300/11}
α_1 (CV%)	4.5 $\pm$ 0.7 (6.7 $\pm$ 1.5)	4.4 $\pm$ 0.7 (7.0 $\pm$ 1.6)	4.1 $\pm$ 0.7 (9.5 $\pm$ 3.2)
α_2 (CV%)	4.9 $\pm$ 1.2 (24 $\pm$ 17)	4.7 $\pm$ 1.1 (15 $\pm$ 6.8)	6.3 $\pm$ 1.0 (13 $\pm$ 2.7)
α_3 (CV%)	6.0 $\pm$ 1.0 (6.2 $\pm$ 1.2)	5.4 $\pm$ 0.9 (11 $\pm$ 2.8)	4.4 $\pm$ 1.7 (13 $\pm$ 1.3)
α_4 (CV%)	6.1 $\pm$ 0.8 (4.6 $\pm$ 0.7)	5.9 $\pm$ 0.8 (7.8 $\pm$ 1.5)	6.3 $\pm$ 0.9 (8.0 $\pm$ 1.9)
α_5 (CV%)	7.0 $\pm$ 1.0 (2.2 $\pm$ 0.2)	6.5 $\pm$ 1.0 (4.0 $\pm$ 0.4)	6.2 $\pm$ 0.9 (4.1 $\pm$ 0.4)
α_6 (CV%)	0.8 $\pm$ 0.3 (14 $\pm$ 4.1)	1.3 $\pm$ 0.2 (11 $\pm$ 2.0)	1.3 $\pm$ 0.2 (11 $\pm$ 2.1)

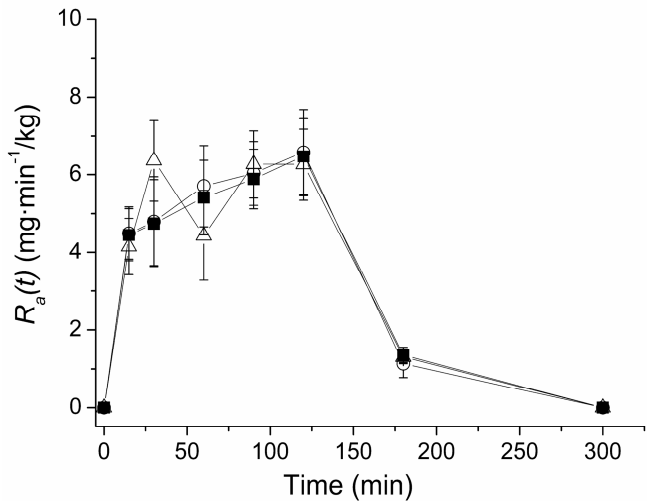


Fig. 4. Mean ($\pm$ SE) values of plasma glucose appearance profiles, $R_a(t)$, predicted in our subjects by the OMM, after application to OGTT_{300/22} (open circles), OGTT_{300/12} (full squares) and OGTT_{300/11} (open triangles)

The estimates of α_i coefficients, $i = 1, \dots, 6$ (α_7 was assumed equal to zero as explained in Methods) from the three different OGTT data sets are reported in

Table 1. The related $R_a(t)$ profiles are shown in Fig. 4. The profiles obtained from the OGTT_{300/22} and the OGTT_{300/12} were practically indistinguishable, whereas the profile obtained from the OGTT_{300/11} was characterized by a 33% overshoot at the 30th minute, followed by a 22% undershoot at the 60th minute.

4 Discussion

It has been demonstrated in previous studies by others [2]-[4] that the oral glucose minimal model (OMM) provides reliable measurements of the overall effect of insulin to stimulate glucose disposal and inhibit glucose production from both meal glucose tolerance test (MGTT) and oral glucose tolerance test (OGTT). For routinely clinical use, an OGTT is more appealing than a MGTT to assess glucose tolerance. The possibility of predicting plasma glucose appearance profile, $R_a(t)$, improves the appeal of the OMM-OGTT method [2]-[4] over other available model based and empiric methods that allow estimation of insulin sensitivity alone [5]. In the present study a full 300 min, 22 blood samples OGTT protocol was administered to test the OMM behaviour in six healthy subjects. Our careful selection criteria limited the number of participants in this study, but allowed us to control for confounding effects of hypertension [5], [11]-[13] and metabolic syndrome, MS, defined according to the ATP III criteria [6].

An index of whole body insulin sensitivity, S_{IW} , and the $R_a(t)$ profile were evaluated after assuming population averages of $S_G=0.028 \text{ min}^{-1}$ and $V=1.34 \text{ dl/kg}$, in accordance with Dalla Man et al. [3]. It has been shown previously that this assumption does not introduce appreciable bias in the estimation of insulin sensitivity [3]. Comparison of different sampling schedules, denoted as OGTT_{300/22}, OGTT_{300/12} and OGTT_{300/11}, was performed here with the aim of finding an easier and less costly protocol that does not bias the estimates of S_{IW} and the predictions of $R_a(t)$.

A substantive concordance was found among the S_{IW} estimates obtained from all three sampling schedules, as shown in Fig. 3, and as supported by intraclass correlation coefficients not lower than 0.74. The mean value of S_{IW}^{11} obtained here from the OGTT_{300/11} ($13.4 \pm 2.5, 10^{-4} \cdot \text{min}^{-1} \cdot \text{dl} \cdot \text{kg}^{-1} / (\mu\text{U} \cdot \text{ml}^{-1})$) is in good agreement with the mean value of $13.7 \pm 0.87 10^{-4} \cdot \text{min}^{-1} \cdot \text{dl} \cdot \text{kg}^{-1} / (\mu\text{U} \cdot \text{ml}^{-1})$ reported by Dalla Man et al. [4] after application of the OMM to nondiabetic subjects submitted to the same OGTT_{300/11} protocol.

A discrepancy characterized the reproduction, in our subjects, of the $R_a(t)$ profile obtained from the OGTT_{300/11}, compared with that obtained from the full OGTT_{300/22} protocol. The former showed a spurious oscillation, with a maximum at the 30th minute and a minimum at the 60th minute, that was caused by the lack of information on insulinemia and glycemia at the 45th minute. Indeed, when this measurement was included (thus giving rise to the OGTT_{300/12} protocol) the OMM-predicted $R_a(t)$ profile resulted practically indistinguishable from that obtained from the full OGTT_{300/22}.

In Fig. 5 the mean $R_a(t)$ profile obtained in our subjects from the OGTT_{300/12} is compared with the mean profile reported by Dalla Man et al. [4] (their Fig. 5A) for nondiabetic subjects with a large spectrum of glucose tolerance. Differences between the two profiles can be explained by some variability, among humans, in gluco-incretins control of insulin secretion and rate of intestinal glucose absorption [1], [14], [15].

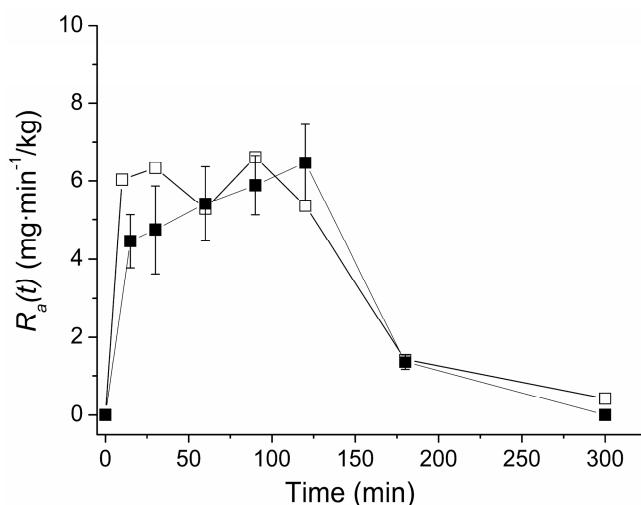


Fig. 5. Mean ($\pm$ SE) values of the rate, $R_a(t)$, of glucose appearance into plasma, as predicted in our subjects by the OMM after fitting to OGTT_{300/12} data (full squares), is compared with the mean $R_a(t)$ profile (open squares) reported by Dalla Man et al. [4].

In conclusion, in our clinical setting, the OMM-OGTT_{300/12} method appears a potentially useful tool for evaluation of insulin sensitivity and plasma glucose appearance profile in pathophysiological studies, from relatively low-cost OGTT.

Acknowledgments. This work was supported in part by the Italian Ministry of Instruction, University and Research (PRIN-COFIN 2004 grant to R. Burattini).

References

1. Breda, E., Cavaghan, M.K., Toffolo, G., Polonsky, K.S., Cobelli, C.: Oral glucose tolerance test minimal model indexes of β -cell function and insulin sensitivity. *Diabetes*, Vol. 50 (2001) 150-158
2. Dalla Man, C., Caumo, A., Cobelli, C.: The oral glucose minimal model: estimation of insulin sensitivity from a meal test. *IEEE Trans Biomed Eng.* Vol. 49 (2002) 419-429
3. Dalla Man, C., Yarasheski, K. E., Caumo, A., Robertson, H., Toffolo, G., Polonsky, K. S., Cobelli, C.: Insulin sensitivity by oral glucose minimal models: validation against clamp. *Am J Physiol Endocrinol Metab*, Vol. 289 (2005) E954-E959
4. Dalla Man, C., Campioni, M., Polonsky, K. S., Basu, R., Rizza, R. A., Toffolo, G., Cobelli, C.: Two-hour seven-sample oral glucose tolerance test and meal protocol: minimal model assessment of beta-cell responsivity and insulin sensitivity in nondiabetic individuals. *Diabetes*, Vol. 54 (2005) 3265-3273
5. Di Nardo, F., Casagrande, F., Boemi, M., Fumelli, P., Morosini, P., Burattini, R.: Insulin resistance in hypertension quantified by oral glucose tolerance test: comparison of methods. *Metabolism*, Vol. 55 (2006) 143-150

6. Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults: Executive Summary of the Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III): JAMA, Vol. 285 (2001) 2486-2497
7. SAAM Institute: SAAM II User Guide. Seattle, WA (1997)
8. Barrett, P. H., Bell, B. M., Cobelli, C., Golde, H., Schumitzky, A., Vicini, P., Foster, D. M.: SAAM II: Simulation, Analysis, and Modeling Software for tracer and pharmacokinetic studies. *Metabolism*, Vol. 47 (1998) 484-492
9. Carson, E.R., Cobelli, C., Finkelstein, L.: *The Mathematical Modeling of Metabolic and Endocrine Systems*. New York: Wiley (1983)
10. Fleiss, J. L.: *Statistical Methods for Rates and Proportion*. New York: Wiley (1980) 212-236
11. Ferrannini, E., Buzzigoli, G., Bonadonna, R., Giorico, M.A., Oleggini, M., Graziadei, L., Pedrinelli, R., Brandi, L., Bevilacqua, S.: Insulin resistance in essential hypertension. *N Engl J Med*, Vol. 317 (1987) 350-357
12. Corry D.B., Tuck M.L.: Glucose and insulin metabolism in hypertension. *Am J Nephrol*, Vol. 16 (1996) 223-236
13. Burattini, R., Di Nardo, F., Boemi, M., Fumelli, P.: Deterioration of insulin sensitivity and glucose effectiveness with age and hypertension. *Am J Hypertens*, Vol. 19 (2006) 98-102
14. Kieffer, T. J., Habener, J. F.: The glucagon-like peptides. *Endocr Rev*, Vol. 20 (1999) 876-913
15. Wackers-Hagedoorn, R.E., Priebe, M.G., Heimweg, J.A., Heiner, A.M., Englyst, K.N., Holst, J.J., Stellaard, F., Vonk, R.J.: The rate of intestinal glucose absorption is correlated with plasma glucose-dependent insulinotropic polypeptide concentrations in healthy men. *J Nutr*, Vol. 136 (2006) 1511-1516

Dynamic Model of Amino Acid and Carbohydrate Metabolism in Primary Human Liver Cells

Reinhard Guthke¹, Wolfgang Schmidt-Heck¹, Gesine Pless², Rolf Gebhardt³,
Michael Pfaff⁴, Joerg C. Gerlach^{2,5}, and Katrin Zeilinger²

¹ Leibniz Institute for Natural Product Research and Infection Biology –
Hans Knoell Institute, Beutenbergstr. 11a, D-07745 Jena, Germany
{reinhard.guthke,wolfgang.schmidt-heck}@hki-jena.de
<http://www.hki-jena.de>

² Division of Experimental Surgery, Charité Campus Virchow, University Medicine Berlin,
Augustenburger Platz 1, D-13353 Berlin, Germany

³ Institute of Biochemistry, Medical Faculty, University Leipzig, Johannisallee 30,
D-04103 Leipzig, Germany
rgebhardt@medizin.uni-leipzig.de

⁴ BioControl Jena GmbH, Wildenbruchstr. 15, D-07745 Jena, Germany
michael.pfaff@biocontrol-jena.com

⁵ Depts of Surgery and Bioengineering, McGowan Institute for Regenerative Medicine,
University of Pittsburgh, PA, USA
{katrin.zeilinger,joerg.gerlach}@charite.de

Abstract. Human liver cell bioreactors are used in extracorporeal liver support therapy. To optimize bioreactor operation with respect to clinical application an understanding of the central metabolism is desired. A two-compartment model consisting of a system of 48 differential equations was fitted to time series data of the concentrations of 18 amino acids, ammonia, urea, glucose, galactose, sorbitol and lactate, measured in the medium outflow of seven liver cell bioreactor runs. Using the presented model, the authors predict an amino acid secretion from proteolytic activities during the first day after inoculation of the bioreactor with primary liver cells. Furthermore, gluconeogenetic activities from amino acids and/or protein were predicted.

1 Introduction

Liver cell bioreactors are being developed and used for temporary extracorporeal liver support [1, 2]. Primary human liver cells isolated from discarded human organs are inoculated and cultured in these bioreactors. The bioreactor provides a valuable tool to analyze the dynamics of the physiological and molecular interactions of liver cells under standardized conditions closely reflecting the situation in the natural organ. The multi-compartment bioreactor analyzed here consists of three interwoven, independent capillary membrane systems. The liver cells are cultivated in the inter-capillary space ('liver cell compartment'). Two of the capillary membrane systems (in the following aggregated to the 'perfusion compartment') provide decentralized plasma flow and the third one provides oxygen supply to the liver cells. The bioreactor is integrated into a perfusion system that enables monitoring and control of system conditions (see Figure 1).

In previous work, data mining and pattern recognition methods were applied to extract knowledge from bioreactor operation data [3-5]. Using fuzzy clustering and rule extraction methods, the kinetics of galactose and urea over the first three culture days were found to be the best single predictors for the bioreactor's long term performance [3]. In addition, kinetic patterns of the amino acid metabolism over the first six culture days were described by different network models (correlation networks, Bayesian networks, differential equation systems) [4, 6]. However, the dynamic models [4, 6] do not include the carbohydrate metabolism, in particular that of galactose. Furthermore, these models do not adequately describe the kinetics of lysine (LYS), which is characterized by a temporal maximum one day after inoculation, and that of alanine (ALA), which is characterized by a temporal minimum at the third day (see Table 3 in [4]). The adequate dynamic modeling of LYS and ALA together with the modeling of the profiles of glucose (GLC), galactose (GAL), sorbitol (SOR) and lactate (LAC) is the goal of the present paper.

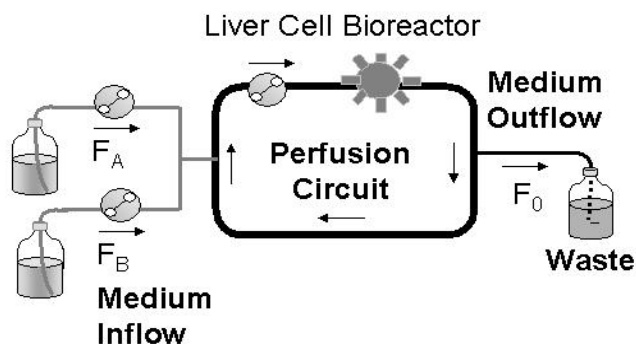


Fig. 1. Scheme of the liver cell bioreactor with the perfusion circuit, the two inflow streams and the outflow stream. Measured data were acquired from the waste.

2 Data and Methods

Primary human liver cells were isolated from human donor livers ($n = 7$) that were not suitable for transplantation due to organ injury. After isolation, cells were cultured within bioreactors over at least seven days. However, only the first six days of culture were investigated in the present study, i.e. the stand-by phase prior to clinical use of bioreactors for liver support therapy. Concentrations of free amino acids, ammonia, urea, GLC, GAL, SOR and LAC in the culture effluates (Waste) were measured and analyzed.

Two compartments of the multi-compartment capillary membrane bioreactor system are considered: The 'liver cell compartment' with the volume $V_2 = 600$ mL contains the liver cells in the extra-capillary space. The 'perfusion compartment' with the volume $V_1 = 900$ mL supplies a stream through the inside of the capillaries. To this perfusion compartment two time-variant inflow streams are added with the flow rates $F_A(t)$ and $F_B(t)$ as defined by Eq. (1). $F_A(t)$ follows a step function from $F_{A1} = 150 \text{ mL}\cdot\text{h}^{-1}$ down to $F_{A2} = 50 \text{ mL}\cdot\text{h}^{-1}$ switching at time $t_A = 1$ d. $F_B(t)$ switches

from $F_{B1} = 0$ up to $F_{B2} = 1 \text{ mL} \cdot \text{h}^{-1}$ at the time $t_B = 3 \text{ d}$. The outflow rate $F_0(t)$ to the waste equals the sum of both inflow rates.

$$\begin{aligned} F_A(t) &= \begin{cases} F_{A1} & \text{for } t < t_A \\ F_{A2} & \text{for } t \geq t_A \end{cases}, \\ F_B(t) &= \begin{cases} F_{B1} & \text{for } t < t_B \\ F_{B2} & \text{for } t \geq t_B \end{cases}, \\ F_0(t) &= F_A(t) + F_B(t). \end{aligned} \quad (1)$$

The inflow rate $F_A(t)$ carries 18 amino acids, NH₃, GLC, GAL, SOR and LAC at the concentrations c_{Ai} (equal to $c_{i,j,0}$ for run j , see Table 1 for $j = 2$). The inflow rate $F_B(t)$ carries only the amino acid aspartate (ASP) at the concentration $c_{B15} = 1,500 \text{ } \mu\text{mol} \cdot \text{L}^{-1}$, i.e. $c_{Bi} = 0$ for all $i \neq 15$. The concentrations $c_{0i}(t)$ in the outflow stream are assumed to be in steady-state prior to the inoculation with cells, i.e. $c_{0i}(0) = c_{Ai}$. They may be considered to describe the response of the medium to the inoculation of the bioreactor with cells. For more detailed information, see [6].

A data set with the elements $c_{i,j,k}$ for 24 kinetic variables ($i = 1, \dots, 24$) at seven time-points t_k ($k = 0, 1, \dots, 6$; $t_k = 0, \dots, 6 \text{ d}$) was analyzed for seven bioreactor runs ($j = 1, \dots, 7$; called ‘high performance runs’ in [4-6]). For instance, Table 1 shows the concentrations $c_{i,2,k}$ of the 24 compounds measured for the second run (R2). The kinetics of run R2 is used as a representative example because its kinetics is most similar to the mean kinetics averaged over the seven runs investigated here.

Amino acid concentrations $c_{i,j,k}$ ($i = 1, \dots, 18$) were determined up to the third day daily and every third day afterwards ($t_k = 0, 1, 2, 3, 6 \text{ d}$). Concentrations of NH₃, UREA, GAL, SOR GLC, and LAC ($i = 19, \dots, 24$) were measured daily. The mean kinetics of GAL, SOR GLC and LAC ($c_{i,j,k}$ versus t_k for $i = 21, \dots, 24$; $k = 0, \dots, 6$) averaged over the seven runs R_j ($j = 1, \dots, 7$) are shown in Figure 2. The mean kinetics of the other 20 compounds ($i = 1, \dots, 20$) were shown in [6].

The samples for the measurement of the $c_{i,j,k}$ were taken from the waste of the liver cell bioreactor system (see Figure 1) which was emptied daily after accumulation of the outflow stream over the period of 24 h. These data measured in the bioreactor outflow were compared with the sliding mean $c^*_i(t)$ of the model simulated kinetics $c_{0,i}(t)$ in order to calculate the mean square error mse defined by Eq. (2.1).

$$mse = \frac{1}{I \cdot K} \sum_{i=1}^I \frac{1}{(\max_k c_{i,j,k})^2} \sum_{k=1}^K (c_{i,j,k} - c^*_i(t_k))^2 \quad (2.1)$$

$$\text{with} \quad c^*_i(t) = \int_{t-24\text{h}}^t c_{0,i}(t) F_0(t) dt / \int_{t-24\text{h}}^t F_0(t) dt \quad (2.2)$$

The kinetics $c_{0,i}(t)$ were obtained from the simulation of the Eqs. (3.1-3.14) together with the initial values $c_i(0) = c_{0,i}(0) = c_{Ai} = c_{i,j,0}$ (see Table 1 for $j = 2$). The variables c_i are the ‘hidden’ (not measured) concentrations in the ‘liver cell compartment’. The differential equations were solved using a Runge-Kutta 4th order algorithm. The parameter identification by minimizing the mse was performed using a simplex search method. MATLAB tools (The MathWorks, Inc., Natick, MA) were used for all calculations.

Table 1. Data $c_{i,2,k}$ [$\mu\text{mol} \cdot \text{L}^{-1}$] measured at time t_k before ($t_0 = 0$) and 1, 2, 3, 4, 5 and 6 days after inoculation of the run R2 (as a representative member of the seven runs); n.d. – not determined. The initial values $c_{i,2,0}$ at $t_0 = 0$ equals the concentrations c_{Ai} in the inflow stream fed with the flow rate $F_A(t)$ according to Eq. (1).

i		$t_0 = 0$ d	$t_1 = 1$ d	$t_2 = 2$ d	$t_3 = 3$ d	$t_4 = 4$ d	$t_5 = 5$ d	$t_6 = 6$ d
1	LEU	2258	2239	1379	1111	n.d.	n.d.	1019
2	HIS	940	462	374	355	n.d.	n.d.	361
3	ARG	2604	n.d.	n.d.	26	n.d.	n.d.	136
4	VAL	987	1046	644	497	n.d.	n.d.	420
5	TRP	464	272	103	55	n.d.	n.d.	27
6	PHE	1300	745	578	532	n.d.	n.d.	628
7	ILE	500	418	134	100	n.d.	n.d.	109
8	TYR	2215	1330	1489	1379	n.d.	n.d.	1509
9	LYS	327	858	481	360	n.d.	n.d.	244
10	MET	259	120	25	3	n.d.	n.d.	16
11	SER	861	533	212	121	n.d.	n.d.	139
12	GLY	1957	1503	910	529	n.d.	n.d.	284
13	THR	859	779	210	87	n.d.	n.d.	77
14	ALA	1825	2153	1302	691	n.d.	n.d.	1824
15	ASP	284	177	42	25	n.d.	n.d.	2817
16	ASN	168	179	74	56	n.d.	n.d.	89
17	GLU	265	1160	774	589	n.d.	n.d.	1585
18	GLN	687	621	671	690	n.d.	n.d.	732
19	NH3	19	265	75	69	81	81	109
20	UREA	0	6835	9335	9669	10836	1286	13169
21	GAL	5049	599	444	694	696	697	699
22	SOR	5376	845	340	307	351	395	439
23	GLC	8159	10490	9879	8270	6882	5994	5661
24	LAC	500	7437	6782	6094	5950	6838	6349

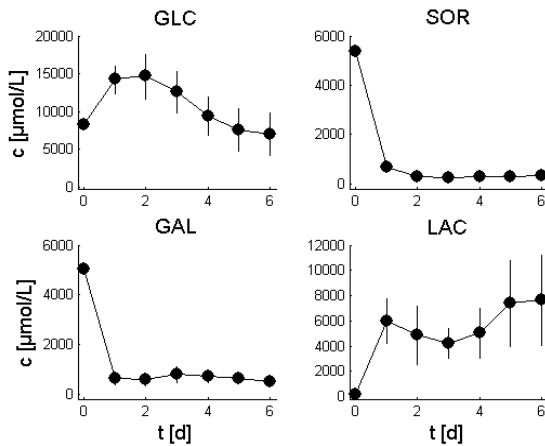


Fig. 2. Mean kinetics (•, $\pm$ standard deviation – shown as vertical lines) of the concentrations of glucose, sorbitol, galactose and lactate averaged over seven runs

3 Results and Discussion

The differential equation system (3.1)-(3.14) was developed to describe the kinetics of the 24 measured time series data $c_{i,j,k}$. Individual models for each of the seven runs were identified.

$$\frac{dc_{0,i}}{dt} = F_A(t)/V_1 \cdot c_{Ai} + F_B(t)/V_1 \cdot c_{Bi} - F_0(t)/V_1 \cdot c_{0,i} - p_0/V_1 \cdot (c_{0,i} - c_i) \quad (3.1)$$

for $i = 1, \dots, 24$

$$\frac{dc_i}{dt} = p_0/V_2 \cdot (c_{0,i} - c_i) - p_i \cdot c_i + p_{AAi} \cdot pp, \quad \text{for } i = 1, \dots, 13 \quad (3.2)$$

$$\frac{dc_{14}}{dt} = p_0/V_2 \cdot (c_{0,14} - c_{14}) - p_{14} \cdot c_{14} + p_{AA14} \cdot pp + p_{29} \cdot c_{15} \quad (3.3)$$

$$\begin{aligned} \frac{dc_{15}}{dt} = & p_0/V_2 \cdot (c_{0,15} - c_{15}) + p_{16} \cdot c_{16} + p_{17} \cdot c_{17} + p_{AA15} \cdot pp - \\ & - (p_{15} \cdot c_{19} + p_{18} + p_{19} \cdot c_{19}/(p_{20} + c_{15}) + p_{29}) \cdot c_{15}, \end{aligned} \quad (3.4)$$

$$\frac{dc_{16}}{dt} = p_0/V_2 \cdot (c_{0,16} - c_{16}) + p_{15} \cdot c_{15} \cdot c_{19} - p_{16} \cdot c_{16} + p_{AA16} \cdot pp, \quad (3.5)$$

$$\frac{dc_{17}}{dt} = p_0/V_2 \cdot (c_{0,17} - c_{17}) + \sum_{i=1}^9 p_i \cdot s_i \cdot c_i + p_{14} \cdot s_{14} \cdot c_{14} + p_{18} \cdot c_{15} + \quad (3.6)$$

$$+ p_{26} \cdot c_{18} - (p_{17} + p_{25} \cdot c_{19} + p_{28}) \cdot c_{17} + p_{AA17} \cdot pp, \quad (3.7)$$

$$\frac{dc_{18}}{dt} = p_0/V_2 \cdot (c_{0,18} - c_{18}) + p_{25} \cdot c_{17} \cdot c_{19} - p_{26} \cdot c_{18} + p_{AA18} \cdot pp, \quad (3.8)$$

$$\begin{aligned} \frac{dc_{19}}{dt} = & p_0/V_2 \cdot (c_{0,19} - c_{19}) + \sum_{i=10}^{13} p_i \cdot s_i \cdot c_i + p_{28} \cdot (1 - g(t)) \cdot c_{17} + p_{16} \cdot c_{16} + p_{26} \cdot c_{18} - \\ & - (p_{19} \cdot c_{15}/(p_{20} + c_{15}) + p_{15} \cdot c_{15} + p_{25} \cdot c_{17}) \cdot c_{19}, \end{aligned} \quad (3.9)$$

$$\frac{dc_{20}}{dt} = p_0/V_2 \cdot (c_{0,20} - c_{20}) + p_{19} \cdot c_{15}/(p_{20} + c_{15}) \cdot c_{19}, \quad (3.10)$$

$$g(t) = 0 \quad \text{for } t < 3d \quad \text{else} \quad g(t) = p_{27} \quad (3.11)$$

$$pp(t) = p_{33} \quad \text{for } t < 1d \quad \text{else} \quad pp(t) = 0 \quad (3.12)$$

$$\frac{dc_i}{dt} = p_0/V_2 \cdot (c_{0,i} - c_i) - p_i \cdot c_i, \quad \text{for } i = 21 \text{ and } 22 \quad (3.13)$$

$$\frac{dc_{23}}{dt} = p_0/V_2 \cdot (c_{0,23} - c_{23}) + p_{21} \cdot c_{21} + p_{22} \cdot c_{22} - p_{23} \cdot c_{23} + p_{31} \cdot c_{15}/(p_{20} + c_{15}) \cdot c_{19}, \quad (3.14)$$

$$\frac{dc_{24}}{dt} = p_0/V_2 \cdot (c_{0,24} - c_{24}) + 2 \cdot p_{30} \cdot p_{23} \cdot c_{23} + p_{32} \cdot c_{15}/(p_{20} + c_{15}) \cdot c_{19} - p_{24} \cdot c_{24}$$

The model with 48 variables takes into account two compartments, i.e. the concentrations $c_{0,i}$ of compound i in the ‘perfusion compartment’ and the concentrations c_i of the compound i in the ‘liver cell compartment’ with $i = 1, \dots, 24$. The initial values of the variables are $c_i(0) = c_{0,i}(0) = c_{Ai} = c_{i,j,0}$ assuming a steady state before inoculation of the bioreactor with liver cells. The Eqs. (3.1) describe the dynamics of the components i in the perfusion compartment by four terms: The first and second term represent the fresh medium inflow with the volumetric rates $F_A(t)$ and $F_B(t)$, respectively, into the perfusion compartment with the volume V_1 . The volumetric rates $F_A(t)$ and $F_B(t)$ are specified by Eqs. (1). The third term in the Eqs.

(3.1) denotes the outflow from the perfusion compartment into the waste with the volumetric rate $F_0(t)$. The last term describes the flow (diffusion) between the perfusion and the liver cell compartment. The corresponding parameter $p_0 = 50$ L/h has been determined in previous studies by model fit to the mean kinetics [6]. In the present identification scheme we fixed the value p_0 for the fit of the other parameters to the individual runs throughout. The reason for this modified identification scheme is, first, that the variability of p_0 could not be explained by any reason in [6] and, second, that no correlations were found between the parameter values p_0 identified in [6] specifically for the seven runs using amino acid kinetics on the one hand and the parameter values p_0 identified specifically here using GAL, SOR, GLC and LAC according to the Eqs. (3.12)–(3.14) on the other hand (data not shown).

The Eqs. (3.2)–(3.10) are the same as used and described in detail in [6], however with two novel features introduced in the present study:

(i) The added term $p_{29}c_{15}$ in the Eqs. (3.3) and (3.4) formulates the reaction from ASP to ALA catalyzed by aspartate decarboxylase (EC 4.1.1.12). The term allows for simulating the observed increase of ALA from the third day after inoculation and, consequently, the temporal minimum of ALA at $t = 3$ d.

(ii) The added term $p_{AAi}pp$ in the Eqs. (3.2)–(3.7) formulates the secretion of amino acids due to proteolysis. The newly introduced term allows the simulation of the temporal maximum of LYS and other amino acids at the first day after inoculation. The function $pp(t)$ defined by Eq. (3.11) describes the secretion of LYS from proteolysis. The parameter p_{33} was identified individually for the seven runs fitting the kinetics simulated by Eq. (3.2) for $i = 1, 4, 7$ and 9 to the time series data of LEU, VAL, ILE, and LYS, which are characterized by a temporal maximum at the first day after inoculation. The result of model fit for LYS is illustrated in Figure 3 (top). Figure 3 (bottom) and Table 3 show the parameters p_9 (decay rate of LYS) and p_{33} (proteolysis rate scaled by the secretion rate of LYS) identified by fitting $c_i^*(t_k)$ as obtained from the Eqs. (2.2), (3.1), (3.2) and (3.11) to randomly disturbed data $c_{ij,k}$ ($i = 1, 4, 7, 9; j = 1, \dots, 7; k = 0, \dots, 6$). The parameter values p_{33} that quantify the proteolysis rate were found to be significantly smaller in the bioreactor cultures R1, R2, R6 and R7, which were inoculated with liver cells from male donors, than the values p_{33} identified for the bioreactors R3, R4 and R5, which were inoculated with cells from female donors (p -value = 0.017).

The parameters p_{AAi} (see Table 2) that describe the protein decomposition profile, i.e., the contribution of proteolysis to the release of amino acids, were estimated commonly for all runs by linear regression over the seven runs ($j = 1, \dots, 7$) using Eq. (4) correlating the concentrations of LYS at the first day (i.e., $c_{9,j,1}$) with the respective value of the other amino acids ($i = 1, \dots, 18$). b_i is an offset parameter. The parameters p_{AAi} were not identifiable individually for the seven runs. We identified a common parameter set p_{AAi} for the seven cultivations.

$$c_{i,j,1} = p_{AAi} \cdot c_{9,j,1} + b_i \quad (4)$$

As shown in Table 2 for 15 of 18 amino acids (with exception of HIS, GLN, and ARG), the correlation coefficients r_{AAi_LYS} are greater than 0.8, the mean of r_{AAi_LYS} is 0.9 and the p -values for testing the hypothesis of no correlation are smaller than 3%. Figure 4a illustrates the high correlation according Eq. (4) for ASN, i.e., $i = 16$ and $j = 1, \dots, 7$.

Table 2. Stoichiometric coefficients s_i for nitrogen (used in Eqs. (3.6) and (3.8) as well as the correlation coefficient r_{AAi_LYS} together with the p -value for testing the hypothesis of no correlation and the parameters p_{AAi} resulting from the linear regression according to Eq. (4); ‘n.d.’ – not determined because not enough data available

i	Name	Symbol	s_i	r_{AAi_LYS}	p	p_{AAi}
1	Leucine	LEU	1	0.91	0.004	1.40
2	Histidine	HIS	3	0.17	0.714	0.18
3	Arginine	ARG	4	n.d.	n.d.	0.00
4	Valine	VAL	1	0.89	0.008	1.19
5	Tryptophan	TRP	2	0.92	0.003	0.45
6	Phenylalanine	PHE	1	0.89	0.007	1.05
7	Isoleucine	ILE	1	0.84	0.017	0.62
8	Tyrosine	TYR	1	0.82	0.025	1.15
9	Lysine	LYS	2	1.00	-	1.00
10	Methionine	MET	1	0.90	0.005	0.15
11	Serine	SER	1	0.95	0.0008	0.77
12	Glycine	GLY	1	0.95	0.001	1.55
13	Threonine	THR	1	0.88	0.009	0.78
14	Alanine	ALA	1	0.93	0.003	1.81
15	Aspartate	ASP	1	0.81	0.025	0.31
16	Asparagine	ASN	2	0.97	0.0003	0.25
17	Glutamate	GLU	1	0.87	0.011	1.27
18	Glutamine	GLN	2	0.12	0.788	0.06
19	Ammonia	NH3	1	-	-	-
20	Urea	UREA	2	-	-	-
21	Galactose	GAL	0	-	-	-
22	Sorbitol	SOR	0	-	-	-
23	Glucose	GLC	0	-	-	-
24	Lactate	LAC	0	-	-	-

Table 3. The gender of the liver donor (f – female, m – male); the model fit error mse (see Eq. (2.1)); the model parameters p_9 and p_{33} identified by the fit of $c^*_i(t)$ calculated by the Eqs. (2.2), (3.1), (3.2) and (3.11) for $i = \{1, 4, 7, 9\}$ to the measured data $c_{i,j,1}$ of the runs $j = 1, \dots, 7$; confidence intervals of the parameters (identified using randomized data; see Figure 3 for p_9)

j	$Gende$ r	mse	p_9 [d ⁻¹]	$[p_{9\_low}, p_{9\_high}]$	p_{33} [μmol L ⁻¹ d ⁻¹]	$[p_{33\_low}, p_{33\_high}]$
1	m	0.221	0.9979	[0.8348, 1.1592]	9290	[8733, 9933]
2	m	0.205	0.4521	[0.3446, 0.5747]	7708	[7163, 8285]
3	f	0.417	2.6025	[2.3363, 2.9263]	1051	[801, 1359]
4	f	0.392	1.9681	[1.7427, 2.1872]	1924	[1649, 2259]
5	f	0.263	2.3576	[2.1141, 2.6734]	1855	[1444, 2294]
6	m	0.314	1.3202	[1.1669, 1.4865]	5241	[4864, 5767]
7	m	0.469	0.2120	[0.0891, 0.3147]	13550	[12673, 14127]

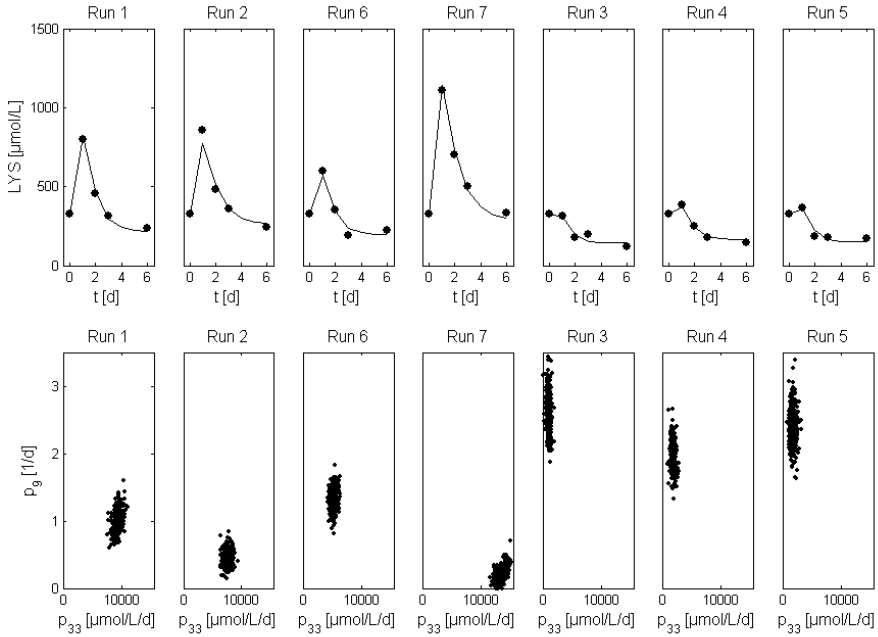


Fig. 3. Top: LYS concentrations $c_{9,j,k}$ measured versus the time t_k ($\bullet$, $k = 0, \dots, 6$; $j = 1, \dots, 7$) and the kinetics (lines) $c_i^*(t)$ obtained using the Eqs. (2.2), (3.1), (3.2) and (3.11) for $i = 9$ after fitting of the parameters p_9 and p_{33} (Table 3). Bottom: Parameters p_9 and p_{33} identified by repeated (250 times) fitting $c_i^*(t)$ to randomly disturbed data (measured data $c_{i,j,1}$ multiplied by $(1+\varepsilon)$; noise ε with the zero mean and standard deviation $\sigma = 0.05$).

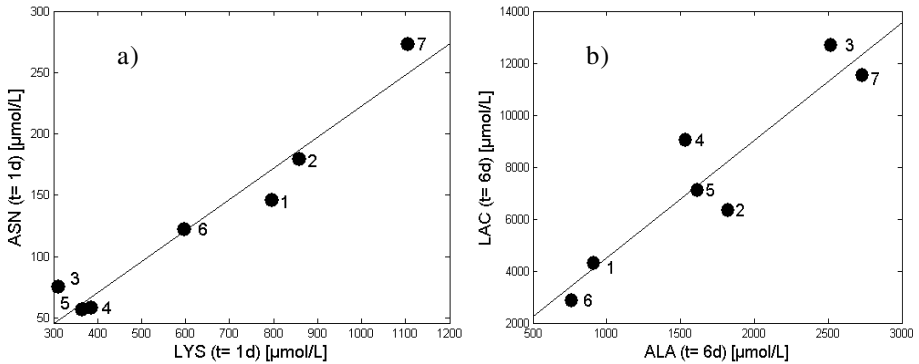


Fig. 4. a) Concentrations of ASN ($c_{16,j,1}$) versus concentrations of LYS ($c_{9,j,1}$) measured one day after inoculation for the seven runs (run numbers indicated; correlation coefficient $r_{AA3\_LYS} = 0.97$). Line according to Eq. (4) with $p_{AA16} = 0.25$ and $b_{16} = -30$ $\mu\text{mol/L}$. b) Concentrations of LAC ($c_{24,j,6}$) versus concentrations of ALA ($c_{14,j,6}$) measured six days after inoculation for the seven runs (correlation coefficient $r = 0.93$).

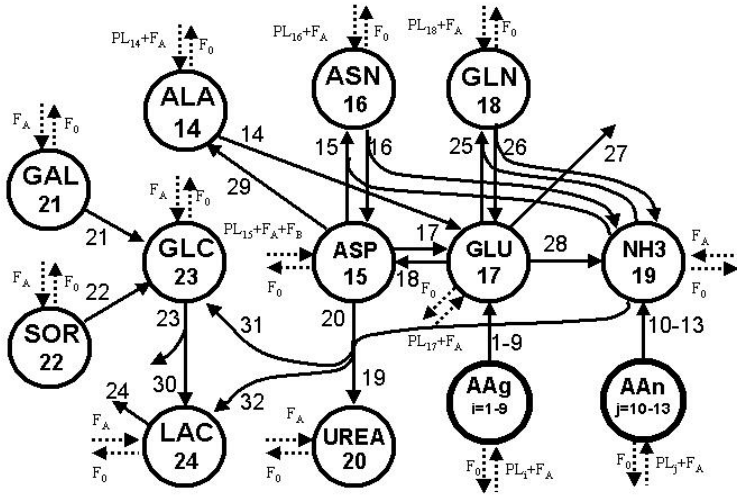


Fig. 5. Structure of the model described by the Eqs. (3.2)-(3.14). The numbers within the circles denote the indices i of variables c_i (see Table 1). AAg and AAn denote individual sets of amino acids {LEU, HIS, ARG, VAL, TRP, PHE, ILE, TYR, LYS} and {MET, SER, GLY, THR}, respectively. The numbers besides the arrows denote the indices of the model parameters p_m .

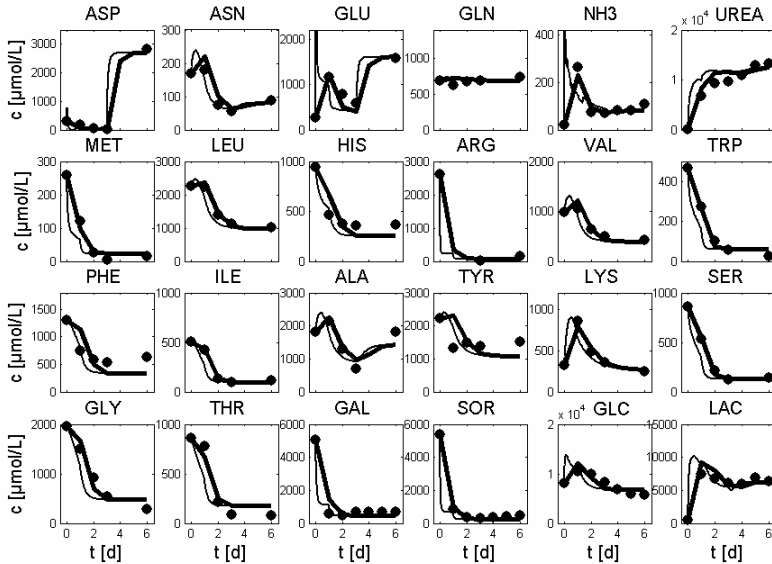


Fig. 6. Concentrations $c_{i,2,k}$ of 24 compounds measured for the run R2 (as a representative example; see Table 1) versus the time t_k (•) and kinetics $c_{0,i}(t)$ obtained by simulation of the Eqs. (3.1)-(3.14) (thin lines) and the kinetics $c^*(i,t)$ averaged according to Eq. (2.2) (thick lines) after fitting the model parameters (Table 4)

The Eq. (3.12) formulates the consumption rates of GAL and SOR (parameters p_{21} and p_{22} , respectively). The Eqs. (3.13) and (3.14) describe both, the synthesis and consumption/degradation of GLC and LAC. Terms in these equations may be interpreted by gluconeogenesis (parameter p_{21}) and glycolysis activities (parameter p_{23}) as well as by reactions catalyzed by sorbitol dehydrogenase (EC1.1.1.14, p_{22}) and lactate dehydrogenase (LDH, EC 1.1.1.27, p_{24}). As a result of this modeling, it was found that the measured increase of GLC and LAC cannot be explained by these reactions alone. Thus, we introduced hypothetical synthesis rates for GLC and LAC from proteins and/or amino acids via the urea cycle and fumarate (p_{31} and p_{32}). The hypothetical link between lactate and amino acid metabolism is supported not only by the improved fit of time course of LAC but also by the correlation between LAC and ALA at the final time ($t = 6d$) as shown in Figure 4b.

Table 4. Model parameters p_m identified by fitting the Eqs. (3) to the measured data of the run R2 (see Table 1, one representative example of the seven runs) as shown in Figure 6. The indices m specify the numbered arrows in the model structure shown in Figure 5.

m	Value	Unit	m	Value	Unit
1	2.638	d ⁻¹	17	55.75	d ⁻¹
2	5.783	d ⁻¹	18	55.86	d ⁻¹
3	253.9	d ⁻¹	19	323.8	L μmol ⁻¹ d ⁻¹
4	3.074	d ⁻¹	20	1.579	μmol L ⁻¹
5	16.46	d ⁻¹	21	27.81	d ⁻¹
6	6.118	d ⁻¹	22	77.00	d ⁻¹
7	8.929	d ⁻¹	23	3.292	d ⁻¹
8	2.917	d ⁻¹	24	4.231	d ⁻¹
9	2.149	d ⁻¹	25	0	L μmol ⁻¹ d ⁻¹
10	0.4579	d ⁻¹	26	0	d ⁻¹
11	26.33	d ⁻¹	27	0.6804	-
12	12.83	d ⁻¹	28	38.05	d ⁻¹
13	6.617	d ⁻¹	29	2.8119	d ⁻¹
14	8.596	d ⁻¹	30	0	-
15	0.0007	L μmol ⁻¹ d ⁻¹	31	1.310	d ⁻¹
16	3.723	d ⁻¹	32	478.7	d ⁻¹
			33	7708	μmol L ⁻¹ d ⁻¹

Figures 5 and 6 illustrate the structure of the dynamic model (Eqs. (3.2)-(3.14)) and the result of model fit to the data of the run R2, respectively. The Tables 3 and 4 show the model fit error for the seven runs and the model parameters identified for run R2, respectively.

The model parameter p_{30} representing the flow from GLC to LAC was found to be low for all runs, i.e. only a small contribution (2% for run R2) of LAC came from SOR, GAL and GLC, whereas the major part (98%) of LAC was obviously produced from other sources, such as nitrogen sources (amino acids and proteins) via the urea cycle or glycogen. Kremling and Gilles [7] estimated a contribution of about 20% of the LAC inflow via the urea cycle.

Figure 7 illustrates in a biplot the first and second principal components calculated from the 33 model parameters identified for the seven runs. The first principal component of the runs R1, R2, R6 and R7 (male donors) are positive and characterized by a high parameter p_{33} (proteolysis rate) and a low parameter p_{10} (methionine uptake), whereas the first principal component of the runs R3, R4, R5 (female donors) is negative and characterized by a low parameter p_{33} and high p_{10} . The parameters p_6 , p_{10} , p_{12} and p_{13} are significantly lower for the runs inoculated with liver from male than from female donor (p-values: 0.004, 0.008, 0.035, 0.044).

The model parameters p_{15} and p_{16} quantifying the asparaginase activity are highly correlated ($r = 0.99$). The parameters p_{25} and p_{26} are positively correlated ($r = 0.95$) and negatively correlated with p_{27} ($r < -0.9$). The model parameters p_{25} and p_{26} quantifying the glutamine synthetase activity are low. They are near to zero (with values $< 0.001 \text{ L } \mu\text{mol}^{-1} \text{d}^{-1}$ and $< 0.001 \text{ d}^{-1}$, respectively) for the runs R1, R2, R4 and R5. When the parameter values are greater, such as for run R3 with the values $p_{25} = 0.011 \text{ L}/\mu\text{mol}/\text{d}$ and $p_{26} = 0.003 \text{ d}^{-1}$, then the parameter p_{27} quantifying the protein synthesis is decreased to 0.2 (whereas p_{27} is greater than 0.6 for the other runs). The parameter p_{27} quantifies the proportion of glutamate efflux to protein (and not to ammonia, see Eqs. (3.8) and (3.10)).

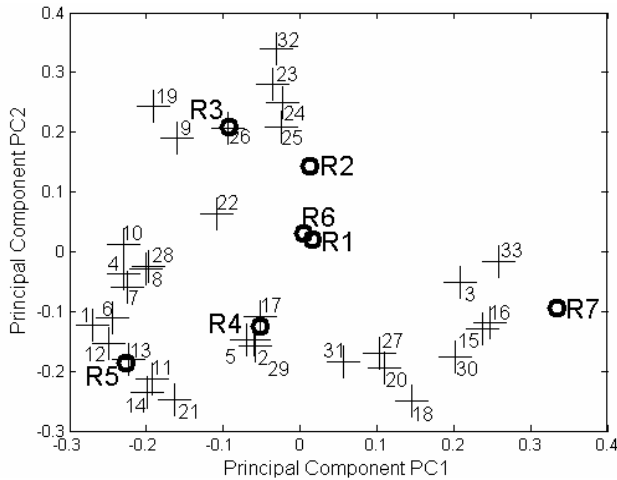


Fig. 7. Biplot of the first and second principal components of the model parameters p_m (centered and scaled; visualized by '+' with indices m) as identified for seven liver cell bioreactor runs R1, ..., R7 (circles; coordinates divided by 20)

4 Conclusion

The kinetics of 24 biochemical compounds in a human liver cell bioreactor was simulated by a system of 48 differential equations. This model extends an already published two-compartment model for 18 amino acids, ammonia and urea [6] by adding of the compounds glucose, galactose, sorbitol and lactate as well as special fluxes. The added fluxes include the secretion of amino acids by proteolysis

(parameter p_{33}) and the synthesis of alanine from aspartate (p_{29}). The model study supports the following hypotheses:

- Enhanced proteolysis during the first day after inoculation of the liver cell bioreactor is responsible for the observed initial increase of the lysine concentration.
- Proteolysis rates (p_{33}) were found to be significantly lower in bioreactors inoculated with liver cells from female donors than from male donors.
- The activities of glutamine synthetase (parameters p_{25} and p_{26}) as well as the flow rate from glucose to lactate (p_{30}) were found to be near to zero.
- Gluconeogenesis (glucose synthesis) and lactate production take place not only from galactose but also from other sources (e.g., glycogen, amino acids, protein).

Thus, the results from modeling analysis show that primary human liver cells cultured in bioreactors follow a metabolic course that is characterized by initial proteolysis, as well as glucose release, indicating a catabolic situation. This could be due to a traumatic effect on the cells of the isolation procedure, comparable to the clinical post-traumatic situation also known as post-aggression or adaption syndrome observed as a response to surgery or trauma [8, 9]. This syndrome is typically characterized by an impaired glucose metabolism and increased rates of proteolysis and lipolysis. Further studies are required to nearer investigate the observation that the proteolysis rate is higher in bioreactors containing cells from male patients than in those containing cells from female patients.

After the first days of culture, a switch of the cell metabolism to an anabolic situation characterized by gluconeogenesis and amino acid uptake took place, which indicates recovery and adaptation of the cells to the culture environment. Thus, the bioreactor provides a representative tool to develop models for the experimental-based mathematical description and simulation of metabolic processes of human liver cells.

Acknowledgement

This work was supported by the German Federal Ministry for Education and Research BMBF within the Programme 'Systems of Life – Systems Biology' (FKZ 0313079B, FKZ 0313079A). The authors thank Susanne Töpfer for proofreading the manuscript.

References

1. Gerlach, J.C., Zeilinger, K., Grebe, A., Puhl, G., Pless, G., Sauer, I., Grunwald, A., Schnoy, N., Muller, C., Neuhaus, P.: Recovery of preservation-injured primary human hepatocytes and nonparenchymal cells to tissuelike structures in large-scale bioreactors for liver support: an initial transmission electron microscopy study. *J Invest Surg.* 16 (2003) 83-92
2. Zeilinger, K., Holland, G., Sauer, I.M., Efimova, E., Kardassis, D., Obermayer, N., Liu, M., Neuhaus, P., Gerlach, J.C.: Time course of primary liver cell reorganization in three-dimensional high-density bioreactors for extracorporeal liver support: an immunohistochemical and ultrastructural study. *Tissue Eng.* 10 (2004) 1113-24

3. Pfaff, M., Toepfer, S., Woetzel, D., Driesch, D., Zeilinger, K., Pless, G., Neuhaus, P., Gerlach, J.C., Schmidt-Heck, W., Guthke, R.: Fuzzy cluster and rule based analysis of the system dynamics of a bioartificial 3D human liver cell bioreactor for liver support therapy. In: Dounias, G., Magoulas, G., Linkens, D. (eds.): *Intelligent Technologies in Bioinformatics and Medicine. Special Session. Proceedings of the EUNITE 2004 Symposium. A Publication of the University of the Aegean* (2004) 57
4. Schmidt-Heck, W., Zeilinger, K., Pfaff, M., Toepfer, S., Driesch, D., Pless, G., Neuhaus, P., Gerlach, J.C., Guthke, R.: Network analysis of the kinetics of amino acid metabolism in a liver cell bioreactor. *Lect. Notes Comput. Sc.* 3337 (2004) 427-38
5. Schmidt-Heck, W., Zeilinger, K., Pless, G., Gerlach, J.C., Pfaff, M., Guthke, R.: Prediction of the Performance of Human Liver Cell Bioreactors by Donor Organ Data. *Lect Notes Bioinformatics*, 3745 (2005) 109-119
6. Guthke, R., Zeilinger, K., Sickinger, S., Schmidt-Heck, W., Buentemeyer, H., Iding, K., Lehmann, J., Pfaff, M., Pless, G., Gerlach, J.C.: Dynamics of Amino Acid Metabolism of Primary Human Liver Cells in 3D Bioreactors. *Bioprocess and Biosystems Engineering*, 28 (2006) 331-340
7. Kremling, A., Gilles, E.D.: Model based measurements for cell culture reactors. *Conference on Systems Biology of Mammalian Cells. Heidelberg, 12.-14.07.2006*
8. Selye, H.: The general adaption syndrome and the disease of adaption. *J Clin Endocrinol* 6 (1946) 117-126
9. Frayn, K.N.: Hormonal control of metabolism in trauma and sepsis. *Clin Endocrinol* 24 (1986) 577-599

The Probabilities Mixture Model for Clustering Flow-Cytometric Data: An Application to Gating Lymphocytes in Peripheral Blood

John Lakoumentas¹, John Drakos¹, Marina Karakantza², Nicolaos Zoumbos²,
George Nikiforidis¹, and George Sakellaropoulos¹

¹ Medical Physics Department, University of Patras, Greece

² Hematology Division, Department of Internal Medicine, University of Patras, Greece

Abstract. Data clustering is a major data mining technique and has been shown to be useful in a wide variety of domains, including medical and biological statistical data analysis. A non trivial application of cluster analysis occurs in the identification of different subpopulations of particles in large-sized heterogeneous flow-cytometric data. Mixture-model based clustering has been several times applied in the past to medical and biological data analysis; to our knowledge, however, non of these applications was involved with flow-cytometric data. We claim, that utilizing the probabilities mixture model offers several advantages compared to other proposed flow-cytometric data clustering approaches. We apply this model in order to gate lymphocytes in peripheral blood, which is a necessary first-step procedure when dealing with various hematological diseases diagnoses, such as lymphocytic leukemias and lymphoma.

1 Introduction

Flow cytometry is a commonly utilized technology in the biological and medical fields, able to provide both quantitative and qualitative information about single particles (cells, nuclei, microorganisms, latex beads, etc.), as they flow in a fluid (for example, blood) stream through a beam of light [1]. In the past decade, flow cytometry has been extensively applied in clinical laboratories for hematological immunophenotyping and corresponding diseases diagnoses, including various types of leukemia and lymphoma [2]. Due to its distributed nature, the hematopoietic system is amenable to flow-cytometric analysis; this analysis produces measurements of physical properties of single cells, such as size (represented by the relative forward angle light scattered intensity) and internal granularity or complexity (represented by the relative 90-degrees-angle or side light scattered intensity), as well as measurements of fluorescent properties of single cells (represented by the relative fluorescence intensity while in presence of certain antigens conjugated to selected fluorescent dyes). A clinician laboratory expert then takes into account such information along with clinical findings and uses his expert knowledge over the considered disease to proceed to a diagnosis.

Despite the recent progress on flow cytometry technology, however, no similar progress seems to have been achieved on flow-cytometric data analysis. Typical flow cytometers are able to provide trivial data analysis tools, like histograms and 2- or 3-parametric scatter plots. The clinician laboratory expert, then, usually performs all the remaining data analysis required by hand, proceeding to conclusions based on his empiric knowledge, instead of utilizing predefined globally accepted algorithmic data analysis techniques. This is mainly due to the fact, that different flow-cytometric data correspond to different underlying medical or biological problems, thus they are characterized by different physical properties and require specific-purpose data analysis algorithms to be handled effectively. In fact, methods of flow-cytometric data analysis are dependent on what exactly the experimenter wants to know about the particles. However, it seems scientifically more accurate, similar data analysis instances to be treated universally by non (or slightly) supervised predefined methods, that may also be able to include and take advantage of all expert information about the underlying medical or biological problem.

A usual arising problem when analyzing heterogeneous (multimodal) medical or biological flow-cytometric data, is the identification of different natural groups of particles (or particle subpopulations) among the whole sample. Under data mining terms, this is a well-defined optimization problem, referred to as unsupervised data clustering: given a set of clustering elements (particles) along with a set of associated vectors of clustering attributes (optical and fluorescent measurements), an optimal partition of the elements into subgroups, called clusters, is requested under some predefined proximity or similarity optimality criteria. Clustering is a hard problem itself, that complicates more when dealing with large-sized heterogeneous flow-cytometric data. Each single dataset is usually described by unique characteristics (for example, the elements' topology as aposed on the Euclidean space of their attributes, the actual number and shape of the clusters formed, etc.), dependent on the nature of the underlying physical problem. Therefore, it seems difficult to obtain a fast clustering algorithm that works well for all flow-cytometric data. Furthermore, the clusters of a single instance are often totally different to each other; that is, they are characterized by various shapes and different variance-covariance intracluster structures. Flow-cytometric data clustering, thus, requires efficient special-purpose sophisticated algorithmic techniques, that are able to include available expert knowledge and respond flexibly in presence of totally different data clustering instances, that correspond to each single underlying medical or biological problem.

We propose a novel procedure for gating the subset of lymphocytes among all cells of peripheral blood flow-cytometric samples, which is a necessary first-step procedure when dealing with various hematological diseases diagnoses, such as lymphocytic leukemias and lymphoma, and forms a well-defined case of (unsupervised) flow-cytometric data clustering. The proposed procedure includes the use of the probabilities mixture model; a relatively new methodology, that has exhibited promising results for clustering in other domains [3]. We claim, that utilizing this model offers several advantages compared to other proposed

flow-cytometric data clustering approaches; this claim is based on experimental observations and also relies on our general belief, that problems generated in nature can be approximated better by using probabilistic methods, instead of geometrical or neural network ones. We also believe, that this model is general enough to be applied to a great variety of unsupervised data clustering instances (and flow-cytometric data clustering instances, among them), each time corresponding to various underlying medical or biological problems, by suitably incorporating experts' prior knowledge over the considered natural problem in an explicit manner.

In this context, gating lymphocytes constitutes the first in a sequence of steps towards a fully computational method of hematological diagnosis, offering an objective perspective to procedures routinely performed by human experts. In addition, most of the consecutive steps of the above mentioned sequence are also formulated as special unsupervised data clustering instances and can be dealt with in a similar uniform manner.

2 Previous Work

The basic principles of data clustering are described in [4]. Many of the existing clustering algorithms are not straightforward applicable to flow-cytometric data analysis, mainly because of two reasons. First, some of those algorithms perform supervised clustering; that is, they initially deal with a pattern set of labeled data (cluster assignments are given), that forms the basis for clustering any other given unlabeled set of similar data. Supervised clustering techniques are usually simple and efficient enough, but, unfortunately, they cannot be used in flow-cytometric data analysis, since rarely such similar data samples are available. Second, flow-cytometric data samples are usually large; some of them may be comprised of thousands or tens of thousands of particles. When considering such large samples, all clustering algorithms of hierarchical philosophy become extremely inefficient and inapplicable, since they require the continuous storage of the whole dataset, in order to maintain a proximity matrix among all possible particle pairs.

The great majority of the initially proposed flow-cytometric data clustering approaches [5,6,7] applied traditional linkage methods, such as the well-known k-means algorithm and few of its variations. K-means is simple and efficient enough; however, it suffers of the main drawback of tending to cluster elements into spherical non-overlapping groups, due to the use of the Euclidean distance as a proximity metric. This is rarely the case when considering flow-cytometric data, however, where clusters tend to be of non canonical shape and are not usually well-separated. K-means variants include the definition of Mahalanobis (instead of the Euclidean) distance as a proximity metric, that is applied to identify different variant-covariant structures and clusters of ellipsoid shape, and the use of a large number of initially requested clusters, that are finally merged into their actual number following some predefined heuristic criteria. An alternative type of clustering approaches for dealing with not well-separated clusters

introduces fuzzy logic, where each element is allowed to belong to many clusters according to a certain probability distribution of membership. An application of fuzzy c-means (fcm) variations to clustering flow-cytometric data is presented in [8]. However, fcm seems to suffer of similar disadvantages to k-means, since both of them are based on common (of linkage nature) clustering principles and require the initial knowledge of the number of existing clusters. Finally, a few newer approaches utilize unsupervised artificial neural network (ANNs) methods for clustering, such as self-organizing maps (SOMs) [9] and adaptive resonance theory (ART) [10]. Our application of clustering peripheral blood cells is also considered there and some of the results provided are comparative to ours, taking into account though all possible attributes a flow cytometer is able to provide (which is not the case in our implementation, as discussed in Sect. 4). To our knowledge, the probabilities mixture model we utilize and describe in the following section, has never been applied to flow-cytometric data clustering in the past.

3 Mixture-Model Based Clustering

The probabilities mixture model for clustering [11] assumes, that the data are generated by a mixture of underlying prior probability distributions, in which each component represents a different cluster. Let *cluster* be a discrete random variable denoting the cluster a randomly chosen element belongs to (taking integer values from 1 to k) and $att_1, att_2, \dots, att_n$ be a set of (generally) continuous random variables obtaining the n attribute values of the randomly chosen element. The joint probability distribution assumed to generate the data is $Pr\{cluster, att_1, att_2, \dots, att_n\}$ and the probabilities mixture model mentioned above is defined by the following equation:

$$Pr\{cluster, att_1, att_2, \dots, att_n\} = Pr\{cluster\} \prod_{i=1}^n Pr\{att_i|cluster\} .$$

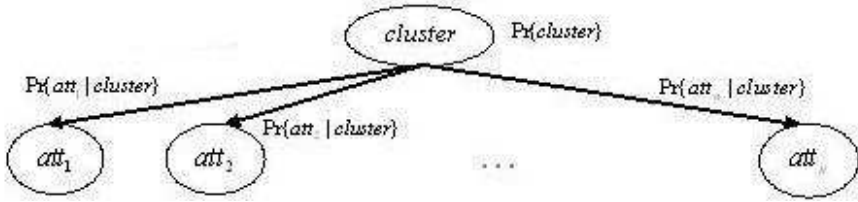
The above equality implies, that if the *cluster* value is given, all attribute variables are pairwise mutually stochastically independent. By applying simple rules of probability, the following relationship also holds:

$$Pr\{cluster|att_1, att_2, \dots, att_n\} \propto Pr\{cluster\} \prod_{i=1}^n Pr\{att_i|cluster\} .$$

Using that equation and given all attribute values, one is able to compute the most probable assignment of each element to a cluster. That is,

$$cluster = \arg \max_{j=1, \dots, k} \left[Pr\{cluster = j\} \prod_{i=1}^n Pr\{att_i|cluster = j\} \right] .$$

Mixture-model based clustering is sometimes also referred to as Bayesian clustering. The probabilities mixture model described above can be alternatively represented by the following (naive) Bayesian network [12]:



Speaking in Bayesian networks terms, all attribute variables are considered to be Gaussian and the *cluster* value is assigned Bayesian-Dirichlet equivalent (BDe) priors. Since a clustering dataset is given, all attribute values become observed, but all *cluster* values remain unknown; then, *cluster* is said to be a hidden variable. In such a case, the unknown prior probability distributions, $Pr\{cluster\}$ and $Pr\{att_i|cluster\}$, cannot be computed accurately and are estimated in a way the data approximate their maximum likelihood (ML) value, which is the optimization criterion of this clustering approach. The estimation is usually done by the well-known Expectation-Maximization (EM) algorithm. EM iteratively assigns *cluster* its expected values given the currently estimated prior probability distributions (Expectation step) and then reestimates these distributions in order to maximize the likelihood of the data and the newly estimated values (Maximization step), i.e., the probability that these values are obtained by the currently estimated underlying joint probability distribution and the data, until a convergence criterion is satisfied.

The unknown prior probability distributions' parameters may obtain initial values with the use of simpler and faster cluster analysis techniques (say, utilizing k-means) or with manual gating by the expert; that latter approach introduces prior expert knowledge over the underlying natural problem to the model. Consequently, since the initial *cluster* assignments are close to optimal, EM is expected to converge faster, escape of local maxima convergence and provide more accurate clustering results. Mixture-model based clustering is efficient enough in terms of time and space, even when considering such large-sized data, that hierarchical clustering algorithms are totally inapplicable. It is also flexible when considering instances of various cluster structures and shapes of the same examined problem; this flexibility lies on the fact, that intracluster variance-covariance structures are estimated through the model, therefore clusters are expected to be identified, no matter what their actual shape is. A worth mentioning theoretical result is that clustering EM is equivalent to k-means, assuming spherical gaussian mixtures [13].

4 The Application: Gating Lymphocytes

The diagnosis of various hematological diseases, such as types of lymphocytic leukemias and lymphoma, requires extensive lymphocytic data analysis. For a clinician laboratory expert using flow cytometry technology, a necessary first-step procedure arising is the selection of the subset of lymphocytes among the

whole sample of peripheral blood (leukocytes). Leukocytes are mainly subdivided into three subpopulations; that is, lymphocytes, monocytes and granulocytes. These subpopulations are usually recognized by their cells' relative size and relative internal granularity or complexity. The former property reflects on the relative intensity value of the forward angle light scattered (FS) on each cell; thus, lymphocytes are expected to have smaller FS values than monocytes, that in turn are expected to have smaller FS values than granulocytes, due to their relative size ordering. The latter property reflects on the relative intensity value of the side angle light scattered (SS) on each cell; therefore, granulocytes are characterized by large SS values, monocytes are characterized by small SS values and lymphocytes are expected to have slightly smaller SS values than monocytes. Consequently, when apposing all leukocytes on the 2-dimensional Euclidean space constructed by attributes FS and SS , lymphocytes are expected to form the subpopulation of high density that lies nearest to the axes cut. Based on that knowledge and using the FS - SS plot provided by the flow cytometer, the laboratory expert easily performs a course gating of the lymphocytes subpopulation by hand.

Gating lymphocytes in an accurate automated (non-supervised) way, however, is not such a trivial task in practice. The combination of attributes FS and SS rarely seems to suffice for achieving a satisfactory distinction of the above subpopulations; i.e., the subpopulations are usually mixed, especially according to the FS attribute (relative size). The SS attribute, on the other hand, is able to distinguish granulocytes from all other leukocytes sufficiently well, but cannot achieve that successfully the distinction of lymphocytes and monocytes. Using additional flow-cytometric (antigen) attributes would be useful; for example, the relative intensity value of the light scattered under the presence of antigen $CD14$ can be used for distinguishing monocytes from all other leukocytes. Unfortunately, most laboratories do not usually provide all necessary (in terms of lymphocytes gating) antigen attributes, as long as they are not required for the considered disease diagnosis. Furthermore, in presence of the considered disease, the clustering structure of a sample may be far different than the expected one or (in extreme cases) some of the above mentioned subpopulations may be totally absent. This latter fact may sometimes yield to better clustering results; however, in such cases, the issue of defining the correct number of existing clusters arises. Finally, beyond the three main leukocyte subpopulations, the FS - SS plot usually contains a great number of isolated points or 'outliers', that correspond to dead cells. All these characteristics affect negatively the optimality of the clustering process, since the applied clustering algorithm has to be flexible in providing optimal results for every possible instance (corresponding to a specific flow-cytometric sample) of the considered general problem.

In our approach, we gate the subpopulation of lymphocytes using FS , SS and additional antigen attributes, taking advantage of the fact, that lymphocytes are mainly subdivided into B-cells and T-cells. For both of those lymphocyte subpopulations, there exist marker antigens that produce corresponding identifying flow-cytometric attributes; specifically, antigen $CD19$ (alternatively, $CD20$) is a

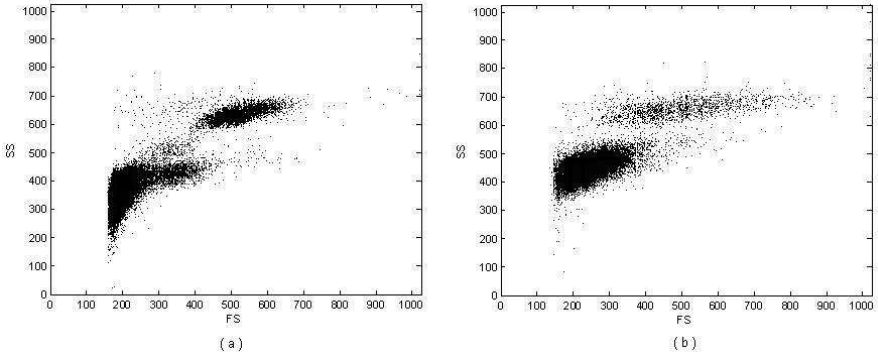


Fig. 1. Two *FS-SS* plots of different patients' peripheral blood samples (lymphocytes, monocytes, granulocytes and dead cells). In plot (a), lymphocytes seem to subdivide into two distinct subpopulations. In plot (b), monocytes are almost totally absent.

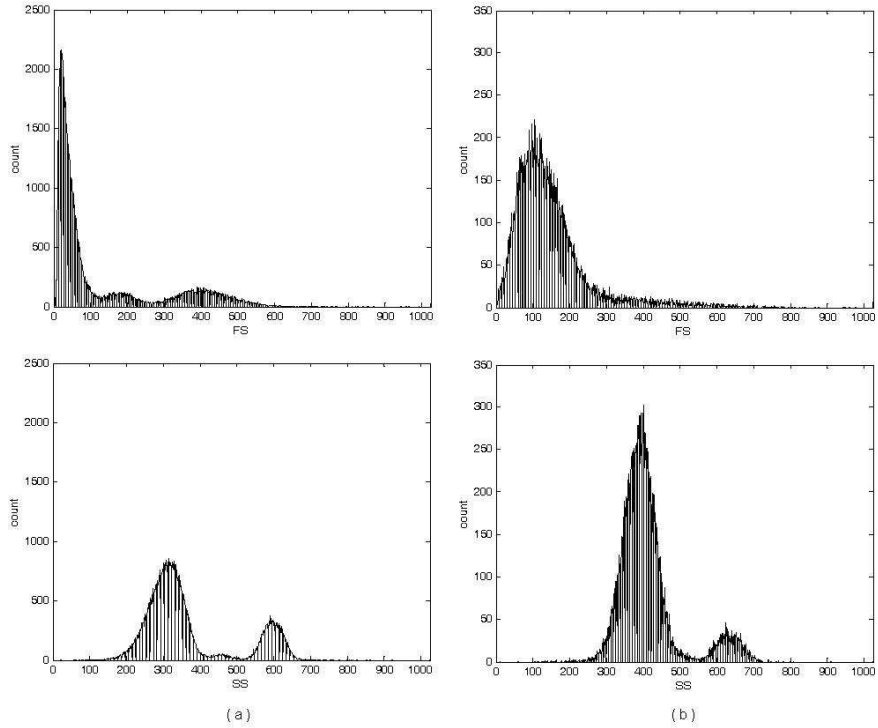


Fig. 2. The histograms corresponding to the samples of Fig. 1

B-cell marker and antigen *CD3* is a T-cell marker. These attributes are always available for B-Chronic Lymphocytic Leukemia diagnosis and flow-cytometric samples of patients suffering from this disease are analyzed in this application. Thus, the *SS-CD19* combination of attributes is used for gating B-cells and the

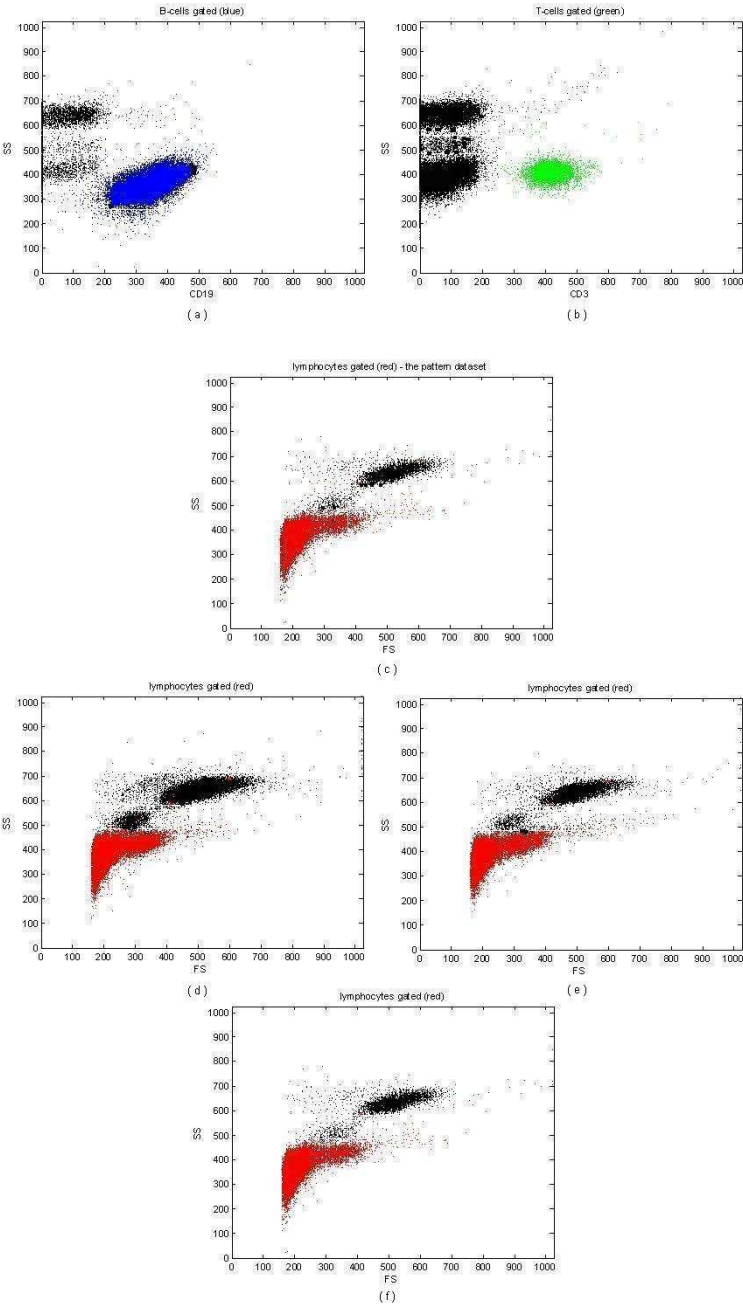


Fig. 3. An example of the whole lymphocytes gating process: Initially, B-cells (a) and T-cells (b) are gated. Then, the results are merged for the construction of a pattern dataset with lymphocytes gated (c). Finally, lymphocytes are gated in each sample of the same patient examination (d,e,f).

SS-CD3 combination of attributes is used for gating T-cells, both by applying the probabilities mixture model, with an initial determination of the requested number of clusters equal to 2.

However, the required antigen attributes, *CD3* and *CD19*, are not available in each flow-cytometric sample of the same patient examination. Therefore, suitable distance-based techniques have to be utilized for combining the above mentioned gating results and achieving lymphocytes gating in every sample of a patient examination. The technique used is a well-known supervised clustering method, usually referred to as *k*-nearest neighbors; each cell is labeled as a lymphocyte or not, depending on the label of the majority of its *k* closest (under Euclidean distance terms) cells on a pattern dataset; i.e., a dataset whose cell labels are already known. The pattern dataset is constructed by merging the results of B-cells gating and T-cells gating, using the same method. For dealing with ‘outliers’, a square window of maximum neighboring distance is posed, into which all *k* neighbors are required to belong; that is, the actual number of a cell’s neighbors is upper bounded by *k* and is lower bounded by the number of cells belonging into the predefined neighboring area. The supervised clustering technique of *k*-nearest neighbors is applied to the *FS-SS* plot of each sample, since the above two attributes are the only ones in common at every sample of the same patient examination; such samples are expected to be similar; that is, to be characterized by similar cluster structure and shapes.

We implemented our clustering approach using the programming environment MATLAB 7.0.1 and analyzed flow-cytometric data of various examinations of patients suffering from B-Chronic Lymphocytic Leukemia. The samples analyzed were of cardinality up to about 30000 cells. An example of gating lymphocytes in peripheral blood (corresponding to the first sample of Fig. 1) with our approach is presented in Fig. 3. The data, originally stored in FCS format in the cytometer, were transformed and exported to a form readable by MATLAB. The functions in MATLAB were developed in-house.

5 Conclusions

We claim, that mixture-model based clustering is an effective method that yields to similar or better results when compared to other approaches currently applied in the domain of flow-cytometric data analysis, like artificial neural networks and fuzzy clustering. We utilized the proposed model for a specific flow-cytometric data clustering instance, as a paradigm. However, it can be generalized for solving any other flow-cytometric clustering instance, by obtaining each time appropriate initial conditions by the laboratory expert. In B-Chronic Lymphocytic Leukemia diagnosis, for example, all rest of the flow-cytometric data analysis required includes similar unsupervised clustering procedures, like gating B-cells and T-cells in the subpopulation of lymphocytes, gating lymphocytes that co-express antigens *CD5* and *CD20*, gating *K*- or *λ*-positive lymphocytes, etc. All these procedures can be effectively dealt with mixture-model based clustering, similarly to the examples presented in our application.

It is mandatory for the proposed clustering approach in order to perform well, each time the correct combination of clustering attributes to be selected. In our application, for example, *FS* is not an attribute able to produce an optimal distinction of the lymphocyte subpopulation among the rest of the leukocytes, while antigen attribute *CD14* in combination with *SS* would be. In that latter case, lymphocytes would be straightforwardly distinguished using mixture-model based clustering with more than two clustering attributes. Given the appropriate set of attributes, mixture-model based clustering is able to perform multidimensional clustering, as long as these attributes are available on a common dataset. It would therefore be useful for medical doctors to preplan a flow-cytometric examination and include critical (in terms of lymphocytes gating) antigens in the same blood sample passing through the cytometer.

Finally, in our application the proposed model was applied to instances with an initially given number of existing clusters; that is, equal to 2 when gating B-cells and T-cells. As presented in Fig. 1, however, the actual number of clusters varies when dealing straightforwardly with the *FS-SS* plot (and probably with a few more attributes available) or in many other possible data clustering analysis instances of various natural problems; in totally unsupervised clustering, the existing number of clusters is initially unknown. We observed experimentally, that the probabilities mixture model is flexible enough to accommodate situations with zero class memberships. That is, when the predefined number of clusters is unrealistically large, the model tends to estimate the actual number of existing clusters (according to the maximum likelihood principle) by assigning zero elements to all redundant clusters. Thus, by initially setting a number of clusters much larger than anticipated, we can use mixture-model based clustering even when the actual number of clusters is unknown and perform fully unsupervised clustering.

References

1. H. M. Shapiro: "Practical flow cytometry, 3rd edition", Wiley Liss, 1994.
2. M. Brown and C.T. Wittwer: "Flow cytometry: principles and clinical applications in hematology", Clinical Chemistry, 46:8(B), 1221-1229, 2000.
3. Y. Barash and N. Friedman: "Context-specific Bayesian clustering for gene expression data", Journal of Computational Biology, 9(2), 169-191, 2002.
4. R. Xu and D. Wunsch II: "Survey of clustering algorithms", IEEE Transactions on Neural Networks, 16(3), 645-678, 2005.
5. R. F. Murphy: "Automatic identification of subpopulations in flow cytometric list mode data using cluster analysis", Cytometry, 6, 302-309, 1985.
6. S. Demers, J. Kim, P. Legendre and L. Legendre: "Analyzing multivariate flow cytometric data in aquatic sciences", Cytometry, 13(3), 291-298, 1992.
7. T. C. Bakker Schut, B. G. De Grooth and J. Greeve: "Cluster analysis of flow cytometric list mode data on a personal computer", Cytometry, 14(6), 649-659, 1993.
8. M. F. Wilkins, S. A. Hardy, L. Boddy and C. W. Morris: "Comparison of five clustering algorithms to classify phytoplankton from flow cytometry data", Cytometry, 44(3), 210-217, 2001.

9. D. S. Frankel, S. L. Frankel, B. J. Binder and R. F. Vogt: "Application of neural networks to flow cytometry data analysis and real-time cell classification", *Cytometry*, 23(4), 290-302, 1996.
10. L. Fu, M. Yang, R. Braylan and N. Benson: "Real-time adaptive clustering of flow cytometric data", *Pattern Recognition*, 26(2), 365-373, 1993.
11. C. Fraley and A. Raftery: "Model-based clustering, discriminant analysis and density estimation", *Journal of the American Statistical Association*, 97, 611-631, 2002.
12. R. E. Neapolitan: "Learning Bayesian Networks", Prentice Hall, 2004.
13. G. Celeux and G. Govaert: "A classification EM algorithm for clustering and two stochastic versions", *Computational Statistics and Data Analysis*, 14, 315-332, 1992.

Integrative Mathematical Modeling for Analysis of Microcirculatory Function

Adam Kapela¹, Anastasios Bezerianos², and Nikolaos Tsoukias¹

¹ Dept. of Biomedical Engineering, Florida International University, Miami, FL, USA

² Dept. of Medical Physics, School of Medicine, University of Patras, Patras, Greece
tsoukias@fiu.edu

Abstract. The microcirculatory vascular tone and the regional blood flow are regulated by an elaborate network of intracellular and extracellular signaling pathways with multiple feedback control loops. This complicates interpretation of experimental data and limits our ability to design appropriate interventions. Mathematical modeling offers a systematic approach for system and data analysis and for guiding new experimentation. We describe here our efforts to model signal transduction events involved in the regulation of blood flow and integrate mechanisms at the cellular level to describe function at the multicellular/whole-vessel level. The model provides a) a working database of rat mesenteric endothelial and smooth muscle physiology where newly acquired experimental information on cell electrophysiology and signal transduction can be incorporated, and b) a tool that will assist investigations on the regulation of vascular resistance in health and disease. An example of model application to the study of the pathogenesis of salt-sensitive hypertension is illustrated.

Keywords: integrative computational physiology, vascular system, calcium dynamics, microcirculation.

1 Introduction

The amount of biological data on the cardiovascular system is growing rapidly and includes electrophysiological, proteomic and genomic information. However, it is now apparent that comprehensive description of individual system components is necessary but not sufficient to elucidate the integrative behavior of the system. Our understanding of physiological behavior at the system level can be facilitated by the development of computational models [1]. For example, vascular tone and regional blood flow in the microcirculation is regulated by an elaborate network of signaling pathways. This network includes intracellular signaling, as well as cell-to-cell communication with paracrine factors or diffusion of species through homo- and hetero-cellular gap junctions. This multitude of signaling pathways create multiple feedback control loops that tightly regulate calcium (Ca^{2+}) homeostasis in vascular smooth muscle cells. Experimentation has begun to untangle this elaborate network and continuously provides new insights about the physiology of blood vessels and the mechanisms that regulate tone and blood flow. Mathematical modeling offers a systematic

approach for the system analysis and can assist in this effort both as a tool for data analysis and for guiding new experimental studies.

The development of multiscale models that will describe function at the tissue level while integrating mechanisms at the subcellular and molecular level holds great promise in the investigations of muscle physiology. Such advancements have been made in modeling cardiac muscle for example. Detailed biophysically-based mathematical models of cardiac myocytes were developed, that include membrane electrophysiology, calcium dynamics, cell mechanics, metabolic and signal-transduction pathways. Multicellular models can account for normal and pathological tissue structure, and allow computing the propagation of activation wavefronts to study the cellular bases of arrhythmogenesis. Cardiac models are also used for functional interpretation of proteomic data and investigation of gene related channelopathies. It is now possible to incorporate information about changes of protein expression or channel mutation and to determine the physiological and pathophysiological function at the cellular and tissue levels [2], [3]. This progress has not been paralleled in the vasculature.

Prior modeling efforts have provided significant insights for the physiology of vascular endothelial and smooth muscle cells and form the basis for the current study to build upon. Previous models have focused mostly on the analysis of a specific physiological behavior, incorporating necessary system components and adjusting unknown parameters to fit the physiological response under investigation. The continuous increase of available experimental data, in particular electrophysiological measurements of membrane currents, allows us to develop models that are tissue specific and to incorporate higher level of detail by integrating new system components and utilizing parameters determined from independent experiments.

This study aims to develop a detailed mathematical model of integrated calcium dynamics in the rat mesenteric microcirculation and describe its regulation by different signaling pathways. The model will provide a working database/benchmark of rat mesenteric endothelial and smooth muscle physiology where newly acquired experimental information on the electrophysiology and signal transduction pathways can be incorporated. Our goal is through modeling to generate experimentally testable hypotheses for the mechanisms that regulate vascular resistance in health and disease. In this paper, we summarize our efforts and results towards achieving these goals.

2 Methods

We first developed theoretical models of isolated endothelial cell (EC) and smooth muscle cell (SMC) and validated them against experimental data from isolated cells under different experimental conditions. Then, the two isolated models were integrated into an EC/SMC model. The integrated model was validated with data from experiments on isolated vessel segments where agonist stimulations could be considered axially homogenous. Finally, the two cell types were combined into a multicellular vessel model. Outputs of the model are compared with experiments on conducted vasomotor responses induced by local stimulations.

To formulate the isolated cell models, the general methodology developed for cardiac myocytes was utilized. In particular, the membrane electrophysiology is described with the lumped Hodgkin-Huxley-type formalism, and changes in membrane potential (V_m) are calculated from:

$$C_m \frac{dV_m}{dt} = -\sum I_m + I_{stim}, \quad (1)$$

where C_m is membrane capacitance, V_m is membrane potential, I_m are different membrane currents described in the following sections, and I_{stim} is an external stimulation current. Two novel approaches are used to build multicellular models. Electrochemical coupling between adjacent cells allows ionic exchange determined by electrochemical gradients. Nitric oxide coupling simulates effect of endothelium-derived nitric oxide on smooth muscle and forms a negative feedback loop through the effect of smooth muscle on the endothelium.

2.1 Endothelial Cell Model

Endothelial cells play a key role in vascular tone modulation by regulating adjacent smooth muscle cell contraction in blood vessel walls. Ca^{2+} has a major role in EC functionality, including production of vasoactive substances such as nitric oxide (NO). EC is an electrically non-excitabile cell and the influence of transmembrane potential (V_m) on EC Ca^{2+} dynamics remains controversial. Nevertheless, EC electrical activity plays an important physiological role in cell-cell communication including myoendothelial communication and conducted responses along the vessel axis. Previous theoretical investigations of EC function have neglected for the most part the role of EC plasma membrane electrophysiology. The EC model integrates both EC Ca^{2+} dynamics and plasmalemmal electrical activity to investigate EC responses to various stimulatory conditions and the interdependency of Ca^{2+} and V_m . The model, unlike previous modeling efforts, contains a detailed description of plasmalemmal electrophysiology (Fig. 1). The plasma membrane includes kinetic descriptions for non-selective cation channels (NSC), store operated cation channels (SOC), small (SK_{Ca}) and intermediate (IK_{Ca}) conductance calcium-activated K^+ channels, inward rectifier K^+ channels (K_{IR}), volume regulated anion channels (VRAC), calcium-activated Cl^- channels (CACC), Na^+ - Ca^{2+} exchanger (NCX) and Na^+ - K^+ - Cl^- cotransporter and Na^+ - K^+ -ATPase pump. It also includes intracellular Ca^{2+} handling components such as IP_3 receptor, sarco/endoplasmic reticulum Ca^{2+} ATPase (SERCA) and plasma membrane Ca^{2+} ATPase (PMCA) mechanisms. Model components are formulated based on recent EC experimental data or adapted from previous EC models. Michaelis-Menten type kinetics describe the steady-state dependence of NO production rate as a function of intracellular calcium concentration ($[Ca^{2+}]_i$). Activation and inactivation time constants express the time required for activation/deactivation of endothelial NO synthase and change in NO release after a change in $[Ca^{2+}]_i$. This formulation allows continuous prediction of NO production rate as the $[Ca^{2+}]_i$ changes with time.

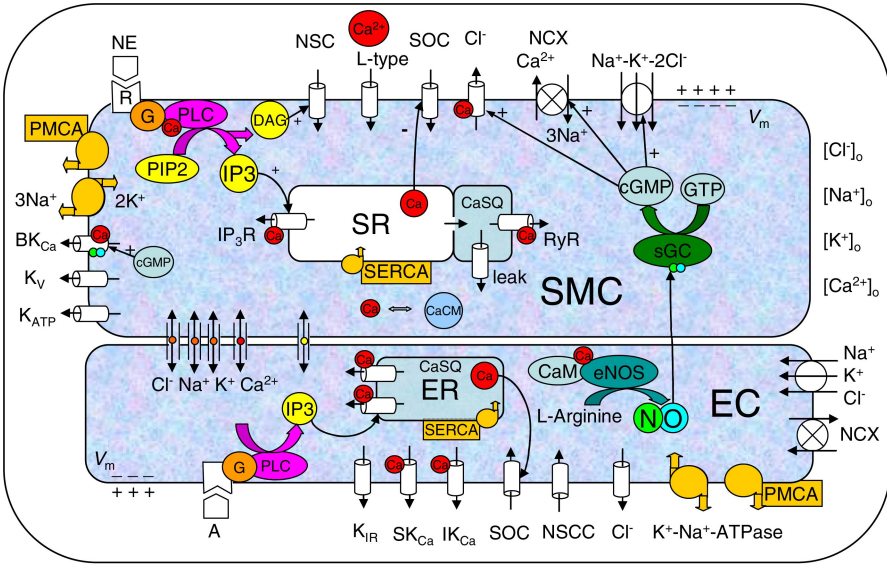


Fig. 1. Integrated calcium dynamics in endothelial and smooth muscle cells

2.2 Smooth Muscle Cell Model

The intracellular concentration of free Ca^{2+} in smooth muscle cells is the main determinant of vascular tone and regional blood flow. The theoretical model of calcium dynamics in SMC from rat mesenteric arterioles is depicted in Fig. 1. The plasma membrane contains ion channels, pumps, exchangers and receptors and includes all the major transmembrane currents that have been identified in SMCs of rat mesenteric arterioles. The ion channels include potassium channels (large conductance Ca^{2+} -activated (BK_{Ca}), ATP-sensitive (K_{ATP}), and delayed rectifier (K_v); cGMP-dependent Ca^{2+} -activated chloride channels (Cl); nonselective Ca^{2+} -impermeable cation channels (NSC); store operated Ca^{2+} -permeable nonselective cation channels (SOC); and L-type voltage-dependent Ca^{2+} channels. The model also contains mathematical descriptions for the Na^+/K^+ -ATPase and plasma membrane Ca^{2+} ATPase (PMCA) pumps, the $\text{Na}^+/\text{K}^+/\text{Cl}^-$ cotransporter and the $\text{Na}^+/\text{Ca}^{2+}$ exchanger (NCX). The intracellular calcium store, mainly representing the sarcoplasmic reticulum, includes uptake and release compartments. The uptake compartment contains IP_3 receptor Ca^{2+} channels (IP_3R), and sarcoplasmic reticulum Ca^{2+} -ATPase pumps (SERCA). In the release compartment Ca^{2+} buffering by calsequestrin (CSQN), ryanodine receptor Ca^{2+} channels (RyR) and a leak current are taken into consideration. A mathematical description for agonist-induced activation of α_1 -adrenoceptor leading to IP_3 formation through the G-protein-phospholipase C (PLC) pathway is incorporated in the model to simulate the effect of NE. Formation of IP_3 and DAG after NE binding to receptor affects release from SR and the NSC channel current. The vasodilatory action of NO is simulated through a direct effect on BK_{Ca} channels or through the formation of cGMP following activation of soluble guanylate cyclase (sGC). In the model four main targets of cGMP are incorporated; the Cl channel, the BK_{Ca} channel, the NCX and the cotransport.

2.3 Integrated EC/SMC Model

EC and SMC models are integrated through the following mechanisms (Fig. 1): a) by the diffusion of IP_3 through the myoendothelial gap junctions, b) by electrochemical coupling and c) by coupling through the paracrine diffusion of NO. Diffusion of IP_3 is proportional to the IP_3 concentration difference in the two cells. Previous models (including cardiac models) have assumed an ohmic behavior of gap junctions. In this study, electrical coupling is accounted through detailed description of ionic (potassium, chloride, sodium and calcium) currents using the Goldman-Hodgkin-Katz formulation. This approach provides ionic currents that depend on gap junctional resistance, electrochemical gradient and ionic concentrations and enables monitoring of ionic exchange in addition to current between the two cells. NO availability in SMC is estimated from the NO concentration profiles computed previously [4]. The integrated model forms a benchmark for investigating feedback control loops in the regulation of Ca^{2+} and NO signaling pathways.

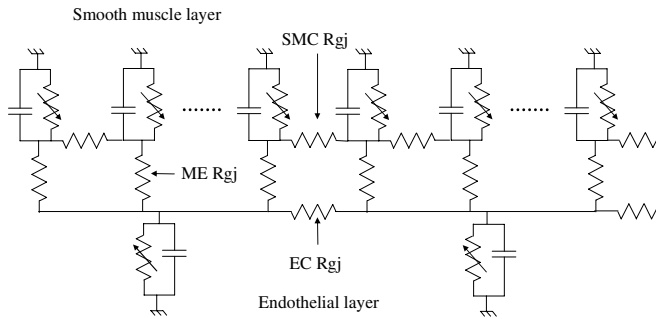


Fig. 2. Electrical equivalent diagram of the multicellular vessel model. SMC, ME and EC Rgj are smooth muscle, myoendothelial and endothelial gap junction resistances, respectively.

2.4 Multicellular Vessel Model

To study conducted vasomotor responses in isolated vessel segments, a multicellular vessel model was developed. A 3mm-long vessel segment spans 30 endothelial cells (each 100 μ m long) arranged serially along the artery's long axis, and 450 smooth muscle cells arranged perpendicularly to the endothelial cells. Each EC is directly coupled through the myoendothelial gap junctions to 15 SMCs. Each SMC was assumed to be connected with 15 ECs arranged parallel under the smooth muscle layer. Assuming circumferential symmetry, this was modeled by rescaling myoendothelial fluxes affecting individual SMCs rather than by addition of actual ECs into the model. Homocellular communication within endothelial and smooth muscle layers was implemented similarly to the myoendothelial coupling as described in the previous section. To simulate NO synthase inhibition applied in the experiments with isolated rat mesenteric arteries [5], the NO pathway was blocked in the simulations. Figure 2 shows electrical equivalent diagram of the multicellular vessel model. The model has electrically sealed ends and negligible extracellular resistance.

3 Results

3.1 EC Calcium Dynamics

The EC model integrates both EC Ca^{2+} dynamics and plasmalemmal electrical activity to investigate EC responses to various stimulatory conditions and the interdependency of Ca^{2+} and V_m . The model reproduces experimentally observed EC V_m responses to volume-sensitive anion channel inhibitors (Fig. 3a)(compare with Fig. 11D in [6]), and hyperpolarizes to moderately increased extracellular potassium concentration (not shown). In addition, simulated Ca^{2+} transients during agonist stimulation agree qualitatively with experimental data, both under control and Ca^{2+} -activated potassium channel blockade conditions (Fig. 3b)(compare with Fig. 6c in [7]). Simulations predict both inward rectifier potassium and volume-sensitive anion channels as major determinants of resting V_m , and intracellular Ca^{2+} transient profiles are modulated but not determined by V_m . Model sensitivity analysis reveals that V_m has distinct effects on each part of the Ca^{2+} signal, contributing significantly to the control of resting and plateau $[\text{Ca}^{2+}]_i$ levels. Furthermore, the model predicts the PMCA Ca^{2+} extrusion pathway is the main regulator of Ca^{2+} homeostasis in the cell and may act as a Ca^{2+} “buffer” that masks $[\text{Ca}^{2+}]_i$ changes due to V_m alterations. This heterogeneous V_m role on Ca^{2+} signaling, which can be largely modulated by PMCA, may shed light on the controversy found experimentally about the interdependency between Ca^{2+} and V_m profiles.

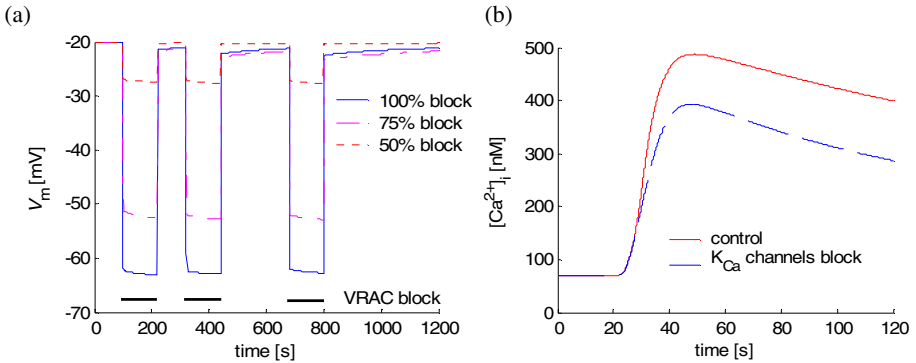


Fig. 3. Simulated isolated EC V_m and Ca^{2+} dynamic responses. (a) Effect of 50, 75 and 100% reduction in VRAC conductance on resting membrane potential. (b) Agonist-induced Ca^{2+} dynamics during control and K_{Ca} channels blockade conditions.

3.2 SMC Calcium Dynamics

Representative simulation results are shown in Fig. 4. Model validation against experimental data from [8] is presented in Fig. 4a. Model simulations are utilized to analyze system’s behavior (Fig. 4b) and to make experimentally testable predictions (Fig. 4c).

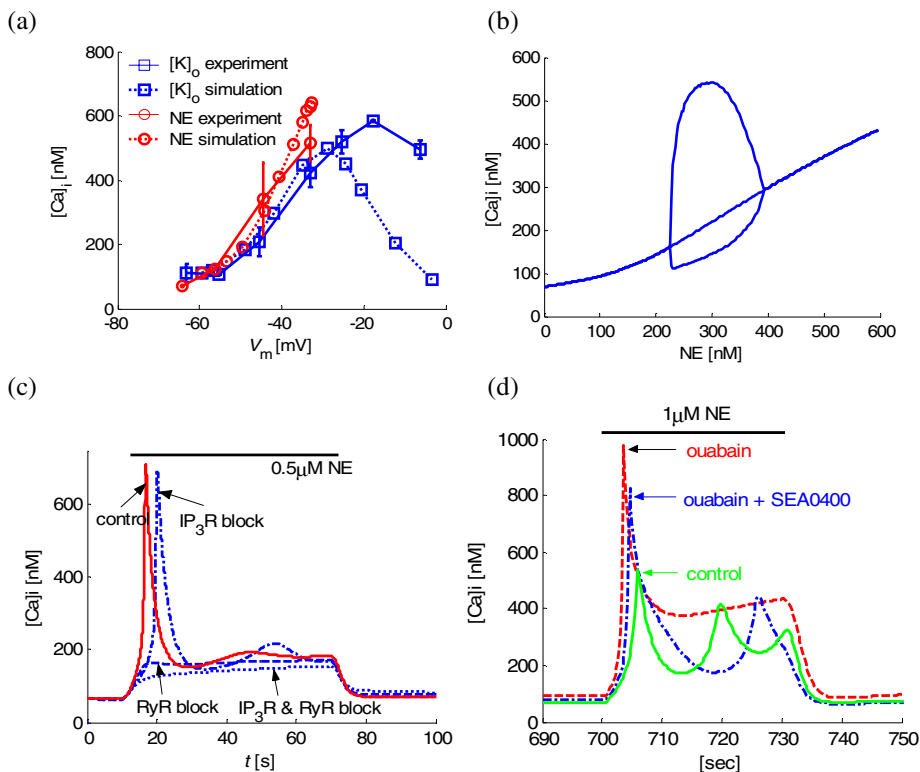


Fig. 4. Representative isolated SMC model simulations. (a) Model validation against experimental data for $[Ca^{2+}]_i$ and membrane potential (V_m) following stimulation with NE and extracellular K^+ . (b) Bifurcation diagram demonstrating Ca^{2+} and NE concentration window for oscillatory behavior. (c) Model predictions for NE induced Ca^{2+} transients after blockade of RyR and/or IP_3R . (d) Theoretical predictions for $[Ca^{2+}]_i$ after NE stimulation in the presence and absence of ouabain, and after inhibition of the reverse mode of NCX with SEA0400 compound. Model predicts a beneficial effect from NCX inhibition in restoring calcium dynamics when ouabain is present.

The SMC model was utilized to investigate Blaustein's hypothesis for the role of ouabain-induced Na^+/K^+ pump inhibition in salt sensitivity [9]. Fig. 4d compares model responses to $1 \mu M$ NE in the presence and absence of ouabain. The effect of ouabain was modeled by 50% reduction of the maximum rate of the Na^+/K^+ pump. The reduction caused intracellular Na^+ accumulation and a $\sim 20\%$ increase in resting $[Ca^{2+}]_i$. This Ca^{2+} rise was mainly due to NCX working now in the reverse mode, and not due to activation of the L-type channels, since the membrane depolarization was insignificant in agreement with data from. This rise in cytoplasmic calcium increased calcium sequestration into the sarcoplasmic reticulum by 22%. This corresponds to 890-fold Ca^{2+} amplification (i.e. due to buffering) and is in the same range with 2500-fold amplification predicted by Blaustein. This results in a NE-induced Ca^{2+} transient (dashed line; Fig. 4d) that is significantly enhanced relative to control (solid line), due

to the increased availability of Ca^{2+} in the stores and lack of extrusion through the NCX. Blockade of the NCX reverse mode nearly eliminated ouabain-induced resting cytosolic Ca^{2+} elevation and hyperresponsiveness to NE (dash-dot line; Fig. 4d).

3.3 Integrated EC/SMC Model

The model simulates experimentally observed KCl and NE-induced oscillations, endothelium-dependent relaxation of prestimulated SMC, and the role of individual coupling components (i.e. NO, IP_3 , gap junctions) in system responses (data not shown). The integrated model can be utilized to analyze the effect of ouabain-induced Na^+/K^+ pump inhibition in salt sensitivity and to investigate optimal therapeutic strategy. Increased myogenic tone, augmented effect of vasoconstrictors, and impaired endothelium-dependent relaxation or response to exogenous NO donors have been observed in different isolated arterial segments after inhibition of Na^+/K^+ pump. Fig. 5 compares model SMC cytosolic calcium concentrations at rest, after stimulation with 300nM NE and during EC-dependent relaxation under control conditions (solid line) and after 20% reduction of Na^+/K^+ pump rate in both EC and SMC (dashed line). As in the case of isolated SMC, this inhibition caused intracellular Na^+ accumulation, reversal of NCX mode and elevation of Ca^{2+} at rest and during NE stimulation. Endothelium-dependent relaxation was significantly compromised.

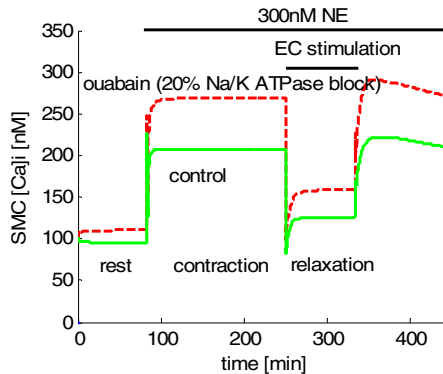


Fig. 5. Integrated EC/SMC model. SMC cytosolic calcium concentrations at rest, after stimulation with 300nM NE and during EC-dependent relaxation under control conditions (solid line) and after 20% reduction of the maximum rate of the Na^+/K^+ pump in EC and SMC (dashed line).

3.4 Vessel Model

Figure 6 shows calcium response of the vessel model to local stimulation of the endothelium. The vessel was uniformly precontracted with 200nM NE, which depolarized smooth muscle membrane potential and increased smooth muscle calcium concentration. Agonist stimulation of a single endothelial cell (EC number 5 at Fig. 6) hyperpolarized the cell and relaxed coupled SMCs, similarly to the integrated EC/SMC

model. In addition, the hyperpolarization was conducted through the endothelial gap junctions to other ECs, and to their adjacent SMCs through the myoendothelial gap junctions. Thus the whole vessel segment was rapidly hyperpolarized and the calcium concentration in the smooth muscle layer was reduced, corresponding to relaxation. Significant calcium elevation in the endothelium was confined to the stimulation site, indicating calcium-independent spread of hyperpolarization. These theoretical results are in agreement with experimental data from [5].

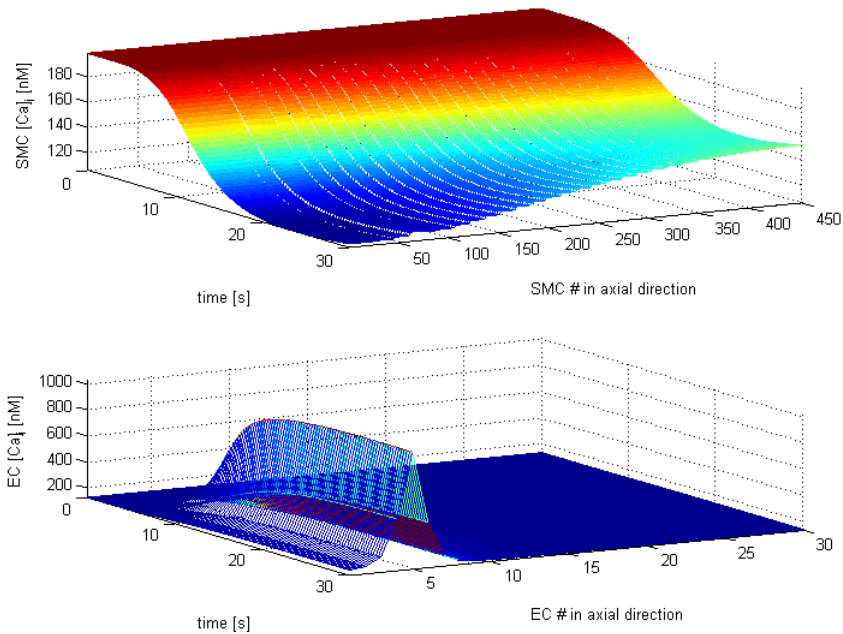


Fig. 6. Smooth muscle (top) and endothelial (bottom) calcium response of the vessel model to a local agonist stimulation of the endothelium

4 Discussion

A significant amount and a wide variety of biological and biomedical data are already available and more are continuously generated at increased rates. This vast amount of information will enable us to increase our understanding of human physiology and improve medical practice. These data can be fully utilized to provide quantitative predictions if they are combined with mathematical modeling. A general strategy for integrative computational physiology that will link model parameters at one scale to more detailed description at the level below is being developed under the Physiome Project [10]. Computational physiology is well advanced in the heart, but much less or no progress has been made for other organs or tissues. In the case of the vascular system, ample experimental studies investigate signaling pathways involved in the regulation of vascular tone, but there are far less theoretical studies to assist in the

analysis of experimental data. Our objective is to further advance currently available vascular models. In this paper, we presented general methodology and some of the results from the isolated, integrated and multicellular vessel models.

A major difficulty in the model development is to obtain all the necessary primary data, such as channel parameters or signaling pathways involved. Retrieval of the information from the literature is now facilitated by the on-line access to the journals and search options in PubMed, but still requires considerable effort and time since the data are highly dispersed. Mathematical models can serve as a database for this kind of biological data. In addition, despite the significant amount of data on the vascular system and particularly on the rat mesenteric microcirculation, not all the necessary information is currently available. The model will highlight important system components and parameter values that merit further experimentation. At the current stage, some parameters had to be assumed or taken from other vascular beds and species. This somewhat limits the model's predictive ability. Nevertheless model outputs were validated and shown to be in good agreement with experimental data from both isolated cells and vessel segments.

Model simulations suggest that ouabain induced inhibition of NaK will result in vessel hyperreactivity to NE (Fig. 4d) while the ability of NO to relax the SMC may be compromised (Fig. 5). Interestingly, blockade of the NCX reverse mode nearly compensated ouabain-impaired SMC Ca^{2+} dynamics (Fig. 4d). This is in agreement with recent data by Iwamoto and colleagues that demonstrate the beneficial effects of SEA0400, a recently developed specific blocker of Ca^{2+} entry mode of NCX type 1, in reversing ouabain-induced $[\text{Ca}^{2+}]_i$ elevation and contraction in mouse small mesenteric arteries [11]. The model predicts also that agents that block both forward and reverse modes of NCX will be less effective at reducing ouabain-induced hyperreactivity, since the forward mode can limit the initial large Ca^{2+} spike. The model also suggests the endothelial layer as the main route for transmission of the hyperpolarizing current in conducted vasomotor responses along the vessel axis (Fig. 6).

Overall, simulations demonstrate system responses with high complexity (e.g. oscillations), that result from the many complex and nonlinear interactions of the individual components. A quantitative theoretical framework, like the one developed in this study, provides us with a significant advantage in system analysis and in predicting physiological behavior under different scenarios. The mathematical model outlined in this study presents a significant step towards the development of an integrative computational vascular model that could be utilized for generating clinically testable hypotheses, interpretation of vascular proteomic data or investigating the impact of gene related channelopathies.

Acknowledgments. This project was supported by the American Heart Association grant NSDG043506N.

References

1. Winslow, R. L., and M. S. Boguski. 2003. Genome informatics: current status and future prospects. *Circ Res* 92(9):953-961.
2. Winslow, R. L., S. Cortassa, and J. L. Greenstein. 2005. Using models of the myocyte for functional interpretation of cardiac proteomic data. *J Physiol* 563(Pt 1):73-81.

3. Rudy, Y. 2006. Modelling and imaging cardiac repolarization abnormalities. *J Intern Med* 259(1):91-106.
4. Tsoukias, N. M., M. Kavdia, and A. S. Popel. 2004. A theoretical model of nitric oxide transport in arterioles: frequency- vs. amplitude-dependent control of cGMP formation. *Am J Physiol Heart Circ Physiol* 286(3):H1043-1056.
5. Takano, H., K. A. Dora, M. M. Spitaler, and C. J. Garland. 2004. Spreading dilatation in rat mesenteric arteries associated with calcium-independent endothelial cell hyperpolarization. *J Physiol* 556(Pt 3):887-903.
6. Nilius, B., and G. Droogmans. 2001. Ion channels and their functional role in vascular endothelium. *Physiol Rev* 81(4):1415-1459.
7. McSherry, I. N., M. M. Spitaler, H. Takano, and K. A. Dora. 2005. Endothelial cell Ca^{2+} increases are independent of membrane potential in pressurized rat mesenteric arteries. *Cell Calcium* 38(1):23-33.
8. Nilsson, H., P. E. Jensen, and M. J. Mulvany. 1994. Minor role for direct adrenoceptor-mediated calcium entry in rat mesenteric small arteries. *J Vasc Res* 31(6):314-321.
9. Blaustein, M. P. 1993. Physiological effects of endogenous ouabain: control of intracellular Ca^{2+} stores and cell responsiveness. *Am J Physiol* 264(6 Pt 1):C1367-1387.
10. Hunter, P., and P. Nielsen. 2005. A strategy for integrative computational physiology. *Physiology (Bethesda)* 20:316-325.
11. Iwamoto, T., S. Kita, J. Zhang, M. P. Blaustein, Y. Arai, S. Yoshida, K. Wakimoto, I. Kuro, and T. Katsuragi. 2004. Salt-sensitive hypertension is triggered by Ca^{2+} entry via $\text{Na}^+/\text{Ca}^{2+}$ exchanger type-1 in vascular smooth muscle. *Nat Med* 10(11):1193-1199.

Searching and Visualizing Brain Networks in Schizophrenia

Theofanis Oikonomou^{1,2}, Vangelis Sakkalis¹,
Ioannis G. Tollis^{1,2}, and Sifis Micheloyannis³

¹ Institute of Computer Science, Foundation for Research and Technology-Hellas,
Vassilika Vouton, P.O. Box 1385, Heraklion, GR-71110 Greece
`{thoikon,sakkalis,tollis}@ics.forth.gr`

<http://www.ics.forth.gr>

² Department of Computer Science, University of Crete,
P.O. Box 2208, Heraklion, Crete, GR-71409 Greece
`{thoikon,tollis}@csd.uoc.gr`

<http://www.csd.uoc.gr>

³ Clinical Neurophysiology Laboratory (L. Widen), Faculty of Medicine,
University of Crete, Heraklion, Crete, GR-71409 Greece
`michelogj@med.uoc.gr`

Abstract. There has been special interest lately in using graph theory to study brain networks, as it provides the theoretic and visualization means to study the "disconnection syndrome" for schizophrenia. In this work we try to visualize the graphs derived from electroencephalographic (EEG) signals using several graph drawing techniques and incorporate them smoothly into an easy-to-use framework. The aim is to reveal and evaluate important properties of brain networks.

