

SOCIETÀ ITALIANA DI FISICA

PROCEEDINGS
OF THE
INTERNATIONAL SCHOOL OF PHYSICS
«ENRICO FERMI»

COURSE CLXV

*Protein Folding and Drug
Design*



SOCIETÀ ITALIANA DI FISICA BOLOGNA, ITALY

This page intentionally left blank

SOCIETÀ ITALIANA DI FISICA

RENDICONTI
DELLA
SCUOLA INTERNAZIONALE DI FISICA
“ENRICO FERMI”

CLXV CORSO

a cura di R. A. BROGLIA and L. SERRANO

Direttori del Corso

e di

G. TIANA

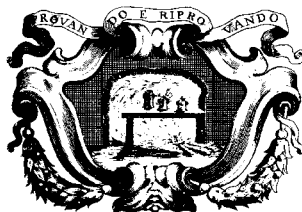
VARENNA SUL LAGO DI COMO

VILLA MONASTERO

4 – 14 Luglio 2006

*Avvolgimento di proteine e
disegno di farmaci*

2007



SOCIETÀ ITALIANA DI FISICA
BOLOGNA-ITALY

ITALIAN PHYSICAL SOCIETY

PROCEEDINGS
OF THE
INTERNATIONAL SCHOOL OF PHYSICS
“ENRICO FERMI”

COURSE CLXV

edited by R. A. BROGLIA and L. SERRANO

Directors of the Course

and

G. TIANA

VARENNA ON LAKE COMO

VILLA MONASTERO

4 – 14 July 2006

*Protein Folding and Drug
Design*

2007

IOS
Press

AMSTERDAM, OXFORD, TOKIO, WASHINGTON DC

Copyright © 2007 by Società Italiana di Fisica

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or any means, electronic, mechanical, photocopying, recording or otherwise, without the prior permission of the copyright owner.

ISBN 978-1-58603-792-5 (IOS)

ISBN 978-88-7438-038-1 (SIF)

Library of Congress Control Number: 2007935265

Production Manager
A. OLEANDRI

Copy Editor
M. MISSIROLI

jointly published and distributed by:

IOS PRESS
Nieuwe Hemweg 6B
1013 BG Amsterdam
The Netherlands
fax: +31 20 620 34 19
order@iopress.nl

SOCIETÀ ITALIANA DI FISICA
Via Saragozza 12
40123 Bologna
Italy
fax: +39 051 581340
order@sif.it

Distributor in the UK and Ireland
Gazelle Books Services Ltd.
White Cross Mills
Hightown
Lancaster LA1 4XS
United Kingdom
fax: +44 1524 63232
sales@gazellebooks.co.uk

Distributor in the USA and Canada
IOS Press, Inc.
4502 Rachael Manor Drive
Fairfax, VA 22032
USA
fax: +1 703 323 3668
iosbooks@iospress.com

Proprietà Letteraria Riservata
Printed in Italy

Supported by

Istituto Nazionale di Fisica Nucleare (INFN)

Università degli Studi di Milano

Regione Lombardia

INFN, Sezione di Milano

This page intentionally left blank

INDICE

R. A. BROGLIA, L. SERRANO and G. TIANA – Preface	pag. XIII
Gruppo fotografico dei partecipanti al Corso	» XVIII
H. A. SCHERAGA – Experimental and theoretical aspects of protein folding and hydrophobic interactions	» 1
1. Introduction	» 1
2. Experimental physical-chemical determination of structure	» 2
3. Experimental physical-chemical determination of oxidative folding pathways of RNase A	» 4
4. Experimental physical-chemical determination of folding pathways of disulfide-intact RNase A	» 9
5. Theoretical determination of structure with an all-atom approach	» 11
6. Theoretical determination of structure with a United-Residue (UNRES) Model	» 11
7. Molecular dynamics with the UNRES force field	» 15
8. Conclusions	» 15
Appendix - Hydrophobic interactions	» 18
P. G. WOLYNES – Lectures on biomolecular energy landscapes	» 27
Introduction	» 27
1. The basis of folding landscapes	» 28
2. Random sequences	» 28
3. The statistical energy landscape	» 29
4. The energy landscape of long evolved proteins	» 33
5. Local <i>vs.</i> global descriptions of the folding landscape	» 40
K. B. ZELDOVICH, I. N. BEREZOVSKY and E. I. SHAKHNOVICH – How organisms adapt to the environment: Sequence determinants of the habitat temperature and its physical rationale	» 47
1. Introduction	» 48
2. Results	» 50
2'1. Sequence determinants of thermal adaptation of soluble proteins ...	» 50
2'2. Statistical tests and controls	» 51

2'3. Membrane proteins and specific folds	»	54
2'4. IVYWREL is not a consequence of nucleotide composition bias	»	56
2'5. Thermal adaptation of DNA	»	57
2'6. Purine loading bias is mainly due to IVYWREL	»	57
2'7. Nearest-neighbor correlation in DNA sequences	»	60
3. Physical principles of protein design and the origin of IVYWREL	»	62
3'1. Discussion	»	72
4. Methods	»	77
4'1. Methods	»	77
R. A. BROGLIA, G. TIANA, D. PROVASI and L. SUTTO – Molecular recog- nition through local elementary structures: Transferability of simple models to real proteins	»	85
1. Introduction	»	85
2. The lattice model	»	88
2'1. Inverse folding problem: the design of good folders	»	89
2'2. Role of the different amino acids in the folding process	»	89
2'3. Extension of the inverse folding strategy	»	92
2'4. How many mutations can a designed protein tolerate?	»	93
3. Hierarchical folding of a model protein	»	103
4. Solving the protein folding problem in the case of a notional protein (three- step-strategy (3SS))	»	104
5. Lattice model design of resistance proof, folding-inhibitor peptides	»	106
6. Drug resistance	»	110
7. Design and folding of dimeric proteins	»	111
8. Conclusions	»	112
G. M. VERKHIVKER – Exploring mechanisms of protein folding and binding in signal transduction networks	»	115
1. Introduction	»	115
2. Simulation methods	»	120
2'1. Monte Carlo simulations: the energy model	»	120
2'2. Monte Carlo simulations: simulated tempering dynamics	»	122
2'3. Binding free energy calculations	»	122
3. Results and discussion	»	123
J. R. BANAVAR, T. X. HOUNG, A. MARITAN and A. TROVATO – Spin glasses, tubes and proteins	»	135
C. CECCONI, E. A. SHANK, S. MARQUSEE and J. BUSTAMANTE – Studying protein folding with laser tweezers	»	145
1. Introduction	»	145
2. Laser tweezers	»	147

3. Synthesis of molecular constructs for use in mechanical manipulation studies	»	150
4. Mechanical manipulation of single RNase H molecules by laser tweezers . .	»	151
5. Conclusion	»	159
G. TIANA, D. PROVASI, A. AMATORI, L. SUTTO and R. A. BROGLIA – Oligarchy in protein folding: The upper and lower classes in protein chains .	»	161
1. History of the hierarchical view of folding	»	161
2. The folding of the SH3 domain: A computational model	»	165
3. Equilibrium thermodynamics	»	167
4. Folding kinetics	»	168
5. The uneven distribution of energy	»	171
6. Inhibition of the folding of SH3	»	172
7. Conclusions	»	174
W. F. VAN GUNSTEREN, D. P. GEERKE, D. TRZESNIAK, C. OOSTENBRINK and N. F. A. VAN DER VEGT – Analysis of the driving forces for biomolecular solvation and association	»	177
1. Introduction	»	177
2. Methods	»	180
3. Results	»	181
3'1. Solvation	»	181
3'1.1. Solvation of polar and apolar solutes in polar and apolar pure solvents	»	181
3'1.2. Solvation of methane in aqueous solutions of different co- solvents	»	181
3'1.3. Solvation of apolar solutes in aqueous solutions of different co-solvents	»	184
3'2. Association	»	184
3'2.1. Association of two hydrophobic or hydrophilic solutes in pure water	»	184
3'2.2. Association of two hydrophobic solutes in water and in aqueous urea	»	188
3'2.3. Association of many hydrophobic solutes in water and in urea solutions	»	189
4. Conclusions	»	190
K. RAHA and K. M. MERZ jr. – Structural basis of dielectric permittivity of proteins: Insights from quantum mechanics	»	193
1. Introduction	»	194
2. Results	»	195
3. Materials and methods	»	200
4. Conclusions	»	203