1 Introduction

Cognitive disorganization is one of the defining, and most disabling, symptoms of schizophrenia. Disturbances in "functional connectivity" have been proposed as a major pathophysiological mechanism for schizophrenia, particularly for cognitive disorganization. The disconnection hypothesis [1,2] and working memory (WM) deficits [3,4] are well established in the literature on schizophrenia. The use of graph theory to study brain networks has drawn much attention recently, since it offers a unique perspective of studying local and distributed brain interactions [5,6]. Both local and long distance functional connectivity in complex networks is evaluated using measures and visualizations derived from graph theory. In this work we study and visualize the graphs derived from EEG signals, of twenty stabilized patients with schizophrenia vs controls using several graph drawing techniques, hoping that this would highlight important properties of brain networks.

The remaining of this paper is organized as follows: We introduce basic information regarding the way we construct the graphs in Sect. 2. In Sect. 3 we discuss the methods we use to analyze and visualize the resulting graphs followed

by some screenshots of these algorithms. Final remarks are discussed in Sect. 4 which concludes this paper.

2 Preliminaries

In this study two different situations are considered: the control (Rest), where subjects had the eyes fixed on a "star" on the computer screen and the cognitive activation during WM while performing a two-back¹ test using capital Greek letters. We tested three different groups. The first group consists of control university educated subjects (CE), the second one consists of control subjects without higher education (CU) and the third one consists of stabilized patients with schizophrenia (P). The testing hypothesis suggests that a WM task requires considerable mental effort and the disconnection on neuronal assemblies in patients could be visible. In order to acquire the raw data the EEG signals in all three groups were recorded from $N = 30$ cap electrodes, according to the 10/20 international system [7], referred to linked earlobe electrodes (Fig. 1).

The spatial pattern of functional connectivity was assessed by computing the Wavelet Coherence (WC) of EEGs [8]. This method yields a statistical coherence measure ranging from 0 to 1, which is an indication of how much a specific electrode is correlated with each of the other electrodes. Thus, we come up with an $N \times N$ coherence matrix (CM) with elements ranging from 0 to 1 formulated per task and subject. In order to obtain a graph from a CM we need to convert it into an $N \times N$ binary adjacency matrix, A . To achieve that we define a variable called threshold T , such that $T \in [0, 1]$. The value $A(i, j)$ is either 1 or 0, indicating the presence or absence of an edge between nodes i and j , respectively. Namely, $A(i, j) = 1$ if $CM(i, j) \geq T$, otherwise $A(i, j) = 0$. Thus we define a graph for each value of T . For the purposes of our work we defined 1000 such graphs, one for every thousandth of T .

3 Methods

In our attempt to visualize the graphs produced by the above-mentioned technique we use several methods and graph visualization algorithms. First, we develop a static method which helps the doctors understand the inter-connectedness of the electrodes. Additionally, we invoke some well known graph algorithms in anticipation of a better, compared to the static method's, visualization outcome. These include force-directed and circular drawing algorithms. Finally, we develop a framework for plotting important properties of the emerged graphs.

3.1 Static Visualization Method

As described in Sect. 2, 30 electrodes were used during the experiments. In order to visualize the topology of the emerged network we create a static framework

¹ In the 2-back condition, the target letter was any letter that was identical to the one presented two trials before it.

where each electrode is depicted by a node placed in a position similar to the actual electrode's position on the human cortex. Thus, we manage to depict the brain network (Fig. 1). This resulted in 1000 graphs to depict for each of the 40 healthy (educated and uneducated) and 20 patient subjects, for each of the two states, that is during Rest and WM, and for each of 7 frequency bands, in terms of functional uniformity, acquired during the experiments, namely $\delta[0.5 - 4Hz]$, $\theta[4 - 8Hz]$, $\alpha_1[8 - 10Hz]$, $\alpha_2[10 - 13Hz]$, $\beta[13 - 30Hz]$, $\gamma_1[30 - 45Hz]$, $\gamma_2[45 - 90Hz]$.

That is a large number of networks to be displayed simultaneously on the screen. This is why we give the user the opportunity to choose which graph he wants to see. Namely, one can decide the subject, state, band and threshold independently (Fig. 2(a)) and in addition one can choose to automatically iterate through thresholds for a given subject, state and band (Fig. 2(b)).

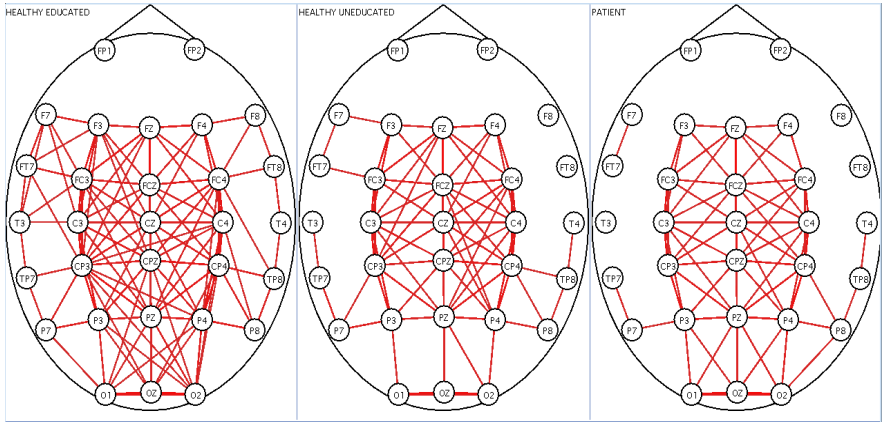


Fig. 1. Static cortex coordinates screenshot

Furthermore, we need a way to visually differentiate the edges between two nodes based on the threshold. To this end we introduce a color coding shown in Fig. 2(b), where an edge's color turns to red if the nodes it connects are highly correlated. On the other hand the edge's color moves to light blue. As an additional aid the more correlated two nodes are, the thicker the edge that connects them. So, when one iterates through the thresholds the most correlated nodes can easily be spotted as the edges that connect them are thicker and more red. The screenshots in the following of the paper are created for the mean case of each subject group, the γ_1 band, during WM and $T = 0.820$.

This visualization is very useful to the doctors as they can see how the interconnectedness in the brain changes and identify the critical thresholds where this happens. Furthermore, they are able to verify already known properties and perhaps discover new ideas concerning the reasons for some disorders.

The proposed graphical framework is particularly useful as compared to the current clinical practice where the doctor can only observe the EEG between two specific nodes and at best have a few such EEGs on the screen. This is a severe drawback as they have to manually search for some prominent pair of nodes that according to the literature are held responsible for a specific disorder. Furthermore they can not focus at a specific threshold and see what the brain network looks like.

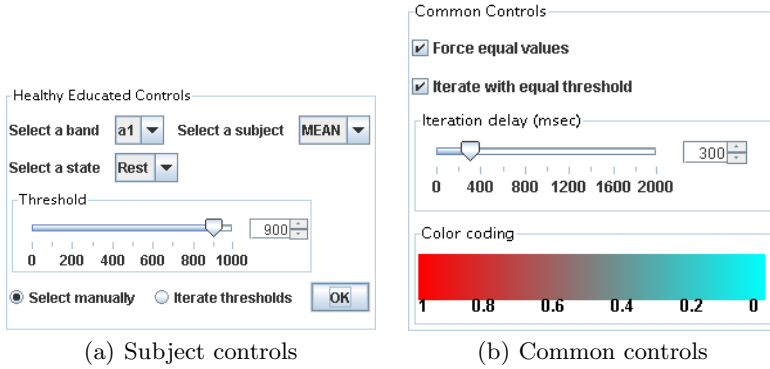


Fig. 2. Means to interact with the framework

3.2 Force Directed Algorithms

The previous approach is very useful for doctors as they can see the topology of their patients' brains and be able to interact in some degree. Our next thought is to use some of the best known force-directed methods hoping for a clearer visualization outcome that could be of greater assistance. First of all, we use the Fruchterman and Reingold method [9], which is a simple model based on a combination of springs and electrical forces. Next we utilize the Kamada and Kawai method [10,11], which is a more complex method that attempts to draw graphs such that the Euclidean distance between two vertices is near to the number of edges on the shortest graph-theoretic path between the vertices. Furthermore, the method of Eades is used [12].

The main drawback of using these methods is that the mapping of the nodes and the exact position of the electrodes on the brain is lost. In order to overcome this flaw, we color nodes that belong to the same lobe, which is a specific area of the human cortex, with the same color. We also introduce a tooltip with useful information about each node when one pauses over a specific node (Fig. 4(a)). What remains to be clarified is whether this approach could be potentially helpful to the doctors. The only knowledge that could be extracted from all the spring-embedder methods is the number of connected components and a nice placement of them.

3.3 Circular Drawing Algorithms

A circular graph drawing is a visualization of a graph with the following characteristics:

1. The graph is partitioned into clusters
2. The nodes of each cluster are placed on the circumference of an embedding circle
3. Each edge is drawn as a straight line segment

The problem of minimizing the number of crossings in a drawing is the well-known NP-Complete crossing number problem [13]. The more restricted problem of finding a minimum crossing embedding such that all the nodes are placed onto the circumference of a circle and all edges are represented with straight lines is also NP-Complete as proven in [14].

In [15,16], a linear time technique, CIRCULAR, to produce circular graph drawings of biconnected graphs on a single embedding circle was introduced. In order to produce circular drawings with fewer crossings the authors presented an algorithm which tends to place edges toward the outside of the embedding circle and nodes are placed near their neighbors. The worst-case time requirement of CIRCULAR is $O(m)$, where m is the number of edges. An important property of this technique is the guarantee that it will find a zero-crossing drawing for a given biconnected graph in case one exists.

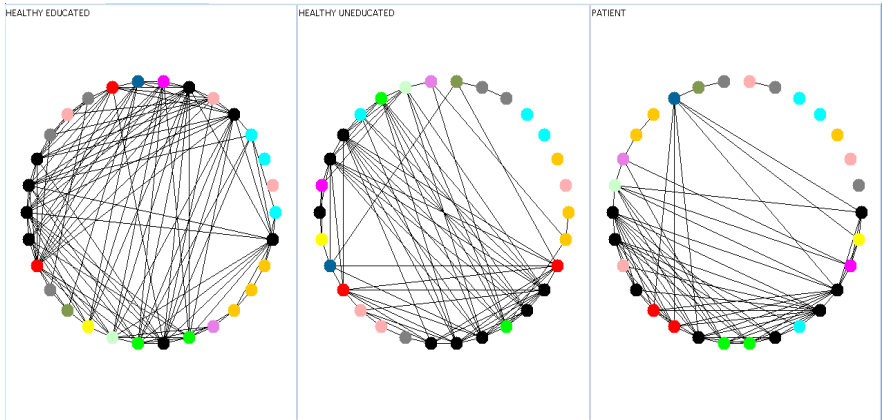


Fig. 3. Simple-Circular screenshot

In [15,17] another linear time algorithm, CIRCULAR-Nonbiconnected, was introduced for producing circular drawings of nonbiconnected graphs on a single embedding circle. Given a nonbiconnected graph G , it was first decomposed into biconnected components. In this technique, the layout of the resulting block-cutpoint tree on a circle was first produced and then the one for each biconnected

component with a variant of CIRCULAR. The worst-case time requirement for CIRCULAR-Nonbiconnected is $O(m)$ if we use a variant of CIRCULAR to layout each biconnected component. The resulting drawings have the property that the nodes of each biconnected component appear consecutively. Furthermore, the order of the biconnected components on the embedding circle are placed according to a layout of the accompanying block-cutpoint tree and therefore the biconnectivity structure of a graph is displayed even though all of the nodes appear on a single circle.

We incorporate the last algorithm in our framework. This is due to the fact that our networks are clustered and these clusters are important in identifying the regions of the brain that are active at the same time while performing a specific task. We implement two variants of this technique. In the first variant (Simple-Circular) we place the nodes according to the position assigned to them by the CIRCULAR-Nonbiconnected algorithm (Fig. 3).

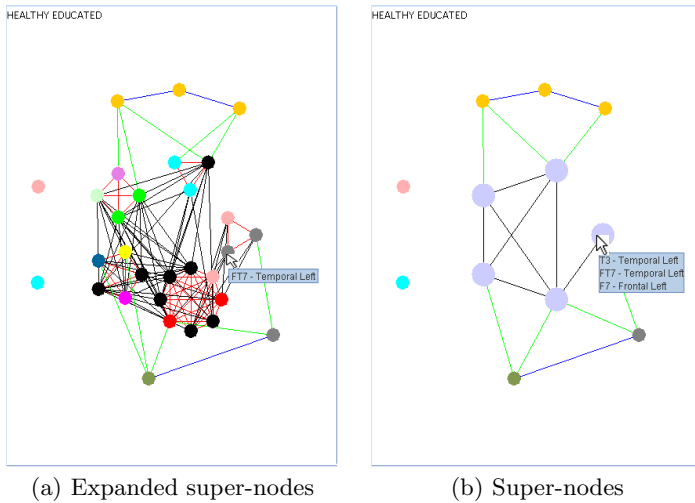


Fig. 4. Super-node screenshots with tooltips

In the second variant (Double-Circular) node-disjoint cliques in our graph are first identified and placed in the circumference of an inner cycle. These cliques can be displayed as super-nodes or can be decomposed to the vertices that form the clique, which are placed in a cycle with center the clique's position (Fig. 4). Additionally, by pausing on a super-node one gets a useful tooltip that gives information about the nodes that form the specific node and the lobe each node belongs to (Fig 4(b)). To find the cliques' positions we apply the CIRCULAR-Nonbiconnected algorithm in an effort to minimize the crossings among these important clusters. This leads to a clearer drawing that helps us understand how tightly connected areas of the brain, that are active while performing a specific WM task, interact with other tightly connected areas. By doing that we are

able to determine a specific node or set of nodes that are responsible for the poor connectivity or the disconnection of certain areas of the brain with each other. The remaining nodes, that do not belong to a clique, are placed in the circumference of an outer cycle. In order to maintain a clear drawing we place the nodes that are adjacent to some cliques in their mean angle (Fig. 5).

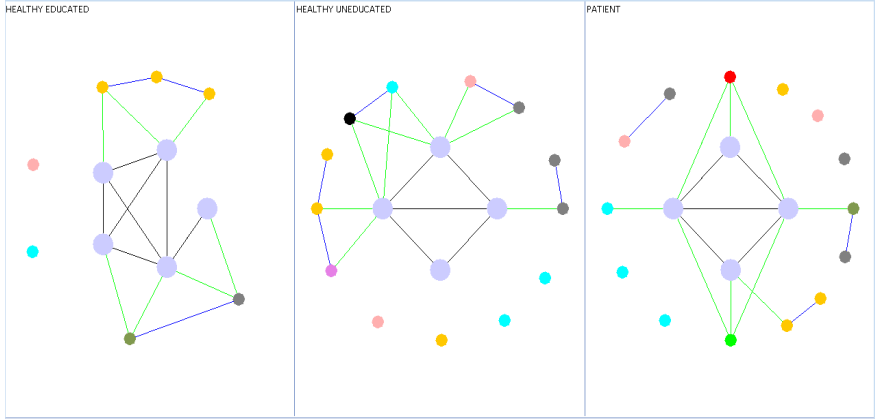


Fig. 5. Double-Circular screenshot

3.4 Charts

In addition to the visualization techniques we also use a framework for plotting several properties of the emerged graphs. These include the mean degree K , the clustering coefficient C and the average minimum path length L of our graphs.

First, let us define a graph in terms of a set of n nodes $V = v_1, v_2, \dots, v_n$ and a set of edges E , where e_{ij} denotes an edge between nodes v_i and v_j . Below we assume v_i, v_j and v_k are members of V . We define the neighborhood for a node v_i as its immediately connected neighbors, namely $N_i = \{v_j\} : e_{ij} \in E$. The degree k_i of a node is the number of vertices in its neighborhood $|N_i|$. The mean degree of a graph is the average degree over all nodes, thus

$$K = \frac{\sum_{i \in V} k_i}{n} \quad (1)$$

The clustering coefficient C_i for a node v_i is the proportion of links between the nodes within its neighborhood divided by the number of links that could possibly exist between them. For an undirected graph, if a node v_i has k_i neighbors, $\frac{k_i(k_i-1)}{2}$ edges could exist among the nodes within the neighborhood, thus

$$C_i = \frac{2|\{e_{jk}\}|}{k_i(k_i-1)} : v_j, v_k \in N_i \quad (2)$$

This measure is 1 if every neighbor connected to v_i is also connected to every other node within the neighborhood, and 0 if no node that is connected to v_i

connects to any other node that is connected to v_i . The clustering coefficient for the whole system is given by Watts and Strogatz [18] as the average of the clustering coefficient for each node,

$$C = \frac{\sum_{i=1}^n C_i}{n}, \quad (3)$$

and is a measure of the tendency of graph nodes to form local clusters.

The shortest path (distance) d_{ij} between two nodes v_i and v_j is the minimum number of links we need to traverse in order to go from node v_i to node v_j . The average shortest path length

$$L = \frac{\sum_{i,j \in V, i \neq j} d_{ij}}{n(n-1)} \quad (4)$$

is the average shortest path (distance) connecting any two nodes of the graph and is a measure of the interconnectedness of the graph. Note that, in our experiments, the absence of a path between v_i and v_j implies $d_{ij} = 1000$.

In order to find out the way the above mentioned properties vary as a function of the threshold T we conducted some experiments. Our study here focuses on *gamma1* band coherence analysis. The average values of K , C and L during WM for $0.75 \leq T \leq 0.85$ with step 0.005 were computed (three selected values shown in Table 1). We concentrate in different values of T , where the values of K and C of schizophrenic patients are equal to those of control subjects. For the above values of T , the respective values of L for patients are much greater than those for controls.

Table 1.

HEALTHY EDUC.				PATIENT			
T	K	C	L	T	K	C	L
0.8	11.03	0.71	100.95	0.755	11.35	0.7	163.55
0.805	10.2	0.68	100.99	0.765	10.05	0.64	163.71
0.84	5.11	0.43	224.78	0.815	5.08	0.43	383.99

The three graph measures K , C , L actually represent an overall signature of the graph topology. Our experiments indicate that K and C are getting lower and L is getting higher while moving from healthy to schizophrenics in the whole threshold range (Fig. 7). But instead of studying each measure independently, we attempt to quantify their interaction. Towards this direction we determine three different values of T (see Table 1), where the values of K and C of patients are almost equal to those of healthy. The physical meaning of this maneuver addresses the question whether the network is proportionally efficient assuming both healthy and schizophrenic populations have the same average degree and clustering coefficient. According to Table 1 the answer is no, which means that for the above values of T the respective values of L of the patients are much

greater than those of healthy. This is also evident by observing Fig. 1. The latter syllogism leads to the suggestion that schizophrenic patients need significantly more direct node (channel) connections in order to perform the same WM task.

This framework is very useful as it provides many functionalities. First of all, one can determine which band, state and subject group to plot (Fig. 6). Additionally, one can plot the above mentioned properties and be able to observe the way they alter as the threshold increases. Furthermore, by performing a statistical test, like t-test, to the input data, one can identify the threshold regions, where a statistically important difference exists. This is indicated by a vertical black marker as shown in Fig. 7.

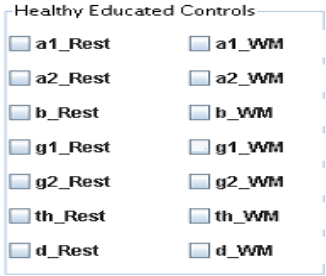


Fig. 6. Chart controls

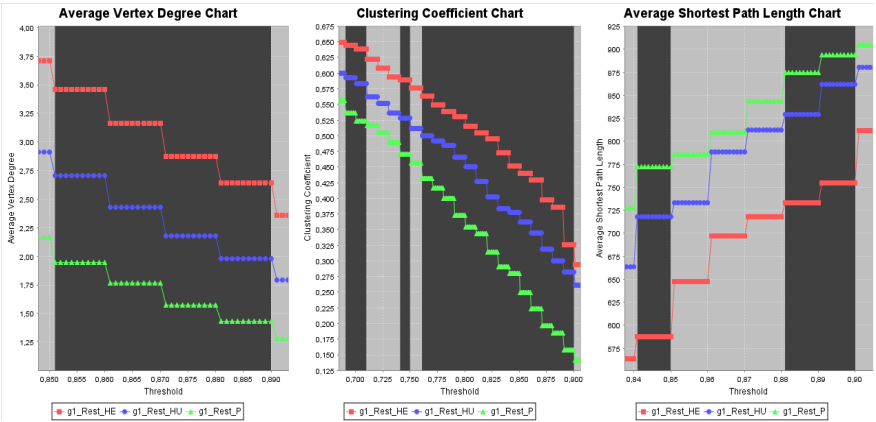


Fig. 7. Chart screenshot. The dark gray columns are markers for indicating the statistically important threshold regions.

The last feature is particularly important. The statistically important threshold regions only offer information about the overall behavior of the network. For example in the case where the graph of the HE group has greater C compared to the graph of the P group for some band, state and threshold, the former

graph has better local organization than the latter. However, we do not get any details as to where this better organization lies in the brain, which nodes make this happen, in what way these nodes are connected to the rest of the graph and generally what the two networks look like. In order to get this information and understand its impact to the topology of the graph, one should be able to see a static or a graph drawing visualization of the graph at these important threshold regions. This is accomplished by clicking on the markers. Getting to the visualization framework, it automatically focuses on the specified threshold. That reveals the exact interconnectedness which results in the aforementioned difference, making it easier for the doctors to discover where the problem might be.

4 Conclusions

An insight into the networks formed in the human brain is given using the graph drawing methods mentioned in the previous section, for the first time. The static method was a good initial attempt. It gave the doctors the opportunity to see how all the nodes (electrodes) correlate with each other in a visual, quick and easy to use way compared to the manual and time consuming techniques used in the past. It would be interesting to see if a 3-dimensional version of the static method could further assist the doctors.

On the other hand, the visualizations yielded from the force directed methods still need to be evaluated, since doctors were only able to identify the structure of the underlying network, as they only revealed the connected components that formed the graphs and made a clearer drawing of them possible.

However, the two circular variants were more useful. The first variant showed with more detail the structure of the network, as doctors could easily identify the biconnected components of the graph. The second variant was of greater help as it revealed the important areas of the brain, which are co-activated. Additionally, it showed the way these areas are linked to each other and gave a hint as to where to search for the disorders present in schizophrenia. An interesting modification of the second variant would be to place in the inner cycle different kind of nodes rather than the node disjoint cliques. These nodes could be the most highly ranked nodes according to some metric or could even be user defined taking advantage of the doctors' expertise.

Finally, the plotting framework gave the doctors the opportunity to see how important metrics alter and how this influences the network's topology as one can change the focus from the graph's overall state to the detailed node interconnectedness.

Acknowledgements

We would like to thank E. Pachou, Th. Adrakta, E. Fazakis and P. Bitsios for acquiring the raw data and performing the detailed psychiatric evaluation of the patients.

References

1. K. H. Lee, L. M. Williams, M. Breakspear, and E. Gordon, "*Synchronous gamma activity: a review and contribution to an integrative neuroscience model of schizophrenia*", Brain Res. Brain Res. Rev., vol. 41, pp. 57-78, 2003.
2. N. C. Andreasen, S. Paradiso, and D. S. O'Leary, "*Cognitive dysmetria as an integrative theory of schizophrenia: a dysfunction in corticallsubcortical- cerebellar circuitry?*", Schizophr Bull, vol. 24, pp. 203- 218, 1998.
3. H. M. Conklin, C. E. Curtis, M. E. Calkins, and W. G. Iacomo, "*Working memory functioning in schizophrenia patients and their first-degree relatives: cognitive functioning shedding light on aetiology*", Neuropsychologia, vol. 43, pp. 930-942, 2005.
4. H. Silver, P. Feldman, W. Bilker, and R. C. Gur, "*Working memory deficit as a core neuropsychological dysfunction in schizophrenia*", Am. J. Psychiatry, vol. 160, pp. 1809-1816, 2003.
5. F. Varela, J. P. Lachaux, E. Rodriguez, and J. Martinerie, "*The brainweb: phase synchronization and large-scale integration*", Nat. Rev. Neurosci., vol. 2, pp. 229-239, 2001.
6. A. A. Fingelkurts and S. Kähkönen, "*Functional connectivity in the brain—is it an elusive concept?*", Neurosci. Biobehav. Reviews, vol. 28, pp. 827-836, 2004.
7. H. H. Jasper, "*The 10-20 electrode system of the International Federation in Electroencephalography and Clinical Neurophysiology.*", EEG Journal, 10, 370-375, 1958.
8. V. Sakkalis, T. Oikonomou, E. Pachou, I. G. Tollis, S. Micheloyannis, and M. Zervakis, "*Time-significant Wavelet Coherence for the Evaluation of Schizophrenic Brain Activity using a Graph theory approach*", accepted for publication, IEEE-EMBS, New York City, USA, 2006.
9. T. Fruchterman, and E. Reingold, "*Graph Drawing by Force-Directed Placement*", Softw.-Pract.Exp., 21, no.11, 1129-1164, 1991.
10. T. Kamada, and S. Kawai, "*An Algorithm for Drawing General Undirected Graphs*", Inform. Process. Lett., 31, 7-15, 1989.
11. T. Kamada, "*Visualizing Abstract Objects and Relations*", World Scientific Series in Computer Science, 1989.
12. P. Eades, "*A Heuristic for Graph Drawing*", Congr. Numer., 42, 149-160, 1984.
13. M. Garey, and D. Johnson, "*Computers and Intractability: A Guide to the Theory of NP-Completeness*", Freeman, 1979.
14. S. Masuda, T. Kashiwabara, K. Nakajima, and T. Fujisawa, "*On the NP-Completeness of a Computer Network Layout Problem*", Proc. IEEE 1987 International Symposium on Circuits and Systems, Philadelphia, PA, pp. 292-295, 1987.
15. J. M. Six (Urquhart), "*Vistool: A Tool For Visualizing Graphs*", Ph.D. Thesis, The University of Texas at Dallas, 2000.
16. J. M. Six, and I. G. Tollis, "*Circular Drawings of Biconnected Graphs*", Proc. of ALENEX '99, LNCS 1619, Springer-Verlag, pp. 57-73, 1999.
17. J. M. Six, and I. G. Tollis, "*Circular Drawings of Telecommunication Networks*", Advances in Informatics, Selected Papers from HCI '99, D. I. Fotiadis and S. D. Nikolopoulos, Eds., World Scientific, pp. 313-323, 2000.
18. D. J. Watts, and S. H. Strogatz, "*Collective dynamics of 'small-world' networks*", Nature, 393, pp. 440-442, 1998.

TRENCADIS – A Grid Architecture for Creating Virtual Repositories of DICOM Objects in an OGSA-Based Ontological Framework

Ignacio Blanquer, Vicente Hernandez, and Damià Segrelles

Instituto de Aplicaciones de las Tecnologías de la Información y de las Comunicaciones
Avanzadas, Universidad Politécnica de Valencia, Camino de Vera S/N, 46022 Valencia Spain
Phone: 963877007 ext.88254
{iblanque,vhernand}@dsic.upv.es, dquilis@itaca.upv.es

Abstract. The creation of virtual repositories of medical data is a very important challenge to ease collaboration, research and training among medical organisations. However, there are several technical and legal problems, such as efficient large data distribution, privacy protection, post-processing and knowledge management. The TRENCADIS project is aiming at the development of an infrastructure able to tackle with such problems using Grid Technologies. This article presents the Grid Software Architecture developed, which is implemented on top of the OGSA specification and defines the mechanisms to create virtual repositories of DICOM objects. It integrates different repositories providing a single-database virtual view through high-level components. The TRENCADIS architecture proposes the use of ontologies and templates to organise the DICOM data of Structured Reports. The TRENCADIS architecture is being used for the development of a cyberinfrastructure for medical imaging on oncology in the land of Valencia, with the participation of seven hospitals.

Keywords: OGSA, DICOM, Grid, Grid Services, Ontology.

1 Introduction and Motivation

The use of digital medical images in hospital environments are changing the way in which radiologist work and cooperate. The generalisation of Digital Imaging and Communications in Medicine [1] (DICOM), as a world-wide standard for the transmission and exchange of medical images, has made it possible to share images across a wide set of users and applications.

Moreover, DICOM does not only imply images coming from the radiology departments, other type of images and movies (dermatology, endoscopes, etc.) and other type of information such as radiology reports are being coded into DICOM. DICOM Structured Reporting [2] (DICOM-SR) codes and integrates radiology reports with seamless references to findings and Regions of Interests on the associated images. Structuring radiology reports offers a comparable way to code reports enhancing the capability of tools to search and to extract knowledge.

The work presented in this article proposes an architecture and a set of services implemented on it, as a solution for interconnecting selected parts of medical data for the development of training and decision support tools. The organisation of the distributed information in virtual repositories is based on semantic criteria. Different groups of researchers could organise themselves to propose a Virtual Organisation (VO). These VOs will be interested in a specific target area (e.g. paediatric oncology), and will share information (studies and reports) concerning this area. Subsets of those images could be obtained for a specific study (e.g. neuroblastoma). Finally, in each subset, users can make complex queries (e.g. male patients above one year with irregular findings of more than 3 mm). Of course, the sharing of the information must be secure and within the VO (even without private data) and images and reports must be structured in a way to enable concept-based searching.

2 Objectives

The main objective of TRENCADIS (Towards a gRid Environment for proCessing and shAring DIcom objectS) is the design and implementation of a middleware that enables sharing DICOM objects (images, reports, movies) located in different geographically-distributed locations, through the virtual view of an integrated repository, organised on an ontology framework. In particular, the main objective of this paper is to show the architecture structure and the parts and requirements of the services that can be implemented above this. The main features of the architecture are the following:

- Service Oriented Architecture (SOA) based in Open Grid Services Architecture [3] (OGSA), that provides the services needed for sharing the information contained in different DICOM repositories using different ontologies.
- Implementation of the services using the infrastructure Web Services Resources Framework [4] (WSRF) of OGSA specifications.

This paper also describes other issues related to the TRENCADIS architecture:

- Design of a language for the specification of the ontological schemas that structure the repositories.
- Definition and implementation of middleware components for integrating DICOM repositories and sharing DICOM objects using different ontologies.

Other objectives of TRENCADIS project are described in other publications [5][6].

3 State of the Art

This section describes the status of the development in the different key blocks of the architecture of TRENCADIS, considering the DICOM standard and metalanguages.

3.1 DICOM

The DICOM standard is a world-wide accepted specification for the connection, transferring and coding of medical images. DICOM was born in the eighties in a

collaboration of the American College of Radiology (ACR) and the National Electrical Manufacturers Association (NEMA) of the United States.

Currently, DICOM standard provides specification to store radiology images and other kind of medical digital information, such as movies (ultrasound, endoscope videos, etc.), biosignals or even structured text.

3.2 Metalanguages (XML)

An important component of this work is the support to a knowledge-oriented organisation of the information. This is achieved through the use of ontologies defined in metalanguages that specify the information of interest of the user communities and the structure of the reports. The use of metalanguages enables reaching a higher degree of independence and eases the interaction among components. In information technologies, ontology is a vocabulary and a set of terms, rules and relations that define with the needed accuracy a set of entities enabling the definition of classes, hierarchies and other relations among them. The ontologies define the terms to be used to describe and represent a knowledge domain. In this sense, the ontologies organise the knowledge in a way that makes it reusable. In the case of TRENCADIS, the ontologies define the fields and the structure in which the information must be specified in the DICOM objects. This structure will enable the creation of index tables that reference subsets of the information according to communities, experiments and searching results. Many languages are used for the specification of ontologies. The Extended Mark-up Language [7] (XML), the Resource Description Framework [8] (RDF), and the Ontology Web Language [9] (OWL) are good examples that enable the web to be used as a global infrastructure, to share data and documents, and to define knowledge-compatible processing services.

The TRENCADIS architecture only needs to define the ontologies and the vocabularies that provide data and documents with a compatible structure. Since services do not need to provide a semantically interoperable interface, RDF or OWL approaches are not necessary and XML can be directly used.

4 Architecture TRENCADIS

The main objective in this paper has been the definition of a secure SOA Architecture. The TRENCADIS architecture is structured into five layers. They are the following:

4.1 Core Middleware Layer

The Core Middleware Layer is related with the lowest abstraction level of the architecture. This layer contains the basic services that directly interact with the physical resources providing the homogeneous interface of a logical virtual resource. This logical resource requires the consideration of the following points:

Definition of the services under the OGSA specification. All basic services are implemented using Grid Services under the OGSA and the WSRF specifications.

Implementation of the resource interaction component. Each type of resource needs a specific component that hides their particularities. These components interact

with the Grid and must translate the higher-level requests into device-understandable actions. The component provides the needed functionality to query and modify the status of the resource, among other functionalities that depends on the device.

Definition, processing and validation of the I/O format of the logic processes. The common interface requires not only the translation of the requests into resource understandable operations, but also the coding and decoding of the information input and output. This requires the definition of schemas (implemented in XML) which are stored and are used to validate the XML-coded input and output data.

Implementation of the logic processes of the Grid Service. The functionality of the logic resources is mainly to query and modify the status of any of the attributes of the physical resource. Thus this functionality must be implemented in all components in an analogous form and using the same interface. There could be, however, different degrees of functionality or operations that are not allowed by nature (e.g. the upload of data on a read-only resource). The functionality offered by the logic processes will be translated by the resource interaction components, coding and decoding the data as defined above. It is important to note that this is the visible part of the Grid Service.

Definition of the Status of the Grid Service. The status is a combination of the functionality that the component will provide and the lower-level status of the physical components. It is homogeneous for all the resources in the same class.

Definition of the clients that interact with the Grid Services. Logic resources could require information coming from other resources. Those could be from the same or different class, but the interaction is only defined for logic resources. These client resources must be clearly defined to integrate the proxies required.

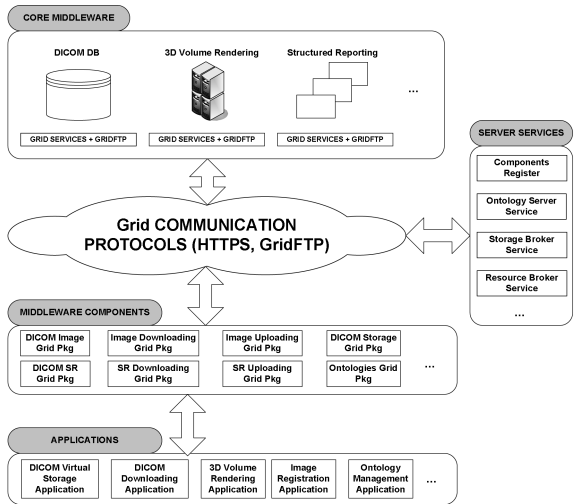


Fig. 1. General schema of the five-layered architecture of TRENCADIS

Definition of the communication interfaces of the Grid Service. Logic resources of the same class will offer the same interface. This will include not only the definition of the interface specification but also the definition of the protocols used for its

communication. The interface is defined using the standard Web Service Description Language [10] (WSDL), compatible with the definitions of OGSA and WSRF. This interface will be used by components defined in the Middleware Components Layer or even in the services of the same Layer.

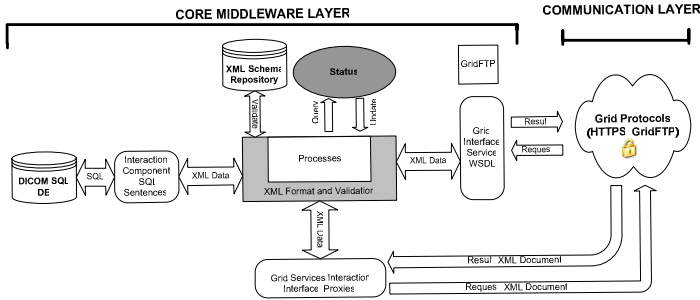


Fig. 2. Schema of a Grid Service implemented in the Core Middleware layer that corresponds to a DICOM Object Storage for an SQL database

4.2 Server Service Layer

The Server Service Layer corresponds to the second level of abstraction in the architecture, although it is more a set of vertical components rather than a horizontal layer. This layer contains the services that interact and manage the other distributed logical services (e.g. to build up the virtual database). The main difference between these services and the ones from the previous layer is that they do not interact directly with any physical resource, but only perform server tasks related with the logical resources. In the definition of the resources of this layer, it is necessary to consider the same points as in the Core Middleware Layer, except from the second point (implementation of the resource interaction component).

4.3 Communication Layer

The Communication layer provides the interconnection protocols that will be used by the Core Middleware and Server Services Layers which act as both clients and users of the services. The protocols used are, on one hand, the Simple Object Access Protocol [11] (SOAP), which is a high-level protocol based in XML that can be supported by many other lower level protocol, such as HTTP, HTTPS, FTP, etc. The communication with the Grid Services is implemented using SOAP under HTTPS to execute the requests to the Grid Services. Transferring of large blocks (as sub-blocks of medical images) is implemented using GridFTP [12] since it is much more efficient. All protocols are defined on a secure framework in which data and connections are encrypted and users are duly identified and authorised.

4.4 Middleware Components Layer

The main objective of this layer is the development of components to solve the main top-level functional objectives proposed. The components constitute the highest

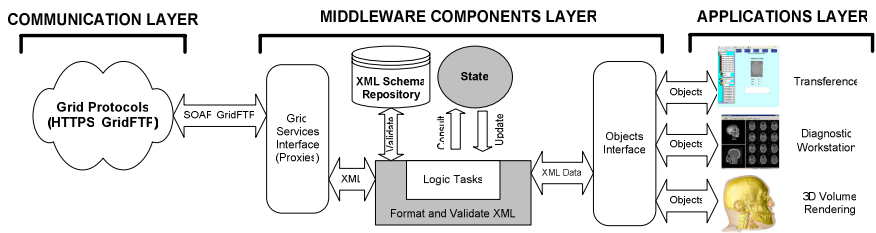


Fig. 3. Scheme of a Middleware Component from the Middleware Components Layer

abstraction level and provide the developers with the components and services to be directly used. They hide not only the implementation details of the physical resources, but also the particularities of the Grid environment used. The components defined take into account the following five points:

Object-oriented interface. The functionality of the system is offered to the users through methods and attributes of a set of classes contained in a software library.

Implementation of the logic processes of the component and the definition of the status. The objects will provide a status that will depend on the values of the attributes and a set of methods that implement the functionality, interacting through proxies with the services of lower level layers when they are needed.

Definition of the input / output data formats. As in the other cases, the components of this layer encode and decode the data into XML format using the schemas stored in a repository.

Interface for the interaction with the Grid Services. The middleware components use the services implemented in the lower layers through the Communication Layer and encoding and decoding the data into XML. The communication could be performed with several objects (as in the case of the virtual repository object, which interacts with server services and different logic resources).

4.5 Applications Layer

This layer comprises the top-level applications that are implemented in the system. Applications mainly use the Middleware Components through the object-oriented interface to provide the functionality requested by the user. An application to create and manage virtual repositories has been created [5][6].

5 Security

One important concern when dealing with medical information is privacy and security. Medical data is highly private, and confidential pieces of information should only be available to the owner, the medical team in charge of his/her treatment and with some restrictions to a medical research community. The distributed architecture proposed in this paper has less security constraints than other distributed approaches. Typical Datagrid approaches focus on the use of distributed resources to maximise the storage capabilities of a virtual organisation. Data security in the Grid Middleware

proposed is based on the Grid Security Infrastructure (GSI), adding some tools and procedures to enhance privacy. Primary motivations behind the GSI are: a) the need for secure communication (authenticated and confidential) between entities of a Grid. b) The need to support security across organizational boundaries, thus preventing a centrally-managed security system. c) The need to support “single sign-on” for users of the Grid, including delegation of credentials for computations, involves multiple resources and/or sites. First level of security is the use of certificates to authenticate users entering the Grid. Transferring of the data is performed through secure protocols (SSL-based) and thus do not present threats. Regarding storage, and since data can be permanently stored on a different organisation than the one that produces it, content must be protected to prevent users, even with administrative privileges, be able to access the content of the files. This approach increases complexity thus requiring secure repositories and encryption keys. Solutions for distributed and shared encrypted repositories do exist such as Perroquet [13] or MDM [14], but in the frame of TRENCADIS a security system has been developed based on [15].

6 Grid Services

This section describes in detail the main Grid Services implemented so far in the different layers of TRENCADIS architecture.

6.1 Core Middleware Layer

6.1.1 DICOM Storage Grid Services

This Grid Service is the interface to those devices or information systems that are part of the storage of DICOM objects. This service provides a common interface to all DICOM object storages of a virtual organisation (SQL databases, DICOM File systems, PACS, DICOM devices, etc.). This service embeds the storage resource in a logical resource that provides a common and homogeneous functionality. However, this service must provide different internal views considering the different devices that integrate the system (interaction component). The information that this device should offer must be organised according to the semantic ontologies defined in the corresponding services.

The interface of the logic service offers access to DICOM data regardless to the final storage. The functionality is defined by four methods: a) DICOM object insertion; b) DICOM object removal; c) DICOM object update, d) Search DICOM objects. All these processes are described in the following subsections.

The Grid Service is identified in a Grid deployment by a unique identifier that also relates to one or several VOs. The status of the resource is defined by the ontologies created in the Grid deployment. Each ontology has a view of the logic resources.

The main logic processes are the following:

Start-Up Process. The process of starting-up of the Grid Service resets the state of the resource. It firstly interacts with the Components Registry Grid Service, registering and enabling the services and retrieving the identifier that has been assigned for it. When the resource obtains the identifier, it request the information of

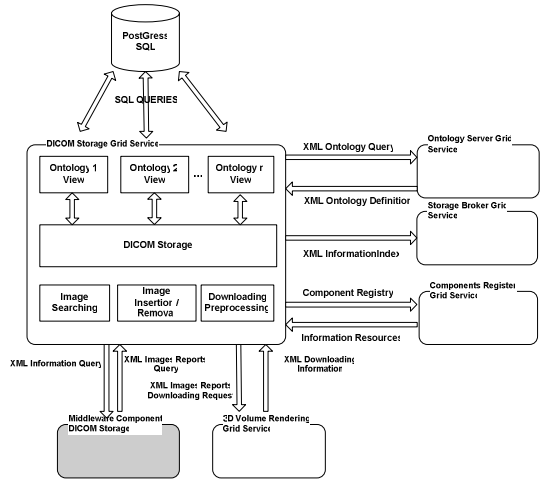


Fig. 4. Interconnection Schema of the DICOM Storage Resource

the ontologies to the Ontology Server Resource (Grid Service Ontology Server), registering the new ontologies in the database. During this process the Grid Service Storage DICOM sends the information to the Storage Broker Grid Service about the new ontologies for the creation of the indexed keys. Finally, the Grid Service waits for the requests of the resource interaction component.

DICOM Objects Searching Process. This process retrieves a searching request and encodes it in an XML document. This document is processed by the DICOM Storage View Manager which translates the request into the implementation native format (Postgress SQL in our case). The results of the searching process are encoded into an XML document and returned back.

Ontologies Creation/Removal Process. This process adds or removes views in the databases of ontologies. All activities in this process imply updating the database of the Storage Broker Service, since this service indexes the fields that have been defined in the ontologies.

DICOM Object Insertion/Removal Process. This process inserts or deletes DICOM objects in the related views. This process also informs the Storage Broker Service of the new data.

The more relevant interfaces for this Grid Services are the following:

INTERFACE	DESCRIPTION
xmlSearch	Searches DICOM objects
xmlAddOntology / xmlRemoveOntology	Adds/Removes the indexes of a new ontology in the DICOM Storage Grid Service

6.2 Server Grid Services Layer

The Grid Services from this layer are services that are located between the Middleware Layer and Core Middleware Layer, interacting with these layers and the Grid

Services of the same layer. The main task of these services is to manage Grid services deployed and provide the information to the Components Middleware Layer. The Grid Services developed in this layer are described in the following subsections.

6.2.1 Grid Service Storage Broker

The Storage Broker saves the information of different DICOM Storage objects related to the fields defined in the corresponding ontology. The main objective is to provide the Middleware Component with the sources (DICOM Storages) where the information matching the ontology searching criteria of the Virtual Storage is located. The state of this Grid Service is equivalent to the Component Registering Server.

The main logic processes are the following:

Start-up Process. Equivalent to other resources.

DICOM Storage Retrieve Process. It returns all the information related to the DICOM Storage Grid services that match a search criterion.

Update Data Process. It updates the database of the Storage Broker with the information of the Storage DICOM encoded in XML.

The more relevant interfaces for this Grid Service are the following:

INTERFACE	DESCRIPTION
xmlGetDICOMStorages	Return DICOM Storage Grid Services matching the searching criteria.
xmlGetAllDICOMStorages	Returns all the information of the DICOM Storage Grid Services.
xmlUpdateStorageBroker	Process that update the database of the Storage Broker.

6.2.2 Ontology Server Grid Service

This Grid Service deals with the storage of the active ontologies in the environment of a Grid deployment. It also updates the views of the DICOM Storage components when the ontologies are added, removed, updated and activated in the Server Grid Service. The ontologies enable showing the same information from different points of view, depending on the work-area or experiment that is being performed.

The main logic processes are the following:

Start-up Process. Equivalent to other resources.

Creation / Removal of Ontologies Process. Creates or removes ontologies in the database. Ontologies are defined using an XML document. When inserting an ontology, this service interacts with all DICOM Storages deployed for creating new views of the new ontology.

Retrieve Ontology Process. This process returns the whole ontology encoded in an XML document. It also consults the ontologies registered in the database.

The more relevant interfaces for this service are the following:

INTERFACE	DESCRIPTION
XmlCreateOntology	Creation/deletion of an ontology in the Grid.
xmlActivateOntology	Activation of an ontology in the Grid.
xmlGetOntology	Retrieval of an ontology.
xmlGetOntologyTypes	Returns the types of ontologies.
xmlGetOntologyIDs	Returns the IDs of the registered ontologies.
xmlGetOntologyData	Returns all data from a given ontology.

7 Ontologies

The ontology of a DICOM object is based on the information considered by that the clinics, researchers and other medical users for their tasks. The ontology contains the attributes, relations and restrictions that are defined among these objects.

Different kinds of information such as images, structured reports, signals, etc... can be stored in a DICOM file. Along with this information, the information related with the owner and the process (patient name, orientation, acquisition time, hospital, etc.) is stored in the DICOM header.

Parts of this information could be of interest for the classification of the medical images. Moreover, the value of some of these fields can define not only the results of a query, but even the information available.

On the other hand, different experiments defined inside the same medical area or research group will require part of the information that has made available to the whole community. Considering those needs, the language that specifies the ontologies in XML contains the following fields:

- **IDOntology:** Unique identifier of the ontology.
- **Description:** Brief description of the ontology.
- **TypeOntology:** These can be DICOM images, Structured Reports, Signals, etc...
- **Restrictive:** DICOM fields that are defined by the DICOM tag that defines the restrictions of the virtual storage. These fields determine the criteria that must match all DICOM objects to be included for a medical area or research group.
- **Creation:** Defined by DICOM tags. These fields define the indexes that will be created in the virtual storages for accessing to the distributed information.
- **Filter:** Fields defined by DICOM tag that will enable the different searching criteria for retrieving the information.

8 Components Middleware

This layer provides with the highest level of abstraction in the architecture. It offers an object-oriented interface to the users (software) for creating applications that manage DICOM objects. The different components interact with the needed Grid Services, described before and deployed in the lower layers. The components developed are organized in different packages, each one related with different issues.

8.1 Ontology Package

This package is in charge of managing the adding, removing and updating of the ontologies. This constitutes the user interface to the Grid Services that are internally used to manage the catalogues. A description of the components is provided next:

C_GRID_OntologyServer. This component creates the object instances that manage the Ontology Server.

C_GRID_Ontology. This component creates object instances that provide the interface for the definition and matching of the ontologies.

8.2 DICOM Medical Image Package

This package implements the objects that manage the DICOM medical images both individually and in groups. The components are the following:

C_GRID_Image_DICOM. This component creates the instances that virtualize a DICOM image.

C_GRID_Set_Images_DICOMs. This component deals with groups of DICOM images, individually managed through a C_GRID_Image_DICOM object.

C_GRID_Serie_DICOM. This component is a specialisation of the C_GRID_Set_Images_DICOM object, considering that all images have the same Series and Study identifier.

C_GRID_Study_DICOM. It is images with the same Study identifier.

8.3 DICOM Storage Package

This component deals with the storage and access to the DICOM objects that there are distributed in the Grid Environment. All DICOM objects are managed as a virtual storage, irrespectively of its location.

The DICOM Virtual Storage offers a single access point to all the DICOM objects fulfilling the restriction of a given ontology. The Virtual Storage only offers searching criteria for the fields that are included in the ontologies matching. The main component is **C_GRID_DICOM_Storage**. This virtually gathers the DICOM Storage objects considering the creation fields of a defined ontology. This component enables searching, filtering, adding and removing DICOM objects in the DICOM Storage. It can be composed by one or several DICOM Storage Grid Services sources.

9 Conclusions and Future Work

The architecture defined has been conceived to tackle with the demand of medical users in sharing and organising medical image databases for training and research, considering the structured report as the key source of knowledge. It is being used for the creation of cyber-infrastructure dedicated to oncology and medical imaging in the Land of Valencia, involving several of the mains hospitals of the region.

The TRENCADIS architecture uses current and standard Grid technologies to develop a framework in which new resources and studies can easily and securely integrated and shared, and in which high-performance computing services can easily be deployed in Grid computing infrastructures. The object-oriented interface provides a useful framework for software developers to speed-up productivity in application developing. The future work is being developed in the following lines:

- Include a workflow specification language to drive the distribution of data, the execution of several services and the gathering of results in a transparent and balanced manner.
- Provide channels for interactive sessions on the Grid.
- Improve the security by the use of encryption and key repositories [15].

Acknowledgment. The authors wish to thank the financial support received from The Spanish Ministry of Science and Technology to develop the project Investigación y Desarrollo de Servicios GRID: Aplicación a Modelos Cliente-Servidor, Colaborativos y de Alta Productividad, with reference TIC2003-01318. This work has been partially supported by the Structural Funds of the European Regional Development Fund (ERDF).

References

1. Digital Imaging and Communications in Medicine (DICOM) Part 10: Media Storage and File Format for Media Interchange. National Electrical Manufacturers Association, 1300 N. 17th Street, Rosslyn, Virginia 22209 USA.
2. DICOM Structured Reporting. Dr. David A. Clunie. ISBN 0-9701369-0-0.
3. "Open Grid Services Architecture (OGSA)", <http://www.globus.org/ogsa>
4. "The WS-Resource Framework". www.globus.org/wsrf
5. Blanquer, V. Hernández, D. Segrelles, "Creating Virtual Storages and Searching DICOM Medical Images through a GRID Middleware based in OGSA". CCGRID 2005. Sponsored by IEEE and IEEE Computer Society IEEE Catalogue 05EX1055C ISBN 0-7803-9075. Proceedings of the 5th IEEE Internacional Symposium on Cluster Computing and the Grid, Cardiff, Wales, UK May 9-12, 2005. Faltan las páginas
6. Blanquer, V. Hernández, D. Segrelles, "An OGSA Middleware for Managing Medical Images using Ontologies", Journal of Computing on Clinical Monitoring, ISSN: 1387-1307. Faltan las páginas y el año
7. Allen Wyke R., Watt A., "XML Schema Essentials". Wiley Computer Pub. ISBN 0-471-412597
8. Eric Miller. An Introduction to the Resource Description Framework. D-Lib Magazine, May 1998.
9. OWL Web Ontology Language Guide, Michael K. Smith, Chris Welty, and Deborah L. McGuinness, Editors, W3C Recommendation, 10 February 2004, 12 <http://www.w3.org/TR/2004/REC-owl-guide-20040210/>. Latest version available at <http://www.w3.org/TR/owlguide>
10. Erik Christensen (Microsoft), Francisco Curbera (IBM Research), Greg Meredith (Microsoft), Sanjiva Weerawarana (IBM Research). "Web Services Description Language (WSDL) 1.1.". <http://www.w3.org/TR/wsdl>
11. "SOAP Version 1.2.". <http://www.w3.org/TR/soap>.
12. "The GridFTP Protocol and Software". <http://wwwfp.globus.org/datagrid/gridftp.html>.
13. C. Blanchet, R. Mollon, "The Biomed/EGEE AuthZ requirements", 16 GGF 2006
14. J. Montagnat, "Requirements of Medical Data Manager Working Group", EGEE project, <http://www.i3s.unice.fr/~johan/mdm/mdm-051013.pdf>
15. E. Torres, C. de Alfonso, I. Blanquer, V. Hernández, "Privacy Protection in HealthGrid: Distributing Encryption Management Over the VO", Proceedings of the HealthGrid 2006 Conference, 2006.

Minimizing Data Size for Efficient Data Reuse in Grid-Enabled Medical Applications^{*}

Fumihiko Ino¹, Katsunori Matsuo¹, Yasuharu Mizutani², and Kenichi Hagihara¹

¹ Graduate School of Information Science and Technology, Osaka University
1-3 Machikaneyama, Toyonaka, Osaka 560-8531, Japan
ino@ist.osaka-u.ac.jp

² Faculty of Information Science and Technology, Osaka Institute of Technology

Abstract. This paper presents a data minimization method that aims at reducing overhead for data reuse in grid environments. The data reuse here is designed to promote efficient use of grid resources by avoiding multiple executions of the same computation in a collaborative community. To promote this at the program block level, our method minimizes the data size of attribute values, which are used for identification of computation products stored in a database (DB) server. Because attribute values are specified in queries used for store, search, or retrieval of computation products, their reduction leads to less communication between computing nodes and the DB server, minimizing the runtime overhead of data reuse. We also show some experimental results obtained using a time-consuming medical application. We find that the method successfully reduces the data size of a query from 683 MB to 52 B. This reduction allows our data reuse framework to reduce execution time from approximately 9 minutes to 27 seconds.

1 Introduction

With the rapid advance of network technology, the computational grid [1] is emerging as an attractive platform for computational scientists. For example, grid technology allows us to build high performance computing environments in virtual organizations. The key role of grid technology here is to make it possible to share computational resources and database (DB) contents in a specific virtual community constructed over the grid.

In contrast, some researchers are trying to share computation products in addition to hardware and software resources mentioned above. This data sharing approach avoids multiple executions of the same computation submitted by (usually different) users in the collaborative community. Therefore, grid resources are dedicated to produce new data, achieving highly efficient use of resources. For example, Quantum Chemistry Grid [2] provides a grid-enabled problem solving environment capable of accumulating computation products obtained by a computational chemistry program. Since this environment focuses on a single program, data reuse can easily be realized by constructing a DB where each of computation products is associated with attributes, namely the inputs given to the program. Attribute values are then specified in queries to store, search, or retrieve computation products in the DB.

^{*} This work was partly supported by JSPS Grant-in-Aid on Priority Areas (170320007), for Scientific Research (B)(2)(18300009), and for Young Researchers (17700060).

On the other hand, Pegasus [3,4] integrates a data reuse functionality into a workflow mapping system. A workflow here abstracts the processing sequence of data in discrete steps. It consists of vertices and edges, representing application components and their flow dependencies, respectively. Before mapping a workflow onto grid resources, the system eliminates vertices in the workflow if the corresponding application components have previously been executed with the same inputs. This elimination achieves higher efficiency by replacing repetitive executions with data retrieval from the DB.

Thus, data reuse capabilities are useful to avoid wasteful executions in a virtual community. However, data reuse is usually done at the program level. Therefore, the efficiency can be further improved if data is reused within a program, for example, at the block level. The problem addressed in the paper is to realize this block-level reuse at a low overhead. We think that data reuse should work at a finer granularity to perform data reuse at the appropriate granularity such as blocks, functions, or programs, chosen according to the tradeoff between its repeatability and time saved by reuse.

In this paper, we present a data minimization method that aims at reducing overhead for block-level reuse in grid environments. Our method requires users to specify the program code for data reuse, and then minimizes the data size of attribute values, which are used for identification of computation products stored in a DB server. The key idea of our method is data dependence analysis that aims at replacing the initially specified code with an extended code that requires less amount of attribute values for data reuse.

2 Related Work

Similar to Pegasus [3,4], the GriPhyN virtual data system (VDS) [5] also provides a data reuse capability to data-intensive applications. The VDS allows users to discover and share virtual data products, compose workflows, and monitor workflow executions. However, their data reuse model does not consider complex workflows having branches or iterations. Dealing with these control flows is required to reuse computation products for program blocks.

Altintas et al. [6] also realize a data reuse functionality for grid workflow systems. Their functionality is designed to reuse workflows rather than computation products. Therefore, their concern is the share of knowledge on effective workflows across different scientific fields.

The AppLeS (Application Level Scheduling) Parameter Sweep Template (APST) [7] maximizes reuse of shared data files by using adaptive scheduling techniques for parameter sweep applications. It employs a replication strategy to minimize data transmission between the client machine and computing nodes. Thus, this reuse functionality works at the scheduling level, trying to place data files to maximize data reuse. With respect to minimization of data transmission, our method also addresses the same problem. However, our method differs from APST, which does not minimize the data size itself. Another replication-based strategy is also presented in [8].

Some researchers resolve the problem of data reuse in caches. Strout et al. [9] performs data reuse to accelerate Gauss-Seidel methods. Their method improves intra-iteration and inter-iteration data locality in iterative solvers. Issenin et al. [10] performs data reuse in a more explicit manner. They make copies of frequently used data in a

Table 1. Notations

Notation	Explanation
V	A set of vertices
E	A set of edges
G	$G = (V, E)$ defines a directed acyclic graph representing a workflow
(u, v)	$(u, v) \in E$ represents an edge from vertex u to vertex v
$label(v)$	The name labeled on vertex v
$label(u, v)$	The name labeled on edge (u, v)
$value(u, v)$	Values (contents) of name $label(u, v)$
I_v	$I_v = \{(u, v) \mid \exists u \text{ such that } (u, v) \in E\}$ defines the set of edges incoming to vertex v
O_v	$O_v = \{(v, w) \mid \exists w \text{ such that } (v, w) \in E\}$ defines the set of edges outgoing from vertex v
S_v	$S_v = \{label(u, v), value(u, v) \mid (u, v) \in I_v\}$ defines the set of pairs of labels and values passed to vertex v
T_v	$T_v = \{label(v, w), value(v, w) \mid (v, w) \in O_v\}$ defines the set of pairs of labels and values passed from vertex v
R	$R \subseteq V$ denotes the set of vertices initially marked as the target code for data reuse
C_R	$C_R \subseteq R$ denotes the set of critical vertices that have attribute information for R , where $C_R = \{v \in R \mid \exists u \text{ such that } u \notin R \wedge (u, v) \in E\}$
$\mathcal{A}_R$	$\mathcal{A}_R = \bigcup_{v \in C_R} S_v$ denotes the initial attributes for user-specified R

small local memory. Although these kinds of data reuse contribute to acceleration of specific applications, their computation products are shared during a single execution. In contrast, we focus on data sharing across multiple executions submitted from different grid users.

Thus, our work tries to enhance data reuse in workflow-based systems and caches by coupling the advantages of these two data reuse concepts.

3 Preliminaries

To describe our problem clearly, we first introduce the program-level reuse addressed by prior systems. Table 1 summarizes notations used in the paper.

Let $G = (V, E)$ be a directed acyclic graph (DAG), where V and E represent a set of vertices and that of edges in the graph, respectively. Graph G here has two special vertices, source and sink, each having only outgoing edges and incoming edges, respectively. Then, as shown in Fig. 1(a), prior systems [3,4] regard a vertex and an edge as an application component and a flow dependency between components, respectively. Each vertex is labeled with a component name while each edge is labeled with a file name given to the next component (destination vertex). In the following, let (u, v) denote the edge connected from vertex u to vertex v .

Note here that communities should have a single name space to give a unique name to identical data. Otherwise, data contents as well as file names must be checked to prevent inappropriate reuse of wrong data. Thus, prior systems assume a single name space where identical components or files have a unique name. Let $label(v)$ and $label(u, v)$ denote the name labeled on vertex v and on edge (u, v) , respectively. Let $value(u, v)$ also denote the values (contents) of the name $label(u, v)$.

Suppose that we have two DAGs $G = (V, E)$ and $G' = (V', E')$, where G and G' represent a workflow being considered for execution and that executed in the past, respectively. Suppose vertex $v \in V$ has a set I_v of incoming edges and that O_v of outgoing edges, where $I_v = \{(u, v) \mid \exists u \text{ such that } (u, v) \in E\}$ and $O_v = \{(v, w) \mid \exists w \text{ such that } (v, w) \in E\}$. Let S_v be the set of pairs of labels and values passed to vertex v :

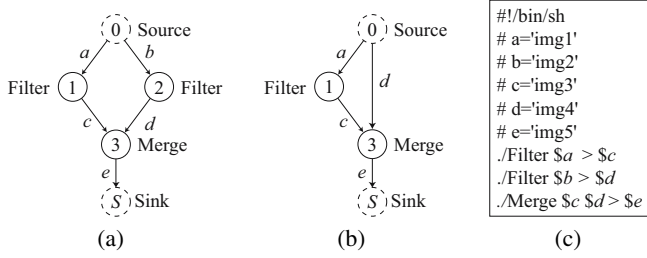


Fig. 1. (a) A directed acyclic graph (DAG) representing a workflow, (b) its reduced DAG for data reuse, and (c) a shell script for the workflow. Vertex 2 and its edges (0, 2) and (2, 3) are replaced with (0, 3), meaning that the second filter program is omitted by reuse of file ‘img4.’

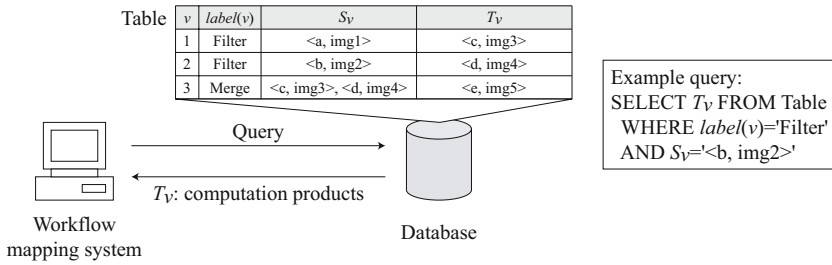


Fig. 2. Overview of data reuse in prior systems

$S_v = \{ \langle label(u, v), value(u, v) \rangle \mid (u, v) \in I_v \}$. Let T_v also be the set of pairs of labels and values passed from vertex v : $T_v = \{ \langle label(v, w), value(v, w) \rangle \mid (v, w) \in O_v \}$. Then, any vertex v and its every edge $(u, w) \in I_v \cup O_v$ can be eliminated if and only if

C1. $\exists x \in V'$ such that $label(x) = label(v) \wedge S_x = S_v$.

Such vertices v and x represent a program with the same inputs, and thus they are identical computation. In this case, we already have computation products of x in the DB. Therefore, graph G can be reduced for data reuse as follows (see Fig. 1(b)). (1) Vertex elimination: Remove vertex v and its every incoming/outgoing edge $(u, w) \in I_v \cup O_v$ from set V and set E , respectively. (2) Vertex connection: For all w such that $(v, w) \in O_v$, add edge $(0, w)$ to E . Edge $(0, w)$ here has $label(v, w)$, representing loading of computation products from the DB.

According to condition C1, data reuse capabilities require a DB that stores a set of 4-tuples $(v, label(v), S_v, T_v)$, as shown in Fig. 2. Given such a DB, identical computations can be detected by comparing the two attributes $label(v)$ and S_v to obtain computation products T_v stored in the DB.

Prior systems typically reuse computation products in the following steps (Fig. 3).

- S1. Workflow submission: The system receives a workflow, namely a DAG, from users.
- S2. Workflow reduction: The DAG is reduced to exploit data reuse based on the DB.

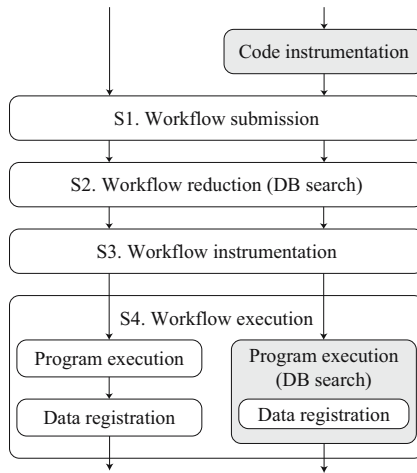


Fig. 3. Procedure for data reuse. Left-hand side shows the procedure for prior program-level reuse while right-hand side shows that for our block-level reuse.

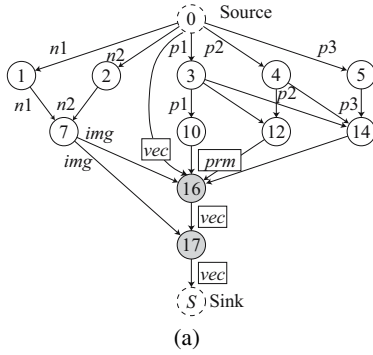
- S3. Workflow instrumentation: Some instrumentation vertices are added to the reduced DAG for later registration of newly produced data.
- S4. Workflow execution: This step consists of two substeps: Program execution step, which executes each application component using grid resources; Data registration step, which registers computation products to the DB after program execution.

4 Problem Description

In contrast to the program-level reuse, which prior systems support, our final goal is to realize the block-level reuse during program execution. To achieve this, we must tackle the following technical issues (see Fig. 3).

- I1. Model extension: We must adapt the data reuse model mentioned in Section 3 to our case, because we focus on blocks rather than programs. The extended model must handle program structures and control-flow dependencies in programs.
- I2. Code instrumentation: In addition to the workflow component, the source code in programs must be instrumented to record computation products for each block. The key issue here is to assist users in selecting the code to be instrumented for data reuse.
- I3. Runtime registration: Since we focus on program blocks, their computation products must be registered during program execution. Therefore, reducing runtime overhead is a critical issue to obtain higher program performance. Note here that prior systems are allowed to register data after program execution.

With respect to issue I2, we assume that users know which code takes most of execution time, and thus such bottleneck code can be initially specified as the instrumented code, which may save significant time and resources. Once such initial code is given



```

1: readln(n1); // File name 1
2: readln(n2); // File name 2
3: readln(p1); // Parameter 1
4: readln(p2); // Parameter 2
5: readln(p3); // Parameter 3
6: h = 0; // Hierarchy
7: img = load(n1, n2);
8: while h < 5 {
9:   if h == 0 {
10:    prm = p1;
11:   } else if h == 1 {
12:    prm = p1 * p2;
13:   } else {
14:    prm = p1 * p2 * p3;
15:   }
16:   vec = registration(prm, img, vec);
17:   vec = resample(img, vec); // Increase resolution
18:   h += 1;
19: }

```

(b)

Fig. 4. Overview of (a) registration algorithm and (b) its DAG. Statements at lines 16–17 are initially specified as the target code for data reuse. Enclosed labels on edges represent flow-influenced variables.

by users, our method tries to suggest better (extended) code with smaller attribute values. Thus, we assist users in selecting appropriate attributes with smaller data size. In summary, the problem of data minimization can be defined as follows:

P1. The data minimization problem addressed in the paper is to minimize the data size of attribute values required for the target code initially specified by users.

Note here that the data minimization contributes to resolve issue I3, because attribute values are transmitted to the DB server as a part of queries. It also minimizes accesses to storage devices on the DB server. Thus, data minimization will reduce runtime overhead of data reuse.

5 Block-Level Data Reuse

We now present our data reuse framework based on an extension of the previous reuse model and the data minimization method.

5.1 Model Extension

To resolve issue I1, we extend the model for our case as follows (see Fig. 4).

- On program structures. In our interpretation, a vertex and an edge in a DAG represent a statement in a program and a flow dependency between statements, respectively. The label on a vertex and that on an edge represent the line number of the corresponding statement and the variable names relevant to the dependency, respectively.

- On control-flow dependencies. Since programs consist of branches and iterations, which workflows in prior systems do not have, we associate labels with additional properties to represent flow-influenced variables. For such variables, we consider variables with loop-carried dependencies [11] or control-flow dependencies.

Figure 4 shows an example of the extended model. In this example, branch statements at lines 10, 12, and 14 depends on runtime conditions, and thus their outputs, namely the labels ‘prm’ on outgoing edges in the DAG, are marked as flow-influenced variables. Similarly, variable ‘vec’ with a loop-carried dependency is also marked as a flow-influenced variable.

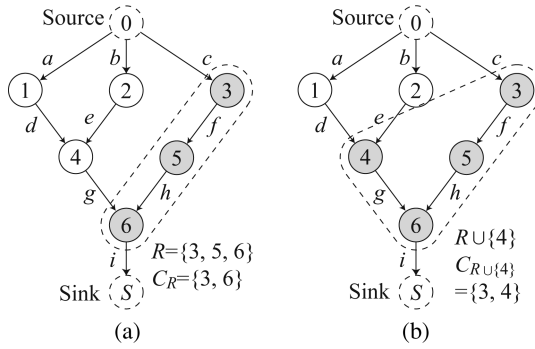


Fig. 5. Data minimization. Given a DAG with a set R of vertices, $\{3, 5, 6\}$, our method returns pairs of labels and values on edges $(0, 3)$ and $(4, 6)$ as the initial attributes. It then tries to replace some edges to others with smaller data size. In this case, (a) edge $(4, 6)$ is replaced with (b) edges $(1, 4)$ and $(2, 4)$ if the data size of attribute g is larger than total size of attributes d and e .

5.2 Principles of Data Minimization

Suppose that we have a DAG $G = (V, E)$ with a set $R \subset V$ of vertices initially marked as the target code for data reuse (see Fig. 5). Then, the initial attributes $\mathcal{A}_R$ for R are given by their inputs:

$$\mathcal{A}_R = \bigcup_{v \in C_R} \mathcal{S}_v, \quad (1)$$

where C_R represents the set of critical vertices that have attribute information for R :

$$C_R = \{v \in R \mid \exists u \text{ such that } u \notin R \wedge (u, v) \in E\}. \quad (2)$$

Let $u \in V$ be a vertex such that $u \notin R$. Given R by users, the proposed method tries to extend R to include u to minimize the data size of inputs, namely attribute values. Such an extension is allowed if

C3. $v \in R$, for all v such that $(u, v) \in O_u$.

This extension tries to replace set O_u of edges outgoing from u with set I_u of edges incoming to u . If this extension is allowed, the initial attributes $\mathcal{A}_R$ can be replaced with those for $R \cup \{u\}$ given by:

$$\mathcal{A}_{R \cup \{u\}} = \bigcup_{v \in C_{R \cup \{u\}}} \mathcal{S}_v, \quad (3)$$

where

$$C_{R \cup \{u\}} = C_R \cup \{u\} - \{w \mid w \in C_R \wedge (u, w) \in O_u\}. \quad (4)$$

Condition C3 is not sufficient if vertex u corresponds to a flow-influenced statement, such as branches and loops. For example, suppose that we have if-else statements followed by the initial target code, as shown in Fig. 4(a). In this case, the extension mentioned above may include branches at lines 9–15 to the target code for data reuse. This implies that computation products vary depending on runtime conditions. In this case, we must record branching results as well as computation products. Otherwise, the replacement may result in wrong data reuse, because it does not have information on the actual flow that have yielded the products. To avoid such wrong reuse, we store actual flows by recording the value history list for flow-influenced variables.

In summary, data minimization is performed by extending R initially given by users. Extended part here is allowed to include flow-influenced statements, however, the actual flow must be recorded as attributes in the DB. We record this by the value history list of flow-influenced variables to perform data reuse only for identical computation.

5.3 Data Minimization Algorithm

Figure 6 shows our algorithm, which tries to extend the target code R to obtain the smallest attributes $\mathcal{A}$. Basically, it backtraces flow dependencies from a vertex in set R . The algorithm consists of the following three phases.

Phase 1. Extract the initial attributes $\mathcal{A}_R$. To do this, the algorithm computes a set C_R of critical vertices according to Eq. (2).

Phase 2. Extend the target code for data reuse. The algorithm then performs the code extension to compute set $\mathcal{L}$ containing possible sets of critical vertices. This code extension is done by recursive calls of function `Extension()`, as shown in Fig. 7. This function backtraces flow dependencies from a start vertex $u \in R$. During the backtracing of flows, candidates for critical vertices are added to set $\mathcal{L}$ at each visited vertex. The backtracing procedure stops when it reaches (1) the source or (2) a flow-influenced vertex. For the former case, it returns the current candidates $\mathcal{L}$. For the latter case, it stops further backtracing of flows because flow-influenced variables cannot be removed from attributes, as we mentioned in Section 5.2.

Phase 3. Select the smallest attributes $\mathcal{A}$. The attributes $\mathcal{A}$ are selected from set $\mathcal{L}$ of sets of critical vertices. This phase is executed at runtime if flow-influenced variables are included in $\mathcal{A}$, because the data size of such variables cannot be determined before program execution. That is, the value history list for such variables grows during program execution.

6 Experimental Results

We now show how the method reduces data size and execution time to evaluate the impact of block-level data reuse.

```

Algorithm DataMinimization( $G, R$ )
// Input #1. DAG  $G = (V, E)$ .
// Input #2. Set  $R$  of vertices representing the initial target code.
// Output. Attributes  $\mathcal{A}$  extended from the initial attributes  $\mathcal{A}_R$ .
begin
  // Phase 1: Extract the initial attributes  $\mathcal{A}_R$ 
   $C_R \leftarrow \emptyset$ ; // initialize set  $C$  as an empty set
  foreach vertex  $v \in R$  do begin
    foreach edge  $(u, v) \in I_v$  do begin
      if vertex  $u \notin R$  then begin
        Add vertex  $v$  to set  $C_R$ ;
      end
    end
  end
  Compute the initial attributes  $\mathcal{A}_R$  by using  $C_R$  // see Eq. (1)
   $\mathcal{A} \leftarrow \mathcal{A}_R$ ;
  // Phase 2: Extend the target code by backtracing
   $\mathcal{L} \leftarrow \emptyset$ ; //  $\mathcal{L} = \{C_R\}$ : A set of sets of critical vertices
  Add  $C_R$  to set  $\mathcal{L}$ ;
  foreach vertex  $v \in R$  do begin
     $\mathcal{L} \leftarrow \text{Extension}(v, C_R, \mathcal{L})$ ; // extend  $R$  from vertex  $v$ 
  end
  // Phase 3: Select the smallest attributes
  foreach set  $C_{R'} \in \mathcal{L}$  do begin
    Compute the attributes  $\mathcal{A}_{R'}$  by using  $C_{R'}$  // see Eq. (1)
    if  $\mathcal{A}_{R'}$  is smaller than  $\mathcal{A}$  then begin
       $\mathcal{A} \leftarrow \mathcal{A}_{R'}$ ;
    end
  end
end

```

Fig. 6. Data minimization algorithm. See Fig. 7 for function Extension()

For experiments, we employ a cluster of 16 PCs, each having two Xeon 2.8 GHz CPUs and 2 GB of main memory. PCs are interconnected by a Myrinet switch [12], providing a bandwidth of 2 Gb/s. A DB system is constructed on a file server. This DB server is accessible from PCs through Gigabit Ethernet network.

The application used for experiments is nonrigid image registration [13], which computes point correspondences between two deformable objects. This application is written using the C++ language and the Message Passing Interface (MPI) standard [14]. It requires three parameters and a pair of three-dimensional images as inputs. The image size is $512 \times 512 \times 295$ voxels, which is equivalent to 148 MB per image in file size. From the viewpoint of program structures, this application have the following characteristics (see Fig. 4).

- Computation intensive. Sequential implementations take several hours to process the core functions, registration() and resample() at lines 16 and 17, respectively.
- Hierarchical algorithm. Registration is hierarchically performed in a coarse-to-fine manner to reduce the computational amount. Each hierarchy requires a parameter to control object deformations.
- Parametric study. While registration is controlled by three parameters ‘p1,’ ‘p2,’ and ‘p3,’ better parameter values are still unknown. Therefore, parametric study is needed to know better parameter values, and thus, registration tasks are repeatedly submitted usually with the same pair of images but with slightly different combinations of parameters.

```

Function Extension( $v, C_R, \mathcal{L}$ )
// Input #1. Vertex  $v$  representing the current point for backtracing.
// Input #2. Current set  $C_R$  of critical vertices.
// Input #3. Current set  $\mathcal{L}$  of sets of critical vertices.
// Output. Updated set  $\mathcal{L}$ .
begin
   $\mathcal{L}_{local} \leftarrow \mathcal{L}$ ;
  if  $I_v \neq \emptyset$  then begin //  $v$  has a predecessor
    foreach edge  $(u, v) \in I_v$  do begin // try to include vertex  $u$ 
      if  $(u, v)$  is not a flow-influenced variable then begin
         $\mathcal{L}_{local} \leftarrow \text{Extension}(u, C_{R \cup \{u\}}, \mathcal{L}_{local})$ ;
        Add  $C_{R \cup \{u\}}$  to  $\mathcal{L}_{local}$ ; // see Eq. (4)
      end
    end
  end
  return  $\mathcal{L}_{local}$ ;
end

```

Fig. 7. Code extension algorithm

Table 2. Data size of attribute values required to reuse data at each of hierarchy H1–H5. Attribute ‘img’ is replaced with ‘n1’ and ‘n2’ while attribute ‘vec’ is replaced with ‘n1,’ ‘n2,’ and ‘prm.’

Prior method w/o data minimization						Our method w/ data minimization					
Attribute	H1	H2	H3	H4	H5	Attribute	H1	H2	H3	H4	H5
img	680 M					n1	6				
vec	58 K	408 K	3 M			n2	6				
prm	8					prm	8	16	24	32	40
Total	680 M	680 M	683 M			Total	20	28	36	44	52

Due to the last characteristic, we think that data reuse is effective to accelerate parametric study of nonrigid registration algorithms on the grid.

We first determine the target code for data reuse. Two functions registration() and resample() are selected for data reuse, because these core functions take most (98%) of execution time. We then manually applied our method to the registration program.

Table 2 shows data reduction results. We can see that the prior method requires three attributes, namely all inputs directly given to the target function: ‘prm,’ ‘img,’ and ‘vec.’ Attributes ‘prm’ and ‘img’ are fixed-size variables containing 8 B and 680 MB of data, respectively. On the other hand, the data size of attribute ‘vec’ increases as the algorithm moves up the hierarchy, and thus it ranges from 58 KB to 3 MB. Summing up these sizes, the prior method requires approximately 680 MB and 683 MB of attribute values for the coarsest hierarchy H1 and for the finest hierarchy H5, respectively.

In contrast to the prior method, our method requires at most 52 B of attribute values. This reduction can be explained as follows.

- Attribute ‘img’ is replaced with ‘n1’ and ‘n2,’ because it has larger data size and satisfies condition C3. Vertex 7 in Fig. 4 intuitively explain this. In the example code, this replacement means that image data can be replaced with its file name.
- Attribute ‘vec’ is replaced with ‘n1,’ ‘n2,’ and the value history list of ‘prm.’ Figure 4 indicates that ‘img’ and ‘prm’ must be included to attributes to remove ‘vec,’ because they are given as inputs to vertices 16 and 17, which output ‘vec.’ The

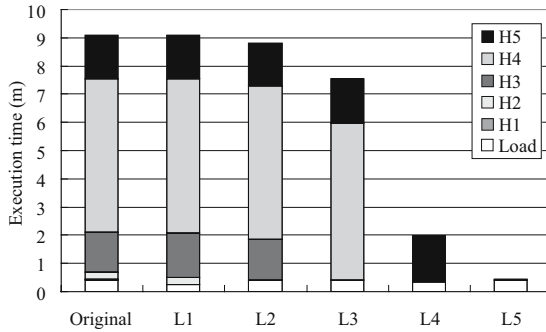


Fig. 8. Timing results. Each bar shows execution time and its breakdown for variation L_i ($1 \leq i \leq 5$). Data reuse is applied to the program in a stepwise manner.

former attribute ‘img’ here is already replaced with ‘n1’ and ‘n2.’ On the other hand, the latter attribute ‘prm’ is marked as a flow-influenced variable, and thus it must be recorded with its value history list. Since ‘prm’ is a double variable, its size is increased by 8 B at each hierarchy.

In order to activate data reuse in the application, we added function calls before and behind the target functions. The added function here requires attributes and their values to deal with 4-tuple data stored in the DB. We specified the smallest attributes in Table 2: ‘n1,’ ‘n2,’ and value history list of ‘prm.’

Figure 8 shows timing results for the original program and its five variations L1–L5. Here, variation L_i perform data reuse from hierarchy H1 to H_i , where $1 \leq i \leq 5$. In this figure, we can see that the original time of 9 minutes is reduced to approximately 30 seconds if all computations are allowed to be omitted by data reuse. Therefore, if all of computation products are already stored in the DB, other programs can use grid resources for approximately 8 minutes in this case.

Note here that hierarchy H3 takes most of execution time in the original program. Therefore, data reuse is effective if the computation products of this hierarchy is reused during program execution. In other words, data reuse at hierarchies H1, H2, and H3 was not so effective in terms of performance.

7 Conclusion

We have presented a data minimization method for efficient data reuse in grid environments. As compared with prior methods, our method focuses on reusing computation products at smaller granularities such as program blocks. The novelty of our method is the data dependence analysis that minimizes the data size of attribute values by extending the target code of data reuse. The proposed method allows us to transmit less amount of data between computing nodes and the DB server, and thus contributes to reduce runtime overhead of data reuse.

In experimental results, we find that the method reduces the data size from 683 MB to 52 B, achieving a small overhead for data reuse in a medical application. Our reuse

framework also accelerates the application from approximately 9 minutes to 27 seconds, achieving shorter response time with higher efficiency.

Future work includes the development of a tool that automates the instrumentation of program code.

References

1. Foster, I., Kesselman, C., Tuecke, S.: The anatomy of the grid: Enabling scalable virtual organizations. *Int'l J. High Performance Computing Applications* **15** (2001) 200–222
2. Nishikawa, T., Nagashima, U., Sekiguchi, S.: Design and implementation of intelligent scheduler for gaussian portal on quantum chemistry grid. In: *Proc. 3rd Int'l Conf. Computational Science (ICCS'03), Part III*. (2003) 244–253
3. Deelman, E., Blythe, J., Gil, Y., Kesselman, C., Mehta, G., Vahi, K., Blackburn, K., Lazzarini, A., Arbre, A., Cavanaugh, R., Koranda, S.: Mapping abstract complex workflows onto grid environments. *J. Grid Computing* **1** (2003) 25–39
4. Deelman, E., Singh, G., Su, M.H., Blythe, J., Gil, Y., Kesselman, C., Mehta, G., Vahi, K., Beriman, G.B., Good, J., Laity, A., Jacob, J.C., Katz, D.S.: Pegasus: a framework for mapping complex scientific workflows onto distributed systems. *Scientific Programming* **13** (2005) 219–237
5. Zhao, Y., Wilde, M., Foster, I., Voeckler, J., Dobson, J., Gilbert, E., Jordan, T., Quigg, E.: Virtual data grid middleware services for data-intensive science. *Concurrency and Computation: Practice and Experience* **18** (2006) 595–608
6. Altintas, I., Birnbaum, A., Baldridge, K.K., Sudholt, W., Miller, M., Amoreira, C., Potier, Y., Ludaescher, B.: A framework for the design and reuse of grid workflows. In: *Proc. 1st Int'l Workshop Scientific Applications of Grid Computing (SAG'04)*. (2004) 120–133
7. Casanova, H., Obertelli, G., Berman, F., Wolski, R.: The AppLeS parameter sweep template: User-level middleware for the Grid. In: *Proc. High Performance Networking and Computing Conf. (SC2000)*. (2000)
8. Santos-Neto, E., Cirne, W., Brasileiro, F., Lima, A.: Exploiting replication and data reuse to efficiently schedule data-intensive applications on grids. In: *Proc. 10th Int'l Workshop Job Scheduling Strategies for Parallel Processing (JSSPP'04)*. (2004) 210–232
9. Strout, M.M., Carter, L., Ferrante, J., Freeman, J., Kreaseck, B.: Combining performance aspects of irregular Gauss-Seidel via sparse tiling. In: *Proc. 15th Workshop Languages and Compilers for Parallel Computing (LCPC'04)*. (2002) 90–110
10. Issenin, I., Brockmeyer, E., Miranda, M., Dutt, N.: Data reuse analysis technique for software-controlled memory hierarchies. In: *Proc. Design, Automation and Test in Europe Conf. and Exhibition (DATE'04)*. (2004) 202–207
11. Bacon, D.F., Graham, S.L., Sharp, O.J.: Compiler transformations for high-performance computing. *ACM Computing Surveys* **26** (1994) 345–420
12. Boden, N.J., Cohen, D., Felderman, R.E., Kulawik, A.E., Seitz, C.L., Seizovic, J.N., Su, W.K.: Myrinet: A gigabit-per-second local area network. *IEEE Micro* **15** (1995) 29–36
13. Ino, F., Ooyama, K., Hagihara, K.: A data distributed parallel algorithm for nonrigid image registration. *Parallel Computing* **31** (2005) 19–43
14. Message Passing Interface Forum: MPI: A message-passing interface standard. *Int'l J. Supercomputer Applications and High Performance Computing* **8** (1994) 159–416

Thinking Precedes Action: Using Software Engineering for the Development of a Terminology Database to Improve Access to Biomedical Documentation*

Antonio Vaquero¹, Fernando Sáenz¹, Francisco Álvarez², and Manuel de Buenaga³

¹ Universidad Complutense de Madrid, Facultad de Informática, Departamento de Ingeniería del Software e Inteligencia Artificial, C/ Prof. José García Santesmases, s/n, E-28040, Madrid, Spain

² Universidad Autónoma de Sinaloa, Ángel Flores y Riva Palacios, s/n, C.P 80000, Culiacán, Sinaloa, México

³ Universidad Europea de Madrid, Departamento de Sistemas Informáticos, 28670 Villaviciosa de Odón. Madrid, Spain

¹ {vaquero, fernan}@sip.ucm.es, ² fjalvare@fdi.ucm.es,
³ buenaga@uem.es

Abstract. Relational databases have been used to represent lexical knowledge since the days of machine-readable dictionaries. However, although software engineering provides a methodological framework for the construction of databases, most developing efforts focus on content, implementation and time-saving issues, and forget about the software engineering aspects of software and database construction. We have defined a methodology for the development of lexical resources that covers this and other aspects, by following a sound software engineering approach to formally represent knowledge. Nonetheless, the conceptual model from which it departs has some major limitations that need to be overcome. Based on a short analysis of common problems in existing lexical resources, we present an upgraded conceptual model as a first step towards the methodological development of a hierarchically organized concept-based terminology database, to improve the access to medical information as part of the SINAMED and ISIS projects.

1 Introduction

Since the days of machine-readable dictionaries (MRD), relational databases (RDB) have been a popular device to store information for linguistic purposes. Relational database technology offers many advantages, being one of its more important ones the existence of a mature software engineering database design methodology. Nevertheless, most of the efforts aimed at developing linguistic resources (LR), whether they used RDB or not, have focused on content, implementation or time-saving issues, putting aside the software engineering aspects of the construction of LR.

* The research described in this paper has been partially supported by the Spanish Ministry of Education and Science and the European Union from the European Regional Development Funds (ERDF) - (TIN2005-08988-C02-01 and TIN2005-08988-C02-02) and (ERDF) - (FIT-350200-2005-16) for SINAMED and ISIS respectively.

Many authors use the term “software engineering” synonymously with “system analysis and design” and other titles, but the underlying point is that any information system requires some process to develop it correctly. The basic idea is that to build software correctly, a series of steps are required. These steps ensure that a process of thinking precedes action: thinking through “what is needed” precedes “what is written”. Although software engineering spans a wide range of problems, we will focus here on the database design aspects.

As it will be seen later, design issues are important when using RDB. Moreover, as we stated in [1], design is also important because in order to develop, reuse and integrate diverse available resources, into a common information system, perhaps distributed, requires compatible software architectures and sound data management from the different databases (DB) to be integrated. With that in mind, we have defined a methodology [1], for the design and implementation of ontology-based LR using RDB and a sound software engineering approach. Nevertheless, the conceptual model we propose as a point of departure of the methodology has some major limitations, which have to be overcome in order to create structurally sound LR.

In this paper, we will focus on the ontology representation limitations of our previous model (leaving the lexical side limitations for a future paper), and create a conceptual model of the ontological part, that overcomes such limitations as part of our efforts to have a solid foundation for action. Our final goal is to create a LR (a hierarchically organized concept-based terminology database) that will be part of an intelligent information access system that integrates text categorization and summarization, to improve information access to patient clinical records and related scientific documentation, as part of the SINAMED and ISIS projects [2].

The rest of the paper is organized as follows. In section 2, we underline the importance of software engineering and software design in the construction of LR and Ontologies, and state that practical reusability can only be obtained by applying software engineering to their construction. In section 3, the advantages and disadvantages of RDB are pointed out, as well as the importance of database design in the construction of ontology-based LR with relational technology. In section 4, some common problems of LR are summarized, and the need to develop methodologically engineered problem-solving LR is signaled. In section 5, the methodological gaps of past developing efforts are highlighted. In section 6, a set of ideas intended to help developers to formally specify and clarify the meaning of concepts and relations are depicted. In section 7, a conceptual model that integrates the aforementioned ideas is introduced and described. Finally, in section 8 some conclusions and future work are outlined.

2 Software Engineering in the Construction of Ontologies and LR

Software can be considered a product of engineering, just like an airplane, automobile, television, or any other object that requires a high degree of skill to turn a raw material into a usable product. Following [3], software: a) is an entity (not a document); b) is generally a component of a larger system (hardware/software); c) must interface with other hardware or software systems; d) is too large and complex to build without a plan (specification) and e) is expensive to build.

Based on the above statements an ontology or a LR are software, and although it might be argued that their nature is different from that of a “piece of code” and that the methods of software engineering cannot be applied to them [4], software engineering not only encompasses the “coding of a solution”. It provides the technical “how to’s” for building software [3]. It is a layered technology that rests on an organizational commitment to quality that includes a broad array of tasks (i.e., requirements analysis, design, program construction, testing and maintenance). Moreover, software engineering offers a variety of different methods to achieve these activities, and tedious debates over which method is best seem to miss the point. Any method, if properly applied within the context of a solid set of software engineering principles, will lead to higher quality software than an undisciplined approach [3].

As pointed out by Guarino in [4], ontologies (and LR) need proper quality control to be effectively deployed in practical applications. Ontologies in particular represent a new technology with relatively few standards, and this causes problems that software engineering has to solve, especially if one looks at the most ambitious application perspective of ontologies: the semantic web. Thus, software engineering is essential if there is to be any hope of meeting the increasing needs for complex, high-quality ontological and linguistic resources.

Nevertheless, the common trend in AI is to develop representation languages, systems, formalisms and even concrete resources, in a rush to implement and have results as soon as possible. We reject this and choose to follow an engineering problem-solving approach. As [3] states, all engineering problem-solving is a process of elaboration where the problem is first represented at a high level of abstraction, and as the process progress, the statement of the problem moves from a representation of the essence of the solution toward implementation-specific detail. Thus, this paper does not present any logico-philosophical, cognitive-linguistic or machine learning ideas for ontology and LR construction. It is a reflection effort in our attempt to develop a methodology for managing LR [1, 5], and obtain a concrete product by following it.

2.1 A Word on Software Design

As we move from analysis to implementation, an important milestone in software development is design. Software design is a set of basic principles that provide the necessary framework for “getting it right” [3]. Nevertheless, many developers begin by “coding the solution”, that is, design begins with the coding or construction of a concrete product, and as a result, interface, architectural, and data design just happen. This approach common among people who insist upon coding or creating the product (in our case a concept-based terminology database) with no explicit design activity invariable leads to low-quality software that is difficult to test, challenging to extend, and frustrating to maintain.

2.2 A Word on Reusability

One of the most cited words in the software development and AI fields is “reusability”. However, do they mean the same?

In software engineering, reusability stands for the repetitive usage of any part of a software system [6]: the documentation, the design, the requirements, test cases, etc. In AI, claiming that a construct (e.g., an ontology, LR, etc.) is reusable entails that it can be used to express knowledge for tasks other than the ones for which it was designed [7].

Since we aim at developing a concept-based LR, it is important to establish where we stand over this issue. Following [7, 8], we believe that the task for which an application is developed fixes a particular point of view on the ontology, and the reusability of this resource for another system (i.e., another task) seems difficult. For instance, in non formal domains like medicine, where knowledge is rather descriptive than formal, this point of view is likely dependent on the application.

Moreover, the definition of an ontology is not the characterization or the determination of primitives that already exist in a domain, but the modeling or construction of primitives for the resolution of a problem. A (non-formal) domain does not have a set of general primitives and it is impossible to find, for example in medicine, the notions from which all other ideas are built [8]. However, it is necessary to have a set of primitives in order to solve a particular set of problems.

Therefore, the generality or universality of an ontology or LR is by definition limited because they are always specific to a given task (the problem(s) at hand) [8]. In spite of this, an ontology or LR can be general for a set of given tasks if and only if they possess a level of granularity that represents the points of view of each task. However, there is not a level of granularity that allows describing all the tasks of a given domain [8]. Consequently, there cannot be a universal ontology, not even for a domain, that is valid for all the possible tasks.

Having said this, we will strive to make our future resource “reusable” in the way software engineering understands this term. In addition, we will remain humble and realistic and state that the resource we will construct will be valid only in the ambit of a given domain and tasks (i.e., text summarization and categorization), and that it will not reflect the universal laws of thought or reality (i.e. it will not be a formal ontology) nor the structure of the mental lexicon (i.e. it will not have a WordNet like Structure).

3 Designing LR Using RDB

RDB present a series of advantages that have been taken into account when used to construct DB for linguistic purposes [1, 9, 10, 11, 12]. From a software engineering point of view, their main advantage is that they provide a mature design methodology, which encompasses several design stages that help designing consistent (from an integrity point of view) DB. This methodology comprises the design of the conceptual scheme (using the Entity/Relationship (E/R) model), the logical scheme (using the relational model), and the physical scheme.

However, RDB have various drawbacks when compared to newer data models (e.g., the object-oriented model): a) Impossibility of representing knowledge in form of rules; b) Inexistence of property inheritance mechanisms; and c) Lack of expressive power to represent hierarchies. In spite of this, by following a software

engineering approach, that is, by paying attention to the database design issues [10], most of these drawbacks can be overcome, and thus, let us take advantage of all the benefits of RDB.

For instance, in [11] we can see how an UML (object-oriented) model is implemented within a RDB in a way that supports inheritance and hierarchy. Another similar example is found in [13], where the authors reproduce the structure of the Mikrokosmos ontology, using the E-R model. Other models [9, 14], although machine translation oriented follow a purely linguistic approach, and are not intended to overcome any of the limitations of the relational data model.

As it can be deduced, we have focused on the limitations of RDB to represent ontologies. There are several reasons why we have done that. First, our work is focused on the design and implementation of ontology-based LR using RDB [1, 5]. Second, it has been proved by [15] that the use of a hierarchically organized concept-based terminology database, improves the results of queries on clinical data, and such is the goal of our projects. Third, we agree with [10, 16, 17, 18], when they state that the computationally proven ontological model, with two separated but linked levels of representation (i.e. the conceptual-semantic level and the lexical-semantic level) is our best choice for linguistic knowledge representation.

We have only found one reference, of a development effort that follows our software engineering approach for the development of ontology-based LR: the aforementioned work of [13]. The difference between our model [1] and the one in [13] is that ours only follows the ontological semantics ideas of Mikrokosmos; it does not recreate its frame-based structure. Nevertheless, although the model in [13] replicates the powerful ontological structure of Mikrokosmos in a RDB, it inherits all its problems (some will be described in the next section). As for the model we present in [1], it has a thesaurus-like structure where the concepts of the ontology are linked by a single implicit and imprecise relation; a situation that is problematic and severely limits the model, as it will be shown next.

4 Some Common Problems in LR

It is relatively easy to create a conceptual model of a LR. As seen in the previous section, this has already been done. However, existing LR (ontology-based or not) are plagued with flaws that severely limit their reuse and negatively impact the quality of results. Thus, it is fundamental to identify these flaws in order to avoid past and present mistakes, and create a sound conceptual model that leads to a LR where some of these errors can be avoided.

Most of the problems of past and present LR have to do with their taxonomic structure. For instance, once a hierarchy is obtained from a Machine-Readable Dictionary (MRD), it is noticed that it contains circular definitions yielding hierarchies containing loops, which are not usable in knowledge bases (KB), and ruptures in knowledge representation (e.g., a utensil is a container) that lead to wrong inferences [19]. WordNet and Mikrokosmos have also well-known problems in their taxonomic structure due to the overload of the is-a relation [20, 21]. In addition, Mikrokosmos represents semantic relations as nodes of the ontology. This entails that such representation approach where relations are embedded as nodes of the ontology is prone to suffer the same is-a overloading problems described in [20, 21], as well as

the well-known multiple inheritance ones (figure 1 illustrates this point by showing part of the Mikrokosmos ontology). In the biomedical domain, the UMLS has circularities in the structure of its Metathesaurus [22], because of its omnivorous policy for integrating hierarchies from diverse controlled medical vocabularies whose hierarchies were built using implicit and imprecise relations. Some of the consequences of these flaws, as well as additional ones have been extensively documented in [16, 17, 20, 23, 24, 25, 26, 27] for these and other main LR.

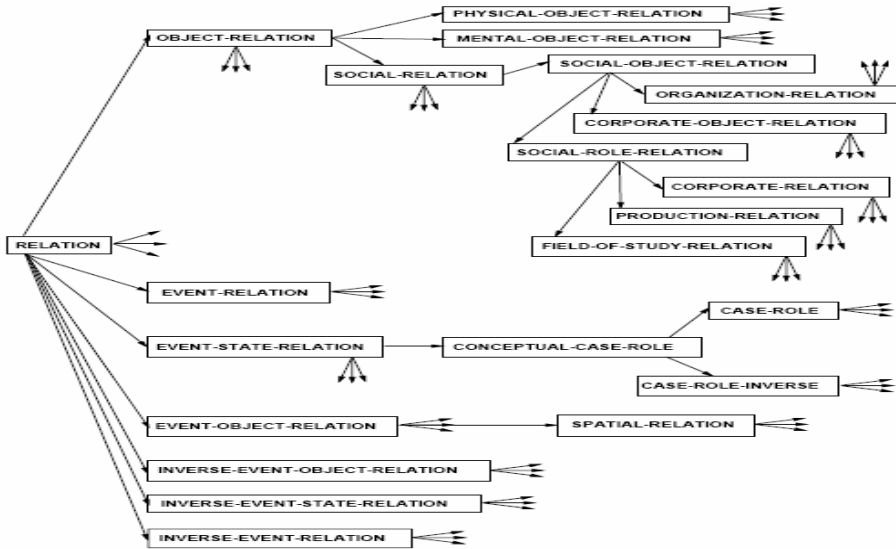


Fig. 1. Embedded Relations in the Mikrokosmos Ontology

4.1 Methodologically Engineered Problem-Solving LR

We have come a long way from the days of MRD. However, still today, the focus is on coverage and time-saving issues, rather than on semantic cleanness and application usefulness. Proof of this are the current different merging and integration efforts aimed at producing wide-coverage general LR [27, 28], and the ones aimed at (semi)automatically constructing them with machine learning methods [29, 30]. However, no amount of broad coverage will help raise the quality of output, if the coverage is prone to error [17]. We should have learned by now that there are no short cuts, and that most experiments aimed at saving time (e.g., automatically merging LR that cover the same domains, or applying resources to NLP that are not built for it, like machine-readable dictionaries and psycholinguistic-oriented word nets) are of limited practical value [31]. Furthermore, in the current trend of LR development, issues such as how to design LR are apparently less urgent. More attention must be paid on how LR are designed and developed, rather than what LR are produced.

The experience gained from past and present efforts clearly points out that a different direction must be taken. As [24] pointed out back in the days of MRD:

“rather than aiming to produce near universal LR, developers must produce application-specific LR, on a case by case basis”. In addition, we claim that these LR must be carefully conceived and designed in a systematic way, according to the principles of a software engineering methodology.

5 Methodological Gaps in the Development of LR Using RDB

Since we are interested in the development of a LR using RDB, it is worth mentioning that all the cited efforts in section 3, although they produced useful resources, forgot about the methodological nature of RDB and stopped at the conceptual design stage. Thus, there is not a complete description of the entities, relationships and constraints involved in the conceptual and logical design of the DB.

The methodology we propose in [1, 5] encompasses all of the database design phases. Nonetheless, the conceptual model from which it departs has several problems with respect to ontology representation; mainly, its does not foresee a way for clarifying the semantics of relations, a problem that as seen in section 4 is of main concern.

Hence, if we are to design a hierarchically organized concept-based terminology database using RDB, our conceptual model must take also into account the semantic relations issue. As a first step, we enhance the conceptual model presented in [1] as shown in the next section.

6 Refining the Semantics of Concepts and Relations

In order to give our first step towards the enhancement of the conceptual model, we need to clearly state what are the elements that will be abstracted and represented in our upgraded conceptual model, that will help us to: a) build application-oriented LR (as pointed out in section 4.1); and b) avoid some of the taxonomic problems present in existing LR described in section 4.

These elements are concepts, properties of concepts, relations, and algebraic and intrinsic properties of relations. They will help an ontology developer to specify for concepts and relations formal and informal semantics that clarify the intended meaning of both entities, and thus help in evaluating the semantics of the ontology [4]. Informal semantics are the textual definitions for both concepts and relations, as opposed to formal semantics that are represented by the properties of concepts and relations.

However, the fact that these elements will be part of the enhanced conceptual model does not imply that they are an imposition but rather a possibility, a recommendation that is given to each ontology developer. In the following, we detail the elements surrounding the basic element of our model: concepts.

6.1 Properties of Concepts

These are formal semantic specifications of those aspects that are of interest to the ontology developer. For instance, these specifications could be the OntoClean metaproperties described in [21] (e.g., R, I, etc.).

6.2 Relations

Instead of relations with an unclear meaning like the ones described in [7, 22, 31], we propose the use of relations with well-defined semantics, up to the granularity needed by the ontology developer. Moreover, we refuse to embed relations as nodes of the ontology (because of the problems commented in section 3) or to implicitly represent any relation as it is done in Mikrokosmos with the is-a relation. We call these, explicit relations. This represents a novelty and an improvement when compared to similar design and implementation efforts as [13] based on RDB and ontological semantics notions. In the next two subsections, we will describe the elements that help clarifying the semantics of relations.

6.3 Algebraic Properties of Relations

The meaning of each relation between two concepts must be established, supported by a set of algebraic properties from which, formal definitions could be obtained (e.g., transitivity, asymmetry, reflexivity, etc.). This will allow reasoning applications to automatically derive information from the resource, or detect errors in the ontology [32]. Moreover, the definitions and algebraic properties will ensure that the corresponding and probably general-purpose relational expressions are used in a uniform way [32]. Tables 1 and 2 (taken from [32]) show a set of relations with their definitions and algebraic properties.

Table 1. Definitions and Examples of Relations

Relations	Definitions	Examples
<i>C is-a C₁</i>	Every <i>C</i> at any time is at the same time a <i>C₁</i>	<i>myelin is-a lipoprotein</i>
<i>C part-of C₁</i>	Every <i>C</i> at any time is part of some <i>C₁</i> at the same time	<i>nucleoplasm part-of nucleus</i>

Table 2. Algebraic Properties of Some Relations

Relations	Transitive	Symmetric	Reflexive
Is-a	+	-	+
part-of	+	-	-

6.4 Intrinsic Properties of Relations

How do we assess, for a given domain, if a specific relation can exist between two concepts? The definitions and algebraic properties of relations, although useful are not enough. As [21] point out, we need something more. Thus, for each relation, there must be a set of properties that both a child and its parent concept must fulfill for a specific relation to exist between them. We call these properties, intrinsic properties of relations. For instance, in [21] the authors give several examples (according to their

methodology) of the properties that two concepts must have so that between them there can be an is-a relation.

7 Designing the Conceptual-Semantic Level of the Concept-Based Terminology Database

In this section, we present the conceptual model (an E/R scheme upgraded from our model in [1]) shown in figure 2, for the conceptual-semantic level of our future terminology database as a result of the first design phase, where all the ideas described in section 6 have been incorporated. However, as it was previously established, the model will reflect only the ontology part of our future hierarchically organized concept-based terminology database.

The entity set Concepts denotes the meaning of words, and it has two attributes: ConceptID (artificial attribute intended only for entity identification), and ConceptDefinition, intended for the textual definition of the meaning (informal semantics). The entity set ConceptProperties represents the set of formal properties described in section 6.1, and it has one attribute: ConceptProperty used to represent each property.

The entity set Relations represents the set of relations that can exist in an ontology, and it has two attributes: Relation that captures the textual name of each relation (e.g., is-a, part-of, etc.), and RelationDefinition for the textual definition of relations (informal semantics) as illustrated in table 1.

The entity set AlgebraicProperties represents the properties of relations (formal semantics) as seen in table 2, and it has one attribute: AlgebraicProperty that denotes each algebraic property. The entity set IntrinsicProperties conveys the set of properties mentioned in section 6.4 and has one attribute: IntrinsicProperty which represents each intrinsic property.

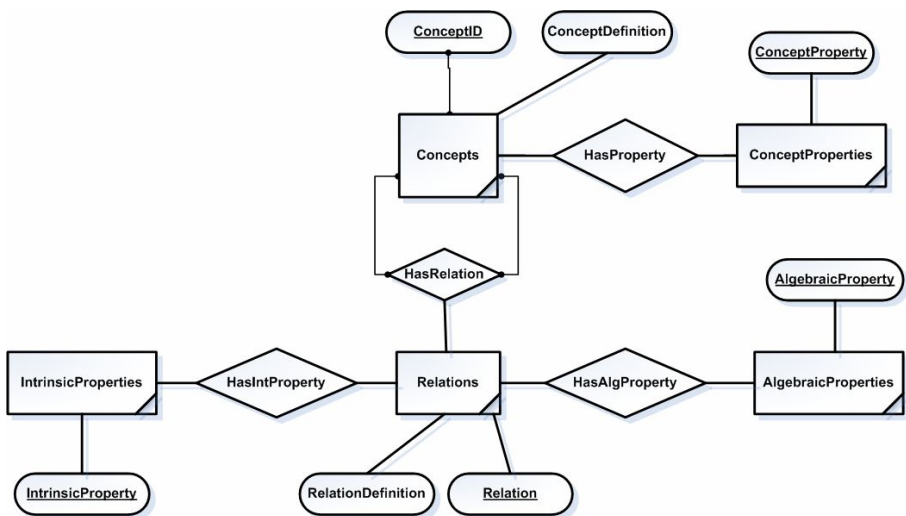


Fig. 2. Conceptual Model for an Ontology-Based LR

The relationship set *HasProperty* is used to assign properties to concepts. The ternary relationship set *HasRelation* is used to represent that two concepts in an ontology can be linked by a given relation. The relationship set *HasAlgProperty* is used to convey that relations could have attached a set of algebraic properties; the same applies for the relationship set *HasIntProperty*, but for intrinsic properties.

8 Conclusions and Future Work

The use of RDB to represent lexical knowledge provides a complete software engineering methodological approach for the design of the database that will contain the LR. However, the approaches that use this technology sometimes only present an E-R schema and forget about the rest of the DB development stages or simply state that they use RDB. This is far from being adequate, as LR to be used by domain specific applications need to be developed in such a way that all the modeling choices are clearly stated and documented.

With that in mind, we have chosen to develop our future terminology database following a sound software engineering methodology. However, the proposed conceptual model of the methodology had some major limitations. In order to overcome them, we modified it based on an analysis of common problems in LR. The new model can now account for any number of ontological relations (as long as they are binary), and we have incorporated a set of ideas that help designing problem-solving ontology-based LR where the semantics of relations is clearly stated and the use of relations can be controlled (e.g., the model allows the integration of the *OntoClean* [21] method for evaluating taxonomies). Moreover, although we have selected RDB to represent lexical and conceptual knowledge, the model is totally independent of any knowledge representation schema (i.e., DB or knowledge bases).

We still have to go through the logical and physical design stages of the database. However, we have taken a first step towards our final goal, by clearly stating and depicting the structure, scope and limitations of our future LR. However, although we have focused on the ontology side of the model, the lexical side of our previous model (see [1]) also needs to be upgraded as it is quite limited. Thus, we are considering the integration of the E-R model for the lexical side of an ontology-based LR proposed and described in [10].

Something that must be clearly understood is that our goal is the establishment of a software engineering methodology for the design and implementation of ontology-based LR using RDB. It is not a methodology aimed at saving time by constructing or extracting a LR from texts using machine learning methods [27, 28] or by merging different LR [29, 30]. We follow a software engineering problem-solving approach that focus on analysis, design and reuse (as understood by software engineering) and apply the principled methods and techniques of software engineering (which guide the development of user-oriented, readable, modular, extensible, and reusable software) to the design and implementation of ontology-based LR with RDB.

Finally, we are also considering coupling our software engineering methodology with the one of [7, 8] to build a prototype of a LR for Pneumonia related documents (i.e. articles) that include a set of software engineered creation and management tools.

References

1. Sáenz, F. and Vaquero, A. Applying Relational Database Development Methodologies to the Design of Lexical Databases. Database Systems 2005, IADIS Virtual Multi Conference on Computer Science and Information Systems (MCCSIS), (2005)
2. Maña, M. et al. Los proyectos SINAMED e ISIS: Mejoras en el Acceso a la Información Biomédica Mediante la Integración de Generación de Resúmenes, Categorización Automática de Textos y Ontologías. En Actas del XXII Congreso de la Sociedad Española de Procesamiento del Lenguaje (SEPLN), (2006)
3. Pressman, R. Software Engineering. In Merlin Dorfman and Richard H. Thayer (Eds.) Software Engineering. Wiley-IEEE Computer Society Press, (1999)
4. York, S., Gómez-Pérez, A., Daelemans, W., Reinberger, M., Guarino, N., Fridman, N. Why Evaluate Ontology Technologies? Because It Works! IEEE Intelligent Systems 19(4): 74-81, (2004)
5. Sáenz, F. and Vaquero, A. Towards a Development Methodology for Managing Linguistic Knowledge Bases. Research and Development in Intelligent Systems XIX. Springer, Cambridge (United Kingdom), (2002)
6. Pfleeger, S. Software Engineering: Theory and Practice, Second Edition. Prentice Hall, (2001)
7. Bouaud, J., Bachimont, B., Charlet, J. and Zweigenbaum, P. Acquisition and Structuring of an Ontology within Conceptual Graphs. In Proceedings of the ICCS'94 Workshop on Knowledge Acquisition using Conceptual Graph Theory, (1994)
8. Bachimont, B. Engagement sémantique et engagement ontologique: conception et réalisation d'ontologies en Ingénierie des connaissances. In J. Charlet, M. Zacklad, G. Kassel & D. Bourigault (Eds.). Ingénierie des connaissances, évolutions récentes et nouveaux défis, (2000)
9. Bläser, B; Schwall, U and Storrer, A. Reusable Lexical Database Tool for Machine Translation. In Proceedings of the International Conference on Computational Linguistics - COLING'92, volume II, (1992) pp. 510-516.
10. Moreno A. Diseño e Implementación de un Lexicón Computacional para Lexicografía y Traducción Automática. Estudios de Lingüística Española, vol(9). (2000)
11. Hayashi, L. S. and Hatton, J. Combining UML, XML and Relational Database Technologies - The Best of all Worlds for Robust Linguistic Databases. In Proceedings of the IRCS Workshop on Linguistic Databases. (2001)
12. Wittenburg, P. et al. Databases for Linguistic Purposes: a case study of being always too early and too late. In Proceedings of the EMELD Workshop. (2004)
13. Moreno, A. and Pérez, C. Reusing the Mikrokosmos Ontology for Concept-Based Multilingual Terminology Databases. In Proc. of the 2nd International Conference on Language Resources and Evaluation, (2000)
14. Tiedemann, J. MatsLex: A multilingual lexical database for machine translation. In Proc. of the 3rd International Conference on Language Resources and Evaluation, (2002)
15. Lieberman, M. The Use of SNOMED to Enhance Querying of a Clinical Data Warehouse. A thesis presented to the Division of Medical Informatics and Outcomes Research and the Oregon Health & Sciences University School of Medicine in partial fulfillment of the requirements for the degree of Master of Science. (2003)
16. Nirenburg, S., McShane, M. and Beale, S. The Rationale for Building Resources Expressly for NLP. In Proc. of the 4th International Conference on Language Resources and Evaluation, (2004)

17. McShane, M.; Nirenburg, S. and Beale, S. An implemented, integrative approach to ontology-based NLP and interlingua . Working Paper #06-05, Institute for Language and Information Technologies, University of Maryland Baltimore County, (2005)
18. Cimino, J. Desiderata for Controlled Medical Vocabularies in the Twenty-first Century. *Methods of Information in Medicine*, 37(4-5):394-403, (1998)
19. Ide, N., and Veronis, J. Extracting Knowledge Bases from Machine-Readable Dictionaries: Have we wasted our time? In *Proc. of the First International Conference on Building and Sharing of Very Large-Scale Knowledge Bases*, (1993)
20. Guarino, N. Some Ontological Principles for Designing Upper Level Lexical Resources. A. Rubio et al. (eds.), In *Proc. of the First International Conference on Language Resources and Evaluation*, (1998)
21. Welty, C. and Guarino, N. Supporting ontological analysis of taxonomic relationships", *Data and Knowledge Engineering* vol. 39(1), (2001)
22. Bodenreider O. Circular Hierarchical Relationships in the UMLS: Etiology, Diagnosis, Treatment, Complications and Prevention. In *Proceedings of the AMIA Symposium*, (2001)
23. Feliu, J.; Vivaldi, J.; Cabré, M.T. Ontologies: a review. Working Paper, 34. Barcelona: Institut Universitari de Lingüística Aplicada, (2002)
24. Evans, R., and Kilgariff, A. MRDs, Standards and How to do Lexical Engineering. *Proc. of 2nd Language Engineering Convention*, (1995)
25. Burgun, A. and Bodenreider, O. Aspects of the Taxonomic Relation in the Biomedical Domain. In *Proc. of the 2nd International Conference on Formal Ontologies in Information Systems*, (2001)
26. Martin, P. Correction and Extension of WordNet 1.7. In *Proc. of the 11th International Conference on Conceptual Structures*, (2003)
27. Oltramari, A.; Prevot, L.; Borgo, S. Theoretical and Practical Aspects of Interfacing Ontologies and Lexical Resources. In *Proc. of the 2nd Italian SWAP workshop*, (2005).
28. Philpot, A., Hovy, E. and Pantel, P. The Omega Ontology. In *IJCNLP Workshop on Ontologies and Lexical Resources*, (2005)
29. Makagonov, P. et al. Learning a Domain Ontology from Hierarchically Structured Texts. In *Proc. of Workshop "Learning and Extending Lexical Ontologies by using Machine Learning Methods"* at 22nd International Conference on Machine Learning, (2005)
30. Makagonov, P. et al. Studying Evolution of a Branch of Knowledge by Constructing and Analyzing Its Ontology. In Christian Kop, Günther Fliedl, Heinrich C. Mayr, Elisabeth Métais (eds.). *Natural Language Processing and Information Systems. 11th International Conference on Applications of Natural Language to Information Systems*, (2006)
31. Nirenburg, S., McShane, M., Zabłudowski, M., Beale, S. and Pfeifer, C. Ontological Semantic text processing in the biomedical domain. Working Paper #03-05, Institute for Language and Information Technologies, University of Maryland Baltimore County, (2005)
32. Smith B, Ceusters W, Klagges B, Köhler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C. Relations in Biomedical Ontologies. *Genome Biology*, 6(5), (2005)

Grid-Based Knowledge Discovery in Clinico-Genomic Data

Michael May¹, George Potamias², and Stefan Rüping¹

¹ Fraunhofer AIS, Schloss Birlinghoven, 53754 St. Augustin, Germany
`{michael.may,stefan.rueping}@ais.fraunhofer.de`
<http://www.ais.fraunhofer.de>

² Institute of Computer Science, FORTH, Heraklion, Crete, Greece
`potamias@ics.forth.gr`
<http://www.ics.forth.gr>

Abstract. Knowledge discovery in clinico-genomic data is a task that requires to integrate not only highly heterogeneous kinds of data, but also the requirements and interests of very different user groups. Technologies of grid computing promise to be an effective tool to combine all these requirements into a single architecture. In this paper, we describe scenarios and future research directions related to grid-based knowledge discovery in clinico-genomic data, and introduce the approach taken by the recently launched ACGT project¹. The whole endeavor is considered in the context of biomedical informatics research and aims towards the realization of an integrated and grid-enabled biomedical infrastructure. The presented integrated clinico-genomics knowledge discovery (ICGKD) scenario and its process realization is based on a multi-strategy data-mining approach that seamlessly integrates three distinct data-mining components: clustering, association rules mining, and feature-selection. Preliminary experimental results are indicative of the rational and reliability of the approach.

1 Introduction

Recent advances in post-genomics research have resulted in an explosion of information, data and knowledge about major diseases, such as cancer, and their treatment. As a result, the application of related technologies to the study of diseases is slowly shifting to the analysis of clinically relevant samples such as fresh biopsy specimens and fluids. The respective scientific and technological challenges push for trans-disciplinary team science and translational research as the means to bring basic discoveries closer to the bedside [1]. In this context the design, development and delivery of up-to-date methods, systems and tools to support knowledge discovery in clinico-genomic data is of major importance. This task is comprised in the research agenda put forward by the scientific discipline of Biomedical Informatics BMI [2], also realized by various EU projects, e.g. INFOBIOMED², and the recently launched ACGT³ project. BMI melds the

¹ <http://www.eu-acgt-org>

² <http://www.infobiomed.org>

³ <http://eu-acgt.org>

study of biomedical computer science with analyses of biomedical information and knowledge, thereby addressing specifically the interface between computer science and biomedical science.

The effective and efficient management and use of stored data, and in particular the transformation of these data into information and knowledge, is a key requirement for success in the biomedical domain. Knowledge Discovery (KD, aka Data Mining) is the de-facto technology addressing this information need. Data mining technology is used for the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. However, this field has mainly been concerned with small to moderately sized data sets, knowledge-weak domains, and within the contest of largely homogeneous and localized computing environments. These assumptions are increasingly not met in modern scientific environments. The shift to large-scale distributed computing has profound implications in terms of the way data are analyzed. Future data mining applications will need to operate on massive data sets and against the backdrop of complex domain knowledge. The domain knowledge (computer-based and human-based), the data sets themselves, and the programs for processing, analyzing, evaluating, and visualizing the data, and other relevant resources will increasingly reside at geographically distributed sites on heterogeneous infrastructures and platforms. Grid computing [3] promises to become an essential technology capable of addressing the changing computing requirements of future distributed knowledge discovery environments. Currently, several project such as DataMiningGrid⁴ [4] and SIMDAT⁵ are concerned with merging generic knowledge discovery services with grid technologies.

In this paper, we will review recent developments on grid-based biomedical research and point out future directions to facilitate an integrated, knowledge-rich, and highly performant analysis of clinical and genomic data. With respect to the last point, we present an Integrated Clinico-Genomics Knowledge Discovery (ICGKD) scenario and its realization be a multi-strategy data-mining process that smoothly integrates three data-mining approaches: clustering, association rules mining, and feature-selection.

The rest of the paper is organized as follows: the next section will introduce the main challenges of KD in clinico-genomic data, while Section 3 introduces grid-enabled Knowledge Discovery. In Section 4 we will present novel research directions targeted at solving the biomedical challenges using grid-based KD, in particular the approach taken by the ACGT project. Section 5 gives an example by describing the ICGKD scenario and its realization. In Section 6 a real-world case study is presented. Finally we conclude in Section 7.

2 Challenges of Knowledge Discovery in Clinico-Genomics

Data mining methodology and technology has been developed for classical business, finance, and customer-oriented application domains. Such domains are

⁴ <http://www.datamininggrid.org/>

⁵ <http://www.scai.fraunhofer.de/simdat.html>

characterized by the availability of large quantities of data in an attribute-value based representation, high ratio of examples over attributes in the data set, and weak background knowledge about the underlying entities and processes.

For biomedical data these conditions do not hold. Although technologies like microarrays for gene expression profiling are rapidly developing, today it still remains an expensive technology. In addition, legal, ethical and practical limitations in clinical trials make it cumbersome to acquire a high number of patients in a clinical trial. As a result, a typical genomic data may contain only about 100 examples. At the same time, the same data sets consists of more than 10^4 attributes (genes). Under these conditions, standard statistical and machine learning methods are likely to over-fit the structures in the data, such that a high amount of domain knowledge is needed to guide the analysis and guarantee the validity of the extracted knowledge.

A specific property of the biomedical domain that make it very challenging for knowledge discovery is its heterogeneity, both in terms of data and in terms of use cases. Concerning the data, next to genomic information very different forms of data, such as classical clinical information (diagnosis, treatments, vital signs) and imaging data (x-rays, CTs) have to be integrated into the analysis. Additionally, most of the high-level knowledge is present in electronic texts, such as journal papers, which can be exploited by methods of text mining. Use cases can differ very much because of the different user groups involved. There are at least three users groups, the clinicians, who want to treat single patients, biomedical researchers which want to acquire new knowledge about genes, and data miners, which are interested in the analysis algorithms per se. All these groups have different interests [5] and very different expertise and views on the same problem. A fruitful collaboration requires that it is easy for each user to benefit from the knowledge of the other user groups without needing to become an expert himself.

3 Grid-Enabled Knowledge Discovery

Grid computing [3] is a generic enabling technology for distributed computing. It is based on a hardware and software infrastructure that provides dependable, consistent, pervasive and inexpensive access to computing resources anywhere and anytime. In their basic form, these resources provide raw compute power and massive storage capacity. These two Grid dimensions were originally dubbed Computational Grid and Data Grid, respectively. However, since the inception of Grid technology, the term resource has evolved to cover a wide spectrum of concepts. Standard grid solutions provide services such as job execution and monitoring, parallelization, distributed data access and security.

3.1 Data Mining on the Grid

The combination of data mining and grid technology offers many interesting scenarios for scaling up data mining tasks and approaching tasks not possible before in stand-alone or cluster environments:

- Distributing data: the integration of relevant, heterogeneous, possibly steadily updated information from distributed sites is practically a very difficult task without standardization. In particular, in the application of models to unseen data, this data may come from very different sites than the original data set that was used for learning.
- Distributing computation: end users often may not have large computational resources at their hands, but may need to rely on other high-performance computing facilities made available to them.
- Flexible combination of both: a particular property of knowledge discovery is that the input data is typically not analyzed in its raw form, but several pre-processing steps are needed. Applying these pre-processing steps directly at the sites where the data is located can result in a massive reduction of the size of the data that needs to be transported.
- Parallel computation: In the majority of cases, data mining algorithms consist of a set of almost identical, independent tasks, e.g. for parameter sweeps, cross-validation, or feature selection. These are trivially parallelizable and can be executed on available low-cost processors. For example, in many organizations, such as a large hospital, the administrative computers are idle during the night and could be integrated for in-house data mining tasks.

All these tasks by themselves are not very complex and appropriate techniques have been well-known for years. The important new contribution of grid technology is to provide a standard architecture that guarantees the correct execution of the jobs, the consistency of the data, and the easy delivery of data and algorithms across different sites.

An approach to provide standard data mining services across a grid infrastructure is the integration of standard statistical and data mining toolkits, such as Weka [6] or R [7]. This approach was followed by the DataMiningGrid project, which allows to integrate Weka operators into a overall grid workflow. Figure 1 presents such an workflow. These workflows are described in the form of an XML document, which is executed by the main grid engine.

4 The ACGT Approach to Grid-Based Clinico-Genomics Knowledge Discovery

The recently started ACGT project aims at:

- the delivery of a European biomedical grid infrastructure offering seamless mediation services for sharing data and data-processing methods and tools, and advanced security;
- the semantic, ontology based integration of clinical and genomic/proteomic information and data;
- the delivery of data-mining grid services in order to support and improve complex knowledge discovery processes.

The main challenge from the knowledge discovery side of the project is the sharing of knowledge, either in the form of the integration of existing knowledge to

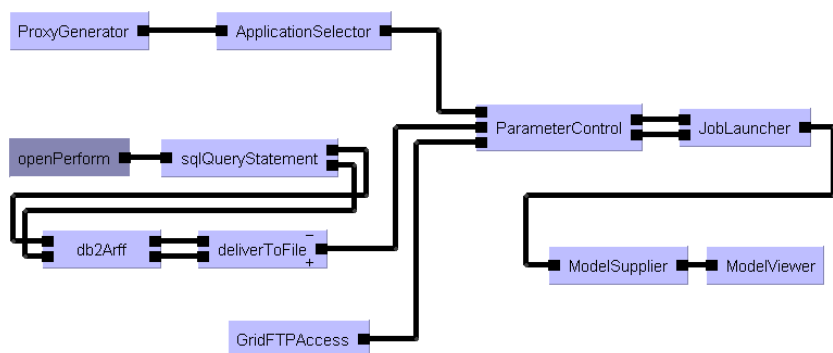


Fig. 1. An integrated grid and data mining workflow, including the selection of a data mining operator (upper left hand side), selection of distributed data (lower left), job execution (upper right) and display of the results (lower right).

design and select appropriate analysis tools, or to manage the discovered knowledge in order to make it available to other researcher. The efficient management of the different views and expertise of clinicians, biologists and data miners will be crucial. For the underlying principles of ACGT and the way it copes with these issues see [8]. In particular, we propose the use of the following techniques

- An ontology-based description of the application domain- taking into account standard clinical and genomic ontologies, nomenclatures and metadata, in order to retain semantic on all steps of the analysis and to guide the construction of data mining workflows.
- A matching ontology to describe data mining tasks and operators, including support to correctly translate research questions of the application domain into specific data mining tasks.
- A database of workflows plus appropriate meta-information to serve as a case base for selecting promising candidate workflows from similar tasks.
- The use of text-mining techniques to extract relevant knowledge from published papers. Text mining can also be used to connect research questions to data mining tasks and algorithms in a more freely structured way, when workflow descriptions and corresponding paper abstracts are joined into one document and are added to the available document corpus.
- The massive parallel use of data mining algorithms to search for dependencies in the cases where no prior expert knowledge is available.

5 An Exemplary Scenario

In order to exemplify our approach, let us discuss an Integrated Clinico-Genomic Knowledge Discovery (ICGKD) scenario that has been discussed in [9]. The scenario is depicted in Figure 2 and consists of three steps, first the clustering of genes based on their gene-expression patterns in order to identify potentially

useful subsets of genes, the discovery of association rules to discover causal relations between genes and clinical attributes (in this case the prognostic status for breast cancer patients), and post-processing by feature selection in order to focus on the most discriminating genes, based on both accuracy and experts' domain knowledge.

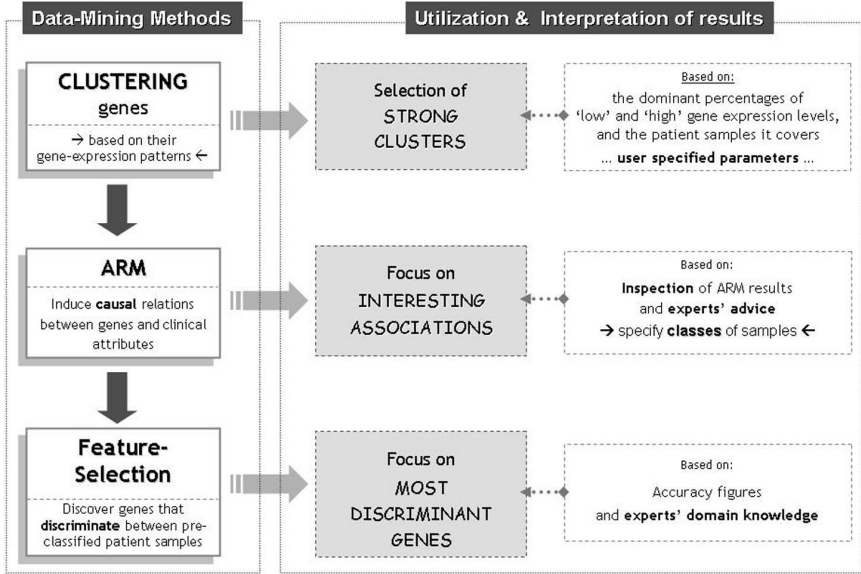


Fig. 2. Integrated Clinico-Genomic Knowledge Discovery Scenario

Using the approaches described in Section 4, this scenario can be extended and realized in multiple ways

- A researcher looking for novel ways to analyze some type of cancer that is known to be related to breast cancer could find the workflow described here by using an ontology-based query to the workflow case base. Executing this workflow on the new set of data would only need to replace the data selection part of the workflow.
- The workflow designer could be interested in improving the analysis by comparing different kinds of *clustering* algorithms in order to identify clusters of genes that exhibit significant relations with specific patient samples. Most standard data mining toolkits already come with a large variety of clustering algorithms, and using a grid environment the parallel execution and comparison of all these alternatives is very easy. In the real-world application presented in the next section the clustering operation is realized by a novel *k-means* clustering algorithm that operates on discretised gene-expression data.

- Next the researcher may be interested to identify ‘strong’ associations (i.e., utilizing a frequent itemset discovery algorithm) between particular patients’ clinical profiles and features with the gene-clusters discovered in the second step. The developer of the association rule algorithm could be interested in developing a parallel version to speed up the computation. This task is very much alleviated by the use of standard components in a grid-aware toolkit. For the real-world experiment presented in the next section a grid-enabled association rules mining system is utilized, the *HealthObs* system [10].
- Focusing on the genes and samples covered by discovered *high-confident* association rules the researcher may be then interested to identify potential *gene-markers*, i.e., genes that best *discriminate* between specific patient status (e.g., good vs. bad prognosis). Particular parallelized and grid-enabled feature-selection and classification algorithms from standard data-mining toolkits could be utilized. The real-world experimental application in the next section utilizes a particular gene-selection system, the MineGene system [11].

5.1 Towards Strong Clinico-Genomic Profiles

A lot of work has been done in identifying co-regulated groups or, clusters of genes [12], clusters of patient samples [13], discriminant set of genes [14], and methods to reduce the dimensionality and complexity of gene-expression data [15]. In this paper we introduce and utilize a discretized two-dimensional k-means clustering algorithm, named *discr-kmeans*, that primary identifies clusters of co-regulated genes. The algorithm resembles similar approaches presented in [16] and [17], and further exemplified in [18]. With a subsequent filtering approach the gene clusters that exhibit, in an adequate number of samples, ‘strong’ gene-expression profiles are selected. A sample with a strong gene-expression profile for a specific cluster of genes is one that exhibits *dominantly* ‘high’ or, ‘low’ gene expression levels. The adequate number of samples, as well as the percentage for considering a sample’s expression profile as strong is set by the user.

Assume s samples, g genes, and a 2-dimensional matrix, $M(s \times g)$ that holds the respective gene-expression matrix. The *discr-kmeans* algorithmic process unfolds into two steps:

- Step-1: *Discretization*. We proceed with a method to overcome the error-prone variance of gene-expression levels by discretizing the respective continuous gene-expression values. A gene-expression value may be assigned to an (ordered) nominal value; assume n such ordered values. In the case of $n = 2$, value 1 is interpreted as of *low*, and value 2 as of *high* expression level. Define,

$$w_i = \frac{\max(g_i) - \min(g_i)}{n}$$

where, $\max(g_i)$ and $\min(g_i)$ the maximum and minimum expression values of gene g_i , respectively. The discretized transform, $V_d(s_j, g_i)$, of gene’s g_i continuous value, $V(s_j, g_i)$ in sample s_j is computed by:

$$V_d(s_j, g_i) = \begin{cases} n & \text{if } V(s_j, g_i) = \max(g_i) \\ \lfloor \frac{V(s_j, g_i) - \min(g_i)}{w_i} \rfloor + 1 & \text{otherwise} \end{cases}$$

where, $\lfloor \text{fraction} \rfloor$ the integer part of the fraction. It can be easily checked that the computed gene's discretized values range from 1 to n .

- Step-2: *Clustering*. The main difference between normal k-means and discr-kmeans is that each cluster's center is not represented by the average value of the cluster's genes values but, by a 2-dimensional matrix that contains the percentage of the discretized cluster's gene values, $C_k(s, n)$. For a sample s_j , and $p \in [1, n]$, $C_k(s_j, p)$ is the percentage of genes in cluster k , the discretized expression-values of which, with respect to sample s_j , is p . For example, in a domain with three samples and discretization value $n = 2$, a cluster's center is an array like the following:

Value $\rightarrow$	1 (<i>low</i>)	2 (<i>high</i>)
Sample $\downarrow$	%	%
1	80	20
2	55	45
3	10	90

In the above example matrix, 80% of all the genes in the cluster exhibit low (discretized value = 1) expression levels in sample 1. Analogously, 90% of the cluster genes exhibit a high expression profile (discretized value = 2) in sample 3. With appropriate (user defined) thresholds, these profiles are assumed to *dominate* the respective samples. In other words, the induced cluster of genes seems to be linked and correlated with dominantly low, or high expression profiles for the specific samples. Clustering unfolds into the standard k-means iterative process. A basic difference is the way that the distance of a gene and the center of a cluster is computed. The distance between a cluster $C_k(s, n)$ and a gene g_i is computed as:

$$\text{dist}(C_k, g_i) = \sum_{l=1}^s C_k(s_l, V_d(s_l, g_i))$$

After clustering process converges, we end-up not only with gene clusters but with an indication of how 'strong' a cluster is. A cluster is considered as 'strong' if it exhibits dominant discretized value percentages in an adequate number of samples, i.e, close to 0%, or close to 100%. We are interested in 'strong' clusters because we want to identify potential subsets of samples that tend to exhibit mainly dominantly high or low expression levels for the respective genes in a cluster. This is why we decide to discretize the continuous gene-expression levels with a discretization value of $n = 2$. For these samples – referred as *strong samples*, the respective cluster's genes tend to be dominantly up- or, down-regulated. The genes of a cluster, accompanied by their respective strong samples may be interpreted as a combined *clinico-genomic attribute* linking patient cases and their genomic (gene-expression) profiles. The quest now is about the *causal* relations that hold between such genomic and clinical profiles.

5.2 Causal Relations in Clinico-Genomic Profiles

In the present study we utilized HealthObs, a system that incorporates Association Rules Mining (ARM) operations specially suited for the clinical domain [10]). HealthObs is able to operate over an integrated electronic health care record environment [19]. The special services that HealthObs brings relate to: (i) ease in query formulation, via a friendly GUI interface- flexible enough to enhance the naturalness of data exploration inquiries; (ii) imposition and utilization of ARM operations directly on-top of XML structures (instead of flat files or, specific databases); and (iii) friendly visualization operations that ease inspection, filtering and interpretation of the discovered association rules.

5.3 Selecting Discriminatory Genes Via Feature-Selection

For the case of the clinico-genomic inquiry and exploration process, each association rule may be taken as a medium to focus on the genes and patient cases covered by it. The expert (molecular biologist or, physician) may inspect the discovered association rules and focus on the ones that seem more interesting for the scope of the inquiry. Then, a *gene-selection* process may be called to operate just on the sets of genes and patient cases being covered by the focused association rules. Provided that a specific clinical feature of interest is targeted (e.g., 'survival over 5 years' vs. 'survival less than 5 years') particular gene-markers may be identified. In the present study we utilize a *feature-selection* method specially suited for the task of selecting discriminant genes, i.e., set of genes able to distinguish between particular pres-classified patient samples. A detailed description of the method may be found in [11]. The method is implemented in an integrated system for mining gene-expression (microarray) data - the MineGene⁶ system. The method is composed by three components: discretization of gene-expression data, ranking of genes, and greedy feature-elimination (or, addition) accompanied with a classification metric to predict patient class categories (e.g., clinical outcome).

6 A Real-World Application

We applied the presented ICGKD scenario on a real world clinico-genomic domain. For the reference study (accompanied with public available data⁷) with which we compare our findings see [20]. The data includes the gene-expression profiles of 24.481 genes over 78 breast-cancer patient samples; 44 of them with a status of *over five years* survival, and 34 with a status of *less than five years* survival. The clinical profiles of the patients are also provided. The clinical data refer to a number of features including: *age* of the patient; *Lymphocytic-Infiltration status* of the tumour; the *estrogen and progesterone receptor profiles* of the patients, as well as their prognostic status - *bad* or, *good*, for less and over five

⁶ <http://dlib.libh.uoc.gr/Dienst/Repository/2.0/Body/uch.csd.msc/2005kanterakis/pdf>

⁷ <http://www.rii.com/publications/2002/vantveer.html>

years survival, respectively. In the reported experiment we focus on the clinical-outcome feature, trying to discover reliable associations between the prognostic profiles of patients and their gene-expression background. Experimental set-up and findings follow.

- *Clustering genes and selection of strong clusters.* For discr-kmeans clustering operation the requested number of clusters was set to 90 (so that all input genes are appropriately covered). In order to consider a cluster of genes as strong (i.e., set of genes which exhibit dominantly up- or, down-regulated profiles in an adequate number of samples) the following parameters were set: *minimum – number – of – genes* ≥ 100 ; *minimum – number – of – samples – with – a – dominant – genes – profile* ≥ 10 ; and *percentage – of – genes – with – dominant – genes – profile – per – sample* $\geq 90\%$ of up- or, under-regulated genes in the respective cluster. We ended up with a set of 13 gene-clusters.
- *Association rules mining and causal clinico-genomic relations.* In order to find informative and highly confident association rules we selected all the genomic features to participate in the *IF* scope of the rule and the *follow-up* (the clinical-outcome or, prognosis) feature to participate in the *THEN* part of them (a service offered by the HealthObs system). We set *minsup* = 10, and *minconf* = 70 for the minimum support and confidence of each rule, respectively. In the resulting association rules only 3 out of the 13 gene clusters appeared, covering 37 (from a total of 78 input) samples, and 5936 genes (5503, 284 and 149 for the three respective clusters).
- *Gene selection.* Applying the gene-selection process of MineGene on the set of 37 sample cases, and the set of 5936 genes we end-up with a set of 100 most-discriminant genes that exhibit an (fitness) accuracy figure of 100% (column 2 and 3 in Table 1, respectively). The gene-selection process was also performed on all 78 patient-samples, as well as on an independent test-set of 19 patient samples (columns 4 and 5 in Table 1, respectively). The presented results are indicative for the rational and reliability of the ICGKD approach.

Table 1. Comparative accuracy results (for gene-selection) after running the presented ICGKD process (*SG: number of selected genes, **samples selected by the ICGKD process, ***NA: not applicable).

	#SG*	37 samples**	78 total samples	19 test samples
ICGKD	100	100%	85.9%	89.5%
Reference study	70	NA***	80.8%	89.5%

7 Conclusions

Recent advances in post-genomics and especially in high-throughput technology (e.g., microarrays) offer the means to examine and profile the expression of all

human genes and relate them with patients' disease profiles. It is an effective approach for developing disease prognostics expected to result into the identification of strong candidate targets for diagnosis and therapeutic intervention. The inherited huge amount of genomic information and respective patients' data calls for advanced data-mining tools and respective high-performing environments.

Grid technology promises to be an effective way to easily combine existing, previously independent approaches for knowledge discovery in clinico-genomic data into a single framework. The recently launched integrated project ACGT aims towards this direction. The provisioned technological platform will be validated in a concrete setting of advanced clinical trials on Cancer.

In this setting solutions to the problem of reducing the dimensionality of the search space, i.e., from thousands of genes to the most disease-status discriminant ones, are crucial in order to cope with the intrinsic noise and deliver reliable diagnostic and prognostic molecular/gene-markers. This is the target of the presented clinico-genomic knowledge discovery scenario and its realization via the smooth integration of different data-mining, namely: *which patients' clinical profiles relate and how with their respective genomic background*.

We expect this research direction to yield important, clinically relevant new results, but also to pose new questions for machine learning, data-mining and knowledge discovery.

Acknowledgments

The financial support of the European Commission (Project INFOBIOMED NoE, FP6/2004/IST-507585, and Project ACGT,FP6/2004/IST-026996) is gratefully acknowledged.

References

1. Sander, C.: Genomic Medicine and the Future of Health Care. *Science* **287**(5460) (2000) 1977–1978
2. Martin-Sanchez, F., et al.: Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care. *Journal of Biomedical Informatics* **37**(1) (2004) 30–42
3. Foster, I., Kesselman, C., eds.: *The Grid: Blueprint for a New Computing Infrastructure*. 2nd edn. Morgan Kaufmann (2004)
4. Stankovski, V., May, M., Franke, J., Schuster, A., McCourt, D., Dubitzky, W.: A service-centric perspective for data mining in complex problem solving environments. In: *Proc. Int. Conf. on Parallel and Distributed Processing Techniques and Applications (PDPTA'04)*. Volume II., Las Vegas, USA (2004) 780–787
5. Parks, M.R., Disis, M.L.: Conflicts of interest in translational research. *Journal of Translational Medicine* **2**(28) (2004) 1–4
6. Witten, I., Frank, E.: *Data Mining – Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann (2000)
7. R Development Core Team: *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (2005) ISBN 3-900051-07-0.

8. Tsiknakis, M., Kafetzopoulos, D., Potamias, G., Analyti, A., Marias, K., Manganas, A.: Building a European Biomedical Grid on Cancer: The ACGT Integrated Project. *Stud Health Technol Inform.* **120** (2006) 247–258
9. Potamias, G., Tsiknakis, M., Papoutsidis, V., Kanterakis, A., Marias, K., Kafetzopoulos, D.: Advancing Clinico-Genomic Research Trials via Integrated Knowledge Discovery Operations. In: MIE2006, (poster presentation). (2006)
10. Potamias, G., Koumakis, L., Moustakis, V.: Mining XML Clinical Data: The HealthObs System. *Ingenierie des systems d'information, special session: Recherche, extraction et exploration d'information* **10**(1) (2004) 59–79
11. Potamias, G., Koumakis, L., Moustakis, V.: Gene Selection via Discretized Gene-Expression Profiles and Greedy Feature-Elimination. *LNAI* **3025** (2004) 256–266
12. Eisen, M., Spellman, P., Botstein, D., Brown, P.: Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.* **96** (1999) 14863–14867
13. Alizadeh, A., et al.: Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403** (2000) 503–511
14. Golub, T., et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* **286** (1999) 531–537
15. Alon, U., et al.: Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proc. Natl. Acad. Sci.* **96** (1999) 6745–6750
16. Gupta, S., Rao, S., Bhatnagar, V.: K-means Clustering Algorithm for Categorical Attributes. *LNCS* **1676** (1999) 203–208
17. San, O.M., Huynh, V., Nakamori, Y.: An alternative extension of the k-means algorithm for clustering categorical data. *Int. J. Appl. Math. Comput. Sci.* **14**(2) (2004) 241–247
18. Kanterakis, A., Potamias, G.: Supporting Clinico-Genomic Knowledge Discovery: A Multi-Strategy Data Mining Process. *LNAI* **3955** (2006) 520–524
19. Katehakis, D., Sfakianaki, S., Tsiknakis, M., Orphanoudakis, S.: An Infrastructure for Integrated Electronic Health Record Services: The Role of XML. *Journal of Medical Internet Research* **3**(1) (2001) E7
20. van't Veer, L., et al.: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415** (2002) 530–536

A Prospective Study on the Integration of Microarray Data in HIS/EPR

Daniel F. Polónia, Joel Arrais, and José Luis Oliveira

University of Aveiro, DETI/IEETA, 3810-193 Aveiro, Portugal
{dpolonia,jpa,jlo}@ieeta.pt

Abstract. The successful completion of the Human Genome Project promised an increase on our knowledge about the way our organism works and therefore would have a major impact in medicine. DNA microarray is one of the techniques that appeared in this “-omic” era and that will certainly change the way diagnosis and disease treatment are made. However, despite the successive scientific breakthroughs the integration of microarrays in clinical practice will face yet the lack of proper information systems and communication standards inside the Health Information Systems (HIS) scenarios. We hereby review current information systems for microarrays’ laboratories and for healthcare institutions and also the latest integration efforts, assessing the shortcomings and structural difficulties derived from integrating two distinct fields. We also present the expected difficulties that may arise from the developments in the genetic diagnosis field and its interactions with other diagnostic areas such as imaging and/or radiology. From this prospective analysis we propose a model where the laboratorial microarray data can be integrated with other diagnostic systems in clinical environments, performing structured diagnostic workflows and integrating information from multiple diagnostic sources onto the HIS.

1 Introduction

In the last few years, microarray technology changed not just the way biologists conduct their laboratorial experiments as also the way clinical diagnosis and disease treatment can be made [1-5]. When it comes to healthcare, microarray promises the personalized medicine where individual genomic data is associated with clinical data in order to support the final clinical decision. Indeed there are a great number of genetic diseases that can be detected by the use of microarrays and many efforts are being done to come with breakthroughs in this area [2, 4, 6].

However, microarrays experiments produce large amounts of data that need to be properly stored, filtered and analysed in order to extract meaningful results [7-9]. At the present moment there are several commercial and public domain applications that suit the purpose of managing the microarray laboratorial data (BASE, ArrayExpress, maxD, ..). In addition, there are also standards that specify how the laboratorial data can be stored and exchanged (MIAME, MAGE, ..) [10]. Concerning the healthcare domain, the development of standards and products along the last decades already allow an adequate management of the clinical data (HL7, DICOM,...) [11].

One of the emerging issues is to achieve the integration of the laboratorial and clinical systems in order to allow the use of the microarray results in the electronic

patient record (EPR) [4, 12]. Several prospective studies [13] have already been published on this theme but the main questions are still unsolved, in particular the lack for interoperability between clinical or laboratorial information systems. Thus the improvement of the available data standards, or eventually the development of new ones, as well as the definition of communication models is required.

In this paper, we present the shortcomings of the integration of microarray information into the HIS/EPR and we also propose a communication model that is mainly supported by available standards, both from the clinical and the laboratorial side.

2 Microarray Impact on Medicine

The principle underlying the DNA microarrays technology is based on the ability of a single-strand nucleic acid fragments to hybridize, with high specificity, to a second complementary single strand to generate a double-stranded DNA molecule. Thus, a microarray is obtained by fixing, in a miniaturized solid support and in an ordered way, thousands of probes constituted by known nucleic acid fragments, such as oligonucleotides or bacterial artificial chromosomes (BACS) [7, 14]. The samples (also called targets in opposition to the probes that rely in the microarray surface) are labelled using either radioactivity or, more often, fluorescent dyes and are hybridized to the array surface. During the hybridization phase, according to the Watson-Crick base pairing rules, the single strand probes try to find a target that matches its sequence. All the probes that do not find a correspondent match are removed through a washing process. Then the amount of the complementary target–probe complexes that remain tightly bound are quantified by fluorescent detectors being the most common the laser scanning digital imaging systems.

This way, the microarray constitutes a highly parallel method of quickly and efficiently analyzing thousands of variables in a single sample with just an experiment.

At the present moment several microarray platforms are available and a big discussion still exist on the best platform to use in each situation [3, 5, 14]. The most prominent methods of fabricating microarrays are DNA spotted and photolithography. As Fadiel *et al* [2] evidence, the major advantages of the spotted arrays are that they are widely accessible to be used by any molecular biology laboratory, and, once the DNA has been prepared, the production costs are low. However this technique also has some drawbacks such as the lack of consistency of spotting and reliable annotation of the DNA. On the other side, the photolithography technique, which uses semiconductor industry principles, has high cost per unit but provides good levels of consistency between arrays and high levels of integration. Affymetrix, the most known enterprise that produce photolithography arrays, provide microarrays with the capability to detect 20,000 genes which corresponds to the double of the maximum value obtained in a spotted array.

While microarrays are being widely used as a tool to conduct several studies in research laboratories, their usage on the medical field is still limited by the high prices and by the lack of consistency necessary to enable the mass usage. In spite of that, some authors defend that clinical diagnosis will be able to take advantage of the benefits offered by this technology in less than one decade with the expected technological enhancements [2, 3, 5, 15]. In the medical field, and although several other techniques

are still being studied, a major benefit will come from the usage of microarrays to allow the detection of Gene Expression Profiles and Genotyping [3].

The microarrays ability to measure the expression levels of thousands of genes on any sample enabled the exploration of relevant metabolic pathways and pathogenic mechanisms as well as new indicators of disease prognosis. Recent studies show that microarrays can be used, for instance, in the diagnosis of acute leukaemia. This diagnosis is based on the discovery of the expression profile, constituted by 6817 genes, that characterizes the disease acute leukaemia. In the future, the development of data mining techniques that allow the discovery of the gene expression profiles associated to the diseases will improve not just the capacity to discovery them as also their efficiency.

The global knowledge about the human genome, provided by the human genome project, enabled the discovery of millions of single nucleotide polymorphisms (SNP's) and therefore opens the doors of the predictive medicine [3]. Those SNP's, can be used to identify genetic variation in individuals and across populations and some of them can be associated with the susceptibility to acquire genetically caused diseases. Therefore those SNP's can be used as genetic markers that can be associated with direct predictors of susceptibility of several genetic disorders such as diabetes, hypertension, coronary, heart disease, asthma, inflammatory bowel disease and breast cancer. With the available microarray technology it is possible to obtain microarrays able to detect up to 2000 SNP's. This way, with single blood sample hybridization, it is possible to obtain a genetic fingerprint of an individual and therefore the susceptibility of developing genetically related diseases.

3 Technological Review of Existing Systems

We now proceed with the analysis of the established state of the art concerning the information systems that support healthcare and microarrays, dealing mainly with the way that standards for data annotation are being implemented and how they can be used to easily interchange information.

3.1 Healthcare Technology Review

From a historical perspective, the primary aim of healthcare information systems, was always patient-centred, with its development points taking in consideration medical, management and administrative tasks involved in patient care [16]. Early functionality was limited to applications in departmental information systems (figuratively denominated as information silos), e.g. at a laboratory, radiology or administration unit where information was exclusively used by health professionals inside the department or, in the best cases, inside the institution [17]. Later on, these applications evolved to Hospital Information Systems, with aggregation of information from silos and processing of patient-centred information in health information systems that allowed some form of connection with the outside world and started to be used by patients as well, either as a stakeholder in the process or as a preventive contact. Nowadays there is a holistic perspective in the use of data, not only for patient care, but also for health care planning and clinical research, influencing the fields of medical statistics and epidemiology [18].

This evolution affected the way Health Information Systems are being regarded nowadays, not only as an isolated clinical tool for patient treatment, but as one of the many sources of information available to provide the best possible treatment to the patient, in the form of preventive and/or healing medicine. It is expected that in the near future all patient information will be accessible/integrated, through the networking of all service providers and the wide adoption of standards such as HL7 and DICOM.

Accompanying the increase in the functionalities that are supported by the HIS, also the types of data are being extended, with the extension from the traditional alphanumeric medical notes to the radiological images and the new types of data on the molecular level, such as DNA or protein data [19], [20].

3.2 Microarrays Technology Review

Microarrays are a prominent tool that have already proven its importance in the biological research by providing genome-wide snapshots of transcriptional networks that are active in the cell. By doing so they opened the opportunity for understanding the global systems-level of cellular process.

However, once that microarray experiments deal with thousands of elements a large volume of output raw data is generated, and so, the task of managing all the data in order to allow the extraction of reliable knowledge is a major challenge. To comply with that several efforts were made to develop systems capable of allowing the proper data acquisition, processing, storage and analysis [7-9, 21].

At the present moment a variety of academic and commercial Laboratory Information Management Systems (LIMS) are available for DNA microarrays, namely MIND, BASE, MiamExpress, Maxd, LIMaS, MADAM and CiBEX. Though, with the increased value of the microarray experiments, new data management challenges have appeared. For instance, the possibility to integrate the data that is present in the available stand-alone LIMS in a unique repository can only be achieved with the development of proper standards for data storage and data communication [10, 12, 22].

The MGED group (Microarray Gene Expression Data) recognised this problem and is working to establish standards for microarray data annotation and exchange, facilitating the creation of microarray databases and related software implementing these standards, and promote the sharing of high quality, well annotated data within the life sciences community. The first standard proposed by the MGED group defines the ‘minimum information about a microarray experiment’ (MIAME). This standard is a checklist that specifies all the information needed to interpret, share and possibly replicate the results of a microarray experiment [10]. Once those common principles in data format and data entry have been used, the data consistency needed to store the results in relational databases and to enable the data sharing is assured.

The other standard proposed by the MGED group, MAGE (Microarray Gene Expression), may be considered as an implementation of the concepts described in MIAME. If the goal of MIAME is to describe what to store, the objective of MAGE is to define how to store and share. MAGE has two variations: MAGE-OM (Object Model), that specifies how the data are stored, and MAGE-ML (Mark-up Language), that specifies how data are transmitted [12, 22].

In addition, and once both MIAME and MAGE refer to ontology entries to describe some data fields, the MGED Ontology Working Group was set out. This way, the MGED Ontology provides terms for annotating all aspects of microarray experiments, from experiment design and array layout to preparation of biological samples and protocols for chip hybridization and data analysis. The MGED ontology terms are a useful tool for removing the gaps or ambiguities that may exist in the MIAME or in the MAGE.

4 Integration of Medical Informatics and Bioinformatics

As previously shown, the microarray technology is expected to have a major impact in the way the clinical diagnosis and treatment will be made in a near future. As Shado [23] points out, the possibility to integrate, in the existing patient health records, genomic relevant data will open the doors of the personalized medicine. At the present moment microarray are being widely used for conducting several genomic experiments. However, there is still a debate in the scientific community about the accuracy, reliability and robustness of the analytical and statistical methods that need to be used to extract the final results. Nevertheless, there are already several microarray products that can be used, with a high level of confidence, to carry routine clinical diagnosis. In addition the developments of more clinical oriented microarray platforms, such as point-of-care for microarrays that are able to miniaturize and automate all the process of the microarray experiment, will increase the interest on this technology.

As the technology evolves, and the integration in the clinical environment seems more prominent, a bigger interest raises on how to develop interfaces between medical informatics and the bioinformatics. Some previous works attained the goal of integrating genomic data with patient data. For instance, García-Hernández *et al* [24] proposed an application that manages and analyzes the data that came from microarray relating to the colorectal cancer disease. In this application a single database is used to store the patient data and the experimental data. However, this is an ad-hoc approach that solve a specific problem but do not really deals with HIS/EPR integration as a whole. Therefore this integration will require that the healthcare information system be able to accept the experiment result in the MAGE-ML standard and that the patient information, usually in through the HL7 data standards could be appended to the experiment data.

Recently, as part of the HL7 group, a clinical genomics interest group was created with the mission of developing and supporting standards by enabling the communication between interested parties of the clinical and personalized genomic data. Some of the aims are to analyse life sciences standards such as BSML (Bioinformatics Sequence Markup Language) and MAGE-ML (Microarray and GeneExpression Markup Language) in order to provide enhancements in the HL7. By storing the patient genomic data with clinical relevance those enhancements in the HL7 will enable the communication between the clinical and the laboratorial systems.

But before that goal could be attained it is needed to specify how the clinical genomics proposal can fit in the existent HL7 data standard. One proposed model defends the agreement on a Health Reference Information Model (RIM) [23]. A RIM is a base grammar that can be shared by several elements and that contains semantic and lexical connections that exist between the information carried. Although after 10 years of development work there isn't any stable version, in theory, the agreement on

a Health RIM, will boost the usability of all specifications derived from that RIM - such as clinical genomics - and therefore enabling the desired goal of semantic interoperability [25].

5 Proposed Model

Considering that we are now evolving from the integration of medical images, namely in the radiological domains, in addition to alphanumeric data, to the existence of new types of data on the molecular level, such as genomic and proteomic data [16], we propose the use of the “Integrating the Healthcare Enterprise (IHE)” for the radiological subset and its expansion to the microarray subset through an analogous applications (Fig. 1).

The IHE integrates the Hospital Information Systems, the digital modalities and the Picture Archiving and Communication Systems (PACS) of the radiological departments, through a comprehensive specification called the Technical Framework [26] as a roadmap for their integration (darker shade of Fig. 1).

When integrating systems with different standards, such as DICOM and HL7 in the case of radiology, overlaps and gaps occurred, however an information mapping was attained, data elements, use cases and scenarios were mapped and its impacts on the transactions and model were identified [27].

The IHE integration profiles define the compatibility between devices since they specify the services, the transactions and the contents of the information exchanged between the systems, solving the overlaps between the standards.

Although the IHE was initially concentrated in the development of specifications for Radiology and Cardiology, it has, in the recent past, began to expand its specification to Laboratory Technical Framework [28], which describes the integration of the clinical laboratory in the healthcare enterprise for the performance of tests on specimens usually collected from the patient.

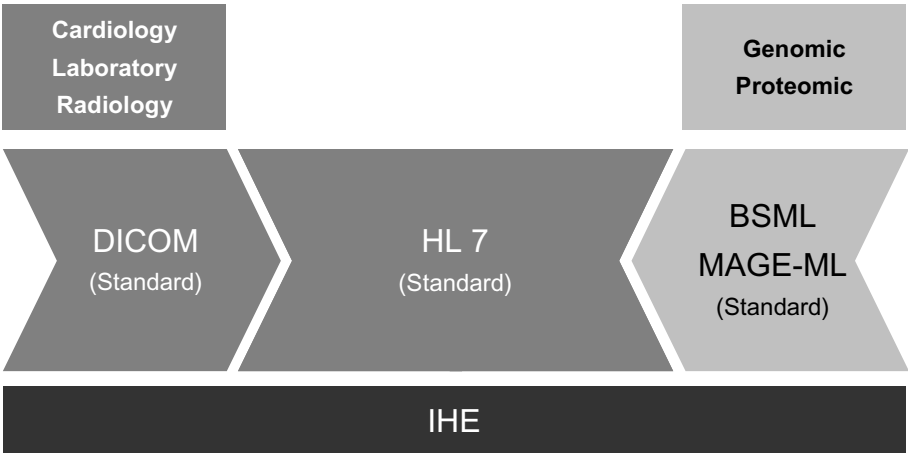


Fig. 1. Standards coexistence in the “Integrating the Healthcare Enterprise” model. Existing one in darker shade and proposed one in lighter shade.

The tests produce observation results which can be of various natures: from simple numeric quantitative measurement such as a blood serum glucose level, to a complex diagnostic pathology report such as a bone marrow biopsy. Some of these results may carry images or graphs, for example blood serum protein electrophoresis. Results are sent to the ordering clinical department; copies may be sent to other physicians or departments, and may also be stored in an electronic healthcare record.

However, not all laboratory specialties are covered in the current framework, and the current IHE development cycle includes the Microbiology discipline, but not specifically the microarrays, supporting the following workflow cases [29]:

- Externally placed order with identified specimens
- Externally placed order with unidentified specimens
- Filler order

The transactions are supported by HL7 and when an application wants to send a message (initiate a transaction), it initiates a network connection.

After establishing the connection, a set of common message segments can be used by the transactions, with some optional fields that encapsulate the laboratorial results and comments into the OBR – Observation Request Segment and the OBX – Observation / Result Segment. Where information is not appropriate into these segments, for example, free text reports, it can be placed in the NTE – Notes and Comment segment of the message [29].

When applying the same concepts underlying Radiological and Laboratorial IHE to microarrays (lighter side of Fig. 1) we question to which extent it would be possible to replicate the workflows and the message exchange/encapsulation principles in order to enable a successful integration between the existing laboratorial information systems and the health/hospital information systems where the microarray exams would be requested.

If we assume that the workflow leading to the execution of a microarray clinical analysis is the same as the one that leads to a blood glucose level analysis or a bone marrow biopsy, for example, including the patient demographic information it is feasible to make an analogy between the IHE Radiology and Laboratory Technical Framework, so that we can reach an integration model for microarrays. A general model is sketched in Figure. 2. On the centre we have the Health Information System, with its applications based on HL7 data standards, and on the right hand side, a Microarray Laboratory Information Management System (LIMS), using in between an adapter for HL7, so that we are able to pass demographic and scheduling information to the LIMS, requiring the execution of a given microarray exam. The information is received by the LIMS that sets up the necessary steps to execute the ordered exam.



Fig. 2. Base modules of a microarrays/HIS/EPR integration solution

Once the exam is executed, the results are stored in the LIMS database and the information stored in MAGE-ML format so that it can be passed to the Health Information System, passing through the adaptation layer placed between them and encapsulating all the relevant information in the OBR – Observation Request Segment or the OBX – Observation / Result Segment, or, as a last resource, in the NTE – Notes and Comment segment of the message.

6 Conclusions

Although one can expect that, in the near future, clinical medicine will benefit from “-omics” tools such as microarrays to complement diagnosis, very little is yet available concerning this subject. Microarrays are still information systems mainly laboratory oriented rather than clinical oriented not being able to provide useful, real-time, clinical information to physicians.

The presented discussion and our prospective model constitute an initial attempt to address the problem of the integration of microarrays results in Health Information Systems. At a broader level, other factors will need to be taken to account. Specific workflow functional requirement for microarray diagnosis will have to be developed and the traditional diagnosis workflows will need to be readdressed.

Acknowledgements

The present work has been funded by the European Commission (FP6, IST thematic area) through the INFOBIOMED NoE (IST-507585).

References

- [1] F. Bertucci, P. Viens, and D. Birnbaum, "[DNA microarrays for gene expression profiling of breast cancer: principles and prognostic applications]," *Pathol Biol (Paris)*, vol. 54, pp. 49-54, 2006.
- [2] A. Fadiel and F. Naftolin, "Microarray applications and challenges: a vast array of possibilities," *Int Arch Biosci*, pp. 1111-1121, 2003.
- [3] A. Kumar, G. Goel, E. Fehrenbach, A. K. Puniya, and S. K., "Microarrays: The Technology, Analysis and Application," *Microarray* vol. 5, pp. 215-222, 2005.
- [4] W. P. Kuo, "Overview of bioinformatics and its application to oral genomics," *Adv Dent Res*, vol. 17, pp. 89-94, 2003.
- [5] N. L. van Berkum and F. C. Holstege, "DNA microarrays: raising the profile," *Curr Opin Biotechnol*, vol. 12, pp. 48-52, 2001.
- [6] K. R. Hess, W. Zhang, K. A. Baggerly, D. N. Stivers, and K. R. Coombes, "Microarrays: handling the deluge of data and extracting reliable information," *Trends Biotechnol*, vol. 19, pp. 463-8, 2001.
- [7] J. Arrais, G. L. Campos, L. Carreto, J. L. Oliveira, and M. A. S. Santos, "Microarray data: from the hybridisation to the analysis," presented at "Best of" European Summer School in Biomedical Informatics Balatonfüred, Hungary, 2006.

- [8] J. Arrais, L. Silva, M. Rodrigues, L. Carreto, J. L. Oliveira, and M. A. S. Santos, "Why another microarray LIMS," presented at EMBEC 2005, Prage, Czech Republic, 2005.
- [9] P. J. Killion, G. Sherlock, and V. R. Iyer, "The Longhorn Array Database (LAD): an open-source, MIAME compliant implementation of the Stanford Microarray Database (SMD)," *BMC Bioinformatics*, vol. 4, pp. 32, 2003.
- [10] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, T. Gaasterland, P. Glenisson, F. C. P. Holstege, I. F. Kim, V. Markowitz, J. C. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron, "Minimum information about a microarray experiment (MIAME) – toward standard for microarray data," *Nat. Genet.*, pp. 29(4) 365-71, 2001.
- [11] B. Vargas and R. Pradeep, "Interoperability of hospital information systems: a case study," 2003.
- [12] J. Arrais, J. L. Oliveira, G. Grimes, S. Moodie, K. Robertson, and P. Ghazal, "Microarray data sharing in BioMedicine," presented at MIE 2006, Maastricht, Netherlands, 2006.
- [13] A. Shabo, "The implications of electronic health records for personalized medicine," *Personalized Medicine*, vol. 2, pp. 251-258, 2005.
- [14] G. Hardiman, "Microarray platforms--comparisons and contrasts," *Pharmacogenomics*, vol. 5, pp. 487-502, 2004.
- [15] P. Jares, "DNA microarray applications in functional genomics," *Ultrastruct Pathol*, vol. 30, pp. 209-19, 2006.
- [16] Haux, "Health information systems - past, present, future," *International Journal of Medical Informatics*, vol. 75, pp. 268-281, 2006.
- [17] M. F. Collen, "General requirements for a Medical Information System (MIS)," *Computers and Biomedical Research*, vol. 3, pp. 393-406, 1970.
- [18] M. Berg, J. Aarts, and J. van der Lei, "ICT in health care: sociotechnical approaches," *Methods Of Information In Medicine*, vol. 42, pp. 297-301, 2003.
- [19] C. A. Kulikowski, "The micro-macro spectrum of medical informatics challenges: from molecular medicine to transforming health care in a globalizing society," *Methods Of Information In Medicine*, vol. 41, pp. 20-24, 2002.
- [20] V. Maojo and F. Martin-Sanchez, "Bioinformatics: towards new directions for public health," *Methods Of Information In Medicine*, vol. 43, pp. 208-214, 2004.
- [21] R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon, "EXPANDER--an integrative program suite for microarray data analysis," *BMC Bioinformatics*, vol. 6, pp. 232, 2005.
- [22] U. Sarkans, "Standardisation of Microarray Data," *Pharmatech*, 2003.
- [23] A. Shado, "The implications of electronic health records for personalized medicine," *Personalized medicine*, vol. 2, pp. 251-258, 2005.
- [24] O. Garcia-Hernandez, G. Lopez-Campos, J. P. Sanchez, R. Blanco, A. Romera-Lopez, B. Perez-Villamil, and F. Martin-Sanchez, "Microarray data analysis and management in colorectal cancer," *BIOLOGICAL AND MEDICAL DATA ANALYSIS, PROCEEDINGS*, vol. 3745, pp. 391-400, 2005.
- [25] B. Smith and W. Ceusters, "HL7 RIM: An Incoherent Standard," presented at MIE 2006, Maastricht, Netherlands, 2006.
- [26] http://www.ihe.net/Technical_Framework/index.cfm.
- [27] M. Henderson, *HL7 Messaging: OTech*, 2003.
- [28] http://www.ihe.net/Technical_Framework/index.cfm#laboratory.
- [29] IHE, "IHE Laboratory Technical Framework," 2006.

Web Services Interface to Run Protein Sequence Tools on Grid, Testcase of Protein Sequence Alignment

Christophe Blanchet, Christophe Combet, Vladimir Daric, and Gilbert Deléage

Institut de Biologie et Chimie des Protéines (IBCP UMR 5086); CNRS; Univ. Lyon 1;
IFR128 BioSciences Lyon-Gerland; 7, passage du Vercors, 69007 Lyon, France
{Christophe.Blanchet,C.Combet,V.Daric, G.Deleage}@ibcp.fr

Abstract. Bioinformatics analysis of data produced by high-throughput biology, for instance genome projects, is one of the major challenges for the next years. Some of the requirements of this analysis are to access up-to-date databanks (of sequences, patterns, 3D structures, etc.) and relevant algorithms (for sequence similarity, multiple alignment, pattern scanning, etc.). GPS@ is a Web portal devoted to bioinformatics applications on the grid (Grid Protein Sequence Analysis, <http://gpsa-pbil.ibcp.fr>). GPS@ is the grid release of the NPS@ bioinformatics portal, and is wrapping the mechanisms required for submitting bioinformatics analyses on the grid infrastructure. For example, we have put online two multiple alignment Web Services that are submitting the computing job on a remote grid environment. One is accessible through a classical Web interface by using a simple Web browser; the other one can be used through a SOAP and workflow client such as Taverna or Triana. These Web services can process the submitted alignment on two different computing environments: a local and classical one which is a cluster of 30 CPUs, but we are also providing biologists with a large-scale distributed one: the grid platform of the EU-EGEE project (more than 20,000 CPUs available at the European scale).

Keywords: Bioinformatics, Grid computing, Web Services, Protein Sequence Analysis.

1 Introduction

Bioinformatics analysis of data produced by high-throughput biology, for instance genome projects [1], is one of the major challenges for the next years. Some of the requirements of this analysis are to access up-to-date databanks (of sequences, patterns, 3D structures, etc.) and relevant algorithms (for sequence similarity, multiple alignment, pattern scanning, etc.) [2]. There are more and more tools that are put online for molecular biology [3]. But most of these portal are proposing isolated bioinformatics methods, some times several analysis methods, but only few of these Web servers are proposing the integration of several program devoted to molecular Biology. Since 1998, we are developing the Network Protein Sequence Analysis (NPS@) Web server [4], that provides the biologist with many of the most common resources for protein sequence analysis, integrated into several pre-defined and connected workflows.

1.1 Related Work: Multiple Sequence Alignment Resources Available Online

Starting in 2003, the Bioinformatics Links Directory [3], [5] is referencing all the online resources, tools and databases devoted to molecular Bioinformatics. This list is curated according to expert recommendations. More than two thousands of links are listed on this molecular bioinformatics Web repository. Among all of them, almost two hundreds are classified as “Sequence comparison” resources, and among these ones, only 39 are furnishing a service for multiple sequence alignments. But all of them are computing the user queries on classical computing resources, local machine or batch cluster, according to their description available on the Bioinformatics Link Directory.

1.2 NPS@, Bioinformatics Web Portal

NPS@ [4] is providing biologist with a Web form to input their data (like protein sequences) in order to run, for example, a BLAST similarity scan against a given protein sequence database, or a multiple alignment of a subset of sequences. In NPS@, user inputs his protein sequences by pasting them in the corresponding field. Then, in case of BLAST, he chooses the database that will be scan with the query sequence. All the protein databases available on NPS@ can be selected through a multi-valued list of the form. These methods and data can be accessed through a classical web browsing and HTTP connection, or through a specialized interface like MPSA [6] or AntheProt [7] programs.

Today, the computing resources available behind the NPS@ Web portal may limit the capabilities put available to the research community. And it is the case also for other genomics and proteomics Web portals. Indeed some methods are very computing-time and memory consuming. Our NPS@ portal is facing an increasing demand of CPU and disk resources and the management of numerous bioinformatics resources (algorithms, databases).

1.3 Testcase: Build Hepatitis C Virus Sequence Alignment

Hepatitis C virus (HCV) causes chronic liver disease in humans, including cirrhosis and hepatocellular carcinoma. The HCV genome shows remarkable sequence variation. Analysis of this variability is essential not only to investigate the correlation between HCV molecular components and diseases expression or antiviral resistance, but also to study the structure-function relationships of these components. To date, more than 40000 HCV sequences are deposited in the generalist databases DDBJ, EMBL, and Genbank [8].

In this testcase, we will consider a common task for bioinformaticians working on Hepatitis C Virus: doing a multiple alignment of sequences issued from different strains. User will upload its own sequence databank (in Pearson/FASTA format) or will extract the sequences from annotated HCV sequence database using a retrieval system (SRS, SQL through a web site). The upload or the query will be done thanks to an integrated Web interface. From the subset of sequences, the user will launch a multiple sequence alignment tools, for example ClustalW [9] or Muscle [10]. A set of tools should be proposed to the user to analyze the computed alignment.

2 Distributed Computing: Web Services and Grid

2.1 Web Services, Weak Connected Concept of Distributed Computing

Web Services (WS) is describing programmatic interfaces that allow different resources, different by location or implementation, to collaborate in a distributed environment base on the Web. Web Services are most of time using three components: WSDL, SOAP and HTTP. Users that will use Web Services will do it through a SOAP client able to create SOAP messages, and able of understanding WSDL file in order to import available WS processors from a remote site.

WSDL. The Web Service Description Language (WSDL) is a language based on XML and aiming to describe Web Services. A WSDL file should define both the resources available in the Web Service, their interfaces and their location (Appendix A). A WSDL file is a directory of several Web Services that could be completely independent and located in different Web places.

SOAP. The Simple Object Access Protocol (SOAP) is a framework that allows describing objects and that they talk together in a distributed and weakly connected Web environment.

HTTP. The Hypertext Transfer Protocol is defining the language and the rules used to exchange messages from clients and servers that are connected to the World-Wide Web (W3). Client are contacting servers, that answer with the same protocol based on messages mainly composed of two parts, a header and a body, containing different command and valued defined with tags.

SOAP client. Taverna [11] and Triana [12] are both combining capabilities of workflow editor, workflow enactor and Web services client, compliant with SOAP and other WS protocols. They provide a visual interface to build simple and complex workflows, providing simple way to link processors available locally or remotely, to datasets or large range of tests. They are both using their own workflow description language, but both based on XML. They are written in Java, and so able to run on most operating systems.

2.2 Grid Computing, Integrated Concept of Distributed Computing

Grid computing concept defines a set of information resources (computers, databases, networks, instruments, etc.) that are integrated to provide users with tools and applications that treat those resources as components within a « virtual » system [13][14][15]. Grid middleware provides the underlying mechanisms necessary to create such systems, including authentication and authorization, resource discovery, network connections, and other kind of components.

The Enabling Grids for E-sciencE project (EGEE [16]), funded by the European Commission, aims to build on recent advances in grid technology and to develop a service grid infrastructure. The EGEE consortium involves 70 leading institutions in 27 countries, federated in regional Grids, with currently a combined capacity of 20,000 CPUs and 5 petabytes of storage. The platform is built on the LCG-2 middleware, inherited from the EDG middleware developed by the European

DataGrid Project [17] (EDG, FP5 2001-2003). The middleware LCG-2 is based upon the Globus toolkit release 2 (GT2) and the Condor middleware [14]. The new middleware gLite [16], that is being developed, have the goals to improve the performances and the services provided by the future EGEE platform.

There are several important components into the EGEE grid: first on the user point of view is the user interface (UI) where the user log in and submit their jobs. These jobs need to be described by JDL files (Job Description Language) with the Condor “ClassAd” formalism. The “workload management system” (WMS) is responsible of the job scheduling on the platform. The scheduler (or “resource broker”, RB) analyzes the JDL file and determines where and when to compute a job: (i) using one “computing element” (CE) near one “storage element” (SE) containing the data in case of simple jobs, or (ii) several CEs and SEs in case of larger jobs. A computing element is a gatekeeper to a cluster of several CPUs, the worker nodes (WN) managed by a batch scheduler system. The “information system” that centralize all parameters raised by the grid components (CPUs, storage, network, ...).

3 Web Resources for Protein Sequence Analysis on the Grid

3.1 GPS@, Web Server for Grid Protein Sequence Analysis

GPS@ grid Web portal (Grid Protein Sequence Analysis, <http://gpsa-pbil.ibcp.fr>) is the grid release of the NPS@ bioinformatics portal. GPS@ portal hides the required mechanisms for submitting bioinformatics analyses on the grid infrastructure. Selecting the “EGEE” check-box will schedule the submission of the ClustalW on the EGEE grid when clicking on the “submit” button. The bioinformatics programs and databases available on GPS@ have been distributed and registered on the grid [18], and GPS@ runs its own EGEE interface to the grid [19].

3.2 gBIO-WS, Interfacing WSDL-Compliant Web Services with the GRID

The Grid capability is added to our Web services by the use of the *bio_launcher* tool (Figure 1). We have developed this *bio_launcher* tool to be able to submit remotely a job to the EGEE Grid.

Indeed, in the normal job submission process on EGEE, user has to connect on a special host of the grid: the user interface. Then, he authenticates itself by creating a proxy certificate signed by his own, valid and recognized by EGEE, electronic certificate. Afterwards, he submits and manages his job by using the appropriate command line interface. These are the commands *edg-job-submit*, *edg-job-status*, *edg-job-get-output*, ...

Bio_launcher gets, as input, an XML file describing the bioinformatics task to compute: which program to use (ClustalW in this case), the data to process, the user values of program options. This XML description file is written by the Web service according to our gBIO DTD, and filled-in with the data provided by the user. Finally, *bio_launcher* connects on our EGEE user interface through a secure connection (SSH), opens an authenticated connection to the EGEE Grid, submits the job on the Grid, and once the job is finished, gets the results back to the Web service, which forwards them to the user.

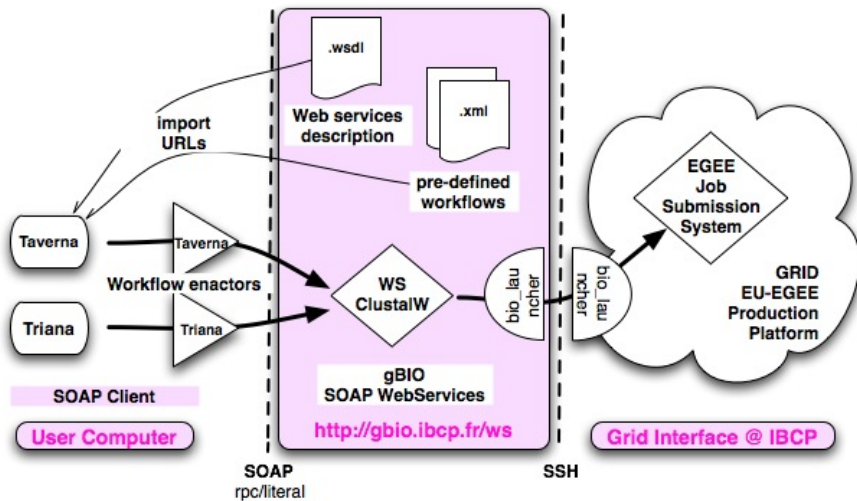


Fig. 1. Architecture of the Bioinformatics Web services at gBIO-WS server, interfaced to the GRID

4 User-Friendly Access to Multiple Alignment Web Services on Grid

We have put online two multiple alignments Web Services on the CNRS IBCP servers. One is accessible through a classical Web interface, the other one can be used through a SOAP client such as Taverna or Triana, but also a user one built with gSOAP, perl SOAP::Lite or Java.

These Web services can process the submitted alignment on two different computing environments: a local and classical one which is a cluster of 30 CPUs, but we are also providing biologist with an original distributed one: the grid platform of the EU-EGEE project (more than 20,000 CPUs available at the European scale)

4.1 Via a Web Browser

Biologist can access this grid service of multiple sequence alignment through a classical Web page on our Grid Protein Sequence Analysis server (<http://gpsa-pbil.ibcp.fr>). The protein sequences of HCV can be pasted from the euHCVdb server to the submission form on the GPS@ Web portal (Figure 2). There, the user can chose to process the alignment on our cluster or on the Grid. When the “Grid” checkbox is checked, the multiple alignment is then processed on the EGEE grid platform.

First, the job description in the Web form is converted to a JDL file that can then be submitted to the workload management system of EGEE. The GPS@ sub-process that have submitted the job, is also checking periodically the status of this job by querying the resource broker with the good commands. All steps are notified to the user through the Web page of the submission, indicating the time and the duration of the current step. When achieved, i.e. reaching the “Done” step, the GPS@ automat

downloads the result file containing the multiple alignment computed by ClustalW. Then this raw result file in ClustalW format is processed and converted into a HTML page showing, in a colored and graphical way, the list of aligned protein sequences, (Figure 2). This formatting process is directly inherited from the original NPS@ portal, providing biologists with a well-known interface and way of displaying results. After biologist has analyzed the alignment thanks the graphical display in colored shape, he can submit new queries to obtain a modified and better alignment.

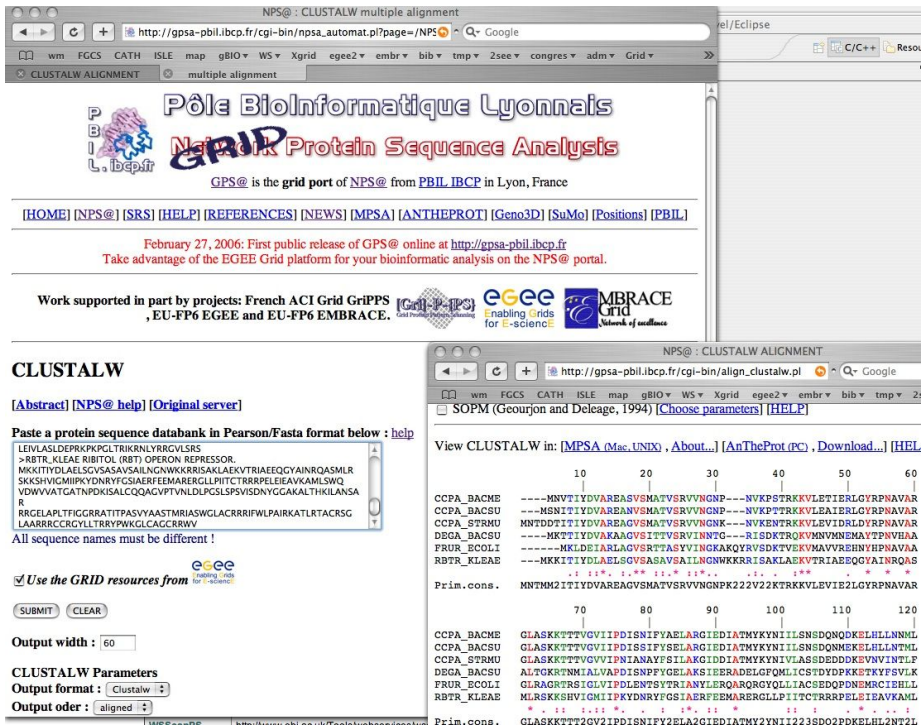


Fig. 2. Multiple alignment of protein sequences obtained through a submission form on the GPS@ Web portal and processed on the EGEE grid platform

4.2 Via a SOAP Client and Workflow Enactor

Biologists and Bioinformaticians can also access this grid service of multiple sequence alignment through Web Services on our Grid Bioinformatics server (<http://gbio.ibcp.fr/ws>). Our Web services are using standard protocols (SOAP, WSDL and HTTP) and have been built with the gSOAP toolkit, and hosted on Apache HTTP server.

Before to be able to use it, user needs obviously to get the SOAP client which may be Taverna [11] or Triana [12]. We have tested only with these two ones, but other client compatible with WSDL and SOAP standard will certainly be able to connect to our Web services.

User then needs to import our Web service within the Taverna tool (Figure 3). He can do it by importing the WSDL file available at <http://gbio.ibcp.fr/ws/gBIO.wsdl> (Figure 1). He also has to build a workflow with the HCV sequences as input of the Web service processor, and the multiple alignment as the output downloaded after computing. We are providing two pre-defined workflows to submit an alignment query on our Web service. These pre-defined workflows can be also imported in Taverna directly from the following URLs (Figure 1): <http://gbio.ibcp.fr/wf/Clustalw.xml> (to import a workflow that process a multiple alignment on our own computing resources), or <http://gbio.ibcp.fr/wf/ClustalwGrid.xml> (to import a workflow that submit the sequence set on the EGEE grid resources).

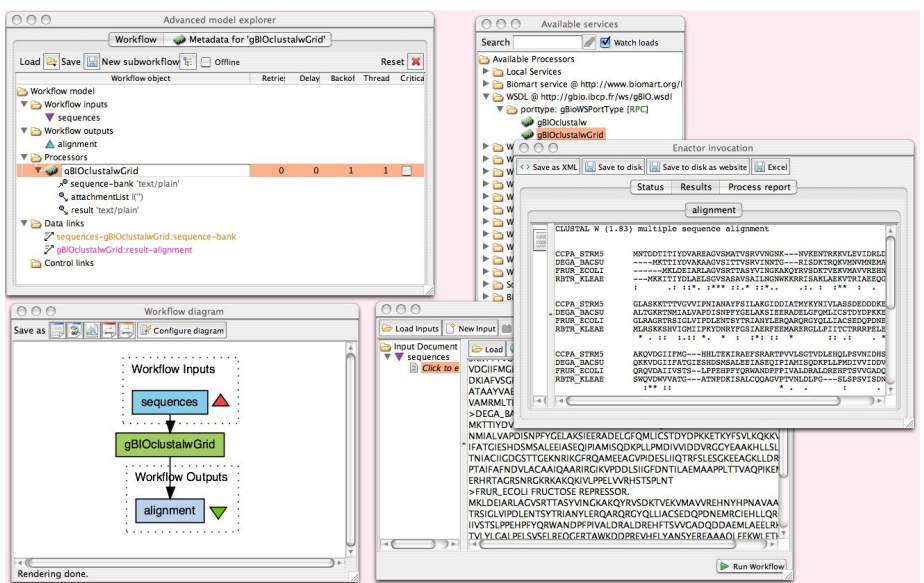


Fig. 3. Multiple alignment of protein sequences obtained through a workflow launched in the Taverna tool, submitted to the IBCP Web Services and processed on the EGEE grid platform.

The protein sequences of HCV can be pasted from the euHCVdb server to the submission field on the Taverna tool (Figure 3). There, the user can choose to process the alignment on our cluster or on the Grid. When the “Grid” processor is used, i.e. the “Grid” workflow ClustalwGrid.xml has been imported, the multiple alignment is then processed on the EGEE grid platform. Afterwards, Biologist analyzes the alignment, and submits new queries to obtain a modified and better alignment.

5 Conclusion

GPS@ Web portal and gBIO-WS make the remote access and bioinformatics job submission easier on the grid. We have used, as testcase, the ClustalW multiple alignment tool run on a remote Grid platform, to analyze the variability of a subset of

sequences. The GPS@ portal and gBIO Web services are compliant with standard protocol, guaranty of a good access with common Web browsers and SOAP clients. Biologists can then submit bioinformatics jobs on the Grid by using their usual Web client, but also integrate these grid services within complex workflow combining different databases and tools. They will then benefit from the large-scale computing resources of the Grid, from their usual and local working environment. Grid computing and storage facilities will also permit GPS@ and gBIO services to scale to thousands of daily user as much as aligning complete genomes or proteomes.

Future works will be done about applying this WebServices-to-Grid interface to other programs. We will, for example, work to put online, as Web Services, a selected panel of other protein alignment methods, but also similarity searching programs, like BLAST or SSEARCH, raising the issues of large and numerous databases management in Grid environment [20].

Acknowledgments. This work was supported in part by projects: French ACI Grid GriPPS (ACI GRID PPL02-05), European projects EGEE (EU FP6, INFSO-508833) and EMBRACE (EU FP6, LHSG-CT-2004-512092).

References

1. Bernal, A., Ear, U., Kyrpides, N. : Genomes OnLine Database (GOLD): a monitor of genome projects world-wide. *NAR* 29 (2001) 126-127
2. G. Perrière, C. Combet, S. Penel, C. Blanchet, J. Thioulouse, C. Geourjon, J. Grassot, C. Charavay, M. Gouy, L. Duret and G. Deléage, Integrated databanks access and sequence/structure analysis services at the PBIL. *Nucleic Acids Res.*, 31:3393-3399, 2003.
3. Fox JA, McMillan S, Ouellette BF. A compilation of molecular biology web servers: 2006 update on the Bioinformatics Links Directory. *Nucleic Acids Res* 34(Web Server Issue) W3-5. (2006).
4. Combet, C., Blanchet, C., Geourjon, C. et Deléage, G. : NPS@: Network Protein Sequence Analysis. *Tibs*, 25 (2000) 147-150.
5. Bioinformatics Links Directory. Online at bioinformatics.ubc.ca/resources/links_directory
6. Blanchet, C., Combet, C., Geourjon, C. et Deléage, G. : MPSA: Integrated System for Multiple Protein Sequence Analysis with client/server capabilities. *Bioinformatics*, 16 (2000) 286-287.
7. Deleage, G, Combet, C, Blanchet, C, Geourjon, C. : ANTHEPROT: an integrated protein sequence analysis software with client/server capabilities. *Comput Biol Med.*, 31 (2001) 259-267
8. Combet C., Penin F., Geourjon C. and Deleage G. HCVDB: Hepatitis C Virus Sequences Database. *Appl. Bioinformatics*, 2004, 3(4):237-240
9. Thompson, JD, Higgins, DG, Gibson, TJ : CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22 (1994) 4673-4680.
10. Robert C Edgar . MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004, 5:113
11. D. Hull, K. Wolstencroft, R. Stevens, C. Goble, M. R. Pocock, P. Li, and T. Oinn. Taverna: a tool for building and running workflows of services. *Nucleic Acids Res.*, July 1, 2006; 34(suppl_2): W729 - W732.

12. I. Taylor, M. Shields, I. Wang, and A. Harrison. Visual Grid Workflow in Triana. In *Journal of Grid Computing*, 3(3-4):153-169, September 2005.
13. Foster, I. And Kesselman, C. (eds.) : *The Grid 2 : Blueprint for a New Computing Infrastructure*, (2004).
14. Thain, D., Tannenbaum, T. Livny, M.: Distributed computing in practice: the Condor experience. *Concurrency and Computation* 17 (2005) 323-356.
15. Vicat-Blanc Primet, P., d'Anfray, P., Blanchet, C., Chanussot, F. : e-Toile : High Performance Grid Middleware. *Proceedings of Cluster'2003* (2003).
16. Enabling Grid for E-science (EGEE). Online at www.eu-eggee.org
17. European DataGrid project (EDG). Online at www.eu-datagrid.org
18. Blanchet, C., Combet, C. and Deléage, G., Integrating Bioinformatics Resources on the EGEE Grid Platform. *ccgrid*, p. 48, Sixth IEEE International Symposium on Cluster Computing and the Grid Workshops (CCGRIDW'06), 2006.
19. Blanchet, C., Lefort, V., Combet, C., Deléage, G., GPS@ Bioinformatics Portal: from Network to EGEE Grid. *Stud Health Technol Inform.* 2006;120:187-93.
20. Desprez, F., Vernois, A., Blanchet, C., Simultaneous Scheduling of Replication and Computation for Bioinformatic Applications on the Grid. *ISBMDA 2005*: 262-273

Appendix A: gBIO WebServices Description (WSDL)

Description of Web Services for Multiple Sequence Alignment, available on the gBIO-WS server (<http://gbio.ibcp.fr/ws>), written according to WSDL standard.

```
<?xml version="1.0" encoding="UTF-8"?>
<definitions name="gBioWS"
  targetNamespace="http://gbio.ibcp.fr:8090/gBioWS.wsdl"
  xmlns:tns="http://gbio.ibcp.fr:8090/gBioWS.wsdl"
  xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
  xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:ns="urn:gBioWS"
  xmlns:SOAP="http://schemas.xmlsoap.org/wsdl/soap/"
  xmlns:MIME="http://schemas.xmlsoap.org/wsdl/mime/"
  xmlns:DIME="http://schemas.xmlsoap.org/ws/2002/04/dime/wsdl/"
  xmlns:WSDL="http://schemas.xmlsoap.org/wsdl/"
  xmlns="http://schemas.xmlsoap.org/wsdl/">

<types>

<schema targetNamespace="urn:gBioWS"
  xmlns:SOAP-ENV="http://schemas.xmlsoap.org/soap/envelope/"
  xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema"
  xmlns:ns="urn:gBioWS"
  xmlns="http://www.w3.org/2001/XMLSchema"
  elementFormDefault="unqualified"
  attributeFormDefault="unqualified">
  <import namespace="http://schemas.xmlsoap.org/soap/encoding/" />
</schema>

</types>

<message name="gBIOclusterRequest">
  <part name="sequence-bank" type="xsd:string"/>
</message>
```

```

<message name="gBIOclustalwResponse">
  <part name="result" type="xsd:string"/>
</message>

<message name="gBIOclustalwGridRequest">
  <part name="sequence-bank" type="xsd:string"/>
</message>

<message name="gBIOclustalwGridResponse">
  <part name="result" type="xsd:string"/>
</message>

<portType name="gBioWSPortType">
  <operation name="gBIOclustalw">
    <documentation>Service definition of function
ns__gBIOclustalw</documentation>
    <input message="tns:gBIOclustalwRequest"/>
    <output message="tns:gBIOclustalwResponse"/>
  </operation>
  <operation name="gBIOclustalwGrid">
    <documentation>Service definition of function
ns__gBIOclustalwGrid</documentation>
    <input message="tns:gBIOclustalwGridRequest"/>
    <output message="tns:gBIOclustalwGridResponse"/>
  </operation>
</portType>

<binding name="gBioWS" type="tns:gBioWSPortType">
  <SOAP:binding style="rpc"
transport="http://schemas.xmlsoap.org/soap/http"/>
  <operation name="gBIOclustalw">
    <SOAP:operation style="rpc" soapAction="" />
    <input>
      <SOAP:body use="encoded" namespace="urn:gBioWS"
encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" />
    </input>
    <output>
      <SOAP:body use="encoded" namespace="urn:gBioWS"
encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" />
    </output>
  </operation>
  <operation name="gBIOclustalwGrid">
    <SOAP:operation style="rpc" soapAction="" />
    <input>
      <SOAP:body use="encoded" namespace="urn:gBioWS"
encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" />
    </input>
    <output>
      <SOAP:body use="encoded" namespace="urn:gBioWS"
encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" />
    </output>
  </operation>
</binding>
<service name="gBioWS">
  <documentation>gSOAP 2.7.8c generated service
definition</documentation>
  <port name="gBioWS" binding="tns:gBioWS">
    <SOAP:address location="http://gbio.ibcp.fr:8090"/>
  </port>
</service>
</definitions>

```


Integrating Clinical and Genomic Information Through the ProgenoChip Mediator

Anastasia Analyti¹, Haridimos Kondylakis¹, Dimitris Manakanatas¹,
Manos Kalaitzakis¹, Dimitris Plexousakis¹, and George Potamias²

¹ Information Systems Lab, Institute of Computer Science, FORTH-ICS, Greece

² Biomedical Informatics Lab, Institute of Computer Science, FORTH-ICS, Greece
{analyti,kondylak,manakan,mkalaitz,dp,potamias}@ics.forth.gr

Abstract. The ultimate goal of the biomedical informatics project ProgenoChip is the identification of classification and prognosis molecular markers for breast cancer. This requires not only an understanding of the genetic basis of the disease, based on the patient's tumor gene expression profiles but also the correlation of this data with knowledge normally processed in the clinical setting. In this paper, we present the Mediator component of the ProgenoChip Integrated Clinico-Genomics Environment (ICGE), through which the integration of the clinical and genomic information subsystems is achieved. The biomedical investigator can form clinico-genomic queries through the web-based graphical user interface of the Mediator. This is split into several query forms, allowing cancerous sample selection (along with their associated gene expression profiles and patient characteristics), based on criteria of interest. After a query is formed, the Mediator translates it into an equivalent set of local subqueries, which are executed directly against the constituent databases. Then, results are combined for presentation to the user and/or transmission to the Data Mining tools for analysis.

1 Introduction

ProgenoChip is an ongoing project that aims at the identification of molecular markers for the classification and prognosis of breast cancer, based on the correlation of patients' clinico-histopathological parameters and therapy response with their tumor gene expression profiles [12]. To achieve this, ProgenoChip joins forces from different scientific disciplines: Molecular Biology (FORTH-IMBB), Medicine (Univ. General Hospital of Heraklion - PAGNH, and PROLIPSIS, a breast cancer center), and Computer Science (FORTH-ICS). The main medical and molecular biology tasks within ProgenoChip are:

Medicine/ Tissue collection & Histopathology: surgical specimens are collected from breast cancer patients that undergo any surgical type of treatment. As soon as the specimen is removed from the patient, it is carried immediately to the histopathology department, where sections are taken from the growing edge of the tumor for (a) histopathological and immunohistochemistry analysis, and (b) identification of their gene expression profiles through DNA Microarray technology.

Molecular Biology/ Microarrays [1,10]: A DNA microarray of long oligonucleotide probes has been designed, representing all known human genes, approximately 35,000 different reporters (oligonucleotides) of 27,000 different genes. Oligonucleotide probes are spotted on four activated glass slides (arrays). A common “reference” material has been decided for the study, consisting from a defined set of cell-line extracts, ensuring accurate quantitation of gene expression for most of the genes. Additionally, RNA extraction, amplification, and fluorescent labeling protocols have been developed, allowing the analysis of small samples. After array hybridization, fluorescence intensity images are acquired. From these images, fluorescence intensities (raw hybridization data) are obtained, using dedicated image analysis software. Raw hybridization data are analyzed to generate *Gene Expression* data, expressing through a *Ratio Value* (per spotted reporter/gene), if the gene in the cancerous tissue is over-expressed, under-expressed, or equally expressed with respect to the “reference” tissue.

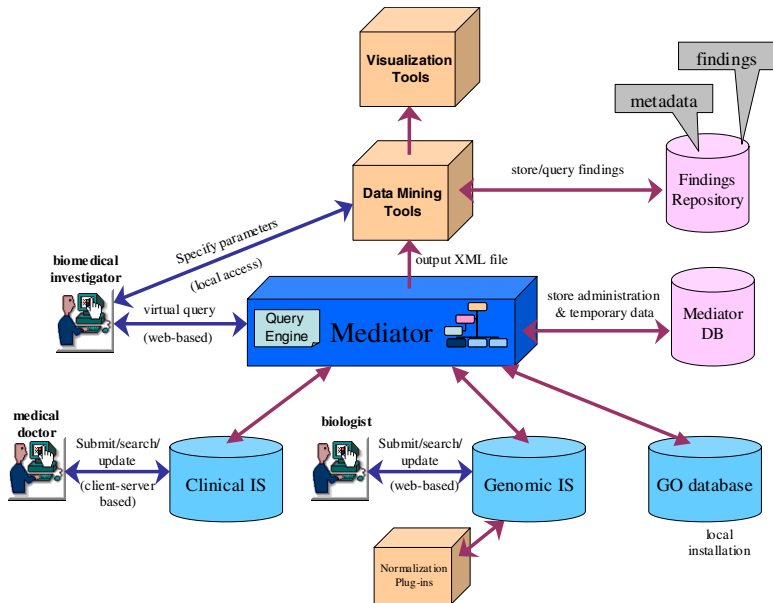


Fig. 1. The Integrated Clinico-Genomics Environment of ProgenoChip

To intelligently correlate clinical and genomic information towards ProgenoChip’s goal of individualized medicine, an *Integrated Clinico-Genomics Environment (ICGE)* has been implemented (see Figure 1), consisting of (a) a *Clinical Information System* to keep patients’ clinical information (i.e., clinical, laboratory, and histopathological information), (b) a *Genomic Information System* to manage the specifications of the respective DNA microarray experiments (i.e., microarray design, hybridizations, etc.), analyze the raw hybridization data, and store the samples’ gene expression profiles, (c) a middleware layer, called the *ProgenoChip Mediator*, for the integration of the Clinical and Genomic Information Systems, and (d) a *Data Mining*

layer, realized by an integrated set of tools, for the intelligent processing of the retrieved clinico-genomic data, knowledge extraction, and visualization.

In this paper, we present the *Prognosis Mediator*, and we give a brief overview of the Clinical and Genomic Information Systems. The Prognosis Mediator offers a powerful web-based graphical user interface for forming clinico-genomic queries. This is split into several query forms, allowing cancerous sample selection (along with their associated gene expression profiles and patient characteristics), based on criteria of interest. After a query is formed, the Mediator translates it into an equivalent set of local subqueries, which are executed directly against the constituent databases. Then, results are combined and an output XML file is formed for presentation to the user and/or transmission to the Data Mining tools for analysis.

The rest of the paper is organized as follows. In Section 2, the subsystems integrated by the Prognosis Mediator are briefly presented. In Section 3, we describe the architecture and interface of the Prognosis Mediator. Finally, Section 4 concludes the paper, including example Data Mining scenarios.

2 The Integrated Subsystems

The Prognosis Mediator integrates the Clinical Information System, the Genomic Information System, and the *Gene Ontology* (GO) Database [2].

GO is a well-known ontology for the annotation of genes/gene products in terms of the biological processes in which they participate, the particular molecular functions that they perform, and the cellular components in which they act. In particular, GO consists of 3 independent taxonomies, namely *GO Biological Process*, *GO Molecular Function*, and *GO Cellular Component*. Ontology terms are associated with a GO id and a human readable GO name. Each gene product annotation should be accompanied by an evidence code, indicating the type of evidence that supports it. The three GO taxonomies are stored in the GO Database, which is publicly available.

The Clinical Information System

The Clinical Information System (*Clinical IS*) of Prognosis manages the Electronic Health Record (EHR) of breast cancer patients, by storing information related to patient identification and demographic information, medical history (past diseases, surgeries, medications, gynecological history), patient risk factors, family history of malignancy, clinical examinations and findings, results of laboratory exams (mammography, ultrasound, hematological and biochemical exams, etc.), histopathologic evaluation and TNM staging (indicating Tumor size, Lymph node involvement, and Metastatic spread), pre-surgical and post-surgical therapies, as well as therapy effectiveness and follow-up exams. The Clinical IS is actually the extension with breast cancer concepts of a patient-oriented Integrated Health Care Environment [16], called *Integrated Care Solutions*, developed by the Biomedical Informatics Lab of FORTH-ICS. The system is based on a client-server architecture, runs on a PC Server, and uses an SQL Server database backend. For the needs of Prognosis, the Clinical IS has been installed at PAGNH and PROLIPSIS.

The Genomic Information System

The Genomic Information System (*Genomic IS*) of PrognoChip is based on the BioArray Software Environment (BASE) [14]. BASE is a MIAME-compliant [3] database and analysis platform designed to be installed in any microarray laboratory and serve many users simultaneously via the web. BASE was developed by Lund University and is a free software release under the GNU General Public License. It runs on a Linux Server using a MySQL database backend.

In short, BASE manages biomaterials (samples, extracts, labeled extracts), reporters/genes and related annotations, as well as array production and hybridizations. When all related information is available, raw hybridization data can be stored. Several scanners and image processors are supported. Each step of a microarray experiment is associated with a protocol description. Raw hybridization data (called, *measured bioassay data* in MAGE-OM: MicroArray and Gene Expression - Object Model [6]) can be organized in Experiments and normalized. Normalization plug-ins are already installed in BASE, but it is also possible to develop and install your own plug-ins. Normalization is performed in a hierarchical way and several normalization methods (such as, lowess, median) can be performed to refine intermediate results.

In PrognoChip, BASE was extended and enhanced in order to ease a biologist's task, and to provide more functionalities (eg., improved result annotation, sorting of results based on user-selected fields, etc.). Moreover, several quality indicators have been added to *Extracts*, *Labeled Extracts*, and *Hybridizations* to provide users with the capability of storing and reviewing the quality of their experiments. Furthermore, new raw hybridization data fields have been added and reporter/gene annotations have been extended with the type of the reporter and ids to public databases, such as Ensembl [4] and EMBL [5], as well as, Gene Ontology (GO) ids/names/evidence codes. Experiments participating in PrognoChip studies, called *PrognoChip Experiments*, are marked with a special flag. For these experiments, an integrity constraint verifies that the same (cancerous and "reference") samples are used in all participating hybridizations and that the designs of the participating arrays are different. This guarantees correspondence of a wet lab experiment to a single dry lab experiment, where four different arrays are used to cover the whole human genome. For the needs of PrognoChip, a *Print-tip Loess – no Background Correction* normalization plug-in [7] (developed by Uppsala University) has been installed and several *PrognoChip normalization procedures* have been defined, which are formed by a fixed sequence of normalization plug-in calls with their associated parameters. The Genomic IS has been installed in the participating molecular biology lab at FORTH-IMBB.

3 The PrognoChip Mediator

The (horizontal¹) integration of the Clinical ISs, the Genomic IS, and the GO database is achieved through the *PrognoChip Mediator* (see Figure 1), which offers a virtual common query model while data is stored only in the constituent heterogeneous

¹ Horizontal integration is the composition of semantically complementary data from multiple heterogeneous sources.

databases. The Mediator is developed using the ASP programming language and ODBC for accessing the corresponding databases.

The (authorized) biomedical investigator can form clinico-genomic queries through the web-based graphical user interface of the Mediator. This is split into several query forms, whose links are found in the left frame of the web interface (see Figure 2). Through these forms, the user can specify criteria for selecting tumors: (i) excised from patients with a clinical profile of interest, (ii) having histopathologic characteristics of interest, and (iii) participating in a PrognoChip microarray experiment of specified quality and characteristics. Returned tumors are accompanied with desirable clinico-genomic information and, in particular, their gene expression profiles. In all forms, numerical and datetime fields are queried as ranges. Additionally, some fields are dependent on a parent field and are appearing only if the parent field takes a specific value.

After the selection criteria of a clinical or genomic form are saved, a complex SQL subquery, or part of it, is formed which is temporarily stored in the Mediator DB. The final SQL subqueries are formed only after the desired output fields are selected by the user. Then, the final subqueries are submitted to the corresponding Clinical ISs or Genomic IS for evaluation. Below, we review the PrognoChip Mediator query forms:

Search by Sample Form

In the *Search by Sample* form, the user specifies the sample names on which the clinico-genomic selection criteria will be applied, as long as these belong to patients participating in the PrognoChip study. If no sample names are specified then all tumors of breast cancer patients participating in the PrognoChip study are considered.

Clinical Query Forms

In the *Past Breast Diseases* form, the user specifies patient's past breast diseases, such as fibroadenoma and fibrocystic changes, selecting these from a predetermined list of breast diseases. For each selected disease, the user can specify if it occurred on the left, right, or both breasts, as well as the number of years that have passed before tumor surgery. Criteria on selected diseases can be combined with either an AND or an OR logical operator.

In the *Gynecological History* form, the user specifies the age of menarche, the patients's menopausal status, the age at menopause (if the patient is postmenopausal), the number of pregnancies, the number of child births, the age at first child birth, and the lactation duration (if the patient has breastfed).

In the *Hormone Intake* form, the user specifies the total duration of hormone intake for conception, regulation, prevention, and contraception, respectively.

In the *Other Risk Factors* form, the user specifies the patient's age, body mass index, and smoking habits, including number of smoked cigarettes per day, years of smoking, and years passed after quitting (if the patient has quit smoking).

In the *Family History of Malignancy* form, the user specifies patient's relatives that have developed cancer sometime in their lives, along with the age that this happened. The various degrees of kinship, as well as types of cancer are selected from corresponding pull-down menus. The specified criteria on the different relatives can be combined by either an AND or an OR logical operator.

PrognoChip Mediator

Institute of Computer Science
Foundation for Research and Technology - Hellas

[Home](#) [About Us](#) [Contact Us](#) [MyAccount](#) [Clear All Forms](#) [Logout](#) [Welcome Mr Kalaitzakis](#)

■ Search By Sample

Clinical Forms

■ Past Breast Diseases

■ Gynecological History

■ Hormone Intake

■ Other Risk Factors

■ Family History of Malignancy

■ Breast - Nipple Findings

■ Blood Tumor Markers

■ Therapies Before Surgery

■ Therapies After Surgery

■ Therapy Effectiveness

Histopathologic Forms

■ Histological Type

■ Carcinoma, Other

■ Lymph Nodes

■ Immunohistochemical Tumor Markers

Genomic Forms

■ MicroArray Experiment

■ Reporter Filtering

Make Query

Histological Type

Ductal Carcinoma

☒ in situ

<= percentage <=

<= extend (mm) <=

cell type

from small cells

necrosis

with necrosis

prevailing hist. type CONTAINS:

☒ microinvasive

<= size (mm) <=

foci type

multifocal

<= foci number <=

<= from <=

<= to <=

☒ invasive

with Invasions satisfying the following constraints

<= number of <=

Invasions

Constraints

<= dimension X <=

(mm)

<= dimension Y <=

(mm)

<= dimension Z <=

(mm)

☒ DCIS <= percentage <=

scirrous reaction

1+

 lymphoplasmacytic reaction

2+

Lobular Carcinoma

☒ in situ

<= percentage <=

☒ invasive

<= diameter (mm) <=

Carcinoma of Specific Character

☒

description CONTAINS:

save

clear form

Fig. 2. The *Histological Type* query form

In the *Breast - Nipple Findings* form, the user specifies patient’s current breast findings (eg., skin invasion) and current nipple findings (eg., retraction), selecting these from a predetermined list of breast and nipple findings. For each selected finding, the user can specify if it occurred on the left, right, or both breasts. Additionally, in this form the user specifies the clinical TNM stage of the patient before pre-surgical chemotherapies and radiotherapies, where the *T*, *N*, *M* fields are queried as ranges.

In the *Blood Tumor Markers* form, the user selects from a predetermined list of blood tumor markers, such as CA15-3 and CEA, the tumor markers of interest. Then, for each selected tumor marker, the user can specify a range of values.

In the *Therapies Before Surgery* form, the user specifies from a predetermined list, the chemotherapies undergone by the patient before surgery (for shrinking tumor size). For each selected chemotherapy, the user specifies if the therapy is intravenous

or local, the number of cycles, and patient's response to the therapy. Similarly, the user specifies from a predetermined list, the radiotherapies undergone by the patient before surgery. In this form, the user also specifies the clinical TNM of the patient after completion of the course of pre-surgical chemotherapies and radiotherapies, respectively.

In the *Therapies After Surgery* form, the user specifies from predetermined lists, the chemotherapies, radiotherapies, hormonotherapies, and immunotherapies undergone by the patient after surgery. Additionally, the user can specify other taken actions.

In the *Therapies Effectiveness* form, the user specifies the Disease Free Survival (DFS) and the Overall Survival (OS) of the patient, as well as if she has presented metastasis or died due to relapse. Of course, if the patient has not presented metastasis (resp. is alive) then only a lower-bound query on DFS (resp. OS) is meaningful.

Histopathologic Query Forms

In the *Histological Type* form, the user specifies the histological type of the tumor, i.e., *Ductal Carcinoma* (in situ, microinvasive, invasive), *Lobular Carcinoma* (in situ, invasive), and *Carcinoma of Specific Character*, by clicking the corresponding checkbox, as shown in Figure 2. When the checkbox is clicked, the fields associated with the particular histological type are appearing. Note that in the case of an invasive ductal carcinoma, the user can specify the number of invasions that satisfy certain criteria.

In the *Carcinoma, Other* form, the user specifies the diameter, grade, and histological TNM of the tumor. It also specifies, (i) if there is lymphatic, venous, perineural, skin, and/or nipple invasion (Paget or non-Paget type), (ii) invasion on the excision margins of the tumor, and (iii) if the tumor is a phylloides tumor or a lesion with atypia. In the latter case, the grade of atypia can also be specified.

In the *Lymph Nodes* form, the user specifies the number of lymph nodes, sentinel lymph nodes, local lymph nodes, and intramammary lymph nodes, respectively, that have X, Y, and Z dimensions within a certain range, present metastatic invasion, capsule invasion, and/or invasion of the adipose tissue of axilla. In the case of metastatic invasion, the user can select the type of metastasis from a pull-down menu.

In the *Immunohistochemical Tumor Markers* form (see Figure 3), the user selects immunohistochemical tumor markers, such as ER, PR, HER-2, etc., from a pull-down menu and for each one of them, he/she specifies the percentage of tumor cell staining, the intensity of staining, the tumor marker score, and internal and external positive control information. Additionally, the user can indicate if the HER-2 FISH score is positive or negative. For example, in Figure 3, the user selects tumors that are both HER-2 and ER positive.

Genomic Query Forms

In the *MicroArray Experiment* form, the user specifies the channel of the cancerous tissue in the microarray experiment (*Channel 1*, if the tumor extract is labeled with the Cy3 fluorescence, or *Channel 2*, if the tumor extract is labeled with the Cy5 fluorescence). As it will become obvious later, this is needed for the correct interpretation of the normalized data. The user also specifies the dates and qualities of the experiment hybridizations, as well as the name of the desired PrognChip normalization procedure (selected from a pull-down menu).

Immunohistochemical Tumor Markers					
Tumor Markers					
Name	Percentage	Intensity of Staining	Score	Internal Positive Control	External Positive Control
HER-2	20 <= % <=	+2	+3		Add new row
ER	30 <= % <=	+3	+3	2+	1+ Remove row
Her 2 FISH					
score: positive					
comments CONTAIN:					
save clear form					

Fig. 3. The *Immunohistochemistry Tumor Markers* query form

In the *Reporter Filtering* form (see Figure 4), the user can filter the reporters appearing in the gene expression profiles of the selected tumors, based on (i) their library name, (ii) their associated GO Biological Process, GO Molecular Function, and GO Cellular Location names, and (iii) their GO annotation evidence codes. GO names can be inserted manually into the corresponding textboxes but they can also be automatically inserted by accessing the local installation of the GO database and retrieving the direct children or descendants of a particular GO name, specified by the user. Criteria on GO Biological Process, GO Molecular Function, and GO Cellular Component annotations, respectively, can be combined by either an AND or an OR operator. Additionally, the user can specify a list of evidence codes of interest. For example, in Figure 4, the user specifies that he/she is interested only in the reporters of the library “*Human Operon v.3*” and annotated by the GO Biological Process term “*signal transduction*” or any of its descendants. In this case, the user enters *signal transduction* in the *GO Biological Process names* textbox, and also enters the same GO name in the *Add all descendants of* textbox, clicking on the *Add* button. At this time, a query asking “for all descendants of *signal transduction* in the GO Biological Process taxonomy” is created and submitted to the local GO database. All GO names in the answer are inserted automatically into the *GO Biological Process names* textbox. Then, the user clicks on the OR radio button (to the right), indicating that annotation criteria should be combined by the OR operator. The user also specifies that he/she is interesting only in annotations, supported by *IC* (inferred by curator) or *IDA* (inferred by direct assay) type of evidence.

Select Page

After patient/tumor criteria specification, the user selects the clinical, histopathologic, and genomic fields of interest that will accompany the cancerous sample in the output. This is done in the *Select Page* (the top half of which is shown in Figure 5), which is appearing after pressing the *Make Query* button, shown in Figure 2. Moreover, in the *Select Page*, the user specifies the gene expression measures of interest and the annotations that should accompany the reporters, selected based on the criteria of the *Reporter Filtering* form. For example, in Figure 5, the user specifies that he/she is interested in the pre-surgical chemotherapies and radiotherapies of the selected patients. Additionally, in the *Select Page*, the user can specify that from the

normalized data fields, he/she is interested in both the $\log_2(intensity1/intensity2)$ and $\log_2(intensity1 + intensity2)$ measures (corresponding to different gene expression measures), where *intensity1* and *intensity2* are the normalized fluorescence intensities of the tissues placed in Channel 1 and Channel 2 of the microarray experiment, respectively. Moreover, the user can specify that, from the reporter annotations, he/she is interested in the *HUGO Gene Nomenclature Committee* name of the corresponding gene, the Ensembl gene id, and the corresponding GO Biological Process annotations.

In the Select Page, the user also specifies the Clinical ISs to which the clinical subqueries will be submitted. For example, in Figure 5, the user selects only the Clinical IS of PAGNH. Though in ProgenoChip, there are only two participating Clinical ISs, our system can handle any number of these. Information on the participating Clinical ISs and Genomic IS is stored in the *Mediator DB* (see Figure 1) and handled by the administrator through the *Administration* form.

Reporter Filtering

Library:

Human Operon v.3

GO Biological Process names: (as ; separated strings)

signal transduction; integrin-mediated signaling pathway; integrin-mediated signaling pathway; glucose mediated signaling; defense response signaling pathway, resistance gene-

AND

OR

Add the direct children of:

Add

Add all descendants of:

signal transduction

Add

GO Molecular Function names: (as ; separated strings)

AND

OR

Add the direct children of:

Add

Add all descendants of:

Add

GO Cellular Component names: (as ; separated strings)

AND

OR

Add the direct children of:

Add

Add all descendants of:

Add

GO annotation evidence codes: (as ; separated strings)

IC; IDA

OR

save

clear form

Fig. 4. The *Reporter Filtering* form

Submission of Final Subqueries & Result Composition

The Mediator is now able to form the final subqueries that will be submitted to Clinical and Genomic ISs. Due to the fact that gene expression data are much larger in size than patient clinical data, the Mediator first submits the clinical subqueries to the Clinical ISs. Subquery results are then combined, providing not only the selected

clinico-histopathologic attributes for the patients satisfying the user-specified, clinico-histopathologic criteria but also their corresponding tumor names. The latter are used in forming the final SQL subqueries that will be submitted to the Genomic IS.

Clinical DB Selection		
Please select the database(s) you want to query		
CLINICAL DATABASES		
Check All Clear All		
☒ PAGNH		
☐ PROLIPSIS		
Please select the fields of interest		
Clinical Fields		
PAST BREAST DISEASES		
Check All Clear All		
☐ disease name	☐ disease position	☐ years before surgery
GYNECOLOGICAL HISTORY		
Check All Clear All		
☐ age of menarche	☐ menopausal status	☐ age at menopause
☐ number of pregnancies	☐ number of childbirths	☐ age at first childbirth
☐ lactation	☐ lactation duration (weeks)	
DURATION OF HORMONE INTAKE		
Check All Clear All		
☐ for conception (months)	☐ for regulation (months)	
☐ for prevention (months)	☐ for contraception (months)	
OTHER RISK FACTORS		
Check All Clear All		
☐ age	☐ body mass index (BMI)	
☐ smoking habits (cigarettes per day, years of smoking, quitted, years after quitting)		
FAMILY HISTORY OF MALIGNANCY		
Check All Clear All		
☐ degree of kinship	☐ at age	☐ type of cancer
BREAST-NIPPLE FINDINGS		
Check All Clear All		
☐ clinical TNM (T,N,M)	☐ breast finding (name, position, date of finding)	☐ nipple finding (name, position, date of finding)
BLOOD TUMOR MARKERS		
Check All Clear All		
☐ tumor marker (name, value, date of exam)		
THERAPIES BEFORE SURGERY		
Check All Clear All		
☒ chemotherapy (start date, intravenous, local, type, number of cycles, response (percentage), clinical TNM after therapy)		
☒ radiotherapy (start date, type, number of cycles, clinical TNM after therapy)		
THERAPIES AFTER SURGERY		
Check All Clear All		
☐ chemotherapy (start date, type, number of cycles, comments)		
☐ radiotherapy (start date, type, number of cycles, comments)		
☐ hormonotherapy (start date, type, comments)		
☐ immunotherapy (start date, type, comments)		
☐ other action (start date, action)		
THERAPY EFFECTIVENESS		
Check All Clear All		
☐ Disease Free Survival (months)	☐ metastasis	☐ DFS comments
☐ Overall Survival (months)	☐ deceased	☐ deceased due to relapse
☐ OS-comments		

Fig. 5. The *Select Page* (top part)

Final results of the Clinical and the Genomic ISs are joined, based on the names of the cancerous samples, and an XML file, called *output XML file*, of predetermined schema is created. Since the set of selected reporters can be very large, reporter/gene annotations are stored in a separate tab-delimited file, called *reporter file*. Similarly, the requested normalized data fields of each cancerous sample are also saved (separately), in a tab-delimited file, called *normalized data file*. This way, the output XML file can contain only the names of the corresponding files, along with their column definition (eg., *<reporter id, gene name (HUGO), Ensemble gene id, GO Biological Process>* for the (single) reporter file, and *<reporter id, $\log_2(intensity1/intensity2)$, $\log_2(intensity1 + intensity2)$ >* for the normalized data files).

The output XML file is viewed through the current web browser, in a separate window, and the complete set of output files can be downloaded to the local machine of the user (as a .zip file). Then, it is given as input to the Data Mining tools for mining interesting clinico-genomic associations between the retrieved attributes of the selected samples. Finally, interesting Data Mining findings can be annotated and stored in the Findings Repository.

The schema of the output XML file can be found at: <http://www.ics.forth.gr/~analyti/ProgenoChip/OutputXMLFileSchema.jpg>.

4 Conclusions

In this paper, we presented the *ProgenoChip Mediator*, which is part of the Integrated Clinico-Genomics Environment (ICGE) of ProgenoChip. The Mediator integrates clinical and genomic data from the respective information systems, through a powerful web-based graphical user interface for submitting clinico-genomic queries. Mediator replies are not only useful by themselves, but they can also support decision making operations, enabled by the Data Mining layer of ICGE, through knowledge extraction and data mining methodologies [7,13]. Example scenarios include:

- Retrieve the tumor gene expression profiles of patients that meet clinical (specifically, clinico-histopathological) profile A and clinical profile B. Then, by the appropriate data-mining and gene-selection methods, find the genes that best discriminate between these two groups (i.e., disease-related gene markers).
- By clustering the tumor gene expression profiles, search and identify the respective clinical description of the corresponding patients. Then, with the application of feature selection and/or classification methods, identify potential interesting and indicative patient clinical profiles.

Arguably, ontology-based integration systems, such as TAMBIS [15], ONTOFUSION [11], and BACIIS [8], are more flexible and scalable than the ProgenoChip Mediator. This is because, considering a single domain conceptualization through an *ontology*, they accept ontology-based queries, which they *dynamically* break into local subqueries, based on pre-defined ontology-to-data source mappings. However, these systems are not powerful enough to (i) offer the user friendliness of our system, and (ii) support the complexity of the local SQL subqueries and additional main-memory processing, required to return the desired data and data associations between the Genomic and Clinical Information Systems of ProgenoChip.

Future work concerns further optimization of the PrognoChip Mediator and performance evaluation based on real data.

Acknowledgements. The authors would like to thank Prof. E. Sanidas , Prof. E. Stathopoulos, M. Michou and M. Nikoloudakis for their help in the design and implementation of the PrognoChip Mediator, T. Margaritis for his help in designing the Genomic IS, and N. Stathiakis, E. Leich, and M. Damianakis for explaining to us the Clinical IS. This work has been supported by the project PrognoChip, funded by the General Secreteriat of Research and Technology, Greece.

References

1. <http://www.ncbi.nlm.nih.gov/About/primer/microarrays.html>
2. <http://www.geneontology.org/>
3. <http://www.mged.org/Workgroups/MIAME/miame.html>
4. <http://www.ensembl.org/>
5. <http://www.ebi.ac.uk/embl/>
6. <http://www.mged.org/Workgroups/MAGE/mage.html>
7. A. Ameur, V. Yankovski, S. Enroth, O. Spjuth, and J. Komorowski: The LCB Data Warehouse. *Bioinformatics*, 22(8), pp. 1024-1026, 2006.
8. Z. Ben Miled, N. Li, and O. Bukhres: BACIIS: Biological and Chemical Information Integration Systems. *Journal of Database Management*, 16(3), pp. 72-85, 2005.
9. A. Kanterakis and G. Potamias : Supporting Clinico-Genomic Knowledge Discovery: A Multi-strategy Data Mining Process. *Procs. of the 4th Hellenic Conference on AI (SETN 2004)*, LNAI 3955, pp. 520-524, 2006.
10. D. J. Lockhart and E. A. Winzeler: Genomics, gene expression and DNA arrays. *Nature*, 405(6788), pp. 827-836, 2000.
11. D. Perez-Rey., V. Maojo, M. Garcia-Remesal, R. Alonso-Calvo, H. Billhardt, F. Martin-Sanchez, A. Sousa: ONTOFUSION: Ontology-based integration of genomic and clinical databases. *Computers in Biology and Medicine*, 36(7-8), pp. 712-730, 2005.
12. G. Potamias, A. Analyti, D. Kafetzopoulos, M. Kafousi, T. Margaritis, D. Plexousakis, P. Poirazi, M. Reczko, I. G. Tollis, E. Sanidas, E. Stathopoulos, M. Tsiknakis, S. Vassilaros: Breast Cancer and Biomedical Informatics: The PrognoChip Project. *Procs. of the 17th IMACS world Congress Scientific Computation, Applied Mathematics and Simulation*, Paris, France, 2005.
13. G. Potamias, L. Koumakis, and V. Moustakis. Gene selection via discretized Gene-Expression Profiles and Greedy Feature-Elimination. *Procs. of the 4th Hellenic Conference on AI (SETN 2004)*, LNAI 3025, pp. 256-266, 2004.
14. L. H. Saal, C. Troein, J. Vallon-Christersson, S. Gruvberger, Å. Borg, and C. Peterson, BioArray Software Environment: A Platform for Comprehensive Management and Analysis of Microarray Data, *Genome Biology*, 3(8): software0003.1-0003.6, 2002.
15. R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N. W. Paton, C. A. Goble, and A. Brass: TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources. *Bioinformatics*, 16(2), pp. 184-186, 2000.
16. M. Tsiknakis, D. G. Katehakis, S. C. Orphanoudakis, An open, component-based information infrastructure for integrated health information networks, *Intern. Journal of Medical Informatics*, 68(1-3), pp. 3-26, 2002.

OntoDataClean: Ontology-Based Integration and Preprocessing of Distributed Data

David Perez-Rey, Alberto Anguita, and Jose Crespo

Biomedical Informatics Group, Artificial Intelligence Laboratory,
School of Computer Science, Universidad Politécnica de Madrid
Campus de Montegancedo, s/n. 28660 Boadilla del Monte, Madrid
dperez@infomed.dia.fi.upm.es

Abstract. Within the knowledge discovery in databases (KDD) process, previous phases to data mining consume most of the time spent analysing data. Few research efforts have been carried out in these steps compared to data mining, suggesting that new approaches and tools are needed to support the preparation of data. As regards, we present in this paper a new methodology of ontology-based KDD adopting a federated approach to database integration and retrieval. Within this model, an ontology-based system called OntoDataClean has been developed dealing with instance-level integration and data preprocessing. Within the OntoDataClean development, a preprocessing ontology was built to store the information about the required transformations. Various biomedical experiments were carried out, showing that data have been correctly transformed using the preprocessing ontology. Although OntoDataClean does not cover every possible data transformation, it suggests that ontologies are a suitable mechanism to improve quality in the various steps of KDD processes.

Keywords: Knowledge Discovery in Databases, Preprocessing, Data Cleaning, Database Integration, Ontologies.

1 Introduction

In most biomedical institutions and organizations, data is being produced and stored at an increasing rate. In this scenario, new models of data analysis are being developed, aiming to support knowledge extraction from different data sources. This process is known as Knowledge Discovery in Databases (KDD). Whereas data mining is usually the main step, the KDD previous phases are crucial to ensure data integration, gathering and quality, preparing the data for further analysis. In fact, experts in charge of analysing the data of a new experiment usually dedicate more efforts to understand and to prepare data than to apply the specific data mining algorithm [1].

Although the relevance of the preprocessing phase has been widely recognized in the literature [2] [3], few research efforts have been carried out to improve this topic. Data cleaning and preprocessing projects seem to have less interest within the

scientific community compared to projects centred on the data mining phase. Current preprocessing tools usually address just some aspects of the problem and require an important user effort. Recently, proposals based on the Semantic Web and its related mechanisms to use metadata support in the KDD process. Ontologies suggest new approaches to enhance the quality of data, facilitate domain understanding and provide formal knowledge support for data preprocessing-related tasks.

According to this scenario, a new approach of ontology-based KDD is proposed in this paper. It adopts a federated approach to access different, remote data sources. Our approach, which has been named OntoDataClean, has been developed to deal with previous stages to data mining. It is also used to enhance information sharing among organizations, aiming to solve main challenges in the field of heterogeneous database integration [4]. Instance-level integration, preprocessing and transformation of records stored in different databases are the focus of the system. To deal with these issues, an ontology-based approach is presented. This approach aims to generate an intuitive framework where data experts model the data and store the information needed to transform such data.