P. J. WINN, M. ZAHARAN and R. C. WADE – Comparisons of protein electrostatic potentials as a means to understanding function	»	207
1. Introduction	»	207
2. Calculation of electrostatic potentials in biomolecular systems	»	208
3. Examples of the use of electrostatic potentials in biomolecular systems ...	»	209
4. Comparison of protein electrostatic potentials: Protein Interaction Property Similarity Analysis (PIPSA)	»	210
5. Electrostatic potentials in the ubiquitin and ubiquitin like systems.....	»	211
6. High-throughput modelling of protein electrostatic potentials	»	216
7. Conclusions	»	217
 G. M. VERKHIVKER – Computational structural proteomics of the kinases binding specificity and drug resistance	»	221
1. Introduction	»	221
2. Materials and methods	»	225
2'1. Structural classification of protein tyrosine kinases	»	225
2'2. A hierarchical model of biomolecular recognition	»	225
2'3. Monte Carlo binding simulations: simulated tempering dynamics ...	»	226
3. Results and discussion	»	227
4. Conclusions	»	235
 G. COLOMBO – Blocking the protein folding machinery. Rational design of inhibitors of the molecular chaperone Hsp90 as new anticancer agents	»	239
1. Materials and methods	»	240
1'1. Peptide molecular-dynamics simulations	»	240
1'2. Docking procedure	»	242
1'3. Pharmacophore generation	»	243
2. Results	»	245
2'1. Shepherdin and Shepherdin RV	»	245
2'2. Simulations of Shepherdin-RV mutants	»	245
2'3. Shepherdin[79-83]	»	246
2'4. Characterization of Hsp90/shepherdin binding interface	»	247
2'5. Pharmacophoric hypotheses and small molecule identification	»	248
3. Discussion	»	249
 G. M. VERKHIVKER – Energy landscapes of bimolecular binding and molecular modulators of protein-protein interactions	»	253
1. Energy landscapes and molecular recognition	»	253
2. Binding hot spots and convergent solutions at protein-protein interfaces ..	»	256
3. Targeting P53-MDM2 interfaces with molecular modulators	»	257
4. Materials and methods	»	260
4'1. Structural analysis	»	260
4'2. Hierarchical models of biomolecular recognition	»	261
5. Results and discussion	»	261

5'1. Conformational landscape of MDM2 and specific binding with small molecular mimics	»	261
5'2. The energy landscape analysis of a hot spot at the consensus binding site of the constant fragment (Fc) of human immunoglobulin G	»	264
S. PANTANO, M. BERRERA, C. ANSELMi and P. CARLONI – Tetrameric voltage-gated ion channels investigated by molecular dynamics and bioinformatics	»	273
1. Introduction	»	274
2. Methods	»	276
3. Results	»	276
4. Concluding remarks	»	280
D. PROVASI, G. TIANA and R. A. BROGLIA – Folding inhibition of HIV Protease: An experimental glance	»	283
1. Introduction	»	283
1'1. The Michaelis-Menten framework	»	284
1'2. Spectrophotometric assay	»	285
1'3. Circular Dichroism	»	286
2. Materials and methods	»	287
3. Discussion	»	287
4. Conclusions	»	291
S. RUSCONI, M. LO CICERO, A. E. LAFACE, S. FERRAMOSCA, E. CESANA, G. TIANA, D. PROVASI, F. SIRIANNI, M. GALLI, M. MORONI, A. CLIVIO and R. A. BROGLIA – Susceptibility to a non-conventional (folding) protease inhibitor of human immunodeficiency virus Type 1 isolates <i>in vitro</i>	»	293
1. Introduction	»	293
2. Experimental strategy	»	294
3. Results	»	295
4. Conclusions	»	297
Elenco dei partecipanti	»	301

This page intentionally left blank

Preface

"One of the great unsolved problems of science is the prediction of the three-dimensional structure of a protein from its amino acid sequence: the folding problem". Thus wrote Sir Alan Fehrst, an illustrious scientist from Cambridge, a few years ago. But because, according to another Cambridge scholar, Lord Rutherford : "science is either physics or it is stamp collection", the "protein folding problem" is also one of the great unsolved problems of physics. This is the reason why the Italian Physical Society organized the present "Enrico Fermi" Summer School, on the premises of Villa Monastero, where Enrico Fermi lectured for the last time in Italy (summer 1954) before his untimely death on November 29th of that year.

It may be stated that the deep connection existing between physics and protein folding is not so much, or in any case not only, through physical methods (experimental: X-rays, NMR, etc, or theoretical: statistical mechanics, spin glasses, etc), but through physical concepts. In particular those associated with the transition of many-body (finite) systems between an initial and a final phase implying breaking of symmetry. In fact, protein folding can be viewed as an emergent property not contained either in the atoms forming the protein or in the forces acting among them, in a similar way as superconductivity emerges as an unexpected coherent phenomenon taking place on a sea of electrons at low temperature.