The paper is structured as follows. First, section 2 presents the state of the art on ontologies applied to KDD, definitions and different approaches. Section 3 presents our ontology-based KDD model. Section 4 describes OntoDataClean, the tool that we have developed for data preprocessing. Finally, various experiments with their results and conclusions are described in sections 5 and 6.

2 State of the Art

Knowledge Discovery in Databases or KDD has been defined as “the process of non trivial and potentially useful knowledge discovery” [5]. It is composed of various steps and can be applied in user hypothesis verification or automatic pattern discovery. KDD methodologies are frequently based on previous proposals [3], adopting a centralized approach concerning data source location, i.e. Data Warehouses. Although this option has a performance advantage, data is not always up to date and this is the key drawback to adopt the federated approach in the biomedical field [6] [7]. Additionally, performance weaknesses are losing significance due to continuous advances in computing speed and data transmission and this trend is expected to continue in the future.

In this framework and modifying classical approaches, six different phases can be considered within the KDD process: (i) data integration, (ii) selection, (iii) preprocessing, (iv) transformation, (v) data mining and (vi) interpretation. As stated above, few efforts have been carried out to support data selection and preprocessing, although some literature is available – e.g. AJAX [8], Potter’s Wheel [9]. In order to provide access to consistent and precise data, it is necessary to homogenize the data representation, eliminate duplicated information and solve inconsistencies. Depending on the number of data sources (one or more sources) different types of inconsistencies can be considered. Regarding the level of these inconsistencies we can distinguish between instance-level and schema-level inconsistencies [1].

Three main goals might be identified in the preprocessing task in KDD processes: (i) data standardization, (ii) data preprocessing to enhance KDD results and (iii) maximum data reduction without loosing necessary information. Including information about data (metadata) in these systems is especially important to achieve these objectives. Ontologies [10] have been used as metadata mechanisms in several data integration projects, proving its suitability in this area [11]. Besides, it has been established that ontologies can be used to support all phases of KDD [12]. The objective of introducing ontologies in the preprocessing phase of the KDD is to produce data improvements with a formal domain knowledge support.

Below, Table 1 presents a review of some ontology-based applications developed for each one of the different KDD phases, as available at the time of writing this paper.

Table 1. Ontology-based KDD review

#	KDD phase	Ontology applications within the KDD phase	References
i	<i>Integration</i>	In general, ontologies are used in this phase as virtual repositories representing heterogeneous data sources. Nowadays, this phase has been the main focus of ontology application in KDD.	[11], ONTOFUSION [13], D2RMAP [14], SEMEDA [15], KAON [16]
ii	<i>Selection</i>		
iii	<i>Preprocessing</i>	Ontologies in this case can be used in two ways: (i) storing the information needed to transform the instances of the data or (ii) actually storing the instances in a formal representation. Some tasks included in these phases can be tackled together with previous ones, or even overlap with them.	[17], [18], ONTOCLEAN [19]
iv	<i>Transformation</i>		
v	<i>Data mining</i>	Users analyzing the data may use ontologies to choose the most suitable algorithm among the huge amount of available ones.	PROTEUS [20], IDEA [21]
vi	<i>Interpretation</i>	Providing a formal representation of the domain of a new knowledge facilitates its understanding and reutilization.	ONTO4KDD4O NTO [22], LISp-Miner [23], MiningMart [24]

Some of the stages are joined due to the scope of the projects supporting them. Integration and selection phases are usually covered by ontology-based database integration projects. They are also the tasks where the research community has found the best synergy between ontologies and KDD. The following phases previous to data mining have been much less investigated, not only regarding the use of ontologies. Some approaches using ontologies and targeting the preprocessing and transformation tasks have been stated theoretically, but none of them have been verified with a complete experiment.

Summarizing, there is a lack of approaches and tools to assist the data processing step before the data mining phase. Anyway, several applications of ontologies in each KDD phase have proved its suitability for these methodologies promising new developments in the near future.

3 Ontology-Based and Federated KDD

Different terms have been used in the literature to describe previous phases to data mining. Some of them are: Data cleaning (or cleansing), preprocessing, homogenization, standardization and others. These terms include differentiated features, challenges and purposes but they also present similarities if they are related to data integration. Although these tasks must be tackled at different stages they may require the same functionalities.

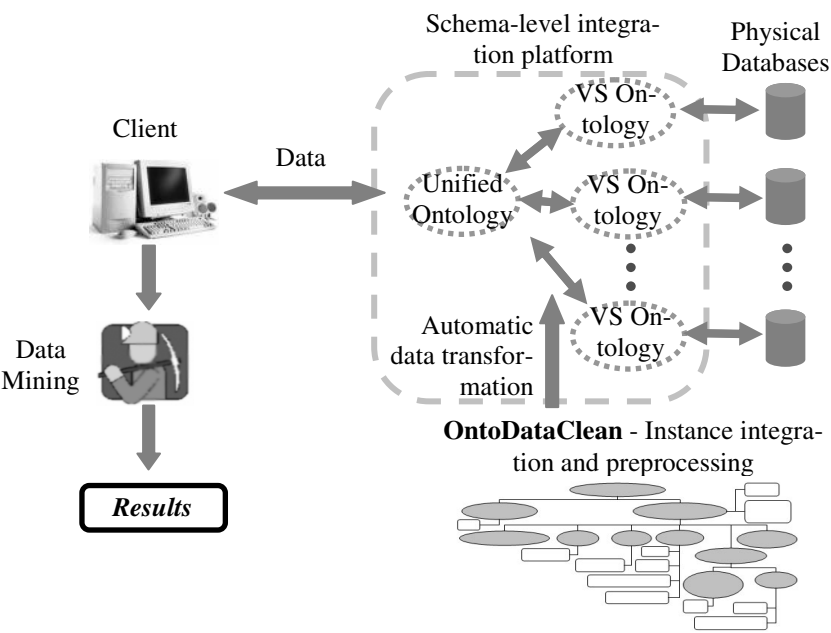


Fig. 1. Federated and ontology-based KDD approach

The new approach for ontology-based KDD presented in this section aims to support the analysis of data stored in heterogeneous sources at different locations. Ontologies are adopted as the metadata mechanism to store the information about the data to be analysed. Using this approach, data experts do not need to be also experts in data cleaning and integration. OntoDataClean aims to be a tool to solve data problems in all these fields. Fig. 1 presents the new proposed KDD process and locates OntoDataClean in this environment.

The new architecture extends the idea of previous work carried out using an ontology-based approach to heterogeneous database integration at the schema-level [12]. In this approach, ontologies are used as virtual schemas representing the different data sources. Although these virtual schemas do not store any data, they are mapped to a database physical schema. Once these repositories are developed, users may query an ontology instead of an entity-relationship schema. Following a similar approach, at the instance-level in this KDD model, ontologies are used as frameworks to store the necessary information to modify the records of the database.

The process of identifying and storing the required transformations has to be always supervised by an expert in the data. These transformations are needed to homogenize and integrate the records so they can be correctly analyzed or unified with other sources. Once the required information is stored in the preprocessing ontology, data transformations can be accomplished automatically. As stated in section 2, few developments have been carried out in this KDD stage using ontologies. Although models regarding ontology-based data transformations have not been fully tested, all of them follow an approach with these two sequential steps: (i) a supervised detection of inconsistencies and (ii) an automatic transformation of data when retrieved by the user.

4 OntoDataClean

OntoDataClean is an ontology-based tool aimed to solve inconsistencies and other issues of former KDD phases. A new structure of preprocessing ontology has been developed, called the OntoDataClean preprocessing ontology. It is presented in Fig. 2.

The Ontology Web Language (OWL) [25] has been chosen to implement this ontology, since it is currently the most widely adopted standard. These ontologies can be visualized and edited using any of the available ontology editors supporting OWL – e.g. Protégé [26], SWOOP [27], KAON2 [28], etc.

Each data source in the system has an associated instance of the preprocessing ontology. These include two main classes, *Data Source* – to specify the source to be preprocessed – and the *Cleaning Model*. From the latter derive six different classes, each one in charge of one type of data transformation. If a specified transformation is needed, an instance of the corresponding class must be created. For a scale transformation, a *Scale* instance must be created with the corresponding formula within the *Expression* property. This process is similar to format, pattern and duplicate transformations but missing values and terminological inconsistencies follow different procedures. To establish a missing value transformation, an instance of the *Detection* class must be included in the ontology. There exist relationships to

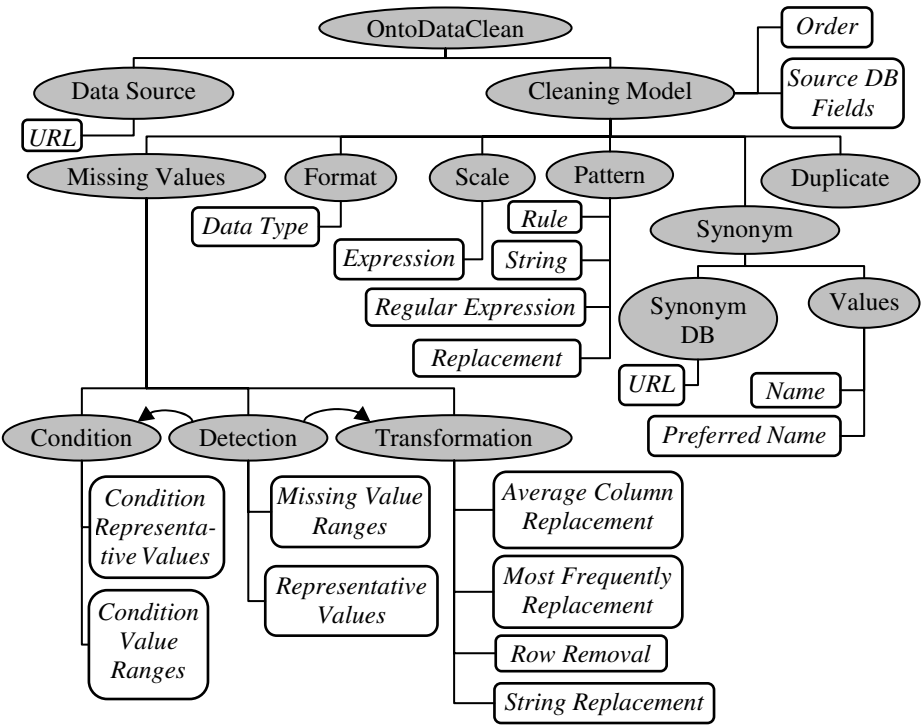


Fig. 2. OntoDataClean Preprocessing Ontology

connect each *Detection* instance with a list of *Condition* instances and one *Transformation* instance. *Condition* instances allow specifying additional conditions referring to other field values, whereas *Transformation* instances specify the type of replacement of the missing value, with four different options: (i) replacing the missing value with the mean value of the column, (ii) replacing with the most frequent value of the column, (iii) replacing with a given string and (iv) deleting the complete record. Finally, terminological transformations are divided into two different cases depending on the dictionary extension. If the needed dictionary is too large to be stored in a single ontology file, a database with a given structure (string – preferred string) can be built and linked to the system by means of the class *SynonymDB*. Otherwise, synonyms can be introduced as instances of *Values*, containing pairs of <string – preferred string> specifying a single terminological transformation.

When a user queries the system, OntoDataClean explores the preprocessing ontology, identifies the corresponding transformations and executes them before sending the data to the user. Each transformation is accomplished targeting the corresponding source field and following a sequential order given by the *Order* and *Source DB Fields*.

5 Results

OntoDataClean can deal with any kind of data, but it is especially suitable for biomedical data. Biomedical repositories usually include textual values containing technical concepts combined with quantitative data. This method can manage both types of data in order to carry out the required preprocessing tasks.

Four different biomedical databases available through the Internet were used to test OntoDataClean – Gepas [29], Reactome [30], BioMérieux [31] – and a local database about rheumatic arthritis [32]. For each data source, the process of inconsistency detection was supervised by an expert, obtaining the corresponding preprocessing ontology. Finally, the automatic mechanism to transform the data according to the information stored in the preprocessing ontology was tested against their corresponding databases.

The experiment using the Arthritis database deals with data from a retrospective cohort epidemiological study. It contains data from patients diagnosed with rheumatoid arthritis. Although all fields are integers, some of them represent categorical variables with only two possible values. The range of these categorical values differs between fields, so a homogenization is required for enabling further

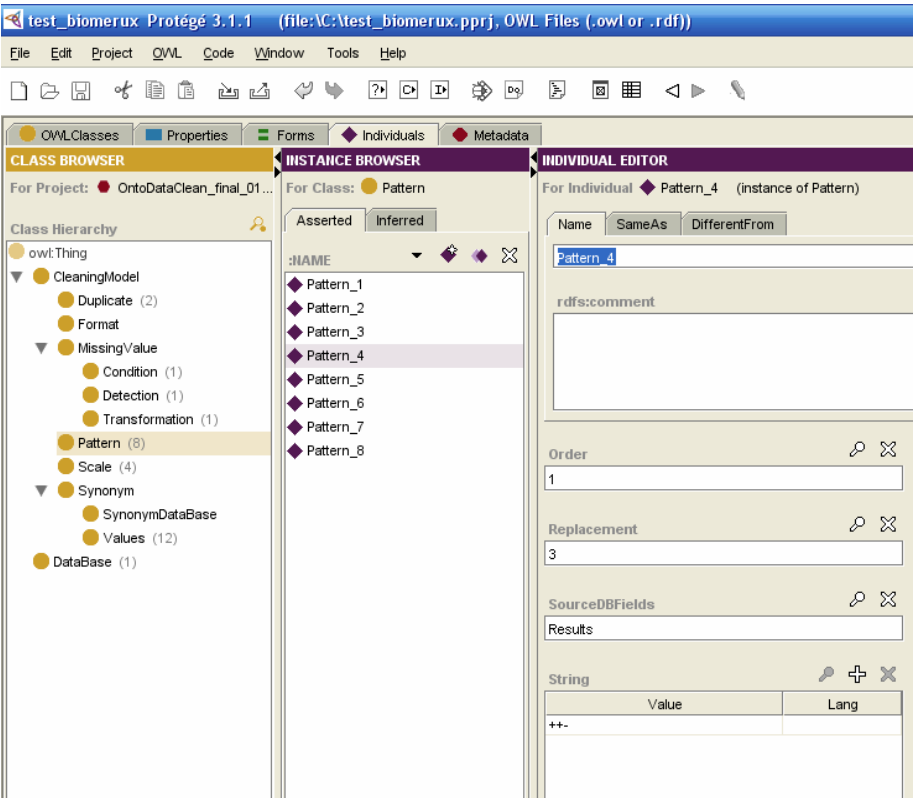


Fig. 3. Preprocessing ontology for BioMérieux experiment, implemented using Protégé

data analysis. This database contains also a great amount of missing values, which are erased in the preprocessing task. The Gepas database handles data concerning human fibroblasts response to serum. Each field represents this response given a specific time. The existing values of time follow an exponential progression. Therefore, response values must be adjusted using a logarithmic function. The experiment with Reactome tests several transformations among the data contained in the database. These include complex pattern modifications on string data concerning urls, erasure of duplicate values, synonym substitutions and missing values transformations. Finally, the BioMérieux database manages data representing biochemical profiles. A numerical representation transformation is required in the *Results* field, together with missing value detections. The pattern and missing value alternatives are applied to carry out a binary to decimal transformation and a deletion of instances with missing values.

Fig. 3 shows the Protégé-based implementation of the OntoDataClean preprocessing ontology employed in the BioMérieux experiment. This ontology consists of several instances, some of them aimed to perform the numerical transformation described above and others related to missing value removal. In Table 2, the results of a BioMérieux query before and after OntoDataClean transformation are shown.

Table 2. An example of BioMérieux data transformation using OntoDataClean

Id	Results	Id'	Results'
r01	++ +- +- +- ++	r01	6 4 2 1 5
r02	+- --- --- --- ---	r02	2 0 0 0 0
r03	+- +- +- --- ++	r03	2 4 2 0 3
r04	++ +- +- +- ++	r04	6 4 2 1 5
r05	++ +- +- +- +?	r07	0 0 4 2 1
r06	+- +- ?+ --- ++	r08	6 4 2 1 5
r07	--- --- +- +- ++	r09	2 0 0 0 0
r08	++ +- +- +- ++	r10	2 4 2 0 3
r09	+- --- --- --- ---	r12	6 4 2 1 5
r10	+- +- +- --- ++	r13	0 0 4 2 1
r11	-?- --- +- +- --	r14	2 0 0 0 0
r12	++ +- +- +- ++	r15	6 4 2 1 5
r13	--- --- +- +- --		
r14	+- --- --- --- ---		
r15	++ +- +- +- ++		

Id – Test identifier
Results – Biochemical profiles using binary codification
Id' – Test identifier
Results' – Biochemical profiles using decimal codification

Table 2 shows a portion of the Biomérieux experiment data before (left) and after (right) applying OntoDataClean. The ‘Biochemical profiles’ field is modified through a pattern transformation. Patterns with ‘+’ and ‘-’ symbols (which codify binary numbers) are transformed into decimal representation. Each chain of three symbols is

translated into a decimal number from 0 to 7. Additionally, each time a ‘?’ symbol is found (meaning that that precise data is missing or unknown), the whole row is deleted from the table. In this fraction of the results for the BioMérieux experiment, it can be observed that each value was transformed according to the given patterns in its corresponding preprocessing ontology. Results provided by OntoDataClean can be obtained in XML or tabulated format files to facilitate further application of data mining algorithms or database storage.

6 Conclusions and Further Research

In this paper, we have presented classical and new ontology-based approaches to the various steps of the KDD process. The results suggest that ontology-based approaches within the KDD process are a suitable mechanism to improve distributed data using formal knowledge support. In the area of genomic medicine, where clinical and biological data need to be integrated, new research approaches are needed to facilitate this work to researchers and users [33] [34] [35].

The federated approach adopted in OntoDataClean and the new KDD approach presented in this paper, solves the main drawback of data warehouses, avoiding that outdated data can be presented to users. Since every query is translated into sub-queries, the distributed approach ensures that the retrieved information is always consistent with the corresponding source. On the other hand, as stated in section 2, this feature may cause higher response times. A cache of results can be used as a balance between a centralized approach with high performance and a federated approach with continuous up-to-date data.

There are ongoing efforts to extend the system with a component for semi-automatic inconsistency detection. This module will aim to facilitate the supervised phase of inconsistency detection and preprocessing ontology construction.

Finally, as stated in section 3, OntoDataClean is intended to work together with a schema-level integration system. Work is already in progress to develop a global preprocessing and integration tool.

Acknowledgements. This work was funded by the INBIOMED research network (Ministry of Health), the INFOBIOMED network of excellence (IST-2002-507585) and the ACGT integrated project (FP6-2005-IST-026996).

References

1. Rahm, E., Hai Do, H.: Data cleaning: problems and current approaches. *IEEE Bulletin of the Technical Committee on Data Engineering* 23(4) (2001) 3-13
2. Dasu, T., Jonson, T.: *Exploratory Data Mining and Data Cleaning*. John Wiley & Sons (2003)
3. Weiss, S.M., Indurkha, N.: *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann (1998)
4. Gurwitz, D., Lunshof, J.E., Altman, R.B.: A call for the creation of personalized medicine database. *Nature Reviews, Drug Discovery* 5 (2006) 23-6

5. Fayyad, U., Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery in databases. *AI Magazine* 17 (1996) 37-54
6. Sujansky, W.: Heterogeneous Database Integration in Biomedicine. *Journal of Biomedical Informatics* 34(4) (2001) 285-98
7. Maojo, V., García-Remesal, M., Billhardt, H., Alonso-Calvo, R., Pérez-Rey, D., Martín-Sánchez, F.: Designing New Methodologies for Integrating Biomedical Information in Clinical Trials. *Methods Inf Med.* 45(2) (2006) 180-5
8. Galhardas, H., Florescu, D., Shasha, D., Simon, E.: AJAX: An Extensible Data Cleaning Tool. *SIGMOD'00 Conf. Management of Data*, Dallas (2000) 590
9. Raman, V., Hellerstein, J.M.: Potter's Wheel: An Interactive Data Cleaning System. *VLDB01, 27th International Conference on Very Large Databases*, Rome (2001) 381-90
10. Gruber, T.R.: A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition* 5(2) (1993) 199-220
11. Silvescu, A., Reinoso-Castillo, J., Honavar, V.: Ontology-Driven information extraction and knowledge acquisition from heterogeneous, distributed, autonomous data sources. *Proceedings of the IJCAI* (2001)
12. Cespivova, H., Rauch, J., Svatek, V., Kejkula, M., Tomeckova, M.: Roles of Medical Ontology in Association Mining CRISP-DM Cycle. In: *ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies (KDO'04)*, Pisa (2004)
13. Pérez-Rey, D., Maojo, V., García-Remesal, M., Alonso-Calvo, R., Billhardt, H., Martín-Sánchez, F., Sousa, A.: ONTOFUSION: Ontology-Based Integration of Genomic and Clinical Databases. *Computers in Biology and Medicine* 36 (2006) 712-30
14. Bizer, C.: D2R MAP - A Database to RDF Mapping Language. In *Proceedings of the International World Wide Web Conference (WWW2003)*, Budapest, Hungary (2003)
15. Köhler, J., Philippi, S., Lange, M.: SEMEDA: ontology based semantic integration of biological databases. *Bioinformatics*, 19(18) (2003) 2420-7
16. <http://kaon.semanticweb.org/alphaworld/reverse/> (last accessed September. 1, 2006)
17. Phillips, J., Buchanan, B.G.: Ontology-guided knowledge discovery in databases. *International Conf. Knowledge Capture Victoria*, Canada (2001)
18. Kedad, Z., Métais, E.: Ontology-based Data Cleaning. In *NLDB'02*, Springer Verlag (2002)
19. Wang, X., Hamilton, H.J., Bither, Y.: An Ontology-Based Approach to Data Cleaning. Technical report. University of Regina. Canada (2005)
20. Cannataro, M., Hiram Guzzi, P., Mazza, T., Tradigo, G., Veltri, P.: Using Ontologies in PROTEUS for Modeling Proteomics Data Mining Applications. *Studies in Health Technology and Informatics* 112 (2005) 17-26
21. Bernstein, A., Provost, F., Hill, S.: Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification. *IEEE Transactions on Knowledge and Data Engineering* 17(4) (2005) 503-18
22. Gottgroy, P., Kasabov, N. MacDonell, S.: An ontology driven approach for knowledge discovery in Biomedicine. *Proceedings of the VIII Pacific Rim International Conferences on Artificial Intelligence (PRICAI)*, Auckland, New Zealand (2004)
23. Svatek, V., Rauch, J., Flek, M.: Ontology-Based Explanation of Discovered Associations in the Domain of Social Reality. In: *ECML/PKDD05 Workshop on Knowledge Discovery and Ontologies*, Porto (2005)
24. Euler, T., Scholz, M.: Using Ontologies in a KDD Workbench. In *Workshop on Knowledge Discovery and Ontologies at ECML/PKDD* (2004)
25. McGuinness, D., van Harmelen, F. (eds.): *OWL Web Ontology Language Overview*. <http://www.w3.org/TR/owl-features/> (last accessed September. 1, 2006) (2003)

26. Knublauch, H., Fergerson, R.W., Noy, N., Musen, M.A.: The Protégé OWL Plugin: An Open Development Environment for Semantic Web Applications. In Third International Semantic Web Conference (2004)
27. Kalyanpur, A., Parsia, B., Sirin, E., Cuenca-Grau, B., Hendler, J.: Swoop: A web ontology editing browser. *Journal of Web Semantics* 4(2) (2005)
28. Volz, R., Oberle, D., Motik, B., Staab, S.: KAON server - a semantic web management system. *Proceedings of the 12th International Conference on World Wide Web (WWW2003). Alternate Tracks - Practice and Experience*, Budapest, Hungary (2003)
29. <http://www.es.embnnet.org/Services/MolBio/gepas/index.html> (last accessed September. 1, 2006)
30. <http://www.reactome.org/cgi-bin/frontpage> (last accessed September. 1, 2006)
31. <http://www.biomerieux.com/servlet/srt/bio/portail/home> (last accessed September. 1, 2006)
32. Sanandr s-Ledesma, J.A., Maojo, V., Crespo, J., Garc a-Remesal, M., G mez de la C mara, A.: A Performance Comparative Analysis Between Rule Induction-Algorithms and Clustering-Based Constructive Induction Algorithms. Application to Rheumatoid Arthritis. *ISMBDA* (2004)
33. Mart n-Sanchez F., Maojo V., L pez-Campos G.: Integrating genomics into health information systems. *Methods Inf Med.* 41 (2002) 25-30.
34. Maojo V., Mart n-Sanchez F.: Bioinformatics: towards new directions for public health. *Methods Inf Med.* 43(3) (2004) 208-14.
35. Maojo V., Kulikowski, C.A.: Bioinformatics and Medical Informatics: Collaborations on the Road to Genomic Medicine? *J Am Med Inform Assoc* 10(6) (2003) 515-22.

Language Modelling for the Needs of OCR of Medical Texts

Maciej Piasecki and Grzegorz Godlewski

Institute of Applied Informatics, Wrocław University of Technology,
Wybrzeże Wyspiańskiego 27, Wrocław, Poland
`maciej.piasecki@pwr.wroc.pl`

Abstract. In the paper different methods of construction of language models are discussed in relation to a corpora of medical texts written in an inflective language, namely Polish. The main result is the proposal of a method of language modelling which sequentially combines tri-grams of morphological base forms with tri-grams of words. The introduction of base form tri-grams increased the overall performance of the combined model, measured as the improvement in the accuracy of OCR of handwriting, as well, as the ability to generalisation. The latter was showed by using corpora of two different types as the training one and the test one. The detailed results of tests run on a large corpora of real life medical language are discussed in the paper. An experimental system of OCR of handwritten epicrisis utilising the proposed model is presented. The proposed language model decreases the overall error of the system by 64.2% (51% in the case of different types of corpora).

1 Introduction

Written medical documents like: accessory examinations, diagnoses, or epicrisis, are a specific but important kind of medical data. They are a rich source of information concerning diseases and applied treatments. This source is especially valuable when perceived from a statistical perspective of thousands of medical cases collected during many years. Only the basic data are stored in tables of a database, a lot of additional information can be extracted from textual descriptions. However, the older documents exist mainly in handwritten forms and are inaccessible for text mining techniques and statistical analyses.

The overall goal of our project is to automatise the process of transferring handwritten medical documents into electronic documents as far as possible. The ideal is the fully automatic procedure. However, all existing methods of OCR of unrestricted handwritten texts express significant level of errors [1,2] when limited only to the recognition on the level of letter signs i.e. recognising sequences of letters as comprising words and using only statistics of letters co-occurrences and a limited lexicon. The improvement can be achieved by applying a *language model* — describing “the distribution of sequences of ‘words’” [3] in a natural language. The language model can be used to predict the next word on the basis of the sequence of the previous words.

The goal of the work presented here was to construct a language model (henceforth LM) on the basis of a corpus of medical texts such that the constructed model maximises the level of reduction of the error of a word-level OCR system. The procedure of the construction of the language model was intended to be easily transferred from one medical corpus to the other with a minimal human effort, e.g. we want to avoid the necessity of introducing some additional annotation to the corpus. The OCR system supported by the constructed LM can be characterised as:

- off-line (static) — recognising cursive handwriting written on paper [1,2,4],
- and a very large vocabulary (possibly open) system — “tens of thousands of words”, called also unrestricted [2].

The only assumption concerning the language is that the documents come from a known domain and a known source, e.g. some hospital. We want to follow the *rejection approach* [5], in which the list of candidates as possible recognition for word positions in the text is gradually reduced by the application of the LM. The rejection approach allows for the separation of the word-level OCR module from the application of a LM. LMs are commonly applied tools in the area of speech recognition, but the number of applications in OCR is limited [1].

The medical texts in the collected corpus were written in an unrestricted Polish, but because of the limited domain, the number of different words is quite small, see Sec. 2. This makes our OCR system an intermediate case between an open and restricted (but still large) vocabulary system.

The Polish language is an inflective language with a large number of morphological forms derived from the same base form, e.g. up to 14 for a noun and 119 for a verb (including participles, gerunds etc.), see Sec. 2. The huge number of possible sequences of word forms can significantly decrease the quality of a LM. This problem is rarely discussed in literature. Existing approaches to inflective languages try to decompose word forms and predict their parts [6,7].

2 Corpus of Medical Texts and Tasks

For the needs of the construction of LMs a corpus (called KorMedIIS — the Polish acronym for “The Medical Corpus of the Institute of Applied Informatics”) [8] of electronic medical texts has been collected from the database of a hospital for which the prototype OCR system is being constructed. The collected texts belong to several categories: *initial diagnosis*, *medical history*, *objective examination*, *accessory examination*, *final diagnosis*, *treatment*, and *epicrisis*. This collection is rather a set of similar but slightly different corpora. The epicrisis is a short description of a patient stay in a hospital. It is being written when a patient is discharged from a hospital. A typical epicrisis includes larger passages of text, consists of several sentences or shorter phrases, reports some details of the patient stay and treatment, and often copies after the other documents. KorMedIIS includes presently 15 961 epicrisis — 1 373 741 words and 1 334 590 words in texts of the other types. These two main parts of KorMedIIS will

be further called, respectively: the *Corpus of Epicrisis* and the *Supplementary Corpus*. The collected electronic texts comes from the last few years (since the introduction of a integrated computer system) but can be treated as representative for the older handwritten texts of the same type for the given hospital, e.g. they possess an identical structure, and were written during similar procedures of treatment.

All texts have been processed by a *morpho-syntactic tagger* called TaKIPI [9]. The tagger first *annotates* each word with the morpho-syntactic description by the application of the morphological analyser Morfeusz [10] (it recognises 115 000 base forms, 1,7 millions of words) and then disambiguates the description choosing one (desirably, 1.03 on average) appropriate *morpho-syntactic tag* for a word. The morpho-syntactic tags are defined according to the standard of the IPI PAN Corpus [11]. Each tag consist of:

- a morphological *base form* of a word,
- its *grammatical class* (a more detailed division than Parts of Speech, there are 32 classes defined for Polish),
- and a sequence of *values* of the *grammatical categories*, e.g. case, number, gender etc.

There are 12 different possible categories. Each category has its set of possible values. A set of categories is assigned to each grammatical class. Sometimes the set is empty, e.g. in the case of conjunctions or adverbs. The order of categories in tags is fixed in relation to the grammatical class, that is why in the XML format of IPI PAN Corpus (and KorMedIIS, too) only values of the categories are stated — the position of a value is determined for a category. An example of the KorMedIIS description of a word is presented below:

```
<tok>
  <orth>kobiety</orth>
    <lex><base>kobieta</base><ctag>subst:sg:gen:f</ctag></lex>
    <lex><base>kobieta</base><ctag>subst:pl:nom:f</ctag></lex>
    <lex><base>kobieta</base><ctag>subst:pl:acc:f</ctag></lex>
    <lex><base>kobieta</base><ctag>subst:pl:voc:f</ctag></lex>
</tok>
```

There are 4 179 theoretically possible tags, but only 1 642 of them occur in the manually disambiguated part of IPI PAN Corpus. TaKIPI selects one tag for a word and in this way determines all three parts of the morphosyntactic information together. However the accuracy of the tagger as tested on the given medical texts is about 93% for all words (including non-ambiguous). In order to construct an error free pattern, a part of the corpus (2 127 epicrisis) was manually disambiguated by a linguist (230 638 words). Epicrisis are larger texts and contain typical elements from other types of the medical documents, including the terminology and the structure of sentences.

The words unrecognised by Morfeusz are annotated as *unknown* words. They form a significant part of the corpus, i.e. about 25%. The set of unknown ‘words’

includes: proper names (e.g. people, medicines, substances etc.), Latin words (or English, too), words with spelling errors (Polish or Latin), abbreviations (very often created *ad hoc*), numbers, alphanumeric symbols, units, ‘bullet-points’, etc. Thus they can be divided into two subclasses: unknown words (e.g. proper names, Latin words) and unknown non-words (e.g. words with spelling errors, symbols). The unknown words can be morphological forms of some base forms, as known words, that creates a possibility for some clustering.

The main task of the constructed OCR system is to transform a handwritten medical document into its electronic version. We assume that the contextual information concerning the type of the document (e.g. recognised from the shape of its form) can be accessible to the system, but is not used yet. We assume also that documents come from one institution (e.g. a hospital ward), that limits their domain, and that they are written in Polish with some possible addition of foreign words.

In order to test the quality of the LMs we constructed a modular version of OCR system. The system is divided into two parts:

- a *word classifier* [12]: isolated characters are recognized (*character level*) and next the results of character classification are used on the higher level (*word level*), where isolated words are recognized, also using soft classification paradigm — Hidden Markov Models (HMM) and Probabilistic Lexical Language Models (PLLM) are used here to improve recognition accuracy.
- a *language modelling level* in which we follow the rejection approach applying a LM to lists of *candidates* generated by the word level classifier, where a candidate is a possible recognition for a word position in the input text.

The simplified version of the word classifier used in experiments requires that word images are correctly segmented into isolated characters and that also words are clearly separated. However we do not deal here with the non trivial problem of word segmentation, but we rather focus our attention on issues related to the efficient construction of a LM for the domain of medical texts written in an inflective language e.g. Polish. Our experiments were performed using the set of texts constituting patient records stored in a real life medical information system. For our experiments we selected epicrisises. The acquired corpus consisted of 15961 texts stored as ASCII strings. The complete lexicon derived from the whole corpus contained more than 34 000 words. The set of texts was divided into the training part consisting of 12 691 epicrisises (1 006 146 words) and the testing part containing the remaining 3 600 epicrisises (367 595 words). These two parts of the Corpus of Epicrisises will be further called, respectively: the *Training Corpus of Epicrisises* and the *Test Corpus*.

Unfortunately, we were not able to collect a sufficiently numerous set of handwritten texts yet. Therefore a simulated experiment was performed, where text images were artificially created using the set of images of 5 080 hand-written characters, manually classified. The procedure of creation of text image consists of the following steps. First, the text to be recognized is randomly drawn from the Test Corpus. Next, for each character in a selected text one image of this character is randomly selected from the set of character samples. Finally, the

drawn character images are arranged side by side into an artificial text image. The real handwritten texts can be slightly different as they were written in the years preceding the creation of the electronic documents from KorMedIIS. However, according to the consistency in the style of work of the given hospital, applied procedures, standards, customs etc. we assume that the changes in the language are gradual in time, slow and concern mainly the set of proper names.

For each written word in the analysed text, the word classifier produces a list of the $k = 10$ most probable *candidate words* (henceforth *candidates*) together with their probabilities. An example of such a list is given below:

$\langle \text{actual} = \text{"KTÓRA"}, 10, \langle$
 $\langle \text{"KTÓRA"}, 0.467746 \rangle, \langle \text{"KTÓRE"}, 0.269696 \rangle,$
 $\langle \text{"KTÓRA"}, 0.254766 \rangle, \langle \text{"KTORA"}, 0.007757 \rangle,$
 $\langle \text{"STÓPW"}, 0.000035 \rangle, \langle \text{"STLEJ"}, 0.000000 \rangle,$
 $\langle \text{"SERCA"}, 0.000000 \rangle, \langle \text{"MOCZU"}, 0.000000 \rangle,$
 $\langle \text{"BADAÑ"}, 0.000000 \rangle, \langle \text{"LEKÓW"}, 0.000000 \rangle \rangle$

On the basis of the LM we want to choose the best candidate among the k candidates, or at least to reject the candidates inconsistent with the context of the surrounding words. The word classifier [12] achieved accuracy 85.9% of the correct recognition in relation to candidates with the highest probability. The correct word was among 10 best words selected by the soft word classifier in 96.6% of cases. These two numbers define the space for the improvement to be achieved by the application of the LM.

3 N-Gram Model

As the first one we applied a simple LM based on probabilities of tri-grams constructed directly from words — morphological forms. The probability of a word w_i in the text is defined conditionally and can be approximated by the *Maximum Likelihood Estimation* (MLE) as following [13,3]:

$$P(w_i | w_{i-2}, w_{i-1}) = \frac{P(w_{i-2}, w_{i-1}, w_i)}{P(w_{i-2}, w_{i-1})} = \frac{c(w_{i-2}, w_{i-1}, w_i)}{c(w_{i-2}, w_{i-1})} \quad (1)$$

where $P(w_{i-2}, w_{i-1}, w_i)$ is the probability of the sequence of words: $w_{i-2} w_{i-1} w_i$, $c(w_{i-2}, w_{i-1}, w_i)$ is the number of occurrences of a sequence of words in the training corpus.

In further processing we apply a *sentencer* (a part of TaKIPI) [9] segmenting sequences of words into sentences or, more precisely, pseudo-sentences, as the error of the sentencer is small, but significant. Anyway, pseudo-sentences are deterministically defined and henceforth, we will call pseudo-sentences simply sentences. Thus, we can assume that each sentence starts with a tri-gram: **none none** w_3 , where **none** represents an ‘empty’ word. In that way, according to the 2nd order Markov model, the probability of the whole pseudo-sentence of n words $w_1 \dots w_n$ is given by:

$$P(w_1, \dots, w_n) = P(w_3 | \text{none}, \text{none}) P(w_4 | \text{none}, w_3) \prod_{5 \leq i \leq n} P(w_i | w_{i-2}, w_{i-1}) \quad (2)$$

For a sentence of n words there are $k^n = 10^n$ different combinations of candidates. In order to efficiently search for the best combination of candidates we treat candidates as states in a stochastic Markov process. As candidates are possible words, the probabilities of state transitions can be estimated on the basis of n -grams. We used tri-grams. As we are interested in the most consistent sequence of candidates from the linguistic point of view we want to look for a sequence of candidates maximising the probability of the whole path across the candidates. Applying the HMM based model of a *trellis* [3,13] and a general scheme of Viterbi algorithm (however modified in order to calculate the maximal path, not maximal subsequent states), we calculate the best maximal path by the algorithm of the *Global Word Consistency* described below, let:

- $\mathbf{V}$ be a matrix of candidates of the dimensions: $n \times k$, where n — number of words in a sentence, k — number of candidates,
 - $\delta_j(i)$ — a function returning the maximal probability of the i -th candidate in the j -th step of the algorithm, i.e. for the j -th position in a sentence,
 - Ψ — a matrix of the dimensions: $n \times k$, where $\Psi(j, i)$ is a backward pointer (a number) to the preceding candidate in the maximal path ending in $\mathbf{V}(j, i)$.
1. $\delta(1, i) = P(\mathbf{V}(1, i) | \text{none}, \text{none})$
 2. $\delta(2, i) = \max_l P(\mathbf{V}(2, i) | \text{none}, \mathbf{V}(2, l)) \delta(1, l)$
 3. $\Psi(2, i) = \arg \max_l P(\mathbf{V}(2, i) | \text{none}, \mathbf{V}(2, l)) \delta(1, l)$
 4. $\delta(j+1, i) = \max_l P(\mathbf{V}(j+1, i) | \mathbf{V}(j-1, \Psi(j, l)), \mathbf{V}(j, l)) \delta(j, l)$
 5. $\Psi(2, i) = \arg \max_l P(\mathbf{V}(2, i) | \text{none}, \mathbf{V}(2, l)) \delta(1, l)$
 6. At the end, starting with the $\arg \max_i \delta(n, i)$ candidate we recover backwards from Ψ the maximal path leading to this candidate.

When constructing tri-grams from the open vocabulary corpus one immediately encounters the problem of data sparseness. We collected only $10^{-7}\%$ of the possible word tri-grams ($365\,288/32\,302^3$) from the training part of the corpus of epicrisis. Thus, we tested several methods of *smoothing*:

- Laplace's law [3, pp. 202],
- Lidstone's law [3, pp. 204] with $\lambda = 0,75$,
- *held out estimator* [3, pp. 206], where 10% of the training corpus was separated as the held out data,
- and the simple linear interpolation:

$$P(w_i | w_{i-2}, w_{i-1}) = \lambda_1 P_1(w_i) + \lambda_2 P_2(w_i | w_{i-1}) + \lambda_3 P_3(w_i | w_{i-2}, w_{i-1}) \quad (3)$$

The λ parameters have been calculated by the algorithm proposed in [14]. The methods of smoothing were tested with the help of the Global Word Consistency algorithm. The tri-grams were collected from the Training Corpus of Epicrisis minus 10% of held out data. The accuracy of the algorithm was tested on the Test Corpus. We achieved the following accuracy of recognition: pure MLE — 92.5%, Laplace's law — 92.82%, Lidstone's law — 92.82%, held out estimator — 91.53% and the simple linear interpolation — 92.43%. In the further experiments we were constantly using smoothing based on the Laplace's law.

As the Training Corpus of Epicrisises and the Supplementary Corpus express some differences in the language used and the structure of texts included, we performed most experiments using three training corpora, namely: the Training Corpus of Epicrisises (TCE), the Supplementary Corpus (SC) and both joined together (TCE+SC). In tests we were using exclusively the Test Corpus (TC) and the results of the application of word classifier to the artificially generated text images, see Sec. 2.

The results achieved by the application of word tri-grams and the Global Word Consistency algorithm are presented in Tab. 1. The *possible improvement* is calculated as the percentage of the maximal theoretically possible error reduction i.e. between 85.9% — the accuracy of the word classifier alone and 96.6% the maximal possible accuracy for the 10 candidates. The *coverage* is calculated as the percentage of the possible tri-grams collected.

Table 1. Word tri-grams and the Global Word Consistency algorithm

corpus	accuracy of recognition [%]	possible improvement [%]	coverage [%]
TCE	92.80	62.4	10^{-7}
TCE+SC	92.66	61	$4 \cdot 10^{-8}$
SC	90.7	41.6	—

The accuracy achieved by the LM of word tri-grams is surprisingly good in relation to the theoretically unrestricted language of TC. It means that the range of language constructions used in the epicrisises is quite limited. A LM model constructed for a set of epicrisises can be successfully applied to the other set of epicrisises from the same hospital. The drop in accuracy observed for the LM trained on SC is natural, but still the improvement is significant in relation to the simplicity of the LM.

4 Generalised Models

The decreased accuracy observed in the case of SC used as the training corpus in Tab. 1 shows that the LM built on word tri-grams lacks generality and even a small difference between the training and the test corpora can be significant for the accuracy. Moreover, the coverage in tri-grams is very low for the open vocabulary. The coverage can be increased by grouping words into classes. In the case of an inflective language like Polish, natural classes arise from the morphological features of word forms.

4.1 HMM Model Based on Classes of Ambiguity

After morphological analysis each word is described by a set of tags expressing the grammatical classes and values of grammatical categories possible for a given word. In the case of a disambiguated corpus, the set of tags is (mostly) reduced

to one. The different possible tags define different possible classes of words, but they are still very numerous, as in the Polish tagset (IPI PAN Corpus standard) there are 4 179 possible tags (but used only 1 642, see Sec. 2). An English tagset has 45–197 tags [3].

However, the worst problem with using tags as word classes in language modelling is that previously one has to prepare a reliable disambiguated corpus. In the case of KorMedIIS the manually disambiguated part is too small and includes only 230 638 words. The construction of a LM on the basis of automatically disambiguated corpus would transfer into the LM the stochastic model of errors of the tagger.

Thus, we decided to explore the possibility of constructing a class LM on the basis of *ambiguity classes*. The ambiguity class is a combination of grammatical classes and/or values of grammatical categories. Examples of classes of ambiguity are: { *adjective*, *noun* } — an ambiguity class defined only according to possible grammatical classes, or { number singular, *m1*, *m2*, *m3* } (possible male genders). A given ambiguity class groups together all words that are ambiguous in a similar way. Thus, they appear in similar morpho-syntactic contexts. While defining ambiguity classes we can freely select parts of tags, e.g. in the examples above only some features of tags are referred to in the definitions of the classes. Different definitions of the ambiguity classes give different granularities of division and different numbers of the classes. A LM based on ambiguity classes can be built from practically ‘any’ corpus, one needs only to morphologically annotate it by Morfeusz.

If we take into account all possible grammatical classes and the values of the all 12 grammatical categories, the number of the ambiguity classes is as large as the number of different tags, i.e. about 1600. That is why we started with the definition of ambiguity classes defined exclusively on the basis of grammatical classes. We will call such classes *PoS ambiguity classes*.

Our main intention was to combine a LM built on ambiguity classes with a word tri-grams model, but in order to make the initial assessment, first we constructed HMM [3,13] on the basis of tri-grams of PoS ambiguity classes:

- a *state* is a pair of PoS ambiguity classes, i.e. the PoS ambiguity classes of the two subsequent candidates,
- a *probability of transition* is estimated as a smoothed MLE probability: $P(K_i|K_{i-2}, K_{i-1})$, where K_i is the PoS ambiguity class of the i -th candidate in a sequence,
- an *observation/emitted symbol* from a state is a word,
- the probability of emission is calculated as MLE probability: $P(w_i|K_{i-1}, K_i)$, where w_i is the word on the i -th position.

All probabilities were smoothed by the Laplace’s law, according to the tests performed in Sec. 3. According to this HMM model, called μ , the probability of a sequence of candidates is equal to the probability of a sequence of observations O in relation to the model μ , i.e. $P(O|\mu)$. This probability can be calculated by a standard *forward procedure* applied on μ (e.g. [3, pp. 327]). In order to efficiently search all possible sequences of candidates we combined the scheme of Global

Word Consistence algorithm with the forward procedure. Avoiding cluttering the presentation with too many details, the general idea of the algorithm is:

- to go recursively across the *outer trellis* of candidates,
- to store for each path of candidates a *inner trellis* of HMM states (pairs of PoS ambiguity classes),
- and for each candidate to go internally across the inner trellises of HMM states calculating the probabilities of different paths of candidates according to the forward procedure (only the best paths of candidates are stored).

The results of the application HMM of PoS ambiguity classes are presented in Tab. 2.

Table 2. HMM model based on classes of ambiguity

corpus	accuracy of recognition [%]	possible improvement [%]	coverage [%]
TCE	45.35	-401.5	1.6
TCE+SC	44.67	-408.2	1.4
SC	43.53	-419.5	—

The results are very discouraging. The general conclusion is that the structures of Polish expressions are determined mainly not by positions of words in an expressions but by their morpho-syntactic *agreement* according to the values of grammatical categories. The applied HMM is blind to the this agreement. Next, we tried to build in a similar way HMM based on ambiguity classes defined on the basis of the values of the categories: *number*, *gender* and *case*. We achieved very similar results, however, quite different errors appeared: the agreement between neighbouring words was good, but the whole expressions looked as built quite randomly. The attempts to join both HMMs by defining ambiguity classes on the basis of all tag parts were stopped by the number of the generated classes and the resulting data sparseness.

4.2 N-Grams of Base Forms

Another possible form of clustering the word forms is according to the base forms shared across the word forms, e.g.

- chodzi* (*walks*), *chodzący* (*walked_{plural, female, progressive}*),
chodząca (*walk_{adjectival progressive participle, neutral gender/female, singular/plural}*),
chodzenie (*walking — gerund*) etc.
- all have the same base form: *chodzić* (*to walk*).

We collected tri-grams of base forms from the corpora automatically disambiguated by TaKIPI. The error of disambiguation of the base forms is much lower than the general error of tagging. The best sequences of candidates were calculated by a modified algorithm of the Global Base Forms Consistency:

1. Each candidate is exchanged on the list with all its possible base forms — very often some new positions are added to the list.

2. The scheme of the Global Word Consistency algorithm is applied to the lists of base forms of candidates; the probabilities are calculated as smoothed MLE: $P(b_i|b_{i-2}, b_{i-1})$, where b_i is a base form (often one of several) of the i -th candidate in a sequence.
3. The best candidates are chosen according to the best base forms, in case of several candidates sharing the same base form, the first one on the initial list is chosen.

The results of the application of the algorithm are presented in Tab. 3.

Table 3. Base form tri-grams and the Global Base Form Consistency algorithm

corpus	accuracy of recognition [%]	possible improvement [%]	coverage [%]
TCE	92.85	62.9	$1.522 \cdot 10^{-6}$
TCE+SC	92.7	61.4	$7.059 \cdot 10^{-7}$
SC	90.98	44.4	—

The achieved results are slightly better than the ones achieved for the word tri-gram LM, see Tab. 1. One can observe that the accuracy of the base form LM trained on SC is significantly better than the corresponding result of the word tri-gram LM trained on SC. In the case of differences between corpora, the generalisation achieved by the application of base forms is helpful.

4.3 Combined Models

In order to improve the result, we tried to combine both successful models, namely the word tri-gram LM and the base form tri-gram LM. However, taking into account the linguistic point of view, we built a two step mechanism of rejection and choice:

1. First, the Global Base Form Consistency algorithm is applied, but this time all candidates sharing the winning base form are preserved, the other ones are rejected.
2. All preserved candidates are restored on the shortened lists.
3. The Global Word Consistency algorithm is applied to the shortened lists of candidates.

The task of the first step is to recognise a typical sequences of words in a text regardless of their exact morphological forms. The first step is based on the generalised (by the means of base forms) stochastic information drawn from the corpus. Next, we apply a kind of ‘fine tuning’ trying to choose the proper morphological word forms according to the agreement expressed in the form of word tri-grams. The results of the combined algorithm are presented in Tab. 4.

We can observe the significant increase of the best result, still achieved for the training on the corpus similar to the test one. But the increase of the result

Table 4. Global Combined Consistency algorithm

corpus	accuracy of recognition [%]	possible improvement [%]
TCE	92.98	64.2
TCE+SC	92.79	62.3
SC	91.65	51

achieved for SC is relatively very high. It shows that the combination of the two steps increased the ability of the LM to make generalisation.

5 Conclusions

The language used in epicrisis is unrestricted but naturally limited by the repetitive character of the reported procedures or histories. This is the source of the good performance of the word tri-gram LM reported in Tab. 1, when it was trained on a corpus of epicrisis and tested on a corpus of epicrisis (there was no intersection of the two corpora). When we use a corpus of different medical texts (still coming from the same hospital ward), the result of the word tri-gram LM is significantly worse. Its ability to generalise is limited.

The PoS HMM LM is too general and records positional sequences of tags which do not describe well an inflective language like Polish. This finding is consistent with problems encountered in the application of stochastic methods of tagging to Polish [15]. The generalisation introduced by the base form tri-gram LM is of much smaller granularity and this model is closer to the lexical level of word co-occurrences. Its performance is comparative (Tab. 3) with the word tri-gram model, but in the case of a different corpus (i.e. SC) it is significantly better. The best result has been achieved by the sequential combination of the two lexical LM, see Tab. 4, where the relatively large increase is noticed for training on a different corpus (i.e. SC).

The proposed method of language modelling can be easily transferred to any other medical corpus, as the preparation of a training corpus requires only a ready to use tagger. In comparison to works on language modelling of inflective languages for the needs of OCR presented in [6,7], our approach is not limited to the recognition of singular words, but is focused on recognition of consistent larger expressions.

A large decrease in the final accuracy of our OCR system was created by unknown words. They are simply undistinguishable for our LMs. We plan to create a kind of ‘morphological guesser’ trying to assign some morphological annotation to unknown words on the basis of their affixes and prefixes. The guesser could also distinguish the unknown words from the unknown non-words or even to correct some simple spelling errors. The assigned morphological description could be next used in the additional tag-based LMs (lexically insensitive).

Acknowledgement. This work was financed by the Ministry of Education and Science project No 3 T11E 005 28.

References

1. Bunke, H.: Recognition of cursive roman handwriting - past, present and future. In: Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR'03). Volume 1., IEEE (2003) 448–460
2. Koerich, A.L., Sabourin, R., Suen, C.Y.: Large vocabulary off-line handwriting recognition: A survey. *Pattern Anal Applic* **6** (2003) 97–121
3. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. The MIT Press (2001)
4. Vinciarelli, A., Bengio, S., Bunke, H.: Off-line recognition of unconstrained handwritten texts using hmms and statistical language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(6) (2004) 709–720
5. Koerich, A.L.: Rejection strategies for handwritten word recognition. In: Proceedings of the 9th International Workshop on Frontiers in Handwriting Recognition October 26–29, 2004, Kokubunji, Tokyo, Japan. (2004)
6. Karacs, K., Prószéky, G., Roska, T.: Intimate integration of shape codes and linguistic framework in handwriting recognition via wave computers. In: Proceedings of the European Conference on Circuit Theory and Design 1 - 4 September 2003, Kraków, Poland. (2003)
7. Pal, U., Kundu, P.K., Chaudhuri, B.B.: Ocr error correction of an inflectional indian language using morphological parsing. *OCR Error Correction Journal of Information Science and Engineering* **16** (2000) 903–922
8. Piasecki, M., Godlewski, G., Pejcz, J.: Corpus of medical texts and tools. In: Proceedings of Medical Informatics and Technologies 2006, Silesian University of Technology (2006)
9. Piasecki, M., Godlewski, G.: Effective architecture of the polish tagger. [16]
10. Woliński, M.: Morfeusz — a practical tool for the morphological analysis of polish. [17]
11. Przepiórkowski, A.: The IPI PAN Corpus Preliminary Version. Institute of Computer Science PAS (2004)
12. Godlewski, G., Piasecki, M., Sas, J.: Application of syntactic properties to three-level recognition of polish hand-written medical texts. In Bulterman, D., Brailsford, D.F., eds.: Proceedings of the 2005 ACM symposium on Document engineering, New York, ACM Press (2006)
13. Jelinek, F.: Statistical Methods for Speech Recognition. The MIT Press (1997)
14. Brants, T.: TnT — a statistical part-of-speech tagger. In: Proceedings of the Sixth Applied Natural Language Processing Conference (ANLP-2000). Seattle. (2000)
15. Łukasz Dębowski: Trigram morphosyntactic tagger for Polish. In Mieczysław A. Kłopotek, Wierzchoń, S.T., Trojanowski, K., eds.: Intelligent Information Processing and Web Mining. Proceedings of the International IIS:IIPWM'04 Conference held in Zakopane, Poland, May 17–20, 2004. Springer Verlag (2004) 409–413
16. Sojka, P., Kopecek, I., Pala, K., eds.: Proceedings of the Text, Speech and Dialog 2006 Conference. Lecture Notes in Artificial Intelligence, Springer (2006)
17. Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K., eds.: Intelligent Information Processing and Web Mining — Proceedings of the International IIS: IIPWM'06 Conference held in Zakopane, Poland, June, 2006. Advances in Soft Computing. Springer, Berlin (2006)

The Use of Multivariate Autoregressive Modelling for Analyzing Dynamical Physiological Responses of Individual Critically Ill Patients

Kristien Van Loon¹, Jean-Marie Aerts¹, Geert Meyfroidt²,
Greta Van den Berghe², and Daniel Berckmans¹

¹ Division Measure, Model & Manage Bioresponses, Katholieke Universiteit Leuven,
Kasteelpark Arenberg 30, B-3001 Leuven, Belgium
daniel.berckmans@biw.kuleuven.be.

² Department of Intensive Care Medicine, University Hospital Gasthuisberg, Herestraat 49,
B-3000 Leuven, Belgium

Abstract. We attempted to find a way to distinguish survivors and non-survivors on the basis of the differences in the dynamics in both patient classes using multivariate autoregressive (MAR) time series analysis techniques. Time series data of 11 physiological variables were used to calculate MAR models. Data were taken from a subset of patients, with an intensive care unit length of stay of at least 20 days, from a database of a previously published randomized controlled trial [1]. The methodology was developed on 20 and validated on 16 patients. Based on the MAR coefficients, impulse response curves were simulated to describe the contributions of a single variable to fluctuations in another. The impulse responses of non-survivors had a tendency to be either more instable or to return to the initial level after a longer time than the responses of survivors did. This allowed us to distinguish survivors from non-survivors in the development cohort with a sensitivity of 0.70 and a selectivity of 1.00. This result was confirmed in the validation set where a sensitivity of 0.63 and a selectivity of 1.00 were reached.

Keywords: critical care, mortality, multivariate time series analysis, outcome prediction.

1 Introduction

From a modelling point of view, biological organisms can be regarded as complex, individually different, time varying and dynamic (CITD) systems that depend on the efficient operation and interaction of a number of regulatory systems to allow them to maintain homeostasis. This is an approximately constant state which varies only within tolerable limits. In humans, in conditions of sepsis, inflammation or critical illness, some regulatory mechanisms that maintain homeostasis in health, become dysfunctional or function in a different way. Each individual responds in a different way to changes in his/her environment or to changes in the set-point of homeostasis. Put another way, the amount of 'effort' to maintain homeostasis is different for each

individual and the individual responses to perturbations in homeostasis are difficult to predict. In general, outcome prediction in critical illness is based on population studies of large databases [2,3]. Risk factors for the outcome parameter (most often mortality) are calculated using traditional statistic methods such as multivariate analysis, or using more recent approaches such as artificial neural networks [4-7] or machine learning techniques [8-10]. Autoregressive modelling techniques approach the problem in a different way because they analyse data of individual systems. In the intensive care unit (ICU) environment, the multitude of clinical and monitoring data, generating a large number of time series of clinical parameters of individual critically ill patients, could be well suited as inputs for this methodology.

In the field of intensive care medicine, mainly univariate retrospective time series analysis of physiological variables [11,12] has been applied. Due to the wide number of regulatory processes and feedback mechanisms in sepsis and critical illness, a multivariate approach might be another valid method. In the late 1960's Akaike developed a practical method for the identification of a multivariate feedback system with the use of multivariate autoregressive (MAR) modelling [13,14].

This research is a first attempt to use multivariate autoregressive analysis techniques to study time series of daily measurements of clinical and laboratory data from individual critically ill patients. More specifically, we wanted to test whether the study of simulated impulse response curves, indicating dependencies and relationships between certain variables, would have different characteristics in patients who survived their ICU stay and patients who did not.

2 Materials and Methods

2.1 Patient Database

We used a database of 1548 patients from a previously published large randomized controlled trial to study the effects of intensive insulin therapy in critically ill patients [1]. The protocol of this trial was approved by the hospital institutional review board. Almost all adults receiving mechanical ventilation who were admitted to our ICU (which is dedicated primarily but not exclusively to surgical patients) between February 2, 2000, and January 18, 2001, were enrolled in this study after written informed consent had been obtained from the closest family member.

In order to have enough data points for time series analysis, we reduced this database by selecting only those patients with a length of stay of at least 20 days because most of the stored data were available only once daily. Next, this reduced database was divided into four subsets: a set of survivors that received intensive insulin treatment, a set of survivors that received conventional insulin treatment, a set of non-survivors that received intensive insulin treatment and a set of non-survivors that received conventional insulin treatment. For developing the method, 20 patients were selected by randomly picking out 5 patients from each subset. In this way, the used data sets were balanced out for intensive and conventional insulin treatment. An overview of the data sets used for development of the method is depicted in Table 1. To evaluate the developed method, 4 other patients were randomly selected from each subset. An overview of the data sets used for evaluating the method is depicted in table 2.

Table 1. Summary of the data used for developing the monitoring method

Survivors			Non-survivors		
Patient	Insulin treatment	Duration of stay (days)	Patient	Insulin treatment	Duration of stay (days)
1	Intensive	39	11	Intensive	34
2	Intensive	34	12	Intensive	28
3	Intensive	44	13	Intensive	43
4	Intensive	49	14	Intensive	42
5	Intensive	28	15	Intensive	32
6	Conventional	41	16	Conventional	25
7	Conventional	21	17	Conventional	122
8	Conventional	34	18	Conventional	110
9	Conventional	37	19	Conventional	40
10	Conventional	41	20	Conventional	64

Table 2. Summary of the data used for evaluating the monitored method

Survivors			Non-survivors		
Patient	Insulin treatment	Duration of stay (days)	Patient	Insulin treatment	Duration of stay (days)
21	Intensive	36	29	Intensive	24
22	Intensive	59	30	Intensive	22
23	Intensive	54	31	Intensive	25
24	Intensive	70	32	Intensive	22
25	Conventional	68	33	Conventional	62
26	Conventional	39	34	Conventional	92
27	Conventional	69	35	Conventional	76
28	Conventional	33	36	Conventional	36

For the data analysis, 11 dynamical variables were selected: minimum and maximum body temperature ($T_{\min}$, $T_{\max}$, $^{\circ}\text{C}$), minimum and maximum glycaemia ($\text{GLYC}_{\min}$, $\text{GLYC}_{\max}$, mg/dl), 24 hour urine output (U_{flow} , ml/day), white-cell count (WBC , $10^9/\text{L}$), BUN (mg/dl), plasma creatinine (CREAT_{pl} , mg/dl), total bilirubine concentration (BILI , mg/dl), C-reactive protein level (CRP , mg/dl) and total protein (P_{tot} , g/l). $T_{\min}$, $T_{\max}$, $\text{GLYC}_{\min}$ and $\text{GLYC}_{\max}$ are the extreme values of 6 temperature and glycaemia measurements that were taken every 4 hours every day. All other variables were measured only once daily, so the sampling frequency of all used signals is one sample a day.

2.2 Multivariate AR Models

A time series is a sequence of observations taken sequentially in time. An intrinsic feature of a time series is that, typically, adjacent observations are dependent [15]. Time series analysis is concerned with techniques for the analysis of this dependence. A common approach for modelling univariate time series is the autoregressive (AR) model [15]. An AR model is a linear regression of the current value of the series against p prior values of the series. The value of p is called the order of the AR model. These univariate AR models can easily be generalized to multivariate AR (MAR) models. The previous scalar values are replaced by vectors of simultaneously observed previous values and the corresponding coefficients are replaced by square matrices.

Suppose there are K variables of interest and that their values at time instant t are denoted by $x_i(t)$, $1 < i < K$. The general equation of the MAR in these variables is then given by

$$x_i(t) = \sum_{m=1}^M \sum_{j=1}^K a_{ij}(m) x_j(t-m) + e_i(t) . \quad (1)$$

where $a_{ij}(m)$ is a weighing coefficient and $e_i(t)$ is white noise for x_i .

In this research the parameters of the AR models are estimated using a stepwise least squares algorithm. With the computed AR coefficient matrices, Akaike's state space representation can be made which works as a kind of differential equation specific for the individual body under study. Using such an equation, one can simulate impulse response functions to describe the contributions of $x_j(t)$ to fluctuations of $x_i(t)$ in the time domain [16-18]. Equation (1) can be rewritten in a matrix form as

$$X(t) = \sum_{m=1}^M A(m) X(t-m) + E(t) . \quad (2)$$

Here, the matrices $A(m)$ are the AR coefficients and $E(t)$ is a prediction error vector: $E(t) = [e_1(t) \ e_2(t) \ \cdots \ e_K(t)]'$. The response of the closed loop system can then be simulated using the state space notation of the MAR model of equation (2), which is given by

$$\mathbf{Z}(t) = \mathbf{\Phi} \times \mathbf{Z}(t-1) + \mathbf{V}(t) . \quad (3)$$

$$\mathbf{X}(t) = \mathbf{H} \times \mathbf{Z}(t) . \quad (4)$$

where $\mathbf{Z}(t) = [\mathbf{X}(t) \ \mathbf{X}(t-1) \ \dots \ \mathbf{X}(t-M+1)]'$ and $\mathbf{Z}(t-1) = [\mathbf{X}(t-1) \ \mathbf{X}(t-2) \ \dots \ \mathbf{X}(t-M)]'$ are the state variables; $\mathbf{V}(t) = [\mathbf{E}(t) \ 0 \ \dots \ 0]'$; $\mathbf{H}(t) = [\mathbf{I} \ 0 \ \dots \ 0]$ and Φ is a so-called transitional matrix, which is expressed as

$$\Phi = \begin{bmatrix} \mathbf{A}(1) & \mathbf{A}(2) & \mathbf{A}(3) & \dots & \mathbf{A}(M-1) & \mathbf{A}(M) \\ \mathbf{I} & 0 & 0 & \dots & 0 & 0 \\ 0 & \mathbf{I} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \mathbf{I} & 0 \end{bmatrix}$$

The vector $\mathbf{Z}(t)$ starts from a zero matrix, simulating the system at standstill, and also $\mathbf{V}(t)$ is $\mathbf{0}$ in the beginning. Then an impulse is given to one of the variables $x_i(t)$ by putting a certain numerical value into the corresponding white noise term $e_i(t)$. By doing this, $\mathbf{V}(t)$ becomes different from $\mathbf{0}$. According to equation (3), $\mathbf{Z}(t)$ can be estimated from its previous step $\mathbf{Z}(t-1)$. Since no white noise is given any more, $\mathbf{Z}(t)$ can simply be computed as $\Phi \times \mathbf{Z}(t-1)$. The predicted values of the variables can be computed by multiplying the $\mathbf{Z}(t)$ with the inverse matrix $\mathbf{H}$ as shown in equation (4).

The same procedure can be used to calculate the one-step-ahead predictions $X(t-1)$ from the observed data $X(t)$ [16]. If all eigenvalues of Φ lie inside the unit circle, then $\mathbf{Z}(t) \rightarrow 0$ if $t \rightarrow \infty$ [19].

3 Results

As a first step, the order of the model had to be determined. Usually, the number of parameters that have to be estimated in each individual model fit should not exceed the limit of 10% of the number of data points [19]. We decided to violate this rule in this case (the number of parameters is 121 and the number of data points is 220 for a length of stay of 20 days) and to use a first order model because prediction accuracy was good, as can be seen in figure 1. The correlation coefficients between the measured variables and the one-step-ahead predictions (table 3) are a measure of how well the measured time series were modelled [20].

In the next step, all pulse responses of each variable to a pulse in another variable were determined in the way described above for the twenty patients. This resulted in eleven plots per patient. Although in some cases a negative impulse would be more meaningful for simulating actual clinical situations [18], in this study all pulses are given a positive direction. This is because the only effect of a giving the pulse the opposite direction is that the response has the reversed sign and that does not influence the remainder steps in our analysis.

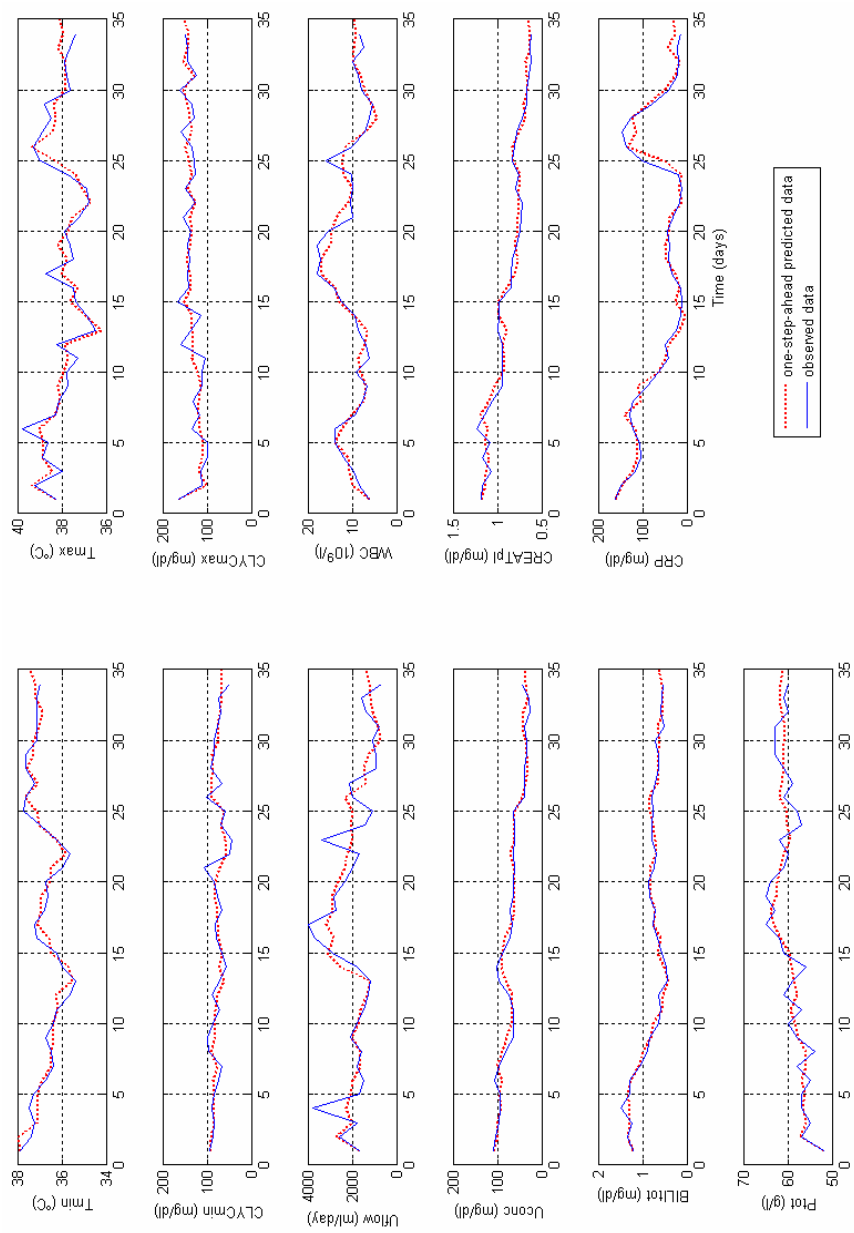


Fig. 1. The predicted data and the real measured data for patient 2. The dotted lines indicate the predicted values, the solid lines the actual data.

The zero levels in the curves correspond to the average levels of the measured time series data for each variable. A typical example of impulse response curves for a

Table 3. Overview of the correlation coefficients (means + SD) between the measured data and the one-step-ahead predictions averaged per variable over twenty patients

Variables	Correlation Coefficients
Tmin	0.78 ± 0.13
Tmax	0.78 ± 0.12
GLYCmin	0.66 ± 0.11
GLYCmax	0.67 ± 0.13
Uflow	0.80 ± 0.11
WBC	0.90 ± 0.06
Uconc	0.92 ± 0.10
CREATpl	0.93 ± 0.07
BILItot	0.93 ± 0.07
CRP	0.91 ± 0.06
Ptot	0.85 ± 0.10

survivor and a non-survivor are shown in figure 2 and figure 3 respectively. The pulses were given at time instant 2 and the responses are plotted until time instant 60.

For 4 of the 10 non-survivors we found some models for which one or more eigenvalues lied outside the unit circle. In some other cases, the impulse responses needed a long time to become zero again. These findings were used to make a distinction between survivors and non-survivors in the following way: if there is one impulse response curve for a given patient where the absolute value of the average of the last 5 values is more than $k\%$ of the maximum of the whole curve in absolute value, this patient is classified as non-survivor. After comparing the results for different values of k , we found that for this study the best results are obtained when k is a number of the interval [15, 25]. All choices for k in this interval led to the same result, namely a true positive fraction (sensitivity) of 7/10 and a true negative fraction (selectivity) of 10/10. The true positive fraction (TPF) is defined as the fraction of patients that are classified as non-survivors and died. The true negative fraction (TNF) is the fraction of patients that are classified as survivor and survived. Accordingly, if the condition to classify a patient as a non-survivor was satisfied, the patient effectively died. Not all impulse response curves of these non-survivors became instable or took a long time to become zero again, and the ones who did satisfy one of the two conditions were not necessarily the same for all patients. Three non-survivors were wrongfully seen as survivors, but it is more acceptable to classify a non-survivor as survivor than the other way around, so these results are acceptable.

After developing the method on 20 patients of the dataset, we validated it on a selection of 16 other patients. For this dataset there were no survivors for whom eigenvalues outside the unit circle were found during modelling. For 4 of the 8 non-survivors we did find eigenvalues outside the unit circle. The total result of the validation was a TPF of 5/8 and a TNF of 8/8.

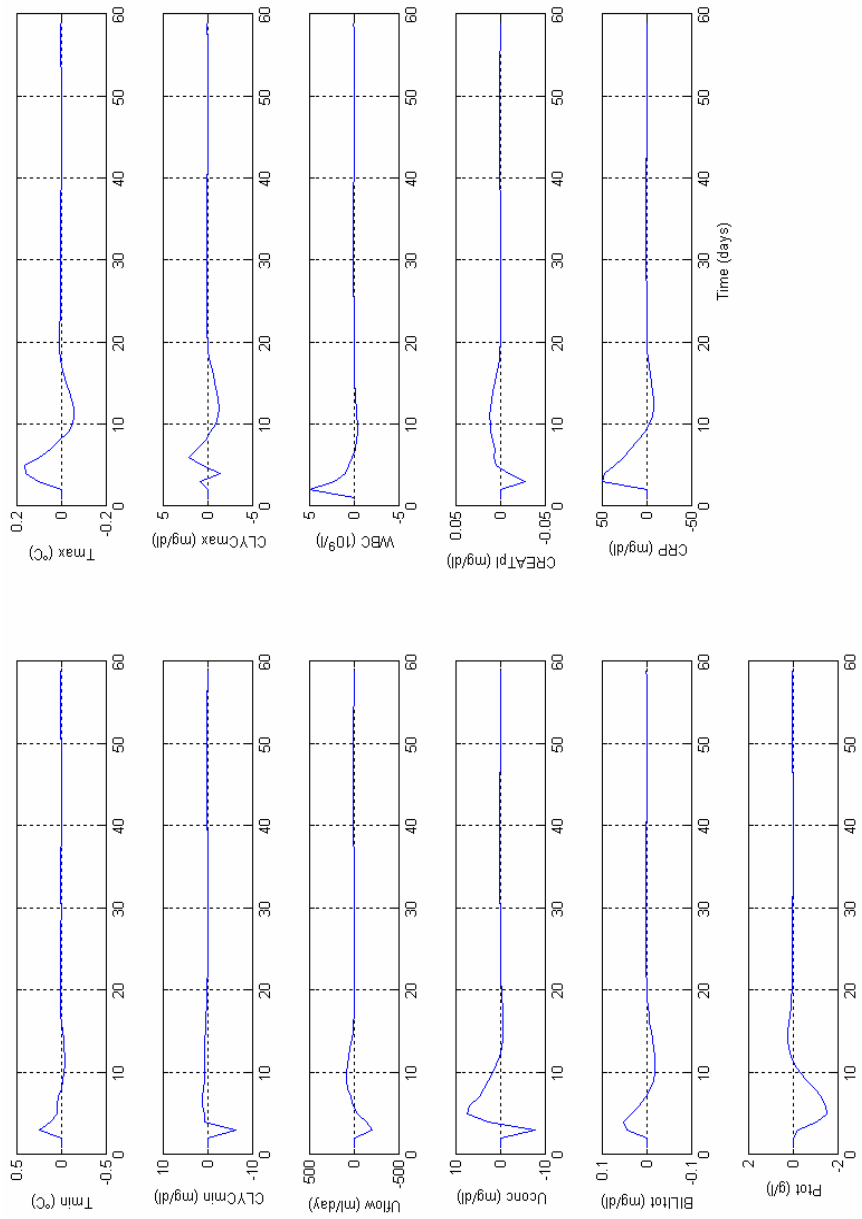


Fig. 2. The impulse response curves of patient 4 when an impulse was given to the WBC

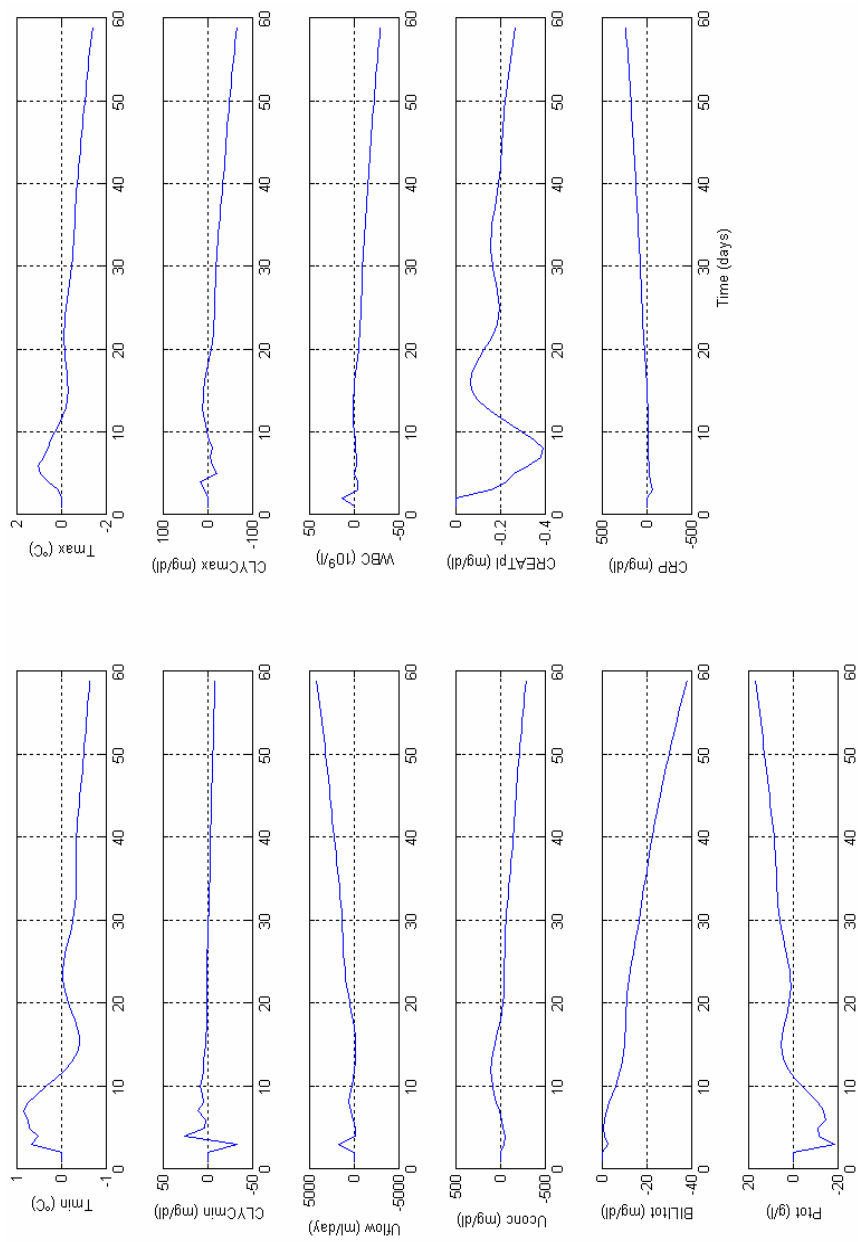


Fig. 3. The impulse response curves of patient 12 when an impulse was given to the WBC

4 Discussion

Akaike’s method is being applied before in the field of medicine and some studies have shown its clinical usefulness in different applications. In one of their studies,

Wada et al. made use of the method for analyzing immunologic networks in man [16]. They simulated the impulse and step responses in the $T4^+/T8^+/IgG$ system. By comparing the response curves of one control subject and one SLE patient, it became possible to understand the abnormality of networks in pathologic conditions. In another study they analyzed the T-lymphocyte subset fluctuations [17]. Here the impulse response curves were useful to demonstrate the features of malfunction of T-lymphocytes in the IgG regulation in different pathologic states. In contrast with the two healthy control subjects, an SLE patient had an instable IgG regulating system and the IgG regulation in the two rheumatoid arthritis and the three hemodialysis patients that were studied, showed a delayed response. The same group also studied the chloride/potassium/bicarbonate relationships in the body by comparing the impulse responses of different groups that all have an acid-base disturbance accompanying hypochloraemia and/or hypokalaemia [18]. In total, data of six humans and eleven dogs was used. This approach enabled them to differentiate between the roles of chloride and potassium in the development of metabolic alkalosis.

Miwakeichi et al. detected that the impulse response function based on a MAR model can differentiate focal hemisphere in temporal lobe epilepsy [19]. Five unilateral focal patients and one bilateral focal patient were studied. In order to detect the location of epileptic foci, linear MAR-models were fitted to electrocorticogram data and the behaviour of the corresponding impulse response functions was studied.

In none of the mentioned articles a quantification is given for how well the measured time series were modelled. Moreover, only in [16] a graphical comparison between the predicted and actual data is shown. Furthermore, these studies employed databases of limited numbers of patients. But although we used a larger set of patients, 20 for developing the method and 16 for validation purposes, the presented results should be confirmed in further studies employing larger databases. We also expect that the classification results for the validation set will be better when more data points of more patients are used in the development cohort. In addition it would be desirable to apply the developed method to data of another type of patients, e.g. patients of a medical intensive care unit, to investigate whether the approach may also be useful in that case.

By making use of MAR models and the corresponding impulse response functions, we studied the relationships and interactions between various parts of a biological system. This is in correspondence with the general concept of systems biology, where one tries to integrate different levels of information to understand how biological systems function. Systems biology offers the opportunity to bridge the gap between scales of description from the genome to the bedside and ultimately to health services research [21]. This will only be possible when computational, experimental and observational enquiries are combined [22].

The impulse response functions of the calculated MAR models give an idea of the dynamics of the system. We were able to distinguish between survivors and non-survivors on the basis of the impulse responses, so this means that the dynamics differ in both patient classes. The finding that changes in dynamics reflect and may even presage the development of clinical critical illness has been reported before [23-25].

A possible explanation for our finding that the impulse responses of non-survivors become unstable or need a long time to become zero again is the loss of the adaptive

capability to keep the critically ill patient in equilibrium in the period before he/she dies. Lipsitz reported that normally, when an organism is perturbed or deviates from a given set of boundary conditions, most physiologic systems evoke closed-loop responses that operate over relatively short periods of time to restore the equilibrium in the organism [26]. Disease can lead to maladaptive responses to perturbations, which subsequently can lead to dynamical instabilities and possibly to death. Probably some patients will be able to recover from an unstable situation and will not die, but on the other hand there will be patients that die without going through an unstable period (e.g. death after a heart attack). Thus far, we did not investigate the changes in the dynamics with time, because of the limited length of the time series data described in this work, but it would be an interesting future analysis to get more insight in the concept of dynamical instabilities.

Other techniques that can be applied to detect and quantify changes in the health condition of a patient are those based on nonlinear dynamics and chaos theory [23-24, 26-30]. Because of our limited dataset, these techniques could not be used during our research.

Very recently a patient data management system (PDMS) is brought into use which provides more high frequently measured physiological variables (e.g. one sample per minute). Using this high frequently sampled data, the described analysis can be repeated in time and a procedure for the real-time detection of changes in health conditions might be developed. It is a challenge to adapt our developed method to an on-line procedure for outcome predictions of individual critical care patients.

5 Conclusion

We developed an a posteriori method to classify a critically ill patient with an ICU stay of at least 20 days as a survivor or a non-survivor based on the impulse responses of eleven daily measured physiological variables calculated by Akaike's method. Because of the limited number of patients studied and the low sampling frequencies these findings are preliminary and will have to be validated in a larger patient group.

Acknowledgments. We wish to thank the Katholieke Universiteit Leuven for funding the research reported in this paper (Interdisciplinary Research project IDO/03/006).

References

1. Van Den Berghe, G., Wouters, P., Weekers, F., Verwaest, C., Bruyninckx, F., Schetz, M., Vlasselaers, D., Ferdinande, P., Lauwers, P., Bouillon, R.: Intensive Insulin Therapy in Critically Ill Patients. *New Engl J Med* 345 (2001) 1359-1367
2. Knaus, W.A., Draper, E.A., Wagner, D.P., Zimmerman, J.E: Apache II: a severity of disease classification system. *Crit Care Med* 13 (1985) 818-829
3. Knaus, W.A., Wagner, D.P., Draper, E.A., Zimmerman, J.E., Bergner, M., Bastos, P.G., Sirio, C.A., Murphy, D.J., Lotring, T., Damiano, A., Harrell, F.E: The APACHE III Prognostic system. *Chest* (1991) 1619-1636
4. Wyatt, J.: Nervous About Artificial Neural Networks. *Lancet* 346 (1995) 1175-1177

5. Dybowski, R., Weller, P., Chang, R., Gant, V.: Prediction of Outcome in Critically Ill Patients Using Artificial Neural Network Synthesised by Genetic Algorithm. *Lancet* 347 (1996) 1146-1150
6. Frize, M., Ennett, C.M., Stevenson, M., Trigg, H.C.E.: Clinical Decision Support Systems for Intensive Care Units: Using Artificial Neural Networks. *Med Eng Phys* 23 (2001) 217-225
7. Ennett, C.M., Frize, M., Charette, E.: Improvement and Automation of Artificial Neural Networks to Estimate Medical Outcomes. *Med Eng Phys* 26 (2004) 321-328
8. Sierra, B., Serrano, N., Larranaga, P., Plasencia, E.J., Inza, I., Jimenez, J.J., Revuelta, P., Mora, M.L.: Using Bayesian Networks in the Construction of a Bi-Level Multi-Classifier. A Case Study Using Intensive Care Unit Patients Data. *Artif Intell Med* 22 (2001) 233-248
9. Frize, M., Walker, R.: Clinical Decision-Support Systems for Intensive Care Units Using Case-Based Reasoning. *Med Eng Phys* 22 (2000) 671-677
10. Hanson, C.W., Marshall, B.E.: Artificial Intelligence Applications in the Intensive Care Unit. *Crit Care Med* 29 (2001) 427-435
11. Lambert, C.R., Raymenants, E., Pepine, C.J.: Time-Series Analysis of Long-Term Ambulatory Myocardial-Ischemia - Effects of Beta-Adrenergic and Calcium-Channel Blockade. *Am Heart J* 129 (1995) 677-684
12. Imhoff, M., Bauer, M., Gather, U., Lohlein, D.: Statistical Pattern Detection in Univariate Time Series of Intensive Care on-Line Monitoring Data. *Intens Care Med* 24 (1998) 1305-1314
13. Akaike, H.: On the use of a linear model for the identification of feedback systems. *Ann I Stat Math* 20 (1968) 425-439
14. Jones, R.W. (ed.): Principles of biological regulation: an introduction to feedback systems. New York, Academic Press Inc. (1973)
15. Box, G.E., Jenkins, G.M., Reinsel, G.C. (ed.): Time series analysis: forecasting and control. New Jersey, Prentice-Hall International (1994)
16. Wada, T., Akaike, H., Yamada, Y., Udagawa, E.: Application of Multivariate Autoregressive Modeling for Analysis of Immunological Networks in Man. *Comput Math Appl* 15 (1988) 713-722
17. Wada, T., Yamada, H., Inoue, H., Iso, T., Udagawa, E., Kuroda, S.: Clinical Usefulness of Multivariate Autoregressive (Ar) Modeling as a Tool for Analyzing Lymphocyte-T Subset Fluctuations. *Math Comput Model* 14 (1990) 610-613
18. Wada, T., Sato, S., Matsuo, N.: Application of Multivariate Autoregressive Modeling for Analyzing Chloride Potassium Bicarbonate Relationship in the Body. *Med Biol Eng Comput* 31 (1993) S99-S107
19. Miwakeichi, F., Galka, A., Uchida, S., Arakaki, H., Hirai, N., Nishida, M., Maehara, T., Kawai, K., Sunaga, S., Shimizu, H.: Impulse Response Function Based on Multivariate Ar Model Can Differentiate Focal Hemisphere in Temporal Lobe Epilepsy. *Epilepsy Res* 61 (2004) 73-8
20. Tschacher, W., Scheier, C., Hashimoto, Y.: Dynamical Analysis of Schizophrenia Courses. *Biol Psychiat* 41 (1997) 428-437
21. Clermont, G., Neugebauer, E.A.M.: Systems Biology and Translational Research. *J Crit Care* 20 (2005) 381-382
22. Kitano, H.: Computational Systems Biology. *Nature* 420 (2002) 206-210
23. Seely, A.J.E., Macklem, P.T.: Complex Systems and the Technology of Variability Analysis. *Crit Care* 8 (2004) R367-R384

24. Buchman, T.G.: Nonlinear dynamics, complex systems, and the pathobiology of critical illness. *Curr Opin Crit Care* 10 (2004) 378-382
25. Glass, L.: Synchronization and Rhythmic Processes in Physiology. *Nature* 410 (2001) 277-284
26. Lipsitz, L.A.: Dynamics of Stability: the Physiologic Basis of Functional Health and Frailty. *J Gerontol A-Biol* 57 (2002) B115-B125
27. Poon, C.S., Merrill, C.K.: Decrease of Cardiac Chaos in Congestive Heart Failure. *Nature* 389 (1997) 492-495
28. Ivanov, P.C., Amaral, L.A.N., Goldberger, A.L.; Havlin, S., Rosenblum, M.G., Struzik, Z.R., Stanley, H.E.: Multifractality in Human Heartbeat Dynamics. *Nature* 399 (1999) 461-465
29. Bruhn, J., Ropcke, H., Hoeft, A.: Approximate Entropy as an Electroencephalographic Measure of Anesthetic Drug Effect During Desflurane Anesthesia. *Anesthesiology* 92 (2000) 715-726
30. Pincus, S.M.: Approximate Entropy as a Measure of System-Complexity. *P Natl Acad Sci USA* 88 (1991) 2297-2301

Time Series Feature Evaluation in Discriminating Preictal EEG States

Dimitris Kugiumtzis¹, Angeliki Papanà¹, Alkiviadis Tsimpiris¹,
Ioannis Vlachos¹, and Pål G. Larsson²

¹ Aristotle University of Thessaloniki, Thessaloniki 54124, Greece
dkugiu@gen.auth.gr

<http://users.auth.gr/dkugiu>

² National Center for Epilepsy, 1303 Sandvika, Norway

Abstract. Statistical discrimination of states in the preictal EEG is attempted using a large number of measures from linear and nonlinear time series analysis. The measures are organized in two categories: correlation measures, such as autocorrelation and mutual information at specific lags and new measures derived from oscillations of the EEG time series, such as mean oscillation peak and mean oscillation period. All measures are computed on successive segments of multichannel EEG windows selected from early, intermediate and late preictal states from four epochs. Hypothesis tests applied for each channel and epoch showed good discrimination of the preictal states and allowed for the selection of optimal measures. These optimal measures, together with other standard measures (skewness, kurtosis, largest Lyapunov exponent) formed the feature set for feature-based clustering and the feature-subset selection procedure showed that the best preictal state classification was obtained with the same optimal features.

1 Introduction

Recent advances in the analysis of epileptic EEG have shown that the change from normal-like behavior (interictal state) to epileptic seizure (ictal state) is not abrupt but there exists a transition period (preictal state), ranging from minutes to hours [1]. Many works have shown evidence for detecting the preictal state from changes of measures of the local brain dynamics, such as multifractal spectrum based measures [2], entropy and symbolic dynamics complexity measures [3,4], and largest Lyapunov exponent [5]. However, subsequent studies have questioned the discriminating power of many of these measures [6,7,8]. In a different approach, changes were attempted to be detected from inter-regional coupling, using measures of synchronization [9,10], entropy transfer or mutual information between two channels, [3,11], and multivariate modeling [12,13]. Beyond the good rates of sensitivity and specificity reported on selected epochs, all the measures have been critically reviewed [14,15,16,17].

The vast majority of the studies use intracranial EEG that do not suffer from artifacts and focus on the seizure-related brain region. In many cases pre-selection of channels relevant to the seizure is done by a physiologist. The aim of

the present study is to assess the discriminating power of a large set of measures, i.e. changes in the preictal state, at the conditions of typical clinical practice, where intracranial and scalp EEG recordings are delivered without preprocessing (e.g. artifact removal). We concentrate on two types of measures of univariate time series analysis, i.e. measures based on correlation, where we also want to compare linear and nonlinear correlation measures used in the literature, and a new set of measures derived from oscillation features. We also select a subset of correlation and oscillation-based measures and evaluate their combined discriminating power using a feature-based clustering approach.

2 Materials and Methods

2.1 Materials

We used 4 EEG epochs, three from scalp EEG of a 10–20 system of 25 channels, referred to as A, B and C, and one from an intracranial EEG recording of 28 channels, referred to as D. All EEG data were band-passed at 0.5–70 Hz and sampled at 100 Hz (actually we used every second value from the data sampled at 200 Hz to speed up calculations). We selected data windows of 10 min duration from early, intermediate and late preictal states, corresponding to periods of approximately 4h, 1h and 10min before seizure onset and denoted as e , i , and l , respectively. Each 10min long data window was split to 20 successive segments of 30 s and the measures were estimated for each of them. For epoch A, we also considered larger data windows of 50 min of state e and the last 60 min of the preictal state (merging states i and l), segmented in the same way.

2.2 Correlation Measures

In our study we assess different correlation measures that have been used in the prediction of seizures and some new correlation measures. Let an EEG segment from a single channel be denoted as $\mathbf{x}$. As a measure of correlation in $\mathbf{x}$ we consider the autocorrelation $r_x(\tau)$ at lags $\tau = 1, 5, 10, 20, 30$ and the cumulative autocorrelation up to a maximum lag τ^* , $Q_x(\tau^*) = \sum_{\tau=1}^{\tau^*} |r_x(\tau)|$. The autocorrelation measure is sensitive to deviations from normal amplitudes and therefore we make the same computations on the transformed to Gaussian time series $\mathbf{y}$, where $y_t = \Phi^{-1}(F_x(x_t))$, Φ is the standard Gaussian cumulative density function (cdf) and F_x is the sample marginal cdf of $\mathbf{x}$. We refer to these measures as "normal" autocorrelations and denote them as $r_y(\tau)$, $\tau = 1, 5, 10, 20, 30$, and $Q_y(\tau^*)$.

The mutual information is an entropy-based measure of both linear and nonlinear correlation that has been applied to delays of a single EEG channels and between channels for the prediction of epileptic seizures [3,11]. We apply the mutual information $I_x(\tau)$ similarly to $r_x(\tau)$, i.e. for the same lags, and compute also the cumulative $I_x(\tau)$ denoted as $M_x(\tau^*)$.

In an attempt to provide a measure of solely nonlinear correlation we consider the difference of $I_y(\tau)$ (the mutual information on $\mathbf{y}$ that attains Gaussian

marginal cdf) from the normal mutual information (as if $\mathbf{y}$ was generated by a linear Gaussian process), defined as $I_y^g(\tau) = -\frac{1}{2} \log(1 - r_y(\tau)^2)$ [18]. The new measure $dI_y(\tau) = I_y(\tau) - I_y^g(\tau)$ is computed for the same lags and for the range of lags up to τ^* that gives the cumulative $dI_y(\tau)$ denoted as $dM_y(\tau^*)$.

We define as τ^* the lag at which $I_x(\tau)$ levels off. Computationally, the leveling of $I_x(\tau)$ is estimated when it first enters and stays (for three consecutive lags) at the zero-correlation area estimated from very large lags. Thus τ^* may vary across the EEG segments and therefore we use τ^* as an additional correlation measure. Moreover, we compute the mean of τ^* over all EEG segments of the epoch, $\langle \tau^* \rangle$, and consider also the cumulative correlation measures based on this, denoted as $Q_x(\langle \tau^* \rangle)$, $Q_y(\langle \tau^* \rangle)$, $M_x(\langle \tau^* \rangle)$, and $dM_y(\langle \tau^* \rangle)$.

2.3 Oscillation-Related Measures

The synchronization measures that have been applied recently in the prediction of seizures assume an oscillating behavior of the EEG signal (though this is not always true). Under the same assumption, we detect the oscillations in the EEG, draw features and compute measures based on these features (a similar approach was followed in [19]). The peak z_1 and valley z_2 of each oscillation are first detected as the local extremes by scanning $\mathbf{x}$ with a running data window of fixed length w and checking at each time step whether its center is a maximum or minimum. Then the time between successive peaks z_3 (oscillation period) and the time from valley to peak z_4 are derived giving 4 time series of oscillation features from $\mathbf{x}$. Statistics of location and scale from these data sets are considered as potential discriminating measures, namely the mean μ_i , the median m_i , the variance s_i^2 and the interquartile range denoted here as IQR_i , for $i = 1, 2, 3, 4$.

Further, we assume that there may be correlation in the feature time series and form a dynamic regression of each feature variable on delays of all four feature variables, i.e.

$$z_{i,t+1} = a_0 + a_1(B)z_{1,t} + a_2(B)z_{2,t} + a_3(B)z_{3,t} + a_4(B)z_{4,t} \quad (1)$$

where i denotes any of the four oscillation features, a_0 is the constant term and $a_1(B)$, $a_2(B)$, $a_3(B)$, $a_4(B)$ are polynomials of the back-shift operator B . We set the order for all polynomials fixed to one according to a pilot study that showed no fit improvement for larger orders (using a Granger causality and a BIC criterion). Opposite to other studies that assume a multivariate autoregressive model for the multichannel EEG [12], we apply here the same model but on the multivariate time series of oscillation features from a single channel.

We consider also the local linear dynamic regression model under the assumption that data points reconstructed from the oscillation features time series have a local structure. Each reconstructed point at time t is comprised of all feature values at times t and $t-1$ (thus the embedding dimension is 8), where all features are first standardized (here a pilot study showed that 8 is a sufficient embedding dimension). Then the prediction of each feature at time $t+1$ is the average of the one step ahead mappings of the 30 closest neighbors of the point at time t . For

both linear and nonlinear dynamic regression the fit of the models for each of the four features is quantified with the normalized mean square error (NMSE), and denoted as El_i and Enl_i , respectively, giving eight more measures. All the measures were computed twice, for features extracted with a data window of length $w = 7$ and $w = 15$.

2.4 Evaluation of Discriminating Power of Measures

The one-way ANOVA test for the means, as well as the Kruskal-Wallis test for the medians, were applied to measure samples from 10 min recordings at the three preictal states for all epochs. Significant difference in at least one pair of the e, i, l preictal states was found with both tests and most measures for all epochs. Therefore detailed results on these tests are not presented. with most of the measures. In line with the follow-up comparisons, we applied the Student test for the mean (equal variances not assumed) and the Wilcoxon rank sum test for the median on the three pairs of preictal states. In order to derive the discriminating power of each measure over all channels, we assigned a score s_q to each measure q for each preictal state pair comparison and epoch as follows.

1. Let p_{qj} , $q = 1, \dots, n$ and $j = 1, \dots, m$ be the p -values of the test for two preictal states of one epoch for all n measures and m channels.
2. Set $p_{qj} = 0$ if $p_{qj} > \alpha$ for a predefined significance level α . For each channel j sort the non-zero p -values at decreasing order and set a score s_{qj} for each q equal to its rank r_{qj} , i.e. if there are n_j non-zero entries for channel j the measure q with the lowest p -value gets the largest score for this channel, equal to n_j . The score of each measure q at each channel j is thus defined as

$$s_{qj} = \begin{cases} r_{qj} & \text{if } p_{qj} < \alpha \\ 0 & \text{if } p_{qj} \geq \alpha \end{cases}$$

3. Average the scores of each measure q over all channels

$$s_q = \frac{1}{m} \sum_{j=1}^m s_{qj}.$$

In this way the score of the performance of each measure in each channel is balanced by the discriminating power of the other measures in the channel. So, a channel that appears to be less relevant to the preictal activity will have less effect on the score of the measures. Finally, the average performance score S_q for each measure is defined as the average of the scores s_q over all three preictal state comparisons and all four epochs.

2.5 Clustering

Besides the evaluation of the discriminating power of each of the correlation and oscillation-related measures, we want to assess whether the combination

of several measures can give better discrimination of the preictal states. For this we let a feature-based clustering algorithm pick-up the feature subset from a set of useful selected measures that gives the best partition of the samples in the epoch (e.g. when the samples are derived from early and late preictal state).

For the clustering we use the k -means partitioning algorithm, the Expectation-Maximization (EM) algorithm and a linear standardization of the feature values. To compare the computed clusters to the original groups of preictal samples we use the similarity measure Corrected Rand Index (CRI) [20]. CRI ranges from -1 to 1 and the closer CRI is to 1 the better is the agreement of the two partitions.

In the computation of the optimal feature subset we adopt the algorithm of Forward Sequential Selection (FSS) of features [21]. Starting with the best single feature clustering, the feature subset is augmented by adding a single feature at a time, but only if the partition accuracy is significantly improved (we require at least 5% increase of CRI). Thus we retain small cardinality of the optimal feature subset by punishing the inclusion of features that give marginal improvement in CRI. This clustering approach was proposed very recently for selecting optimal features of oscillating time series [22].

3 Results

We estimate all the correlation measures and the oscillation-related measures on the 20 multi-channel EEG segments at each of the three preictal states (e, i, l) and for each of the four epochs (A,B,C,D). Sample sets of 20 measure values were selected randomly from the whole ensemble and the estimated autocorrelation showed no evidence of serial dependence. The variance and shape of the distribution for each measure changed a lot across the sample sets, also due to the existence of artifacts that were not removed. Besides the aforementioned violations of hypothesis testing assumptions, we applied hypothesis tests for the means and medians on the samples of 20 measure values from each group (preictal state) in order to discriminate the three preictal states.

3.1 Evaluation of Single Measures

The Student test and the Wilcoxon rank sum test gave similar results, but there was large variation of the test results across preictal state comparisons, channels, measures, and epochs, as expected. In Fig. 1 the Student test results are shown for all three pair comparisons with all measures and for the epoch B. For this epoch it seems that the late preictal state is distinguished well and with many measures whereas the early and intermediate preictal states do not bear a clear difference. This can be observed in all channels using the oscillation-related measures, but mostly in the right temporal, occipital and middle brain areas when using the correlation measures. Most interesting for our study is that

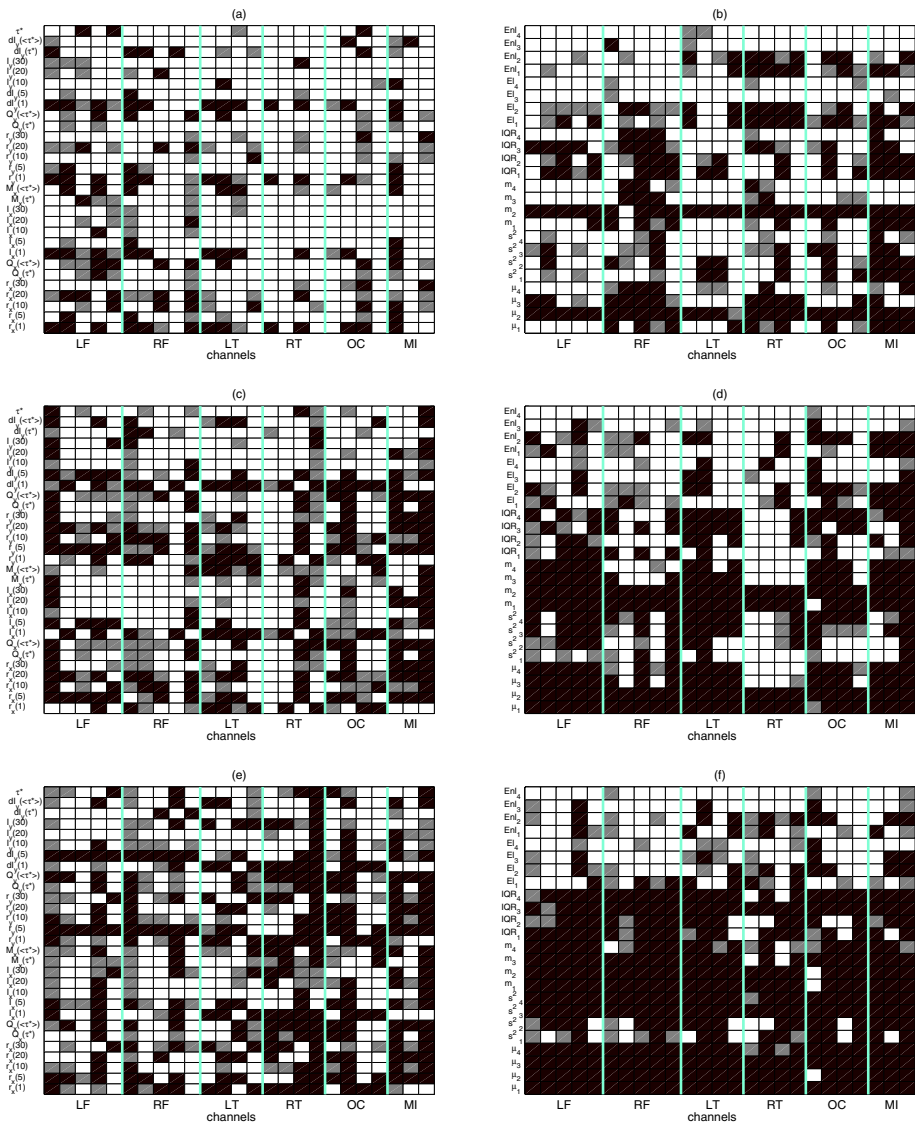


Fig. 1. The Student test results for epoch B based on the p -values: white cells when $p \geq 0.5$, grey cells when $0.01 \leq p < 0.05$, and black cells when $p < 0.01$. The channels are organized in brain areas as left and right frontal (LF and RF), left and right temporal (LT and RT), occipital (OC) and middle (MI). The correlation and oscillation measures are shown for the $e-i$ comparison in (a) and (b), respectively, for the $e-l$ comparison in (c) and (d), and for the $i-l$ comparison in (e) and (f).

the oscillation-related measures give better discrimination, and also for the $e-i$ comparison, where the correlation measures seem to fail (see Fig. 1a and b).

Moreover, for both measure types, some measures seem to detect better and more consistently the differences across channels.

The performance of the measures varied in the other three epochs, so that conclusive results could not be obtained by simple eye-ball judgement. Therefore we report summary results of the average performance score S_q . The 10 best S_q from the p -values of the Student test in the set of 29 correlation measures, the set of 24 oscillation-related measures, and the merged set of all 53 measures, are shown in Table 1.

Table 1. The 10 best measures with their scores S_q in the set of correlation measures, the set of oscillation-related measures, and the merged set of all measures taken over all pairs of preictal states and epochs

type		Measure and score									
correlation	q	$r_y(5)$	$Q_y(\langle \tau^* \rangle)$	$r_x(5)$	$r_y(10)$	$Q_x(\langle \tau^* \rangle)$	$r_x(10)$	$r_y(20)$	$r_x(20)$	$r_y(30)$	$dI_y(5)$
	S_q	76.8	72.0	69.6	66.1	65.6	59.9	57.5	50.2	47.6	47.0
oscillation	q	m_2	μ_2	m_1	μ_1	μ_3	μ_4	m_3	s_3^2	IQR_3	IQR_2
	S_q	59.6	57.3	56.0	52.1	47.6	35.6	28.0	27.7	26.8	26.2
all	q	m_1	m_2	$r_y(5)$	μ_2	$Q_y(\langle \tau^* \rangle)$	$r_x(5)$	μ_1	$Q_x(\langle \tau^* \rangle)$	$r_y(10)$	μ_3
	S_q	124.7	120.6	119.5	111.6	110.2	108.8	106.8	101.0	97.7	90.4

For the oscillation measures the shown results are for $w = 15$. For $w = 7$ the results were similar. The scores cannot be compared across the rows of the table as the score scale depends on the number of measures. For the correlation measures, we note that the autocorrelation measures outperformed the mutual information measures, with the "normal" autocorrelation measures scoring always better than the respective autocorrelation measures on the original time series. For the oscillation-related measures the median and mean values of the oscillation features, and in particular the local minimum and local maximum, gave the best discriminating performance. The measures of dynamic regression fit performed worst for all epochs (see Fig. 1 for epoch B).

When we compared the discriminating performance of all the measures together, the median local maxima and local minima scored best with the "normal" autocorrelation at lag 5 following close. The overall results show that the average local minima and maxima and the "normal" autocorrelation for small lags (5 and 10) as well as the cumulative autocorrelation (but only when a fixed τ^* is used) have the best discriminating power for the setting of preictal comparisons we used here.

The performance of the measures appeared to be dependent on the choice of the time window in each preictal state. To investigate this effect we made the same analysis on 5 data windows of 10min each for e (± 20 min to the initial selected 10 min data window), and 5 data windows of 10 min each for i (from 60 min to 10 min prior to seizure onset) from epoch A. The measures were

estimated on successive segments of 30 s duration and the tests were applied on samples of size 20 as before. The measure profiles varied across channels as shown in Fig. 2 for two adjacent channels in the middle brain area and the measures m_1 and $r_y(5)$. There seems to be a change in the profile of both measures and for both channels, with m_1 giving better discrimination of e from i and l .

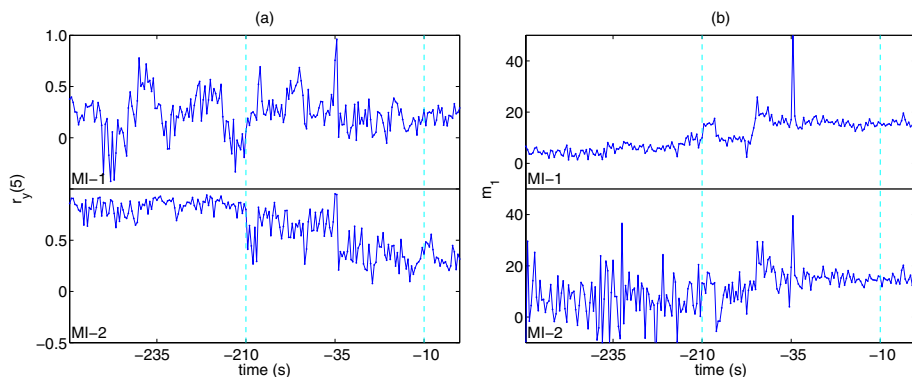


Fig. 2. Measure profiles for two channels in the middle area of the brain (upper and lower panel) over an early preictal state ($[-260,-210]$ min with respect to seizure onset) followed by an intermediate and late preictal state ($[-60,-10]$ min and $[-10,0]$ min), separated by vertical lines. In (a) the measure is $r_y(5)$ and in (b) m_1 .

We applied the Student tests with all measures on all 25 combinations of the 5 data windows in e and i for the A epoch. The average performance scores S_q were computed on basis of the 25 cases and the best 10 correlation measures, oscillation measures, as well as the overall 10 best measures are shown in Table 3.1. The results are in agreement to those in Table 1. The "normal" autocorrelation measures are still the best among the correlation measures (with $r_y(10)$ now being the best) and the average local maxima among the oscillation-related measures and among all measures, too. Here also the mutual information for lag 1 is among the best correlation measures and the linear and nonlinear dynamic regression fit on local maxima among the 10 best oscillation-related measures.

The analysis on the selected data windows from the e, i and l states from 4 epochs and on several data windows from the e and i states from one epoch suggests that the measures with the best discriminating power are the average (median first and then mean) of the local maxima (and minima) and the "normal" autocorrelation at small lags as well as the cumulative "normal" autocorrelation.

Table 2. As for Table 1 but for the 25 e - i comparisons of the epoch A

type		Measure and score									
correlation	q	$r_y(10)$	$r_y(5)$	$r_x(10)$	$Q_y(\langle\tau^*\rangle)$	$r_x(5)$	$I_x(1)$	$r_y(1)$	$r_y(20)$	$dI_y(10)$	$Q_x(\langle\tau^*\rangle)$
	S_q	166.5	145.4	134.5	133.6	131.5	130.6	115.6	111.5	108.6	105.8
oscillation	q	m_1	μ_1	μ_3	m_2	μ_2	μ_4	m_3	El_1	IQR_1	Enl_1
	S_q	153.6	119.2	63.3	58.0	51.7	50.8	49.6	41.3	39.9	36.1
all	q	m_1	μ_1	$r_y(10)$	$r_y(5)$	$I_x(1)$	$r_x(10)$	$Q_y(\langle\tau^*\rangle)$	$r_x(5)$	$r_y(1)$	$dI_y(10)$
	S_q	338.7	251.8	232.2	207.2	191.7	191.0	190.4	185.7	177.0	156.2

3.2 Clustering Accuracy of Feature Set

We want to assess whether the measures studied above can have a combined discrimination power when they are used together in a feature-based clustering. We form a feature set of 18 features, comprised of 8 correlation measures ($r_y(5)$, $r_y(10)$, $r_y(20)$, $Q_y(\langle\tau^*\rangle)$, $I_x(5)$, $M_x(\langle\tau^*\rangle)$, $dI_y(5)$ and τ^*), 6 oscillation-based measures (m_1 , m_2 , m_3 , IQR_1 , IQR_2 , and IQR_3), two scalar higher order moments (skewness λ and kurtosis κ), the bicornelation (or third order moment) at lags 1 and 2, denoted r_3 , and the largest Lyapunov exponent λ_1 estimated from the Lyapunov spectrum as implemented in the TISEAN package [23]. In this way we want to investigate whether other features, possibly with less discriminating power, can add to better clustering accuracy.

For each clustering problem, we give as input to the clustering algorithm the features estimated on the 20 time series of each preictal state and the number of clusters to be formed. For example, for the discrimination of e and i we form a data base of 40 records of features (20 records for each preictal state) and the algorithm gives out the optimal feature subset that attains best clustering accuracy, i.e. the estimated partition classifies best the original groups. This procedure was applied for all clustering tasks (groups (e,i) , (e,l) , (i,l) and (e,i,l)), all channels and all epochs.

A summary of the best clustering results is presented in Table 3. The clustering accuracy attained with only one or two features was high for all epochs and especially for epochs A and B that regard scalp recordings with artifacts as opposed to the intracranial recording of epoch D. There is a large variation of feature occurrence in the feature subsets that gives the best CRI across channels, as can be seen in the third column of Table 3. To get a better picture of the most useful features for clustering, the frequency of occurrence of features in the selected feature subset of cardinality 1 to 4 is shown in Table 4. Again, the median local minima and maxima turn out to be included most of the times in the optimal feature subset followed by the "normal" autocorrelation for lag 5. Thus the findings of the evaluation of the discriminating power of the measures are validated also from clustering.

Table 3. The best clustering results for each epoch and clustering task, as given in the first column. For each case the best CRI over all channels is given in the second column and in the third column the channel is given followed with the corresponding feature subset in brackets. In case the best CRI is attained for more than one channels they are all listed in the third column.

clusters	CRI	channel [features]
epoch A		
(e, i)	1.00	15[m_1], 16[m_1], 19[m_1]
(e, l)	1.00	many channels, features: $m_1, r_y(10), r_y(5), I_x(5)$
(i, l)	0.81	19[$m_1, Q_y(\langle \tau^* \rangle)$]
(e, i, l)	0.80	13[$m_1, r_y(10)$]
epoch B		
(e, i)	0.72	14[$m_2, dI_y(5)$], 20[m_2]
(e, l)	1.00	many channels, features: m_1, m_2
(i, l)	1.00	many channels, features: $m_1, r_y(5)$
(e, i, l)	0.86	14[m_2], 20[m_2]
epoch C		
(e, i)	0.72	15[r_3, m_1, λ_1]
(e, l)	0.55	17[λ_1], 24[$r_y(20), \text{IQR}_3$]
(i, l)	0.55	4[m_2, λ_1, κ]
(e, i, l)	0.45	15[$m_1, r_y(10), r_y(20), \tau^*$]
epoch D		
(e, i)	0.81	17[$dI_y(5)$], 18[$I_x(5)$]
(e, l)	0.72	2[$I_x(5), \text{IQR}_3, Q_y(\langle \tau^* \rangle)$]
(i, l)	0.55	17[$r_y(10), \lambda, M_x(\langle \tau^* \rangle)$], 20[$r_y(20), M_x(\langle \tau^* \rangle)$]
(e, i, l)	0.53	2[$r_y(5), \text{IQR}_3, r_y(20), dI_y(5), \text{IQR}_2$]

Table 4. Frequency of occurrence of features in the selected feature subset of cardinality 1 to 4. The frequency of a subset of a given cardinality is shown in column two and in column three the most frequent features are shown with the respective frequency in brackets.

subset	frequency	feature/frequency
1 feature	207	$m_1[71]$, $m_2[39]$, $r_y(5)[14]$, $r_y(10)[12]$, $r_y(20)[10]$, $m_3[9]$
2 features	130	$m_1[41]$, $m_2[29]$, $r_y(5)[29]$, $r_y(10)[18]$, $r_y(20)[19]$, $\text{IQR}_3[12]$, $\tau^*[11]$
3 features	58	$m_1[22]$, $m_2[24]$, $r_y(5)[13]$, $r_y(10)[7]$, $r_y(20)[9]$, $dI_y(5)[9]$, $\lambda[9]$
4 features	16	$m_1[8]$, $m_2[6]$, $r_y(5)[5]$, $r_y(10)[5]$, $r_y(20)[2]$, $m_1[4]$, $r_3[4]$

4 Discussion

The evaluation of a number of correlation and oscillation-related measures in detecting statistically significant differences between early, intermediate and late preictal states, showed that simple and computationally effective measures, such as the average of local maxima and minima, perform better than other well-known computationally intensive measures, such as the mutual information.

Certainly, the evaluation of the measures is not complete and comparison to other measures that are reported to discriminate preictal states is missing here. For example, we considered the largest Lyapunov exponent only for clustering, but still its absence in the optimal single feature subsets suggests worse performance than the optimal correlation and oscillation-related measures.

The analysis was done on small data windows pruned to artifacts that could affect the results on the measure evaluation. However, when the same analysis was done on larger data windows for one epoch the results were about the same.

The EEG records were analyzed without preprocessing or physiological expert knowledge about the type of epilepsy seizure and the active channels. Therefore we used an automatic scoring approach for the evaluation of the measures that downweighs the effect of channels when they do not exhibit changes in the measure values across the preictal states. We do not claim that this is the best way to score the measure performance but this task is inherently difficult as there are many factors to account for, i.e. measure related parameters (such as the running time window for the extraction of local extremes), channels, preictal states and epochs. We have tried alternative scoring schemes and they all tend to give the same results. Moreover, the scores from the Student tests and the feature subsets from clustering show a rather consistent tendency as to the performance of the measures ranking the average of local minima and maxima first and then the "normal" autocorrelation for lag 5, followed by the cumulative "normal" autocorrelation.

With regard to clustering, the high CRI values show that, given an ensemble of EEG recordings, one can use suitably selected features to identify the different preictal states. This approach may be further used to classify a freshly recorded EEG signal to one of the two or three classes, including also the interictal (normal-like) class that was not considered here.

Acknowledgments. This work was supported by the joint grant No 03ED748 from the General Secretariat of Research and Technology, Greece, and the National Center for Epilepsy, Norway.

References

1. E. Hirsch, F. Andermann, P. Chauvel, J. Engel, F. Lopes da Silva, and H. Luders. *Generalized Seizures: from Clinical Phenomenology to Underlying Systems and Networks*. Elsevier, Paris, 2006.
2. I.-H. Song, S.-M. Lee, I.-Y. Kim, D.-S. Lee, and S.I. Kim. Multifractal analysis of electroencephalogram time series in humans. *Lecture Notes in Computer Science*, 3512:921–926, 2005.
3. M. Paluš, V. Komárek, T. Procházka, Z. Hrnčíř, and K. Šterbová. Synchronization and information flow in EEGs of epileptic patients. *IEEE Engineering in Medicine and Biology Magazine*, 20(5):65–71, 2001.

4. R. Steuer, W. Ebeling, T. Bengner, C. Dehnicke, H. Hättig, and H.-J. Meencke. Entropy and complexity analysis of intracranially recorded EEG. *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering*, 14(2): 815–823, 2004.
5. L. D. Iasemidis, D.-S. Shiau, P. M. Pardalos, W. Chaovalitwongse, K. Narayanan, A. Prasad, K. Tsakalis, P. R. Carney, and J. C. Sackellares. Long-term prospective on-line real-time seizure prediction. *Clinical Neurophysiology*, 116(3):532–544, 2005.
6. Y.-C. Lai, M. A. F. Harrison, M. G. Frei, and I. Osorio. Controlled test for predictive power of Lyapunov exponents: Their inability to predict epileptic seizures. *Chaos*, 14(3):630–642, 2004.
7. T. Maiwald, M. Winterhalder, R. Aschenbrenner-Scheibe, H. U. Voss, A. Schulze-Bonhage, and J. Timmer. Comparison of three nonlinear seizure prediction methods by means of the seizure prediction characteristic. *Physica D*, 194(3-4):357–368, 2004.
8. D. Kugiumtzis and P. G. Larsson. Linear and nonlinear analysis of EEG for the prediction of epileptic seizures. In K. Lehnertz, J. Arnhold, P. Grassberger, and C. E. Elger, editors, *Chaos in Brain?*, Proceedings of the 1999 Workshop, pages 329–332, Singapore, 2000. World Scientific.
9. F. Mormann, T. Kreuz, R. G. Andrzejak, P. David, K. Lehnertz, and C. E. Elger. Epileptic seizures are preceded by a decrease in synchronization. *Epilepsy Research*, 53(3):173–185, 2003.
10. B. Schelter, M. Winterhalder, T. Maiwald, A. Brandt, A. Schad, A. Schulze-Bonhage, and J. Timmer. Testing statistical significance of multivariate time series analysis techniques for epileptic seizure prediction. *Chaos*, 16(1):013108, 2006.
11. S. Chillemi, R. Balocchi, A. Di Garbo, C. E. D’Attellis, S. Gigola, S. Kochen, and W. Silva. Discriminating preictal from interictal states by using coherence measures. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology*, volume 3, pages 2319–2322, 2003.
12. C. C. Jouny, P. J. Franaszczuk, and G. K. Bergey. Signal complexity and synchrony of epileptic seizures: Is there an identifiable preictal period? *Clinical Neurophysiology*, 116(3):552–558, 2005.
13. T. Dikanev, D. Smirnov, R. Wennberg, J. L. P. Velazquez, and B. Bezruchko. EEG nonstationarity during intracranially recorded seizures: Statistical and dynamical analysis. *Clinical Neurophysiology*, 116(8):1796–1807, 2005.
14. C. J. Stam. Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *Clinical Neurophysiology*, 116:2266–2301, 2005.
15. J. L. P. Velazquez. Brain, behaviour and mathematics: Are we using the right approaches? *Physica D*, 212(3-4):161–182, 2005.
16. F. Mormann, T. Kreuz, R. G. Rieke, C. Andrzejak, A. Kraskov, P. David, C. E. Elger, and K. Lehnertz. On the predictability of epileptic seizures. *Clinical Neurophysiology*, 116(3):569–587, 2005.
17. M. Chavez, M. Besserve, C. Adam, and J. Martinerie. Towards a proper estimation of phase synchronization from time series. *Journal of Neuroscience Methods*, 154(1-2):149–160, 2006.
18. G. A. Darbellay. An estimator of the mutual information based on a criterion for conditional independence. *Computational Statistics and Data Analysis*, 32(1):1–17, 1999.
19. A. Kugiumtzis, D. and. Kehagias, E. C. Aifantis, and H. Neuhaüser. Statistical analysis of the extreme values of stress time series from the Portevin-Le Châtelier effect. *Physical Review E*, 70(3):036110, 2004.

20. L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2: 193–218, 1985.
21. D. W. Aha and R. L. Bankert. A comparative evaluation of sequential feature selection algorithms. In D. Fisher and H. Lenz, editors, *Proceedings of the Fifth International Workshop on Artificial Intelligence and Statistics*, pages 1–7, 1995.
22. A. Tsimpiris and D. Kugiumtzis. Clustering of oscillating dynamical systems from time series data bases. In *Electronic Proceedings of the International Workshop on Knowledge Extraction and Modeling*, Capri, Italy, 2006.
23. R. Hegger, H. Kantz, and T. Schreiber. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos*, 9:413, 1999.

Symbol Extraction Method and Symbolic Distance for Analysing Medical Time Series

Fernando Alonso, Loïc Martínez, Aurora Pérez, Agustín Santamaría,
and Juan Pedro Valente

Facultad de Informática. Universidad Politécnica de Madrid. Campus de Montegancedo.
28660 Boadilla del Monte. Madrid. Spain
{falonso, loic, aurora, jpvalente}@fi.upm.es,
Agustin.Santamaria@Sun.COM

Abstract. The analysis of time series databases is very important in the area of medicine. Most of the approaches that address this problem are based on numerical algorithms that calculate distances, clusters, index trees, etc. However, a symbolic rather than numerical analysis is sometimes needed to search for the characteristics of the time series. Symbolic information helps users to efficiently analyse and compare time series in the same or in a similar way as a domain expert would. This paper focuses on the process of transforming numerical time series into a symbolic domain and on the definition of both this domain and a distance for comparing symbolic temporal sequences. The work is applied to the isokinetics domain within an application called I4.

Keywords: Time series characterization, isokinetics, symbolic distance, information extraction and text mining.

1 Introduction

An important domain for the application of time series analysis in the medical field is physiotherapy and, more specifically, muscle function assessment based on isokinetics data.

Isokinetics data is retrieved by an isokinetics machine (Fig. 1a), on which patients perform exercises at maximum strength. To assure that the patient performs exercises at constant speed, the machine puts up the required resistance to the strength the patient exerts. Our patients are chiefly sportspeople. Therefore, we decided to focus on knee exercises (extensions and flexions) since most of the data and knowledge gathered by sports physicians is related to this joint. The data takes the form of a strength curve with additional information on the angle of the knee (Fig. 1b). The positive values of the curve represent extensions (knee angle from 90° to 0°) and the negative values represent flexions (knee angle from 0° to 90°).

After observing experts at work, we found that they apply their knowledge and expertise to focus on certain sections of the isokinetics time series and ignore others. Therefore, we looked for a way of bringing the system output closer to the information sports physicians deal with in their routine work, since they demand a representation related to their own way of thinking and operating. Hence, symbolic series have been used as an alternative that more closely resembles an expert's conceptual mechanisms.

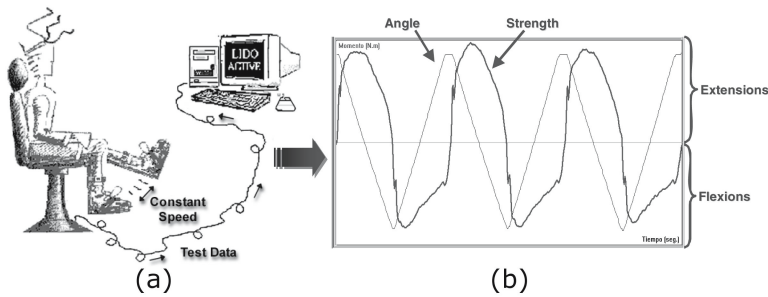


Fig. 1. Isokinetics machine (a) and collected data (b)

To do this, our research focused primarily on the design of the symbols extraction method that translates numerical time series into symbolic temporal series. An early version of this method was described in [1]. Second, we designed a distance measure to indicate how similar two symbolic time series are. This way, symbolic sequences can be automatically compared to detect similarities, classify patients, etc.

In this paper, section 2 describes the I4 system of which this research is part. Sections 3 and 4 describe, respectively, time series comparison issues and the semantic extraction method. Section 5 introduces the isokinetics symbolic distance, the proposed metric for comparing symbolic series. This metric is an extension of the Needleman-Wunch distance [2]. Section 6 shows the visualization provided to physicians. Section 7 shows the research results and evaluation and, finally, section 8 presents some conclusions and mentions future lines of research.

2 I4 System

This work is part of the I4 Project (Intelligent Interpretation of Isokinetics Information) [3], which provides sports physicians with a set of tools to visually analyse patient strength data output by an isokinetics machine (Figure 2).

I4 is composed of several subsystems. First, there are data preparation tasks, which include translation, formatting, cleaning and pre-processing. These tasks use expert knowledge and generate a database in which data are homogeneous, consistent and noise-free. The second subsystem is a knowledge-based system (KBS) that analyses expert data to make it easier for novice users and also blind physiotherapists to interpret the isokinetics curves. Third, there is a knowledge discovery in databases (KDD) system that performs numerical comparisons of isokinetics data to define reference models for patient groups and to identify injury patterns. Finally, there is a visualization module that displays exercises, injury patterns, reference models, etc.

Many of these functionalities are used on a daily basis by specialized physicians to assess their patients' (mostly top-competition sportsmen and women) potential, diagnose injuries and analyse what progress patients have made in injury recovery. The I4 system is reliable and outputs equivalent results to what an expert would do. However, it has failed to gain experts' total confidence. This is because the information the expert receives from the system does not highlight the significant aspects of the

isokinetics series in a language that he or she can easily understand. It is this state of affairs that has led to the need to build a symbolic comparison method into the I4 symbolic data subsystem. Not only should this symbolic method produce equally reliable results, but it should also provide a reasonable explanation of the results in terms of the domain under study.

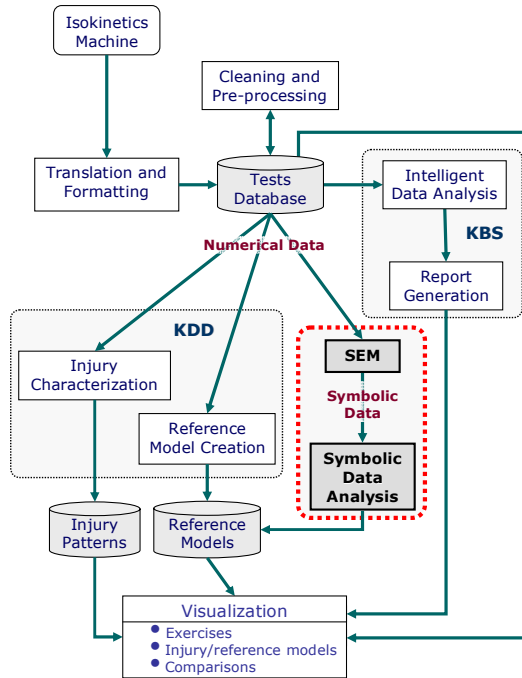


Fig. 2. I4 System Overview

3 Isokinetics Time Series Comparison Issues

There has been a lot of research in the area of numerical time series comparison [4,5,6,7]. Most of these methods are based on comparing the values of separate points in each series rather than on the overall appearance of these series. In the case of Fig. 3, for example, they would indicate that series b1 and b2 resemble each other more closely than a1 and a2.

We, however, are interested in the morphology of the isokinetics curves rather than in the strength value exerted at any given point in time. Although a simple time translation would solve the problem for the example in Fig. 3, this translation would overlook the patients' strength values (which is not unimportant) and would not be a valid solution in all cases or for all parts of the sequence.

In the field of morphological comparison, there is a shape definition language (SDL) [8] for retrieving objects based on shapes contained in the histories associated

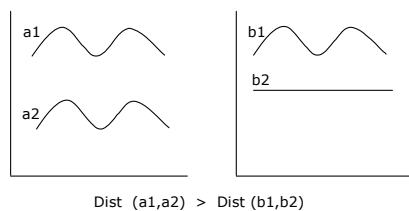


Fig. 3. Example of traditional similarity methods

with these objects. This language is, however, domain independent, which is one of its main differences from the metric that we propose.

In the isokinetics domain, time series should mostly be analysed by a specialist who has the expertise to interpret the series' different features. When analysing a sequence, most experts instinctively split the temporal sequence into parts that are clearly significant to their analysis and ignore other parts that provide no information. Accordingly, the expert identifies a number of concepts based on the features present in each part of the time series that are relevant for explaining its behaviour.

After observing isokinetics domain experts at work, we found that they focus on sections like “ascent, curvature, peaks...” These are the sections that contain the concepts that have to be extracted from the data. Therefore, we developed the symbol extraction method (SEM) to translate numerical into symbolic time series that include expert knowledge. Our experience suggests that physiotherapists are better at interpreting symbolic data because it is more akin to their way of reasoning.

The next point was to find a way of comparing symbolic series and automatically evaluating how similar two isokinetics series are. To do this, we had to define what we called the isokinetics symbolic distance (ISD).

4 Isokinetics Symbolic Domain

To analyse isokinetics data symbolically, we first defined our vocabulary, called isokinetics symbols alphabet (ISA). We then developed a method to extract symbolic information from the numerical data. This section just gives the details required to gain an overall understanding of the article. For further information see [1].

4.1 Isokinetics Symbols Alphabet

Any exercise successively chains regions corresponding to a knee extension and flexion, both with a similar morphology (shown in Fig. 4). After a number of interviews with the expert, we were able to identify the following symbols that capture the meaningful information contained in the curves:

- *Ascent*: part of the curve where the patient gradually increases the applied strength.
- *Descent*: the patient gradually decreases the applied strength.
- *Peak*: a spike in any part of the sequence.

- *Trough*: a valley in any part of the sequence.
- *Curvature*: the upper section of a region.
- *Transition*: the changeover from extension to flexion (or vice versa).

The symbols are labelled with the region to which they belong (extension or flexion), taking into account that they have been considered as absolute values and, therefore, the flexions also represent positive values.

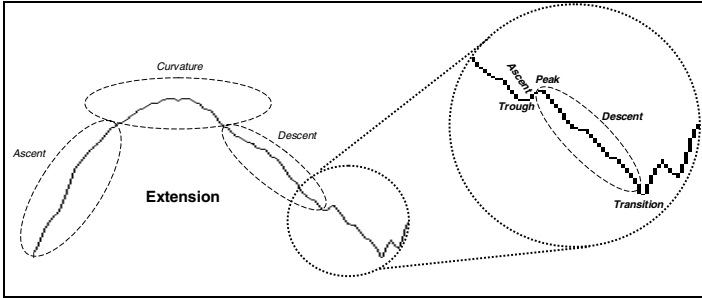


Fig. 4. Symbols of an isokinetics curve

After identifying the symbols used by the expert, we needed to know what each symbol could be like, that is, its type. These types were elicited directly from the expert as she analysed a set of supplied sequences that constituted a significant sample of the whole database. The set of symbols, types and regions form the ISA, which is shown in Table 1.

Table 1. Isokinetics Symbols Alphabet

Region	Symbol	Type	
EXT	<i>Ascent</i>	Sharp	Gentle
	<i>Descent</i>	Sharp	Gentle
	<i>Trough</i>	Big	Small
	<i>Peak</i>	Big	Small
	<i>Curvature</i>	Sharp	Flat
FLEX	<i>Transition</i>	-	

4.2 Symbols Extraction Method

SEM, whose architecture is shown in Fig. 5, was designed to transform the isokinetics curves into symbolic sequences represented according to ISA.

First, a prepared numerical sequence is put through the domain-independent module (DIM), which outputs a set of domain independent features, that is, peaks and troughs. Both the features output by the DIM and the actual numerical sequence are used as input for the domain-dependent module (DDM), which outputs all the domain-dependent data of the sequence. This module is divided into three submodules:

1. *Output of domain-dependent features.* The aim of this submodule is to get all the symbols that characterize the given numerical sequence. To do this, the module

selects the relevant peaks and troughs and identifies the ascents, descents and curvatures.

2. *Filter*. The set of symbols output by the above submodule is put through a filtering stage. Apart from other filtering processes, this filter checks that there are no consecutive symbols that are equal. For example, it makes no sense to have two ascents one after the other, because they would really be just one ascent.
3. *Assign types to symbols*. The goal of this submodule is to label each symbol with a type. This will provide more precise information about the original temporal sequence. This process is based on a set of rules that use a number of thresholds to define the symbol type in each case.

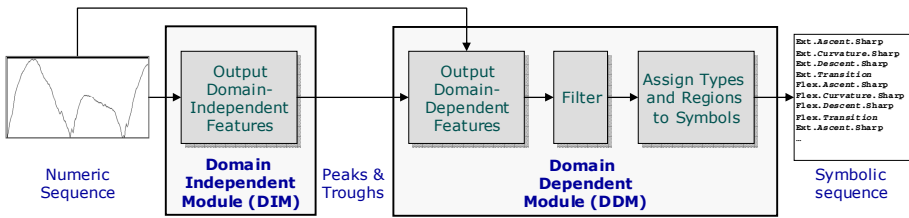


Fig. 5. Architecture of SEM

5 Comparing Symbolic Series: Isokinetics Symbolic Distance

5.1 String Measures

Our goal is to find a similarity measure that can be used to compare isokinetics symbolic sequences and perform data mining tasks.

After a thorough study to select the best similarity measure for the medical field of isokinetics during which we analysed the string metrics listed in Fig. 6, we reached the conclusion that a new measure needed to be designed. This measure is based on edit distances and, specifically, on the Needleman-Wunch distance [2]. The analysis we conducted rejected the transposition- and term-based distances and hybrid measures for the reasons explained below.

Transposition-based distances use the transposition of the sequence elements. The order of the elements is unimportant in this type of distance, which merely ascertains whether two given sequences have the same (or similar) values in an interval with a previously defined range. Isokinetics symbolic sequences have an ordered structure that should be taken into account to examine the similarity between two given sequences. Therefore, we were unable to use this type of distance for our purposes.

Term-based distances consider each sequence simply as a set of elements (terms or tokens) so the order in which the elements are arranged in the sequence is lost when the distance is analysed. Therefore, they are not applicable in the isokinetics domain either.

Hybrid measures are distances defined from other distances, by mixing edit distances with term-based distances, for example. They cannot be used to implement a distance in the isokinetics domain either, as they also use the term-based or

transposition-based distances for their definition and, therefore, do not take into account the order of the sequence elements.

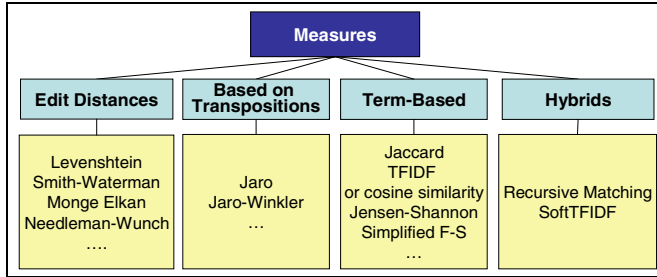


Fig. 6. Types of Strings Measures

5.2 Isokinetics Symbolic Distance

The family that best meet the needs of the isokinetics domain is edit distances, as it takes into account the order of the components and the morphology of the sequence. However, none of the edit distances we examined exactly fits our problem, because the symbols used in the isokinetics domain also have an associated type that needs to be taken into account to calculate the distances. This led us to propose a variation on the Needleman-Wunch distance. The suggested distance, the isokinetics symbolic distance (ISD), allocates a variable cost to the *insert* and *delete* operations depending on the symbol and symbol type to be inserted or deleted. It also allocates a variable cost to the *substitute* operation depending on the symbol and symbol type that are substituted.

The researched isokinetics sequences are composed of three repetitions, and each repetition is composed of an extension and a flexion. Therefore, an isokinetics sequence contains six parts, each of which is represented by the notation shown in (1).

$$\langle \text{Zone} \rangle \langle \text{Repetition} \rangle \langle \text{Sequence} \rangle \quad (1)$$

where $\langle \text{Zone} \rangle$ can take the value E (for Extension) or F (for Flexion), $\langle \text{Repetition} \rangle$ can take the value R^1 , R^2 or R^3 depending on whether it is repetition 1, 2 or 3, and $\langle \text{Sequence} \rangle$ can take the value S^1 or S^2 depending on the sequence 1 or 2.

Fig. 7 shows the three steps required to calculate the ISD of two symbolic sequences: calculate the ISD between each pair of subsequences, normalize these distances and calculate the arithmetic mean to get the total distance.

The ISD between two series, S_1 , of length n , and S_2 , of length m , is calculated by building a matrix of $m \times n$ elements. This matrix includes the accumulated costs of the *insert*, *delete* or *substitute* operations, always calculating the best alignment between the two symbolic sequences for comparison. This prevents trapping in local minima. The value of each matrix element is calculated using equation (2): element (i, j) indicates the ISD between S_1' and S_2' (the subsequences —prefixes— of S_1 and S_2 ending in elements j and i , respectively); element (m, n) indicates the ISD between S_1 and S_2 . This way, the ISD can be used to get the least costly edit command sequence (delete, insert and substitute) for transforming S_1 into S_2 .

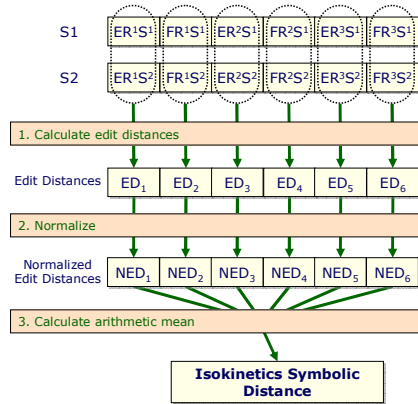


Fig. 7. Computing the Isokinetics Symbolic Distance

$$D(i, j) = \min \begin{cases} D(i-1, j-1) & \text{if } s_i = t_j & // \text{copy} \\ D(i-1, j-1) + \text{SubstituteGapCost} & \text{if } s_i \neq t_j & // \text{substitute} \\ D(i-1, j) + \text{InsertGapCost} & & // \text{insert} \\ D(i, j-1) + \text{DeleteGapCost} & & // \text{delete} \end{cases} \quad (2)$$

Due partly to qualitative aspects (each symbol has a different structural weight) and partly to quantitative issues, not all the operations or all the symbols can be allocated an identical gapcost in the isokinetics field. For example, curvatures are symbols that are part of any repetition, whereas peaks and troughs are circumstantial symbols, usually induced by minor patient injuries and, therefore, may or may not appear. Additionally, a large peak cannot be considered the same as a small peak. Therefore, each symbol has to be allocated a different weight, and a distinction has to be made depending on the symbol type.

We had to define both the cost of substituting one symbol-type by another and the cost of inserting or deleting a particular symbol-type. This was done with the help of an isokinetics expert. The *insert* and *delete* costs were unified to assure that the comparison of two series is symmetric.

As regards the *substitute* cost, several possibilities were weighed up. Initially, we designed a tabular structure, where the table rows and columns included all the symbols-types and the cell (i, j) represented the cost of substituting the symbol-type i by the symbol-type j . However, this table was hard work for the expert to build. For instance, the expert would have to define $(n \times m)^2 / 2 - (n \times m)$ cells if the number of symbols is n and the mean number of types per symbol is m (the table is symmetric and the cost will always be 0 along the main diagonal). Additionally, this table is not very open to the entry of any change in the symbols alphabet, as the expert would have to put in a lot of work to reformulate the table to accommodate the changes.

To overcome these two problems, we opted for a graph structure, where the principal cost of substituting two symbols is determined mainly by the symbol, whereas the symbol type serves to refine that cost. Fig 8b shows this substitution graph.

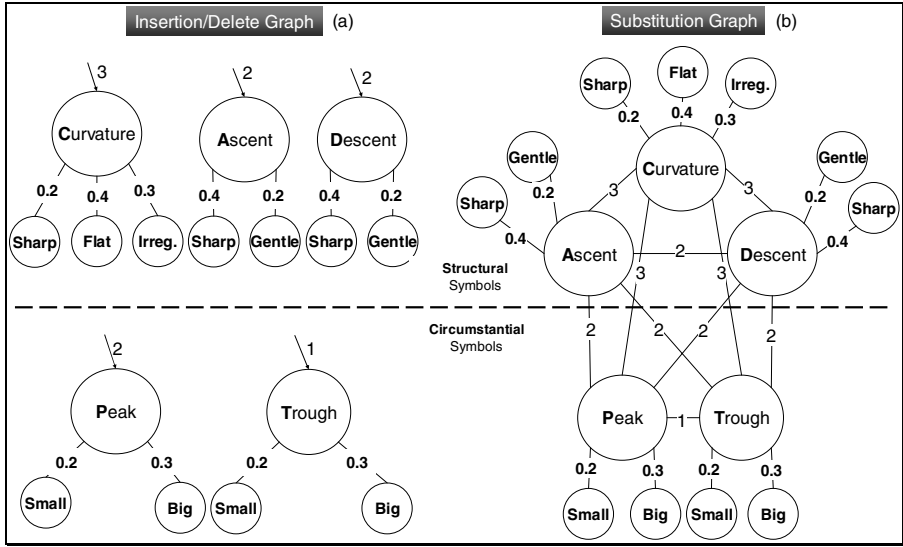


Fig. 8. Insertion/Deletion and Substitution Graph

The expert will have to define $n^2/2 - n + nxm$ values, which is clearly fewer than for the table. Additionally, this structure is much more open to the entry of any change in the symbols alphabet and it is also more self-explanatory for the expert.

For the sake of coherency, we have used a similar representation for the *insert* and *delete* costs (Fig 8a), although, in this case, there is no difference in the number of values that the expert has to define for the graph and for the table.

To make things easier for the expert, we took the graphical representation for each symbol-type and defined some initial costs by comparing the area each symbol covered. These initial values were presented to the expert and proved to be a good starting point.

The gapcosts plotted in the graphs of Fig. 8 are the ones to be used in (2). It is clear from these graphs that there is a cost per symbol to which a cost per type associated with each symbol is added.

Having obtained the distances between each of the six components of the two sequences for comparison (in Fig 7 these distances are denoted ED_x , where x is the number of the component that has been compared), these values go through the normalization process after which all the distances are defined in the interval $[0, 1]$. The normalization is based on dividing the obtained distance value between what would have been output in the worst case. In our domain, as all the sequences have six curvatures (two for each repetition), the worst case would be to have *substitute* operations for the curvatures ($WorstGapCostCurvature$) and have *substitute* operations for ascents or descents with the worst gapcost ($WorstGapCostAscent\_Descent$). Therefore, the value by which the ISD has to be divided is (3).

$$(Size\_of\_S_x - 6) \cdot WorstGapCostAscent\_Descent + 6 * WorstGapCostCurvature \quad (3)$$

Once the normalized distances have been obtained for each component, their arithmetic mean is calculated. This process outputs the isokinetics symbolic distance between the two compared sequences.

6 Symbolic Visualization

Fig 9 shows a prototype evaluator interface. The original numerical series (at the top of the interface) is translated into a symbolic series (centre right). This second series is equivalent to the first (although it includes the significant and omits the irrelevant aspects), and is processed internally by the I4 system. This symbolic series is also displayed graphically as a curve (at the bottom) to give the user a more intelligible view of the data. The central part of the interface shows the intermediate steps in the translation of the numerical sequence into symbols (i.e. the SEM stages).

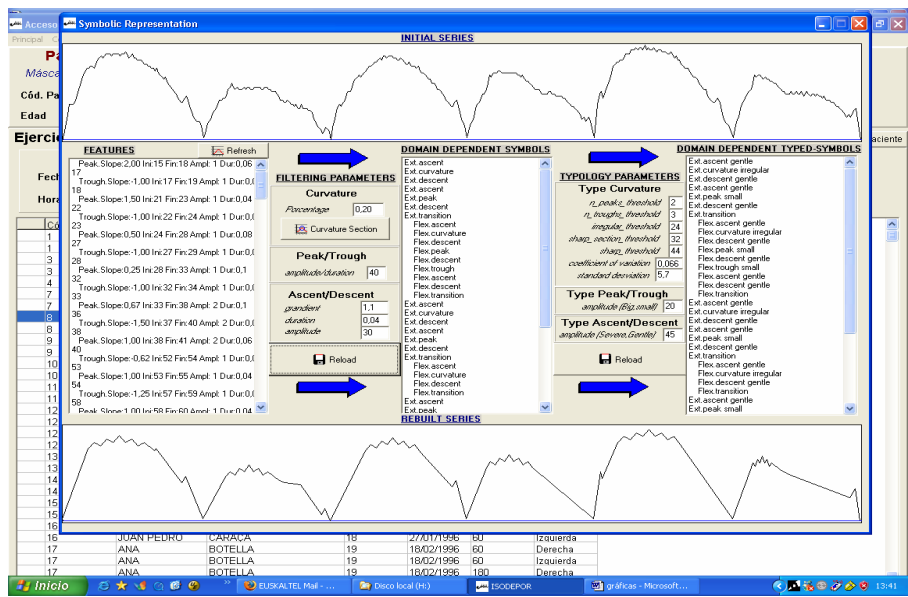


Fig. 9. I4 prototype evaluator interface

As one of the goals of the system is for the user to compare models and exercises on the basis of their symbolic features, the interface can also receive two input curves, translate them into symbolic curves and compare these curves.

7 Results and Evaluation

The evaluation focused on two points: a) check whether the physiotherapist achieved more efficient results by analysing symbolic isokinetics symbols (SIS) than using numerical isokinetics symbols (NIS); b) check whether the results achieved by the

system comparing symbolic sequences using the symbolic distance were more significant than comparing their respective numerical sequences using the Fourier transform.

For point a) the expert and novice physicians who participated in I4 project development were given the same information: an isokinetics test. The test was repeated for 34 occurrences (20 with no injuries at all, 8 with common injuries and 6 with unusual injuries). The results are shown in Table 2.

Table 2. Evaluation of injury detection

	<i>NIS</i>		<i>SIS</i>	
	<i>Expert</i>	<i>Novice</i>	<i>Expert</i>	<i>Novice</i>
<i>20 uninjured</i>	20 OK	Failed 4	20 OK	Failed 2
<i>8 common injuries</i>	8 OK	5 OK (3 mistakes)	8 OK	7 OK (1 mistake)
<i>6 unusual injuries</i>	2 mistakes and 1 don't know	2 mistakes and 4 don't knows	1 don't know	2 mistakes and 4 don't knows

We found that the symbolic sequence yielded better results than the numerical sequence for both the expert and the novice physicians, but the results were more significant in the latter case.

With respect to point b), the I4 system was fed a knee isokinetics test with 28 occurrences, each performed by a different sportsperson: 20 had no injury, 5 had a common knee injury (torn ligament), and 3 had an unusual injury (osteocondritis). The system was also given 3 reference models: 1 without injuries and 2 with the above-mentioned injury types.

The results gathered for the comparison between each reference model and each of its respective occurrences using the distance provided by the Fourier transform (FT), for the numerical sequences and the ISD for the symbolic sequences are listed in Table 3. We found that the ISD distance is more discriminative than the FT because it focuses more on the singular points (peaks and troughs) that define the injury.

Table 3. FT and SD distances evaluation.

	FT distance			SD distance		
	0-0.33	0.34-0.66	0.67-1	0-0.33	0.34-0.66	0.67-1
<i>20 uninjured</i>	16	2	2	16	2	2
<i>5 common injures</i>	3	2	-	4	1	-
<i>3 unusual injuries</i>	2	-	1	2	1	-

8 Conclusions

In the field of isokinetics, the automatic analysis of time series is an essential tool for the physiotherapist. This paper has presented ongoing work on the development of a comprehensive system to deal with isokinetics data, including symbolic data analysis.

Our previous experience with numerical methods has been very positive, but experts did not have enough confidence in the system, because the information they received from I4 did not highlight the relevant aspects of the isokinetics series in a

language they found easy to understand. This is the reason that led us to introduce symbolic methods, which use the same language as our experts.

This paper presented SEM, which extracts symbolic information from numerical isokinetics data, using an alphabet defined by our experts. SEM contains a domain independent module, which can be used in other domains. Additionally, we defined a symbolic distance, based on edit operations on an isokinetics symbolic sequence.

As the evaluation has shown, both the symbolic sequence generated by SEM and the comparison of isokinetics data using the ISD have proved helpful for sports physicians. Given those encouraging results, we are continuing our research in the field of symbolic data analysis to build new functionalities into I4 and add symbolic injury characterization and symbolic reference model creation to the numerical KDD subsystem.

References

1. Alonso F., Martínez, L., Montes, C., Pérez, A., Santamaría, A., Valente, J.P. (2004) Semantic Reference Model in Medical Time Series. International Symposium on Biological and Medical Data Analysis ISBMDA 2004: In Lecture Notes in Computer Science, no. 3337, pp. 344-355.
2. Needleman, S. B. & Wunsch, C. D. (1970) A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, 48, 443-453.
3. Alonso F, Valente J P, Martínez L and Montes C. (2005) Discovering Patterns and Reference Models in the Medical Domain of Isokinetics. In: J. M. Zurada, editor, *New Generations of Data Mining Applications*, IEEE Press/Wiley.
4. Agrawal R, Faloutsos C, and Swam A N. (1993) Efficient Similarity Search In Sequence Databases. In D. Lomet, editor, *Proceedings of the 4th International Conference of Foundations of Data Organization and Algorithms (FODO)*, pages 69-84, Chicago, Illinois, Springer Verlag.
5. Faloutsos C, Ranganathan M, Manolopoulos Y (1994) Fast subsequence matching in time series databases. In *Proceedings of SIGMOD'94*, Minneapolis, MN, pp 419-429.
6. Rafei D, Mendelzo A. (1997) Similarity-Based Queries for Time Series Data. In *Proceedings of SIGMOD*, Arizona
7. Han J, Dong G, Yin Y (1998) Efficient mining of partial periodic patterns in time series database. In *Proceedings of the 4th international conference on knowledge discovery and data mining*. AAAI Press, Menlo Park, CA, pp 214-218.
8. R. Agrawal, G. Psaila, E. L. Wimmers, and M. Zaït (1995) Querying shapes of histories. IBM Research Report RJ 9962 (87921), IBM Almaden Research Center, San Jose, California.

A Wavelet Tool to Discriminate Imagery Versus Actual Finger Movements Towards a Brain–Computer Interface

Maria L. Stavrinou¹, Liviu Moraru¹, Polyxeni Pelekouda², Vasileios Kokkinos²,
and Anastasios Bezerianos¹

¹ Dept. of Medical Physics,

² Department of Physiology, School of Medicine University of Patras,
26500 University Campus, Rio, Greece
bezer@patreas.upatras.gr

Abstract. The present work explores the spatiotemporal aspects of the event-related desynchronization (ERD) and synchronization (ERS) during rhythmic finger tapping execution and imagery task. High resolution event related brain potentials were recorded to capture the brain activation underlying the motor execution and motor imagery. ERS and ERD were studied using a complex morlet wavelet decomposition of EEG responses. The results show similar patterns of beta ERD/ERS after the stimulus onset, for both the actual and imagery finger tapping task. This approach and results can be regarded as indicative evidences of a new strategy for recognizing imagined movements in EEG-based brain computer interface research. The long-term objective of this study is to create a multiposition brain controlled switch that is activated by signals that are measured directly from a human's brain.

Keywords: EEG, Brain-Computer Interface, finger-tapping, imagery, beta rhythm, wavelet, Event Related Synchronization (ERS) -Desynchronization (ERD).

1 Introduction

The electroencephalogram (EEG) based Brain-Computer Interface (BCI) is a communication system which represents a direct connection between the human brain and the computer. The general idea behind any BCI research is to establish patterns of activation that can be used to help the disabled people perform the desired action [1]. The BCI research revolves around the design of effective experimental protocols, the development of efficient methods for feature extraction of brain activation and the evaluation of algorithms for translating these features into commands. Nowadays, a great variety of EEG-based BCI systems are in use [2, 3]. The methodology used for these machines make use of slow potentials as P300 or beta (14-30 Hz) and mu (8-12 Hz) rhythm detection. The detection of such signals is then transformed into commands that operate a computer display or other device [4].

Motor imagery has become the newest trend in BCI research [5, 6, 7]. This is due to the fact that imagination of movements appears to recruit similar -or the same- neural networks in the brain, to those used to perform actually the same movements [8].

In our work, we investigated the spatio-temporal EEG brain activity during a real and an imaginary rhythmic finger tapping task. Previous studies using relative longer inter-stimulus interval protocols (of 4 to 10 seconds) have reported characteristic synchronization and desynchronization timecourses, following the onset of the action (from 4 to 7 sec). We investigated whether we could see similar activation patterns for an imagery task using an 1.5 sec inter-stimulus interval protocol.

Time-frequency decomposition of brain electrical signals by means of wavelet transform has been widely employed in the study of brain rhythms. The major advantage of the wavelet transform is that it allows the decomposition and manipulation of time-varying non-stationary signals, being particularly suited to the analysis of ERPs. In our study, we first applied the continuous wavelet transform (CWT) in order to calculate the power spectra of various frequency bands for the each single trial, using a complex Morlet mother wavelet. Next, task related neural responses have been uncovered by averaging single trial time-frequency representations.

It is widely accepted that while brain processes certain events the ongoing brain rhythmical activity can be blocked or desynchronized. These types of changes are better detected by frequency analysis because they represent frequency specific changes of the ongoing EEG activity. They consist, in general of an amplitude attenuation or power decrease and/or of an amplitude/power enhancement in certain frequency bands. This is considered to be due to a decrease or an increase in synchrony of the underlying neuronal populations. The former case is called event-related desynchronization (ERD), and the latter event-related synchronization (ERS) [9].

In this paper we focus on the detection of beta rhythms and associated ERD/ERS patterns, previously described to occur with initiation and execution of motor actions [10] as well as with motor imagination [11]. Discrimination of short time activations during motor imagery based on the frequency content may improve decision making and enhance performance of a BCI system.

2 Methods

Subjects and Experimental Paradigm

Two healthy volunteers participated in this study (2 males) with age range 26-28 years. Subjects were strongly right-handed according to the Edinburgh Inventory [12]. None of them had a previous history of neurological disease and took no medication at the time of the experiment. All subjects gave their written informed consent. The protocol and experimental procedures were approved by the local ethics committee and were in compliance with the declaration of Helsinki. The experiments have been conducted in the EEG Laboratory of Neurophysiology Unit, Department of Physiology, Medical School, University of Patras, Greece.

Subjects sat on a comfortable armchair in an electrically isolated room, dimly illuminated. A small led light was adjusted on the wall in front of the subjects, in order to fixate their sight on it, to avoid ocular movements. The experimental session consisted of four parts. In the first part, median nerve stimulation of subject's right wrist, above the motor threshold where a definite twitch of the thumb was visible, was performed. The ISI of the electric stimulation was 1500 msec and stimulus duration

200 microseconds, with a total number of trials 250. For the second part the subject was instructed to make a right index finger task, tapping a key of a keyboard, externally paced by an auditory signal (1000 Hz, 64dB max arranged to be heard but not annoying to the subject, 50microseconds duration and ISI 1500 msec). The next session consisted of a sub-session where the subjects practiced right index finger tapping for another 250 trials before the right index finger imagery task begun. Subjects were instructed to imagine the right index finger tapping task with the auditory stimulation providing the pace. Subjects were instructed to imagine the kinesthetic of the movement and not the visual image of the movement itself. After training, approximately 130 trials of right index finger imagery were recorded. The last session consisted of a control auditory stimulation, where the subject was instructed to hear passively the auditory stimulation without executing any movement, while being relaxed.

Data Acquisition

EEG signals were recorded from 60-electrodes mounted on an elastic cap (Electrocap International, Ohio, USA), and acquired with a SynAmps amplifier (Neuroscan, USA). The Neuroscan software was used for recording. Impedances were kept below 5 K Ω . Linked earlobes were used as a reference and AFZ electrode as ground. The signals, were filtered between 0.1 and 200 Hz, with a sampling frequency 1000 Hz. The positions of all the electrodes as well as of four anatomical landmarks (nasion,inion and the two preauricular points) and points on the head were digitized with a 3d Digitizer (Pohlemus 3Dspace Fastrack, Colchester Vt, USA).

Data Analysis

The datasets were visually checked for noisy epochs which were excluded from further analysis. Approximately 220 artifact-free trials were selected for each session and each subject for the actual finger tapping task and about 120 for the imagery task. Each epoch consisted of a time window of 300 ms pre- and 1200 ms post- stimulus, while a (-100 -80) interval was used for baseline correction. Because we were interested in the activity in the sensory and motor cortex we selected for time-frequency analysis only the corresponding electrodes. The selection of electrodes was based also on the detection of maximum activity during the actual finger tapping task and the control median nerve recording which served to provide a landmark for the sensory cortex. Therefore the cluster of electrodes selected for further analysis was FC3, FC1, FCZ, C5, C3, C1, CZ, CP3, CP1, CPZ, for the left (contralateral) hemisphere and for the right hemisphere FCZ, FC2, FC4, CZ, C2, C4, C6, CPZ, CP2, CP4.

Laplacian Filtering

The electric potential recorded on the brain cortex results from the generation of sources of current. However, it depends on the reference electrode chosen and devices used. Laplacian transformation of scalp surface potentials is commonly used as a reference-free method to attenuate low spatial frequencies ('smearing') introduced into

the scalp potential distribution due to volume conduction, thus sharpen ERP topographies in a physiologically meaningful way [13, 14]. 2-D Laplacian has been computed by using the electrode locations from the digitization procedure. In this way we reduced the contribution of distant sources at each scalp location. This transformation has been applied before further signal processing and feature extraction.

Time – Frequency Representations

Single trials were further detrended in order to remove DC-offset and slow drifts (< 1 Hz). The time-frequency maps were constructed by the calculation of the power spectra of the detrended single trials, by squaring the convolution of them to the Morlet complex mother wavelet, which in our case is defined as follows:

$$u(t, f_0) = A \exp(-t^2 / 2\sigma_f^2) \exp(2\pi f_0 t) \quad (1)$$

where $\sigma_f = 1 / 2\pi\sigma_t$ and σ_t are the frequency and time resolution respectively, around the central frequency f_0 , is a frequency inside the frequency band of our interest for which we calculate the power spectrum. $A = (\sigma_t \sqrt{\pi})^{-1/2}$ is a normalization factor which ensures that the wavelet has unit energy [15, 16, 17]. The energy for each frequency analyzed, is the absolute value of the convolution of this mother wavelet with the signal (in our case, each single trial).

$$E(t, f_0) = |u(t, f_0) * s(t)|^2 \quad (2)$$

The time-frequency maps are the result of the average of this energy averaged over time for all the single trials.

Quantification of ERD/ERS

The time-frequency maps were calculated for the frequencies from 4 to 44 Hz. Based on the time-frequency plots we selected the most reactive individual frequency band in the beta range with 2 Hz span [9]. The ERD/ERS is defined as percentage power decrease (ERD) or power increase (ERS) in relation to a baseline time interval, in our case from -120 to -20 ms before stimulus onset (which is assumed as time zero). Due to inter-individual differences in the peak-frequency activity, ERD/ERS is calculated within individually determined frequency bands. A wavelet-based estimator is therefore calculated in order to determine the most reactive frequency, based on the average of the power of the wavelet coefficients for each frequency inside the entire frequency band under analysis. The estimators in this study were derived from the average of wavelets coefficients for the four following time intervals: from -200 to -50 ms, the pre-stimulus interval, and post-stimulus 10 to 200, 205 to 400 and 400 to 900 ms. An algorithm providing the quantitative outcome regarding the central frequency of the frequency band under consideration, measures the occurrence of a maximum of frequency in the 4 pre-selected time periods for the electrode selected, and provide us with the value having the most occurrences. The ERD/ERS then is calculated on 2 Hz interval around this frequency (most reactive frequency band).

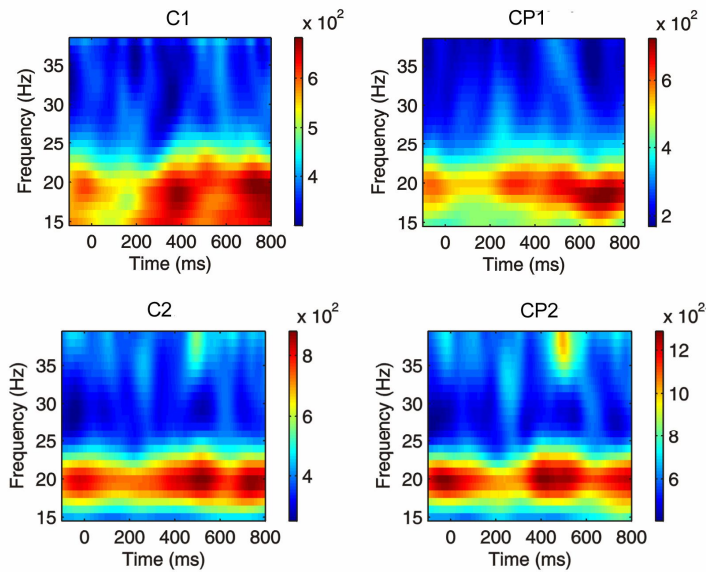


Fig. 1. Time-frequency representations of signals recorded for actual movement at electrodes of the contralateral hemisphere (C1, CP1) top row and ipsilateral hemisphere (C2, CP2)

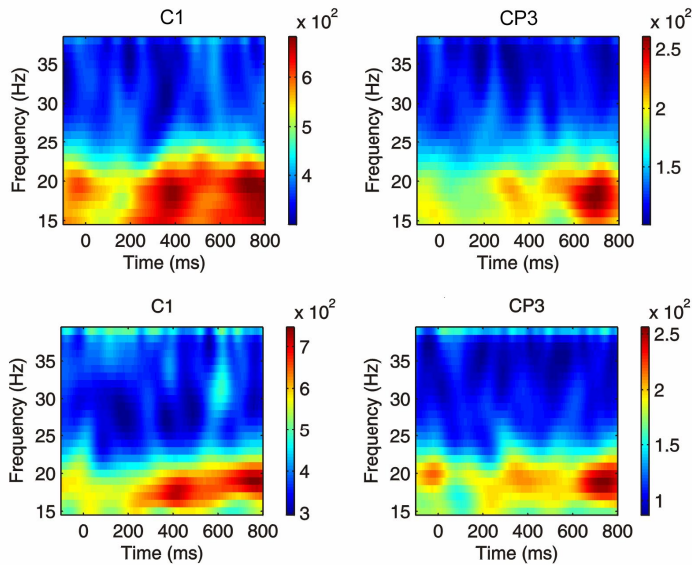


Fig. 2. Time-frequency representations of signals recorded at electrodes of the contralateral hemisphere (C1, CP3) for actual movement (top row), and during imagery (bottom row)

3 Results

Examples of time-frequency ERD/ERS maps during the actual execution of the movement, from one subject are displayed in Figure 1. As it was expected, we found in all channels prominent alpha (μ) and beta band rhythmical activity. We focused our analysis of the beta band, as a dominant beta rhythm has been found in most of the selected EEG electrodes and it displayed for our subjects a more distinct ERD pattern than the alpha (or μ) rhythm. In Figure 1, on top plot, we can see the pattern of activation for electrodes C1 and CP1, for the execution of movement. The maximum average activation analysis revealed maximum activation at the C1 electrode. A clear decrease in the signal energy can be see after the presentation of the stimulus at C1 and CP1 (contralateral hemisphere), ending at about 200 ms post-stimulus. After, the 200 ms energy power rebounds and an event related synchronization (ERS) occurs again. However, the ipsilateral hemisphere (e.g., at C2 and CP2) displays beta synchronization as well, but no similar to contralateral hemisphere beta power decrease is observed.

During imagination of finger tapping, time-frequency energy maps reveals similar energy patterns in the contralateral hemisphere, at the same electrode sites as during the

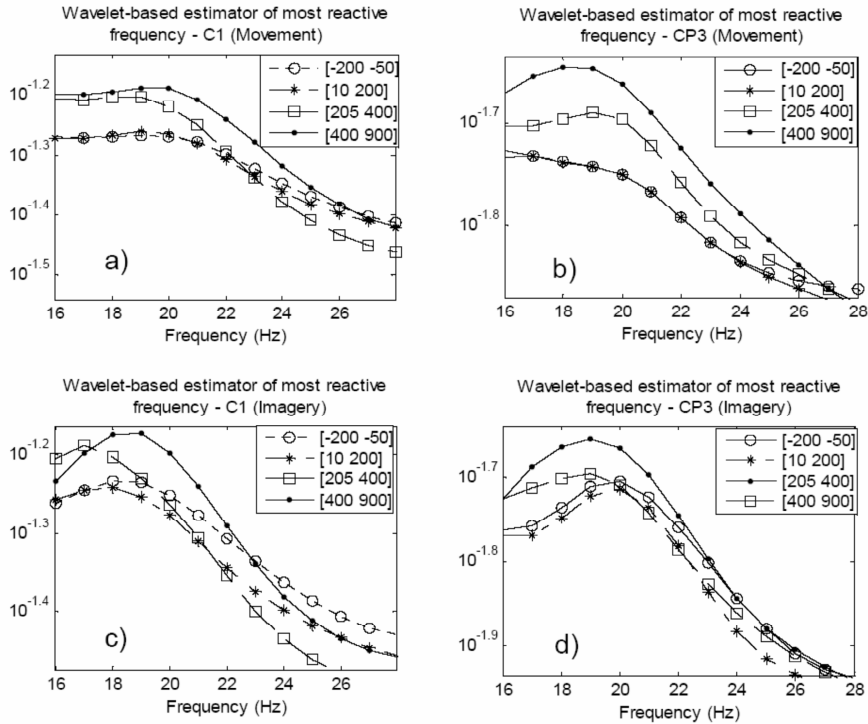


Fig. 3. Wavelet-based estimators for C1 (a, c) and CP3 (b, d) electrodes, for movement and imagination of movement

actual execution of the movement (see Figure 2, for C1 and CP3). Moreover, at the majority of electrodes over the contralateral sensorimotor cortex, the desynchronization seems to have a longer duration comparing to the actual movement.

For the auditory control experiment, a constant beta ERS was observed, with no ERD pattern either for contralateral or ipsilateral hemisphere (results not shown here).

In Fig. 3, we present the wavelet-based estimators for C1 and CP3 electrodes, for movement and imagination of movement. The results, as shown and in the above figures, indicate that the most reactive frequencies for these time intervals are in the range of 18-20 Hz. Moreover, they also show that the decrease in this beta sub-band occurs between 10-200 ms post-stimulus, as can be seen from the time-frequency maps (Fig. 2).

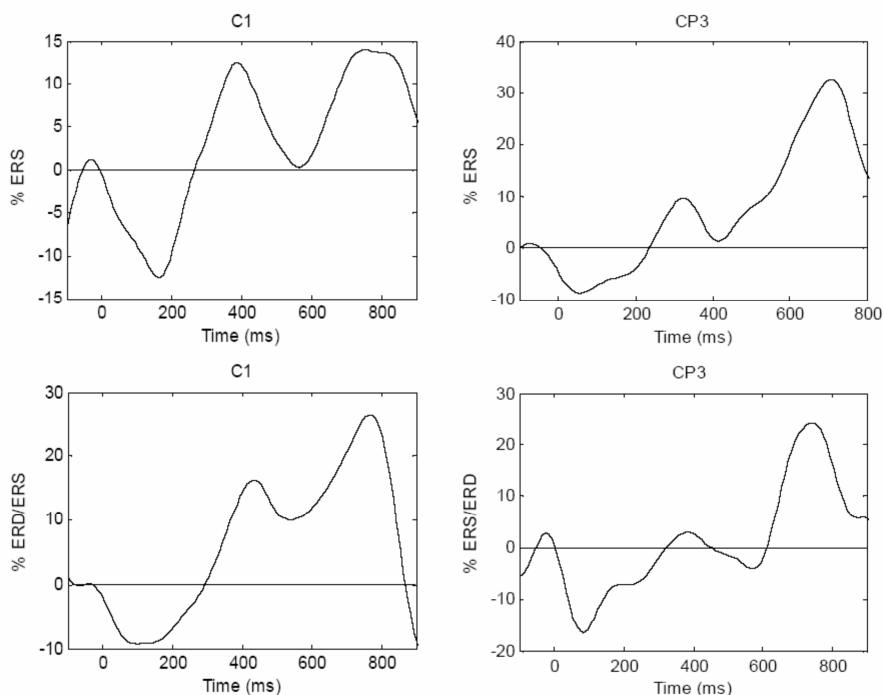


Fig. 4. Percentage ERD/ERS time courses of the activity recorded in C1 and CP3, for the actual (top) and imagination (bottom) of finger tapping movement

In Fig. 4 are shown the percentage ERD/ERS time courses of the activity recorded in C1 and CP3, for the actual and imagery of movement. In this case, both ERD and ERS were located inside the 18-20 Hz (beta) sub-band. The ERD, present in the first 200 msec, reaches a percentage decrease in the order of -10% (-12.5 % for C1 and -8.7 % for CP3), for this subject, while the percentage increase for the rebound of ERS is 14% for C1 and 32.6% for CP3. For the imagination of movement, the percentage

of ERD is in the order of -10% (-9.21% for C1 and -16.39% for CP3) and the increase of the ERS rebound +25% (26.42% for C1 and 24.14% for CP3).

4 Discussion and Conclusions

As it established from previous studies, the beta rhythm represents activity of the motor cortices including planning and execution of movements [8, 19, 9]. Thus, the desynchronization/synchronization properties of beta rhythm are expected to play a major role, in the detection of movement intention of disabled people and imagination of movement. Therefore, for the BCI research, detection of ERD/ERS patterns and especially quantification of ERD can serve as an effective and indicative parameter.

Desynchronization of rhythmical activity dominantly on the contralateral hemisphere has been also reported in other studies for the mu rhythm [5, 7] after execution of movement. It has been reported however also in the beta band during imagery tasks [11] and during actual motor execution [10, 18]. As has been reported, these beta oscillations are highly somatotopically localized in the sensorimotor hand representational areas [19, 20]. On the other hand the alpha rhythms demonstrate a more widespread desynchronization after various cognitive and motor tasks. In our study we investigated the prominent beta band (18-20 Hz) desynchronization over the sensorimotor cortex that occurred shortly after the stimulus onset. This ERD lasted for about 200 ms for the actual finger movement. A similar ERD at the same frequency band was observed during the imagery of the same task. Functional Brain Imaging studies have indicated almost the same areas where activated during imagination or actual execution of hand movement [21]. A prominent ERS rebound occurred after the end of the ERD (Fig. 2, 3) for both cases. Similar results for a beta rebound were reported after 200 ms post stimulus [17], as in our results. These results indicate a similarity in the processing of the brain between imagery and actual movement that exists both anatomically and functionally. Through quantitative analysis of ERD and ERS, features efficient to detect an intention of movement could be extracted. Moreover, another criterion is the frequency for which less or no ERD occurs in the ipsilateral hemisphere.

The ERD/ERS quantification is based on the subtraction of a pre-selected baseline period from the energy of the wavelet coefficients. One could argue that this could not be valid due to a different level of activation in the preparation and execution of a task. However, it has been found that at time scales similar to ours, this activation is similar in both actual and imagery tasks [22], and is in agreement with the result of the wavelet-based estimator for the pre and post-stimulus period.

The pattern of activation between subjects can be different, as inter-individual differences exist for the most reactive ERD/ERS frequency band [9]. Thus the wavelet-based estimator is a first step to quantitatively detect exactly this parameter (the most reactive frequency band) for every subject. However, the percentage of ERD/ERS for the individualized most reactive frequency band, must exhibit a similar form. And this is what one should detect and extract as a feature for the BCI system. Here results are shown from a representative subject therefore, necessity of further investigation including a larger population study will provide us with statistically

valuable results. However, in studies with higher ISIs stability in the ERD/ERS patterns was observed [23]. The same methodology can be performed for another frequency band like alpha, or both alpha and beta, depending on the response of the subject.

The ability to perform an imagery task relies on the ability to suppress and enhance the hand cortical rhythms, in order for them to be detected from the BCI feature extraction application. Thus concentration, alertness and good performance are critical parameters for the efficacy of the method. In our study, the similarity in the patterns of imagery and actual movement shows that subjects could actually efficiently direct and maintain their attention in order to mentally realize the movement. It is important to note here that stability of intra-subject beta ERD/ERS during imagery has been previously reported [23]. Moreover, indications exist that higher attentional activation is achieved for self-paced than externally paced responses to a stimulus [24]. In addition with short interstimulus intervals, around 1 Hz, attention seems to be not necessary to recruit or re-activate the neuronal circuits that were previously –not so long ago activated [25, 26, 27]. Thus we could assume that the task itself activates attention. It has been reported that not only attention, but also age and intelligence level, contribute to the enhancement of the beta ERD [9]. In addition as we have expressed before, alertness and attention have been related to spatiotemporal changes for the alpha band; lately they have been found to affect the gamma band while for the beta band, results are still controversial [28].

It should be kept in mind that for the imagined movement we cannot exclude some missed trials. Then the more smooth less power of the wavelet coefficients could be due to this fact or to a lesser degree of excitability from the underlying neuronal networks responsible for the preparation but not execution of the movement.

Possible effects of the auditory stimulation on the beta oscillations and their ERD/ERS properties can be excluded. Even though beta oscillations have been detected during the auditory control experiment they did not exhibit any desynchronization. Moreover, as this experiment was performed last, these beta oscillations during the auditory control stimulation could reflect evaluation of auditory information in order to prepare motor responses in posterior parietal cortex [29].

The promising results we have obtained in this pilot study suggest that ERD/ERS phenomena detected in this short interstimulus interval could help design a BCI application by recognizing imagination of a movement. High-quality features from wavelet coefficients can be used as input to a machine learning technique for classification of motor imagery EEG signals. Clearly, the present offline analysis results have to be further investigated during online settings, which consists the topic of our long term BCI research.

Acknowledgments. We thank the European Social Fund (ESF), Operational Program for Educational and Vocational Training II (EPEAEK II), and particularly the Program PYTHAGORAS II, for funding the above work.

Authors would like to thank Dr. Stefania Della Penna and Dr. Laura Cimponeriu.

References

1. Nicoletis, M.A.L.: Actions from thoughts. *Nature*. 409(2001) 403-407
2. Birbaumer, N., Kubler, A., Ghanayim, N., Hinterberger, T., Perelmouter, J., Kaiser, J., Iversen, I., Kotchoubey, B., Neumann, N., Flor, H.: The thought translation device (TTD) for completely paralyzed patients. *IEEE Trans Rehabil Eng*. 2000 8(2)190-193
3. Pfurtscheller, G., Neuper, C., Muller, G.R., Obermaier, B., Krausz, G., Schlogl, A., Scherer, R., Graimann, B., Keinrath, C., Skliris, D., Wortz, M., Supp, G., Schrank, C.: Graz-BCI: state of the art and clinical applications. *IEEE Trans Neural Syst Rehabil Eng*. 11 (2003) 177-180
4. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain-Computer Interfaces for communication and control. *Clin Neurophysiol*. (2002) 113: 767-791
5. Pfurtscheller, G., Brunner, C., Schlogl, A., Lopes da Silva, F.H.: Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *Neuroimage*. 31 (2006) 153-159
6. Hung, C.I., Lee, P.L., Chen, L.F., Yeh, T.C., Hsieh, H.C.: Recognition of Motor Imagery Electroencephalography Using Independent Component Analysis and Machine Classifiers *Ann Biomed Engineer*. 33 (2005) 1053-1070
7. Qin, L., Ding, L., He, B.: Motor Imagery classification by means of source analysis for brain-computer interface applications. *J Neural Eng.* 1 (2004) 135:141
8. Neuper, C., Pfurtscheller, G.: Motor imagery and ERD. In: Pfurtscheller G., Lopes da Silva F. (eds.): *Event-Related Desynchronization. Handbook of Electroencephalography and Clinical Neurophysiology*. 6th Revised edition, Elsevier, Amsterdam (1999) 303-325
9. Pfurtscheller, G., Lopes da Silva, F.H.: Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*. 110 (1999) 1842-1857.
10. Della Penna, S., Torquati, K., Pizzella, V., Babiloni, C., Franciotti, R., Rossini, P.M., Romani G.L.: Temporal dynamics of alpha and beta rhythms in human SI and SII after galvanic median nerve stimulation A MEG study. *NeuroImage*. 22 (2004) 1438-1446
11. Pfurtscheller, G., Neuper, C., Brunner, C., Lopes da Silva, F.: Beta rebound after different types of motor imagery in man. *Neurosci Lett* 378 (2005) 156-159
12. Oldfield ,R.C.: The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*. 9 (1971) 97-113
13. Kayser, J., Tenke, C.E.: Principal components analysis of Laplacian waveforms as a generic method for identifying ERP generator patterns: I. Evaluation with auditory oddball tasks. *Clin Neurophysiol* 117 (2006) 348-368
14. Nunez, P.L., Westdorp, A.F.: The surface Laplacian, high resolution EEG and controversies. *Brain Topogr*. 6 (1994) 221-226
15. Tallon-Baudry, C., Bertrand, O., Delpuech, C., Pernier, J.: Oscillatory γ -Band (30-70 Hz) activity Induced by a Visual Search Task in Humans. *J Neurosci*. 17(1997) 722-734
16. Jensen, O., Tesche, C.D.: Frontal theta activity in humans increases with memory load in a working memory task. *Eur J Neurosci*. 15 (2002) 1395-1399
17. Jensen, O. 4-D toolbox, version 1.1, A Matlab toolbox for the analysis of Neuromag Data.
18. Jurkiewicz, M.T., Gaetz, W.C., Bostan, A.C., Cheyne, D.: Post-movement beta rebound is generated in the motor cortex: Evidence from neuromagnetic recordings. *NeuroImage* 32 (2006) 1281-1289
19. Salmelin, R., Hamalainen, M., Kajola, M., Hari, R.: Functional segregation of movement-related rhythmic activity in the human brain. *Neuroimage*. 2(1995) 237-243

20. Neuper, C., Pfurtscheller, G.: Post-movement synchronization of beta rhythms in the EEG over the cortical foot area in man. *Neurosci Lett* 216(1996) 17-20
21. Michelon, P., Vettel, J.M., Zacks, J.M.: Lateral somatotopic organization during imagined and prepared movements. *J Neurophysiol.* 95 (2006) 811-822
22. Caldara, R., Deiber, M.P., Andrey C., Michel, M.C., Thut, G., Hauert, C.A.: Actual and mental motor preparation and execution: a spatiotemporal ERP study. *Exp Brain Res.* 159 (2004) 389-399
23. Pfurtscheller, G., Neuper, C., Flotzinger, D., Pergenzer, M.: EEG-based discrimination between imagination of right and left hand movement. *Electroencephalogr Clin Neurophysiol.* 103(1997) 642-651
24. Kincade, J.M., Abrams, R.A., Astafiev, S.V., Shulman, G.L., Corbetta, M.: An event-related functional magnetic resonance imaging study of voluntary and stimulus-driven orienting of attention. *J Neurosci.* 25(2005) 4593-4604
25. Miyake, Y., Onishi, Y., Poppel, E.: Two types of anticipation in synchronization tapping. *Acta Neurobiol Exp. (Wars)* 64 (2004) 415-426
26. Heuer, H., Spijkers, W., Kleinsorge, T., van der Loo, H.: Period Duration of Physical and Imagery Movement Sequences Affects Contralateral Amplitude Modulation. *Q J Exp Psychol A.* 51 (1998) 755-779
27. Woodrow, H.: The effect of rate of sequence upon the accuracy of synchronization. *J Exp Psychol* 15(1932) 357-379
28. Bauer, M., Oostenveld, R., Peeters, M., Fries, M.: Tactile Spatial Attention Enhances Gamma-Band Activity in Somatosensory Cortex and Reduces Low-Frequency Activity in Parieto-Occipital Areas. *J Neurosci.* 26 (2006) 490-501
29. Pesonen M., Bjornberg CH., Hamalainen H., Krause CM.: Brain Oscillatory 1-30 Hz EEG ERD/ERS responses during the different stages of an auditory memory search task, *Neurosci Lett.* 399 (2006) 45-50

A Fully Bayesian Two-Stage Model for Detecting Brain Activity in fMRI

Alicia Quirós, Raquel Montes Diez, and Juan A. Hernández

University Rey Juan Carlos, Madrid, Spain

Abstract. Functional Magnetic Resonance Imaging (fMRI) is a non-invasive technique for obtaining a series of images over time under a certain stimulation paradigm. We are interested in identifying regions of brain activity by observing differences in blood magnetism due to haemodynamic response to such stimulus.

Here, we extend Kornak (2000) work by proposing a fully Bayesian two-stage model for detecting brain activity in fMRI. The only assumptions that the model makes about the activated areas is that they emit higher signals in response to an stimulus than non-activated areas do, and that they form connected regions, providing a framework for detecting activity much as a neurologist might.

Due to the model complexity and following the Bayesian paradigm, we use Markov chain Monte Carlo (MCMC) methods to make inference over the parameters. A simulated study is used to check the model applicability and sensitivity.

1 Introduction

Magnetic resonance imaging (MRI) is a method used to visualise the inside of living organisms which does not require painful interventions and is widely used in medical diagnosis support. The fundamental research to develop MRI techniques started at the beginning of 19th century. However the image obtaining technology could not be developed until the appearance of high speed computers. The sensitivity of magnetic resonance to changes in blood oxygen level was discovered at the beginning of the 1980's and this great advance caused the advent of functional magnetic resonance imaging. As it was already known that brain haemodynamics changes (blood flow and blood oxygenation) were closely linked to neural activity, fMRI appeared to be a promising tool to use on to track physiological activity. Image acquisition with BOLD (blood oxygen level dependency) contrast is based on that principle. Since 1990, fMRI has contributed with a deeper knowledge of the brain functions to neuroscience; and in other fields, a better understanding of the physiology of other organs. Scientists use it to analyze changes in the brain activity in patients with certain pathologies with the objective of developing treatments and more effective therapies. It also serves as a guide to surgeons during operations for the pain to be minimised.

Essentially, magnetic resonance machines build the image by analysing structural changes in an electromagnetic wave sent to hydrogen atom protons in the

different tissues of the human body. The grey level of each voxel in the image corresponds to the behavior of lots of protons and it allows us to distinguish between tissues since it defines the image contrast.

Once they are generated, magnetic resonance images are digitally treated to improve its quality and to minimise error sources that could bias the subsequent statistical analysis. Such analysis is therefore posterior to pre-processing and its goal is to localise activity regions (in presence of noise). Currently, the most popular and complete statistical package, available in neuroscience is SPM 14, created by Frackowiak et al. (1997). Voxel classification is carried out by thresholding the value of a certain statistic in each voxel. In fMRI, the general linear model (GLM) is used, in combination with a convolution based temporal model, to identify functionally specialized brain responses. Descombes (1998) proposes a two stage procedure to analyse fMRI data. In the first stage a image restoration is performed by the means of a MRF that plays the role of a smoothing filter. The second stage is a spatio-temporal procedure based on MRFs applied to the restored data to make inference over the activity patterns. A high level analysis model to determine the location and magnitude of activity regions is the one proposed by Hartvig (1999) in which the activity pattern is modeled considering that an activity region takes the form of a bivariate Gaussian density function and the placement of the modes of these regions are represented by a Strauss point process. Hartvig and Jensen (2000) suggest a model in which the activity in a voxel depends, probabilistically and in a Markovian way, on its neighbours' responses. From 2001 on, the fMRI analysis came to be considered as a non linear dynamics problem and, thanks to the increasing ability of dealing with more complex models, the model of Wang et al (2003), based on Support Vector Machines (SVMs), appeared. In the same line, but using wavelets, there is the work developed by Meyer (2003).

Taking into account the high dimensions of data and that, in most cases, we have prior information about activity (as location, shape, magnitude, ...), Bayesian statistics constitutes an ideal framework to carry out neuroimage analysis. Kornak (2000) suggests a two stage model to localise activity regions in fMRI. At the first stage, the observed haemodynamic response function is fitted, pixelwise, with the help of least squares technique. The resulting map of activity values is past on to the second stage where a spatial model is fixed by using a Bayesian approach. This spatial model is based on Markov random fields in order to set restrictions on the activity patterns smoothness as prior information. The formulation of the model is made in a simple and natural way thanks to a small number of parameters.

2 Fully Bayesian Two-Stage Model for fMRI

In this work, we extend Kornak's model by introducing a totally Bayesian temporal analysis of the fMRI data at the first stage.

In particular, we analyse box-car shaped fMRI data to find areas of brain activity. A box-car shaped fMRI experiment consists of several scans taken across

time under two different conditions: the stimulus is on during $T/2$ images, it is off during the next $T/2$ ones, on again, off again and so on. It is well known that activated areas emit higher signals in response to a stimulus than non-activated do, which may be used to find activity into the brain.

Let us denote the resulting magnetic resonance images by y_{stk} , where $s = 1, \dots, N \times M$; $t = 0, \dots, T - 1$ and $k = 1, \dots, K$ are the indexes for pixels, scans in a cycle and cycles, respectively. Thus there are K cycles of T images of dimension $N \times M$, half of them taken while the stimulus is on.

2.1 First Stage: Temporal Inference

At the (temporal) first-stage, we are interested in modelling the haemodynamic response, independently for each pixel s , without taken in to account the spatial dependence of our fMRI data. So that for each pixel, the corresponding data y_s describe a time series $\{y_{s01}, y_{s11}, \dots, y_{s(T-1)1}, \dots, y_{s0K}, y_{s1K}, \dots, y_{s(T-1)K}\}$, corresponding to the T images of each one of the K cycles in the experiment.

Figure 1 represents common time series corresponding to an activated and a non-activated pixels.

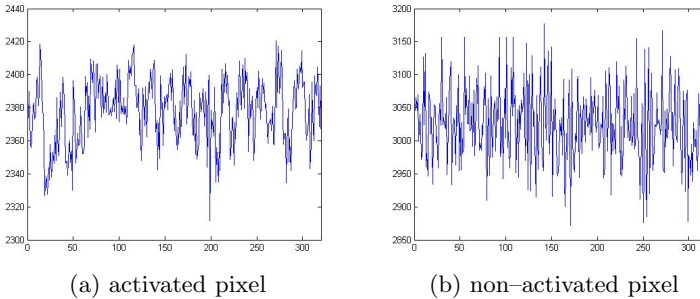


Fig. 1. Examples of two time series from an fMRI experiment

The archetypical haemodynamic response (HDR) function (see figure 2) describes the local response to the oxygen utilisation (by nerve cells) and it consists of an increase in blood flow to regions of increased neural activity, occurring after a delay of approximately 1-5 seconds. This haemodynamic response rises to a peak over 4-8 seconds, before falling back to baseline (and typically under-shooting slightly). This leads to local changes in the relative concentration of oxyhaemoglobin and deoxyhaemoglobin as well as local changes in brain blood volume and flow. This delayed response is characterized by $h(t)$, where t is the time since activity started.

The optimum choice to parametrise the HDR is open to debate. The general opinion is that the most efficient way to describe this curve is to model it by a function containing few parameters, but with the ability to extract adequately the most important features of the signal. Friston and Turner (1994) suggested the Poisson curve; Lange and Zeger (1997), the Gamma curve and Rajapakse et al.

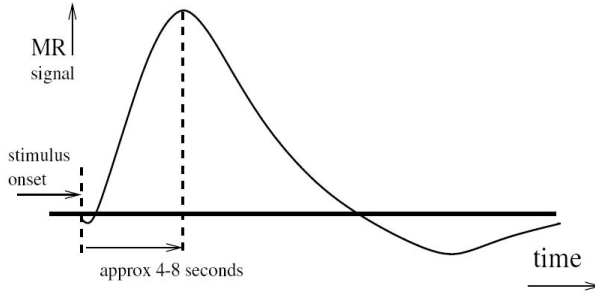


Fig. 2. Archetypal haemodynamic response curve

(1998), the Gaussian one. Other authors employed not that simple curves, as combination of cosines (see, for instance, Woolrich et al. (2004)).

Here, we adopt the Friston and Turner (1994) proposal and we model the haemodynamic response, by the Poisson distribution density function, scaled and shifted, this is

$$h_s(t) = \begin{cases} c_s \frac{\lambda_s^{t-1} e^{-\lambda_s}}{(t-1)!} & t = 1, 2, \dots, T-1 \\ 0 & t = 0 \end{cases} \quad (1)$$

So that, for each $k = 1, \dots, K$, our model may be express by

$$y_{stk} = h_s(t) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \delta^2) \quad (2)$$

where $c_s \in [-\infty, \infty]$, $\lambda_s \in [0, \infty]$ and $\delta^2 \in [0, \infty]$ are the unknown parameters.

Following a Bayesian approach, the joint posterior distribution of the unknown parameters is given by

$$p(c_s, \lambda_s, \delta^2 | y) \propto p(y | c, \lambda_s, \delta^2) \pi(c_s) \pi(\lambda_s) \pi(\delta^2) \quad (3)$$

where the likelihood function is given by

$$p(y_s | c_s, \lambda_s, \delta^2) = \left(\frac{1}{2\pi\delta^2} \right)^{\frac{KT}{2}} \exp \left\{ -\frac{1}{2\delta^2} \sum_{k=1}^K \sum_{t=0}^{T-1} [y_{s(t+1)k} - h_s(t)]^2 \right\} \quad (4)$$

and where $\pi(c_s)$, $\pi(\lambda_s)$ and $\pi(\delta^2)$ denote the corresponding prior distributions, which according to the Bayesian paradigm allow us to introduce any available prior information about the parameters. Here we adopt the following general forms for each one of the prior distributions:

$$\begin{aligned} c_s &\sim N(\mu_c, \sigma_c^2) &\Rightarrow &\pi(c_s) \propto \exp \left\{ -\frac{(c_s - \mu_c)^2}{\sigma_c^2} \right\} \\ \lambda_s &\sim G(l_1, l_2) &\Rightarrow &\pi(\lambda_s) \propto \lambda_s^{l_1-1} e^{-\lambda_s/l_2} \\ \delta^2 &\sim IG(d_1, d_2) &\Rightarrow &\pi(\delta^2) \propto \left(\frac{1}{\delta^2} \right)^{d_1-1} \exp \left\{ \frac{-1}{d_2\delta^2} \right\} \end{aligned}$$

Thus, the posterior (3) becomes

$$p(c_s, \lambda_s, \delta^2 | y) \propto \left(\frac{1}{\delta^2} \right)^{\frac{KT}{2}} \exp \left\{ -\frac{1}{2\delta^2} \sum_{k=1}^K \sum_{t=0}^{T-1} \left(y_{s(t+1)k} - c_s \frac{\lambda_s^{t-1} e^{-\lambda_s}}{(t-1)!} \right)^2 \right\} \\ \left(\frac{1}{\delta^2} \right)^{d_1-1} \exp \left\{ -\frac{1}{d_2 \delta^2} \right\} \lambda_s^{l_1-1} \exp \left\{ -\frac{\lambda_s}{l_2} \right\}$$

In order to make inference about the unknown quantities in our model and given the complex expression of the posterior distribution, MCMC techniques are required here. In Particular, by using the prior distributions proposed above, we can automatically employed the Gibbs algorithm for the c_s and δ^2 parameters. The marginal distribution of λ_s , however, does not have a known form, so that a Metropolis Hasting algorithm is required, (see Gilks et.al. (1996) for details).

MCMC algorithm. We propose the following hybrid MCMC posterior sampling scheme to implement inference in the proposed model 2.

1. Initialize parameters c_s , λ_s and δ^2 .
2. Given current values of c_s and λ_s , generate new values for parameters δ^2 , drawing from its full conditional posterior:

$$\delta^2 | y, c_s, \lambda_s \sim IG \left(\frac{KT}{2} + d_1, \left\{ \frac{1}{d_2} + \frac{1}{2} \sum_{k=1}^K \sum_{t=0}^{T-1} [y_{stk} - h_s(t)]^2 \right\}^{-1} \right)$$

3. Given current values of λ_s and δ^2 , generate new values for parameters c_s , drawing from their full conditional posteriors:

$$c_s | y, \lambda_s, \delta^2 \sim \mathcal{N} \left(\frac{\sum_{k=1}^K \sum_{t=0}^{T-1} \left[\frac{\lambda_s^{t-1} e^{-\lambda_s}}{(t-1)!} \right] y_{stk}}{K \sum_{t=0}^{T-1} \left[\frac{\lambda_s^{t-1} e^{-\lambda_s}}{(t-1)!} \right]^2}, \frac{\delta^2}{K \sum_{t=0}^{T-1} \left[\frac{\lambda_s^{t-1} e^{-\lambda_s}}{(t-1)!} \right]^2} \right)$$

4. Finally, given current values of δ^2 and c_s , for each λ_s , $s = 1, \dots, M \times N$ generate a proposal $\tilde{\lambda}_s \sim N(\lambda_s, \sigma_{\lambda_s})$, and calculate the acceptance probability

$$\alpha_s = \min \left[1, \frac{p(\tilde{\lambda}_s | y, c_s, \delta^2)}{p(\lambda_s | y, c_s, \delta^2)} \right]$$

where the marginal posterior distribution for λ_s is

$$p(\lambda_s | y, c_s, \delta^2) \propto \lambda_s^{l_1-1} \exp \left\{ -\frac{\lambda_s}{l_2} - \frac{1}{2\delta^2} \sum_{k=1}^K \sum_{t=0}^{T-1} \left(y_{s(t+1)k} - c_s \frac{\lambda_s^{t-1} e^{-\lambda_s}}{(t-1)!} \right)^2 \right\}$$

With probability α_s replace λ_s by $\tilde{\lambda}_s$, otherwise, leave λ_s unchanged.

It is of interest to observe here that by implementing the above MCMC algorithm for the proposed model and according to the Bayesian approach, we obtain posterior estimation maps of the brain activity in the temporal dimension, as well as a measured of the uncertainty involved in such estimation.

2.2 Second Stage: Spatial Inference

The MCMC algorithm described in the previous section provides an estimative map of the activation areas in the brain. Note, however that such estimation procedure has taken no account of the (known) spatial structure inherent in neural activation on the BOLD effect.

At the second stage of the procedure, we are interested in incorporating this spatial auto-correlation structure into inference about activation patterns, by following the model and Bayesian approach proposed by Kornak (2000),

$$a = z \odot x + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

where

- a is the activation map estimated in the first stage
- z is a binary indicator map of activation defined by

$$z|w = I_{w>0}$$

where I is the indicator function and w is modelled as an improper Gaussian Markov random field (GMRF) of first order, defined by

$$\pi(w) \propto \exp\left\{-\frac{1}{2} \sum_{\langle s, t \rangle} (w_s - w_t)^2\right\}$$

- x is the activity level pattern, also modelled as the natural logarithm of a proper GMRF of the first-order conditional auto-regressive (CAR):

$$\log(x) \sim NMV(\mu 1, \kappa^2(I - \beta N)^{-1})$$

Making use of MRFs as prior distributions for the image of interest, activity location in a pixel is determined by the response magnitude in it and also by the activity evidence in its neighbouring pixels. That is, we place, a priori, the expectation that activity takes the form of regions conformed of several neighbour pixels (but not that it is in isolated pixels).

Again, MCMC techniques are required here in order to perform Bayesian inference, (see Kornak (2000)).

3 Simulated Study

In order to check the model applicability and sensitivity, we perform a fMRI data simulation. We create this simulated study from a real fMRI rest study provided by the Ruber Internacional hospital of Madrid with a 3T machine.

A central slice was selected from the study and convolved with an activation pattern of the form of a Gamma distribution in the temporal dimension and of different geometric figures submitted to different signal intensity (2% y 3% of the signal) to model spatial activation regions. Figure 3 shows the activity phantom

created for our experiment. By choosing this phantom, we are interested in observing that our Bayesian two-stage approach is able to discriminate pixels submitted to the activation convolution, identifying them as non-active pixels (if they are isolated) or active pixels (if they form part of a convex form), in the same way a neurologist would, (see section 2.2).

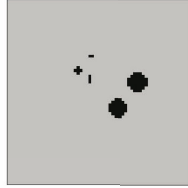


Fig. 3. Activity phantom used to simulate the response paradigm to an stimulus in the simulated study

The simulated study resembles a epoch experiment with 310 scans, corresponding to $K = 15$ cycles of $T = 20$ bidimensional images of $N \times M = 64 \times 64$ pixels each. In each cycle, the tenth first images are taken while the stimulus is on and the ten last images when it ceases. Figure 4 shows several slices of the

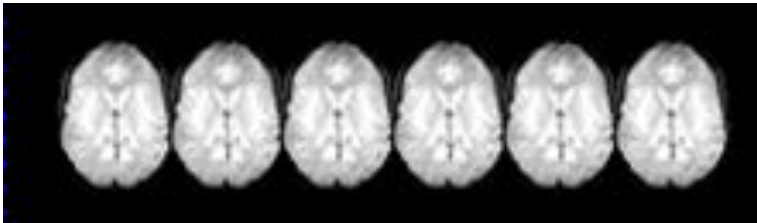


Fig. 4. Several slices of the study

simulated study, taken during both the rest and the activity periods. We observe that, to human eye, there are no notable differences between them, so that, we cannot discriminate activity regions at sight.

4 Results

In order to fit the haemodynamic response curve in the temporal dimension, independently for each one of the $N \times M$ pixels, we use the (temporal) MCMC algorithm described in section 2.1. A total of 20000 simulations were performed, 5000 of which were discarded as burn-in for the algorithm. Subsequently, 15000 samples were used for making posterior inference.

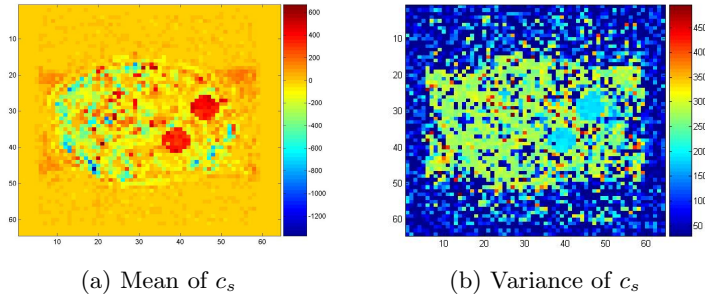


Fig. 5. Estimation for c_s for $s = 1, \dots, M \times N$

As mentioned before, Kornak (2000) proposed the mean estimated map of c , obtained at the first stage of the procedure, to be past on to the second stage as the input of his (spatial) MCMC algorithm (section 2.2). By using the Bayesian MCMC algorithm proposed here, we are able to obtain not just a point estimation of the c parameter but also an estimation of its variance, which give us an idea of the present uncertainty (see figure 5).

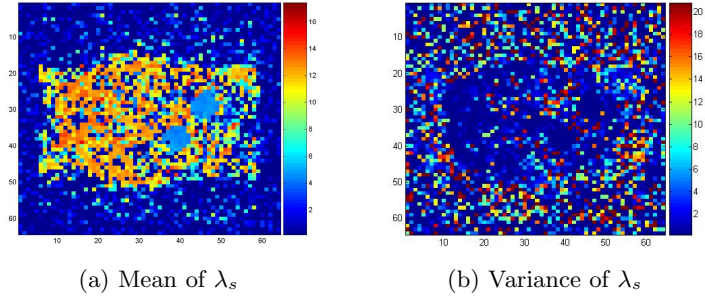


Fig. 6. Estimation for λ_s for $s = 1, \dots, M \times N$

It is also of interest to take into account the λ mean and variance estimation maps. Looking at figure 6 we observe that this provides similar information to the estimation maps of c . We consider this information should therefore be past on to the second stage of the procedure.

Again, by following a Bayesian approach, and by averaging the values taken by the model (2) over the 15000 MCMC iterations, we are able to obtain the estimated mean for the haemodynamic response curve, as well as the corresponding 95% high probability interval (HPI). Figure 7 shows the Haemodynamic response curve fitting for two (activated and non-activated) pixels. It can be observed that the empirical curve (in blue) is essentially noisy and that it adopt the archetypical shape of the haemodynamic response described above. The approximated haemodynamic response curve is shown in red.

For illustration purposes, figure 8 shows the mean and 95% HPI, which provides a measure of the uncertainty present in our fitting procedure, for the last

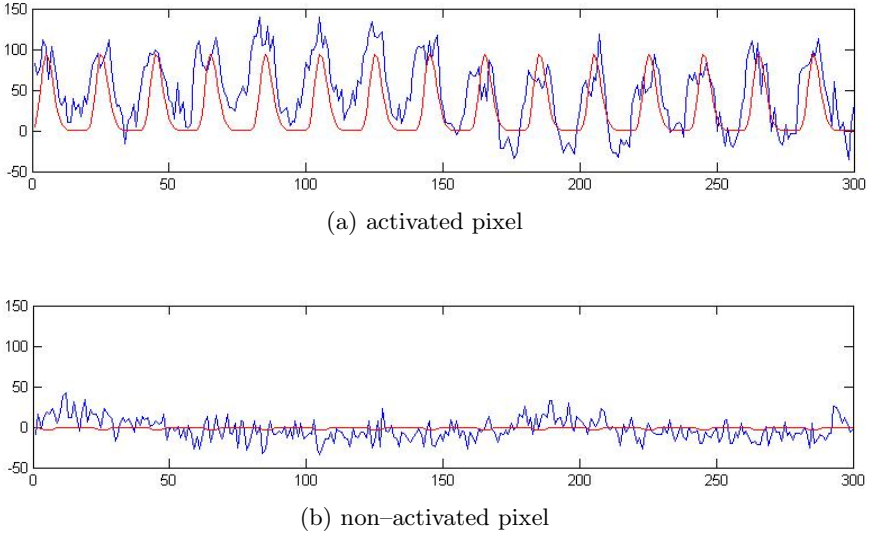


Fig. 7. Haemodynamic response curve fitting

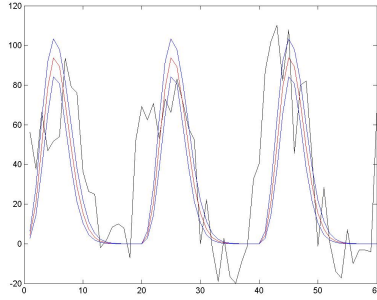


Fig. 8. Mean and HPI estimators for the haemodynamic response curve fitting in an activated pixel

three cycles of the activated pixel (so it can be graphically appreciated). Note that the uncertainty observed in the estimation of the unknown parameters is obviously transmitted when fitting the empirical time series for each one of the $N \times M$ pixels.

In order to use both the information provided by the parameter c and the parameter λ , we define the activity a_s for a given pixel s , as the maximum value of estimated haemodynamic response curve, over the T images of a cycle, this is

$$a_s = \max_t h_s(t)$$

Again, by averaging the values taken by the activity, a_s over the 15000 MCMC iterations, we obtain the estimated mean and variance of the activity in each pixel s .

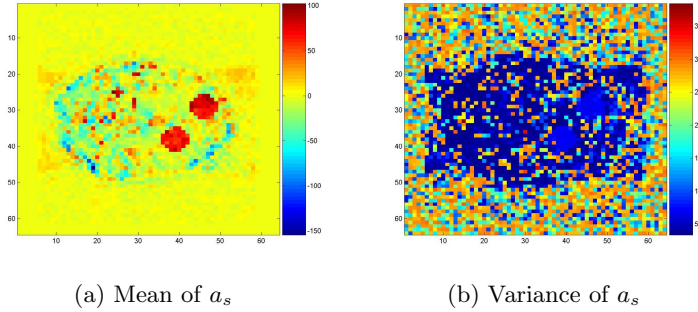


Fig. 9. Estimation of the activity $a_s = \max_t h_s(t)$ for $s = 1, \dots, M \times N$

We then use the mean activated map (figure 9 (a)) as the input to be introduced in Kornak (2000) second-stage model. As a final result, figure 10 shows a binary indicator map of activation defined by those pixels for which the average value of z is greater than 0.80, this is with a high probability of being activated pixels. Comparing the activity regions shown in figure 10 with the activity phantom (figure 3) used to simulate the response paradigm, we observe how the isolated pixels introduced by the phantom have been discarded as non-active pixels, whereas the connected regions are identified as activity regions.

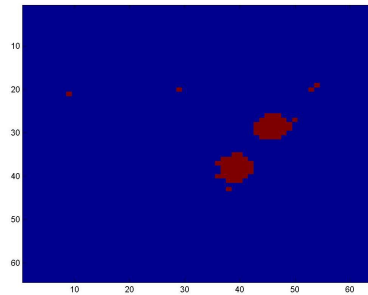


Fig. 10. Binary activation map

5 Conclusions

In this work, we extend Kornak (2000) two-stage procedure to identifying brain activity regions in fMRI data. Essentially, our approach, introduces a Bayesian approach at the first stage of the procedure, which allows us to obtain measures

of the uncertainty present in the estimated parameters and estimated responses of the model. Due to the model complexity and following the Bayesian paradigm, we use a hybrid Markov chain Monte Carlo algorithm which combines Gibbs and Metropolis Hastings steps to make inference over the parameters.

The two-stage procedure avoids decisions based on thresholding of the value of certain estimated statistic in each pixel, and it proves to be able to identify activity regions, and to reject possible activation in isolated pixels, in the same way that a neurologist might judge.

Acknowledgments

This research was partially supported by grants from URJC, CAM and MCYT. One of the authors (AQC) gratefully acknowledges receipt of a Research Studentship from the University Rey Juan Carlos.

References

- DESCOMBES, X., KRUGGEL, F. AND VON CRAMON, D.Y., *Spatio-temporal fMRI analysis using Markov random fields*, IEEE Transactions on Medical Imaging, 17:1028-1039, 1998.
- FRACKOWIAK, R.S.J., FRISTON, K.J., FRITH, C.D., DOLAN, R.J. AND MAZZIOTA, J.C., *Human Brain Function*, Academic Press, 1997.
- FRISTON, K.J. AND TURNER, R., *Analysis of functional MRI time series*, Human Brain Mapping, 1:153-171, 1994.
- GILKS, W.R., RICHARDSON, S., SPIEGELHALTER, D.J., *Markov Chain Monte Carlo in Practice*, Chapman & Hall, 1996.
- HARTVIG, N.V., *A stochastic geometry model for fMRI data*, Technical Report 410, University of Aarhus, 1999.
- HARTVIG, N.V. AND JENSEN, J.L., *Spatial mixture modelling of fMRI data*, Technical Report 414, University of Aarhus, 2000.
- KORNAK, J., *Bayesian Spatial Inference from Haemodynamic Response Parameters in Functional Magnetic Resonance Imaging*, University of Nottingham, PhD. Thesis, 2000.
- LANGE, N. AND ZEGER, S.L., *Non-linear Fourier time series analysis for human brain mapping by functional magnetic resonance imaging (with discussion)*, Journal of Applied Statistics, 46(1): 1-29, 1997.
- MEYER, F.G. AND CHINRUNGRUENG, J., *Clustering of spatiotemporal signals: application to the analysis of fMRI data*, IEEE Transactions on Medical Imaging, 22(8), 933-939, 2003.
- RAJAPAKSE, J.C., KRUGGEL, F., MAISOG, J.M. AND VON CRAMON, D.Y., *Modeling hemodynamic response for analysis of functional MRI time-series*, Human Brain Mapping, 6:283-300, 1998.
- WANG, Y., SCHULTZ, R.T., CONSTABLE, R.T., STAIB, L.H., *Nonlinear estimation and modeling of fMRI data using spatio-temporal support vector regression*, Information Processing in Medical Imaging Proceedings, 647-659, 2003.
- WINKLER, G., *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods. A Mathematical Introduction*, Berlin [etc.] : Springer, cop. 2003.

WOOLRICH, M.W., JENKINSON, J.M., BRADY, J.M. AND SMITH, S.M., *Fully bayesian spatio-temporal modeling of fMRI data*, IEEE Transactions on Medical Imaging, 23(2), 2004.
<http://www.fil.ion.ucl.ac.uk/spm/>

A Novel Algorithm for Segmentation of Lung Images

Aamir Saeed Malik¹ and Tae-Sun Choi²

^{1,2} Gwangju Institute of Science and Technology,
1 Oryong-Dong, Buk-Gu, Gwangju, 500712, Korea
¹ aamir@gist.ac.kr, ² tschoi@gist.ac.kr

Abstract. Several image segmentation techniques have been presented in the literature applied in the medical domain. However, there are few multiscale segmentation methods that can segment the medical image so that various components within the image could be separated at multiple scales. In this paper, we present a new segmentation method based on an optical transfer function implemented in the Frequency domain. With this new segmentation technique, we demonstrate that it is possible to segment the High Resolution Computed Tomographic (HRCT) images into its various components at multiple scales hence separating the information available in HRCT image. We show that the HRCT image can be segmented such that we get separate images for bones, tissues, lungs and anatomical structures within lungs. The processing is done in frequency domain using the Fast Fourier Transform and Discrete Cosine Transform. Further, we propose an algorithm for extraction of anatomical structures from the segmented image.

Keywords: Segmentation, optical transfer function.

1 Introduction

Segmentation methods can be broadly categorized into three major classes, that is, supervised segmentation, unsupervised segmentation (automated segmentation) and semi-supervised (semi-automated) segmentation. The application of these methods largely depends on the various medical imaging modalities and the complexity of diagnosis. Various imaging modalities include but are not limited to X-Ray, CT, MRI, fMRI, PET, SPECT etc. Some of them provide images in 2D domain while some result in 3D domain. Similarly, some just provide result in one image while some provide a sequence of images. There is a wide range of resolution, information per pixel etc considering all the imaging modalities. Similarly, the diagnosis can vary from very simple task to a very complex one hence again providing a very wide range.

In clinical practice, the physicians have access to large amount of data. This large amount of data has been made available because of the medical imaging instrumentation. Different anatomical features are captured in various orientations using different imaging procedures. Sometimes physicians must combine the information from different images to fully visualize the imaged anatomical structure.

Considering the High Resolution Computed Tomographic (HRCT) images, the objective is to assess a variety of lung diseases, for example, pulmonary emphysema, nodules, interstitial lung disease etc. The advantage of HRCT image is increasingly better anatomic resolution because of very thin image slices. However, as a consequence of thin slices, we get a sequence of images consisting of large number of slices for a single patient. This fact results in time consuming visual assessment of all the images. According to [1], experienced observers typically make correct global diagnosis of parenchymal lung diseases in 40% to 70% cases.

Therefore, the need arises for some type of automated segmentation technique to analyze the HRCT images. Although this problem has been addressed by many in the literature, segmentation of HRCT images into homogeneous regions still remains an open issue. Some have proposed semi-automated segmentation methods while others have demonstrated automated segmentation techniques therefore making it an active area of research today.

The objective of image segmentation is to partition the image into smaller regions. These regions are formed based on some type of feature and characteristic which is same for all the pixels lying within the same region and different for pixels among different regions. One way of segmentation is to first segment the whole image into smaller regions based on some unique feature or attribute of the pixels and then label each region in the second step based on the difference of segmented values of the pixels lying in the different regions. Finally separate each region from the image. Each of the regions corresponds to separate anatomical structure present within the whole image.

The other way is to devise some method so that the image can be segmented based on its various scales or resolutions. The result of such segmentation is that the various anatomical structures within the same image are extracted separately. This type of segmentation is classified as multiscale segmentation. In this paper, we present a technique for multiscale segmentation based on an optical transfer function implemented in the frequency domain.

With above considerations in mind, this paper presents our research related with the multiscale segmentation of HRCT images with the aim to segment various anatomical structures present within the image. The rest of the paper is organized as follows: the related work done in this field is discussed in section 2. Section 3 outlines the technique and method proposed for multiscale segmentation of HRCT images. Section 4 gives description of the experiments by employing the method described in section 3. Section 5 discusses an algorithm for the extraction of lungs and bones from the HRCT image that is segmented in section 4. Section 6 concludes this paper followed by references.

2 Previous Work

As mentioned in section 1, there are various methods and techniques leading to supervised as well as semi-supervised segmentation of medical images. The aim of all these methods is to finally extract various anatomical structures that are generally present within an image so that such segmentation can assist the doctors and radiologists for diagnosis of diseases.

Brown et. al. [2] used explicit anatomical knowledge to generate an anatomical model. They incorporated anatomical knowledge such as expected size, shape, relative position of objects in the same or adjacent slices etc within their model. This model was developed with the guidance from experienced radiologists. The image processing routines were guided using this model to differentiate between various objects. They showed [3] that 86% of the lung segmentation was correct. On the other hand, 14% required manual corrections. 104 data series were considered comprising of 1313 images.

Hu et. al. [4] used an iterative searching method to compute an optimal threshold value for each CT case and use conditional morphological operations to segment lung regions. The basic morphological operations of dilation and erosion were used. First the algorithm checks whether the left and right lungs are connected or not. If they are connected then the algorithm iteratively applies morphological filters till the two lungs are separated. Also, a separate set of morphological filters were used to delete the major pulmonary vessels or airways connected to or inside the lung areas. However, appropriate selection of kernel sizes of the filters may require human intervention due to the large variation of lung structures.

Zheng et. al. [5] suggested using artificial neural networks to classify each pixel in the CT slice into different anatomical structure. For each pixel, they used 48 pixel values around the POI (Pixel Of Interest), that is, 7x7 neighborhood, as an input to the artificial neural network. The distribution pattern of the pixel values in a local region surrounding the POI were used for training of the artificial neural network.

Zheng et. al. [6] suggested a fully automated segmentation process that included a series of six simple steps: (1) filtering and removing pixels outside the scanned anatomic structures, (2) segmenting the potential lung areas using an adaptive threshold based on pixel value distribution in each CT slice, (3) labeling all selected pixels into segmented regions and deleting isolated regions in the non-lung area, (4) labeling and filling interior cavities (e.g., pleural nodules, airway wall, major blood vessels etc) inside lung areas, (5) detecting and deleting the main airways (e.g., trachea, central bronchi etc) connected to the segmented lung areas, and (6) detecting and separating possible anterior or posterior junctions between the lungs. The method was applied to CT scans of 50 patients.

Kuhnigk et. al. [7] used anatomy guided 3D watershed transform for lung lobe segmentation. To make use of image regions with visible fissures, they linearly combined the original data with distance map. The segmentation itself was performed on the combined image using an interactive 3D watershed algorithm which allows an iterative refinement of the results. The method was applied to CT scans of 24 patients.

Lobar segmentation techniques that rely on fissure detection dominate the literature. Zhang et. al. [8] proposed an interactive method to extract the oblique fissures using a fuzzy reasoning system followed by a graph search. They later [9] extended the same system to include 3D shape constraints.

Wie [10] proposed using a DCT descriptor for each pixel of the image. An adaptive k-means clustering algorithm was used for assignment of labels to each pixel. Jei Wei did some adjustment to get the appropriate number of clusters in the proposed algorithm. In the algorithm, clusters of size less than one percent of the number of pixels in the image were removed. This approach may remove small clusters which may have importance in medical applications.

3 Method

In this paper, we introduce a new method for segmentation of HRCT images with the aim to separate the various anatomical structures within the image.

The success of any segmentation method depends on how accurate the sharpness in image pixels could be detected. Therefore, algorithms and techniques based on calculating sharpness and edges in an image automatically become potential candidates for the segmentation process.

We introduce a new segmentation method based on bipolar incoherent image processing and we call it Optical Segmentation Technique and denote it as OS_O. T.-C. Poon and P. P. Banerjee [11] has discussed bipolar incoherent image processing in detail. Generally, there are severe limitations of incoherent processing with standard incoherent systems in that the Optical Transfer Function achievable is the autocorrelation of the pupil function. Or equivalently the Point Spread Function is real non-negative. Among the acousto-optic heterodyning image processing, a number of novel techniques have been devised to implement bipolar point spread functions in incoherent systems. These techniques are usually referred to as bipolar incoherent image processing in the literature. Our proposed segmentation technique is based on bipolar incoherent image processing.

The sharpness of pixel values in the image is found by convolving the spectrum of the intensity image with the transfer function which in our case is Optical Transfer Function (OTF). The computed image [$i_c(x, y)$] is given as:

$$i_c(x, y) = \text{Re} \{ |\Pi_0(x, y)|^2 * h_\Omega(x, y) \}, \quad (1)$$

where ‘*’ indicates convolution and:

$$\begin{aligned} |\Pi_0(x, y)|^2 &= \text{Spectrum of the Intensity Image} \\ h_\Omega(x, y) &= \text{Transfer Function} \end{aligned}$$

Transfer function is basically the OTF which is calculated in Fourier domain hence also simplifying the convolution operation to simple multiplication. The transfer function $h_\Omega(x, y)$ is given as:

$$h_\Omega(x, y) = F^{-1} \{ \text{OTF}_\Omega(k_x, k_y) \},$$

where:

$$\begin{aligned} \text{OTF}_\Omega(k_x, k_y) &= \text{Optical Transfer Function} \\ k_x, k_y &= \text{Spatial frequencies} \end{aligned}$$

So finally we can write the computed image from equation (1) as:

$$i_c(x, y) = \text{Re} \{ F^{-1} \{ F \{ |\Pi_0(x, y)|^2 \} \text{OTF}_\Omega(k_x, k_y) \} \}, \quad (2)$$

where F is for Fourier Transform and F^{-1} is for Inverse Fourier Transform. The OTF itself is calculated as:

$$\text{OTF}_\Omega(k_x, k_y) = \iint p_1(x', y') p_2\left(x' + \frac{fk_x}{k_0}, y' + \frac{fk_y}{k_0}\right) dx' dy' \quad (3)$$

Where f is the focal length of the lenses and k_0 is the wave number of light. The OTF is the cross correlation of the two pupils (p_1 and p_2) in the incoherent optical system. Therefore, the point spread function becomes bipolar.

In equation (3) above, p_1 is a difference of Gaussian aperture function and p_2 is a small pin hole aperture. p_1 is given as:

$$p_1 = \exp[-a_1(x^2 + y^2)] - \exp[-a_2(x^2 + y^2)],$$

where a_1 and a_2 are constants. p_2 is given as:

$$p_2 = \delta(x, y)$$

For implementation purposes, equation (3) can be rewritten as:

$$\text{OTF}_\Omega(k_x, k_y) = \exp[-\sigma_1(k_x^2 + k_y^2)] - \exp[-\sigma_2(k_x^2 + k_y^2)], \quad (4)$$

Where:

$$\begin{aligned} \sigma_1 &= a_1 (f/k_0)^2 \\ \sigma_2 &= a_2 (f/k_0)^2 \end{aligned}$$

So, now the OTF becomes a filtering operation that provides the sharpness at pixel points in an image. The filtering operation depends upon σ_1 and σ_2 . These values are adjusted to provide the desired filter shape and it can be adjusted to low pass, band pass and high pass filter operations. Therefore, the operator parameters can easily be adjusted to respond to the high frequency variations in the image intensity. The high frequency component of an image area is determined by processing in the Fourier domain and analyzing the frequency distribution. Fourier transform used to be computationally expensive but with high speed personal computers available today, this computational complexity has decreased exponentially and it is not a matter of concern anymore. The processing in the frequency domain is particularly useful for noise reduction also as the noise frequencies are easily filtered out. Fig 1 shows an HRCT image, its corresponding Fourier spectrum and the filter with $\sigma_1 = 0.01$ and $\sigma_2 = 0.1$.

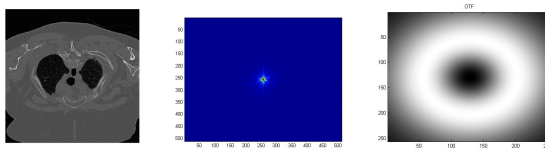


Fig. 1. HRCT Image, its Fourier Spectrum & filter designed with $\sigma_1 = 0.01$ & $\sigma_2 = 0.1$

The high frequency component of an image area can also be determined by processing with Discrete Cosine Transform (DCT) and analyzing the frequency distribution. So we can rewrite the computed image from equation (2) as:

$$i_c(x, y) = \text{Re} [\text{DCT}^{-1} \{ \text{DCT} \{ |I_0(x, y)|^2 \} \text{OTF}_\Omega(k_x, k_y) \}] \quad (5)$$

Fig 2 shows the spectrum obtained by using DCT for image shown in Fig 1 and the filter that is designed with parameter values of $\sigma_1 = 0.01$ and $\sigma_2 = 0.05$.

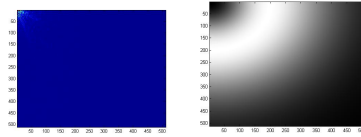


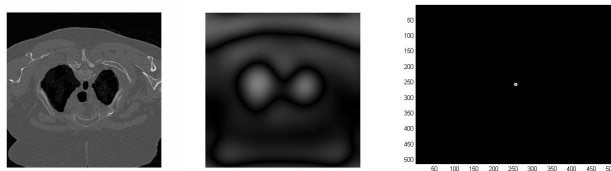
Fig. 2. Spectrum obtained using DCT & filter designed with $\delta_1=0.01$ and $\delta_2=0.05$

Once the OTF is applied to the HRCT images, we can select the boundary points for the segmented regions in the HRCT segmented image. So the selection at a point (i,j) can be computed in a small window around (i,j) and the value at (i,j) can be replaced by the sum of computed values (by equation 4 above) of all pixels in that window only after enhancing the extracted edges and removing noise, if any. This operation is similar to that used for Sum of Modified Laplacian [12,13]. We should use small window size of 3×3 because larger window size results in smoothing of the image and hence losing the actual sharp point. Therefore, the segmented point can also be given by the following equation where OS_O stands for Optical Segmentation technique:

$$OS_O(i, j) = \sum_{x=i-N}^{i+N} \sum_{y=j-N}^{j+N} i_c(x, y) \quad (6)$$

4 HRCT Image Filtering

A database of HRCT lung images of resolution 512×512 pixels were used to test the algorithm. All the images were of DICOM format. In fig 3, the results are shown when the HRCT image is processed using Fourier transform. Fig 3(a) is the original image. We demonstrate the results with one slice only for clarity purposes. It is the same image as shown in fig 1. So in all of the subsequent images, reference for the original image should be taken from fig 3(a). Fig 3(b) shows the extracted region when we design the filter with its parameters of $\sigma_1=10$ and $\sigma_2=20$. The filter designed with these values is shown in fig 3(c) and it can be seen that it is a low pass filter but with a lower cut off frequency too in addition to higher cut off frequency. Therefore, the DC component is not taken into account when the image is filtered using this filter.



(a) Original Image (b) extracted region (c) Filter ($\delta_1=10, \delta_2=20$)

Fig. 3. Original image with extracted region and corresponding filter

Fig 4 shows the different anatomical structures segmented by using a set of filters with different parameter values for σ_1 and σ_2 . See the initial three images in fig 4.

These images are the result of allowing only the high frequency part of the image to remain within the image while eliminating all other frequency components. Therefore, these images are filtered with high pass filter. There is one very interesting point that can be observed from the images, i.e., only the bones structure is visible in these images and we cannot observe lungs, tissues and other anatomical structures within lungs. So we can easily separate the bones from the rest of the anatomical structures present within the image. Hence, we can deduce that high pass filtering separates bones in a HRCT image. On the other hand, low pass filtering results in separation of lungs and tissues from the image.

The last image in fig 4 is low pass filtered with no information from medium and high frequencies. As a result, the image does not have sharp edges and it has the blurring effect. As we move towards the first image, edge information is now accounted for in those images. Therefore, the result is less blurring and increased sharpness of pixel values. Another point is quite clear from the last image, i.e., the lungs are quite pronounced in these set of images while the other anatomical structures (bones, structures within the lungs etc) are not obvious. Hence, using this image, we can easily separate lungs from the rest of the image. Also, considering last three images, we can see that there is some tissue information present in the image.

Now consider images in between the first and the last images. These are band pass filtered images. So they have medium frequency component present in the images. We can see from the figures that the bones start to appear in the images while lungs remain present in the image too. Hence, by considering all the images in fig 4, we can see that it is possible to separate the tissues, lungs, bones and the anatomical structures within the lungs from the HRCT images. Therefore, we can perform the multiscale segmentation by using the filter based on optical transfer function to segment the HRCT image at various levels or scales.

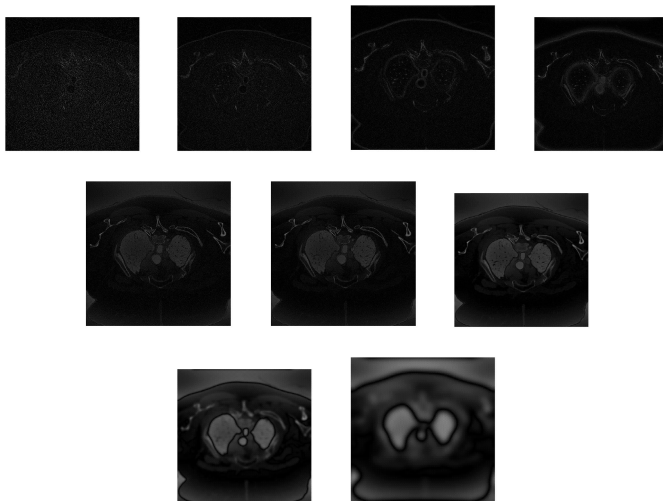


Fig. 4. Extracted regions from the HRCT image when another set of filters with different values of σ_1 and σ_2 is applied

The Results in fig 5 are similar to those in fig 4. However, the main difference is that we have used DCT instead of FFT for processing using the optical segmentation technique. It can be seen that the results follow the same characteristics as discussed earlier for FFT case. As we know that the DC component is on the top left corner followed by the low frequency component, then medium frequency and finally high frequency component in a semi ring fashion. Filter values are adjusted such that the unique semi-ring pattern is obtained for low pass, band pass and high pass filtering.