Let us recall that in spite of the fact that one does not yet know how to read the 3D structure of a protein from its 1D structure, much is known about the protein folding problem, thanks, among other things, to protein engineering experiments (φ -values determination, Alan Fehrst and Luis Serrano) as well as from a variety of theoretical inputs: inverse folding problem (Eugene Shakhnovich), funnel-like energy landscapes (Peter Wolynes), helix-coil transitions (Harold Scheraga), etc.

Although quite different in appearance, the fact that the variety of models can account for much of the experimental findings is likely due to the fact that they contain much of the same (right) physics. A physics which is related to the important role played

by selected highly conserved, “hot”, amino acids which participate in the stability of independent folding units which, upon docking, give rise to a (post-critical) folding nucleus lying beyond the highest maximum of the free energy associated to the process. In the same way as Heisenberg (matrix) and Schrödinger (differential equation) versions of quantum mechanics have been shown to contain the same physics, it is highly likely that the physics which is at the basis of the different views presented by the lecturers of the phenomenon of protein folding is, to a large extent, equivalent.

This impression also emerged from the answers given by the lecturers to many questions and comments put forward by the lively group of students which attended the School. Within this context, we want to thank them for their attendance and acknowledge the assiduity of their interventions in terms of questions and comments, the high level of the ten minutes talks many of them gave, as well as the high level of the posters presented.

At the basis of their collective response were the high level of the lecturers and seminar speakers presentations. In particular, Harold Scheraga, Peter Wolynes, Eugene Shakhnovich, Amos Maritan, Luis Serrano, Leonid Mirny and Guido Tiana reminded us that while we still do not know how to solve the “protein folding problem”, one has developed a series of powerful methods (like, *e.g.*, chain initiation folding events, foldons and folding funnels, local elementary structures, etc.) which have shed much light on the mechanism which is the basis of the folding of proteins.

It is remarkable that the read thread going through these concepts and corresponding results, starting from those associated with the simplified lattice models discussed by R. A. Broglia, seem to extend all the way to HIV-1 reproduction in infected cell, opening the way for the development of non-conventional (folding) inhibitors, as was reported by Stefano Rusconi, a non-obvious consequence of the *in vitro* experiments reported by Davide Provasi. A new interdisciplinarity embracing not only physicists, chemists and biologists, but also medical doctors which is likely to be needed to solve such formidable problems as those created by HIV, seems to be in the nascent stage.

The remarkable advances in the field of *ab initio* studies which has taken place during the last years were reported by Michele Parrinello, Wilfred Van Gunsteren, Paolo Carloni and Peter Winn. Peter Wolynes, Gennady Verkhivker and Giorgio Colombo updated students and lecturers alike on the latest developments on drug design and on the many successes as well as challenges facing this exciting field lying at the borderline between pure and applied research. The role quantum mechanics plays within this context, as well as within the framework of protein folding, was discussed by Kenneth Merz.

The School could have not been possible without the indefatigable support of the President of the Italian Physical Society Professor Franco Bassani, who very early in the programming of the Enrico Fermi School realized the relevance, for physicists, of the subject of protein folding and drug design. To him and to the Secretarial staff (Ramona Brigatti and Giovanna Bianchi Bazzi) headed by Barbara Alzani, as well as to Villa Monastero housekeeper, Antonio Cintorino, our warmest thanks.

Aside from the economical support provided by the Italian Physical Society, we acknowledge the support coming from the University of Milan. The presence in the concluding Session of the School of the Vice President of research Prof. Giampiero Sironi

and of the Dean of the Faculty of Sciences, Prof. M. Pignanelli testifies to the importance adscribed by our University to the interdisciplinary field of protein folding and drug design. Within this context, the presence of Prof. Mauro Moroni (Head of the Department of Clinical Sciences, Division of Infective Diseases) and of Prof. Massimo Galli (Director of the Institute of Infective Diseases) of the Faculty of Medicine (and Sacco Hospital), of the University of Milan, was only natural. We also thank the support of INFN (Istituto Nazionale di Fisica Nucleare).