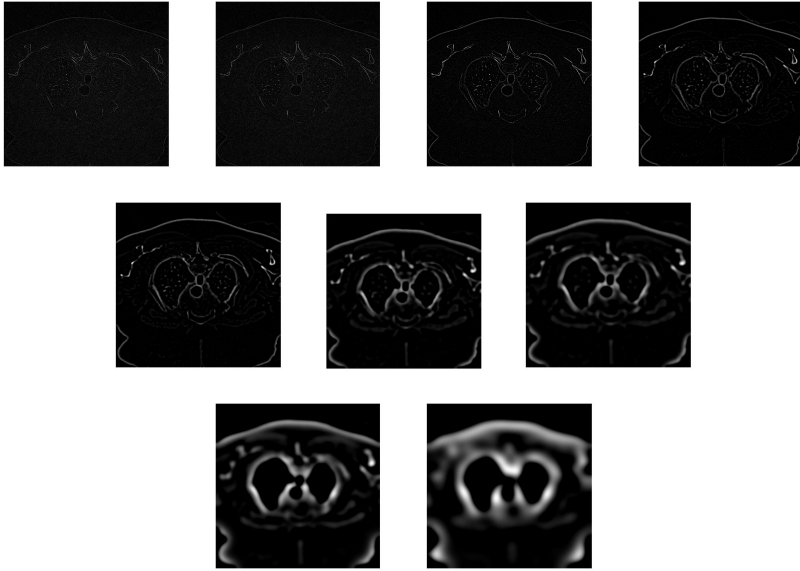


Fig. 5. Extracted regions from the HRCT image using DCT instead of FFT

5 Segmentation of Lungs and Bones

For extraction of lungs and bones, we selected to use the images processed by Optical Transfer Function (OTF) in DCT domain. After experimentation, we concluded that filter should use following parameters:

- $\sigma_1 = 0.05$ and $\sigma_2 = 1$ for lungs segmentation
- $\sigma_1 = 0.007$ and $\sigma_2 = 0.008$ for bones segmentation

Lets first consider the extraction of lungs from the image. The algorithm for extraction of lungs is as follows:

1. Process original HRCT slice using OTF in DCT domain ($\sigma_1 = 0.05$ and $\sigma_2 = 1$) and invert it (hence removing thorax). The image obtained after OTF processing is inverted so as to highlight the edges detected by OTF. The inverted image $I_v(x, y)$ is obtained from the computed image in equation (5) as:

$$I_v(x, y) = \max(i_c(x, y)) - i_c(x, y) \quad (7)$$

2. Convert the image to a binary image (I_B) using thresholding. (Since the edges are now inverted so they will have the maximum value in the inverted image. Hence, threshold value of 0.99 is used for converting to the binary image.)

$$I_B = \begin{cases} 0, & I_v(x, y) < 0.99 \\ 1, & I_v(x, y) \geq 0.99 \end{cases} \quad (8)$$

3. Label the regions as separate independent objects and remove small objects from the image based on the area of each object. (We used 8-connected neighborhood method to determine the area of each object/region and after experimentation we found that objects/ regions with 8000 or less pixels are a good estimate for removing small objects/regions in the image.) Let R_L be region with label L then:

$$R_L = \begin{cases} 0, & \text{Area}(R_L) \leq 8000 \text{ pixels} \\ 1, & \text{Area}(R_L) > 8000 \text{ pixels} \end{cases}, L = 1, 2, 3, 4, \dots \quad (9)$$

4. Remove the regions that connect to any of the boundary in the image. (We utilize the fact that the lungs are in the center of the image and they are not connected with any of the boundary of the complete image slice.)

$$R_L = \begin{cases} 0, & \text{boundary}(R_L) = B \text{ pixels} \\ 1, & \text{boundary}(R_L) \neq B \text{ pixels} \end{cases}, L = 1, 2, 3, 4, \dots \quad (10)$$

$\text{boundary}(R_L)$ = Any boundary pixel of the region L

B = Any boundary pixel of the image slice

5. Fill the holes within the remaining regions (the holes are filled with the boundary pixel value of the region.)
6. Extract the lungs based on the area calculated in step 2. (We exploit the fact that we have already removed small objects and objects connected with the boundary. Therefore, the regions with maximum area are now the lung regions.)

$$R_{Lung} = \begin{cases} \max(\text{area}(R_L)), & Lung = 1 \\ (\max - 1)(\text{area}(R_L)), & Lung = 2 \end{cases} \quad (11)$$

Where $Lung = 1, 2; L = 1, 2, 3, 4, \dots$

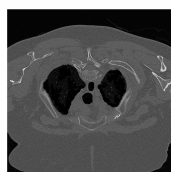
After extraction of the lungs, we extract bones from the HRCT image slice using the following algorithm:

1. Create a binary mask from the result in step 6 of the above algorithm

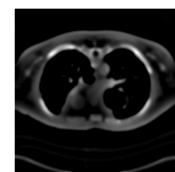
$$\text{mask} = \begin{cases} 1, & \text{mask} = R_{Lung} \\ 0, & \text{otherwise} \end{cases} \quad (12)$$

Where $Lung = 1, 2$

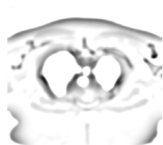
2. Mask original HRCT image slice. (Result is intensity image with no information in lungs region.)
3. Process the resultant image slice using OTF in DCT domain ($\sigma_1 = 0.007$ and $\sigma_2 = 0.008$)



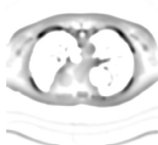
(a)



(b)



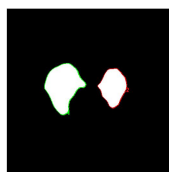
(c)



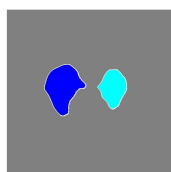
(d)



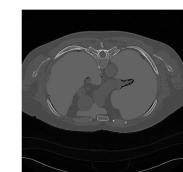
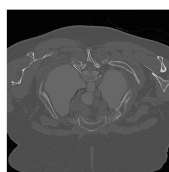
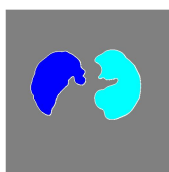
(e)



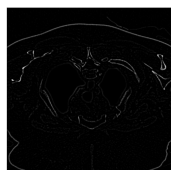
(f)



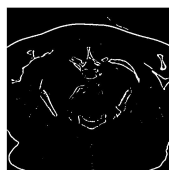
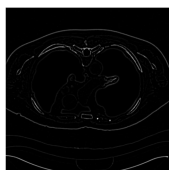
(g)



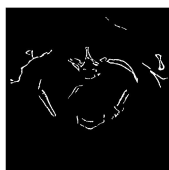
(h)



(i)



(j)



(k)

Fig. 6. Step by step application of the algorithms for extraction of lungs and bones, (a) HRCT slices of 2 different patients, (b) Process using OTF ($\delta_1=0.05$ & $\delta_2=1$), (c) Inverting the OTF processed image, (d) Convert to binary using thresholding, (e) Removing small objects based on area, (f) Removing boundary connect regions, (g) Extracting lung based on area, (h) Masking the original image slice, (i) Process using OTF ($\delta_1 = 0.007$ and $\delta_2 = 0.008$), (j) Thresholding, (k) Extracting bones by removing small objects & those connected with boundary.

4. Steps 2 to 6 of above mentioned algorithm are repeated and the bones are extracted based on the area. (Difference is that regions of 20 or less pixels are removed in step 3 because now objects of very small size are present in the image.)

Fig 6 shows all the steps of above mentioned algorithms for two HRCT image slices of different patients. Left hand column shows a normal HRCT image and right hand column shows a honeycombed HRCT image. Fig 6(g) shows the segmented lungs and fig 6(k) shows the segmented bones. During all the simulations, we have taken into account the optimum window size calculations [14].

6 Conclusions

The objective of this paper was to show the multiscale segmentation of HRCT images and we demonstrated this fact by using a filter whose parameters can be adjusted to make it a low pass or band pass or high pass filter. As we know that all the edge information and sharpness of the objects is in the band pass or high frequency region. So we did the processing in the frequency domain. We applied this filter using both the Fast Fourier transform and Discrete Cosine Transform. The results are almost similar for both the transforms. We successfully demonstrated the segmentation of tissues, lungs, bones and anatomical structures within the lung using the filter designed by optical transfer function at different scales or resolutions. Further we proposed an algorithm for extraction of lungs and bones from the HRCT image and we successfully demonstrated the extraction in this paper.

Acknowledgement

This work was supported by Korea Research Foundation Grant under Grant No. KRF-2004-041-D00640. Also we wish to acknowledge our collaboration with Tatjana Zremic, University of New South Wales.

References

1. Nishimura K., Izumi T., Kitaichi M., Nagai S., Itoh H., The Diagnostic Accuracy of High Resolution Computed Tomography in Diffuse Infiltrative Lung Diseases, *Chest* Vol. 104 (1993) 1149-1155.
2. Brown M.S., McNitt-Gray M.F., Mankovich N.J., Goldin J.G., Hiller J., Wilson L. S., Aberle D.R., Method for Segmentation Chest CT Image Data using an Anatomical Model: Preliminary Results, *IEEE Transactions on Medical Imaging*, Vol. 16 (1997) 828-839.

3. Brown M.S., McNitt-Gray M.F., Mankovich N.J., Goldin J.G., Hiller J., Wilson L. S., Aberle D.R., Knowledge based Segmentation of Thoracic Computed Tomography Images for Assessment of Split Lung Function, *Medical Physics* Vol. 27 (2000) 592-598.
4. Hu S., Hoffman E.A., Reinhardt J.M., Automatic Segmentation of Accurate Quantitation of Volumetric X-ray CT Images, *IEEE Transactions on Medical Imaging*, Vol. 20 (2001) 490-498.
5. Zhang D., Valentino D.J., Segmentation of Anatomical Structures in X-Ray Computed Tomography Images using Artificial Neural Networks, *Proc SPIE* Vol. 4684 (2001) 1640-1652.
6. Zheng B., Leader J. K., Maitz G. S., Chapman B. E., Fuhrman C. R., Rogers R. M., Scurba F. C., Perez A., Thompson P., Good W. F., Gur D., A Simple Method for Automated Lung Segmentation in X-Ray CT Images, *Medical Imaging*, *Proc SPIE*, Vol. 5032 (2003) 1455-1463.
7. Kuhnigk J. M., Hahn H. K., Hindennach M., Dicken V., Krass S., Peitgen H.-O., Lung Lobe Segmentation by Anatomy Guided 3D Watershed Transform, *Medical Imaging*, *Proc SPIE* Vol. 5032 (2003) 1482-1490.
8. Zhang L., Reinhardt J.M., Detection of Lung Lobar Fissures using Fuzzy Logic, *Physiology and Function from Multidimensional Images*, *Proc SPIE* Vol. 3660 (1999) 188-199.
9. Zhang L., Hoffman E.A., Reinhardt J.M., Lung Lobe Segmentation by Graph Search with 3D Shape Constraints, *Physiology and Function from Multidimensional Images*, *Proc SPIE*, Vol. 4321 (2001) 204-215.
10. Wei J., Image Segmentation based on Situational DCT Descriptors, *Pattern Recognition Letters* Vol. 23 (2002) 295-302.
11. Poon T.-C., Banerjee P. P., *Contemporary Optical Image Processing*, 1st ed., Elsevier Science Ltd. New York (2001).
12. Nayar S. K., Nakagawa Y., Shape from Focus: An Effective Approach for Rough Surfaces, *CRA90* (1990) 218-225.
13. Nayar S. K., Nakagawa Y., Shape from Focus, *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 16, No. 8 (1994) 824-831.
14. Malik A. S., Choi T.-S., Consideration of Illumination Effects and Optimization of Window Size for Accurate Calculation of Depth Map for 3D Shape Recovery, *Pattern Recognition* Vol. 40/1, 154-170.

An Evaluation of Image Compression Algorithms for Colour Retinal Images

Gerald Schaefer¹ and Roman Starosolski²

¹ School of Engineering and Applied Science, Aston University, U.K.

² Institute of Computer Science, Silesian University of Technology, Poland

Abstract. Diabetic retinopathy is the leading cause of blindness in the adult population. Mass-screening efforts, during which high resolution images of the retina are captured, are therefore underway in order to detect the disease in its early stages. In this paper we evaluate the compression performance of several lossless image compression algorithms that could be employed in a retina Picture Archiving and Communications System to lessen the demand on computing resources. The algorithms we analyse are TIFF PackBits, Lossless JPEG, JPEG-LS, and JPEG2000 all of which are incorporated in the current DICOM standard together with the non-standard CALIC algorithm for benchmark comparison. Compression performance is evaluated in terms of compression ratio, compression speed, and decompression speed. Based on a large dataset of more than 800 colour retinal images, divided into groups according to retinal region (nasal, posterior, and temporal) and image size, JPEG-LS is found to be the most suitable compression algorithm, offering good compression ratios combined with high compression and decompression speed. Compression ratios can be further improved through the application of a reversible colour space transformation prior to compression as a second set of experiments show.

Keywords: retinopathy, retinal images, lossless image compression, colour space transform, DICOM, PACS.

1 Introduction

Diabetic retinopathy is the leading cause of blindness in the adult population. In order to effectively identify patients suffering from the disease, mass-screening efforts are underway during which digital images of the retina are captured and then assessed by an ophthalmologist. In order to identify features such as exudates and microaneurysms, which are typically very small in extent, retinal images are captured at high resolutions. This in turn means large file sizes and, considering the archival of typically thousands of records, a high demand on computational resources, in particular storage space as well as bandwidth when used in a Picture Archiving and Communications System (PACS). Image compression therefore seems a necessary step. Image compression algorithms can be divided into two groups: lossy techniques where some of the visually less important image data is discarded in order to improve compression ratios, and lossless

methods which allow the restoration of the original data. As the features that indicate retinopathy are very small in size and following legislation in several countries, only lossless compression seems suitable for retinal images.

In this paper we build upon our earlier work [1] and aim to identify a suitable compression algorithm for colour retinal images. Such an algorithm, in order to prove useful in a real-life PACS, should not only reduce the file size of the images significantly but also has to be fast enough, both for compression and decompression. Furthermore, it should be covered by international standards such as ISO standards and, in particular for medical imaging, the Digital Imaging and COmmunication in Medicine (DICOM) standard [2,3]. For our study we therefore selected those compression algorithms that are supported in DICOM, namely TIFF PackBits [4], Lossless JPEG [5], JPEG-LS [6], and JPEG2000 [7]. For comparison we also included CALIC [8] which is often employed for benchmarking compression algorithms. All algorithms were evaluated in terms of compression ratio which describes the reduction of file size, and speed. For speed, we consider both the time it takes to encode an image (compression speed) and to decode (decompression speed) as both are relevant within a PACS.

Experiments were performed on a large dataset of more than 800 colour retinal images, which were also divided into subgroups according to retinal region (nasal, posterior, and temporal) and images size. Overall, JPEG-LS was found to be the best performing algorithm as it provides good compression ratios coupled with high speed. Little variation was found between the individual image categories. In a second set of experiments we tested whether the application of a reversible colour space transformation prior to compression can improve the compression performance. Based on the same image set, an improvement of about 5% in terms of compression ratio was found to be achievable.

The rest of the paper is organised as follows: Section 2 introduces the compression algorithms that were evaluated while Section 3 gives details of the dataset of images we used. The experiments together with the resulting evaluation of compression performance is provided in Section 4. Section 5 explains the application of colour space transformations and gives experimental results for this method. Section 6 concludes the paper.

2 Image Compression Algorithms

In this section we give a brief overview of the lossless compression algorithms that we have evaluated. For further details on the algorithms we refer the reader to the original references.

- TIFF PackBits - a simple RLE compression algorithm [4]. The Tag Image File Format (TIFF) standard specifies this simple runlength (RLE) coding technique; we used the LibTIFF implementation by Leffler (version 3.6.1, <http://www.remotesensing.org/libtiff/>).
- Lossless JPEG - former JPEG (Joint Photographic Experts Group) committee standard for lossless image compression [5]. The standard describes

predictive image compression algorithm with Huffman or arithmetic entropy coder. We used the Cornell University implementation (version 1.0, <ftp://ftp.cs.cornell.edu/pub/multimed/ljpg.tar.Z>) which applies Huffman coding. The results are reported for the predictor function SV7 which resulted in the best average compression ratio for the dataset.

- JPEG-LS - standard of the JPEG committee for lossless and near-lossless compression of still images [6]. The standard describes low-complexity predictive image compression algorithm with entropy coding using modified Golomb-Rice family. The algorithm is based on the LOCO-I algorithm [9]. We used the University of British Columbia implementation (version 2.2, ftp://ftp.netbsd.org/pub/NetBSD/packages/distfiles/jpeg_ls_v2.2.tar.gz).
- JPEG2000 - a more recent JPEG committee standard describing an algorithm based on wavelet transform image decomposition and arithmetic coding [7]. Apart from lossy and lossless compressing and decompressing of whole images it delivers many interesting features (progressive transmission, region of interest coding, etc.) [10]. We used the JasPer implementation by Adams (version 1.700.0, <http://www.ece.uvic.ca/~mdadams/jasper/>).
- CALIC (Context-based Adaptive Lossless Image Compression) - a relatively complex predictive image compression algorithm using arithmetic entropy coding, which because of its usually good compression ratios is commonly used as a reference for other image compression algorithms [8,11]. We used the implementation by Yang. In contrast to the other algorithms CALIC is designed for grayscale images only. We therefore apply CALIC to each of the individual channels separately.

Lossless JPEG, JPEG-LS, and JPEG2000 are covered by international ISO standards whereas TIFF represents an industry standard. All algorithms except CALIC are incorporated into the medical imaging DICOM [2,3] standard.

3 Retinal Image Dataset

A large set of over 800 colour retinal images captured at various ophthalmology centres was used in our experiments. The set is large both with regards to the number of images as well as with regards to the actual sizes of individual images. All images were initially obtained in uncompressed 24-bit RGB format. Images contain between 1.4 and 3.5 millions pixels and hence require, between 4 and 10 MB of storage space.

In order to verify whether there are certain image classes that are especially susceptible to compression (or especially hard to compress) we divided the dataset into categories according to the following criteria:

1. Retinal region: the whole set is divided into three groups: *nasal*, *posterior*, and *temporal*. Evaluating the compression performance for individual subgroups will highlight whether any algorithms work especially well on any of these categories. An example of the three images for the same patient is given in Figure 1

2. Image size: here the images fall mainly into two categories: those images with about 1.4 million pixels (*small*) and those with about 3.4 million pixels (*large*). Compression ratios and compression speed are sometimes dependent on the image size, evaluating the individual size categories will hence confirm whether any such dependency exists for retinal images.



Fig. 1. *Nasal* (l), *posterior* (m), and *temporal* (r) retinal image of a patient

The two criteria are independent of each other (i.e. there exists both *small* and *large* images of all three regions in the dataset) and are hence tested separately. Details on how the images are divided into the individual categories are given in Table 1.

Table 1. Test dataset of retinal images

Category	Number of images	Average size [pixels]
nasal	252	2214864
posterior	301	2203523
temporal	250	2213720
small	468	1383184
large	335	3365689
all	803	2210257

4 Compression Performance Evaluation

4.1 Experimental Procedure

Experimental results were obtained on a HP Proliant ML350G3 computer equipped with two Intel Xeon 3.06 GHz (512 KB cache memory) processors and Windows 2003 operating system. Single-threaded applications of algorithms used for comparisons were compiled using Intel C++ 8.1 compiler. To minimise effects of the system load and the input-output subsystem performance, the algorithms were run several times; the time of the first run was ignored while the collective time of other runs (executed for at least one second, and at least 3 times) was measured and then averaged. The time measured is the sum of time spent by the processor in application code and in kernel functions called by the application, as reported by the operating system after

application execution. The speed of implementations is reported in megabytes (uncompressed) per second [MB/s], where $1 \text{ MB} = 2^{20}$ bytes, both for compression and decompression speeds. We note that we actually measure the speed of the specific implementation of the given algorithm on the particular computer system, not the absolute speed of the algorithm itself. The computer system we used, in terms of the processor speed, amount of the installed memory etc. is similar to machines currently employed in PACSs. The compression ratios are reported as bitrates, expressed in bits per pixel [bpp] $8e/n$, where e is the size in bytes of the compressed image including the header and n is the number of pixels in the image. We note that smaller bitrates mean better compression and that uncompressed images are stored using 24 bpp.

4.2 Experimental Results

As mentioned above all algorithms were evaluated in terms of compression ratio, compression speed and decompression speed. Results were obtained both for the full dataset and for the individual categories outlined in Section 3 and are given in Tables 2 to 4. The numbers are calculated as the averages for all images contained in a category; since not all groups contain the same number of images the average results for all images are slightly different from the average over all groups.

Table 2. Compression ratio results [bpp]

Category	PackBits	L-JPEG	JPEG-LS	JPEG2000	CALIC
nasal	19.38	8.12	6.83	7.14	6.61
posterior	19.25	8.08	6.80	7.11	6.56
temporal	19.48	8.10	6.84	7.16	6.63
small	18.32	7.92	6.49	6.84	6.25
large	20.78	8.34	7.26	7.53	7.07
all	19.36	8.10	6.82	7.14	6.60

Table 2 lists the compression ratio results for all categories. From there we can see immediately that the compression performance of the PackBits methods is very different from those of the other algorithms. Considering that uncompressed images require 24 bits per pixel, the 19.36 bpp achieved by PackBits is fairly feeble yet not surprisingly so as runlength coding is only suitable for images with large uniform patches to produce reasonably compressible runlengths of pixels. As this is not the case for retinal images (apart from the fairly uniform black background) the achieved compression ratios alone disqualify PackBits as a suitable method for compressing retinal images. Among the remaining algorithms, Lossless JPEG is consistently the worst performing algorithm which again is expected as it is based on relatively simple predictive coding. The best performing algorithm is CALIC with an average bitrate of 6.60 bpp. That CALIC provides the best compression ratio was to be expected and was indeed the reason for

including the algorithm in the evaluation in the first place. However, JPEG-LS performs only slightly worse than CALIC with an average bpp value of 6.82. Somewhat worse than JPEG-LS is the performance of JPEG2000 with a compression ratio of 7.14 bits per pixel.

Looking at the variation of results between the image categories we see that there is very little difference between the compression ratios for the *nasal*, *posterior*, and *temporal* image groups. This suggests that these groups share similar image characteristics as can indeed be seen by looking Figure 1. Inspecting the *small* and *large* image categories we can see that all algorithms are more effective for smaller images. The larger images of about 3.4 million pixels are compressed with an on average about 10% higher bitrate compared to the smaller 1.4 megapixel images; the highest differences are observed for PackBits while Lossless JPEG is the algorithm that performs most constant across different image sizes.

Table 3. Compression speed results [MB/s]

Category	PackBits	L-JPEG	JPEG-LS	JPEG2000	CALIC
nasal	51.2	16.4	14.7	3.3	2.7
posterior	51.4	16.5	14.7	3.2	2.7
temporal	51.0	16.4	14.6	3.3	2.7
small	50.2	16.5	15.3	3.4	2.8
large	52.6	16.4	13.8	3.0	2.6
all	51.2	16.5	14.7	3.3	2.7

We now turn our attention to Table 3 which lists the compression speeds expressed in terms of MB/s of all algorithms. Here, PackBits is clearly the best performing technique affording compression speeds more than three times higher the next best competitor. However, as has been pointed out above, this high speed comes at the expense of an unsatisfyingly high bitrate which rules out the algorithm as being useful in practise. Lossless JPEG is the next fastest algorithm with an average compression speed of 16.5 MB/s which is again due to the relatively simple compression technique. Only slightly slower is JPEG-LS with 14.7 MB/s which hence provides fast compression coupled with good compression ratios. JPEG2000 and CALIC are much slower than the other algorithms with compressions speeds of only about 3 MB/s which is about 5 times slower than Lossless JPEG and JPEG-LS. This low speed results from the relatively complex compression paradigms involved which are based on wavelets and complex predictive and arithmetic coding respectively. Looking at the individual image groups, the results across the different region category images are again very uniform while larger images allow higher compression speeds with PackBits but lower speeds with JPEG-LS, JPEG2000, and CALIC.

While the compression speed is obviously of importance at the time of image capture, in a PACS the decompression time might be of higher interest as medical images are typically stored (and hence compressed) only once whereas

Table 4. Decompression speed results [MB/s]

Category	PackBits	L-JPEG	JPEG-LS	JPEG2000	CALIC
nasal	68.7	24.3	14.4	3.1	2.4
posterior	69.1	24.5	14.4	3.0	2.4
temporal	68.4	24.3	14.3	3.1	2.3
small	64.8	24.3	15.0	3.2	2.4
large	74.2	24.5	13.5	2.7	2.2
all	68.7	24.4	14.4	3.1	2.4

they are read (and hence decompressed) several times, in particular in the context of medical image retrieval applications [12]. In Table 4 we therefore list decompression speeds, again expressed in terms of MB/s, for the algorithms on all image categories. Since there is always a certain symmetry in the operations involved in encoding and decoding images it is not surprising to see that overall the decompression speed results are not very different from the compression speed ones. Lossless JPEG represents a notable exception as for this algorithm the decompression speed is much higher (by about 2/3) than the compression speed. PackBits also decompresses faster than it encodes whereas for the rest of the algorithms encoding and decoding speeds are fairly close to each other. Hence, PackBits is by far the fastest algorithm in terms of decompression speed, followed by Lossless JPEG, JPEG-LS, JPEG2000, and CALIC. Again, there is no noticeable difference between the different region categories while PackBits decompresses larger images faster still, and JPEG-LS, JPEG2000, and CALIC have slightly higher decompression speeds for smaller images.

Integrating all experimental results, i.e. compression ratios, compression and decompression speeds, should enable us to judge each algorithm's suitability for the task of being employed as a compression algorithm for colour retinal images in a medical Picture Archiving and Communications System. While TIFF PackBits provides by far the highest encoding and decoding speeds its performance in terms of compression ratios is also significantly worse than those of all others. The fact, that it reduces file sizes only by slightly more than 20% rules it out as an algorithm to be used in practise. CALIC on the other hand provides the best compression rates but is at the same time the slowest algorithm. In fact, both CALIC and JPEG2000 are significantly slower than the rest of the algorithms, both for compression and decompression. This slow decompression speed makes them unsuitable to be employed in PACSs; in addition CALIC is not covered by any standard, in particular not by the DICOM standard. The most suitable algorithm for lossless compression of colour retinal images hence seems to be the JPEG-LS standard. It provides good compression ratios (only slightly worse than CALIC) combined with fast compression and decompression performance. Furthermore, JPEG-LS represents an ISO standard and is included in the DICOM standard for medical imaging. In systems where decompression speed is significantly more important than compression ratio, Lossless JPEG might be an alternative as it decodes faster, yet at the expense of higher bitrates.

5 Colour Space Transformation

In a second set of experiments we set out to investigate whether the application of a reversible colour space conversion can improve the compression ratios achieved by the algorithms. It is well known that the Red, Green and Blue channels of the RGB colour space are correlated. To remove this inter-channel correlation one could use the optimal PCA (principal component transform), but as the PCA coefficients are calculated for each image individually based on the analysis of the whole image the use of PCA would slow down the compression process noticeably due to the high computational complexity involved. Therefore, in lossy compression algorithms like JPEG or JPEG2000 (lossy mode), a standard transform to a different color space such as YC_bC_r is employed. In the case of lossless coding, clearly a reversible transform must be employed to retain all image information. In this paper we report results on applying such a transform during the compression of colour retinal images.

The Reversible Color Transform (RCT) of the JPEG-LS standard [13] is a modulo-arithmetic integer approximation of the YC_bC_r transform defined as

$$\begin{aligned} C_1 &= (R - G + 2^{N-1}) \bmod 2^N \\ C_2 &= (B - G + 2^{N-1}) \bmod 2^N \end{aligned} \quad (1)$$

and

$$Y = \left(G + \left\lfloor \frac{C_1 + C_2}{4} \right\rfloor - 2^{N-2} \right) \bmod 2^N \quad (2)$$

where R , G , and B are the pixel values of the red, green and blue channels respectively, and Y , C_1 , and C_2 describe the transformed components. N is the colour channel bit depth. The required precision for the transformed components is equal to that of the RGB representation, i.e. there is no expansion of the dynamic range. Due to modulo clipping the transform may introduce rapid changes of intensities in the transformed channels even if the intensities in RGB were changing gradually, as is demonstrated in Figure 2.

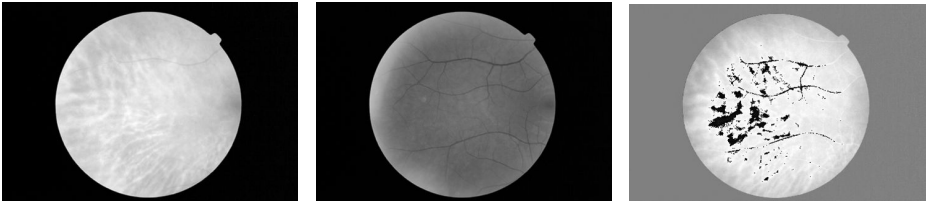


Fig. 2. Red (l), green (m), and transformed C_1 (r) channel of a retinal image

The colour space transform is applied prior to compression while the inverse transform is applied after the decompression stage. In Table 5 we list the achieved compression ratios when applying the RCT transform. Again, results are provided for the complete dataset as well as for all categories.

Table 5. Compression ratio results after colour space conversion [bpp]

Category	PackBits	L-JPEG	JPEG-LS	JPEG2000	CALIC
nasal	18.83	7.94	6.54	6.91	6.33
posterior	18.54	7.86	6.45	6.86	6.22
temporal	18.66	7.82	6.46	6.78	6.21
small	18.17	8.12	6.68	6.94	6.45
large	19.33	7.51	6.20	6.72	5.97
all	18.67	7.87	6.48	6.84	6.25

Applying a colour space conversion does not change the relative performance among the tested algorithms. On average an improvement of about 5% in terms of compression rate is observed for the whole dataset. However, while there is little variation among the region image categories, the image size seems to play a more important role here. While for the *large* images a significant improvement can be observed the opposite is true for the *small* image category, here the application of a colour space transform actually worsens the compression ratio for all algorithms except PackBits.

6 Conclusions

We have analysed the performance of several standard lossless image compression algorithms, namely TIFF PackBits, Lossless JPEG, JPEG-LS, and JPEG2000 as well as the non-standard “benchmark” CALIC algorithm for a large set of colour retinal images. The compression performance was measured in terms of compression ratio, compression speed and decompression speed. JPEG-LS was found to be the best performing algorithm offering good compression ratios combined with high compression and decompression speed. In addition, apart from being a standard itself, JPEG-LS is also incorporated in the DICOM standard and is hence readily available to be employed in Picture Archiving and Communications Systems. We have also analysed the application of a reversible color space transformation as part of the compression process and have found that while on average improved compression ratios are achieved this is not consistently so as for all image classes.

References

1. Starosolski, R., Schaefer, G.: Lossless compression of color medical retinal images. In: 20th European Multiconference. (2006) 79–84
2. National Electrical Manufacturers Association: Digital Imaging and Communications in Medicine (DICOM). Standards Publication PS 3.1-2004 (2004)
3. Mildenerger, P., Eichelberg, M., Martin, E.: Introduction to the DICOM standard. *European Radiology* **12** (2002) 920–927
4. Adobe Systems Inc.: TIFF 6.0.1 specification (1995)

5. Langdon, G., Gulati, A., Seiler, E.: On the JPEG model for lossless image compression. In: 2nd Data Compression Conference. (1992) 172–180
6. ISO/IEC: Lossless and near-lossless compression of continuous-tone images - baseline. ISO/IEC International Standard 14495-1 (1999)
7. ISO/IEC: JPEG2000 image coding system: Core coding system. ISO/IEC International Standard 15444-1 (2002)
8. Wu, X., Memon, N.: Context-based adaptive lossless image codec. *IEEE Trans. Communications* **45**(4) (1997) 437–444
9. Weinberger, M., Seroussi, G., Sapiro, G.: The LOCO-I lossless image compression algorithm: Principles and standardization into JPEG-LS. *IEEE Trans. Image Processing* **9**(8) (1996) 1309–1324
10. Christopoulos, C., Skodras, A., Ebrahimi, T.: The JPEG2000 still image coding system: An overview. *IEEE Trans. Consumer Electronics* **46**(4) (2000) 1103–1127
11. Wu, X.: Lossless compression of continuous-tone images via context selection, quantization, and modeling. *IEEE Trans. Image Processing* **6**(5) (1997) 656–664
12. Mueller, H., Michoux, N., Bandon, D., Geissbuhler, A.: A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *Int. Journal of Medical Informatics* **73**(1) (2004) 1–23
13. ISO/IEC: Lossless and near-lossless compression of continuous-tone images - extensions. ISO/IEC International Standard 14495-2 (2002)

An Automated Model for Rapid and Reliable Segmentation of Intravascular Ultrasound Images

Eirini Parissi¹, Yiannis Kompatsiaris¹, Yiannis S. Chatzizisis², Vassilis Koutkias³,
Nicos Maglaveras³, M.G. Strintzis¹, and George D. Giannoglou²

¹ Informatics and Telematics Institute, Centre for Research and Technology-Hellas, 57001,
Thessaloniki, Greece

{parissi,ikom, strintzi}@iti.gr

² Cardiovascular Engineering and Atherosclerosis Laboratory, 1st Cardiology Department,
AHEPA University Hospital, Faculty of Medicine, Aristotle University of Thessaloniki, 54636,
Thessaloniki, Greece

{joc, yan}@med.auth.gr

³ Laboratory of Medical Informatics, Faculty of Medicine, Aristotle University of
Thessaloniki, 54124, P.O.Box 323, Thessaloniki, Greece

{bikout, nicmag}@med.auth.gr

Abstract. The detection of lumen and media-adventitia borders in intravascular ultrasound (IVUS) images constitutes a necessary step for accurate morphometric analyses of coronary plaques and accordingly assessment of the atherosclerotic lesion length. Aiming to tackle this issue, an automated model for lumen and media-adventitia border detection is presented, which is based on active contour models. The proposed approach enables extraction of the corresponding boundaries in sequential IVUS frames by applying an iterative procedure, in which initialization of the two contours in each frame is performed automatically, based on the segmentation of its previous frame. The above procedure is implemented through a user-friendly interface, permitting the interaction of the user when needed. The in vivo application and evaluation of our model in sequential IVUS images indicated that the proposed approach is capable of accurately and rapidly segmenting hundreds of IVUS images.

Keywords: image segmentation, intravascular ultrasound (IVUS), snakes, active contours.

1 Introduction

Imaging modalities constitute a primary data source in medicine. Medical images originate from diagnostic technologies (e.g., X-ray, ultrasound, computed tomography, magnetic resonance, nuclear imaging), surgical procedures, pathology (e.g., light and electron microscopy) and the research laboratory. Advances in computational technology are making complex image processing techniques widely available for medical use, targeting not only to improve the existing diagnostic systems for clinicians, but also to facilitate medical research.

Intravascular Ultrasound (IVUS) constitutes a powerful imaging modality for the analysis and diagnosis of coronary artery disease [1]. It comprises of a catheter-based

technique, which provides high-resolution tomographic images of both the arterial lumen and wall (Fig. 1), in contrast to X-ray angiography, which provides a shadow image of the lumen only (luminogram). IVUS cross-sectional images are generated by detecting the scattered waves of the ultrasound signal transmitted by the probe of the IVUS catheter as it is pulled back along the vessel. The introduction of IVUS in the routine clinical practice the last decade changed our understanding about the progression of atherosclerosis and the concomitant vascular remodeling [2].

In order to evaluate the plaque or lumen area, IVUS images should be segmented; two borders should be identified, i.e., the interface between the lumen and the wall and the leading edge between the media and adventitia (Fig. 1) [1]. In the current clinical use of IVUS, segmentation is manually performed, and sometimes even visually. Taking into account that the acquired images of a typical IVUS pullback may be a few hundreds, it is obvious that manual segmentation is quite laborious and time-consuming, subjected at the same time to high inter- and intra-observer variability [3].

The development of a sophisticated computer model that could ideally support an automated segmentation of IVUS images in a reliable and quick manner, allowing user interaction would be of great importance. Such a system is presented in this work characterized by rapid and efficient analysis in a user-friendly and easily applicable computational environment. The development of this model has been based on the theory of active contours, also known as snakes. Comparing to other snake – based medical image segmentation systems and algorithms [4], [5], our approach has the advantage that it is fully integrated with a user – friendly graphical interface. Through this interface, the user has straightforward interaction with the system and as a result total control at each step of the procedure is achieved. This functionality leads to higher accuracy in the detection results.

2 Methods

2.1 Theoretical Background

Snakes were originally presented as a regularization step in edge detection algorithms [6]. In particular, they are flexible models based on object detection techniques by the use of parametric curves, which can deform their shape under the action of restrictive internal and external forces, which eventually guide the snake to significant features of the image, such as the lumen and media-adventitia borders [7]. To date, several active contour models have been described, originated from the initial proposal, some of them targeting medical imaging related problems [8], [4].

A snake is an ordered set of points, also called snaxels [6], constituting an energy-minimizing parametric closed curve guided by external forces, which has to be initially defined on the image plane. Specifically, the aim is to minimize an energy function defined as:

$$E_{\text{snake}} = E_{\text{int}} + E_{\text{ext}}, \quad (1)$$

where E_{int} and E_{ext} are the internal energy formed by the snake configuration and the external energy formed by external forces affecting the snake, respectively. In other

words, the aim of the initially defined contour is to find a location that minimizes equation (1). In our approach, this energy functional is defined as:

$$E_{\text{snake}} = E_{\text{cont}} + E_{\text{curv}} + E_{\text{image}}, \quad (2)$$

where the first two terms correspond to the internal energy, while the third term corresponds to the external one. Specifically, E_{cont} is the contour continuity energy (causes the snake points to become more equidistant), E_{curv} is the contour curvature energy (the smoother the contour is, the less is the curvature energy), while E_{image} is the image energy (forces the snake to be attracted by image features). More specifically, considering a set of snaxels $S = \{s_1, s_2, \dots, s_N\}$ the partial energy terms are defined as:

$$E_{\text{cont}} = \left| \bar{\delta} - |s_i - s_{i-1}| \right| \quad (3)$$

where $\bar{\delta}$ is the average distance between all pairs $|s_i - s_{i-1}|$,

$$E_{\text{curv}} = \left| s_{i-1} - 2s_i + s_{i+1} \right|^2, \quad (4)$$

$$E_{\text{image}} = - \left| \nabla I \right|^2, \quad (5)$$

where I is the image intensity.

In more detail, summary energy at every point is calculated as a linear combination of the aforementioned terms, i.e.:

$$E_{\text{snake}, i} = \alpha E_{\text{cont}, i} + \beta E_{\text{curv}, i} + \gamma E_{\text{image}, i}, \quad (6)$$

where α , β , and γ are appropriate weighting factors, which control the relative influence between the terms. In particular, α is responsible for contour continuity,

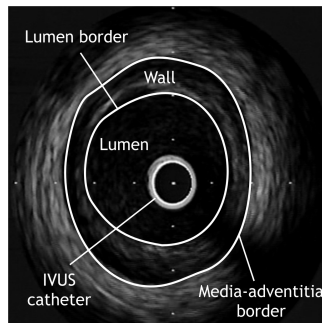


Fig. 1. An IVUS image depicting the lumen and media-adventitia borders

i.e., a high value makes snaxels more evenly spaced, β is responsible for snake corners, i.e., a high value for a specific snaxel makes the angle between snake edges

more obtuse, while γ is responsible for making snaxels more sensitive to the image energy, rather than to continuity or curvature.

2.2 General Structure of the Segmentation Model

Fig. 2 schematically illustrates the structure of the proposed model. Our approach enables extraction of the lumen and media-adventitia boundaries in sequential IVUS frames by applying an iterative procedure, in which initialization of the two contours in each frame is automatically performed based on the segmentation of the previous image. The whole procedure evolves automatically, permitting the user to interact when needed.

More specifically, the user initializes the snake in the first frame, by providing an initial closed contour near the media-adventitia border, and launches the detection mechanism. The snake deforms by minimizing its energy function in order to capture the corresponding boundary. The user can further deform the snake curve, by adjusting α , β and γ parameters appropriately (Eq. (6)). If the result is not satisfactory, the user may re-initialize the snake contour and launch the detection mechanism again. The same procedure is iterated for the detection of the lumen wall. The initialization curves are then applied to the whole sequence under user supervision.

2.3 Implementation of the Segmentation Model

The segmentation model is integrated through a friendly graphical user interface (GUI) (Fig. 3), implemented using the Borland C++Builder®, while the contour detection algorithms are based on the Intel-OpenCV libraries [9]. The GUI features a modular, platform-independent design and includes a full set of components presented below:

- Multiple image analysis
- Automated lumen and media-adventitia contour detection
- Total control of the snake's parameters
- Spontaneous user interaction
- Contour improvement or redraw, wherever necessary
- Automated storage of the segmented images
- Re-examination of the already detected images

In the left panel a portion of the segmented frames is presented, while in the right panel the user may adjust the segmentation parameters, as well as the w parameter. The w parameter specifies and controls the neighborhood of pixels in which the snake contour deforms in order to minimize its energy function. Finally, the user may browse the segmented frame sequence using the arrows and even achieve more accurate detection results by re-initializing the inner- and outer-curves on several frames, where the results were not satisfactory enough.

The whole application can be installed and executed on any PC with simple system requirements.

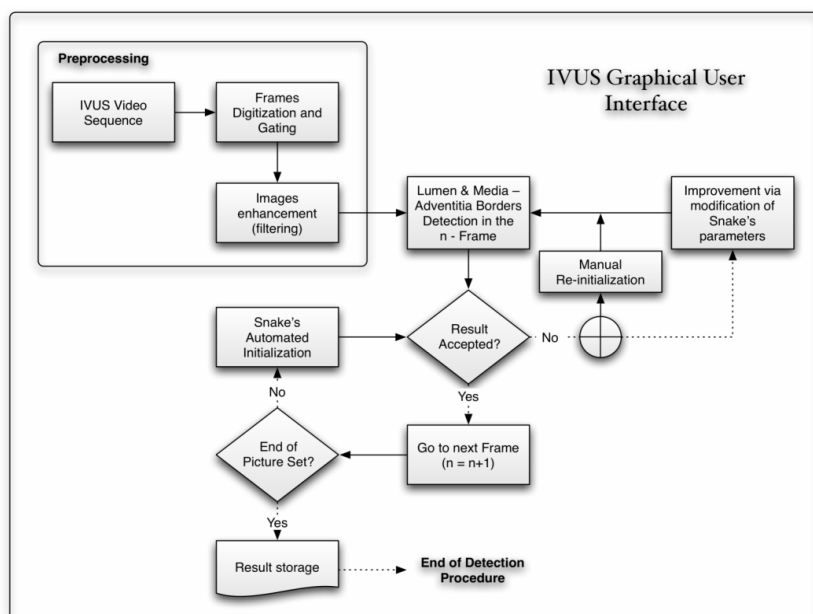


Fig. 2. Overview of the structure of the segmentation model

Pre-processing Step. The first step of the segmentation procedure includes image pre-processing, in order to prepare appropriate versions of images, increasing the detection efficiency. Initially, the IVUS S-VHS video sequence is digitalized by an integrated to the IVUS console frame-grabber at 512x512 pixels, with 8-bit grey scale, and the end-diastolic images are selected (peak of R-wave on ECG) [3]. From each ECG-gated IVUS image, a 340x340 pixels sub-image is extracted, including the region of interest and the transducer of the catheter in the centre of the sub-image. The resulting image sequence undergoes noise reduction by the application of a custom-developed filter [10].

Detection Procedure. As it can be seen in Fig. 2, the proposed model is divided in two major steps, namely, the manual determination of the initial contours and the application of the snake's algorithm to achieve the segmentation. Specifically, when the user starts the application, the snakes in the first image are initialized by providing two initial closed contours near the lumen and media-adventitia borders respectively and launching the detection algorithm. Then the snakes deform themselves by minimizing their energy function to capture the corresponding boundaries. Upon satisfactory results in the first IVUS image, the detection mechanism is applied in the next images of the sequence, under user supervision. The detected boundaries of the previous image are utilized as initialization scheme for the next ones.

Parameters Control and Re-examination. There are two ways, in which the user can modify and improve the snake algorithm's detection result, in order to achieve more accurate output:

- *Snake's parameters*: The user can change the values of the snake's parameters. Snakes can be seen as energy minimizing curves. In brief, the parameters that define the snake's behavior are three, namely, α , β and γ (Eq. (6)).
- *Re-examination*: The obtained results are reviewed by the user and, in case these are not satisfactory in a part of the image sequence, the above steps can be iterated, starting from the image where the detected boundary diverges from the actual one.

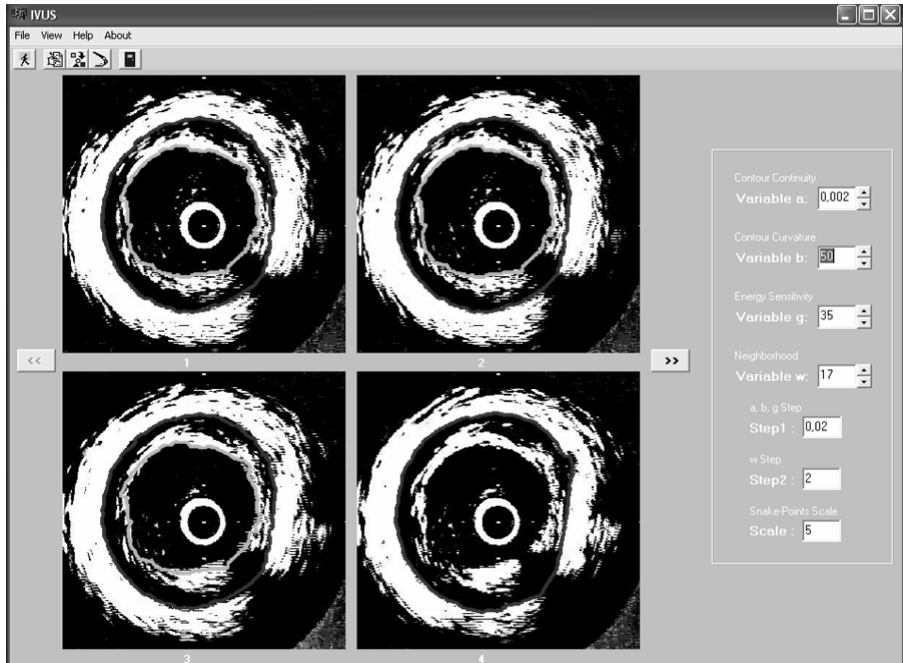


Fig. 3. The graphical user interface (GUI)

3 Clinical Evaluation of the Segmentation Model

3.1 General Performance of the Segmentation Model

In Fig. 4, three consecutive IVUS images are depicted, where the proposed border detection method was applied. The first two frames constitute cases of successful automated detection, while the third one illustrates the case where the automated detection did not result satisfactorily; hence, the user had to re-initialize the contours.

The reconstruction of a representative part of a vessel requires 200-400 gated IVUS images (Fig. 5). We tested in real medical environment the application of both manual and automated segmentation in 400 IVUS images. For the former, more than 15 hours of processing were needed, while for the latter the procedure was accomplished in less than 2 hours, without compromising accuracy.

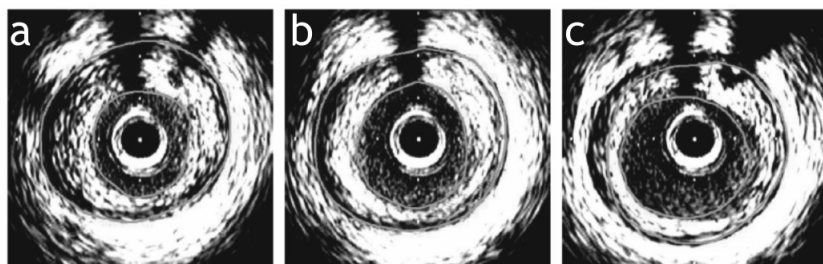


Fig. 4. Three consecutive segmented IVUS images: a, b, cases of successful border detection, c, not satisfactory border detection requiring manual re-initialization of contours

3.2 In-vivo Validation of the Segmentation Model

To evaluate the in-vivo applicability and reproducibility of the proposed segmentation model, 17 arterial segments from 9 patients, who gave written informed consent, were investigated with IVUS. The in-vivo IVUS procedures were performed with a 40 MHz, 2.5F sheath-based catheter at an automated pullback speed of 0.5 mm/sec.

The intra-observer agreement (IOA), which constitutes a measurement of reproducibility and the inter-observer agreement (INA) of manual and semi-automated segmentation were investigated in 50 randomly selected images by two independent experts. Each image was traced manually, according to the accepted international standards [1] and semi-automatically by both experts initially, and only by the first expert a month apart; thus, three manual and three semi-automated segmentations were performed in each image. The computed parameters for IOA and INA of manual and semi-automated tracing were luminal cross-sectional area (LCSA, mm^2), vessel cross-sectional area (VCSA, mm^2), maximum luminal diameter (MLD, mm) and maximum vessel diameter (MVD, mm). Additionally, the IOA and INA of semi-automated tracing were evaluated for lumen volume (LV, mm^3), vessel volume (VV, mm^3) and wall volume (WV, mm^3) in five randomly selected pullbacks.

To assess the overall effectiveness of the proposed segmentation model, the semi-automatically determined borders were compared to the reference manual tracing [1]. The compared parameters were cross-sectional areas (LCSA, VCSA; $n=50$) and maximum diameters (MLD, MVD; $n=50$). The between-experts average of the above parameters for the initial manual segmentations was considered as reference value.

All results were expressed as mean $\pm$ SD and $p<0.05$ were considered as the level of significance. IOA and INA were investigated by the calculation of Pearson's product-moment correlation coefficient and Bland-Altman analysis. The same tests were utilized for the comparison of the proposed method with the reference one. The utilized sample of 50 images was considered adequate for the reliability of the method comparison study.

Both manual and semi-automated segmentation showed significantly high IOA ($r\geq 0.98$ and $r\geq 0.97$ respectively, $p<0.01$, $n=50$) and INA ($r\geq 0.95$ and $r\geq 0.95$ respectively, $p<0.01$, $n=50$) for areas and maximum diameters, with slightly higher values of agreement for VCSA and MVD against LCSA and MLD, respectively. Likewise, the IOA and INA of our method for volumetric analyses were high ($r\geq 0.95$, $p<0.01$, $n=5$).

The Bland-Altman plots of differences between semi-automated and manual tracing against their means, revealed that the proposed method had minor differences as compared with the reference manual for all the calculated parameters. Specifically, the $Md \pm 2SD$ for LCSA, VCSA, MLD and MVD were $0.23 \pm 1.24 \text{ mm}^2$, $0.00 \pm 0.89 \text{ mm}^2$, $-0.01 \pm 0.39 \text{ mm}$ and $-0.07 \pm 0.37 \text{ mm}$ respectively. Furthermore, all differences were concentrated within the limits of agreement in the corresponding plots of differences against means.

4 Discussion

In this work we described a model for automated segmentation of sequential IVUS images. In Table 1 the advantages and the limitations of the proposed approach are presented, while at the same time our method is compared to the manual approach. The main drawback of our method is that, initially, it requires user interaction to initialize the snake. Further improvement of our model may eliminate the user implication even in the early stages of the process.

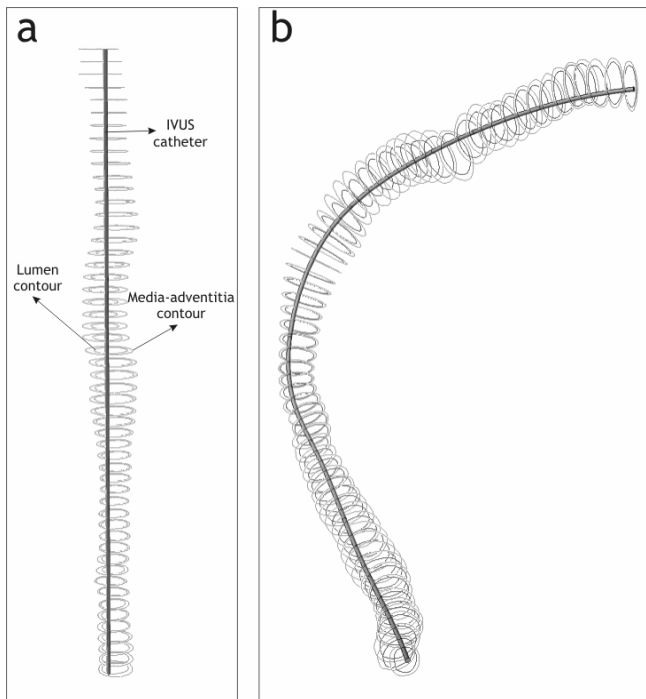


Fig. 5. A typical IVUS pullback resulting in a sequence of IVUS images which constitute the input of our automated segmentation model. Left-side, the detected lumen and media-adventitia borders positioned linearly along the IVUS catheter; right-side, the same sequence of contours after integration with biplane angiogram which provides the spatially correct 3D reconstruction of the catheter.

Provided that the proposed method facilitates the rapid and accurate contour detection in hundreds of IVUS images acquired during a routine pullback, it could potentially constitute a valuable tool for both diagnostic and research purposes. First, it could contribute in rapid plaque morphometric analyses including planimetric, volumetric and wall thickness measurements contributing to clinical decision-making [2], [3]. In the same manner, it could be utilized for the evaluation of plaque progression in follow-up studies after pharmaceutical or mechanical (e.g., stent) interventions [11]. In addition, this method could be integrated with biplane angiography for spatially correct 3D reconstruction of coronary arteries [12], [13], [5] enabling a more reliable evaluation of plaque spatial distribution, coronary 3D geometry estimation [14] and intracoronary flow simulation, in conjunction with computational fluid dynamics rules [12].

In future work, our interest will be focused on the development of a fully automated system for 3D reconstruction of coronary arteries integrating our segmentation model on it.

Table 1. Comparison of the proposed segmentation model with the manual approach

	Manual segmentation	Automated segmentation
Flexibility	Very low	High
Procedure complexity	Very high	Low
Time consuming	>15 hours for a set of 400 images	2 hours for a set of 400 images
Efficiency	Subjected to high inter- and intra-observer variability	Subjected to user initialization and snake's detection accuracy

Acknowledgments. This work was supported by the Greek State Scholarships Foundation and the Aristotle University Research Committee.

References

1. Mintz, G.S., Nissen, S.E., Anderson, W.D. et al.: American College of Cardiology clinical expert consensus document on standards for acquisition, measurement and reporting of intravascular ultrasound studies (IVUS). A report of the American College of Cardiology Task Force on clinical expert consensus documents. *J. Am. Coll. Cardiol.* 37 (2001) 1478-1492
2. Schoenhagen, P., Nissen, S.: Understanding coronary artery disease: tomographic imaging with intravascular ultrasound. *Heart* 88 (2002) 91-96
3. von Birgelen, C., de Very, E.A., Mintz, G.S., et al.: ECG-gated three-dimensional intravascular ultrasound: feasibility and reproducibility of the automated analysis of coronary lumen and atherosclerotic plaque dimensions in humans. *Circulation* 96 (1997) 2944-2952
4. Kirbas, C., Quek, F.: A review of vessel extraction techniques and algorithms. *ACM Computing Surveys* 36 (2004) 81-121

5. Giannoglou, G.D., Chatzizisis, Y.S., Sianos, G., Tsikaderis, D., Matakos, A., Koutkias, V., Diamantopoulos, P., Maglaveras, N., Parcharidis, G.E., Louridas, G.E.: In-vivo validation of spatially correct three-dimensional reconstruction of human coronary arteries by integrating intravascular ultrasound and biplane angiography. *Coron. Artery Dis.* 17 (2006) 533-543
6. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. *Int. J. Comp. Vision* 1 (1987) 321-331
7. Millman, R.S., Parker, G.D.: *Elements of differential geometry*. Prentice Hall, New Jersey, USA (1997)
8. Kompatsiaris, I., Tzovaras, D., Koutkias, V., Strintzis, M.G.: Deformable boundary detection of stents in angiographic images. *IEEE Trans Med Imaging* 19 (2000) 652-662
9. Intel Corporation. Open Source Computer Vision Library Reference Manual. December 8, 2000. Available at: <http://sourceforge.net/projects/opencvlibrary>
10. Pratt, W.K.: *Digital image processing: PIKS inside*. J Wiley and Sons (2001)
11. Klingensmith, J.D., Schoenhagen, P., Tajaddini, A. et al.: Automated three-dimensional assessment of coronary artery anatomy with intravascular ultrasound scanning. *Am. Heart J.* 145 (2003) 795-805
12. Stone, P.H., Coskun, A.U., Kinlay, S. et al.: Effect of endothelial shear stress on the progression of coronary artery disease, vascular remodeling and in-stent restenosis in humans. *Circulation* 108 (2003) 438-444
13. Slager, C.J., Wentzel, J.J., Schuurbers, J.C.H. et al.: True 3-dimensional reconstruction of coronary arteries in patients by fusion of angiography and IVUS (ANGUS) and its quantitative validation. *Circulation* 102 (2000) 511-516
14. Krams, R., Wentzel, J.J., Oomen, J.A. et al.: Evaluation of endothelial shear stress and 3D geometry as factors determining the development of atherosclerosis and remodeling in human coronary arteries in vivo. Combining 3D reconstruction from angiography and IVUS (ANGUS) with computational fluid dynamics. *Arterioscler. Thromb. Vasc. Biol.* 17 (1997) 2061-2065

Supervised Neuro-fuzzy Clustering for Life Science Applications

Jürgen Paetz

J.W. Goethe-Universität Frankfurt am Main,
60439 Frankfurt am Main, Germany

Abstract. Classification, clustering and rule generation are important tasks in multidimensional data analysis. The combination of clustering or classification with rule generation gives an explanation for the achieved results. Especially in life science applications experts are interested in explanations to understand the underlying data. The usage of supervised neuro-fuzzy systems is a suitable approach for this combined task. Not always classification labels are available for the data when considering new problem areas in life science. Since we had already used a supervised neuro-fuzzy system for some applications, our aim in the case studies was to use the same neuro-fuzzy classifier for clustering, generating understandable rules also for clusters. To do so, we added Monte-Carlo random data to the original data and performed the clustering task with the present classifier in the medical, chemical, and biological domain.

1 Introduction

The demand for explanations of underlying patterns in data analysis increases. Complex datasets cannot be *explained* simply by visualization or black box models like statistical regression or neural networks. Especially life science experts as medical doctors, chemists, or biologists are relied on explanations, e.g. for therapy planning [1], for designing new drugs [2], or for understanding biological processes [3] or sequence data [4]. Having class labels for datasets at hand, such as patient outcome or biological activity of molecules, classifiers can be used for performing the *classification* task. A supervised neural network approach is suitable for this task. However, not for every datasets class labels are available. Nevertheless, patient, molecule, and sequence patterns can be grouped to find common properties of data subgroups.

In *cluster analysis* mainly two general approaches are in use, some considering geometrical properties (distances of data samples), and other considering statistical properties (densities). For classification as well as for clustering the extraction of rules is beneficial. Rules explain the structure of the class or cluster regions, respectively. Many algorithms for clustering have been developed, for example the standard algorithms *k*-means clustering, and hierarchical linkage [5]. More recent methods are dynamical variant grids [6], neighborgrams for analysing chemical data [7], or self-organizing maps for clustering chemical

compounds [2] for example. A cluster analysis that is combined with fuzzy technology is fuzzy- k -means [8]. Another approach combining self-organizing maps with rule generation was proposed [9]. For using these clustering methods several parameters (e.g. minimum distances) have to be chosen. Our approach is different from other cluster methodologies because we do not propose a new cluster method but we use a present neuro-fuzzy classifier to perform cluster analysis. At first sight this seems to be an inconsistent approach since we have no class labels at hand when considering cluster problems. Using the neuro-fuzzy classifier [10,11], we demonstrate that the cluster task becomes solvable by adding random Monte-Carlo data (MC data) to the original data. This can be interpreted as adding artificial noise to the data. Usually, such noise is not useful at all and unwanted but we will see that additional MC data enables a classifier to perform the clustering task [12] with the benefit of having rules for regions where *no* original data is located.

In the next section we describe the main idea of the neuro-fuzzy classifier and the technical reasons why such a system can be enabled for clustering. We give a short overview of Monte-Carlo methods (MC methods) for our purposes. In section 3 we present the ideas of our clustering by neuro-fuzzy classification with added MC data. In section 4 we give several results on datasets that were chosen for case studies. We present an introductory example to make clear our ideas, then we present results on life science data based on three datasets stemming from medicine, chemistry, and biology. We end up with a conclusion.

2 Basics

At first we explain the geometrical idea of the neuro-fuzzy classification method [11] that was applied for example in the medical area [13], [14] for classification tasks. In section 2.2 we explain the idea behind MC methods to understand the following clustering process in section 3. We assume that we have a set of data samples $M := \{x_1, \dots, x_n\}$. Having class labels c_i , $i = 1, \dots, n$ at hand we write $M = \{(x_1, c_1), \dots, (x_n, c_n)\}$.

2.1 Neuro-fuzzy Classifier

The supervised neuro-fuzzy algorithm [11] is based on an adaptation process using the geometrical properties of the data $M = \{(x_1, c_1), \dots, (x_n, c_n)\}$. The basic neural network has two layers. It has neurons in the hidden layer with n -dimensional asymmetrical trapezoidal fuzzy activation functions. Every neuron in the first layer belongs to only one class and represents a fuzzy rule. During the learning phase these neurons p are adapted, i.e. the sides of the upper, smaller rectangles (called core rules) and the sides of the lower, larger rectangles (called support rules) of the trapezoids are adapted to the data. For every new training data point (x, c) this happens in four phases, initialized by the first training sample x_1 for which one neuron is committed with infinite side expansions in every dimension for the support rule and no (zero) expansion for the core rule (core rule = x_1):

1. *cover*: if x is an element of the region of a support rule of the same class c as x , expand one side of the corresponding core rule to cover x and increment the weight of the neuron.
2. *commit*: if no such support rule covers x , insert a new neuron p at point x of the same class and set its weight to one and its center $z := x$. The expansions of the sides of the support rule, associated with the new neuron, are set to infinite. The expansions of the sides of the core rule are set to zero.
3. *shrink committed neuron*: for a committed neuron shrink the volume of the support region within one heuristically chosen dimension of the neuron in relation to the neurons belonging to other classes.
4. *shrink conflict neurons*: for all the neurons belonging to other classes $\neq c$, heuristically shrink the volume of these support regions within one dimension in relation to x .

For performing the adaptation process class labels are needed. A clustering using only the data $\{x_1, \dots, x_n\}$ without class labels is not feasible directly.

2.2 Monte-Carlo Methods

MC methods are common statistical methods that use random data for heuristic data analysis. If a problem is too complex to solve it analytically such as optimal rule generation in high-dimensional spaces one can try to approach the problem by generating random data in the problem space. MC methods are invented in the 1940s where game outcomes, e.g. for roulette wheels, were studied [15].

We will give an example to introduce the MC method. We assume that we have an unit rectangle R with area $A(R)$. We assume that in this unit rectangle we have painted the figure of an elephant E . If we want to determine the area $A(E)$ of the elephant drawing, we can hardly use exact integration methods. If we generate uniform random data (MC data) within R we can count the data in the region E and in the region R . The value $|E|/|R|$ is an approximation of $A(E)$. The more MC data we generate the more exact becomes the result. This is an easy example where random data can be used for performing a data analysis task.

3 Monte-Carlo Clustering

In this section we explain the combination of the neuro-fuzzy classification process with the MC method to perform clustering. Then, we discuss advantages and disadvantages of our approach. A first theoretical insight is given to the problem of how much MC data should be generated when considering the rule generation problem. In section 4 we give a detailed example and application results on real world data.

3.1 The Algorithm

Let us consider the dataset $M = \{x_1, \dots, x_n\}$. We assign the artificial class label “original” (“G”) to the data and obtain $M := \{(x_1, G), \dots, (x_n, G)\}$ with

$c_1 = \dots = c_n = G$. All data are labeled with the same class label. The neuro-fuzzy classifier cannot be used without data having another different class label. Thus, we generate MC data, i.e. uniform distributed data within a region S where $M \subset S$. The set S should not be chosen geometrically much larger than M what can be controlled by minimum and maximum values. We label this data with “C”, and obtain the set $N := \{(y_1, C), \dots, (y_m, C)\}$. The number m of MC data can be chosen similar to n , i.e. $m \sim n$ for small datasets. It is $N \subset S$, too. Now, we have a second, different class label, and we can use the classifier above. To simplify the whole clustering procedure, we assume that MC data were not randomly generated within a small range of the present data samples x_i . This can be achieved by checking the Euclidean distance d between x_i and y_j that should not become too small. Without this restriction the core and support regions of several generated rules during classifier adaptation may become small, so that an additional pruning step has to be done. We will discuss this problem in section 4.1 again. The algorithm is summarized in Algorithm 3.1.

Algorithm 3.1: (MC Clustering)

Label the data samples with “G” to obtain $M := \{(x_1, G), \dots, (x_n, G)\}$;
 Generate MC data and assign class label “C”: $N := \{(y_1, C), \dots, (y_m, C)\}$;
 Restriction: Do not generate y_j with $d(x_i, y_j) < \delta$ for all $i = 1, \dots, m$;
 Perform classification task with the neuro-fuzzy classifier;
 Only when *not* using the restriction (optional):
 Prune too small rules of class “C” with corresponding MC data;
 Train again the classifier without the pruned data;
 Identify generated rules with neighboring core regions as one cluster;
end Algorithm 3.1

Advantages of Algorithm 3.1 are:

- Cluster membership is not 0 or 1 only, but a fuzzy membership $\in [0, 1]$ (fuzzy clustering). The core elements can be identified as representative elements for the clusters.
- We have both clusters for the data, and for regions where originally no data was given (clusters of empty regions). This is helpful for identifying new clusters if data changes over time.
- We have (fuzzy) rule descriptions for the clusters what is important for high-dimensional data that cannot be visualized easily.
- We can use the same method for classification and clustering (increasing convenience for users). The description of rules for classification and clustering is of an identical format.

The advantages of our approach are significant but problems are caused by the MC data:

1. Additional MC data increases run time.
2. If MC data is generated within a data cluster more rules will be generated for one cluster.
3. We do not know exactly how many MC data samples should be generated.

The first point is a fact that is directly connected to the third point. The more MC data points we add the higher the run time becomes. Our aim is to generate only as many MC data points as needed for clustering. If a cluster can be embedded in one hyper rectangle, then it is represented by only one hyper rectangle with a center. Large clusters, that have not a rectangular shape, are described by more than one hyper rectangle. This assures the finding of representative data points for diverse regions within a cluster. Thus, additional rules caused by adding MC data is not a general problem if the clusters are not splitted too intensely.

The question for the number of MC data samples remains open. For smaller datasets with only few dimensions we propose that the number of MC data can be chosen similar to the number of the original data samples. If the restriction distance is difficult to determine, e.g. if data clusters are overlapping, one can omit the restriction step and perform an additional pruning and training step or use an online pruning technique [16]. The run time, using about the same number of MC data as the number of data samples, is maintainable. We will use this heuristic in our introductory example. Nevertheless, for large datasets with a high number of dimensions we need a finer heuristic for MC data generation. Let us consider two extreme cases for two clusters in a unit hyper cube.

- 1) We have two dense clusters C_1, C_2 with only a small hyper diagonal stripe as cluster border where no data are located. The probability for a uniform random MC data point to be located in a cluster and not on the hyper diagonal is $p(x \in C_i) \approx 1$.
- 2) We have two point-like dense clusters in two different edges of the hyper cube. Then, the probability for a uniform random MC data point to be located in a point-like cluster and not in the large area between the clusters is $p(x \in C_i) \approx 0$.

Thus, it is not exactly possible to determining a-priori a suitable number of MC data points without including knowledge about the data. In the second case the number of MC data is not critical while in the first case almost every MC data point cause rule splitting. Thus, in our first analyses of real life datasets in the Sections 4.2 - 4.4 we tried out the adding of the same number of MC data samples as data samples are available in the original datasets.

4 Experiments on Artificial and Life Science Data

We present results on four datasets, one experiment with artificial data, one with septic shock patient data, one with chemical data, and one on malaria pathogen sequence data. For every real life dataset we will give a short introduction in the expert domain.

4.1 Introductory Example

For our introductory example we generated data from two normal distributions: 100 samples from a normal distribution $\mathcal{N}(0, 10)$ (with mean 0 and S.D. 10) and

100 samples from $\mathcal{N}(60, 10)$. The union of these data samples is noted as set M . Our task is to find the two clusters in M , cf. Fig. 1.

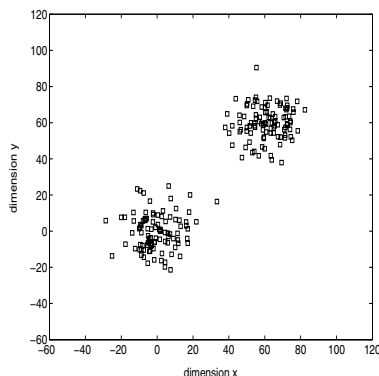


Fig. 1. The dataset M : data randomly generated by two normal distributions

For using Algorithm 3.1 we generated 250 uniformly distributed MC data samples in the region $S := [-60, 120] \times [-60, 120]$. Thirty MC data samples were removed due to the restriction criteria with $\delta := 5$. The set of 220 samples is noted as N , cf. Fig. 2.

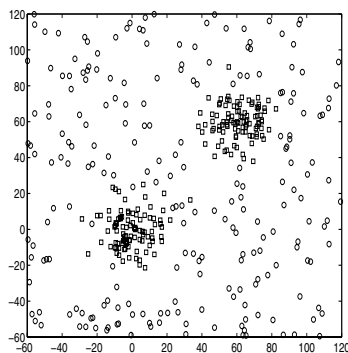


Fig. 2. The dataset $M \cup N$: data of Fig. 1 (squares) with added MC data (circles)

The results of Algorithm 3.1 (neuro-fuzzy classification with data and MC data) are depicted in Fig. 3. We remark that due to a technical transformation of S to $[0, 1] \times [0, 1]$ the labels of the axes are within the unit cube. An expansion of one rule to 1.2 means an infinite expansion, made finite for visualization ($-0.2 = -\infty$). The rectangles of the core rules are filled. One projected trapezoid corresponds to one rule or one neuron, respectively.

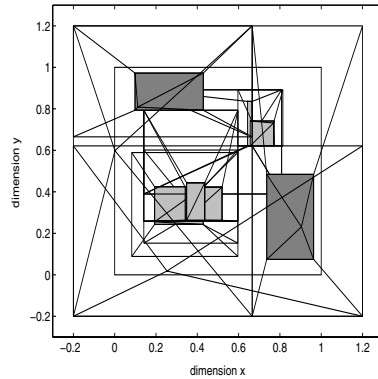


Fig. 3. Generated rules. Rules with overlapping core regions can be considered as one cluster. The value 1 in the figure corresponds to 120, value 0 to -60 , cf. the visualization remark in the text.

In Fig. 3 we see in the upper left area one rule for class “C”, meaning that there were originally no data. We see another rule in the lower right area. We have another two smaller rules of class “C” that cannot be seen well in Fig. 3 since they are too small. Such rules could be pruned. In the lower left area we see three rules of class “G” that represent one cluster. In the upper right area there are two rules of class “G”, representing another cluster. Two rules of class “G” are too small and cannot be seen easily in Fig. 3. They are candidates for pruning again. In total we have four clusters with non-overlapping core regions, two of class “G” and two of class “C”. The upper left core region can be described by the following rule:

If x_1 in $[72.12, 112.77]$ and x_2 in $[-46.22, 27.37]$ then class “C” (empty region).

One larger core region of the right upper cluster can be described by:

If x_1 in $[57.78, 78.38]$ and x_2 in $[52.00, 73.55]$ then class “G” (cluster 1).

As we have already remarked, it is generally not possible to describe one cluster only by one rule since one cluster needs not to have a rectangular shape.

4.2 Medical Septic Shock Patient Data

Septic shock is a severe problem in intensive care medicine. Surgical patients suffering from immune system reactions and are extensively medicated [1]. In the MEDAN project (www.medan.de) we collected data of 382 abdominal septic shock patients. We documented and analyzed most of the commonly documented vital parameters and doses of medicine (metric variables) beside others. Data of 382 patients were collected in German hospitals from 1998 to 2001. The mortality rate was 49%, i.e. 187 of the 382 patients died during their stay

at the ICU. We found out that the three parameters systolic blood pressure [mmHg], diastolic blood pressure [mmHg], and thrombocytes [$1000/\mu\text{l}$] are the most relevant variables for outcome prediction [17]. We used 1100 samples for 382 patients, one for every day of the last three days in the ICU. Some patients have less than three days of ICU stay. Here, we take these data to perform the clustering task. We train the system with 50% of the samples and test it with the remaining 50% to calculate the performance measures. We added 1100 uniformly distributed MC data samples to the 1100 original data samples. The two out of 45 resulting rules with the best test confidence are:

If systolic blood pressure **in** (78.61, 117.16) **and** diastolic blood pressure **in** (32.29, 57.77) **and** thrombocytes **in** (32, 216) **then** class “patient data” (test frequency 9.3%, test confidence 95.1%).

If systolic blood pressure < 122.19 **and** diastolic blood pressure > 75.97 **and** thrombocytes > 489 **then** class “no patient data” (test frequency 3.9%, test confidence 95.3%).

The results show that many different clusters have been generated for septic shock patients. Overall we have 45 clusters, 23 for regions with septic shock patient data, and 22 for empty regions. The diversity of the clusters demonstrates the individuality of the patients. In this sense the results confirm earlier results where we found no significant clusters for survived and deceased patients with linkage clustering [18].

4.3 Chemical Compound Data

In drug design it is of prime importance to identify molecules with desired properties like biological activity or non-toxicity. This task, virtual screening, can be performed using intelligent algorithms [19]. Usually, molecular properties are encoded in a vector representation, a so called descriptor vector. For experimentation we used the CATS2D (Chemically Advanced Template Search for two-dimensional structures) descriptor which encodes 2D topological information about the atom types in the molecular graph [20]. The CATS2D descriptor is a 150-dimensional vector, but after feature selection, using an importance measure [14], only less dimensions remained.

For a first experiment we used all entries in the COBRA (Collection of Bioactive Reference Analogues) compound library [21], including 4705 compounds in its version 2.1, but restricted to 21 dimensions that were relevant for MMP (matrix metalloproteinase) ligands. For a second experiment we used only the 35 PPAR (peroxisome proliferator-activated) ligand molecules, restricted to the relevant 13 dimensions. In the following rules the letters d, a, p, n, l are assigned for the membership of an atom to one of the five groups of hydrogen-bond donors (d), hydrogen-bond acceptors (a), positively charged atoms/groups (p), negatively charged atoms/groups (n), and lipophilic atoms/groups (l), respectively. For example the pair “aa5” corresponds to a shortest path in the molecular graph from “a” to “a” with length five.

First experiment: We added 4705 uniformly distributed 21-dimensional MC data samples (100%), having 9410 data samples in total. Then, 4705 data samples of both classes were randomly chosen for training, the rest for testing. Two out of nine resulting rules are:

If “aa1” < 0.22 and “al5” < 0.66 and “dd6” < 0.67 and “dd8” < 0.35 then class “not empty cluster” (test frequency 42.6%, test confidence 100.0%).

If “dd8” > 0.27 then class “empty region” (test frequency 51.2%, test confidence 85.5%).

This result is plausible since most of the one-dimensional distributions of the original data are skewed with a lot of zero values. Empty regions are regions where a value for one or more variables are greater than thresholds. We have not found a meaningful, specific cluster for MMP ligands. A training with a self-organizing map with all dimensions in [21], Fig. 5a showed that there are no interesting clusters when using all the data due to the similarity of different ligands.

Second experiment: The 35 PPAR samples were taken, adding 35 uniformly distributed 13-dimensional MC data samples (100%), having 70 data samples in total. We present all three rules that we have found for these data:

If “dp0” < 1.84 and “aa3” < 0.27 then class “PPAR data” (test frequency 50.0%, test confidence 100.0%).

If “aa3” > 0.27 then class “not PPAR data” (test frequency 66.7%, test confidence 75.0%).

If “dp0” > 0.00 then class “not PPAR data” (test frequency 55.6%, test confidence 90.0%).

In a trained SOM [21], Fig. 5b only one statistically meaningful cluster was found in the PPAR data with 100% test confidence. Thus, we have a reasonable description of the data cluster and of the empty regions.

4.4 Biological Malaria Pathogen Sequence Data

Human malaria is caused by the parasite *Plasmodium falciparum*. The lethality rate is about 30% together in the American, Asian, and African continent. A problem is caused by drug-resistance of malaria pathogens so that new drugs need to be found. Bioinformatics is supposed to be helpful in finding useful hints in the *Plasmodium falciparum* data. The genome of *Plasmodium falciparum* is composed of 14 chromosomes [22]. For analysis, we have taken the 5287 known peptide sequences from web pages of The Institute for Genomic Research (TIGR), <http://www.tigr.org/>, located at ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/p_falciparum/annotation_dbs/. At the TIGR web pages more information about the sequencing methodology and the involved groups can be found. We determined the percentage of the 20 amino acids in each one

of the sequences that are stored in the FastA format. We added 5287 uniformly distributed 20-dimensional MC data samples, having then 10574 data samples. We give two of the ten resulting rules with letters coding the amino acids:

If “A” < 16% and “D” < 11% and “F” < 10% and “M” < 15% and “Y” < 11% then class “data” (test frequency 47.7%, test confidence 100.0%).
If “D” > 7% and “Y” > 4% then class “no data” (test frequency 50.9%, test confidence 80.9%).

Most of the data is located in regions where the percentage is moderate for some amino acids. The regions with a higher amino acid percentage where no data is located, are not confident with 100%, i.e. there are samples in the “no data” region with a higher amino acid percentage but there seem to be no significant clusters of these samples.

5 Conclusion

Motivated by classification tasks that we had analysed with a neuro-fuzzy system we approached the clustering problem by supervised neuro-fuzzy classification. We have performed cluster analysis without introducing a new cluster method, but by using the present neuro-fuzzy classifier. This was possible only by generating randomly additional MC data with an additional class label. Then, clusters can be identified by considering the resulting classification rules. An advantage of supervised clustering with a classifier and MC data are the additional rules for empty regions. If one has non-stationary data, then one can use the information about empty clusters to detect changes in the data space over time. Additional MC data can be interpreted as artificial noise. Contrary to the fact that usually noise is unwanted and filtered, MC data makes this clustering possible. We explained the whole procedure by an introductory example. Then, we applied the method to three life science datasets to demonstrate the general practicability of the Monte Carlo method. We analysed medical septic shock data where many clusters were found, chemical compound data and biological sequence data, where only few clusters were found.

The approach could be promising for high dimensional sparse data without class labels where it seems reliable to add additional MC data, so that cluster borders can be learned. An open problem is still the efficient adaptation or calculation of an optimal number of MC data, based on properties of the data since we demonstrated that it is in general not possible to calculate such a number without a-priori knowledge about the data. Here, we have used about the number of the original data samples for MC data as a first approach. The influence of MC data to the results needs to be studied in greater detail in further experiments. It can be studied if the MC data approach can be used for other classifiers as well. Another part of the future work is the more theoretical determination of an optimal number and location of MC data points with a-priori knowledge.

Acknowledgement. We thank the MEDAN team for collecting and discussing septic shock patient data and the Chair for Bio- and Cheminformatics at the J.W. Goethe University Frankfurt for let me have the chemical data for research.

References

1. R. M. Hardaway: A review of septic shock. *American Surgeon* **66** (2000) 22–29
2. G. Schneider, P. Wrede: Artificial neural networks for computer-based molecular design. *Biophysics & Molecular Biology* **70** (1998) 175–222
3. R. Brause: Adaptive modeling of biochemical pathways. In: *Proc. of the 15th IEEE Int. Conf on Tools with Artificial Intelligence*, Sacramento, CA, USA (2003) 62–68
4. M.J. Gardner: The genome of the malaria parasite. *Current Opinion in Genetics and Development* **9** (1999) 704–708
5. R.O. Duda, D.G. Stork, P.E. Hart: *Pattern Classification and Scene Analysis Part 1: Pattern Classification*. Wiley & Sons, New York, 2000
6. E. Schikuta, M. Erhart: The BANG-clustering system: Grid-based data analysis. In: *Proc. of the 2nd Int. Symp. on Intelligent Data Analysis*, London, Great Britain (1997) 513–524
7. M.R. Berthold, B. Wiswedel, D.E. Patterson: Neighborgram clustering interactive exploration of cluster neighborhoods. In: *Proc. of the 2nd IEEE Int. Conf. on Data Mining*, San Jose, CA, USA (2002) 581–584
8. J.C. Bezdek: *Pattern Recognition with Fuzzy Objective Function*. Plenum Press, New York, 1981
9. A. Ultsch, H.P. Siemon: Kohonen's self-organizing feature maps for exploratory data analysis. In: *Proc. of the Int. Conf. on Neural Networks*, Paris, France (1990) 305–308
10. J. Paetz: Metric rule generation with septic shock patient data. In: *Proc. of the 1st IEEE Int. Conf. on Data Mining*, San Jose, CA, USA (2001) 637–638
11. K.-P. Huber, M.R. Berthold: Building precise classifiers with automatic rule extraction. In: *Proc. of the IEEE Int. Conf. on Neural Networks*, Perth, Western Australia (1995) 1263–1268
12. J. Paetz: Monte-Carlo clustering by neuro-fuzzy classification. In: *Proc. of the 1st Indian Int. Conf. on Artificial Intelligence*, Hyderabad, India (2003) 66–72
13. R. Silipo, M.R. Berthold: Discriminative power of input features in a fuzzy model. In: *Proc. of the 3rd Int. Symp. on Intelligent Data Analysis*, Amsterdam, The Netherlands (1999) 87–98
14. J. Paetz: Knowledge based approach to septic shock patient data using a neural network with trapezoidal activation functions. *Artificial Intelligence in Medicine* **28** (2003) 207–230
15. N. Metropolis, S. Ulam: The Monte Carlo method. *J. Amer. Stat. Assoc.* **44** (1949) 335–341
16. J. Paetz: Reducing the number of neurons in radial basis function networks with dynamic decay adjustment. *Neurocomputing* **62** (2004), 79–91
17. J. Paetz, B. Arlt: A neuro-fuzzy based alarm system for septic shock patients with a comparison to medical scores. In: *Proc. of the 3rd Int. Symp. on Medical Data Analysis*, Rome, Italy (2002) 42–52
18. F. Hamker, J. Paetz, S. Thöne, R. Brause, E. Hanisch: Erkennung kritischer Zustände von Patienten mit der Diagnose Septischer Schock mit einem RBF-Netz. *Interner Bericht 4/00*, Fachbereich Informatik, J.W. Goethe-Universität Frankfurt am Main, 2000

19. H.J. Böhm, G. Schneider: Virtual Screening for Bioactive Molecules. Wiley VCH, Weinheim, 2000
20. G. Schneider, W. Neidhart, T. Giller, G. Schmid: Scaffold hopping by topological pharmacophore search: a contribution to virtual screening. *Angewandte Chemie, International Edition* **38** (1999) 2894–2895
21. P. Schneider, G. Schneider: Collection of Bioactive Reference Compounds for Focused Library Design. *QSAR & Combinatorial Science* **22** (2003) 713–718
22. M.J. Gardner, N. Hall, E. Fung, O. White, M. Berriman and others: Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419** (2002) 498–511

Study on Preprocessing and Classifying Mass Spectral Raw Data Concerning Human Normal and Disease Cases

Xenofon E. Floros¹, George M. Spyrou², Konstantinos N. Vougas²,
George T. Tsangaris², and Konstantina S. Nikita¹

¹ National Technical University of Athens
Electrical and Computer Engineering Faculty
9, Iroon Polytechniou str, 15773, Athens, Greece
xfloros@biosim.ntua.gr, knikita@cc.ece.ntua.gr

² Foundation for Biomedical Research of the Academy of Athens
4, Soranou Efessiou str, 11527, Athens, Greece
{gs Spyrou, gt Tsangaris, kvougas}@bioacademy.gr

Abstract. Mass spectrometry is becoming an important tool in biological sciences. Tissue samples or easily obtained biological fluids (serum, plasma, urine) are analysed by a variety of mass spectrometry methods, producing spectra characterized by very high dimensionality and a high level of noise. Here we address a feature extraction method for mass spectra which consists of two main steps : In the first step an algorithm for low level preprocessing of mass spectra is applied, including denoising with the Shift-Invariant Discrete Wavelet Transform (SIDWT), smoothing, baseline correction, peak detection and normalization of the resulting peak-lists. After this step, we claim to have reduced dimensionality and redundancy of the initial mass spectra representation while keeping all the meaningful features (potential biomarkers) required for disease related proteomic patterns to be identified. In the second step, the peak-lists are aligned and fed to a Support Vector Machine (SVM) which classifies the mass spectra. This procedure was applied to SELDI-QqTOF spectral data collected from normal and ovarian cancer serum samples. The classification performance was assessed for distinct values of the parameters involved in the feature extraction pipeline. The method described here for low-level preprocessing of mass spectra results in 98.3% sensitivity, 98.3% specificity and an AUC (Area Under Curve) of 0.981 in spectra classification.

Keywords: ovarian cancer, mass spectra preprocessing, biomarkers, feature extraction, early diagnosis, classification.

1 Introduction

When we refer to the proteome, we refer to the entirety of proteins in existence within an organism at a given point in time. While the genome is a rather constant entity, the proteome differs from cell to cell and is constantly changing through its biochemical interactions with the genome and the environment.

Thus, studying the proteome captures not only the genomic background of the cell, but also the impact of its environment. Such a knowledge enables the discovery of disease biomarkers and drug targets. One of the most important tools in proteomics is mass spectrometry, as it can supply the researchers with quantitative data on the proteome itself. New techniques and algorithms on processing and classifying mass spectral raw data are constantly emerging, towards effective and accurate biomarker discovery and early diagnosis models.

In this paper algorithms for preprocessing mass spectra are proposed, extracting meaningful features that capture most of the biological information hidden in the original noisy spectra and finally performing high-accuracy classification to unknown (blind) spectra. This work aims, initially, to the development of an open-source clinical decision support system capable of classifying any kind of mass spectra (SELDI, MALDI-TOF etc.), having as goal optimized biomarker discovery and early diagnosis.

Although the basic ideas highlighted in this paper are not introduced for the first time in handling mass spectra, the approach followed differs from other works on the following issues: The main tool used is denoising based on Shift-Invariant Discrete Wavelet Transform (SIDWT). Wavelets have been widely used before in mass spectrometry studies, [1,2,6,7]. Coombes *et al* use the undecimated decomposition in order to extract peaks from spectra, addressing the problem of appropriate threshold setting through visual inspection, [1]. Nevertheless, they do not follow up with classification. Kalousis *et al* investigate various classification schemes but denoising is carried out by means of a simple decimated wavelet transform,[2], which theoretically is outperformed by our SIDWT approach, [9,10]. Finally, Qu *et al* and Lee *et al* analyse the mass spectra data in the wavelet domain, [6,7]. On the contrary, we choose to return to the initial representation of the mass spectra and extract features (peaks) that have a profound biological meaning in proteomics.

Apart from the wavelet denoising we discuss the step of peak alignment across spectra, highlighting a problem that is not clarified in other approaches. One of the main problems in preprocessing and classifying mass spectra is the appropriate selection of the parameters involved in the whole procedure. Furthermore, this selection has to be automatic and based on objective metrics assessing the performance of the model. Thus, the problem of defining a measure to be optimized is arisen. Kalousis *et al*, propose the obvious measure of classification accuracy, since the goal is classification and biomarker discovery, which is being maximized through the proper selection of only one parameter of the preprocessing pipeline, [2]. In our approach, we study the effect of all the parameters involved in preprocessing and feature extraction on the resulting classification, reaching a critical subset of the parameters that strongly affect our model performance. Then, we propose a feedback model coupling the classification block to the preprocessing and feature extraction procedure, in order to automatically optimize the critical parameters according to the estimated classification performance. We provide two metrics to estimate that performance, accuracy and the more robust AUC metric of ROC analysis.

2 Approach

2.1 Biological Samples and Generation of Spectra

For our study we used the high-resolution SELDI-TOF ovarian cancer dataset provided in [17] and used by Conrads *et al.*, [8]. Serum samples were obtained from the National Ovarian Cancer Early Detection Program (NOCEDP) and gynecologic oncology clinic at Northwestern University (Chicago, IL, USA). Ninety (90) specimens from women enrolled in the NOCEDP who had no evidence of any cancer for 5 years and were evaluated as being healthy, were used. Furthermore, 90 preoperative specimens were used from women who were surgically staged and found to have epithelial ovarian carcinoma. The mass spectrometer was externally calibrated with a mixture of known peptides. For the spectra evaluation an error tolerance of $\pm 400ppm$ was used.

2.2 Raw Mass Spectra Preprocessing Pipeline

A mass spectrum, produced by any kind of mass spectrometry instrument, can be viewed as a vector with dimensionality equal to the number of distinct m/z values recorded by the instrument and the value of each dimension is the intensity of the corresponding m/z value. Preprocessing of each spectrum is of vital importance, [1,2]. In our approach we divide the pipeline in the following subtasks: *Resampling* $\rightarrow$ *Denoising* $\rightarrow$ *Smoothing* $\rightarrow$ *Baseline subtraction* $\rightarrow$ *Peak detection* $\rightarrow$ *Peak Normalization* $\rightarrow$ *Peak Allignment*. In the following subsections we describe how we tackled each of these issues. Let $y^{raw} \in \mathbf{R}^{N_1}$ be the raw mass spectrum, produced by the spectrometer, that enters the pipeline.

Resampling Spectra. In discrete signal processing, the time interval between samples is kept constant (for example, sample every millisecond) unless externally clocked. However, the raw mass spectral data's mass steps (the mass differences of two data points which are next to each other) are not uniform at different positions in a spectrum. For the signal itself, no matter whether the x-axis means the flight time or the mass weight, the intervals are not constant for the reason that the ions would not reach the detector in a constant time interval. Therefore, the inconstant property is inconvenient for discrete signal processing. For instance, when we apply discrete wavelet decomposition to the vector of discrete intensities in wavelet denoising, we transform the vector from the time/mass domain to the wavelet domain by the assumption that the sample intervals are constant. Furthermore the SIDWT we use to perform denoising requires vector length of a power of two. Thus, we chose a piecewise cubic spline interpolation with a resampling step resulting in a signal $y^{resampled}$ with length $N = 2^d$, where $d: 2^{d-1} < N_1 < 2^d$. The spectrum m/z range is [1000 12000] Da.

Denoising - Reducing High-Frequency Noise. We follow the approach proposed by Lang *et al.*, [9,10], and used in mass spectra denoising by Coombes *et al.*, [1]. In terms of implementation, our method relies on the SIDWT as implemented in v2.4 of the Rice Wavelet Toolbox (RWT), [18].

More specifically we propose the following scheme for denoising:

1. Compute the SIDWT (using Beylkin algorithm) $Y = Wy^{resampled}$, where W is the left invertible transformation matrix of the SIDWT.
2. Perform hard thresholding in the wavelet domain on the wavelet coefficients :

$$\hat{X} = T_h(Y, thresh) = \begin{cases} Y & \text{if } |Y| \geq thresh \\ 0 & \text{if } |Y| < thresh \end{cases} \quad (1)$$

where $\hat{X}$ is an estimate of the DWT transform of the true signal x and $thresh$ the selected threshold value.

3. Compute the inverse DWT $\hat{x} = M\hat{X} = y^{denoised}$
4. Smooth the denoised spectrum $y^{smoothed} = \mathcal{S}(y^{denoised})$, where $\mathcal{S}$ is the selected smoothing function

The smoothing step inserted in the procedure is essential for the whole pipeline and the feature extraction as it eliminates the small number of high frequency oscillations, still present, in the denoised spectrum. In order to get a completely smooth signal we coupled the denoising procedure with a locally weighted linear regression method using a tricubic kernel. From another point of view, the above denoising scheme simply averages the result of classic wavelet denoising for all possible shifts of the input signal. Making the assumption that true peaks in the time domain are represented by a small number of (relatively large) coefficients in the wavelet domain, while noise is distributed over most wavelet coefficients, we expect that this averaging (thresholding in SIDWT domain) would eliminate the noise component, leading to a more robust result.

In the SIDWT-based denoising Daubechies of order 8 have been chosen as wavelet basis, which have been reported previously to have a good performance on mass spectral data, [1,2]. Using this wavelet basis leads to a maximum decomposition level of 15, hence a reasonable choice is a level choice of 10 which results in a smooth signal while preserving true peak information. Concerning the threshold scheme, hard thresholding was selected, as Lang *et al* have shown that, in SIDWT-based denoising, it provides both good l_2 performance and smoothness properties and outperforms soft thresholding in most cases, [9,10]. In order to define the threshold value several approaches were studied including a heuristic selection based on Stein's unbiased risk estimate, adaptive thresholding on each decomposition level etc. but a simple global threshold was finally selected, as for the data we studied it outperformed the other schemes. Thus, in the rest of the paper we define wavelet threshold $waveThres$: $thresh = waveThres \times MAD/0.6745$, where MAD is the mean absolute deviation of the wavelet coefficients and $thresh$ the threshold appearing in equation 1 .

Estimating and Subtracting Baseline. One of the artifacts affecting the spectra is the baseline, which is a mass-to-charge dependent low-frequency offset on which the information-bearing component of the spectra is superimposed. Generally, mass spectra exhibit a monotonically decreasing baseline, originated

by ionized matrix material clusters. In order to be able to compare peak intensities across spectra we have to estimate the baseline and then subtract it from the original mass spectrum. To estimate the baseline we used a previous proposed algorithm calculating baseline points within multiple shifted windows, [14]. We consider that in each window the estimated baseline point is the p -quantile of the intensities distribution, i.e if $Y(i), i = 1, \dots, W_j$ the vector of mass spectra intensities in the specified window j , of length W_j , then the baseline point $B(j)$ satisfies : $\text{Prob}(Y < B(j)) = p$. In order to ensure that points lying on peaks are not selected, we set p to 7% and window size to 50Da. Such a small window size, leads to 240 estimated points and possible outliers may still be present. Hence, to eliminate their effect on the resulting curve we perform a locally weighted smoothing using least squares linear polynomial fitting. Finally, we regress the estimated points across all windows to a smooth curve, using a shape-preserving piecewise cubic interpolation. Having estimated the baseline, we subtract it from the smoothed spectrum getting $y^{baseCorrected}$.

Detecting Peaks. Peak picking relies on the assumption that, in the previous preprocessing steps, the noise level has been estimated and removed from the signal. Thus, a possible algorithm simply has to detect all local maxima in the mass spectrum, leading to an over-estimated number of peaks. We enhance this approach considering as outliers all the peaks with a SNR below a pre-defined threshold. In order to retain the maximum possible amount of information, while eliminating peaks that seem to belong in the noise component, we keep the SNR threshold in low level (around 2). The peak detection algorithm, described above, receives as input $y^{baseCorrected}$ and generates a peak-list $\{p_1, \dots, p_n\}$, with each peak p_i being described by a tuple (MZ_i, I_i, SNR_i) where MZ_i the peak m/z , I_i the peak intensity corresponding to $y^{baseCorrected}$ and SNR_i the peak SNR.

Normalizing Peak Intensities. Intensities in different peak-lists cannot be compared directly, as large experimental variations still remain in the data due to different conditions during the spectra preparation. Thus, we need to normalize the peak-lists in order to even out these variations and allow comparisons across spectra, [3,15]. In our pipeline we provide the following normalization schemes:

1. Total Ion Current (TIC) normalization

$$I'_i = \frac{I_i}{\sum_{i=1}^n I_i} \quad . \quad (2)$$

2. Unity Vector Length normalization

$$I'_i = \frac{I_i}{\sum_{i=1}^n I_i^2} \quad . \quad (3)$$

3. Root Mean Square normalization

$$I'_i = \frac{I_i}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n I_i^2}} \quad . \quad (4)$$

4. Z-Score normalization

$$I'_i = \frac{I_i - \bar{I}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (I_i - \bar{I})^2}} . \quad (5)$$

where I'_i the normalized intensity of the peak p_i , n the length of the peak-list and $\bar{I}$ the average intensity of the peak-list.

Aligning Peaks. The peak-alignment step tries to compensate for the spectrometer measurement error present in every spectrum. This error results in an uncertainty whether observed peaks across spectra are the same or not, i.e if they reflect the presence of the same proteomic marker. In order to align peaks across a group of peak-lists we designed a greedy algorithm which performs clustering of the peaks on their m/z dimension. More specifically, the peak-lists are used to fill a $M \times 3$ matrix, where M is the number of all peaks and the three columns contain the metrics characterizing each peak. The matrix is sorted along the m/z dimension (with cost $\mathcal{O}(M \log M)$). Then starting from the smaller m/z we add peaks in clusters under the following constraints: Two peaks cannot belong in the same cluster if their m/z distance is greater than the measurement error or if they originate from the same spectrum. The final clusters contain masses from different spectra that correspond to the same peak, represented by the median of the cluster. Although we refer to the peak alignment step as part of the preprocessing pipeline, in practise we apply aligning, just before classification, as we will describe in the following section.

2.3 Feature Extraction

Having implemented the preprocessing pipeline, we have, at the same time, completed feature extraction from the raw mass spectrum. We have to mention that, while for a single spectrum features are the peaks detected by the algorithm, when we examine a set of spectra, we also consider as features the flat signal areas where the spectrum doesn't have a peak but, at least, one of the other spectra presents a peak. Thus, the dimensionality of the feature space and the result of the classification highly depends on three of the preprocessing steps: denoising-smoothing, peak detection and peak alignment. In other words, the parameters associated with each step affect the dimensionality of the produced feature space. These key-parameters are: $waveThres \in \mathbf{R}^+$, $smoothed \in \{true, false\}$, $snrThres \in \mathbf{R}^+$, and $alignError \in \mathbf{N}$.

2.4 Classification

In a real diagnosis model, a classifier is trained on a specific training set and unknown spectra, consisting the test set, are preprocessed independently, i.e without adding any knowledge to the whole model. Then, the classifier is responsible for the disease identification across the test set. For the classification we used an SVM with a simple linear kernel, as there are promising studies showing

that SVMs outperform other learning algorithms (such as decision trees, nearest neighbours algorithms etc) in classifying mass spectral data, [2,4,5]. Let N training data points $\{x_i\}_{i=1}^N \in \mathbb{R}^d$, where d the reduced dimensionality of the feature space, and $y_i \in \{-1, 1\}$ their labels. Using a linear kernel we are trying to specify a linear hyperplane $f(x) = 0$ separating the two classes while at the same time having the maximum separating margin with respect to the two classes, where $f(x) = \langle w, x \rangle + b$ and $w, x \in \mathbb{R}^d, b \in \mathbb{R}$. The class associated to a new example $x \in \mathbb{R}^d$ is given by:

$$class = \text{sig}(f(x)) = \begin{cases} 1 & \text{if } f(x) \geq 0 \\ -1 & \text{if } f(x) < 0 \end{cases}$$

At this point we would like to discuss about the peak alignment step that, in a sense, remains a "black box" in other approaches. So far, this step is performed during the preprocessing and it is not clarified whether peaks are aligned across all spectra of the dataset, or not. It is easy to verify that different clusters are formed if we perform alignment across the whole dataset and individually across subsets of the dataset. We claim that, when the peak lists of a dataset are aligned together, we extract knowledge from the whole dataset reflected in the formed clusters. Thus, in a real diagnosis case, we cannot align spectra across the whole dataset (consisting of the training and the test set). This is the reason why we don't consider alignment as part of the preprocessing pipeline. In our approach, we align separately the normal peak-lists and the cancer peak-lists of the training set, before training the SVM classifier. Then, for each peak-list of the test set, alignment is performed across the peak-lists consisting the training set and the current peak-list to be classified. Finally, the aligned features of the current spectrum are used for the classification.

3 Results - Evaluating a Real Diagnosis Case

3.1 Evaluation Metrics

In order to evaluate the performance of the proposed algorithm, we adopt metrics widely used in the assessment of classification models. Besides True Positive Ratio (TPR), True Negative Ratio (TNR) and their weighted average, Accuracy, we plot Receiver Operating Characteristic (ROC) curves and calculate the Area Under Curve (AUC). A common weakness of the first three metrics is that they are not robust to the change of the class distribution, i.e when the proportion of positive to negative instances changes in a test set, they may no longer perform optimally. On the other hand, the ROC curve is insensitive to such changes, while it is proved that AUC is a better measure for model evaluation than accuracy, [16]. As the depiction of a ROC curve for an SVM classifier is not trivial, we describe the method we used to plot it, addressing the use of a threshold value $t \in (-\infty, +\infty)$ so that :

$$class = \text{sig}(f(x)) = \begin{cases} 1 & \text{if } f(x) \geq t \\ -1 & \text{if } f(x) < t \end{cases}$$

When t goes from $-\infty$ to $+\infty$ we obtain a set of classifiers whose performances (True Positive Ratio and False Positive Ratio) are used to draw the ROC curve. For example, when t moves from 0 to $-\infty$ it is clear that positive examples are most likely to be correctly classified at the expense of adding more false positive errors and these errors are increased as we move towards $-\infty$ (where $\text{TPR}=\text{FPR}=1$). On the contrary, when t moves from 0 to $+\infty$ we increase the true negative ratio, increasing at the same time the false negative ratio towards an all-negative classifier (where $\text{TPR}=\text{FPR}=0$). In fact, we do not have to test threshold values in the whole $\mathbb{R}$ but only in a subrange specified by the training data points. Specifically, when we train the SVM we calculate $f(x_i) = \langle w, x_i \rangle + b \quad \forall i = 1, \dots, N$. Then it is sufficient to study t in the range $[\min\{f(x_i)\}, \max\{f(x_i)\}]$. For each threshold value t , in this range, we obtain a set of classifiers (from all-positives to all-negatives), each classifier is used to classify all the example data points x according to $\text{sig}(f(x) - t)$ and calculate the TPR and FPR which define a single point on the ROC curve.

In order to calculate the AUC we followed the nonparametric estimate, i.e the summation of the areas of the trapezoids formed by the points on the ROC curve. It represents the probability that a randomly selected positive instance will score higher than a randomly selected negative instance.

3.2 Studying the Influence of Critical Parameters

The alignment error is set equal to the mass spectrometer error tolerance, as defined in the spectra generating procedure. For the obtained dataset, this error is considered to be 400 ppm, thus *alignError* = 400ppm. Initially, we study the effect of the normalization scheme used and also the effect of the smoothing step. In the rest of the results section, when we study a subset of the variables, the remaining variables are considered fixed ¹, unless otherwise stated.

From the classification results we conclude that the four normalization approaches do not seriously affect the performance, although normalizing with respect to Root Mean Square performs better than the other schemes both in terms of accuracy and AUC. Thus, in the following, we fix the normalization scheme to RMS. Concerning the smoothing step, we processed the spectra for 6 cases, combining 3 values of the *waveThres* (*waveThres* = 10, 20, 30) and performing or not smoothing. From the classification results, listed in Table 1, it's obvious that not applying smoothing, after the denoising step, seriously deteriorates the model performance. This is an expected result, as we observed that, after denoising the spectrum, there was still a high-frequency noise component present, which is eliminated with the smoothing procedure. Thus, we conclude that smoothing is an essential step for a succesful mass spectra classification and we fix the *smoothed* variable to true. After fixing the values of the three parameters mentioned above, two critical parameters, *waveThres* and *snrThres*, are left to be studied.

¹ Fixed values are *smoothed* = true, *waveThres* = 30, *snrThres* = 2, *alignError* = 400.

Table 1. Effect of the Smoothing step on performance

[waveThres,smoothed]	TPR	TNR	Accuracy %	AUC	features
[10,true]	0.883	0.883	88.333	0.977	491
[10,false]	0.667	0.650	65.833	0.773	1852
[20,true]	0.917	1.000	95.833	0.916	283
[20,false]	0.750	0.417	58.333	0.841	463
[30,true]	0.983	0.983	98.333	0.975	206
[30,false]	0.567	0.750	65.833	0.638	290

As we saw in the section of feature extraction, the SNR threshold highly affects the dimensionality of the resulting feature space, thus we expect to affect the classification performance. Indeed, Figure 1 confirms that relation, where we see that $snrThres > 7$ seriously deteriorates the performance. Another result that depicts from this figure is that the $snrThres$ is strongly connected with the value of the $waveThres$ and those two variables cannot be treated independently. We also observe that the AUC metric exhibits a much smaller deviation than accuracy, a result connected to the robustness of the metric.

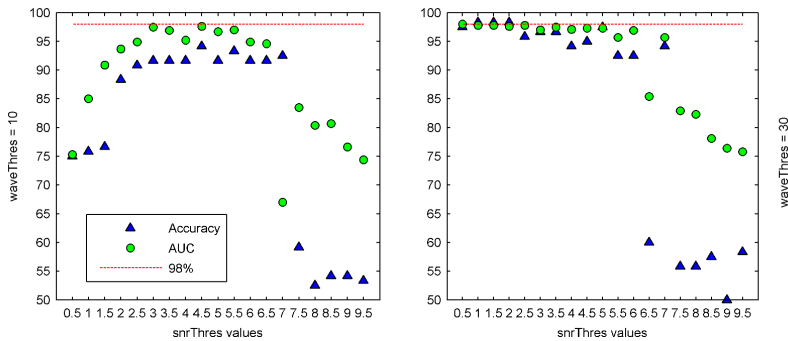


Fig. 1. Accuracy and AUC results for varying $snrThres$ and $waveThres$ values

3.3 Optimizing Critical Parameters

Classification performance is measured by accuracy and AUC, so the problem arising is the selection of the parameters that would maximize one of the metrics. Thus, the problem reduces to the maximization of the function $y = f(\mathbf{x})$, where $y = AUC$ or $y = Accuracy$, $y \in [0, 100]$, and $\mathbf{x} = (waveThres, snrThres)$, $\mathbf{x} \in A_1 \times A_2$ ². The function to be maximized depicts after coupling the entire preprocessing pipeline with the classification algorithm, letting as free variables the two critical parameters and returning the selected classification performance

² $A_1 \times A_2$ is the search subspace for the two parameters, $A_1 \times A_2 \subset \mathbb{R}^+ \times \mathbb{R}^+$.

metric. It is obvious that calculating a single value $y = f(\mathbf{x})$ is computationally expensive, as all spectra have to be preprocessed and classified. Thus, a parameter range, $A_1 \times A_2$, has to be specified by the user. The maximization was carried out using a simple exhaustive search algorithm based on a 10-fold cross validation scheme which provides a robust estimation of the performance. The algorithm takes as input the search subspace for the critical parameters and returns $\mathbf{x}_{opt}$, maximizing the classification performance. The optimal critical parameters selected by the algorithm, for both metrics, are listed in Table 2, while in 2 we can see the ROC curve corresponding to the parameters maximizing AUC.

As described before, we assess the model performance for a wide range of the critical parameters and then, we define the range where the model seems to perform better as the search subspace for the optimum settings that maximize the performance. This approach promises to find with a great propability the parameters optimizing the classification. Furthermore, we have to mention that if the algorithm finds two combinations that result in the same performance, it selects the one that guarantees for a lower False Negative Ratio, as an ideal diagnosis system has to eliminate all the FP cases, i.e all the patients that are diseased and the algorithm doesn't detect it.

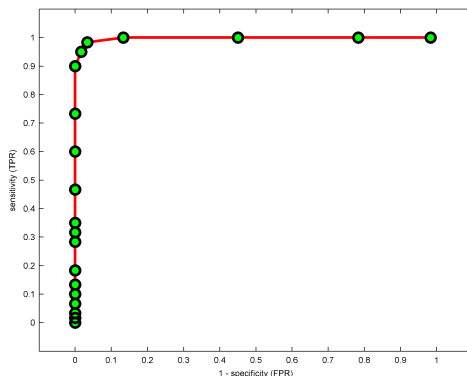


Fig. 2. Optimal ROC curve for $waveThres = 25$ and $snrThres = 2$

Table 2. Parameters optimizing the classification performance

Chosen metric	<i>waveThres</i>	<i>snrThres</i>	Accuracy	AUC
Accuracy	30	2	98.33	0.976
AUC	25	2	97.50	0.981

4 Conclusion

The present study describes a work in progress towards the development of an open-source decision support system that will be able to classify any kind of

mass-spectra in a high throughput environment (clinical routine). This system is being developed in a modular way in order to be easily executed in platforms that support parallel computing (grid-computing). Actually, the whole preprocessing pipeline is performed independently for each spectrum, thus can be executed in parallel.

One main issue highlighted in the study concerns the parameters engaged in the proposed model. Our goal is to discover the parameter values that maximize the performance. Thus, from a huge search space consisting of all possible parameter values we have to end up with strictly defined values. The approach followed tries to prune the search space by dividing the variables in two main categories: The critical and the non-critical parameters. The non-critical parameters are fixed after a short inspection of their effect on classification. The critical parameters, *waveThres* = 30 and *snrThres*, are estimated within a 10-fold cross validation scheme searching for the values maximizing the classification performance. We concluded that the optimum critical values that maximize accuracy are : *waveThres* = 30, *snrThres* = 2 which drive to 98.3 % specificity and 98.3 % sensitivity (resulting from a FPR=FNR=1/60), while the optimal values maximizing AUC are : *waveThres* = 25, *snrThres* = 2 which lead to an AUC of 0.981.

We have to note here that the primary goal of the proposed algorithm is to extract a compact, less redundant representation of the mass spectra that retains as much as possible the initial discriminatory content. In current work this representation is used to effectively classify normal and diseased spectra in a real diagnosis case. Future work will focus on feature selection methods that use this representation to discover biomarkers, in cases where proper and adequate initial datasets are available.

References

1. Kevin R. Coombes, Spiridon Tsavachidis, Jeffrey S. Morris, Keith A. Baggerly, Mien-Chie Hung, and Henry M. Kuerer.: Improved Peak Detection and Quantification of Mass Spectrometry Data Acquired from Surface-Enhanced Laser Desorption and Ionization by Denoising Spectra with the Undecimated Discrete Wavelet Transform. *Proteomics*. **Nov;5(16):4107-17** (2005)
2. Alexandros Kalousis, Jullien Prados, Elton Rexhepaj and Melanie Hilario : Feature extraction from mass spectral data for the classification of pathological states. In *Principles of Data Mining and Knowledge Discovery*, Ninth European Conference. Springer (2005)
3. Witold E Wolski, Maciej Lalowski, Peter Martus, Ralf Herwig, Patrick Giavalisco, Johan Gobom, Albert Sickmann, Hans Lehrach, Knut Reinert.: Transformation and other factors of the peptide mass spectrometry pairwise peak-list comparison process. *BMC Bioinformatics*. **6: 285** (2005)
4. Zhang X, Lu X, Shi Q, Xu XQ, Leung HC, Harris LN, Iglehart JD, Miron A, Liu JS, Wong WH.: Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics*. **Apr 10;7:197** (2006)

5. Wagner M, Naik D, Pothén A.: Protocols for disease classification from mass spectrometry data. *Proteomics*. **Sep;3(9):1692-8** (2003)
6. Qu, Y., Adam, B.I., Thornquist, M., Potter, J.D., Thompson, M.L., Yasui, Y., Davis, J., Schellhammer, P.F., Cazares, L., Clements, M., Wright, G.L., Feng, Z.: Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensional data. *Biometrics*. **59 : 143-151**(2003)
7. Lee, K.R., Lin, X., Park, D., Eslava, S.: Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. *Proteomics* **3** (2003)
8. Conrads TP, Fusaro VA, Ross S, Johann D, Rajapakse V, Hitt BA, Steinberg SM, Kohn EC, Fishman DA, Whitely G, Barrett JC, Liotta LA, Petricoin EF 3rd, Veenstra TD.: High-resolution serum proteomic features for ovarian cancer detection. *Endocrine-Related Cancer*. **11:163-178** (2004)
9. Lang M, Guo H, Odegard JE, Burrus CS, Wells RO Jr.: Nonlinear processing of a shift invariant DWT for noise reduction. *Mathematical Imaging: Wavelet Applications for Dual Use, SPIE Proceedings*, **vol. 2491**, Orlando FL (1995)
10. Lang M, Guo H, Odegard JE, Burrus CS, Wells RO Jr.: Noise Reduction Using an Undecimated Discrete Wavelet Transform. *IEEE Signal Processing Letters*. **3, 10-12** (1996)
11. Donoho D.L.: De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*. **41(3):613-627** (1995)
12. Donoho D.L, Johnstone I.M.: Ideal spatial adaptation by wavelet shrinkage. *Biometrika*. **81:425-455** (1994) Also Tech. Report 400, Department of Statistics, Stanford University, July (1992)
13. Beylkin G.: On the representation of operators in bases of compactly supported wavelets. *SIAM J. Numer. Anal.* **29(6):1716-1740** (1996)
14. Lucio Andrade and Elias Manolakos.: Signal Background Estimation and Baseline Correction Algorithms for Accurate DNA Sequencing. *Journal of VLSI, special issue on Bioinformatics* **35:3 pp 229-243** (2003)
15. Alfassi Zeen B.: On the normalization of a mass spectrum for comparison of two spectra (2004)
16. J. Huang and C.X. Ling.: Using AUC and Accuracy in Evaluating Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*. **17(3) 299-310** (2005)
17. Ovarian Cancer DataSet <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>
18. Rice Wavelet Toolbox Licence <http://www.dsp.rice.edu/software/RWT/LICENSE>

Non-repetitive DNA Sequence Compression Using Memoization

K.G. Srinivasa¹, M. Jagadish², K.R. Venugopal³, and L.M. Patnaik⁴

¹ Data Mining Laboratory, M S Ramaiah Institute of Technology, Bangalore
kgsrinivas@msrit.edu

² Software Engineer, MindTree Consulting, Bangalore
jagadish88@gmail.com

³ Professor, University of Visvesvaraya College of Engineering,
Bangalore University, Bangalore 560 001

⁴ Professor, Microprocessor Application Laboratory,
Indian Institute of Science, Bangalore 560 012

Abstract. With increasing number of DNA sequences being discovered the problem of storing and using genomic databases has become vital. Since DNA sequences consist of only four letters, two bits are sufficient to store each base. Many algorithms have been proposed in the recent past that push the bits/base limit further. The subtle patterns in DNA along with statistical inferences have been exploited to increase the compression ratio. From the compression perspective, the entire DNA sequences can be considered to be made of two types of sequences: repetitive and non-repetitive. The repetitive parts are compressed using dictionary-based schemes and non-repetitive sequences of DNA are usually compressed using general text compression schemes. In this paper, we present a memoization based encoding scheme for non-repeat DNA sequences. This scheme is incorporated with a DNA-specific compression algorithm, *DNAPack*, which is used for compression of DNA sequences. The results show that our method noticeably performs better than other techniques of its kind.

Keywords: DNA Compression, Memoization, Text Compression.

1 Introduction

The bases found in DNA come in four varieties: adenine, cytosine, guanine, and thymine often abbreviated as A, C, G, and T, the letters of the genetic alphabet. Genome sequencing is finding the order of DNA nucleotides, or bases, in a genome the order of As, Cs, Gs, and Ts that make up an organism's DNA. Sequencing the genome is an important step towards understanding it. A genome sequence does contain some clues about where genes are, even though scientists are just learning to interpret these clues. The human genome is made up of over 3 billion of these genetic letters. The human genome is about 20-40 percent repetitive DNA, but bacterial and viral genomes contain almost no repetition. In repetitive DNA, the same short sequence is repeated over and over again.

Somewhere in the genome the sequence GCA may be repeated 100 times in a row; elsewhere there may be 30 consecutive copies of the sequence ACTTCTG. For example, the following DNA sequence is just a small part of telomere located at the ends of each human chromosome:

```
...GGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGT
TAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTT
AGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGG...
```

An entire telomere, about 15 kb, is constituted by thousands of the repeated sequence "GGGTTA".

Based on the repetition rate, DNA sequences are divided into three classes:

- Highly repetitive: About 10-15% of mammalian DNA reassociates very rapidly. This class includes tandem repeats.
- Moderately repetitive: Roughly 25-40% of mammalian DNA reassociates at an intermediate rate. This class includes interspersed repeats.
- Single copy : This class accounts for 50-60% of mammalian DNA.

Deriving meaningful knowledge from DNA sequence will define biological research through the coming decades and require the combined effort biologists, chemists, engineers, and computational scientists, among others. Some of the research challenges in genetics include Gene regulation, DNA sequence organization, Chromosomal structure and organization, Non-coding DNA types and functions, coordination of gene expression, protein synthesis, and post-translational events interaction of proteins in complex molecular machines, evolutionary conservation among organisms, Protein conservation, correlation of SNPs (single-base DNA variations) with health and disease, etc.

With increasing number of genome sequences being made available, the problem of storing and using databases has to be addressed. Conventional text compression schemes are not efficient when DNA sequences are involved. Since DNA sequences contain only 4 bases *A, G, T, C*, each base can be represented by 2 bits. However standard compression tools like *compress*, *gzip* and *bzip2* have more than 2 bits per base when compressing DNA data. Consequently, DNA compression has become a challenge. Algorithms like *GenCompress* [1], *Biocompress* [2], *Biocompress-2* [3], that use the characteristics of DNA like point mutation or reverse complement achieve a compression rate about 1.76 bits per base [4]. Many compression methods have been discovered to compress DNA sequences. Invariably, all the methods found so far take advantage of the fact that DNA sequences are made of only 4 alphabets, together with techniques to exploit the repetitive nature of DNA [5]. The algorithm given here is used to encode non-repetitive regions. The popular techniques that have been shown to be efficient in compressing non-repeat regions are Order-2 Arithmetic Coding and Context Tree Weighting Coding. Order-2 coding overcomes the constraint of Huffman coding that the symbol to be encoded has to be coded by round number of bits. The adaptive nature of coding has been an advantage. The adaptive probability of a symbol is computed from the context after which it appears. Order-2 algorithm usually have better compression ratios with high efficiency in general. The

Context Tree Weighting Coding was proposed by Willems [6] and has a good compression ratio for and unknown model. The CTW encoder has two parts: a source modeler which is the actual CTW algorithm, which receives the uncompressed data and estimates the probability of the next symbol and an encoder which uses the estimated probabilities to compress the data. The context tree is built dynamically during the encoding decoding process. All of the visited substring of shorter size than a fixed bound, exist as a path in the tree. Each node of the tree contains a probability. In order to encode a given bit, the following steps are performed: the path in the context tree which coincides with the current context is searched and if needed extended. For every node in this context path, an estimated probability of the next symbol is computed using weighting function on all the estimated probability values. The weighted probability is sent to the arithmetic encoder which encodes the symbol, and the encoder goes to the next symbol [7]. The repetitive regions are compressed using methods given in [8,9].

2 Related Work

The first DNA-specific algorithms were given by Grumbach and Tahi [2,3]. Two algorithm namely, *BioCompress* and *BioCompress - 2*, that were based on Ziv and Lempel data compression method [10,11] were proposed. *BioCompress - 2* detects exact repeats and complementary palindromes located earlier in the target sequence, and then encodes them by repeat length and the position of a previous repeat occurrence. If no significant repetition was found then the arithmetic coding of order-2 was used. The use of arithmetic encoding was the only difference between *BioCompress* and *BioCompress - 2*.

Cfact algorithm was proposed by E. Rivals et al, which searches the longest exact matching repeat using suffix tree data in an entire sequence. *Cfact* was similar to *BioCompress - 2* except for being a two-pass algorithm, where the first pass involved building the suffix tree. In the second phase, repetitions are coded with guaranteed gain; else, two-bit per base encoding was used. The compression algorithm to detect the approximate tandem repeats in DNA sequences was later given in [8]. Approximate string matching was used to provide some lossy compression algorithms by [12]. However, lossy algorithms were of little use in DNA compression.

A better compression algorithm than *BioCompress* and *BioCompress - 2* is GenCompress [1,13,14]. The basic idea is to approximate repetitions. There are two variants of *GenCompress*-one that uses hamming distance for repeats and the other uses the edition distance (deletion, insertion and substitution) for the encoding of the repeats.

CTW - LZ [4] algorithm is based on context tree weighting method. It basically works by combining LZ-77 type algorithm like *GenCompress* with *CTW* algorithm. The long repeats are coded by LZ77 while the short repeats are encoded by CTW. The execution time of *CTW + LZ* is not impressive although it does achieve good compression ratios.

DNACompress [15] is a two-phase algorithm that employs the Ziv-Lempel compression scheme as *BioCompress* – 2 and *GenCompress*. The first phase finds all approximate repeats including complementary palindromes, using a special software, PatternHunter [16]. The second phase involves coding of non-repeat regions and approximate repeat regions. The *DNACompress* achieved a better execution time in general than *GenCompress*.

DNAC [17] is another DNA compression algorithm that works in four phases. The first phase involved building of suffix tree to find exact repeats. The second phase involved extending exact matches to approximate matches using dynamic programming. In the third phase, it extracts the optimal non-overlapping repeats from overlapping ones. The repeats are encoded in the last phase. Some of the recent lossless compression for large DNA microimages compression is given in [18,19,20]. *DNAPack* that uses hamming distance for repeats and complementary palindromes, and either CTW or Arth-2 compression for non-repeat regions is proposed in [21]. The algorithm marginally performs better than the earlier mentioned algorithms due to selection of repeat regions using dynamic programming method rather than greedy approach. In this paper, we propose another encoding algorithm based on memoization that is useful in coding non-repeats. Section 3 describes the idea behind the algorithm along with pseudocode. In Section 4, we describe the setup and evaluate the performance of the proposed encoding scheme by comparing the results obtained by incorporating our encoding scheme into *GenCompress* and *DNAPack* algorithms along with results obtained by other algorithms.

3 Algorithm

The encoding scheme works in two passes. Each pass is identical, except that the symbols being encoded are different. The following method is used to represent the bases:

- In the first pass alphabets A and G are represented by A; T and C are represented by T.
- In the second pass alphabets A and C are represented as A; G and T are represented by T.

For example, the sequence **AGTC** would be taken as **AATT** and **ATTA**. The decoding procedure requires both the sequences, with four possible combinations of A and T representing each base as shown in Table 1.

The above substitutions is made to the entire DNA sequence. Let the pass 1 substitution sequence be S_1 and pass 2 sequence be S_2 . If the length of the sequences is l then the each linear sequence is transformed to a matrix of dimension $\alpha \times \beta$ where $\alpha * \beta = l$. The transformation is done row-wise and hence the entry in i th row and j th column (zero-indexed) would correspond to alphabet in $i * \beta + j$ position in the sequence.

The choice of α and β can be made in different ways. A simple approach would be to break the sequence into multiple sequences each of length of a perfect

Table 1. Mapping scheme for decoding

Sequence 1	Sequence 2	Base
A	A	A
A	T	G
T	T	T
T	A	C

square, chosen greedily, and represent each sub-sequence as a square matrix. This is always possible since any number can be represented as sum of squares. A more efficient way to determine the split is to consider all possible combinations of α and β and take the combination that leads to maximum compression ratio. This increases the complexity considerably but can be made non-prohibitive if the length of the longest sequence is restricted to a certain maximum value. The encoding and decoding of S_1 and S_2 are done independently. The compression technique works on the matrices obtained by the sequences. The encoding idea is based on the idea of recursively dividing the matrix into sub-matrices until each sub-matrix is composed of a single alphabet. The division of matrix can be done in one of the following ways:

- L : The matrix can be divided into left half and right half. If the number of columns is odd the center column is included in the left half.
- U : The matrix can be divided into upper half and lower half. If the number of rows is odd the center row is included in the upper half.
- C : The matrix can be split into even columns and odd columns. The leftmost column is considered even.
- R : The matrix can be split into even and odd rows. The first row is considered as even.

For example, the matrix

AAAATTTT
AAAATTTT
AAAATTTT

can be divided into two 3×4 sub-matrices:

AAAA TTTT
AAAA TTTT
AAAA TTTT

the left sub-matrix can be encoded as A and the right sub-matrix can be encoded as T. Hence the entire division of matrix would be represented by “LAT”. The letter A and T denote that the each sub-matrix is composed only of alphabet A and T, respectively. If the matrix is a combination of both the alphabets, it is divided into one of the four ways mentioned above. The first letter gives the division type followed by encoding of left/upper/even submatrix followed by corresponding right/lower/odd submatrix. For example, the matrix :

```

ATATATAT
TATATATA
ATATATAT
TATATATA

```

would be encoded the best as CRATRTA. The first column split produces two submatrices:

```

AAAA    TTTT
TTTT    AAAA
AAAA    TTTT
TTTT    AAAA

```

The subsequent letters (RAT and RTA) further describe each of the subimages, until only 2×4 images of As and Ts are left.

```

AAAA    TTTT
AAAA    TTTT

TTTT    AAAA
TTTT    AAAA

```

Decompressing the encoded string requires the knowledge of the original size of the matrix. The dimension is not encoded in the string. This is stored as a separate array of integers and is used at the time of decompression. The compression ratio is given taking into account the space needed to store the sizes of each matrix.

The simplest implementation (Algorithm 1) is a recursive function that tries all possible ways of dividing the matrix and keeps the solution having the minimum encoded length, with memoization to keep it efficient. The main function is the *compress* which takes the matrix as the argument and returns the best encoding string possible. The matrix is considered as array of strings. The termination condition occurs when all the entries in the matrix have the same alphabet. Otherwise *compress* makes use of four functions *splitrow*, *splitcolumn*, *splitupper* and *splitleft* in order to try all possible combinations of division of matrix recursively. *mem* is the map data structure that holds the values for each matrix for which the best encoding string has been found.

For simplicity, the algorithm is shown without incorporating any of the optimizations that reduce the running time significantly. The functions in algorithm takes the matrices themselves as arguments thus maintaining copies of matrices of their own. This can be avoided by representing the sub-matrices by 6 parameters and keeping a single copy of the matrix.

- row: the topmost row in the sub-matrix
- col: the leftmost column in the sub-matrix
- rowC: the number of rows in the sub-matrix
- colC: the number of columns in the sub-matrix
- rowS: the distance between adjacent rows in the sub-matrix, relative to the original matrix
- colS: the distance between adjacent columns in the sub-matrix, relative to the original matrix

Algorithm 1. Encoding algorithm

Input : Matrix consisting of As and Ts

Output: Encoded string corresponding to input matrix

mem is the data structure used to store matrices and their corresponding encoded strings if already found

t_s is local string ; s_1 and s_2 are matrices local to the functions

function *compress* (matrix *mat*)

if *mem*[*mat*] not empty **then**

 return *mem*[*mat*]

end if

if *mat* contains only single alphabet(τ) **then**

mem[*mat*] $\leftarrow \tau$

 return *mem*[*mat*]

else

if *mat* has more than one column **then**

$t_s \leftarrow \text{"C"} + \textit{splitcolumn}(\textit{mat})$

if *mem*[*mat*] is empty OR $\text{len}(t_s) < \text{len}(\textit{mem}[\textit{mat}])$ **then**

mem[*mat*] $\leftarrow t_s$

end if

$t_s \leftarrow \text{"L"} + \textit{splitlower}(\textit{mat})$

if *mem*[*mat*] is empty OR $\text{len}(t_s) < \text{len}(\textit{mem}[\textit{mat}])$ **then**

mem[*mat*] $\leftarrow t_s$

end if

end if

if *mat* has more than one row **then**

$t_s \leftarrow \text{"R"} + \textit{splitrow}(\textit{mat})$

if *mem*[*mat*] is empty OR $\text{len}(t_s) < \text{len}(\textit{mem}[\textit{mat}])$ **then**

mem[*mat*] $\leftarrow t_s$

end if

$t_s \leftarrow \text{"U"} + \textit{splitupper}(\textit{mat})$

if *mem*[*mat*] is empty OR $\text{len}(t_s) < \text{len}(\textit{mem}[\textit{mat}])$ **then**

mem[*mat*] $\leftarrow t_s$

end if

end if

 return *mem*[*mat*]

end if

function *splitcolumn* (matrix *mat*)

return *compress*(even columns of *mat*)+*compress*(odd columns of *mat*)

function *splitrow* (matrix *mat*)

return *compress*(even rows of *mat*)+*compress*(odd rows of *mat*)

function *splitleft* (matrix *mat*)

return *compress*(first half columns of *mat*)+*compress*(second half columns of *mat*)

function *splitupper* (matrix *mat*)

return *compress*(first half rows of *mat*)+*compress*(second half rows of *mat*)

Initialization: $BestComp[0] = 0$
Recurrence:

$$BestComp[i] = \min \begin{cases} BestComp[j] + CopyCost(j, i, k) & \forall k \forall 0 < j < i \\ BestComp[j] + PalinCost(j, i, k) & \forall k \forall 0 < j < i \\ BestCopy[j] + MinCost(j + 1, i) & \forall 0 < j < i \end{cases}$$

Fig. 1. Dynamic programming scheme for finding best compression

Initially, row and col are the top left corner of the original matrix, and both step sizes are 1.

- Left-Right: Sets colC to half for the left sub-matrix and colC-half for the right sub-matrix, where half is $(colC+1)/2$. Also sets col to $col+half*colS$ for the right sub-matrix. Similar method is used for rowC for Upper-lower split.
- Even-Odd Columns: Sets colC to half for the even sub-matrix and colC-half for the odd sub-matrix, where half is $(colC+1)/2$. Also sets col to $col+colS$ for the odd sub-matrix, and colS to $2*colS$ for both sub-matrices. Similar method is used for rowC and rowS in even-odd row split.

4 Experimental Results

4.1 Setup

Since the compression algorithm proposed is specifically made to compress non-repeat DNA sequences, a fair method of evaluation of performance can be made only by combining the compression scheme with another DNA-specific algorithm that exploits repeating sequences. For this purpose we use *DNAPack*, which is found to outperform most other methods available. We briefly describe the working of *DNAPack* and the method of incorporating memoization algorithm into it to achieve better results. *DNAPack* is based on dynamic programming for selection of segments as opposed to greedy methods of selection [22]. Let s be the input sequence. Let $BestComp[i]$ be the smallest compressed size of prefix $s[1 \dots i]$. The recurrence given in Fig 1 is the general scheme of dynamic programming.

$CopyCost(j, i, k)$ is the number of bits needed to encode the substring of size k starting at position i if it is an approximate repeat of the substring of size k starting at j . The $PalinCost$ is similarly defined for reverse complementary substrings. The function $MinCost(j + 1, i)$ is the number of bits needed for compression of the segment $s[j + 1, i]$. It depends on the size of the substring as well as the compression ratio obtained for the algorithm by arithmetic coding or CTW. $MinCost$ allows the creation of repeat segment if it would yield a benefit in the compression ratio [21]. We replace the CTW or arithmetic coding with our compression algorithm. The modified algorithm is referred as *DNAMem*. The performance of the *DNAMem* is evaluated along with other popular algorithms of its kind.

4.2 Comparison

The comparison of results were made between the conventional text compression algorithms with *DNAMem*. Table 2 gives the compression ratios expressed as *bitsperbase*. The first column gives the DNA sequence name. The second column gives the sequence length. Columns from 3-10 gives the *bitsperbase* value of all the algorithms. The *DNAMem* algorithm performs better than all the algorithms in all cases. Table 3 shows the comparison of DNA-specific algorithms. The *DNAMem* performs slightly better than all others in 7 out of 11 sequences considered. BC2, GC and DNAC refer to BioCompress2, GenCompress, DNACompress respectively.

Table 2. Comparison with text-compression algorithms

DNA Sequence name	sequence length	gzip-9	lz(1M)	arith(1M)	PPMD+	adapted PPMD+	normal CTW	CTW-4	DNAMem
CHNTXX	121024	2.220	2.234	1.866	1.977	1.840	1.879	1.838	1.6601
CHNTXX	155844	2.291	2.300	1.956	2.062	1.934	1.974	1.933	1.6101
HEHCMVCG	229354	2.279	2.286	1.985	2.053	1.965	1.997	1.958	1.8349
HUMDYSTROP	38770	2.377	2.427	1.948	2.237	1.921	1.960	1.920	1.9084
HUMGHCSA	66495	1.551	1.580	1.438	2.077	1.694	1.376	1.363	1.0311
HUMHBB	73308	2.228	2.255	1.911	2.116	1.921	1.917	1.892	1.7765
HUMHDABCD	58864	2.209	2.241	1.950	2.130	1.948	1.909	1.897	1.7395
HUMHPRTB	56737	2.232	2.269	1.942	2.130	1.932	1.922	1.913	1.7884
MPOMTCG	186609	2.280	2.289	1.961	2.075	1.966	1.989	1.962	1.8925
PANMTPACGA	100314	2.232	2.249	1.873	2.018	1.872	1.902	1.866	1.8533
SCCHRII	315339	2.265	2.268	1.935	2.023	1.950	1.976	1.945	1.8331
VACCG	191737	2.190	2.194	1.862	2.002	1.910	1.897	1.857	1.7582

Table 3. Comparison with DNA-specific compression algorithms

DNA Sequence name	sequence length	BC2	GC	CTW-LZ	DNAC	DNAPack	DNAMem
CHMPXX	121024	1.6848	1.6730	1.6690	1.6716	1.6602	1.6601
CHNTXX	155844	1.6172	1.6146	1.6120	1.6127	1.6103	1.6101
HEHCMVCG	229354	1.8480	1.8470	1.8414	1.8492	1.8346	1.8349
HUMDYSTROP	33770	1.9262	1.9231	1.9175	1.9116	1.9088	1.9084
HUMGHCSA	66495	1.3074	1.0969	1.0972	1.0272	1.0390	1.0311
HUMHBB	73308	1.8800	1.8204	1.8082	1.7897	1.7771	1.7765
HUMHDABCD	58864	1.8770	1.8192	1.8218	1.7951	1.7394	1.7395
HUMHPRTB	56737	1.9066	1.8466	1.8433	1.8165	1.7886	1.7884
MPOMTCG	186609	1.9378	1.9058	1.9000	1.8920	1.8932	1.8925
PANMTPACGA	100314	1.8752	1.8624	1.8555	1.8556	1.8535	1.8533
VACCG	191737	1.7614	1.7614	1.7616	1.7580	1.7583	1.7582

4.3 Execution Time Analysis

The experiment was conducted on Pentium 4 machine with 256MB RAM running Red Hat Linux 9. On an average the execution time taken by *DNAMem*

was 26mins, while that of *DNAPack* was 1min. The high execution time is the result of high asymptotic complexity of the memoization algorithm. Despite the improvements made the complexity remains $O(n^5)$, where n is the size of longest non-repeat sequence.

5 Conclusion

We have presented a new algorithm for encoding non-repeat parts of DNA sequences. Twelve sequences were used for experimentation. The algorithm proposed was combined with *DNAPack* compression algorithm to evaluate the performance of compression with other algorithms of its kind. The results obtained show that *DNAMem* clearly outperforms conventional text compression algorithms and marginally does better than DNA-specific algorithms. The compression ratio of *DNAMem* was the best for 7 out of 11 sequences considered. However, the time taken by *DNAMem* was around 25 times slower than other similar algorithms.

Acknowledgments

The Project is partially supported by the AICTE, as a part of Career Award of Young Teachers(AICTE File No.: F. No.1-51/FD/CA/(9)2005-06) to Mr.K.G. Srinivasa, who is presently working as a faculty in Department of Computer Science and Engineering, M. S. Ramaiah Institute of Technology, Bangalore – 560 054, India.

References

1. Xin Chen, Sam Kwong, and Ming Li. A compression algorithm for dna sequences and its application in genome comparison. *genomic*, 12:512–514, 2001.
2. Grumbach S and F. Tahi. Compression of dna sequences. *Data compression conference*, pages 340–350, 1993.
3. Grumbach S and F. Tahi. A new challenge for compression algorithms genetic sequences. *Journal of Information processing and Management*, 30:866–875, 1994.
4. Matsumoto T, Sadakane K, and Imai H. Biological sequences compression algorithms. *Genome Information Ser. Workshop Genome Inform*, 11:43–52, 2000.
5. E. Rivals, J-P. Delahaye, M. Dauchet, and O. Delgrange. A guaranteed compression scheme for repetitive dna sequences. *LIFL Lille I Univerisity technical report*, page 285, 1995.
6. F. M. J. Willems, Y. M.Shtralov, and T. J. Tjalkens. The context tree weighting method:basic properties. *IEE trans Inform Theory*, 41(3):653–664, 1995.
7. Kunihiro Sadakane, Takumi Okazaki, and Hiroshi Imai. Implementing the context tree weighting method for text compression. In *DCC '00: Proceedings of the Conference on Data Compression*, page 123, Washington, DC, USA, 2000. IEEE Computer Society.
8. E. Rivals and M. Dauchet. Fast discerning repeats in DNA sequences with a compression algorithm. In *Proc. Genome Informatics Workshop*, pages 215–226. Universal Academy Press, Tokyo, 1997.

9. Hisahiko Sata, Takashi Yoshioka, Akihiko Konagaya, and Tetsuro Toyoda. Dna compression in the post genomic era. *Genome Informatics*, 12:512–514, 2001.
10. J. Ziv and A. Lempel. Compression of individual sequences using variable-rate encoding. *IEEE trans Inform Theory*, 24:530–536, 1978.
11. J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE trans Inform Theory*, 23(3):337–343, 1977.
12. I. Sadel. Universal data compression algorithm based on approximate string matching. In *Probability in the Engineering and Informational Sciences*, pages 465–486, 1996.
13. Xin Chen, Sam Kwong, and Ming Li. A compression algorithm for dna sequences. *IEEE Engineering in Medicine and biology Magazine*, 20(4):61–66, 2001.
14. Li M, J. H Badger, J. H.Chen, S. Kwong, P. Kerney, and H. Zhang. An information based sequences distance and its application to whole mitochondrial genome. *Bioinformatics*, 17(2):149–154, 2001.
15. Xin Chen, M La, B Ma, and J Tromp. Dnacompress: fast and effective dna sequence compression. *Bioinformatics*, 18:1696–1698, 2002.
16. B Ma, J Tromp, and M. Li. Patternhunter-faster and more sensitive homology search. *Bioinformatics*, 18:440–445, 2002.
17. Chang C. Dnac: A compression algorithm of dna sequences by non-overlapping approximate repeats. *Master Thesis*, 2004.
18. Toshio Modegi. Development of lossless compression techniques for biology information and its application for bioinformatics database retrieval. *Genome Informatics*, (14):695–696, 2003.
19. Yong Zhang, Rahul Parthe, and Don Adjeroh. Lossless compression of dna microarray images. *csbw*, 0:128–132, 2005.
20. Zhenqiang Tan, Xia Cao, Beng Chin Ooi, and Anthony K. H. Tung. The ed-tree: An index for large dna sequence databases. *ssdbm*, 00:151, 2003.
21. Behshad Behzadi and Fabrice Le Fessant. Dna compression challenge revisited:a dynamic programming approach. In *CPM*, pages 190–200, 2005.
22. Alberto Apostolico and Stefano Lonardi. Compression of biological sequences by greedy off-line textual substitution. *dcc*, 00:143, 2000.

Application of Rough Sets Theory to the Sequential Diagnosis

Andrzej Zolnierrek

Wroclaw University of Technology, Faculty of Electronics, Chair of Systems and
Computer Networks, Wyb. Wyspińskiego 27, 50-370 Wroclaw, Poland
`andrzej.zolnierrek@pwr.wroc.pl`

Abstract. Sequential classification task is typical in medical diagnosis, when the investigations of the patient's state are repeated several times. Such situation takes place in controlling of the drug therapy efficacy. In this paper the methods of sequential classification using rough sets theory are developed and evaluated. The proposed algorithms, using the set of learning sequences, calculate the lower and upper approximations of the set of proper decision formulas and then use them to make final decision. Depending on the input data different algorithms are derived. Next, all presented algorithms were practically applied in computer-aided recognition of the human acid-base state balance and the results of comparative experimental analysis of in respect of classification accuracy are also presented and discussed.

1 Introduction

In many computer-aided medical diagnosis tasks there exist dependencies among the patterns to be recognized. Such a task, henceforth called the sequential classification (SC) task, involves dealing with a complex decision problem in which the sequences of patterns should be recognized. We deal with such a situation in the medical diagnosis when we have to control the course of therapy. In order to recognize the current patient's state, except for the current medical observations (*feature vector*) the medical doctor takes into account during his decision process the available medical information from the past. Such information can be of various origin, for example: patient's preceding states, previously observed feature vectors or applied therapies. It means that the patient's state at a given time depends on his so far observed history (*the context*) [15]. From the theoretical point of view, during construction of an appropriate decision algorithm we must not limit our approach only to the current feature vector but we have to consider all available measurements and applied therapies as input data instead, as they may contain important information about the recognized pattern at a given instant. In such a situation the amount of data is very large and grows over the time from one instant to another then performing of SC task various simplifications and compromises must be made. The dependence can be included in an early stage as formulating a mathematical model for the SC task, or in the later one as selecting the appropriate input data set in the decision algorithm

which otherwise does not differ from the classical recognition task. An example of the first case can be the probabilistic approach which offers the description of the dependencies in the form of a controlled Markov chain [17]. However, in this approach the assumption about existence of *prior* or *posterior* probabilities is made. Moreover, this approach requires a priori knowledge of probability characteristics of compound statistical process what seriously restrict its practical usefulness. Additionally, in order to obtain appropriate algorithm in this attempt to SC problem so-called naive Bayes assumption (the conditional independence of patterns for given the class) ([17], [16]) must be made. The second approach can be called *data-oriented*, while it uses methods developed in the field of computational intelligence such as fuzzy logic ([4], [5], [6], [11], [12]), rough sets ([3], [8], [9]) theory and genetic algorithms [7]. These methods, in particular based on the rough sets theory, are recently becoming increasingly popular in the pattern recognition applied to the problem of medical diagnosis (e.g. [1], [2], [9], [13],[14]) as an attractive alternative to statistical approach. They can perform classification from both labeled and unlabeled training sets as well as acquire and explore the human expert knowledge. They have been successfully applied in classical pattern recognition tasks, i.e. without taking into account the context, then application of such methodology to the problem of sequential classification is described in this paper. After preliminaries and problem statement, we present several algorithms of SC using rough sets methodology, which differ from each other using different kind of input data. All presented algorithms were practically applied to the problem of medical diagnosis (classification of states of acid-base balance) and results are presented and discussed at the end of this paper.

2 Preliminaries and the Problem Statement

We will treat the sequential classification (SC) task as a discrete controlled dynamical process. The pattern (patient's state) is at the n -th instant in the state $j_n \in \mathcal{M}$, where $\mathcal{M}$ is an m -element set of possible states numbered with the successive natural numbers, thus

$$j_n \in \mathcal{M} = \{1, 2, \dots, m\}. \quad (1)$$

Obviously, the notion of instant has no specific temporal meaning here, as its interpretation depends on the character of the medical case under consideration. The actual used measure may be minutes, hours, days or even weeks. The patient's state j_n is unknown and does not undergo our direct observation. What we can only observe are the symptoms by which a state manifests itself. We will denote an d -dimensional symptom vector measured at the n -th instant by $x_n \in \mathcal{X}$ (thus $\mathcal{X}$ is the observation space). While at same time the patient is treated let us also denote by u_{n-1} the therapy chosen from the discrete set of possible therapies $\mathcal{U}$ which was applied during the n -th instant. The set $\mathcal{U}$ is an r -element set of possible therapies numbered with the successive natural numbers, thus

$$u_{n-1} \in \mathcal{U} = \{1, 2, \dots, r\}. \quad (2)$$

As already mentioned, the patient's current state depends on the history and thus in the general case the decision algorithm must take into account the whole sequence of the preceding symptom vectors $\bar{x}_n = \{x_1, x_2, \dots, x_n\}$ and the sequence of applied therapies $\bar{u}_{n-1} = \{u_1, u_2, \dots, u_{n-1}\}$. It must be underlined here that sometimes it may be difficult to include all the available data, especially for bigger n . In such cases we have to allow some simplifications (e.g. take into account only several recent values in the sequence of symptom vectors $\bar{x}_n$ and applied treatment $\bar{u}_{n-1}$ or compromises (e.g. substituting the whole classification history segment that spreads as far back as the k -th instant with data processed in the form of a decision established at that instant, say i_k). In order to classify such sequences of patterns we need some more general information to take a valid decision, namely the *a priori* knowledge concerning the general associations that hold between patient's state on the one hand, and the sequence of symptom vectors and the sequence of applied therapies on the other. This knowledge may have multifarious forms and various origins. From now on we assume that it has the form of so called *training set*, which in the investigated decision task consists of N training sequences:

$$\mathcal{S} = \{S_1, S_2, \dots, S_N\}, \quad (3)$$

A single sequence:

$$S_k = ((x_{1,k}, j_{1,k}, u_{1,k}), (x_{2,k}, j_{2,k}, u_{2,k}), \dots, (x_{L,k}, j_{L,k})) \quad (4)$$

denotes a single controlled dynamic process (the course of disease) that comprises L feature symptom observations and $L - 1$ observed therapies as well as the consequent patient's states. Analysis of the SC task implies that, when considered in its most general form, the explored decision algorithm should use in the n -th instant the whole available observed data i.e. the sequences of input information as well as the knowledge included in the training set $\mathcal{S}$. In consequence, the algorithm is of the following form:

$$i_n = \Psi_n(\mathcal{S}, \bar{u}_{n-1}, \bar{x}_n) \quad (5)$$

The next section describes in depth the construction of the sequential classification algorithms (5) using various concepts based on the rough sets theory.

3 Algorithms of SC Based on Rough Sets Theory

In this section we will apply the rough sets theory [8], [9]) to the construction of SC algorithm (5). Now, the training set (3) is considered as an *information system* $Is = (Un, At)$, where Un and At , are finite sets called *universe* and the set of *attributes*, respectively. For every attribute $a \in At$ we determine its set of possible values V_a , called *domain* of a . Such information system can be represented as a table, in which every row represents a single sequence (4). In successive column of k -th row of this table we have values of the following attributes:

$$x_{1,k}^{(1)}, \dots, x_{1,k}^{(d)}, u_{1,k}, j_{1,k}, x_{2,k}^{(1)}, \dots, x_{2,k}^{(d)}, u_{2,k}, j_{2,k}, \dots, x_{L,k}^{(1)}, \dots, x_{L,k}^{(d)}, j_{L,k}. \quad (6)$$

In such an information system we can define in different way the subset $C \subseteq At$ of *condition attributes* and the single-element set $M \subseteq At$ which will be the *decision attribute*. Consequently, we obtain the *decision system* $Ds = (Un, C, M)$ in which, knowing the values of condition attributes, our task is to find the value of decision attribute, i.e. to find appropriate pattern recognition algorithm of sequential classification. As it was mentioned above, we can choose the subset of condition attributes in different way. Taking into account the set of condition attributes C , let us denote by X_j the subset of Un for which the decision attribute is equal to $j, j = 1, \dots, m$. Then, for every j we can defined respectively the *C-lower approximation* and the *C-upper approximation* of set X_j i.e.:

$$C_*(X_j) = \bigcup_{x \in Un} [C(x) : C(x) \subseteq X_j], \quad (7)$$

$$C^*(X_j) = \bigcup_{x \in Un} [C(x) : C(x) \cap X_j \neq \emptyset]. \quad (8)$$

Hence, the lower approximation of set X_j is the set of objects $x \in Un$, for which if we know values of condition attributes C , we can for sure say that they are belonging to the set X_j . Moreover, the upper approximation of set X_j is the set of objects $x \in Un$, for which if we know values of condition attributes C , we can not say for sure that they are not belonging to the set X_j . Consequently, we can define *C-boundary region* of as follows:

$$CN_B(X_j) = C^*(X_j) - C_*(X_j). \quad (9)$$

If for any j the boundary region of X_j is the empty set, i.e. $CN_B(X_j) = \emptyset$, then X_j is *crisp*, while in the opposite case, i.e. $CN_B(X_j) \neq \emptyset$ we deal with *rough set*. For every decision system we can formulate its equivalent description in the form of set of decision formulas $For(C)$. Each row of the decision table will be represented by single if-then formula, where on the left side of this implication we have logical product (*and*) of all expressions from C such that every attribute is equal to its value. On its right side we have expression that decision attribute is equal to the one number of class set (1). These formulas are necessary for constructing different pattern recognition algorithms for sequential classification.

3.1 Algorithm Without Context (R-0)

At first we start with the algorithm without the context which is well known in literature ([1], [2], [9]). In this case our decision table contains $N \times L$ patterns, each having d condition attributes (features) and one decision attribute (the class to which the pattern belongs), so the the algorithm is of the following form:

$$i_n = \Psi(S, x_n), n = 1, 2, \dots, i_n \in \mathcal{M}. \quad (10)$$

Application of rough set theory to the construction of classifier (10) from the learning set (3) can be presented according to the following items:

1. If the attributes are the real numbers then the discretization preprocessing is needed first. After this step, the value of each attribute is represented by the interval number in which this attribute is included. Of course for different attributes we can choose the different numbers of intervals in order to obtain their proper covering and let us denote for l -th attribute ($l = 1, \dots, d$) by ν_p^l its p_l -th value or interval. In this case each attribute is equivalent to corresponding symptom.
2. The next step consists in finding the set $For(C)$ of all decision formulas from (3), which have the following form:

$$IF(x^{(1)} = \nu_p^1)and(x^{(2)} = \nu_p^2)and...and(x^{(d)} = \nu_p^d)THEN\Psi(S, x) = j \quad (11)$$

Of course, it can happen that from the learning set (3) we obtain more than one rule for particular case. Then for such a formula (11) we determine its *strength* factor [3], which is the number of correct classified patterns during learning procedure. If any case in (3) is single then the strength factor of corresponding rule is equal to one.

3. For the set of formulas $For(C)$, for every $j = 1, \dots, m$ we calculate their C -lower approximation $C_*(X_j)$ and their boundary regions $CN_B(X_j)$.
4. In order to classify the n -th pattern x_n (after discretization its attributes if it is necessary) we look for matching rules in the set $For(C)$, i.e. we take into account such rules in which the left condition is fulfilled by the attributes of recognized pattern.
5. If there is only one matching rule, then we classify this pattern to the class which is indicated by its decision attribute j , because for sure such a rule belongs to the lower approximation of all rules indicating j , i.e. this rule is *certain*.
6. If there is more then one matching rule in the set $For(C)$, it means that the recognized pattern should be classified by the rules from the boundary regions $CN_B(X_j)$, $j = 1, \dots, m$ and in this case as a decision we take the index of boundary region for which the strength of corresponding rule is maximal. In such a case we take into account the rules which are *possible*.

3.2 Algorithm with k -th Order Context (R-k)

Although, we could take into account at n -th instant whole available information about the state of recognized sequential process i.e. $\bar{x}_n$, in order to simplify the form of SC algorithm we take into account, except the current symptom vector, only $k + 1$ recent observations of symptom vectors

$\bar{x}_n^{(k)} = (x_{n-k}, x_{n-k+1}, \dots, x_{n-1}, x_n)$ and k previously applied therapies

$\bar{u}_{n-1}^{(k)} = (u_{n-k}, u_{n-k+1}, \dots, u_{n-1})$. In such a situation we have the following decision attributes in our decision table:

$$x_{n-k}^{(1)}, \dots, x_{n-k}^{(d)}, u_{n-k}, x_{n-1}^{(1)}, \dots, x_{n-1}^{(d)}, u_{n-1}, \dots, x_n^{(1)}, \dots, x_n^{(d)}. \quad (12)$$

This means that algorithm includes k -instant-backwards-dependence ($k < L$) with full measurement data. Let us denote by D the total number of decision

attributes (in the algorithm **R-0** was $D = d$ and now $D = (k + 1) \times d + k$). Next, from the learning set (3), we can create the decision table which will have column $D + 1$ (the last one is the true classification of n -th recognized pattern) and consequently the number of rows will be equal to $N \times (L - k)$, because from each learning sequence (4) we can obtain $L - k$ subsequences of the length $k + 1$. The main idea of the proposed methods of SC is the same as for independent patterns but there are differences concerning details in procedure of construction of the set of decision formulas $For(C)$. If for simplicity we denote by the same letter $x^{(\alpha)}$, $\alpha = 1, \dots, D$ all condition attributes (i.e. features of symptom vectors $\bar{x}_n^{(k)}$ and successive therapies $\bar{u}_{n-1}^{(k)}$ then the decision formulas are of the following form:

$$IF(x^{(1)} = \nu_p^1) \text{ and } (x^{(2)} = \nu_p^2) \text{ and } \dots \text{ and } (x^{(D)} = \nu_p^D) \text{ THEN } \Psi(\mathcal{S}, \bar{x}_n^{(k)}) = j_n \quad (13)$$

The next steps of SC are the same as previously, i.e. we calculate $C_*(X_{j_n})$ and $CN_B(X_{j_n})$ and finally, the decision is made according to same procedure (steps 4, 5, 6).

3.3 Reduced Algorithm with k th Order Context(RR-K)

As it was mentioned in problem statement sometimes, we can accept the compromise in reduction of amount of input data, which consists in substituting them by data processed in the form of a decision established at that instant. As usual the main idea consists in adequate selection of decision attributes in the decision table which are now of the following form:

$$i_{n-k}, i_{n-k+1}, \dots, i_{n-1}, x_n. \quad (14)$$

For such decision attributes, we determine the set of decision formulas from the training set (3), as previously:

$$\begin{aligned} IF(j_{n-k} = \nu_p^l) \dots \text{ and } (j_{n-1} = \nu_p^k) \text{ and } (x_n^{(1)} = \nu_p^{k+1}) \dots \\ \text{ and } (x_n^{(d)} = \nu_p^{k+d}) \quad THEN \quad \Psi(\mathcal{S}, \bar{i}_n^{(k)}, x_n) = j_n \end{aligned} \quad (15)$$

where $D = k + d$ and $\nu_p^l, \dots, \nu_p^k \in \mathcal{M}$. Let us notice that like in fuzzy relation method during learning procedure, i.e. in finding formulas (15), we look for such situation in (3) where the left side condition of formulas (15) are fulfilled taking into account the true classifications. The next steps of SC are the same, but now we take into account the previous decisions $\bar{i}_n^{(k)} = (i_{n-k}, i_{n-k+1}, \dots, i_{n-1})$ as they were correct. All the decision algorithms that are depicted in the previous sections have been experimentally tested in respect of the decision quality (frequency of correct classifications) for real data that concern recognition of human acid-base equilibrium states (ABE). Results of experimental investigations are presented in the next section.

4 Practical Example: Sequential Diagnosis of Acid-Base Equilibrium States

4.1 Material

In the course of many pathological states, there occur anomalies in patient's organism as far as both hydrogen ion and carbon dioxide production and elimination are concerned, what leads to disorders in the acid-base equilibrium (ABE). Thus we can distinguish acidosis and alkalosis disorders here. Each of them can be of metabolic or respiratory origin, what leads to the following ABE state classification: metabolic acidosis, respiratory acidosis, metabolic alkalosis, respiratory alkalosis, correct state. In the process of treatment, correct recognition of these anomalies is indispensable, because the maintenance of the acid-base equilibrium, e.g. the pH stability of the fluids is the essential condition for correct organism functioning. Moreover, the correction of acid-base anomalies is indispensable for obtaining the desired treatment effects. In medical practice, only the gasometric examination results are made to establish fast diagnosis, although the symptom set needed for correct ABE estimation is quite large. The utilized results are: the pH of blood, the pressure of carbon dioxide, the current dioxide concentration. An anomalous acid-base equilibrium has a dynamic character and its changes depend on the previous state, and in consequence they require frequent examinations in order to estimate the current ABE state. It is clear now that the sequential decision methodology presented above suits well the needs of computer aided ABE diagnosing. The current formalization of the medical problem leads to the task of the ABE series recognition, in which the classification basis in the n -th moment constitutes the quality feature consisting of three gasometric examinations. And the set of diagnostic results M is represented by 5 mentioned acid-base equilibrium states. This model should be completed also with therapeutic possibilities (controls) which patient might undergo. Assuming certain simplification, these therapies can be divided into three following categories: respiratory treatment, pharmaceutical treatment, no treatment. The diagnostic algorithms applied to the ABE which state sequential diagnosis task have been worked out on the basis of evidence material that was collected in Neurosurgery Clinic of Medical Academy of Wroclaw and constitutes the set of training sequences (3). The material comprises 78 patients (78 sequences) with ABE disorders caused by intracranial pathological states for whom the gasometric examination results and the correct ABE state diacrisis were regularly put down on the 12-hour basis. There were around 20 examination cycles for each patient, yielding the total of 1416 single examination instances.

4.2 Results

In testing of algorithms based on rough sets theory, the cross validation method was used, i.e. for every trial ten testing sequences were chosen randomly and results are shown in Table 1.

Table 1. Frequency of correct diagnosis for various diagnostic algorithms

Trial	R-0	R-1	R-2	RR1	RR2
1	83.5%	87.7%	92.1%	85.2%	86.2%
2	85.1%	86.4%	89.7%	85.6%	85.9%
3	83.8%	88.1%	91.6%	85.1%	86.9%
4	84.0%	89.5%	92.8%	86.8%	87.3%
5	83.6%	87.6%	89.8%	85.9%	87.2%
6	82.1%	88.8%	89.9%	87.1%	89.1%
7	83.1%	87.1%	93.2%	86.9%	88.2%
8	85.9%	89.9%	91.5%	84.4%	89.2%
9	83.4%	88.9%	91.3%	87.2%	88.3%
10	86.9%	88.0%	91.9%	87.6%	89.0%
Best	86.9%	89.9%	93.2%	87.6%	89.2%
Mean	84.5%	88.2%	91.4%	86.2%	87.7%
SD	1.23%	3.37%	3.70%	1.03%	1.05%

These results imply the following conclusions:

1. Algorithm **R-0** that do not include the inter-state dependencies and treat the sequence of states as independent objects is worse than those that have been purposefully designed for the sequential medical diagnosis task, even for the least effective selection of input data.
2. There occurs a common effect that the model of the second order dependency (**R-2**) turns out to be more effective than the first order dependence approach (**R-1**).
3. Algorithms **R-1**, **R-2** that utilize the original data (i.e. gasometric examinations) always yield better results than those which substitute the data with diagnoses, i.e. **RR-1**, **RR-2**

5 Conclusions

The aim of this work was the application of soft computing methods to the sequential classification tasks in which there exist dependencies among the patterns to be recognized. In this paper new approach to SC task was considered, i.e. using rough sets theory. On the base of it several algorithms were proposed, which use the input data in different way, i.e. in different way take into account the dependencies in the sequence of recognized patterns. The empirical results show that taking into account such dependencies the accuracy of classification can be improved however, more empirical studies are required. This confirms the effectiveness and usefulness of the conceptions and algorithm construction principles presented above for the needs of sequential diagnosis. Moreover, analyzing results presented in the paper ([4]) for the same practical example, we can see no significant differences. It means that soft computing methods in SC

task basing on fuzzy sets theory or on rough sets theory can be considered as complementary.

References

1. Fang, J., Grzymala-Busse, J.: Leukemia Prediction from Gene Expression Data-a Rough Set Approach. In: L. Rutkowski, R. Tadeusiewicz, L. Zadeh, J. Zurada (Eds.), *Artificial Intelligence and Soft Computing*, Berlin, Heidelberg, New York, Springer Verlag (2006) 899-908
2. Fibak, J., Pawlak, Z., Slowinski, K., Slowinski, R.: Rough Set Based Decision Algorithm for Treatment of Duodenal Ulcer by HSV, *Bull. of the Polish Acad.Sci., Bio Sci.* Vol. 34 (1986) 227-246
3. Grzymala-Busse, J.: A System for Learning from Examples Based on Rough Sets. In: R. Slowinski (Ed.), *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, Dordrecht, Kluwer Academic Publishers. (1992) 3-18
4. Kurzynski, M., Zolnierek, A.: Computer-Aided Sequential Diagnosis Using Fuzzy Relations - Comparative Analysis of Methods. *Lecture Notes Computer Science., Lecture Notes Bioinformatics*, Vol. 3745 (2005) 242-251
5. Kurzynski, M., Zolnierek, A.: Sequential Pattern Recognition: Naive Bayes Versus Fuzzy Relation Method. In: M. Mohammadian (Ed.), *Proceedings International Conference on Computational Intelligence for Modelling, Control and Automation CIMCA 2005*, Los Alamitos, IEEE Computer Society Press (2005)
6. Kurzynski, M., Zolnierek, A.: Sequential Classification via Fuzzy Relations. In: L. Rutkowski, R. Tadeusiewicz, L. Zadeh, J. Zurada (Eds.), *Artificial Intelligence and Soft Computing*, Berlin, Heidelberg, New York, Springer Verlag (2006) 623-632
7. Michalewicz, Z.: *Genetic Algorithms + Data Structure = Evolution Programs*. Berlin, Heidelberg, New York, Springer Verlag (1996)
8. Pawlak, Z.: *Rough Sets-Theoretical Aspect of Reasoning About Data*. Dordrecht, Kluwer Academic Publishers (1991)
9. Pawlak, Z.: Rough Sets, Decision Algorithms and Bayes' Theorem. *European Journal of Operational Research*, Vol. 136 (2002) 181-189
10. Pawlak, Z., Slowinski, K., Slowinski, R.: Rough Classification of Patients after Highly Selective Vagotomy for Duodenal Ulcer. *Int. J. Man-Machine Studies*, Vol. 24 (1986) 413
11. Pedrycz, W.: Fuzzy Sets in Pattern Recognition: Methodology and Methods. *Pattern Recognition*, Vol. 23 (1990) 121-146
12. Pedrycz, W.: Genetic Algorithms for Learning in Fuzzy Relation Structures. *Fuzzy Sets and Systems*, Vol. 69 (1995) 37-45
13. Slowinski, K., Slowinski, R.: Sensitivity of Rough Classification to Changes in Norms of Attributes. In: R. Slowinski (Ed.) *Intelligent Decision Support, Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer Academic Publishers (1992) 363-373
14. Slowinski, K., Stefanowski, J., Siwinski, D.: Application of Rule Induction and Rough Sets to Verification of Magnetic Resonance Diagnosis. In: *Rough Set Data Analysis on Bio-medicine and Public Health*, Berlin, Heidelberg, New York, Springer Verlag (2001)
15. Toussaint G.: The Use of Context in Pattern Recognition, *Pattern Recognition*, Vol. 10 (1978) 189-204

16. Zolnierek, A.: The Empirical Study of the Naive Bayes Classifier in the Case of Markov Chain Recognition Task. In: M. Kurzynski, E. Puchala, M. Wozniak, A. Zolnierek (Eds.) Computer Recognition Systems CORES 05, Berlin, Heidelberg, New York, Springer Verlag (2005) 329-336
17. Zolnierek, A.: Pattern Recognition Algorithms for Controlled Markov Chains and their Application to Medical Diagnosis. Pattern Recognition Letters Vol. 1(1983) 299-303

Data Integration in Multi-dimensional Data Sets: Informational Asymmetry in the Valid Correlation of Subdivided Samples

Qing T. Zeng^{1,3}, Juan Pablo Pratt², Jane Pak^{1,2}, Eun-Young Kim^{1,4}, Dino Ravnic²,
Harold Huss², and Steven J. Mentzer²

¹ Decision Systems Group

² Department of Surgery, Brigham and Women's Hospital,
Harvard Medical School, Boston, MA

³ Harvard-MIT Division of Human Sciences and Technology, Cambridge, MA
qzeng@dsg.harvard.edu, {jppratt, jpak, drevnic, hhuss,
smentzer}@partners.org

⁴ Department of Clinical Pharmacology, Inje University Busan Paik Hospital
eykim@inje.ac.kr

Abstract. Background: Flow cytometry is the only currently available high throughput technology that can measure multiple physical and molecular characteristics of individual cells. It is common in flow cytometry to measure a relatively large number of characteristics or features by performing separate experiments on subdivided samples. Correlating data from multiple experiments using certain shared features (e.g. cell size) could provide useful information on the combination pattern of the not shared features. Such correlation, however, are not always reliable. Methods: We developed a method to assess the correlation reliability by estimating the percentage of cells that can be unambiguously correlated between two samples. This method was evaluated using 81 pairs of subdivided samples of microspheres (artificial cells) with known molecular characteristics. Results: Strong correlation ($R=0.85$) was found between the estimated and actual percentage of unambiguous correlation. Conclusion: The correlation reliability we developed can be used to support data integration of experiments on subdivided samples.

Keywords: correlation, data integration, subdivided sample, flow cytometry.

1 Introduction

Sampling infers information about an entire population by observing a fraction or subset of the overall population. To ensure that the sampling is representative, standard sampling theory relies on random selection of a sample from the population. For example, clinical and research studies will infer the state of a patient's hematologic system from the study of a relatively small sample of blood. Further, the original blood sample may be divided into subsamples with laboratory tests applied to each tube. In most cases, the information obtained from the analyses is relevant to the patient, but independent of the results in the other tubes.

There are special circumstances in which the information from one tube is not independent from the results of another tube. When the subsample is nonuniform with respect to the test being performed, it is possible that the results of the test will reflect properties of a subset or subpopulation within the tube. Tests performed on other tubes may reflect additional properties of the same subset or subpopulation. Linking the information from the separate tubes provides an opportunity for data integration and knowledge discovery. For example, flow cytometry analyses of a blood sample may identify molecular expression data specific to a subset of T lymphocytes [1]. By linking the information about these T lymphocyte populations between tubes, a more complete description of the molecular expression of T lymphocytes can be obtained.

The utility of data integration depends upon both the shared and uniquely measured properties of the subdivided samples. In this report, we describe a method for estimating the reliability of the correlation. When applied to a data set generated using artificial cells (microspheres) with known molecular characteristics, we found strong correlation ($R=0.85$) between the estimated and actual percentage of cells that could be unambiguously correlated. The result suggests that our reliability estimation method is a valid tool for the integration of multi-dimensional data obtained through multi-samples.

2 Background

Multidimensional data sets are common in both clinical patient care and biomedical research. A logical view of the derived multidimensional data is that it exists in a large Cartesian space bounded by all the dimensions of the data set. The data within this “data space” generally form clusters that can be defined by conventional clustering techniques. The informatics approaches to these data sets have focused on storage, retrieval and visualization of the data. For example, data warehouses are large repositories that integrate data from several sources for analysis. Online analytical processing (OLAP) systems provide rapid answers for queries by aggregating large amounts of data to discover trends. Data mining applications search data repositories for previously unknown patterns and relationships in the data set.

Unlike typical multidimensional data sets that derive from one sample, data sets from partitioned or subdivided samples have features that can only be realized through integration. The process of integration, commonly referred to as correlation, provides the opportunity to discover previously unknown properties of the original sample. Correlating multiple observations of the same population is routine when we have a unique identifier for each subject. In relational database operations, we often “join” two or more data tables using common fields [2, 3]. However, were such identifiers existent for each individual cell, they would not be helpful because different subsamples contain totally different cells.

In the absence of a unique identifier, correlation uses shared properties of subdivided samples to discover relationships among properties that are measured in only one of the samples. In the example of blood cells subdivided into numerous tubes, the shared properties of the T lymphocytes identified in all tubes can be used to correlate or integrate the T lymphocyte properties measured in individual tubes.

A useful example of multidimensional data sets containing both shared and uniquely measured properties can be found in the results of flow cytometry analyses. Flow cytometry is a widely used clinical and research tool that measures multiple parameters of individual cells. Flow cytometry is commonly used in such diverse biomedical application as the diagnosis of leukemia and the detection of gene expression in basic research. Despite the appealing multidimensional data obtained in flow cytometry, previous attempts to correlate cytometric data have been infrequent (U.S. Pat. No. 5,605,805). We suspect that data correlation in flow cytometry has been limited by significant variability in the validity of the correlated data. In some of our experiments, the correlated data provides striking insights into properties of blood cell populations. In other experiments, the results appear to be largely random without physical reality.

Our theory is that effective data integration depends upon the relationship between the shared and uniquely measured properties. In valid correlations, populations uniform with respect to both the shared properties and the uniquely measured properties are correlated. In contrast, invalid correlations occur with populations uniform with respect to the shared properties and yet nonuniform with respect to the uniquely measured properties. Because the shared properties cannot inform the process of correlating the nonuniform properties, the subsamples are randomly combined.

The associated schematic (Figure 1) illustrates these data relationships. In the example, experiment 1 measured properties x , y and z_1 of a subsample and experiment 2 measured properties x , y and z_2 of another subsample. There are 5 populations (A-E) in the sample. Prior to correlation, the combined z_1 and z_2 properties of any of the populations are unknown. Using the shared x and y properties, population C can be correlated from the two experiments. Population DE in experiment 1 can also be unambiguously linked with that of population D and E in experiment 2. Because subsamples are inherently similar, the size of DE will roughly equal the sum of D and E in experiment 2, allowing us to confidently predict two combinations of z_1 and z_2 properties. Populations A and B, however, can not be differentiated based on x and y . If A and B are correlated using just x and y , it would produce some non-existing combinations z_1 and z_2 properties.

To examine the populations in a sample requires clustering analysis without a priori knowledge of the number of clusters. Because the task of unsupervised clustering is difficult, no perfect solution exists and a number of proposed methods for cluster number determination could be found in literature [4-7]. After experimenting with several existing methods such as the Partition Index [8] with less than satisfactory results, we developed a histogram feature guided (FG) clustering algorithm that uses SiZer (a kernel smoothing method) [9] to extract histogram features and k-means [10] to partition the data into clusters. The FG algorithm was evaluated on a flow cytometry data set of microspheres and successfully identified all of the experimental populations.

Figure 1. A schematic demonstrating the results of three parameters (x , y and z) measured for each cell in tube #1 (X, Y , and Z_1) and tube #2 (X, Y , and Z_2). 3 properties are measured. In this illustration, there are 5 populations (A-E). Without correlation, we could not know the combined z_1 and z_2 properties of any of the populations. By correlating the populations based on the shared X and Y properties,

however, population C can be unambiguously correlated between the two tubes. Similarly, populations D and E in tube #1 can be unambiguously linked to D and E in tube #2. In contrast, populations A and B, can not be differentiated based on x and y . The correlation of A and B based on X and Y would produce non-existing combinations of z_1 and z_2 properties.

The clustering algorithm used in this paper is a modification of the previously published FG algorithm. The previous algorithm starts with a small number of clusters, attempts to match histogram peak locations with cluster centroid locations in each dimension, and gradually increases the cluster number until all peaks found a match. The revised algorithm also starts with a small number of clusters, but in place of matching, checks if the histograms of the clusters have multiple peaks (suggesting multiple populations), and gradually increases the cluster number until no cluster has multiple peaks in any of the dimensions. The modification eliminated the need to set a matching criterion between peak and centroid locations. When tested on the microsphere data, the performance of the revised algorithm was equal to that of the published algorithm.

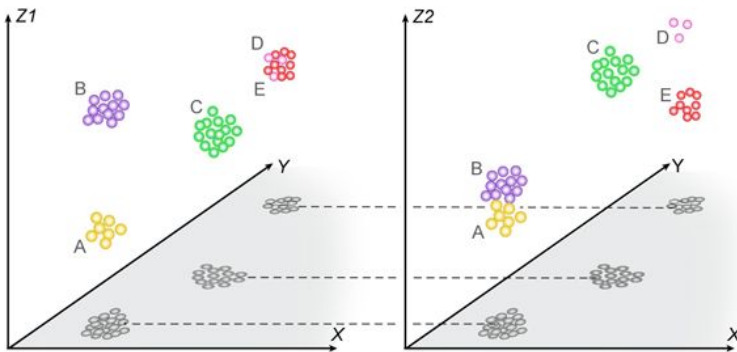


Fig. 1. A schematic demonstrating the results of three parameters (x , y and z) measured for each cell in tube #1 (X , Y , and Z_1) and tube #2 (X , Y , and Z_2)

3 Methods

3.1 Artificial Cells

The artificial cells were developed from commercially available microspheres (Bangs laboratories, Fishers, IN, USA; Polysciences, Inc, Warrington, PA, USA) formulated with carboxylic acid groups incorporated into the microsphere surface. To provide a defined protein surface, murine immunoglobulin (IgM; eBioscience, San Diego, CA, USA) was covalently linked to 6 μ m, 3 μ m, 2 μ m, 1 μ m and 0.75 μ m microspheres using a water soluble carbodiimide (1-ethyl-3-(3-dimethylaminopropyl) carbodiimide hydrochloride (EDAC)) reaction. Briefly, the microspheres were washed in an activation buffer (MES, pH 6) and reacted with EDAC (PolyLink; Bangs) and various concentrations of IgM for 30 minutes at 25°C. The reaction was

quenched with a 35uM glycine buffer. The microspheres were washed in a 1% (w/v) bovine serum albumin blocking buffer (pH 7) and stored at 4oC. To confirm the reaction stoichiometry, the density of covalently linked protein was confirmed by quantitative flow cytometry.

3.2 Determining Microsphere Number

Electronic counting of the microspheres was performed using a Coulter Z2 Particle Analyzer (Beckman Coulter, Miami, USA). The Coulter Z2, based on the Coulter principle [11], counted the microspheres by measuring changes in electrical resistance produced by the nonconductive microspheres suspended in a standard electrolyte solution (Isoton II; Beckman Coulter). A 100um aperture was used with constant voltage settings (gain 128, current 0.707, preamp gain 179.20). Cell minimum and maximum diameter settings were modified for the analysis of the various sized microspheres.

3.3 Detection Antibodies

The surface IgM was detected using a goat anti-mouse Ig fluorescein isothiocyanate (FITC) antibody (Southern Biotech, Birmingham, AL, USA), goat anti- mouse Ig phycoerythrin (PE) antibody (Southern Biotech) and a rat anti-mouse IgM PE-Cy5 conjugate (eBioscience, San Diego, CA, USA).

3.4 Flow Cytometry

Samples were analyzed using an Epics XL-MCL flow cytometer (Beckman Coulter, Miami, FL, USA). The artificial cells were analyzed on an Epics XL (Beckman Coulter; Maimi, FL, USA) equipped with a single laser with excitation wavelength at 488nm and three emission detectors. Gain settings were calibrated to 4 peak Rainbow calibration particles (Spherotech; Libertyville, IL, USA). During the experiments the laser power, photomultiplier tube voltage, light scatter and fluorescent gains were kept constant. A total of 10,000 events were acquired from each sample. The data was processed using WinList 5.0 (Verity; Topsham, ME, USA) and exported to Microsoft Excel (Redmond, WA, USA) for further analysis.

3.5 Cell Population Simulation

The artificial cells of known size, concentration and protein surface density—detected with one of three different fluorophores—provided 81 pairs (162 total) of subdivided samples to test the correlation reliability algorithm. Each sample contained 3 to 7 cell populations.

3.6 Data Analysis

We developed a two step approach to calculate the reliability of correlating two data sets: First, each set of data are clustered; Second, the clusters from each data set are compared to estimate the portion of the data that could be unambiguously correlated.

Assessment of the reliability of correlating n ($n > 2$) data sets can be conducted as ${}_nC_2$ pair-wise correlation assessments.

Take two data sets A and B , they are first partitioned respectively into multiple clusters $A_{1 \sim n}$ and $B_{1 \sim m}$. Each cluster can be represented as a unique combination of the cluster features (e.g. centroid location) in the multi-dimensional space: $\{A_{i,1} \dots A_{i,x}\}$. Assume A has g dimensions, B has h dimensions and they share k dimensions, cluster A_i may be represented as $\{A_{i,1} \dots A_{i,k}, A_{i,k+1} \dots A_{i,g}\}$ and cluster B_j may be represented as $\{B_{j,1} \dots B_{j,k}, B_{j,k+1} \dots B_{j,h}\}$.

A_i and B_j are matched for correlation if their shared dimensions features are similar: $\text{distance}(\{A_{i,1} \dots A_{i,k}\}, \{B_{j,1} \dots B_{j,k}\}) < t$, t is calculated based on the statistical characteristics of A_i and B_j . Between the clusters in A and B , there could be 1-to-0, 1-to-1, 1-to- n and n -to- n matches:

- The 1-to-0 match does not lead to any correlation – due to variance in subdivided samples, one may have small populations that do not exist in another.
- The 1-to-1 match produces a single combination of unique features $\{A_{i,k+1} \dots A_{i,g}, B_{j,k+1} \dots B_{j,h}\}$.
- The 1-to- n match produces n combinations of unique features $\{A_{i,k+1} \dots A_{i,g}, B_{j,k+1} \dots B_{j,h}\}$. We can be certain of the actual existence of these feature combinations because the number of cells with the same shared features are roughly equal in subdivided samples and each cluster from clusters n will be matched with part of the “1” cluster.
- The n -to- n match produces possible feature combinations that may or may not actually exist: for example, when equal-sized clusters A_1 and A_2 match B_1 and B_2 , it is not certain that a cluster of cells with the feature combination $\{A_{1,k+1} \dots A_{1,g}, B_{2,k+1} \dots B_{2,h}\}$ exist, for A_1 and B_1 instead of A_1 and B_2 might be the correct match.

We consider only cells involved in 1-to-1 and 1-to- n matches to be unambiguously correlated; the percentage of such cells are used as a reliability measure for correlating A and B using k shared dimensions.

Detailed steps of the algorithm are as follows:

3.6.1 Clustering

1. Obtain smoothed histogram h_i from each dimension i of a multidimensional data set A .
2. Identify peak locations x_q ($h_i'(x_q)=0$ and $h_i(x_q)>h_i(x_q-1)$) in the smoothed histograms and count the total number of peaks (i.e. number of x_q) pn_i in each histogram.
3. Set the initial cluster number n to be the maximum number of peaks in a dimension: $n = \max(pn_i)$.
4. Use the k-means algorithm to partition A into n clusters.
5. For each dimension i of each cluster j , calculate the number of peaks pn_{ji} as in step 2. If any $pn_{ji} > 1$, c_j is considered nonuniform. Otherwise, c_j is considered uniform.
6. Add uniform clusters to a set F .
7. Let k be the number of nonuniform clusters. If $k=0$, terminate the clustering process and return F ; otherwise, continue to step 8.
8. Merge all nonuniform clusters into the new data set A . The new cluster number $n = k+1$. Repeat Steps 4-7.

3.6.2 Correlation Ambiguity Checking

To correlate two sample data sets A and B , we compare their component clusters S_a and S_b :

1. For each cluster C_i in S_a and S_b , calculate the median Euclidean distance D_i between the data points in C_i and its centroid CT_i . Please note that in this and following steps, distance calculation only involves the shared dimensions.
2. Match clusters between S_a and S_b . Let C_{ai} be a cluster in S_a and C_{bj} be a cluster in S_b ; C_{ai} is considered to match C_{bj} if the Euclidean distance D_{ai-bj} of their centroid locations is smaller than one of their cluster's median Euclidean distance: $D_{ai-bj} < \max(D_{ai}, D_{bj})$.
3. Count the number of cells that cannot be unambiguously correlated (ncc) in each match between S_a and S_b .
 - **1 to 0** match (i.e. a C_{ai} matches to no C_{bj}). no correlation: $ncc = |C_{ai}|$
 - **1 to 1** match (i.e. a C_{ai} matches to one C_{bj} and vice versa). Cells in C_{ai} can be unambiguously correlated with C_{bj} except the extra cells C_{ai} contains if any: $ncc = \max(|C_{ai}| - |C_{bj}|, 0)$
 - **1 to n** match (i.e. a C_{ai} matches to multiple C_{bj}). Cells in C_{ai} can be unambiguously correlated with the multiple C_{bj} except the extra cells C_{ai} contains if any: $ncc = \max(|C_{ai}| - \sum |C_{bj}|, 0)$
 - **n to 1** match (i.e. multiple C_{ai} matches to one C_{bj}). Cells in C_{ai} can be unambiguously correlated with C_{bj} except the extra cells C_{ai} contains if any: $ncc = \max(\sum |C_{ai}| - |C_{bj}|, 0)$
 - **n to n** match (i.e. multiple C_{ai} matches to multiple C_{bj}). No cells matching C_{ai} can be unambiguously correlated: $ncc = \sum |C_{ai}|$
4. Calculate the total percentage of data points that cannot be unambiguously correlated (pcm): $pcm = \frac{\sum ncc}{|S_a|}$

In the study, we used SiZer to smooth the histograms. In the default setting, SiZer smoothes data into 11 even levels, we empirically chose the mid-range smoothing level 5 in this study. In the clustering analysis, very small peaks and clusters (defined as size smaller than 5% of the total data points) were presumed to be noise biologically, though they may not statistically be considered as such. The very small peaks were ignored and the very small clusters were merged with its closest neighbor based on the Euclidean distance between centroids.

4 Results

The use of artificial cells (microspheres) allowed for the experimental manipulation of the shared and uniquely measured properties. In figures 2A and 2B, shared properties are illustrated by FSC and SSC; likewise, unique properties are shown by FL1 and FL2.

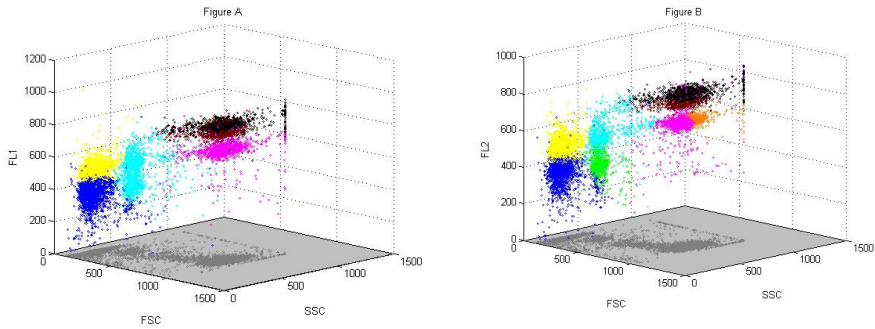


Fig. 2. 3D scatter plot of microsphere subsample A and B. FSC and SSC are shared properties and FL1 and FL2 properties unique to A and B respectively. There are 7 actual populations in samples. The clusters identified by our algorithm are highlighted in different colors. Correlating A and B clusters based on the share properties FSC and SSC results in a couple of n-n matches.

Based on the properties of the artificial cells, the percentage of unambiguous correlation was estimated using our algorithm for 81 pairs of subdivided samples. Each pair had 2 to 3 shared parameters, 2 to 3 unique parameters and contained 3 to 7 actual populations. The actual percentage of cells that could be unambiguously correlated between the subdivided samples was deduced from the known co-expression characteristics of cells populations in the samples.

The estimated and actual percentages were compared (Figure 3) and strong correlation was found ($R=0.85$). This suggests the estimation to be fairly accurate. A closer examination of the difference between estimate and actual found that the mains causes

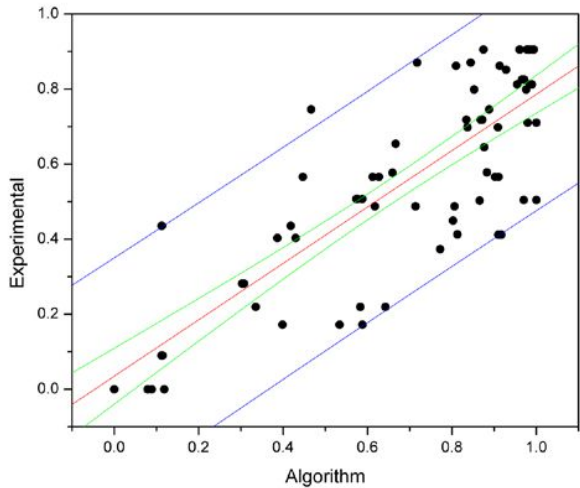


Fig. 3. Comparison of number of cells that could be reliably correlated experimentally or using the algorithm. Linear regression is shown in the red lines. The confidence limits (green lines) and prediction bands (blue lines) are also shown.

were 1) the computer algorithm treated noise as cells and identified some noise clusters; 2) in several samples, some microspheres attached to other microspheres and formed unexpected clusters; 3) the smoothing level was too high for a few samples.

5 Discussion

We have developed a method to support the integration of data from multiple flow cytometry experiments by assessing the reliability of correlating the data using common/shared features. To our knowledge, the only prior claim to correlate data from multiple flow cytometry experiments is a patent by Becton, Dickinson and Company (BD). The patent described an automated diagnostic method for acute leukemia (U.S. Pat. No. 5,605,805) that involved correlating data and differentiating normal and abnormal subpopulations. The need to measure co-expression of multiple cell surface molecules, however, has been demonstrated by the substantial research effort on increasing the number of simultaneously measured fluorochromes [1].

While standard cytometers measure three distinct fluorochromes, prototype flow cytometers can now measure eight or more [1]. This approach has the advantage of unequivocally measuring co-expressions, though it also leads to problems associated with spectral overlap of the multiple fluorochromes (so-called “compensation” problems) [12]. Regardless how many parameters are measured physically, computational correlation of data sets could analyze parameters that exceed this number.

Since correlated data do not come directly from observation, it is important for us to investigate if the correlated patterns of features reflect actual co-expressions. Our method carries out this task by analyzing if clusters of cells in one sample can be unambiguously matched to clusters in another sample according to certain shared parameter features; such matches allow us to predict the feature combination with 100% certainty. In evaluation, the method worked well: the predicted and actual percentage of unambiguously correlatable cells showed strong correlation ($R = 0.85$).

We view the percentage of unambiguous match 1-pcm as a reliability measure: only when 1-pcm is equal or close to 100%, all feature combinations in the correlated data set could be considered actual. Depending on the number and size of the clusters involved in the ambiguous ($n\text{-}n$) matches, sometimes it is possible to infer the actual existence of feature combinations: assume clusters A_1 and A_2 match B_1 and B_2 , and A_1 and B_1 each contain 1000 cells while A_2 and B_2 each contain 100 cells, then at least 900 cells in A_1 and B_1 must correlated. On the other hand, if A_1 , A_2 , B_1 and B_2 are equal sized, we would not know if any cells in A_1 and B_1 should be correlated. One limitation of our algorithm is that it treats all correlations among ambiguously matched clusters as unreliable.

Another limitation of the algorithm is the need to select a smoothing level (bandwidth). Because a data-driven approach to the determination of optimal smoothing level is still a research topic, currently we have to empirically choose a level based on the noise level and desired sensitivity. The kernel smoothing method SiZer does offer a bandwidth-free solution by providing a whole family of histograms with least to most amount of smoothing. Nevertheless, we guide the clustering process with peak features, and could not identify peaks without setting some kind of threshold or smoothing level.

This study is conducted using idealized data, though artificial cell samples and correlation tasks were designed to simulate a range of difficulties. We plan to apply and validate the method to lymphocyte samples in the future.

References

1. Shapiro HM. Practical Flow Cytometry. 4 ed. New York: Wiley-Liss Inc.; 2002.
2. Saeed M, Mark RG. Efficient hemodynamic event detection utilizing relational databases and wavelet analysis. *Comput Cardiol* 2001;28:153-6.
3. Chu SC, Thom JB. Database issues in object-oriented clinical information systems design. *Stud Health Technol Inform* 1997;46:376-82.
4. Montgomery EB, Jr., Huang H, Assadi A. Unsupervised clustering algorithm for N-dimensional data. *J Neurosci Methods* 2005;144(1):19-24.
5. Ben-Hur A, Elisseeff A, Guyon I. A stability based method for discovering structure in clustered data. *Pac Symp Biocomput* 2002:6-17.
6. Dudoit S, Fridlyand J. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biol* 2002;3(7):RESEARCH0036.
7. Lange T, Roth V, Braun ML, Buhmann JM. Stability-based validation of clustering solutions. *Neural Comput* 2004;16(6):1299-323.
8. Lee S, Crawford MM. Unsupervised multistage image classification using hierarchical clustering with a Bayesian similarity measure. *IEEE Trans Image Process* 2005;14(3):312-20.
9. Chaudhuri P, Marron JS. SiZer for exploration of structures in curves. *Journal of the American Statistical Association* 1999;94:807-823.
10. O.Duda R, Hart PE, Stork DG. Pattern Classification. second ed: John Wiley & Sons, Inc; 2001.
11. Brecher G, M. MS, Williams GZ. Evaluation of electronic red blood cell counter. *Am.J.Clin.Pathol.* 1956;26:1439-1449.
12. Young IT. Proof without prejudice: use of the Kolmogorov-Smirnov test for the analysis of histograms from flow systems and other sources. *J Histochem Cytochem* 1977;25(7):935-41.

Two-Stage Classifier for Diagnosis of Hypertension Type

Michał Wozniak

Chair of Systems and Computer Networks, Wrocław University of Technology
Wybrzeże Wyspiańskiego 27, 50-370 Wrocław, Poland
michal.wozniak@pwr.wroc.pl

Abstract. The inductive learning approach could be immensely useful as the method generating effective classifiers. This paper presents idea of constructing two-stage classifier for diagnosis of the type of hypertension (essential hypertension and five type of secondary one: fibroplastic renal artery stenosis, atheromatous renal artery stenosis, Conn's syndrome, renal cystic disease and pheochromocytoma). The first step decides if patient suffers from essential hypertension or secondary one. This decision is made on the base on the decision of classifier obtained by boosted version of additive tree algorithm. The second step of classification decides which type of secondary hypertension patient is suffering from. The second step of classifier makes its own decision using human expert rules. The decisions of these classifiers are made only on base on blood pressure, general information and basis biochemical data.

1 Introduction

Machine learning [10] is the attractive approach for building decision support systems because in many cases we can find following problem:

- the experts can not formulate the rules for decision problem, because they might not have the knowledge needed to develop effective algorithms (e.g. human face recognition from images),
- we want to discover the rules in the large databases (data mining) e.g. to analyze outcomes of medical treatments from patient databases; this situation is typical for designing telemedical decision support system, which knowledge base is generated on the base on the large number of hospital databases,
- decision support system has to adapt dynamically to changing conditions.

These situations are typical for the medical knowledge acquisition also. For many cases the physician can not formulate the rules, which are used to make decision or set of rules given by expert is incomplete.

In the paper we present continuation of our work on the effective classifier for recognition of the type of hypertension problem. In [18] we shown idea of two-stage classifier for the problem under consideration. We concentrate our attention on classifier selection for the two stages of proposed recognizer. We present the results of some known algorithms like decision tree induction algorithms and we discuss if boosting methods can improve their qualities for the medical problem under

consideration. Finally we obtain two-stage classifier where the decision about the nature of hypertension (if it is essential type of secondary one) is made on the base of classifier obtained via machine learning procedure. For the diagnosis of the type of secondary hypertension (5 classes) we use human expert rules.

The content of the work is as follows. Section 2 introduces idea of the inductive decision tree algorithms and concept of boosting. In Section 3 we describe mathematical model of the hypertension's type diagnosis. Next section presents results of the experimental investigations of the algorithms. Section 5 concludes the paper.

2 Algorithms

2.1 Decision Tree Induction

The most of algorithm as C4.5 given by R. J. Quinlan [12] or ADTree (Alternative Decision Tree) [6] are based on the idea of "Top Down Induction of Decision Tree". The central choice in the TDIDT algorithm is selecting "the best" attribute (which attribute to test at each node in the tree). Family of algorithm based on ID3 method [10] (e.g. C4.5) uses the information gain that measures how well the given attribute separates the training examples according to the target classification. The future implementations of decision tree induction algorithm use measure based on previously defined information gain (e.g. information ratio [12]).

2.2 Boosting

Boosting is general method of producing an accurate classifier on base of weak and unstable one [15-16]. It is often called metaclassifier. The idea of boosting has its root in PAC (*Probably Approximately Correct*) theory. The underlying idea of boosting is to combine simple classifiers to form an ensemble such that the performance of the single member of ensemble is improved [8, 14]. As we see the main problem of the boosting is how to construct ensemble. The one of the most popular algorithm AdaBoost [4, 5] produces at every stage, a classifier which is trained with the modified learning set. The output of the classifier is then added to the output of classifier ensemble, with the strength proportional to how accurate obtained classifier is. Then, the elements of learning set are reweighted: examples that the current learned function gets wrong are "boosted" in importance, so that the classifier obtained at the next stage will attempt to fix the errors. The main advantage of boosting is that it often does not suffer from overfitting.

3 Model of Type of Hypertension (HT) Diagnosis

During the hypertension's therapy is very important to recognize state of patient and the correct treatment. The physician is responsible for deciding if the hypertension is of an essential or a secondary type (so called the first level diagnosis). The senior physicians from the Broussais Hospital of Hypertension Clinic and Wroclaw Medical Academy suggest 30% as an acceptable error rate for the first level diagnosis. The

presented project was developed together with Service d'Informatique Médicale from the University Paris VI. All data was getting from the medical database *ARTEMIS*, which contains the data of the patients with hypertension, whose have been treated in Hôpital Broussais in Paris.

The mathematical model was simplified. Hover our experts from the Broussais Hospital, Wroclaw Medical Academy, regarded that stated problem of diagnosis as very useful. It leads to the following classification of type of hypertension:

1. essential hypertension (abbreviation: essential),
2. fibroplastic renal artery stenosis (abbreviation: fibro),
3. atheromatous renal artery stenosis (abbreviation: athero),
4. Conn's syndrome (abbreviation: conn),
5. renal cystic disease (abbreviation: poly),
6. pheochromocytoma (abbreviation: pheo).

Although the set of symptoms necessary to correctly assess the existing HT is pretty wide, in practice for the diagnosis, results of 18 examinations (which came from general information about patient, blood pressure measurements and basis biochemical data) are used, whose are presented in table 1.

Table 1. Clinical features considered

No	Feature	No	Feature
1	sex	10	effusion
2	body weight	11	artery stenosis
3	high	12	heart failure
4	cigarette smoker	13	palpitation
5	limb ache	14	carotid or lumbar murmur
6	alcohol	15	serum creatinine
7	systolic blood pressure	16	serum potassium
8	diastolic blood pressure	17	serum sodium
9	maximal systolic blood pressure	18	uric acid

4 Experimental Investigation

All learning examples were getting from medical database *ARTEMIS*, which contains the data of 1425 patients with hypertension (912 with essential hypertension and the rest of them with secondary ones), whose have been treated in Hôpital Broussais.

We used WEKA systems [17] and our own software. Quality of correct classification was estimated using 10 folds cross-validation tests.

4.1 Experiment A

The main goal of experiment was to find quality of recognition the C4.5 algorithm and its boosted form. The obtained decision tree is shown in Fig.1. Its frequency of correct classification is 67,79% and its confusion matrix looks as follow

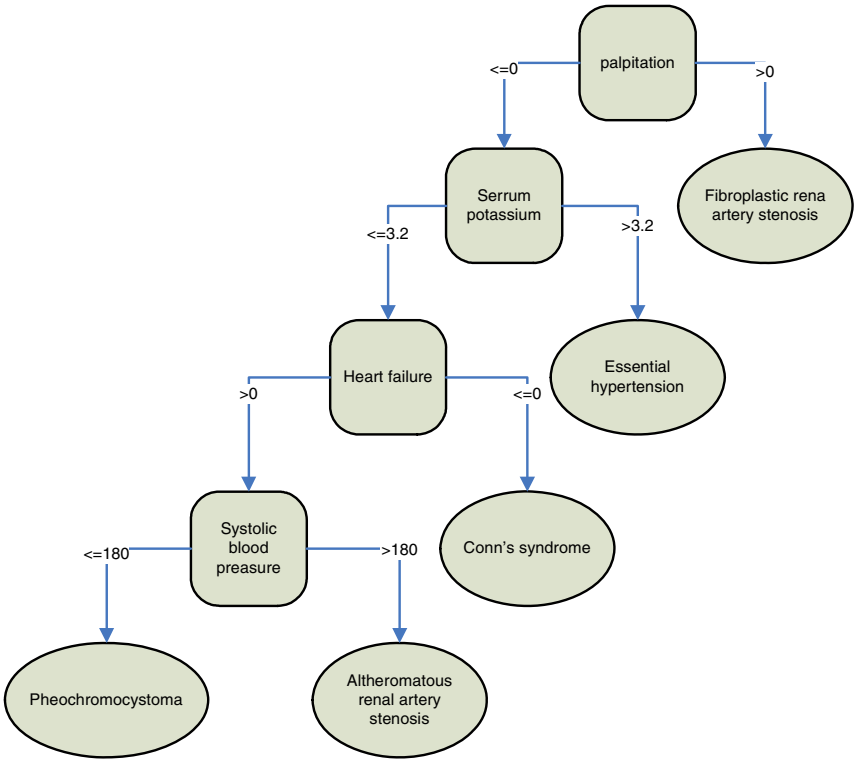


Fig. 1. Decision tree for hypertension diagnosis given by C4.5 algorithm

Table 2. Confusion matrix for decision tree

Real diagnosis						Recognized class
athero	conn	essent	fibro	pheo	poly	
4	3	54	16	0	0	athero
0	44	92	11	0	0	conn
2	25	878	7	0	0	essent
4	5	64	40	0	0	fibro
2	3	80	2	0	0	pheo
0	2	81	6	0	0	poly

We rejected the classifier because its quality did not satisfy expert. But we have to note that advantage of this tree is that the essential hypertension was recognized pretty good (96,26%).

We tried to improve quality of obtained classifier using boosting concept. Unfortunately new classifier had worse quality than original one (59,30%). The confusion matrix of the boosted C4.5 is presented in Tab.2

Let us note that essential hypertension is represented by 912 object. It means that if any classifier always points at this class that its quality is about 64%. This result is only slightly worse than the quality of the best classifier in experiment A (67,79%).

Table 3. Confusion matrix for boosted decision tree

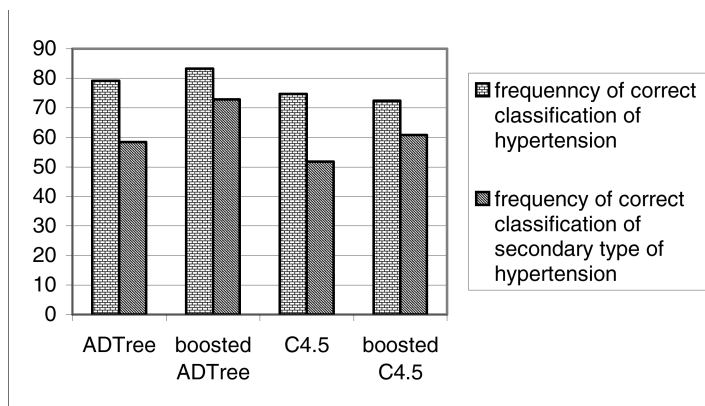
Real diagnosis						Recognized class
athero	conn	essent	fibro	pheo	poly	
8	4	52	7	2	4	athero
2	22	106	10	4	3	conn
19	47	790	34	12	10	essent
3	6	86	9	8	1	fibro
1	3	73	5	4	1	pheo
1	4	70	0	2	12	poly

4.2 Experiment B

Physician-experts did not accept classifiers obtained in Experiment A. After the discussion we simplified the problem once more. We were trying to construct classifiers which would point at essential type of hypertension or secondary one. We used two methods to obtain the classifiers:

1. Alternative Decision tree (ADTree),
2. C4.5 algorithm.

For each classifier we check its boosted form also. The results of tests are shown in Fig.2.

**Fig. 2.** Quality of recognition the essential and secondary type of hypertension

As we see the frequency of correct classification of ADTree algorithm is 79,16%. Unfortunately the quality of recognition the secondary hypertension is only 58,48%. We tried improve the quality of classifier by AdaBoost.M1 procedure and we obtained new classifier based on ADTree concept (we use 10 iterations), which frequency of correct classification grew to 83,30% and 72,90% of correct classified secondary type of hypertension. This results satisfied our experts.

Additionally we check the quality of C4.5 for the same dichotomy problem. We obtained the decision tree similar to tree in Fig.1 which quality is 74,74% and

frequency of correct recognized secondary type of hypertension is 51,85%. The boosting procedure did not improve the average quality of C4.5 (72,42%) but strongly improved the recognition secondary type of hypertension (60,81%).

4.3 Experiment C

Most of obtained classifiers (especially based on C4.5 method) did not satisfy experts. The best classifier (obtained for simplified decision problem) was accepted by our expert. Now we want to construct classifier ensemble on the based on the two-stage classifier concept [2, 11] which idea is depicted in Fig.3. Obtained boosted ADTree classifier can be use for the first stage of recognition. We try to find any algorithm which may be used on the second stage. Lets note that the probability of correct decision for two-stage classifier presented in the Fig.3 is given by the following formula

$$P_{correct} = P_c(essential\_HT) * P(essential\_HT) + P_c(sec\ ondary\_HT) * P(sec\ ondary) * P_c(sec\ ond\_stage) \quad (1)$$

where

P_c is the probability of correct decision of the two stage classifier,

$P_c(essential\_HT)$ is the probability of correct classification of essential hypertension (on the first stage),

$P(essential\_HT)$ is prior probability of essential hypertension,

$P_c(sec\ ondary\_HT)$ is the probability of correct classification of secondary hypertension (on the first stage),

$P(sec\ ondary\_HT)$ is prior probability of secondary hypertension,

$P_c(sec\ ond\_stage)$ is the probability of correct classification of secondary hypertension (on the second stage).

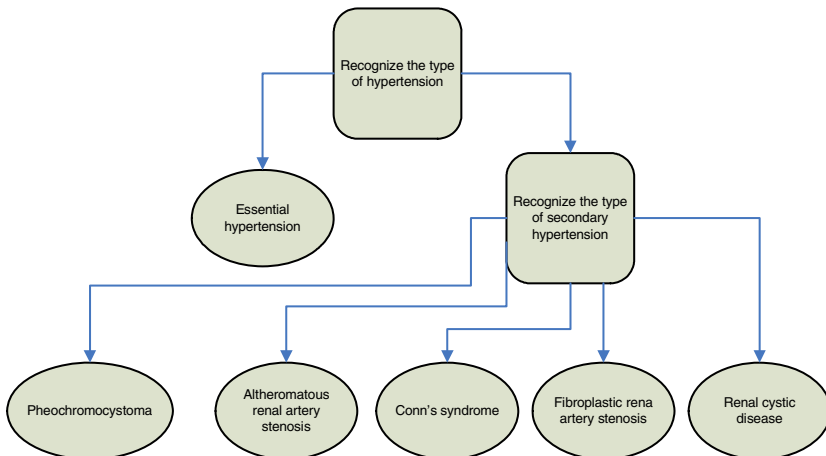


Fig. 3. Two-stage classifier of hypertension's type

Lets note that quality of the best one stage classifier is 67,79%. The quality of classifier of the first step is 83,3% for essential hypertension recognition and 72,9% for the secondary one. The estimations of the prior probabilities for essential and secondary type of hypertension are 64% and 36%. Therefore the two-stage classifier gives better results of recognition than one stage one if the quality of second-stage classifier is better than 55,17%.

We were testing group of classifier but their quality of recognition for the second stage is about 50%. The list of some tested classifiers and their frequencies of correct classification are shown in Tab.4.

Table 4. Frequencies of correct classification for the second stage of recognition

Classifier	Frequency of correct classification
DECORATE with C4.5[9]	31,77%
Naïve Bayes[7]	54,485
C4.5	29,43%
RIPPER[3]	42,10%

As we see no one may be use for decision making in the second stage. We asked human experts for the set of rules for this problem and we obtained 17 logical rules. The quality of correct classification for the second step achieved 73,07% and finally quality of two-stage classifier about 72,50%. Its quality is quite better than qualities of the one-stage classifier (67,79%).

5 Discussion and Conclusions

The methods of inductive learning were presented. The classifiers generated by these algorithms were applied to the medical decision problem (recognition of the type of hypertension). For the real decision problem we have to compare many classifiers and their *boosted* version. The general conclusion is that *boosting* does not improve each classifier for each decision task. The similar observations were described by Quinlan in [13] where he did not observe quality improvements of boosted C4.5 for some of databases.

Additionally we describe the concept of two-stage classifier for the diagnosis of hypertension type task which uses method based on decision tree concept on the first stage and human expert rules on the second one. The quality of the above method is definitely better than any methods testes as the one-stage recognition algorithm.

The similar problem of computer-aided diagnosis of hypertension's type was described in [1] but authors used another mathematical model and implement Bayes decision rule. They obtained slightly better classifier than our, its frequency of correct classification of secondary type of hypertension is about 85% (our 83,30%). Advantage of our proposition is simplified and cheaper model than presented in [1] (we use 18 features, authors of mentioned work use 28 ones).

Advantages of the proposed methods make it attractive for a wide range of applications in medicine, which might significantly improve the quality of the care that the clinicians can give to their patients.

Acknowledgement

This work is supported by The Polish State Committee for Scientific Research under the grant which is realizing in years 2006-2009.

References

1. Blinowska A. et al., Bayesian Statistics as Applied to Hypertension Diagnosis, *IEEE Trans. on Biomed. Eng.*, vol. 38, no. 7, July 1991, pp. 699-706.
2. Burduk R., Case of Fuzzy Loss Function in Multistage Recognition Algorithm, *Journal of Medical Informatics & Technologies*, vol. 5, 2003, pp. MI 107-112.
3. Cohen W.W., Fast Effective Rule Induction, *Proc. of the 12th International Conference on Machine Learning*, Tahoe City, pp.115-123.
4. Freund Y., Schapire R.E., A decision-theoretic generalization of on-line learning and application to boosting, *Journal of Computer and System Science*, 55(1), 1997, pp. 119-139.
5. Freund Y., Schapire R.E., Experiments with a New Boosting Algorithm, *Proceedings of the International Conference on Machine Learning*, 1996, pp. 148-156.
6. Hastie T., Tibshirani R., Friedman J., *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer Series in Statistics, Springer Verlag, New York 2001.
7. Jain A.K., Duin P.W., Mao J., Statistical Pattern Recognition: A Review, *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol 22., No. 1, January 2000, pp. 4-37.
8. Meir R. and Rätsch G. An introduction to boosting and leveraging, *Lecture Notes in Artificial Intelligence*, vol. 2600, Springer, 2003, pp. 119-184.
9. Melville P., Mooney R., Constructing diverse classifier ensembles using artificial training examples, *Proc. of 18th Intl. Joint Conf. on Artificial Intelligence*, Acapulco, Mexico, August 2003, pp. 505-510.
10. Mitchell T., *Machine Learning*, McGraw Hill, 1997.
11. Opitz D., Maclin R., Popular Ensemble Methods: An Empirical Study, *Journal of Artificial Intelligence Research*, 11 (1999), pp. 169-198
12. Quinlan J.R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
13. Quinlan J.R., Bagging, Boosting, and C4.5, *Proc. AAAI 96 and IAAI 96 conferences*, vol. 1, Portland, Oregon, August 4-8, 1996, pp. 725-230.
14. Schapire R. E., The boosting approach to machine learning: An overview. *Proc. Of MSRI Workshop on Nonlinear Estimation and Classification*, Berkeley, CA, 2001.
15. Shapire R.E., The Strength of Weak Learnability, *Machine Learning*, no. 5, 1990, pp. 197-227.
16. Schapire R.E., A Brief Introduction to Boosting, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999.
17. Witten I.H., Frank E., *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann Pub., 2000.
18. Wozniak M., Boosted decision trees for diagnosis type of hypertension, *Lecture Notes in Bioinformatics*, vol. 3745, Springer, 2005, pp. 223-230.

Handwriting Analysis for Diagnosis and Prognosis of Parkinson's Disease

Atilla Ünlü¹, Rüdiger Brause¹, and Karsten Krakow²

¹ Institute of Informatics, Johann Wolfgang Goethe-Universität,
60054 Frankfurt a.M., Germany
{Atilla,RBrause}@informatik.uni-frankfurt.de

² Neurological Clinics, Johann Wolfgang Goethe-Universität,
Theodor-Stern-Kai 7, 60596 Frankfurt a. M.
k.krakow@em.uni-frankfurt.de

Abstract. At present, there are no quantitative, objective methods for diagnosing the Parkinson disease. Existing methods of quantitative analysis by myograms suffer by inaccuracy and patient strain; electronic tablet analysis is limited to the visible drawing, not including the writing forces and hand movements. In our paper we show how handwriting analysis can be obtained by a new electronic pen and new features of the recorded signals. This gives good results for diagnostics.

Keywords: Parkinson diagnosis, electronic pen, automatic handwriting analysis.

1 Introduction

The Parkinson disease is a degenerative disorder of the central nervous system that affects the control of muscles, and so may affect movement, speech and posture [1]. It is often characterized by muscle rigidity, tremor, a slowing of physical movement (*bradykinesia*), and in extreme cases, a loss of physical movement (*akinesia*). The primary symptoms are the results of irregular muscle contraction, caused by the insufficient formation and action of dopamine, which is produced in the dopaminergic neurons (*substantia nigra*) of the brain.

From the pathological view of the Parkinson disease there is no reliable method for an objective, quantitative diagnosis. Clinically, one distinguishes the tremor between type I Parkinson tremor of 4-7 Hz, which can be observed when the hand is not moving, but shaking, and a higher frequency type II Parkinson tremor of 5-12 Hz which is observed during movement [2].

There are two methods known for quantitative analysis of tremors. Classically, the muscle activity is recorded from electrodes (skin surface electrodes, or steel needles pinned into the muscle through the skin), and printed in the form of myograms EMG [3]. This is not only a tedious procedure for the patient, but also not very accurate and therefore not used in the normal case. The second method has been pioneered by Marquard and Mai [4] and had been widely accepted by research [5],[6]. The procedure records the handwriting of the patient on a graphical tablet, especially loops like . These loops are slightly deformed by patients with tremor. The deformation can be used as a feature for diagnostics.

The theory that lies behind this, says that normal handwriting is marked by automation; the movements are so fast that normal feedback loop by visual perception and muscle control is disabled. This results in an open loop configuration. For Parkinson patients, the automation is no longer valid; their handwriting depends on the visual closed loop.

This approach is questioned by researchers who claim that for Parkinson tremor not the absolute positioning but the grip forces are important [7][8][9]. Thus, a force sensitive device is needed for recording the pressure and not the pen location. This is provided by an electronic pen. Since there are no known features for pressure analysis of Parkinson handwritings, in this paper we investigate different kind of features and characterize them by their diagnostic value.

In all experiments, an electronic pen was used for recording the samples for our analysis.

2 The Electronic Pen Recordings

Let us first introduce how the data were obtained.

2.1 The Electronic Pen

The recording device is an electronic pen, composed of several sensors and integrating a real writing ball pen which was build by the BiSP-Project [10]. This design gives the user the feeling of a real ball pen and facilitates its use by patients. In Fig. 1 the pen architecture and the pen usage are shown.

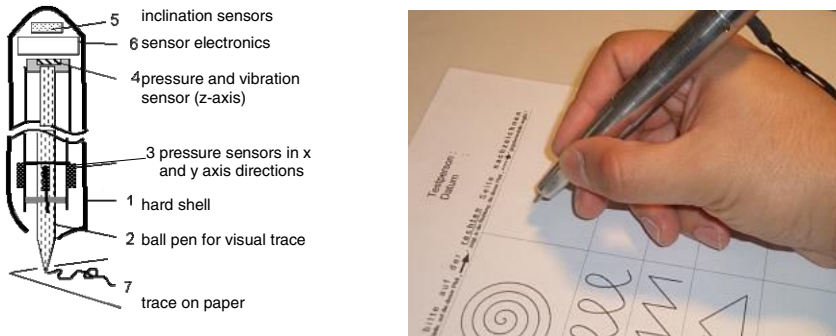


Fig. 1. (a) The architecture of the electronic pen (b) The test writing and recording

In difference to graphic tablets, the recording does not record the absolute position, but only the pressure in x, y and z direction. Additionally, there are two tilt sensors which measure the inclination relative to the x-y plane.

The writing signals are sampled with a sample frequency of $f = 500$ Hz and cut off at 200 Hz.

2.2 The Recordings

In Fig. 2(a), a sample drawing and in Fig. 2(b) its corresponding pressure signals x, y, z and tilt signals α and β are shown.

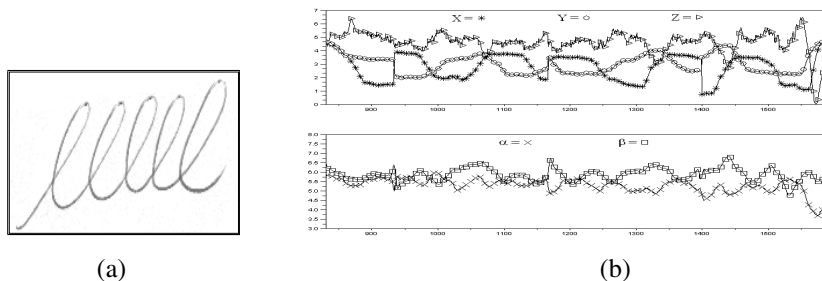


Fig. 2. (a) Handwriting sample. (b) The corresponding recordings of 3D pressure and tilt.

For 28 Parkinson patients and 28 control persons we recorded different patterns: the *lllll* loops, meanders, words and a complete sentence. As preprocessing, all signal noise and artifacts are filtered by *sinc* filters.

For the loop analysis, we selected one segment of $N = 1000$ samples from the recordings of each individual. The same segment was tested by different features.

3 The Analysis Methods and Results

Our analysis concentrated on finding different relevant features for a high separation of the handwriting of Parkinson patients from those of control persons. Since most of the patients received a medical treatment which compensated the tremor, the signal energy in the high frequency band was not a salient feature for itself. Instead, we had to devise more sophisticated features.

The evaluation of the chosen features was done by a receiver operating characteristic (ROC) analysis in order to characterize the diagnostic possibilities not only by their sensitivity, but also by the specificity of the diagnostic approach. For this, the diagnosis $D(u)$ of feature $u(t)$ was defined by

$$D(u) = \begin{cases} \text{ill} & u \geq \theta \\ \text{not ill} & u < \theta \end{cases} \quad (1)$$

The threshold θ was chosen in discrete steps over the whole range of the feature variable.

The first attempt tried to check the validity of the criteria described in literature.

3.1 Relative Number of Loop Extremes

As criteria we chose the relative number of loop extremes e which were already defined in the literature for tablet handwriting input, see [4]. Here, people with Parkinson tremor tend to have more extremes due to a more irregular movement at

writing loops. The number of extremes e depends heavily on the bandwidth of the signal. The increase in $e(f_g)$ with increasing frequency f_g in signals x and y is more rapid for Parkinson patients than for control people. Therefore, for detection we might use the increase from frequency $f_1 = 1$ Hz to frequency $f_{30} = 30$ Hz reflected by the quotients

$$x_e = e_x(f_{30})/e_x(f_1), \quad y_e = e_y(f_{30})/e_y(f_1),$$

$$\alpha_e = e_\alpha(f_{30})/e_\alpha(f_1), \quad \beta_e = e_\beta(f_{30})/e_\beta(f_1).$$

The feature u_1 is then defined by the averages of pressure signal behaviour to tilt signal behaviour

$$u_1 \equiv \frac{x_e + y_e}{\alpha_e + \beta_e}. \quad (2)$$

This gives us an area under the ROC curve of $AUC = 0.896$. The corresponding ROC is shown in Fig. 3

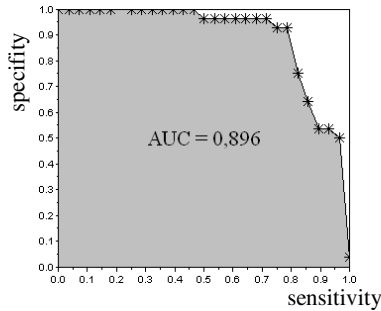


Fig. 3. The receiving operation characteristic for feature u_1

Choosing directly the quotient of extreme values of pressure and tilt for a fixed frequency, say $f = 1$ Hz, gives us feature u_2

$$u_2 \equiv \frac{e_\alpha(f_1) \cdot e_\beta(f_1)}{e_x(f_1)} \quad (3)$$

This has an $AUC = 0.933$ which is even better than that of the first, more complicated feature.

3.2 The Impulse Correlation Coefficient

Regular loops of healthy people tend to be sine-like functions. If we take the derivative of it, the product $-s(t) \cdot s''(t)$ will become maximal. For irregular loops of ill patients, this will decrease. This idea gives us a new feature u_3 which uses the definition of the impulse correlation coefficient p_{s1s2} of two signals $s_1(t)$ and $s_2(t)$

$$p_{s_1 s_2}^E = \frac{\sum_{t=-\infty}^{\infty} s_1(t) \cdot s_2(t)}{\sqrt{E_{s_1} \cdot E_{s_2}}} \quad (4)$$

with energies E_{s_1} and E_{s_2} of the signals. The new feature is then defined for the tilt $\alpha(t)$. Since we are not interested in the offset but in the dynamical behaviour, we take not α and α'' but their derivatives α' and α''' which produces an impulse correlation coefficient of -1 instead of $+1$ as best value

$$u_3 \equiv \frac{\sum_{t=-\infty}^{\infty} \alpha'(t) \cdot \alpha'''(t)}{\sqrt{E_{\alpha'} \cdot E_{\alpha'''}}} \quad (5)$$

Using the tilt α gives the best value from the set of all five sensor signals and an AUC = 0.86. Compared to our first two features this result is not so good. So, let us investigate another set of features by a completely different idea.

3.3 Approximative Entropy

The tremor signals with their irregularities have some similarities to chaotic signals investigated in chaos theory. There, a measure called "approximative entropy" has been introduced [11]. It takes advantage of the fact that very regular signals of automated movements have a high self-similar degree, but those of irregular movements have not. The number of similar points can be counted and gives an occurrence frequency. The average of the logarithm of this gives us an entropy or average information of the time series. The higher the entropy, the more irregular are the movements.

The computation starts with a sliding window of m samples of signal $s(t)$, giving tuples

$$r(i) \equiv (s(i), s(i+1), \dots, s(i+m-1)) \quad (6)$$

The relative number of tuples which are not so different, i.e. which have all entries with a difference smaller than a given number ϵ , is

$$C_i^m(\epsilon) = \frac{|\{(j, k) \mid |r(i+k-1) - r(j+k-1)| \leq \epsilon \text{ for } k=1..m, j=i..N-m\}|}{N-m+1} \quad (7)$$

This is called a *correlation integral*. The averaged logarithm is

$$\Phi_m(\epsilon) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log C_i^m(\epsilon) \quad (8)$$

Then, the approximative entropy of a time series of N samples is defined as the change (tangent) of the entropy due to a variation in the sliding window width m

$$\text{ApEn}(m, \epsilon, N) = \Phi_m(\epsilon) - \Phi_{m+1}(\epsilon) \quad (9)$$

For $m = 2$, $\epsilon = 3$ and $N = 4$ the ApEn of the primary five signals were computed as demonstration. In Fig. 4 a sample plot of the x signal of a control person writing loops and its $ApEn$ are shown. The x pressure signal has been filtered before with a cut off frequency of 100 Hz.

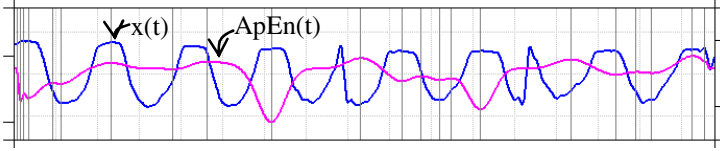


Fig. 4. x -recording of a loop signal and its corresponding ApEn

For computing the feature value of a handwriting, the whole time series $N=1000$ and $\epsilon = 12$ were used. As good feature we found

$$u_4 \equiv \frac{ApEn(z) \cdot ApEn(\alpha)}{ApEn(x) \cdot ApEn(y)} \quad (10)$$

This feature u_4 gives us a AUC of 0.905. Although this is not bad, it is still worse than the second feature u_2 .

3.4 Multiscale Entropy

A remarkable variant of the approximative entropy is the entropy behaviour due to different time scales. Irregular signals can be characterized by a time scaled entropy analysis called multiscale entropy [12]. This obtained by first computing the average and then the logarithm of the correlation integral of eq.(7)

$$\Theta(m, \epsilon, N) = \log \left(\frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} C_i^m(\epsilon) \right) \quad (11)$$

The change in this entropy of averaged probabilities due to changing the window width ϵ is called *sample entropy* $S_E(\cdot)$

$$S_E(m, \epsilon, N) = \Theta(m, \epsilon, N) - \Theta(m+1, \epsilon, N) \quad (12)$$

We might merge several samples of a time series $s(t)$ together into one sample of a new time series $r(t)$. For instance, for a scale factor of $\tau = 3$ we average every three samples into one by

$$r_3(i) = \frac{s(t) + s(t+1) + s(t+2)}{3} \quad \text{with } i = 0, \dots, \lfloor N/3 \rfloor - 1 \quad t = 3i \quad (13)$$

For every scaled time series with scale τ , the sample entropy $S_{E\tau}(\cdot)$ can be computed. The resulting time series shows if there are irregular, more chaotic components which will remain also in higher time scales. In Fig. 5 the different behaviour of natural signals which have a $1/f$ characteristic to not natural ones, e.g. white noise.

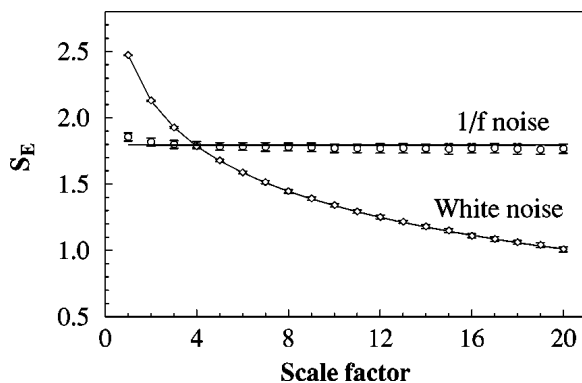


Fig. 5. The multiscale entropy of 1/f noise and white noise (after [12])

We reflect the different signal slopes by the difference $S_E(\tau=4) - S_E(\tau=1)$. As new feature u_5 , let us define the ratio of the pressure behavior to the tilt behavior

$$u_5 \equiv \frac{S_E(x, \tau=4) - S_E(x, \tau=1)}{(S_E(\alpha, \tau=4) - S_E(\alpha, \tau=1))(S_E(\beta, \tau=4) - S_E(\beta, \tau=1))} \quad (14)$$

All signals were filtered by a 100 Hz windowed *sinc* cut off filter before processing, and S_E was computed with $m = 3$ and $\varepsilon = 2$, using the N samples of the scaled time series.

The resulting ROC was not very smooth (see Fig. 6) and the corresponding AUC = 0.859 not very convincing.

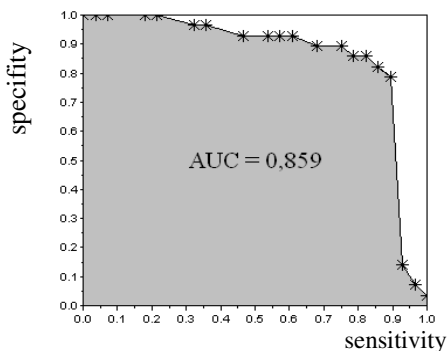


Fig. 6. The ROC of the multiscale feature u_5

Therefore, we looked for a more intrinsic discriminating feature for handwriting between Parkinson patients and control persons.

3.5 Writing Power Rate Coefficient

One of the most apparent characteristics of the Parkinson disease is the muscle rigidity and tremor. Compared to control persons who have a highly optimized and automated handwriting, Parkinson patients use remarkably more effort and stress for compensating their tremor while producing readable handwriting. This results in less power for the handwriting task.

For the meander pattern we got the mean values of the pressure signals and then computed the signal energy per time unit

$$u_6 \equiv \left(\frac{1}{N} \sum_{t=1}^N (x(t))^2 + \frac{1}{N} \sum_{t=1}^N (y(t))^2 \right)^{-1}. \quad (15)$$

This feature gave an only moderate AUC of 0.84.

3.6 Combining the Features

Now, since we have found no ideal feature, what about combining the most successful ones? Since they have different information sources, do this multi-expert team allow a better diagnosis? The AUC values provides us with a ranking of the possible candidates of u_2 , u_4 , u_1 , u_3 , u_5 , u_6 . In fact, taking the best ones gives us a new multi-expert feature

$$u_7 \equiv u_2 u_4 u_1 = \frac{e_\alpha(f_1) \cdot e_\beta(f_1)}{e_x(f_1)} \cdot \frac{\text{ApEn}(z) \cdot \text{ApEn}(\alpha)}{\text{ApEn}(x) \cdot \text{ApEn}(y)} \cdot u_1 \quad (16)$$

Since the quotient $\text{ApEn}(z)/e_x(f_1)$ seemed to be roughly constant for our data, we simplified this to

$$u_8 \equiv e_\alpha(f_1) \cdot e_\beta(f_1) \cdot \frac{\text{ApEn}(\alpha)}{\text{ApEn}(x) \cdot \text{ApEn}(y)} \cdot u_1. \quad (17)$$

obtaining a very nice ROC (see Fig. 7) and a very good AUC of 0.963.

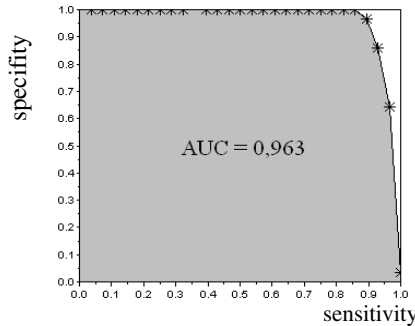


Fig. 7. The ROC of the multi-expert feature u_8

4 Discussion

For the quantitative analysis of muscle anomalies, especially those of Parkinson's disease, there are not many analytic possibilities. Here, we discussed a new method based on pressure and movement features recorded by handwriting with an electronic pen. We evaluated different approaches, some of them based on the tremor characteristics for position pressure and tilt of the pen, and some based on methods for quantification of chaotic signals.

It turned out that the most salient feature (u_2) is based on the difference between the controlled writing pressure in x-y direction and the tilt tremor of the pen. It seems that for these medicated patients the tremor control is better achieved for movements (handwriting) than for constant pressure (pen tilt). The handwriting of Parkinson patients itself have less chaotic characteristics than suspected. Nevertheless, combining all features gives promising indications for diagnosis.

Our results give a base for quantitative evaluation of the state of a Parkinson patient. This can be used in several fashions.

- The features may be used individually in a long term quantitative recording for the treating doctor in order to detect and predict long term changes in the individual disease history.
- Another application may be the adjustment of the medication regime. The quantitative, easy and quick diagnosis provides a basis for a frequent estimation of the proper drug dosage getting the desired effects and avoiding unwanted side effects.
- By a very sensitive diagnosis, Parkinson symptoms may be detected very early on persons with a high risk background. This can be used for early medication avoiding subsequent development of the disease.
- Subsequent studies may use these features not only for a quantitative refinement of a already stated principal diagnosis "Parkinson's disease", but for a diagnosis discriminating general tremor symptoms from special Parkinson ones.

In conclusion, this study showed that there are several features available for good Parkinson diagnosis which can be obtained by simple, cheap and noninvasive handwriting measurements.

Acknowledgments. We want to thank all people who have supported us by their work, especially the BiSP team, university of applied sciences Regensburg, Dr. Kessler of the Maria-Hilf clinic, Mönchengladbach, and Dr. Korchounov of the Parkinson clinic, Bad Nauheim.

References

1. Rajesh Pahwa, Kelly E. Lyons, William C. Koller: *Handbook of Parkinson's Disease: Neurological Disease & Therapy*, Marcel Dekker Inc, 3rd ed., New York 2003
2. Andres Ceballos-Baumann, Bastian Conrad: *Bewegungsstörungen*, Georg Thieme Verlag, Stuttgart 2005

3. Roberto Merletti, Philip M. Parker: *Electromyography: Physiology, Engineering and Non-Invasive Applications*, John Wiley & Sons Inc, New York 2004
4. Marquardt C., Mai N.: *A computational procedure for movement analysis in handwriting*. J. Neurosci Methods (1994) 52:39-45
5. Pullmann S.: *Spiral Analysis: A New Technique for Measuring Tremor With a Digitizing Tablet*, Movement Disorders, vol.13, Suppl.3, (1998), pp85-89
6. Eichhorn T.E., Gasser T., Mai N., Marquardt C., Arnold G., Schwarz J., Oertel W.H. *Computational analysis of open loop handwriting movements in Parkinson's disease: a rapid method to detect dopaminergic effects*. Mov Disord (1996); 11: 289-297.
7. Ingvarsson P.E., Gordon A.: *Coordination of Manipulative Forces in Parkinson's Disease*, Exp. Neurology 145, 489-501 (1997)
8. Fellows S., Noth J.: *Grip Force Abnormalities in De Novo Parkinson's Disease*, Movement Disorders, Vol 19,(5) 560-565 (2003)
9. Fellows S.J., Noth J., Schwarz M. *Precision grip and Parkinson's disease*. Brain 1998; 121: 1771.1784
10. <http://www.bisp-regensburg.de/>
11. Steven M.Pincus: *Approximative entropy as a measure of system complexity*, Proc.Natl. Acad. Sci. USA Vol. 88, pp. 2297-2301
12. Madalena Costa, Ary L. Goldberger, C.-K. Peng: *Multiscale entropy analysis of biological signals*, Physical Review E 71, 021906 (2005)

A Decision Support System for the Automatic Assessment of Hip Osteoarthritis Severity by Hip Joint Space Contour Spectral Analysis

Ioannis Boniatis¹, Dionisis Cavouras², Lena Costaridou¹, Ioannis Kalatzis²,
Elias Panagiotopoulos³, and George Panayiotakis¹

¹ University of Patras, School of Medicine, Department of Medical Physics, 265 00 Patras, Greece

iboniati@yahoo.com, {costarid,panayiot}@upatras.gr

² Technological Institute of Athens, Department of Medical Instrumentation Technology, 122 10 Athens, Greece

{ikalatzis,cavouras}@teiath.gr

³ University of Patras, School of Medicine, Department of Orthopaedics, 265 00 Patras, Greece

ecpanagi@med.upatras.gr

Abstract. A decision support system was developed for the grading of hip osteoarthritis (OA) severity. Sixty four hips (18 normal, 46 osteoarthritic) were studied from the digitized radiographs of 32 patients with unilateral or bilateral hip-OA. Hips were allocated into three OA-severity categories, formed accordingly to the Kellgren and Lawrence scale: “Normal”, “Mild-Moderate”, and “Severe”. Employing custom developed algorithms: (i) the radiographic contrast was enhanced, (ii) 64 ROIs, corresponding to patients’ radiographic Hip Joint Spaces (HJSs), were determined, and (iii) Fourier descriptors of the HJS-ROIs boundary were generated. These descriptors were used in the design of a two-level hierarchical decision tree structure, employed for the discrimination of the OA-severity categories. The overall classification accuracies accomplished by the system, regarding the discrimination between: (i) Normal and osteoarthritic hips, and (ii) hips of “Mild-Moderate” OA and of “Severe” OA were 92.2% and 86.0%, respectively. The proposed system may contribute to osteoarthritic patients management.

1 Introduction

Osteoarthritis (OA) is a rheumatologic disease, the pathophysiology of which is associated with biochemical, cellular, and mechanical processes [1]. The condition is characterized by degeneration of the cartilage with associated underlying bony alterations [2]. Magnetic resonance imaging is characterized by its high sensitivity in detecting soft tissue alterations [3], however, its availability is limited due to its expense. On the other hand, plain film radiography is considered as the imaging modality of reference for the investigation of the disease in daily clinical routine [4]. The characteristic radiographic findings of hip OA include non-uniform Hip Joint Space (HJS) narrowing, subchondral sclerosis, osteophyte formation, and development of subchondral cysts [5]. In the context of the radiographic assessment of hip

OA, the characterization of the severity of the disease concerns mostly the utilization of qualitative grading scales. The latter provide a qualitative assessment of the extent of osteoarthritic joint alterations through subjectively assigned severity grades. The latter are defined on the basis of the characteristic radiographic findings of the disease [6]. The Kellgren and Lawrence (KL) grading system [7] has been considered as the golden-standard for epidemiological studies of OA, despite its limitations [8].

Among the characteristic radiographic findings of hip OA, the narrowing of radiographic HJS has been accepted as a defining criterion of the disease [9], while it has been considered as the most reliable index for the monitoring of the disease progression [5]. The specific feature reflects, indirectly, the progressive loss of the articular cartilage due to OA [4]. Previous studies have introduced thresholds of manually measured HJS-width, for characterizing a hip as normal or osteoarthritic [9, 10]. In previous studies performed by our group, the radiographic texture of HJS has been utilized for the discrimination among OA-severity categories, as well as for the quantification of the severity of the disease [11-14]. However, to the best of our knowledge, a computer-based approach for the assessment of osteoarthritic alterations utilizing information associated with the spectral content of the periarticular contour of the hip joint has not been reported.

The objective of the present study was to develop a decision support system for the grading of hip OA severity from pelvic radiographs. In order to accomplish our goal, Contour Spectral Features (CSFs) were generated from the region of radiographic HJS and were utilized in the design of a computer-based classification scheme. The latter was structured as a hierarchical decision tree, designed so as to characterize hips as: "Normal", of "Mild / Moderate" OA or of "Severe" OA.

2 Materials and Methods

2.1 Patients and Clinical Assessment

Thirty two patients (mean age: 66.7 years, range: 49 years to 83 years) with verified hip OA were included in the study. All patients were candidates for total hip arthroplasty at the University Hospital of Patras. OA diagnosis was based on the clinical and radiographic American College of Rheumatology criteria [15]. In particular, a hip was characterized as osteoarthritic if pain (associated with hip joint use) and limited mobility of the joint were reported in combination with the presence of osteophytes (femoral or acetabular) and joint space narrowing on radiographs. Accordingly, 18 patients were characterized as of unilateral-OA, while 14 as of bilateral-OA.

2.2 Imaging and Radiographs

A pelvic radiograph was available for each patient of the sample, while all radiographs were obtained following a specific radiographic protocol. The latter comprised the following parameters: use of the same X-ray unit (Siemens, Polydoros 50, Erlangen, Germany), tube voltage 70-80 kVp, 100 cm focus to film distance, alignment of the X-ray beam 2 cm above the pubic symphysis, use of a fast screen and film cassette (30 cm x 40 cm). Pelvic radiographs were digitized employing a laser digitizer, suitable for medical applications (Lumiscan 75, Lumisys, Sunnyvale, CA, USA) [16].

2.3 Radiographic Assessment of Hip Osteoarthritis Severity

The severity of OA was assessed by three experienced orthopaedists, who employed the KL scale [7]. The specific grading system defines five OA-severity categories via an equal number of grades ranging between 0 and 4. Grade 0 is assigned to a normal hip joint, while grade 4 indicates a severe osteoarthritic condition. Intermediate levels of OA-severity, characterized as “Doubtful”, “Mild”, and “Moderate” are described by the grades 1–3, respectively [7]. Each of the orthopaedists graded OA by assigning a KL grade to the 64 hips of the sample, while in order to establish a golden-standard, only those exams of common consent were retained for the purposes of the present study. Based on the KL grades, three major OA-severity categories were formed, in which the hips of the sample were allocated into: “Normal (KL=0, 1)”, “Mild / Moderate (KL=2, 3)”, and “Severe (KL=4)”. In this context, 18 hips were assigned to the “Normal / Doubtful”, 16 to the “Mild / Moderate”, and 30 to the “Severe” category.

2.4 Determination of Radiographic Hip Joint Space

A custom developed contrast enhancement algorithm, based on the adaptive wavelet transform, [17] was applied on the digitized radiographs in order to emphasize the articular margins of the hip joint.

On each enhanced pelvic radiograph, two Regions Of Interest (ROIs), corresponding to patient’s both HJSs, were determined employing custom developed software [18, 19]. In particular, by utilizing patient’s standard anatomical landmarks [20] an acute angle encompassing the weight-bearing portion of the hip joint was formed. As it can be observed in Fig. 1, the summit of the angle was determined by the centre of the femoral head (O), the medial limit was defined by the line joining the centre of the femoral head and the highest point of the homolateral sacral wing (OR), while the lateral limit was provided by the line joining the centre of the femoral head and the

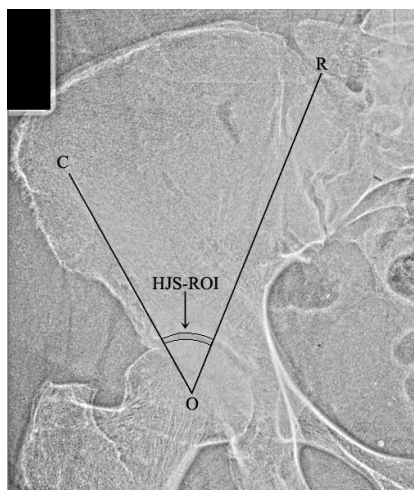


Fig. 1. Determination of Hip Joint Space (HJS) Region of Interest (ROI) by utilizing patient’s anatomical landmarks



Fig. 2. Segmented Hip Joint Space (HJS) Region Of Interest (ROI), corresponding to Fig. 1

lateral rim of the acetabulum (OC). Within this angle, the operator delineated manually the articular margins of the joint (edge of the femoral head, inferior margin of the acetabulum). This HJS-ROI (Fig. 2) was subjected to further Contour Spectral Analysis (CSA).

2.5 Contour Spectral Analysis of Radiographic Hip Joint Space

For the needs of the present study, a number of computational descriptors of the radiographic HJS, associated with the spectral content of its periarticular margins, were introduced. These descriptors, labeled as CSFs, were defined on the basis of a Contour Spectral Analysis (CSA) approach. The latter concerned the suitable combination of contour-based shape representation and 1-D digital signal processing techniques. In particular, the computation of CSFs involved a two-step process: (i) the generation of the outline profile of the HJS-ROI, which provided an one-dimensional representation of radiographic HJS boundary, and ii) the computation of the Discrete Fourier Transform (DFT) of the generated 1-D signal [21, 22].

In this context, the following tasks were performed, employing custom developed algorithms in Matlab (The Mathworks Inc, Natick, MA, USA):

1. Determination of the centre of “mass” (“centroid”) of the HJS-ROI
2. Tracing of the exterior boundary of the HJS-ROI
3. Calculation of the radial Euclidean distances between the centroid and each point of the exterior boundary of the HJS-ROI. In particular, assuming that the exterior boundary of the ROI comprises N pixels, the coordinates of the centroid $(\bar{x}, \bar{y})$ are given by:

$$\bar{x} = \frac{1}{N} \sum_{n=0}^{N-1} x(n) \quad (1)$$

and

$$\bar{y} = \frac{1}{N} \sum_{n=0}^{N-1} y(n) \quad (2)$$

where $x(n)$ and $y(n)$ are the discrete coordinates of each boundary pixel [21]. The outline profile is then defined as a 1-D discrete signal of length N :

$$d(n) = \sqrt{[x(n) - \bar{x}]^2 + [y(n) - \bar{y}]^2} \quad (3)$$

An example of a generated outline profile is presented in Fig. 3.

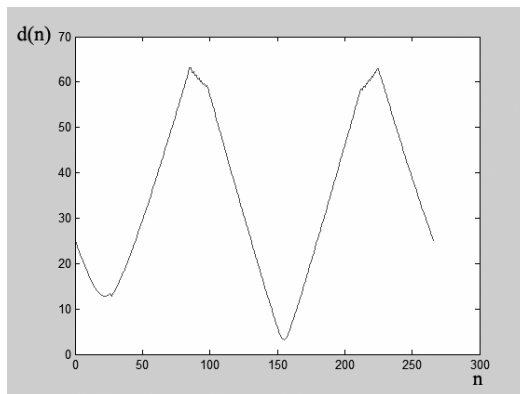


Fig. 3. Outline profile of the Hip Joint Space Region Of Interest, corresponding to Fig. 2

4. Computation of the DFT [22] of the generated 1-D profile signal:

$$D(k) = \frac{1}{N} \sum_{n=0}^{N-1} d(n) \cdot e^{-j \frac{2\pi kn}{N}}, \quad 0 \leq k \leq N-1 \quad (4)$$

The amplitude was computed by $|D(k)|$ and it was normalized to the region $[0, 1]$. For the needs of the present study, and due to the symmetry of the DFT [22], the $|D(k)|$ signal corresponding to frequencies in the interval $[0, \frac{N}{2} - 1]$ (see Fig. 4) was used for the generation of the CSFs.

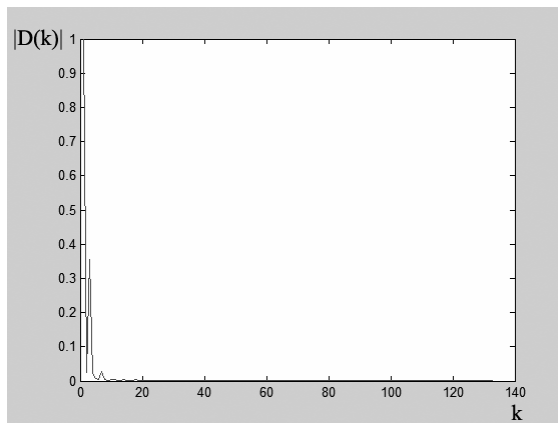


Fig. 4. The Amplitude of the Discrete Fourier Transform of the Outline Profile, corresponding to Fig. 3

2.6 Generation of the Contour Spectral Features

Two sets of CSFs were generated from the $|D(k)|$:

- Normalized moments [21, 23] of the $|D(k)|$. The specific features ($CSF\_ \overline{m}_p$) were generated according to:

$$CSF\_ \overline{m}_p = \frac{m_p}{(M_2)^{p/2}} = \frac{\frac{1}{N/2} \sum_{k=0}^{\frac{N}{2}-1} [|D(k)|]^p}{\left\{ \frac{1}{N/2} \sum_{k=0}^{\frac{N}{2}-1} [|D(k)| - m_1]^2 \right\}^{p/2}} \quad (5)$$

where, $CSF\_ \overline{m}_p$ is the p^{th} normalized moment of $|D(k)|$, m_p is the p^{th} moment of $|D(k)|$, defined according to:

$$m_p = \frac{1}{N/2} \sum_{k=0}^{\frac{N}{2}-1} [|D(k)|]^p \quad (6)$$

while, M_2 is the second central moment of $|D(k)|$, defined as:

$$M_2 = \frac{1}{N/2} \sum_{k=0}^{\frac{N}{2}-1} [|D(k)| - m_1]^2 \quad (7)$$

Five moment-based CSFs ($CSF\_ \overline{m}_1 \div CSF\_ \overline{m}_5$) were generated, corresponding to p values ranging between 1 and 5.

- Descriptors of frequency bands

The spectrum of frequencies was divided into three bands, corresponding to Low, Medium, and High frequencies (see Fig. 5).

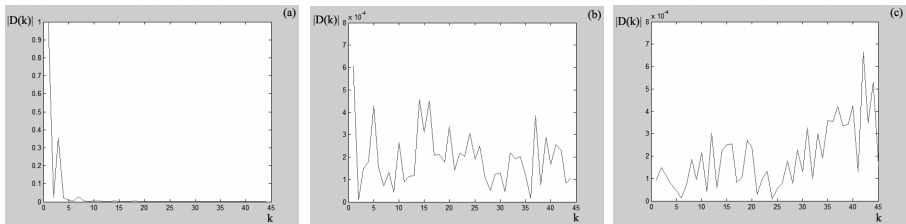


Fig. 5. Bands of: Low (a), Medium (b), and High (c) frequencies

For each frequency zone x (x : Low, Medium, High), the following CSFs were generated:

- CSF_{6_x} : the mean value of $|D(k)|$
- CSF_{7_x} : the maximum value of $|D(k)|$
- CSF_{8_x} : the minimum value of $|D(k)|$
- CSF_{9_x} : the position (frequency) of the maximum value of $|D(k)|$
- CSF_{10_x} : the position (frequency) of the minimum value of $|D(k)|$
- CSF_{11_x} : the absolute value of the difference between the mean and the maximum value of $|D(k)|$

Summarizing, for each HJS-ROI a pattern vector of 23 CSFs was formed, which comprised: 5 $\overline{m}_p$ moment-based descriptors ($CSF\_ \overline{m}_p, p=1..5$) and 18 frequency bands descriptors ($CSF_{6x} \div CSF_{11x}$, 6 features for each one of the 3 bands).

2.7 Design of the Grading Classification Scheme

For the computer-based grading of hip OA-severity, a classification system was implemented as a two-level hierarchical decision tree (see Fig. 6). The first level of the system was designed so as to discriminate between normal and osteoarthritic hips. At the second level, the hips that had correctly been characterized as osteoarthritic at the first level were further classified as of Mild / Moderate OA or of Severe OA.

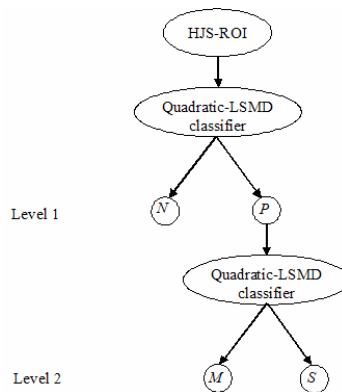


Fig. 6. Hierarchical decision tree structure for the discrimination between: (i) Normal (N) and Osteoarthritic hips (P) at Level 1, and (ii) hips of “Mild / Moderate” OA (M) and of “Severe” OA (S) at Level 2. HJS-ROI: Hip Joint Space – Region Of Interest, Quadratic-LSMD: Quadratic Least Squares Minimum Distance.

As it can be observed in Fig. 6, the classification task at each level of the tree structure was performed by the Quadratic Least Squares Minimum Distance (Quadratic-LSMD) classifier [24].

2.7.1 The Quadratic-Least Squares Minimum Distance Classifier

The Quadratic-Least Squares Minimum Distance (Quadratic-LSMD) classifier discriminates among classes by implementing quadratic decision boundaries. The specific classification algorithm is a modified version of the Least Squares Minimum Distance (LSMD) classifier, in the sense that employs a quadratic equation within the Least Squares method. In the Quadratic approach, for a pattern vector $\mathbf{x}$ of p -dimensionality, a pattern vector $\mathbf{y}$ of augmented dimensionality q is formed, through a one – to – one transformation. Due to the latter, the quadratic decision functions involving the elements of pattern $\mathbf{x}$, can be written in the form of linear equations employing the elements of the augmented pattern $\mathbf{y}$. Thus, the classifier finally opines similarly to the LSMD classifier. Accordingly, the augmented pattern $\mathbf{y}$ is mapped from the augmented feature space into a decision space wherein the patterns of a class are clustered around a pre-selected point. The transformation that implements the mapping from the feature space to the decision space is chosen so as the overall mean-square mapping error is minimized (Least Squares). An unknown pattern is assigned to a class if it is closest (Minimum Distance) to the predefined point of the decision space, corresponding to the class [24].

2.7.2 Feature Selection and Evaluation of System Performance

In order to determine the feature combination providing the highest classification accuracy with the minimum number of features (“optimum” or “best” feature combination) the exhaustive search procedure was followed in conjunction with the Leave One Out (LOO) performance evaluation method [25]. In particular, the generated features were exhaustively combined with each other (i.e. combinations of two, three, four, etc. features) in order to form a pattern vector. For every feature combination, the performance of the classifier, expressed in terms of overall accuracy, was evaluated according to the LOO method [25]. The above described procedures were followed at each level of the decision tree-structure. In order to safeguard against variations in the dynamic range of the generated features, a fact that could result in inaccurate classification scores, the features were normalized to zero mean and unit standard deviation [25] according to:

$$\tilde{f} = \frac{f_i - \mu}{\sigma} \quad (8)$$

where $\tilde{f}$ is the normalized value of the i^{th} feature (f_i), while μ and σ are the mean value and standard deviation, respectively, of f_i feature over all HJS-ROIs.

2.8 Statistical Analysis

The student’s t-test was used in order to investigate the existence of statistically significant differences ($p < 0.05$) between normal and osteoarthritic hips for the generated

feature values. The Coefficient of Variation (CV) [26] was used in order to assess the reproducibility of the HJS-ROI determination process. In particular, each of the orthopaedists segmented the HJS-ROIs twice, according to the previously described procedure. A time interval of about a month was intervened between the two determinations, while the evaluation scores were employed for the calculation of the CV. High degree of reproducibility is indicated by low values of the CV coefficient, and vice versa. Student's paired t-test was used in order to investigate whether features extracted from the two determinations differed significantly ($p < 0.05$). All statistical processing was performed utilizing the "Matlab Statistics Toolbox".

3 Results and Discussion

In the present study, a decision support system was designed with the intension to be used by orthopaedists as an assistance tool for the grading of hip OA-severity.

In radiographic images of the hip joint, the loss of articular cartilage in osteoarthritic hips is indicated by the narrowing of radiographic HJS [4]. In the present study, osteoarthritic alterations of radiographic HJS were assessed by means of the introduced contour spectral features. The latter provide information regarding the frequency content of a signal, which represents the specific anatomical region through its boundary (periarticular contours of the hip joint). Statistical analysis revealed the existence of statistically significant differences ($p < 0.05$) between normal and osteoarthritic hips for the CSF values. This finding indicates that osteoarthritic alterations, perceived in the 2-dimensional representation (image) of radiographic HJS, are associated with spectral differentiations concerning a 1-dimensional (signal) representation of the outline of the specific anatomical region. The determination of radiographic HJS-ROI was found to be reproducible. Regarding the intra-observer reproducibility, the CV was found equal to 3.4%, on average, indicating the reliability of the segmentation process. Inter-observer reproducibility was also high, since the corresponding value for the CV was 4.2%, on average. In addition, the feature values that were generated from the twice-determined HJS-ROIs were found not to differ significantly ($p > 0.05$).

At the first level of the hierarchical decision tree structure, the overall classification accuracy achieved by the Quadratic-LSMD classifier was 92.2%, since 59 out of 64 hips were correctly classified. As it can be observed from Table 1, all the normal hips but two were properly characterized, resulting in a discrimination accuracy of 88.9%. On the other hand, only three osteoarthritic hip were misclassified (93.5% accuracy). The optimum feature combination provided the aforementioned scores comprised the features [CSF_{9_HIGH} , CSF_{11_HIGH} , CSF_m_4 , CSF_m_5].

Table 1. Truth table tabulating classification scores for the discrimination between normal and osteoarthritic hips at the first level of the hierarchical decision tree

Hip characterization	Normal	Osteoarthritic	Accuracy (%)
Normal	16	2	88.9
Osteoarthritic	3	43	93.5
Overall accuracy (%)			92.2

At the second level of the hierarchical tree, the hips, which had correctly been classified as osteoarthritic at the first level were further characterized as of Mild / Moderate or of Severe OA. The Quadratic-LSMD classifier performed the specific discrimination task accomplishing an overall accuracy of 86.0%. As it can be observed from Table 2, only two hips of Mild / Moderate OA were incorrectly assigned to the 'Severe' class, resulting in 86.7% accuracy. Referring to the hips of 'Severe OA', the Quadratic-LSMD classifier discriminated correctly 24 out of 28 hips (85.7% accuracy), employing the optimum feature combination comprising the features $[CSF_{8\_MEDIUM}, CSF_{10\_LOW}, CSF_{-\bar{m}_1}, CSF_{-\bar{m}_2}]$.

Table 2. Truth table for the discrimination between hips of Mild / Moderate osteoarthritis and of Severe osteoarthritis at the second level of the hierarchical decision tree

Osteoarthritis severity category	Mild / Moderate	Severe	Accuracy (%)
Mild / Moderate	13	2	86.7
Severe	4	24	85.7
Overall accuracy (%)			86.0

The classification scores accomplished in the present study, may be indicative of the capacity of the proposed system to assess hip OA. However, in order to arrive at even more reliable conclusions regarding the utility of the suggested approach, the employment of a larger dataset is considered as necessary.

4 Conclusion

In conclusion, the introduced by the present study CSFs can be utilized for the assessment of structural alterations in osteoarthritic joints. The specific features, which provide information regarding the spectral content of the articular margins of the hip joint, were employed in the design of a computer-based system that discriminated efficiently normal from osteoarthritic hips, while it graded reliably the severity of OA. Taking into consideration that the system is compatible with the KL scale, it could be used as a second opinion diagnosis decision support tool, contributing to the management of osteoarthritic patients.

Acknowledgements. Ioannis Boniatis was supported by a grant by the State Scholarship Foundation (SSF), Greece. The authors thank the staff of the Departments of Orthopaedics and Radiology of the University Hospital of Patras for their contribution to this work.

References

1. Hinton, R., Moody, R.L., Davis, A.W., Thomas, S.F.: Osteoarthritis: Diagnosis and Therapeutic Considerations. *Am. Fam. Physician* 65 (2002) 841-848
2. Iannone, F., Lapadula, G.: The Pathophysiology of Osteoarthritis. *Aging Clin. Exp. Res.* 15 (2003) 364-372

3. Peterfy, C.G.: Imaging of the Disease Process. *Curr. Opin. Rheumatol.* 14 (2002) 590-596
4. Ory, P.A.: Radiography in the Assessment of Musculoskeletal Conditions. *Best. Pract. Res. Clin. Rheumatol.* 17 (2003) 495-512
5. Altman, R.D., Fries, J.F., Bloch, D.A., et al.: Radiographic Assessment of Progression in Osteoarthritis. *Arthritis Rheum.* 30 (1987) 1214-1225
6. Sun, Y., Günther, K.P., Brenner, H.: Reliability of Radiographic Grading of Osteoarthritis of the Hip and Knee. *Scand. J. Rheumatol.* 26 (1997) 155-165
7. Kellgren, J.H., Lawrence, J.S.: Radiological Assessment of Osteoarthrosis. *Ann. Rheum. Dis.* 16 (1957) 494-501
8. Spector, T.D., Cooper, C.: Radiographic Assessment of Osteoarthritis in Population Studies: Whither Kellgren and Lawrence? *Osteoarthritis Cartilage* 1 (1993) 203-206
9. Croft, P., Cooper, C., Wickham, C., Coggon, D.: Defining Osteoarthritis of the Hip for Epidemiologic Studies. *Am. J. Epidemiol.* 132 (1990) 514-522
10. Ingvarsson, T., Hägglund, G., Lindberg, H., Lohmander, L.S.: Assessment of Primary Osteoarthritis: Comparison of Radiographic Methods Using Colon Radiographs. *Ann. Rheum. Dis.* 59 (2000) 650-653
11. Boniatis, I., Costaridou, L., Cavouras, D., Panagiotopoulos, E., Panayiotakis, G.: Quantitative Assessment of Hip Osteoarthritis Based on Image Texture Analysis. *Br. J. Radiol.* 79 (2006) 232-238
12. Boniatis, I., Costaridou, L., Cavouras, D., Kalatzis, I., Panagiotopoulos, E., Panayiotakis, G.: Assessing Hip Osteoarthritis Severity Utilizing a Probabilistic Neural Network Based Classification Scheme. *Med. Eng. Phys., In Press*
13. Boniatis, I., Costaridou, L., Cavouras, D., Kalatzis, I., Panagiotopoulos, E., Panayiotakis, G.: Osteoarthritis Severity of the Hip by Computer-Aided Grading of Radiographic Images. *Med. Bio. Eng. Comput., In Press*
14. Boniatis, I., Costaridou, L., Cavouras, D., Panagiotopoulos, E., Panayiotakis, G.: A Computer-Based Image Analysis Method for Assessing the Severity of Hip Joint Osteoarthritis. *Nucl. Instrum. Meth. A., In Press*
15. Altman, R., Alarcón, G., Apperlouth, D., et al.: The American College of Rheumatology Criteria for the Classification and Reporting of Osteoarthritis of the Hip. *Arthritis Rheum.* 34 (1991) 505-514
16. Lumiscan 75, system specifications. Lumisys Inc. 1998; <http://www.lumisys.com/support/manuals.html>
17. Sakellaropoulos, P., Costaridou, L., Panayiotakis, G.: A Wavelet Based Spatially Adaptive Method for Mammographic Contrast Enhancement. *Phys. Med. Biol.* 48 (2003) 787-803
18. Sakellaropoulos, P., Costaridou, L., Panayiotakis, G.: An Image Visualisation Tool in Mammography. *Med. Inform. Internet. Med.* 24 (1999) 53-73
19. Sakellaropoulos, P., Costaridou, L., Panayiotakis, G.: Using Component Technologies for Web - Based Wavelet Enhanced Mammographic Image Visualization. *Med. Inform. Internet Med.* 25 (2000) 171-181
20. Conrozier, T., Tron, A.M., Balblanc, J.C., et al. : Measurement of the Hip Joint Space Using Computerized Image Analysis. *Rev. Rhum. Engl. Ed.* 60 (1993) 105-111
21. Rangayyan, R.M.: Biomedical image analysis. CRC Press LLC, Boca Raton (2005)
22. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. 2nd edn. Prentice Hall, Inc., New Jersey (2002)
23. Gupta, L., Srinath, M.D.: Contour Sequence Moments for the Classification of Closed Planar Shapes. *Pattern Recognition* 20 (1987) 267-272
24. Ahmed, N., Rao, K.R.: Orthogonal Transforms for Digital Signal Processing. Springer Verlag, Berlin Heidelberg New York (1975)

25. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. 2nd edn. Elsevier Academic Press, Amsterdam (2003)
26. van Belle, G., Fisher, L.D., Heagerty, P.J., Lumley, T.: Biostatistics. A methodology for the health sciences. 2nd edn. Wiley-Interscience, New Jersey (2004)

Modeling for Missing Tissue Compensator Fabrication Using RFID Tag in U-Health*

O-Hoon Choi¹, Jung-Eun Lim¹, Hong-Seok Na², and Doo-Kwon Baik¹

¹ Dept. of Computer Science and Engineering, Korea University, 136-701, Seoul, Korea

² Dept. of Computer and Information Science, Korea Digital University, Seoul, Korea
{pens,jelim,baik}@software.korea.ac.kr, hsna99@kdu.edu

Abstract. U-Health (Ubiquitous based Healthcare System that supports medical services) is one of the technology areas proposed to realize the vision of ubiquitous computing. A plethora of different alternative or complementary RFID sensing technologies and RFID management systems are available. And mostly RFID technologies in medical facilities are applied for tracking a patient's location, storing medical equipment, and keeping on patient's record. In this paper, we primarily apply RFID technology to measure the affected part which is needed to gain information about its volume, size and mass. Thus we will propose modeling method with using RFID tags for making missing tissue compensator which is used in Radiation Therapy. The missing tissue compensator is commonly used to maximize the effect of skin protection and to irradiate an even dose on tumor tissue. Existing missing tissue compensator marked the contour of the body surface directly on the patient's skin using a curved ruler or used medical images such as computerized tomography images and magnetic resonance images. In addition, the application of medical images is expensive. In this paper we will obtain necessary 3 dimension location information using RFID technology which is fixed on the surface of the patient's affected parts using a rubber mask. The rubber mask has RFID tags on its surface. So RFID readers to detect RFID tags on the mask obtain each of tags' location information, and we calculate them to make a missing tissue compensator. According to the result, the missing tissue compensator modeled in this research compensated defective tissue and protected normal tissue, so it was considered clinically applicable.

Keywords: RFID, Compensator, Radiation Therapy.

1 Introduction

The purpose of radiotherapy is to maximize the dose on the tumor and minimize the dose on normal tissue. Uneven dose appearing in radiotherapy is caused by the change of the thickness of the body surface due to the loss of weight, by the loss of tissue in operation, or by a bent or slope on the body surface due to difference in thickness of the body surface. In order to compensate such defective tissue, we use wedge filter, bolus, missing tissue compensator that can absorb unnecessary

* This work is supported by the second Brain Korea 21 project.

radioactive rays and protect a normal cell in a patient. [1][2] Among them, the missing tissue compensator is commonly used to maximize the effect of skin protection and, at the same time, to irradiate an even dose on tumor tissue. However, existing missing tissue compensator marked the contour of the body surface directly on the patient's skin using a curved ruler or used medical images such as computerized tomography images and magnetic resonance images. [3][4] The patient may feel uncomfortable and there may be an error caused by the patient's movement if information on the contour of the body surface is obtained using a curved ruler directly the patient's skin. In addition, the use of medical images is expensive. Thus we will propose modeling method with using RFID tags for making missing tissue compensator which is used in Radiation Therapy. In this paper, we primarily apply RFID technology to measure the affected part which is needed to gain its information about its volume, size and mass. We will acquire basic 3 dimension location information using RFID tags which is fixed on the surface of the patient's body. We use 13.56MHz RFID reader to recognize RFID tags response time. And RFID readers detect ranges are 50 cm from surface the patient's face.

2 Related Works

2.1 Technology of Location Recognition

Measure the distance by calculating the position of the object from various base points. To calculate the position of the object 2-dimensionally, measurements from 3 base points not on the same line are needed. To calculate the position of the object 3-dimensionally, measurements from 4 base points not on the same line are needed. Methods of measuring the distance are generally divided into 3 types. First, direct measurement of distance, which is simple but as the measurement has to be taken through physically moving around, it is difficult to automate the movements. Second, measure the time taken to move in constant velocity between the object and set point, thus calculating the distance between the object and set point. [5][6] Third, measure the distance by employing the fact that signal intensity decreases as the distance increases.

In this case, the decrease of signal relative to the original signal is called diminution. If the diminution and relative function between the distances are given, the distance between the object and a set point can be calculated by measuring the signal intensity at a set point. Angle measurement is similar to distance measurements, but instead of distance angle is used to determine the position of an object. For 2-dimensional angle measurement, 2 angles and a distance between base points is needed. For 3-dimensional angle measurement, 2 angles, a distance between base points and also a point of compass. Image Analysis Position Recognition technique uses characteristics of an image observed at specific point. The image observed is simplified to abstract characteristics prone to comparison and expression. In Static Image Analysis, a pre-defined data table is mapped to the position of an object, thus by searching the observed characteristics within the data table the position of an object can be determined. Automatic Image Analysis tracks the differences between two consecutive images and this difference corresponds to the movement of the object.

2.2 Measurement of Relate Location with Using RFID Tags

Objects which are attached RFID tag exist irregularly within 3 dimension space like as indoor. RFID reader will reactivate to recognize each of RFID tag identities. In this paper we use 4 RFID readers to acquire response time per each RFID tags. In Fig. 1, each RFID tags reacts in the different RFID reader and that reaction time is use to calculate 3 dimension location measurement methods and a relative distance.

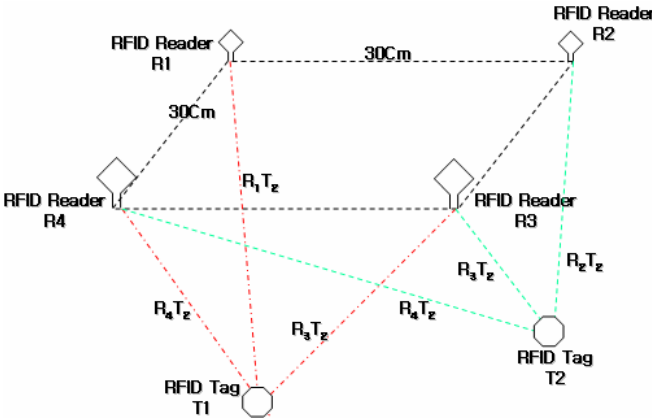


Fig. 1. Recognition of multi RFID Tags

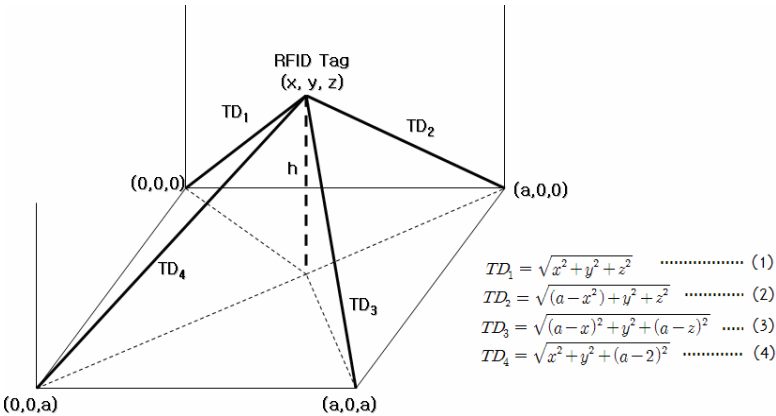


Fig. 2. Location Formula using 3 Dimensional coordinates

- RFID Tag Response Time ($R_i T_n$) : Measuring RFID tag's response time by ms(10-3 second) with each RFID leader.
- RFID Tag Distance ($T_i D_n$) : Using RFID Tag Response Time which is measured by ms unit, we calculate the direct distance among RFID leader and RFID tags. Because the speed of RFID is too fast, we don't calculate an absolute distance. So we uses only the difference of distance with reaction velocity

- RFID Tag 3 Dimensional Location : (x,y,z) Calculating RFID Tag Distance (TiDn) measured from over 3 RFID readers. Fig. 2 shows a location formula using 3 dimensional coordinates. It helps to making missing tissue compensator modeling. The Location of RFID Tag (x, y, z) is represented by Formula (1), (2), (4) and (4) and h(eight) computed by the Pythagorean theorem.

3 Missing Tissue Compensator Modeling

3.1 Measuring Surface Information

By using 13.56MHz RFID reader we detect RFID tag on surface of Humanoid Cranial Phantom. Fig. 3 shows we acquire tag information on Gantry as X-ray equipment with using RFID. We get the area information of same height surface from 3 dimensional location information. Recognized IDs from each RFID reader makes an image model by the location formula using 3 dimensional coordinates in Fig. 2.



Fig. 3. Location Measurement on Gantry

In table 1, X-axis represents Humanoid Cranial Phantom's horizontal axis. And it does treatment center axis with 0cm, left range with 4.5cm, and right range with 5 cm. Y-axis covers range with 8.6 cm from upper layer with 5cm to upper layer with 3.5cm. We assume the model of treat-

ment is head cancer, which range is 7cm x 6.5cm. However, in this paper, we enlarge the range of treatment to cover the cancer adding X-axis with 2.5cm, and Y-axis with 2cm. Finally, we acquire the information of surface depth which is covered 9.5cm x 8.5cm range. The use of information of surface depth in table 1, we make a map of surface contour, which is separated by color. Thickness data of surface contour is basic model to realize the missing tissue compensator.

3.2 Modeling of Missing Tissue Compensator

Based on table 1, we produce the surface contour by 5mm, and made the Map of surface contour to compensate thickness with contour information from each section. (Fig. 4) The process of missing tissue compensator is generally divided 3 parts. First part is exact measuring the missing tissue for care within treatment range. Second part is deciding the material thickness of missing tissue compensator which is complement. [7] We designed the compensator which is packed with wedge filter in X-ray accelerator, and fixed a distance as 49cm from x-ray source to missing tissue compensator. The material of missing tissue compensator is lead (Bb, density:11.35 g/cm³) and its thickness calculation equation is (5)

$$\text{Lead thickness ratio } (\tau)/\rho_{\text{com}} = 0.7/\rho_{\text{com}} = 0.061 \quad (1)$$

$$(\rho_{\text{com}} : \text{density of compensator material } (\rho_{\text{Pb}} = 11.35 \text{ g/cm}^3))$$

Table 1. Thickness data of surface contour

		X-axis (cm)																			
Y-axis (cm)	5.5	-4.5	-4.0	-3.5	-3.0	-2.5	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
	5.0	2.1	3.8	5.7	6.7	7.9	8.5	9.0	9.3	9.5	9.8	9.8	9.9	9.9	9.9	10.0	9.9	9.9	9.9	9.8	9.6
	4.5	1.9	3.9	5.7	7.1	8.2	8.6	9.1	9.4	9.5	9.6	9.8	9.9	10.0	10.0	10.1	10.0	10.0	10.0	9.8	9.7
	4.0	2.0	4.0	5.6	7.2	8.3	8.7	9.1	9.4	9.5	9.7	9.9	10.0	10.0	10.1	10.1	10.1	10.0	10.0	9.8	9.7
	3.5	2.2	4.2	6.3	7.4	8.3	8.7	9.1	9.3	9.5	9.5	9.8	9.9	10.0	10.1	10.1	10.1	10.1	9.9	9.8	9.7
	3.0	2.7	4.4	6.5	7.5	8.2	8.6	9.1	9.3	9.4	9.3	9.5	9.7	9.8	10.0	10.1	10.1	9.9	9.8	9.8	9.5
	2.5	2.5	4.5	6.3	7.6	8.0	8.5	8.8	9.0	9.1	9.2	9.3	9.2	9.4	9.7	9.9	9.7	9.3	9.0	9.0	9.4
	2.0	2.3	4.9	6.5	7.5	7.9	8.5	8.8	8.9	9.0	9.0	9.2	9.2	9.5	10.0	10.1	9.7	9.4	9.0	9.0	8.9
	1.5	1.4	4.0	5.5	7.0	7.8	8.3	8.5	8.8	8.9	9.0	9.3	9.4	9.8	10.2	10.3	10.3	9.6	9.3	9.0	9.0
	1.0	2.1	5.3	6.7	7.4	7.7	8.2	8.5	8.7	8.9	9.0	9.2	9.4	10.0	10.4	10.6	10.3	9.7	9.4	8.9	9.0
	0.5	3.2	5.0	6.2	7.0	7.8	8.3	8.6	8.8	8.9	9.1	9.3	9.6	10.3	10.6	11.0	10.6	10.3	9.6	9.3	9.2
	0.0	2.3	4.3	6.1	7.0	7.9	8.2	8.6	8.8	9.0	9.3	9.6	10.0	10.5	11.2	11.4	11.2	10.5	10.0	9.5	9.3
	-0.5	2.4	4.3	6.2	7.0	7.9	8.1	8.5	8.8	9.0	9.4	9.5	10.1	10.8	11.4	11.7	11.4	10.8	10.1	9.6	9.4
	-1.0	2.0	3.6	5.7	6.7	7.5	7.9	8.3	8.6	8.9	9.4	9.6	10.4	11.0	11.5	11.9	11.7	11.2	10.0	9.6	9.4
	-1.5	3.7	5.9	6.7	7.3	7.6	8.1	8.2	8.7	8.9	9.3	9.5	10.5	11.1	11.8	12.0	11.8	11.5	10.5	9.6	9.3
	-2.0	0.7	3.9	5.3	6.3	7.2	7.6	8.1	8.5	8.8	9.3	9.5	10.6	11.1	11.7	11.9	11.8	11.2	10.4	9.5	9.1
	-2.5	0.7	3.1	4.7	6.1	7.1	7.6	8.1	8.5	8.7	9.2	9.5	9.9	11.0	11.7	11.8	11.8	11.1	10.2	9.4	9.1
	-3.0	0.0	2.8	4.9	6.0	6.8	7.2	7.8	8.3	8.8	9.2	9.5	10.0	10.5	10.5	10.7	10.3	10.0	9.7	9.3	9.0
	-3.5	0.0	1.5	5.8	5.2	6.4	7.1	7.6	8.3	8.7	9.1	9.5	9.8	10.1	10.3	10.2	10.2	9.9	9.7	9.3	8.8

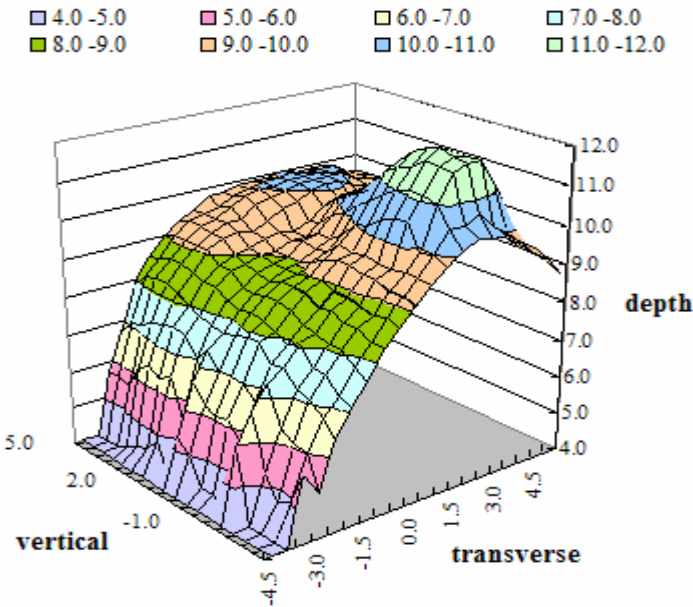


Fig. 4. Map of Surface Contour

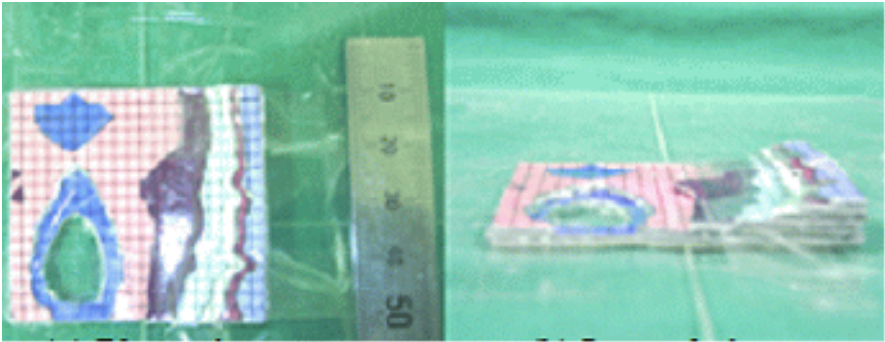


Fig. 5. Fabrication of custom-made compensator: (a) plane view (b) lateral view

4 Evaluation

4.1 Measurement of Output for Missing Tissue Compensator

The result of output measurement based on the depth of lead is measured 9.39 nC (numerical control) without adding 9.39 nC and 8.87 nC with adding 0.5 mm lead in case that source-axis distance is set 100 cm, surveyed area is 10 cm x 10 cm. The result value when changing lead depth from 1 mm to 10 mm with giving 1 mm depth change are measured from 8.74 nC to 5.13 nC and generating 3.87nC, 2,94 nC in each when having depth of 15 mm and 20mm. The result value of when changing SAD to 120 cm is measured 6.53 nC with 0.1 mm lead depth, 6.18 nC with 0.5 mm lead depth, 6.07 nC to 3.55 nC when measuring by 1 mm from 1 mm to 10 mm and 2.71 nC, 2.04 nC in each when measuring 15 mm and 20 mm. When increasing lead depth by 1 mm, the change of SAD and Surveyed Area were not having much difference and since the change of generations are measured from 5.5% to 6.8%, just less than 1 % was resulted compared to rate of lead depth. [8]

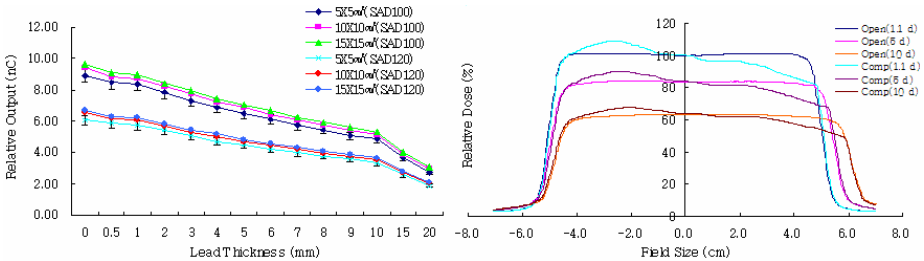


Fig. 6. (a) It shows that relationship between output and lead thickness. (b) Comparison of relative dose profiles between with and without compensator by collimator angle 0 degree (comp.: compensator, d: depth).

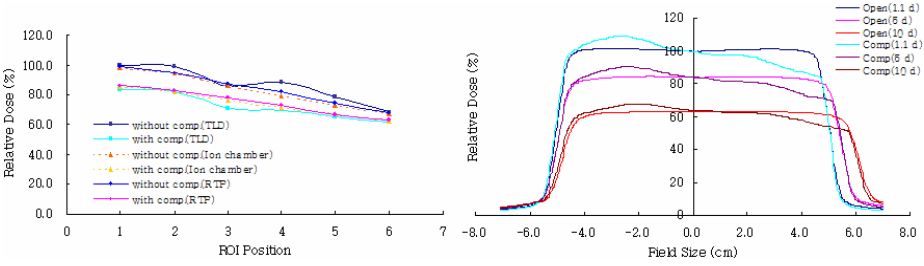


Fig. 7. (a) Comparison of relative dose measurements between with and without compensator by ion chamber, TLD and RTP data. (b) Comparison of relative dose profiles between with and without compensator by collimator angle 90 degree (*comp.:* compensator, *d:* depth).

4.2 Geometric Evaluation

Based on use or non-use for missing tissue equivalent compensator, the characteristics changes of energy are measured for performance evaluation of Particle Accelerator. Geometric evaluation list is symmetry, flatness and after test, as you can see table 2. symmetry is measured 0.5%, 0.8%, 0.6% in each with Dmax(maximum depth, 1.1cm) and 5 cm, 10 cm in depth in case of collimator 0° and 0.6%, 0.7%, 0.8% in case Collimator 90°

The flatness is measured with an error between 0.7% and 2.6% and if the change of center point off-axis is showed less than 0.02 cm change from the central axis. From the this result, we can understand the treatment machine allow just less than ± 2 □ error, thus, we can confirm energy characteristics are accurately materialized according to use or un-use of missing tissue equivalent compensator.

Table 2. Displacement of isocenter by collimator rotation

		Collimator 0°			Collimator 90°		
Depth		Dmax	5 cm	10 cm	Dmax	5cm	10cm
Symmetry (%)	open	0.5	0.8	0.6	0.6	0.7	0.8
Flatness (%)	open	0.7	1.4	2.5	0.6	1.2	2.6
Center Point off-axis(cm)	open	0	0.02	0.03	0.02	0.01	0.01
	Comp.	0.01	0.02	0.01	0.02	0.02	0.01

4.3 Radiation Physics Evaluation

The cooperation of radiation beam profile based on use and un-use of missing tissue equivalent compensator should have no Radiation Physics change despite the change of Collimator, so when Collimator set as 0 and 90, we tested two case. With no use of missing tissue equivalent compensator, we measured 1.1cm, 5cm, 10cm according to central axis and with using information of surface contour from MRI, we can see that the difference of dose is resulted less than $\pm 2\%$.

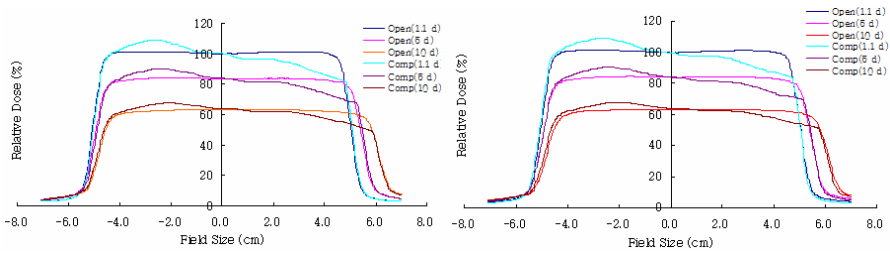


Fig. 8. (a) Comparison of relative dose profiles between with and without compensator by collimator angle 0 degree (*comp.*: compensator, *d*: depth). (b) Comparison of relative dose profiles between with and without compensator by collimator angle 90 degree (*comp.*: compensator, *d*: depth).

5 Conclusion

Under Ubiquitous environment, end user should be provided all necessary data within reasonable time. Also, in the area of U-Health, correction of Data and management has been researched for treating a patient to satisfy this kind of demand. This study applied to the manufacture of missing tissue equivalent compensator for protecting normal cell with using RFID Tag when having X-ray treatment. Currently, high-expense equipment like MRI is used for compensating the change of surface contour caused by face curve or operation. The methods for compensating current surface contour are getting information with using a lead ruler to patient directly and MRI image. However, these methods are giving unpleasant feeling, burdening economical anxious to patient due to high expense medical equipment and having demerit that is not properly responding to radical change of surface contour, but the method of using RFID Tag which is given by this study has the merit that shortening the time of manufacturing and convenience, accuracy and low-cost. Also, the generation which is emphasized on actual clinical demonstration and mechanical accuracy is having less than $\pm 2\%$ and that is agreed by the standard of AAPM. From now on, the research is necessary for compensating for missing tissue equivalent compensator which is cause by RFID Tag's incorrectness of response time, limitation of frequency range and distortion of RF signal.

References

1. Faiz M, Khan, Ph.D, "The Physics of Radiation Therapy", 2nd Ed, pp.299-307, 1994.
2. Harold Elford Hohns, PhD, John Robert Cunningham, PhD, "The Physics of Radiology", 4th Ed, pp.380-390, 1983.
3. John Robert Cunningham, Ph.D, "The Physics of Radiology", 4th Ed, pp.389-390, 1983.
4. S.C. Sharma, M.W. Johnsom, "Clinical considerations in the use of missing tissue compensators for Head and Neck cases", Medical Dosimetry, Vol.23, No.4, pp.267-270, 1998.
5. Jeffrey Hightower and Gaetano Borriello, A Survey and Taxonomy of Location Systems for Ubiquitous Computing, Technical Report, Computer Science and Engineering, University of Washington, Aug. 2001.

6. Parambir Bahl and Venkata N. Padmanabhan, "RADAR : An in-building RF-based user location and tracking system", INFOCOM, March, 2000, pp. 75-784
7. Arthur L. Boyer and Michael Goitein, "Simulator mounted Moire topography camera for constructing compensator filters", Med. Phys, 7(1), 19-25, 1980.
8. AAPM, Radiation Therapy Committee Task Group 40, Med. Phys, 21, pp.581-618, 1994.

The Effect of User Factors on Consumer Familiarity with Health Terms: Using Gender as a Proxy for Background Knowledge About Gender-Specific Illnesses

Alla Keselman^{1,2}, Lisa Massengale¹, Long Ngo³, Allen Browne¹, and Qing Zeng⁴

¹ LHCNCB, National Library of Medicine, NIH, DHHS, Bethesda, MD

² Aquilent, Inc., Laurel, MD

³ DSG, Brigham and Women's Hospital, Harvard Medical School, Boston, MA

⁴ DSG, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA
keselmana@mail.nih.gov

Abstract. An algorithm estimating vocabulary complexity of a consumer health text can help improve readability of consumer health materials. We had previously developed and validated an algorithm predicting lay familiarity with health terms on the basis of the terms' frequency in consumer health texts and experimental data. Present study is part of the program studying the influence of reader factors on familiarity with health terms and concepts. Using gender as a proxy for background knowledge, the study evaluates male and female participants' familiarity with terms and concepts pertaining to three types of health topics: male-specific, female-specific and gender-neutral. Of the terms / concepts of equal predicted difficulty, males were more familiar with those pertaining to neutral and male-specific topics (the effect was especially pronounced for "difficult" terms); no topic effect was observed for females. The implications for tailoring health readability formulas to various target populations are discussed.

Keywords: consumer health informatics; readability formulas; consumer health vocabularies.

1 Introduction

Studies suggest that individuals frequently have difficulties reading health texts, and that the readability of most consumer health websites are beyond the reading level of the average consumer [1]. Vocabulary complexity is one of the text factors that contribute to this difficulty [2]. An informatics tool that could evaluate vocabulary complexity of a health text and suggest consumer-friendly synonyms for "difficult" medical terms could help address this problem. The development of such a tool, however, is a challenging task. First of all, a definition of a "difficult" health term is required. Many general-purpose readability formulas estimate word "difficulty" in terms of their length [3]. This approach may not be appropriate for consumer health domain, ridden with many short technical terms that are likely to be unfamiliar to lay health consumers (e.g., "myelin", "apnea"). Moreover, reader's ability to comprehend a text is affected by many factors that are located with the reader (e.g., prior

knowledge, motivation), rather than with the text [4]. The effect of various reader factors on comprehension in consumer health texts domain may be somewhat different from other domains. For example, in many “general” domains individuals with high levels of educational attainment are more likely to comprehend texts with complex vocabulary. In the consumer health domain, however, experience with a particular disease may override “insufficient” education level.

We have previously developed a regression model for predicting consumers’ “familiarity likelihood scores” with health terms. The model relies on two sources of information: 1) empirical data from user studies evaluating “Consumer-Friendly Display” names for medical concepts [5] and (2) term frequency counts from consumer health corpora [6]. The algorithm assigns each consumer health term a predicted familiarity likelihood score from 0-1 range. Terms with scores in the 0.8-1 sub-range are categorized as “likely” to be familiar to health consumers, scores in the 0.5-0.8 sub-range are categorized as “somewhat likely” to be familiar, and scores in the 0-0.5 sub-range are categorized as “not likely” to be familiar. A validation study with 52 participants showed that model-based scores were indeed predictive of consumer recognition and understanding of health terms [7]. The validation study also pointed to two reader factors that could mediate familiarity with health terms: health literacy and English proficiency.

Present study continues to explore reader factors influencing familiarity with consumer health terms and concepts. General comprehension literature contains many testimonies of the effect of background knowledge on text comprehension [8]. Part of the positive effect of background knowledge on comprehension has to do with the fact that background knowledge broadens vocabulary knowledge [9]. Matching vocabulary complexity of consumer health materials to the level of background knowledge of potential readers may therefore improve readability. Web designers and writers rarely conceptualize consumer health audiences in terms of their background knowledge. Instead, audiences are usually defined in terms of some demographic and/or experiential factors (e.g., patients with a specific disease, women, seniors). A match between topic and reader characteristics, however, is likely to influence background knowledge, as individuals are more likely to have knowledge of issues that they have personally experienced and that are specific to their group. We may expect women to be more familiar than men with terminology pertaining to female-specific diseases, and diabetes patients to be more familiar than the general public with diabetes terminology. Identifying demographic factors that are likely to affect term familiarity may allow us to make the predictive model more sensitive by adjusting it to various target population groups.

The general hypothesis underlying this study is that readers’ background knowledge influences their familiarity with health terms and concepts. The specific hypotheses concern the effect of gender on consumer familiarity with terms related to gender-specific health issues. They are the following:

1. Participants’ gender will affect their familiarity with terms related to different health topics. Given comparable familiarity likelihood scores of the terms, men will be more familiar with terms pertaining to male-specific and neutral health topics than with terms pertaining to female-specific topics. Similarly, women will be more familiar with terms pertaining to female-specific and neutral topics than with those pertaining to male-specific topics.

2. The relationship between gender and topic may differ for terms with different predicted familiarity likelihood scores. Relatively common terms that are predicted as highly likely to be familiar may be equally familiar to both genders regardless of the topic. However, familiarity with terms predicted as unlikely to be familiar may be more affected by the gender-topic match.

Gender in this study was chosen as a proxy for better knowledge of gender specific health issues. The study was not concerned with the general effect of gender on the knowledge of health terminology and concepts.

Based on our previous findings, we also expected that regression model-based familiarity likelihood level would be predictive of consumers' actual term familiarity.

2 Methods

2.1 Participants

Convenience sample of 50 employees of the US National Library of Medicine was recruited for the study. Twenty five of the participants were males, and twenty five were females. All had adequate health literacy skills (scores in the 23-36 range out of 36, comparable average scores for both groups), according to Short Test of Functional Health Literacy in Adults (S-TOFHLA) [10]. Male and female groups had comparable educational levels. For each gender group, seven participants had high school level of education (possibly with some college work, but without college diploma), nine were college graduates, and nine had graduate degrees. Female participants were slightly younger than male participants (3 males and 8 females in the 18-25 year old category, 12 males and 9 females in the 26-39 year old category, 10 males and 7 females in the 40-59 year old category, and 1 female in the over 60 years old category).

2.2 Instrument

The survey instrument used in this study tested consumer familiarity with 27 health-related terms. Nine of these terms pertained to conditions that were prevalent among or specific to males (baldness and prostate cancer); nine terms pertained to conditions specific to females (menopause and pregnancy). The terms were extracted from consumer health websites on these four topics, linked to MedlinePlus consumer health portal of the US National Library of Medicine. The remaining nine terms were gender neutral terms extracted from MedlinePlus-linked consumer health website on the topic of gastroesophageal reflux disorder (GERD). From now on, for simplicity, we will refer to the terms as "male", "female" and "neutral." These labels, however, refer to the topics of the texts from which the terms were extracted, rather than to the terms themselves.

The terms were selected to be comparable in "familiarity likelihood scores", as computed by our regression model algorithm [5, 6]. In each group of nine terms ("male", "female" and "neutral"), three terms were categorized as "likely" to be familiar to health consumers, three were predicted as "somewhat likely" to be familiar, and three were predicted as "not likely" to be familiar.

Alopecia: __nails __teeth __hair __don't know

(stem) (2 distractors) (key)

Fig. 1. Sample item from the familiarity test

The layout of the survey was modeled on the Short Assessment of Health Literacy for Spanish-speaking Adults (SAHLSA) [11], which in turn is based on the Rapid Estimate of Adult Literacy in Medicine (REALM) health literacy test for English speakers [12]. SAHLSA consists of 50 items, each with a “stem” or target term, “key” or semantically-related term, “distractor,” and a “don’t know” option to discourage guessing. The goal of SALHSA is to measure both reading ability and comprehension, and the task is to both correctly select and correctly pronounced the key answer option. Since we were interested in evaluating participants’ familiarity with health terms in written consumer health materials, the SALHSA requirement to pronounce the key answer was replaced with a second distractor (Figure 1).

Table 1. Familiarity instrument terms

Familiarity Likelihood	“Male”	“Female”	Neutral
Likely to be familiar	baldness prostate prostatitis	folic acid osteoporosis menopause	asthma acid reflux biopsy
Somewhat likely to be familiar	scalp testosterone urethra	ovaries uterus prenatal	pulmonary fibrosis esophagus antacids
Not likely to be familiar	rogaine hematuria alopecia	perimenopause phytoestrogens blastocyst	heartburn sphincter internist

Two types of questions were developed for each term:

- Surface-level familiarity questions assessed the ability to match written health terms with basic relevant associated terms at the super-category, location or function level (eg, alopecia → hair) (Figure 1).
- Concept familiarity questions assessed the ability to associate written terms with brief phrases describing the meaning or “gist” (e.g., alopecia → hair loss).

The final instrument consisted of 54 questions (27 surface level familiarity questions and 27 concept familiarity questions). Table 1 presents distribution of items among topics and predicted difficulty scores.

2.3 Administration and Scoring

Participants first completed the demographic survey, followed by S-TOFHLA and familiarity survey, with surface-level items followed by concept familiarity items.

Surface-level familiarity and concept familiarity scores were calculated separately, in the following way. First, for each type of familiarity, correct answers were assigned the score of 1, while incorrect answers were assigned the score of 0. Next, for each of the three categories of familiarity likelihood (likely, somewhat likely and not likely) within each of the three categories of topic (“male”, “female” and “neutral”), the sum of the three answers was computed. Thus, for each type of familiarity score, there were 9 measurements for each subject. Each measurement represented a score for a difficulty level within a topic, and ranged from 0 to 3.

2.4 Statistical Analysis

We used linear mixed-effects models [13] to estimate the quantities of interest. We first checked the distribution of the Surface-Level Term Familiarity Score, and the Concept Familiarity Score. The distributions of these two outcome variables appeared to be reasonably normal. Since there were multiple measurements for each subject, the models took into account the within-subject correlation by treating the within-subject measurements using compound symmetry variance-covariance matrix structure. Linear contrasts were then used to obtain the linear combination of parameters of interest (e.g. the estimated mean difference between male and female score for male participants, or for female participants).

To model Surface-Level Term Familiarity, the independent variables Predicted Familiarity Likelihood Score (raw scores from the 0-1 range), Gender, Highest Education Level, Age, and Topic (“male”, “female” or “neutral”) were used as dependent variables in the linear mixed-effects model. Similarly, the same independent variables were used to model Concept Familiarity Score. As health literacy scores of all participants were in the “adequate” range as measures by S-TOFHLA, health literacy was not used as a variable in the analysis due to lack of variation. Education and age were included in the models as potential confounders, as well as to detect potentially meaningful trends for future studies.

3 Results

3.1 Overall Patterns

Both regressions found statistically significant effects ($P < .001$) of predicted familiarity likelihood score on surface-level term familiarity and concept familiarity. Surface-level familiarity model also found statistically significant effect of topic, with participants appearing most familiar with neutral terms, followed by “male” and then “female” terms. The effect of female specific vs. neutral terms was -0.58 ($P < .001$), and the effect of male specific vs. neutral terms was -0.34 ($P = .007$). This may be due to the fact that while familiarity likelihood scores for the three topics were evenly distributed among the three familiarity likelihood categories, the raw scores were somewhat higher for “female” terms. No significant topic effects were found for the concept familiarity model.

3.2 Hypothesis 1: Gender Differences in Mean Familiarity Scores for Different Topics

Means and standard deviations of participants' surface level term and concept familiarity scores by gender and topic are presented in Table 2.

Table 2. Mean surface-level and concept familiarity scores

	Surface-Level Familiarity mean (SD)			Concept Familiarity mean (SD)		
Gender	"Male" terms	"Neut" terms	"Female" terms	"Male" terms	"Neut" terms	"Female" terms
Male (n=75)	2.17 (0.79)	2.45 (0.66)	1.85 (1.23)	2.18 (0.82)	2.26 (0.78)	1.68 (1.08)
Female (n=75)	2.57 (0.64)	2.64 (0.69)	2.36 (0.95)	2.56 (0.66)	2.41 (0.79)	2.33 (0.95)

n=75 refers to the number of observations used in the analysis (3 data points per participant) rather than participants

Male Participants' Performance. Male participants showed greater surface-level familiarity with "male" and "neutral" terms than with "female" terms (Table 2). The estimated mean difference between familiarity with "male" vs. "female" terms (corrected for differences in predicted familiarity likelihood scores for "male" and "female" terms, Age, and Education) via linear contrast from the linear mixed-effects model was 0.24 (SE=0.12, $P=0.059$). The estimated mean corrected difference between familiarity with neutral vs. "female" terms was 0.58 (SE=0.12, $P<.001$). Male participants also showed greater surface-level familiarity with neutral terms than "male" terms, mean corrected difference 0.34 (SE=0.12, $P=0.007$).

Concept familiarity was similarly greater for "male" and "neutral" terms than for "female" terms among male participants. The corrected difference between familiarity with "male" vs. "female" concepts was 0.45 (SE=0.13, $P<.001$). The mean corrected difference between familiarity with neutral vs. "female" terms was 0.57 (SE=0.13, $P<.001$).

Female Participants' Performance. No statistically significant effect of topic on surface-level and concept familiarity was found for female participants.

3.3 Hypothesis 2: The Effect of Predicted Familiarity Likelihood Scores on the Relationship Between Gender and Topic (for Male Participants)

We hypothesized that the relationship between gender and topic may differ for terms with different predicted familiarity likelihood scores. As the overall effect of topic on familiarity was not significant for female participants, the analysis for Hypothesis 2 was conducted for male participants only. Table 3 presents the relationship between predicted familiarity likelihood (based on the ranges used in our previous work, see Introduction) and actual familiarity scores for the three topics for male participants. Examination of the data suggests that the greatest difference in mean familiarity

scores between “male” vs. “female” terms and “male” vs. “neutral” terms lies at the level of “difficult” terms predicted not likely to be familiar. A linear mixed-effects model was used to estimate the mean corrected difference between “male” and “female” terms (1.32, SE=0.19, $P<0.0001$), and “male” and “neutral” terms (-0.44, SE=0.19, $P=0.02$).

Table 3. Male participants’ performance by term difficulty level

	Surface-level term familiarity Mean (SD)			Concept familiarity Mean (SD)		
<i>Familiarity likelihood</i>	“Male” terms	“Neut” terms	“Female” terms	“Male” terms	“Neut” terms	“Female” terms
Likely (n=25)	2.28 (0.54)	2.88 (0.33)	2.44 (0.58)	2.20 (0.70)	2.68 (0.56)	1.84 (0.80)
Somewhat (n=25)	2.60 (0.58)	2.40 (0.65)	2.80 (0.50)	2.64 (0.64)	2.16 (0.75)	2.64 (0.49)
Not likely (n=25)	1.64 (0.91)	2.08 (0.70)	0.32 (0.56)	1.72 (0.84)	1.96 (0.84)	0.56 (0.65)

The pattern was similar for concept familiarity. The estimated corrected mean difference between “male” and “female” terms was 1.16 (SE=0.21, $P<0.0001$), and between “male” and “neutral” terms was -0.14 (SE=0.21, $P<0.0001$).

4 Discussion

The study supported our notion that while a primarily frequency-based algorithm for estimating consumer familiarity with health terms has significant predictive power, some reader factors may also carry predictive weight. Including these reader factors into the regression model algorithm can potentially make the model more powerful. In particular, this study pointed to the effect of gender on familiarity with health terms that pertain to gender-specific topics.

The findings of the study also suggest that the relationship between gender and knowledge of terminology related to gender-specific health topics may be less straightforward than we had expected. As expected, men were more likely to be familiar with “male” and “neutral” terms than with “female” terms of comparable predicted difficulty. Also as expected, the relationship mostly existed at the level of low frequency terms, predicted to be largely unfamiliar. The findings for the female participants, however, were unexpected, as no difference was found in women’s familiarity with terms pertaining to different topics. One possible explanation for this is that women are more likely than men to play the role of family caregivers and therefore be familiar with health issues that are not directly relevant to them [14]. We should also keep in mind that this study only tested a small set of terms related to four gender specific health issues, presented to the participants out-of-context, rather than within a passage. Finally, the relationship may be somewhat obscured by the ambiguity of the concept of gender-specific terminology. While all the terms used in

the study pertained to gender-specific health issues, some of them denoted anatomical structures and attributes that were common to both males and females (e.g., “scalp”, “urethra”). It is conceivable that the effects of the study would be stronger and would generalize to female participants, had we chosen a different definition of “gender-specificity.”

In the present study, gender was used as a proxy for background knowledge. The findings support the idea that background knowledge and experience are likely to affect individuals’ familiarity with health-related terminology. A follow-up study could validate these findings by including a direct measure of background knowledge and correlating it with gender. Other (perhaps more clinically promising) proxies of background knowledge that deserve research attention are health status, diagnosis and time since diagnosis. For example, patients with chronic illnesses and experience with managing their health status are likely to be more knowledgeable about terms and concepts related to their condition than the newly diagnosed. This, in turn, will have implications for setting the optimal terminological and conceptual complexity of e-health websites targeting various specific populations.

The ultimate goal of our research agenda is to develop an algorithm that could predict readability and comprehensibility of consumer health materials for individuals. Part of the challenge lies in identifying the factors that are likely to affect readability and term familiarity. While the specific hypotheses of the present study addressed the effect of gender, level of education was also included in the regression analysis. The lack of education effect is counter-intuitive, may be due to the limited value range of the variable (high school to graduate school), and perhaps warrants a more thorough investigation. Another variable that is likely to affect term familiarity is health literacy. Studying the effect of health literacy is methodologically difficult, because most existing tests (e.g., S-TOFHLA) have low ceiling and are not sensitive enough to detect health literacy variations in a typical convenience sample. Yet another part of the challenge lies in accurately estimating the size of the effect of various factors, and then incorporating these effect sizes into the formula. This task requires collecting the data on large samples of participants, using a wide range of health terms.

When thinking about tailoring health materials to individuals’ characteristics, it is important to distinguish between stable and transient features of readers and users. Transient features are those that exhibit significant fluctuations as a function of time and context. These may include the users’ mood, level of fatigue and current blood glucose level. Stable features are those that can only be changed with a significant investment of time and effort (if at all), including for example presence of a chronic disease that requires continuous management, caregiver status, and age. While both factors may affect reading comprehension, it is presently unrealistic for us to talk about tailoring messages to accommodate transient states. Instead, we can focus on stable characteristics that constitute membership in the targeted audience group for the health materials in question. For example, if our hypothesis of the relationship between background knowledge and terminological knowledge is true, we can further hypothesize that patients’ knowledge about their disease increases with time since diagnosis. We can then envision two versions of a website dedicated to providing information about a specific disease. One version would be for the newly diagnosed, the other for individuals who have lived with this diagnosis for a period of time.

Vocabulary complexity could then constitute one of the differences between the two versions. Similarly, a website providing support to caregivers can have some information specifically tailored for male and female caregivers. Findings of our study (if confirmed by subsequent research) would suggest that the information for female caregivers could accommodate more complex terminology with less detriment to comprehension.

This study looked at the effect of gender on two types of consumer familiarity with health terms: surface-level familiarity and concept familiarity. While the findings for both types of familiarity were similar, we should not assume that these results would generalize to all contexts. Consumer health term and concept familiarity has more inherent complexity than the present survey captures. Historically, health literacy studies do not distinguish among different levels of familiarity, from associating the term with a broad health area it belongs to, to deep understanding of the underlying concept. The ability to associate the term with a related term or use it in a sentence correctly is often viewed as an indicator of understanding the underlying concept. However, the relationship between surface level familiarity and conceptual knowledge may be non-linear. In a previous study we have shown that conceptual knowledge may lag behind terminological familiarity, and that the gap may be greater for frequent terms that are more likely to be familiar [7]. Our current algorithm was not specifically designed to predict conceptual knowledge. Further work is knowledge assessment is necessary for optimize the algorithm for predicting understanding.

In summary, this paper presents a step in a research program, intended to accumulate knowledge for developing a formula for predicting readability of health materials for various consumer groups. Follow-up work should address the limitations of this study by increasing the scope of terms in the study, including additional individual factors, defining the continuum of term/concept familiarity and developing methodology for assessing various stages of familiarity. Findings of such research program can be used in the design of tools for assisting consumers with information seeking and comprehension of health materials.

Acknowledgments. This research was supported by the Intramural Research Program and the Association Fellowship of the US National Library of Medicine, US National Institutes of Health (AK, LM, AB) and NIH grant R01 LM007222-05 (LN, QZ). The authors thank Tony Tse and Guy Divita for their comments on the earlier versions of this manuscript.

References

1. Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. *JAMA* 2002 May 22-29;287(20): 2691-2700.
2. Berland GK, Elliott MN, Morales LS, et al. Health information on the Internet: accessibility, quality, and readability in English and Spanish. *JAMA* 2001 May 23-30;285(20): 2612-2621.
3. Flesch JR, Kincaid C. Flesch-Kincaid Readability Formula. Boston: Houghton Mifflin. 1965.

4. Guthrie TJ, Wigfield A, Metsala JL, Cox KE. Motivational and Cognitive Predictors of Text Comprehension and Reading Amount. *Scientific Studies of Reading* 1999; 3 (3): 231-56
5. Zeng Q, Tse T, Crowell J, Divita G, Roth L, Browne AC. Identifying consumer-friendly display (CFD) names for health concepts. *Proc AMIA Symp* 2005: 859-63.
6. Zeng Q, Kim E, Crowell J, Tse T. A text corpora-based estimation of the familiarity of health terminology. *Proc ISBMDA* 2005: 184-92.
7. Keselman A, Tse T, Crowell J, Browne A, Ngo L, Zeng Q. Assessing Consumer Health Vocabulary Familiarity: An Exploratory Study. *Proc MEDNET* 2006.
8. Langer JA. Examining Background Knowledge and Text Comprehension Reading Research Quarterly 1984; 19 (4) : 468-81
9. Anderson RC, Freebody P. Vocabulary knowledge. In J.T. Guthrie (Ed.), *Comprehension and teaching: Research reviews*. Newark: International Reading Association. 1981.
10. Baker DW, Williams MV, Parker RM, Gazmararian JA, Nurss J. Development of a brief test to measure functional health literacy. *Patient Educ Couns* 1999 Sep;38(1): 33-42.
11. Lee S-YD, Bender DE, Ruiz RE, Cho YI. Development of an easy-to-use Spanish health literacy test. *Health Serv Res*. 2006;41(4):1392-1412.
12. Davis TC, Long SW, Jackson RH, Mayeaux EJ, George RB, Murphy PW, et al. Rapid estimate of adult literacy in medicine: a shortened screening instrument. *Fam Med*. 1993;25(6): 391-5.
13. Laird MN, Ware HJ. (1982). *Random-Effects Models for Longitudinal Data*. *Biometrics* 38: 963-974.
14. Bull MJ. Interventions for women as family caregivers. *Annu Rev Nurs Res*. 2001;19:125-42

ICT for Patient Safety: Towards a European Research Roadmap

Veli N. Stroetmann¹, Daniel Spichtinger¹, Karl A. Stroetmann¹,
and Jean Pierre Thierry²

¹ empirica Communication and Technology Research, Oxfordstrasse 2,
D-53111 Bonn, Germany

² Symbion, Maisons-Laffitte, France
veli.stroetmann@empirica.com

This paper analyses key issues towards a research roadmap for eHealth-supported patient safety. The *raison d'être* for research in this area is the high number of adverse patient events and deaths that could be avoided if better safety and risk management mechanisms were in place. The benefits that ICT applications can bring for increased patient safety are briefly reviewed, complemented by an analysis of key ICT tools in this domain. The paper outlines the impact of decision support tools, CPOE, as well as incident reporting systems. Some key research trends and foci like data mining, ontologies, modelling and simulation, virtual clinical trials, preparedness for large-scale events are touched upon. Finally, the synthesis points to the fact that only a multilevel analysis of ICT in patient safety will be able to address this complex issue adequately. The eHealth for Safety study will give insights into the structure of such an analysis in its lifetime and arrive at a vision and roadmap for more detailed research on increasing patient safety through ICT.

Healthcare as a Risky Endeavour

Reflecting on more than a decade of global research, two, by now famous, USA Institute of Medicine (IOM) reports, *To Err Is Human* [1] and *Crossing the Quality Chasm* [2] highlighted the risks of modern healthcare. The first report included an estimate that organisational systems failures in healthcare delivery (i.e., poorly designed or “broken” care processes) were responsible for at least 90,000 deaths each year in the USA. The second report revealed a wide “chasm” between the quality of care the health system should be capable of delivering today (given the astounding advances in medical science and technology in the past half century) and the quality of care most Americans received. In its recent report *Ending the Document Game: Connecting and Transforming Your Healthcare Through Information Technology* [3], the US Commission on Systemic Interoperability pointed out that medical errors are killing more people each year than breast cancer, AIDS, or motor vehicle accidents.

It is widely believed that the situation in many, if not all European health delivery contexts is characterised by similar, if not the same deficiencies. Of activities seen as potentially risky, travel by rail in Europe or commercial air travel are actually among the safest activities, with fewer than one in 100,000 fatalities per personal encounter

or trip. Driving is far more dangerous as Fig. 1 shows. It is no surprise that statistically, mountain climbing and bungee jumping are among the most dangerous activities. But a great surprise is that there are more deaths per encounter with the healthcare system than for any of the other activities [4].

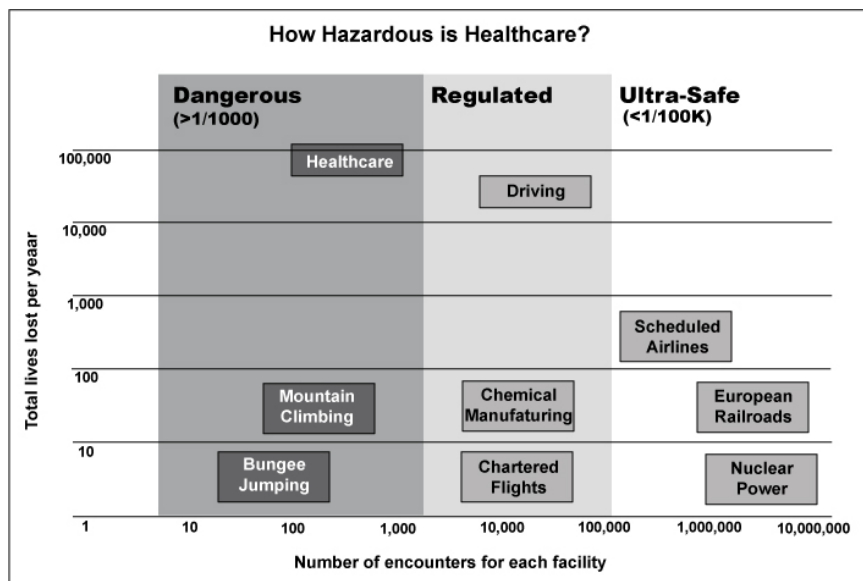


Fig. 1. Risk of fatality in different domains (Source: AHRQ, Commission on Systemic Interoperability, USA, 2005)

ICT in Healthcare: Current State of Play

The benefits that information and communications technologies (ICT) can bring for improved quality of care and increased patient safety are briefly reviewed in this section, complemented by a short analysis of the state of play in the implementation of some key ICT tools.

ICT applications can be useful in almost every aspect of healthcare, including the delivery of care to remote locations, reducing costs, increasing the efficiency of delivery, facilitating information and communication within and among healthcare organisations, simplifying diagnostic and therapeutic processes and, last but not least, increasing the quality of care provided to patient, including improvements in patient safety [5]. ICTs are expected to help relieve the strain that healthcare systems experience: the pressure to increase the quality of care and decrease costs simultaneously [6]. The recent IOM/NAE report, *Building a Better Delivery System: A New Engineering/Health Care Partnership* [7] underscores the importance of information and communications technologies for meeting multidimensional performance challenges. It also identified proven, fundamental engineering concepts, such as designing for safety, mass customisation, continuous flow, and production

planning, that could be brought to bear immediately to redesign and improve care processes to facilitate risk management, deliver greater patient safety and better quality.

Furthermore, Wachter [8] indicates that “it seems self-evident that many, perhaps most, of the solutions to medical mistakes will ultimately come through better information technology. We may finally be nearing the time when institutions and providers will not be seen as credible providers of safe, high-quality care if they lack a strong IT backbone.” This development, he adds, is fuelled by the activities of the *Leapfrog Group*, a business coalition that promotes patient safety through public reporting and pay for performance initiatives [9]. The *National Audit Office* in the UK also sees the preventing of errors by the appropriate use of information technology as a well established fact [10].

A report on a workshop about the use of ICT for patient safety and risk management [11], organised by the *European Commission* in 2004, outlines as a key finding that information society technology can reduce the rate of errors in three ways: by preventing errors and adverse events, by facilitating a rapid response after an adverse event has occurred, and by tracking and providing feedback about adverse events. However, one should also mention some concerning reports of multiple errors actually introduced by IT systems themselves [12]. For risk and safety management, ICT applications have a certain "Janus" characteristic: On the one hand, they develop into the key tool to improve safety in health systems. On the other hand, they themselves may become the cause of pertinent risks.

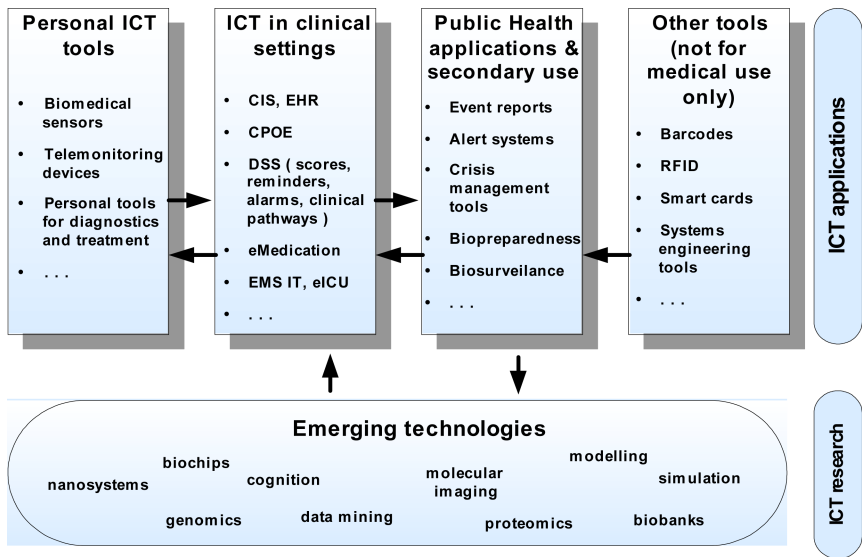


Fig. 2. ICT in support of patient safety and risk management in healthcare (Source: *empirica, eHealth for Safety study*, 2005)

Fig. 2 provides an overview of the areas where ICT can support patient safety and risk management. The *eHealth for Safety* study is currently reviewing the state of the

art in some of these application fields, and will outline major opportunities and challenges as well as to identify in a next step important research aspects and innovative approaches to address patient safety issues. In many of the fields in the Fig. 2, only experimental or pilot systems and applications are as-of-yet available. In this paper we briefly review only a few well known ICT tools and some emerging technologies.

One of the most important developments in eHealth in recent years in many countries has been the ongoing spread of activities concerned with the implementation of *Electronic Health Records* (EHR) on the national, regional and local level. The IOM has advised that moving from a paper to an electronic based patient record system would be the single step that would most improve patient safety. In UK, the National Programme for Information Technology in the NHS being delivered by the Department's agency, NHS Connecting for Health, has begun to roll out its National Care Record system and expects it to have full functionality by 2010. An evaluation of the activities conducted so far in the UK states that "the National Care Record has significant potential to improve safety as lost or poorly completed records are a major contributory factor to patient safety incidents." [10] It is likely that these large scale developments of eHealth infrastructure in many countries will lead to broader implementation of other well known ICT tools, like the ones addressed below.

According to Coiera et al. [13] there is a clear consensus that the use of *Decision Support Systems* (DSS) can improve patient outcomes and make clinical services more effective. DSS are broad solutions, which are often incorporated in a variety of eHealth applications. They go back as far as 1974, and evidence indicates that they can indeed enhance clinical performance for drug dosing, preventive care and other aspects of care, but so far not really convincingly for diagnoses. This is the main finding of Hunt et al.'s review [14]. Several other reviews of the evidence collected so far have taken place. A study by Sintchenko et al. [15] notes that the use of DSS plus microbiology report improved the agreement of decisions by clinicians with those of an expert panel from 65% to 97% ($p=0.0002$) or to 67% ($p=0.02$) when only antibiotic guidelines were accessed.

In their assessment of computer-based cardiac care suggestions Tierney et al. [16] found that the intervention had no effect on physicians' adherence to care suggestions. Physicians viewed guidelines as providing helpful information but containing their practice. They suggest that future studies must weigh the costs and benefits of different (perhaps more draconian) methods of affecting clinician behaviour. Rousseau et al. [17] report primarily negative comments about a DSS. The three main concerns voiced by clinicians were: timing of the guideline trigger, ease of use of the system, and helpfulness of the content.

In Garg et al.'s [18] systematic review of controlled trials of DSSs, about two thirds of these are effective at narrowing knowledge gaps, improving decisions, clinical practice or patient outcomes, but many are not (e.g. computer-based guidelines on the management of angina and asthma) [19]. Ash et al. [20] identify instances where DSS (or patient care information systems, PCIS, as they call it) foster errors rather than reducing them. They distinguish between errors in the process of entering and retrieving information, and errors in the communication and coordination process.

A recent report by Kawamoto et al. [21] reviewed seventy studies and concluded that decision support systems significantly improved clinical practice in 68% of trials. Most notably, 75% of interventions succeeded when the decision support was provided to clinicians automatically, whereas none succeeded when clinicians were required to seek out the advice of the DSS. Similarly, systems that were provided as an integrated component of charting, or order entry systems, were significantly more likely to succeed than stand alone systems.

Coiera et al. [13] conclude that “the use of clinical decision support systems (CDSS) can improve the overall safety and quality of healthcare delivery, but may also introduce machine-related errors. Recent concerns about the potential for CDSS to harm patients have generated much debate, but there is little research available to identify the nature of such errors, or quantify their frequency or clinical impact.”

Computerized Physician Order Entry systems (CPOE) have received considerable attention in the USA as a key technology to help realize the goal of reducing medical errors. CPOEs are defined as a process whereby the instructions of physicians regarding the treatment of patients under their care are entered electronically and communicated directly to responsible individuals or services [22]. Clinical decision support systems are built into almost all CPOE systems to varying degrees, providing basic computerised advice regarding drug doses, routes and frequencies, as well as more sophisticated data such as drug allergy, drug-laboratory values, drug-drug interactions, checks and guidelines [23]. CPOE are applied in a variety of physical and technical environments using currently available vendor software but CPOE is also very resource-intensive, time consuming, and expensive.

Proponents of CPOE systems argue that they have led to reductions in transcription errors, which in turn have led to demonstrable improvements in patient safety. Furthermore, CPOE systems that include data on patient diagnoses, current medications, and history of drug interactions or allergies can significantly reduce prescribing errors [24]. CPOE systems also improve the quality of care by increasing clinician compliance with standard guidelines of care, thereby reducing variations in care.

From four studies on CPOE with DSS, analysed by Kaushal and Bates [25] - three of which were conducted at Brigham and Women’s Hospital (BWH) - the first study (from BWH) found a 55% decrease in serious medication error. As a secondary outcome this study found a 17% decrease in preventable Adverse Drug Events (ADE). In their analysis of CPOE implementations Sittig and Stead [26] point out that key ingredients must be present for a system to work. These include: the system must be fast and easy to use, the user interface must behave consistently in all situations, the institutions must have broad and committed involvement and directions by clinicians prior to implementation, the top leadership of the organisation must be committed to the project and a group of problem solvers and users must meet regularly to work out procedural issues.

However, some authors have also drawn attention to the potential danger of CPOE use. Studies in the US, UK and Australia have found that “commercial prescribing systems often fail to uniformly detect significant drug interactions, probably because of errors in their knowledge base. Electronic medication management systems may generate new types of error because of user-interface design, but also because of

events in the workplace such as distraction affecting the actions of system users.” [22] Han et al. [27] recently reported about an unexpected increase in child mortality coincident with CPOE implementation. While the exact reason for this correlation remains unclear, it underlines that institutions should evaluate mortality effects, in addition to medication error rates, for patients who are dependent on time-sensitive therapies.

Whereas CPOE systems aim to prevent errors, *computerized adverse event systems* aim to monitor the occurrence of instances that could be adverse events and alert the clinician when certain indicators are present. The most common adverse events are nosocomial infections and Adverse Drug Events (ADE) and consequently IT systems have been tested primarily in these areas. [28] Most institutions use spontaneous incident reporting (relies exclusively on voluntary reports from nurses, pharmacists and physicians focused on direct patient care) to detect ADEs; however, this method is generally regarded as rather ineffective and only identifies about one in 20 ADEs.

Conversely, most IT trials have found a significant increase in the number of ADEs reported. *Automatic alerts* can also reduce the time until treatment is ordered for patients with critical laboratory results. [29] This already works well for some types of adverse events, including adverse drug events and nosocomial infections, and are in routine use in some hospitals. In addition, these techniques seem to be well adaptable for the detection of broad arrays of adverse events, in particular as more information becomes computerised.

In their review Gandhi and Bates [30] report one study demonstrating significant decreases in adverse clinical outcome with alert systems, in particular regarding allergic reactions. Significant improvements in response times concerning lab values were reported by several studies, and one study reported significant decrease in the risk of serious renal impairment. Furthermore, noteworthy changes in physician behaviour and modification of therapy based on alerts with recommended actions were reported.

On a larger scale, several countries have already implemented or are considering *national or regional incident or event reporting system* (a concept that is also used in a variety of non-health related areas). By accumulating patient data from a variety of local sources such systems can be used for biosurveillance, such fast alert and pattern tracking as in case of a bioterrorism attack or an epidemic outbreak. In Australia, for instance, an incident reporting system – AIMS – was already set up in 1987, initially only in the field of anaesthesia. [31] Until 1992, 2000 incidents had been collected and reviewed, leading to significant changes at the local and national level.

Ideally, these and other applications will become part of an integrated system, for instance a combination of DSS, CPOE and alerting. Actually, in some cases such integration has already been achieved.

Research has also shown how important it is to design systems with the end-user, the clinician, in mind. If systems are not fast and do not display all relevant information in a coherent and easy to use manner, they will be rejected by the professionals and can even lead to more errors, not less. As Coiera et al. conclude, a deeper understanding of the “complex set of cognitive and socio-technical interactions” can result in the “design of systems which are not just intrinsically ‘safe’ but which also result in safe outcomes in the hands of busy or poorly resourced

clinicians.” Furthermore, the organisational culture, including barriers to reporting errors, will play a key role in the acceptance of electronic tools such as incident reporting systems.

Some Research Trends and Emerging Technologies

New and developing technologies also have a significant patient safety component, either because they pose risks or because they may offer benefits in their application to patient safety – or both. In this section we provide an overview of such emerging technologies and their (potential) application to patient safety and risk management in healthcare.

Towards a culture of safety in eHealth RTD

Whereas eHealth tools and services are intended to have a beneficial impact on citizens' health, recent research has shown that some of these tools and services may under certain circumstances also be potentially harmful to citizens' health. New technologies inherently pose new risks. Health risk and patient safety aspects should therefore be taken into account by all health ICT RTD from EHR integration, home monitoring and assistive living, to bio-medical informatics, nano-devices and grid computing.

Data mining for improved patient safety

Data mining techniques can be applied to emerging electronic health record and clinical research databases to push forward knowledge of risks associated with unique patient characteristics and treatment patterns. Such tools need to be developed to discover, for example, instances where patient safety has been endangered, and identify the causes.

Ontology of patient safety

For exchanging information on patient safety, as a common framework for modelling threats to safety and to support communication between clinicians and others on patient safety issues, a taxonomy and ontology covering healthcare risks and safety considerations should be developed.

Mathematical modelling and simulation

Modelling and simulation tools are anticipated to have significant impact on patient safety especially through advancing prediction, prevention and personalisation of healthcare. The European Information Society Technologies Advisory Group (ISTAG) proposed in 2004/2005 to stimulate research in the area of “The Disease and Treatment Simulator”: a computational platform for simulating the function of a concrete disease. [32] “This simulator will enable medicines to be tested without putting people at risk, and will accelerate research into damaging diseases such as heart disease and cancer.” The Advisory Group also suggested that the disease and treatment model should interface directly with other projects of human benefit, such as the *Physiome* project [33] and the modelling of whole organs. In this context the European Commission (EC) is supporting research on the *Virtual Physiological*

Human (VPH) which is expected to accelerate knowledge discovery leading to improved disease prevention, early diagnosis and individuals' health risk management. [34] To reduce risks to citizens participating in clinical research, to enable a radical expansion of the volume of research into clinical outcomes to the full range of treatments and to significantly accelerate production of results from clinical research it appears important to support research into tools to implement virtual clinical trials. According to the *Academy of Medical Sciences* in UK [35], "sophisticated modelling has great potential and it is possible to envisage a time when models could be used to test a greater range of possible situations than it is practical to address in affordable clinical trials" which also "permits the evaluation of heterogeneity and the active exploration of those who may be at risk." Using simulation has already enabled pharmaceutical companies to eliminate four-fifths of a clinical trial, reducing the total number of recruited patients by 60% and shortening the trial's duration by 40%. "Virtual patient" engines [36] are helping researchers and physicians select the best among existing therapies, e.g., for breast cancer, and to develop optimal dosing regimes. So-called "computer-assisted trial design" systems - a field in which models have become so useful that the FDA itself is adopting them [37] - model and simulate clinical trials to determine the optimal number of patients, dose amounts, and dosing frequency, all of which have for years mostly been determined through time-consuming and costly trial and error.

Medical simulation and virtual reality

This is already being used as a training and feedback method in which learners practice tasks and processes in lifelike circumstances using models or virtual reality (VR), with feedback from observers, peers, actor-patients, and video cameras to assist improvement in skills. Medical simulators allow individuals to review and practice procedures as often as required to reach proficiency without harming patients. VR simulations are revolutionising surgical training [38] (e.g., for laparoscopic, gastrointestinal, plastic, ophthalmological, dermatological, and some laryngological procedures), and error reporting [39] in the healthcare field.

Pathways and health pathway risk models

Pathways are generally multidisciplinary by design and may incorporate the responsibilities of physicians and nurses with those of ancillary medical providers including pharmacists, physical therapists and social workers. In the future, it may be possible to build health pathway models which encompass citizen / patient passage through clinical pathways, with predictive ability, focussing on the prior identification of potential risks to a citizen's future health.

Socio-economic and behavioural aspects

Research into how eHealth applications and the concomitant re-engineering of healthcare processes may change the behaviour of health professionals, care personnel, citizens and patients to improve system safety and performance is a promising field. This should also involve analysing the impact of medico-cultural, legal/regulatory and socio-economic factors. Assessing the risk and developing guidelines and certification procedures for Decision Support and Expert Systems and other tools need also to be mentioned here.

Monitoring and risk management of large-scale events

Research into strategies and ICT support for preparedness for large-scale events like pandemics (e.g. avian flu) or bio-terrorism attacks (e.g. epidemiological modelling of regional events) is an important challenge. It may allow a better response to threats through better information but also could play a key role in resource planning and management. ICT should also be exploited as a means to inform and reach professionals and the public on a large scale and help adapt responses. The use of Geographical Information Systems in healthcare appeared recently as a promising field and research should be conducted involving epidemiologists, managers of health resources and policy makers.

Summary - Key Issues for a Research Roadmap

To summarise our initial discussions, Fig. 3 illustrates and delimits an initial model for the patient and health system risk domain. This model will not only allow for different types of risks and ICT applications relevant to improve safety management to be related to the corresponding meta categories, but it may also direct the research towards other innovative fields which may be critical and important.

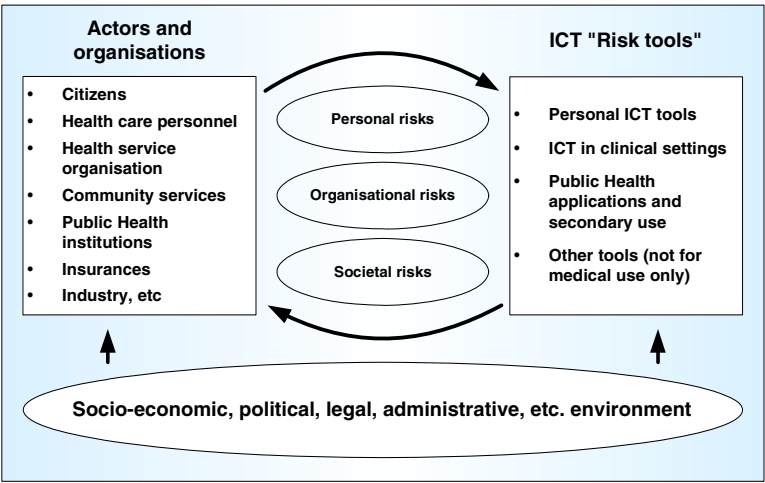


Fig. 3. ICT in support of patient safety and risk management in healthcare (*Source: empirica, eHealth for Safety study, 2005*)

Our recent research showed that it is vital not only to look at the issues from their technical point of view but also to take organisational and political factors into account, which will play a key role if patient safety is to be strengthened. The following table gives an overview of the components of each level:

Table 1. Components in a multi-level approach to patient safety

Level	Component
Policy level (regional, national, European level)	• Patient safety policies • Implementation measures • Socio-economic and health policy framework conditions • Legal and ethical issues • Funding, clinical and economic evaluation
Organisational level	• Organisational structure and culture • Work processes • Change management • Training and learning
Technical & RTD level / applications	• Personal ICT tools, e.g., biomedical sensors • ICT in clinical settings, incl. EHR, DSS, CPOE • Public health applications & secondary use, e.g., event reporting, alert systems • Semantic aspects / ontologies • Emerging technologies

In its work the *eHealth for Safety* study is applying this multilevel analysis of ICT in patient safety in order to arrive at a vision and roadmap for future research which will ultimately benefit European citizens and healthcare providers.

Acknowledgements. This paper was written as part of the *eHealth for Safety - Study on the impact of ICT on patient safety and risk management in healthcare* (www.ehealth-for-safety.org) commissioned by the European Commission, Directorate General Information Society and Media, Brussels. This paper reflects solely the views of its authors. The European Community is not liable for any use that may be made of the information contained therein.

References

1. IOM Report (2000): To err is human: Building a safer health system. Institute of Medicine, 287 p. Available at: <http://books.nap.edu/books/0309068371/html/index.html>.
2. IOM Report (2001): Crossing the Quality Chasm: A New Health System for the 21st Century, Institute of Medicine, 364 p., <http://books.nap.edu/catalog/10027.html>
3. Commission on Systemic Interoperability (2005): Ending the Document Game: Connecting and Transforming Your Healthcare Through Information Technology, U.S. Government Printing Office (GPO), Washington, 2005, 249 p. <http://endingthedocumentgame.gov>
4. Young, S.: The Role of Health IT in Reducing Medical Errors and Improving Healthcare Quality & Patient Safety. Agency for Healthcare Research and Quality. August 2005. http://www.ehealthinitiative.org/assets/documents/Capitol_Hill_Briefings/Young9-22-04.PPT
5. JRC/IPTS (2004): eHealth in the Context of a European Ageing Society. A Prospective Study. P.17.
6. European Commission (2003): The Social Situation in the European Union 2003. Luxembourg: Office for Official Publications. P. 69.

7. Proctor P. R. et al., Editors, (2005): Building a Better Delivery System: A New Engineering/Health Care Partnership. Committee on Engineering and the Health Care System, National Academies Press, 276 p., <http://www.nap.edu/catalog/11378.html>
8. Wachter, R (2004): The End Of The Beginning: Patient Safety Five Years After 'To Err Is Human'. Health Affairs Web Exclusive. W4- 539.
9. Milstein A. et al.: "Improving the Safety of Health Care: The Leapfrog Initiative," Effective Clinical Practice 3, no. 6 (2000): 313–316; and N. Versel, "Performance Driving Investment Up," Modern Physician (November 2003): 15, 23.
10. NAO (National Audit Office) (2005): "A Safer Place for Patients: Learning to improve patient safety", Department of Health, 86 p., www.nao.org.uk/publications/nao_reports/05-06/0506456.pdf
11. Lessens, V., Lloyd-Williams, D. (2004): Workshop on Risk Management for Health Professionals. Use of ICT, Workshop report, p. 14.
12. Ash, J.S., M. Berg, and E. Coiera, (2004): "Some Unintended Consequences of Information Technology in Health Care: The Nature of Patient Care Information System–related Errors," Journal of the American Informatics Association, 11, no. 2 (2004): 104–112.
13. E. Coiera, J.I. Westbrook, J.C. Wyatt (2006): The Safety and Quality of Decision Support Systems. In Haux R, Kulikowski C. (ed.) IMIA Yearbook of Medical Informatics 2006. Methods Inf Med 2006; 45 Suppl 1
14. Hunt et al. (1998): Effects of computer-based clinical decision support system on physicians performance and patient outcomes: a systematic review. JAMA 280: 1339-1346
15. Sintchenko et al. (2004): Comparative impact of guidelines, clinical data and decision support on prescribing decision: an interactive web experiment with simulated cases. J Am Med Inform Assoc 11 (1) 71-7.
16. Tierney et al (2003): Effects of computerized guidelines for managing heart disease in primary care. J Gen Intern Med 18 (12): 967-76.
17. Rousseau et al (2003): Practice based, longitudinal, qualitative interview study of computerised evidence based guidelines in primary care. BMJ. 2003 Feb 8;326(7384):314.
18. Garg AX, Adhikari NK, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene S, Sam J, Haynes RB. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA. 2005;293:1223–1238.
19. Eccles, M. et al. (2002): Effect of computerised evidence based guidelines on management of asthma and angina in adults in primary care: cluster randomised controlled trial. BMJ. 2002;325:941–944.
20. Ash et al. (2004): Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. Journal of the American Medical Informatics Ass. 11:104-112
21. Kawamoto et al. (2005): Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ, April 2, 2005; 330(7494): 765.
22. FCG (2003): Computerized Physician Order Entry: Costs, Benefits and Challenges. A Case Study Approach
23. Bonnabry, P. (2003): Information Technologies for the Prevention of Medication Errors. Business Briefing: European Pharmacotherapy 2003 1-5.
24. Bates, D.W. et al. (1998): Effect of computerized physician order entry and a team intervention on prevention of serious medication errors. Journal of the American Medical Association 280(15): 1311–1316. Leapfrog Group. 2000. Leapfrog Patient Safety Standards: The Potential Benefit of Universal Adoption. <http://www.leapfroggroup.org>.

25. Kaushal Rainu, Bates, D. W. (2003): Computerized Physician Order Entry (CPOE) with Clinical Decision Support Systems (CDSSs). In: Making Health Care Safer: A Critical Analysis of Patient Safety Practices. Evidence Report/Technology Assessment: Number 43. AHRQ Publication No. 01-E058, July 2001. www.ahrq.gov/clinic/ptsafety/, p.59-70
26. Sittig and Stead (1994): Computer based physician Order Entry: the state of the art; in: Journal of the American Medical Informatics Association. 108-123
27. Han, Y. et al. (2005): Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system in: Pediatrics Vol.116 No.6 (12/2005) 1506-1512
28. Bates, D.W. et al. (2003): Detecting Adverse Events Using Information Technology JAMIA 10 p.119
29. Kuperman et al. (1999): Improving Response to Critical Laboratory Results with Automation: Results of a Randomized Controlled Trial. JAMIA 6 512-22
30. Gandhi T. K., Bates D. W. : Computer Adverse Drug Event (ADE) Detection and Alerts. In: Making Health Care Safer: A Critical Analysis of Patient Safety Practices. p.81
31. Runciman W.B. (2002) Lessons from the Australian Patient Safety Foundation: setting up a national patient safety surveillance system – is this the right model? Quality and Safety in Health Care 11/2002: 250
32. Information Society Technologies Advisory Group (ISTAG), (2004): “Grand Challenges in the Evolution of the Information Society”, W. Wahlster (ed.) p. 26-29
33. IUPS Physiome Project Roadmap (2005), www.physiome.org.nz/roadmap/roadmap-mar05
34. Towards Virtual Physiological Human (2005): Multilevel modelling and simulation of the human anatomy and physiology, White Paper, http://europa.eu.int/information_society/activities/health/docs/events/barcelona2005/ec-draft-vph-white-paper-v2.8.pdf, p. 3
35. Safer Medicines, Academy of Medical Sciences (2005) A report from the Academy's FORUM with industry, <http://www.acmedsci.ac.uk/p99puid61.html>, p. 22
36. Agur, Z. (2006): "Biomathematics in the development of personalized medicine in oncology" Future Oncology, Feb 2006, Vol. 2, No. 1, pp 39-42.
37. Models that take drugs. Biosimulation: Designing drugs in computers is still some way off. But software is starting to change the way drugs are tested, The Economist, June 9th 2005
38. Gorman PJ, Meier AH, Krummel TM (2000): Computer-assisted training and learning in surgery. Comput Aided Surg 2000;5:120–30.
39. Fried M P et al. (2004): Identifying and reducing errors with surgical simulation, Qual Saf Health Care 2004;13(Suppl 1):i19–i26. doi: 10.1136/qshc.2004.009969, http://qhc.bmjournals.com/cgi/content/full/13/suppl_1/i19

Author Index

- Aerts, Jean-Marie 285
Alonso, Fernando 311
Álvarez, Francisco 207
Analyti, Anastasia 250
Anguita, Alberto 262
Arrais, Joel 231
- Baik, Doo-Kwon 463
Beigi, Majid 25, 104
Berberidis, Christos 92
Berckmans, Daniel 285
Bezerianos, Anastasios 161, 323
Blanchet, Christophe 240
Blanquer, Ignacio 183
Boemi, Massimo 128
Boniatis, Ioannis 451
Brause, Rüdiger 441
Browne, Allen 472
Burattini, Roberto 128
- Casagrande, Fabrizio 128
Cavouras, Dionisis 451
Chatzizisis, Yiannis S. 368
Cho, Sung-Bum 37
Choi, O-Hoon 463
Choi, Tae-Sun 346
Combet, Christophe 240
Costaridou, Lena 451
Crespo, Jose 262
- Daric, Vladimir 240
de Buenaga, Manuel 207
Deléage, Gilbert 240
Di Nardo, Francesco 128
Diez, Raquel Montes 334
Drakos, John 150
Dubitzky, Werner 116
- Erlich, Henry A. 1
- Floros, Xenofon E. 390
- Gao, Xiaofeng 49
Gebhardt, Rolf 137
Gerlach, Joerg C. 137
- Giannoglou, George D. 368
Godlewski, Grzegorz 273
Guthke, Reinhard 137
- Hagihara, Kenichi 195
Hernández, Juan A. 334
Hernandez, Vicente 183
Hraber, Peter T. 1
Huss, Harold 423
- Ino, Fumihiko 195
- Jagadish, M. 402
- Kalaitzakis, Manos 250
Kalatzis, Ioannis 451
Kapela, Adam 161
Karakantza, Marina 150
Kepler, Thomas B. 1
Keselman, Alla 472
Kim, Eun-Young 423
Kim, Ju Han 37
Kokkinos, Vasileios 323
Kompatsiaris, Yiannis 368
Kondylakis, Haridimos 250
Korber, Bette T. 1
Kouidou, Sofia 60
Koutkias, Vassilis 60, 368
Krakow, Karsten 441
Krause, Antje 116
Kugiumtzis, Dimitris 298
Kurzynski, Marek 83
- Lakoumentas, John 150
Larsson, Pål G. 298
Lee, Sang-Ho 37
Lim, Jung-Eun 463
- Maglaveras, Nicos 60, 368
Malik, Aamir Saeed 346
Malousi, Andigoni 60
Manakanatas, Dimitris 250
Martínez, Loïc 311
Massengale, Lisa 472
Matsuo, Katsunori 195

- May, Michael 219
 Mentzer, Steven J. 423
 Meyfroidt, Geert 285
 Micheloyannis, Sifis 172
 Mizutani, Yasuharu 195
 Moraru, Liviu 323
 Morosini, Pierpaolo 128
- Na, Hong-Seok 463
 Ngo, Long 472
 Nikiforidis, George 150
 Nikita, Konstantina S. 390
 Nugent, Chris 116
- Oikonomou, Theofanis 172
 Oliveira, José Luis 231
- Paetz, Jürgen 72, 378
 Pak, Jane 423
 Panagiotopoulos, Elias 451
 Panayiotakis, George 451
 Papana, Angeliki 298
 Parissi, Eirini 368
 Patnaik, L.M. 402
 Pelekouda, Polyxeni 323
 Pérez, Aurora 311
 Perez-Rey, David 262
 Pfaff, Michael 137
 Piasecki, Maciej 273
 Pless, Gesine 137
 Plexousakis, Dimitris 250
 Polónia, Daniel F. 231
 Potamias, George 219, 250
 Pratt, Juan Pablo 423
- Quirós, Alicia 334
- Ravnic, Dino 423
 Reczko, Martin 13
 Rüping, Stefan 219
- Sáenz, Fernando 207
 Sakellaropoulos, George 150
 Sakkalis, Vangelis 172
 Santamaría, Agustín 311
 Schaefer, Gerald 358
- Schmidt-Heck, Wolfgang 137
 Segrelles, Damià 183
 Spichtinger, Daniel 482
 Spyrou, George M. 390
 Srinivasa, K.G. 402
 Starosolski, Roman 358
 Stavrinou, Maria L. 323
 Strintzis, M.G. 368
 Stroetmann, Karl A. 482
 Stroetmann, Veli N. 482
 Symeonidis, Alkiviadis 13
- Thierry, Jean Pierre 482
 Tollis, Ioannis G. 13, 172
 Trachtenberg, Elizabeth A. 1
 Tsangaris, George T. 390
 Tsimpiris, Alkiviadis 298
 Tsoukias, Nikolaos 161
 Tzanis, George 92
- Ünlü, Atilla 441
- Valente, Juan Pedro 311
 Van den Berghe, Greta 285
 Van Loon, Kristien 285
 Vaquero, Antonio 207
 Venugopal, K.R. 402
 Vlachos, Ioannis 298
 Vlahavas, Ioannis 92
 Vougas, Konstantinos N. 390
- Wan, Baikun 49
 Wang, Chong 116
 Wolinsky, Steven 1
 Woloszynski, Tomasz 83
 Wozniak, Michal 433
- Yang, Chunmei 49
 Yoon, Hye-Sung 37
- Zeilinger, Katrin 137
 Zell, Andreas 25, 104
 Zeng, Qing 472
 Zeng, Qing T. 423
 Zolnierrek, Andrzej 413
 Zoumbos, Nicolaos 150