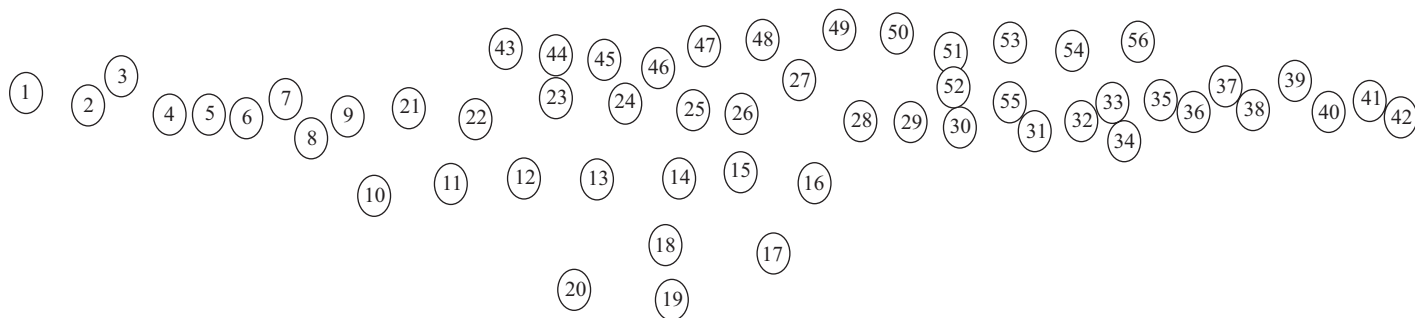
Last, but not least, we acknowledge the privilege of having held this School in the suggestive premises of Villa Monastero, at Varenna, on Como Lake. As one of the lecturers vividly put it, it felt almost surreal to be able to carry out business as usual, namely discuss with each other what we understood, as well as what we do not understand about protein folding, in such beautiful surroundings.

R. A. BROGLIA, L. SERRANO and G. TIANA

This page intentionally left blank

Società Italiana di Fisica
SCUOLA INTERNAZIONALE DI FISICA «E. FERMI»
CLXV CORSO - VARENNA SUL LAGO DI COMO
VILLA MONASTERO 4 - 14 Luglio 2006





1) M. Hatlo	11) G. Verkhivker	21) C. C. Ngang	30) F. Marini	39) S. Trygubenko	48) G. Nico
2) L. Lanzaò	12) A. Maritan	22) J. Siltberg Liberles	31) C. Camilloni	40) L. Nardo	49) L. Mirny
3) A. Trovato	13) G. Tiana	23) G. Grazioso	32) M. Caldanini	41) D. Branduardi	50) R. Travasso
4) P. Cerri	14) L. Serrano	24) F. E. Herrera	33) A. Soranno	42) M. Bonomi	51) I. Moreira
5) L. Sutto	15) P. Wolynes	25) M. Zamparo	34) T. Bogdan	43) I. Mesropyan	52) A. Sahakyan
6) D. Provasi	16) R. Broglia	26) M. Rakonjac	35) S. Luccioli	44) S. Khatuntseva	53) S. Kmiecik
7) S. Yesylevskyy	17) G. Bianchi Bazzi	27) A. Barducci	36) C. Guardiani	45) M. Cotallo Alban	54) M. Kurcinski
8) E. Mirmontaz	18) R. Brigatti	28) V. Vetri	37) M. Buscaglia	46) E. F. De Sousa Henriques	55) F. Pullara
9) Y. Ivarsson	19) B. Alzani	29) L. Dandrea	38) O. Obolensky	47) F. Neri	56) S. Mitternacht
10) P. De Los Rios	20) F. Franciamore				

Experimental and theoretical aspects of protein folding and hydrophobic interactions

H. A. SCHERAGA

*Baker Laboratory of Chemistry and Chemical Biology, Cornell University
Ithaca, New York 14853-1301, USA*

A description is provided of experimental studies of the folding pathways of bovine pancreatic ribonuclease A (RNase A) and of a physics-based theoretical approach to compute both the folded native structure of a globular protein and the pathways leading to it, using the information contained in the amino acid sequence and an empirical potential energy function. A brief discussion of hydrophobic interactions is also provided.

1. – Introduction

Ever since Anfinsen demonstrated that polypeptide chains can fold spontaneously into the three-dimensional structure of a globular protein [1], without ancillary help from enzymes or templates, much activity has been devoted to determining the three-dimensional structure of a protein, and the pathways leading to it. The guiding principle behind this activity is the thermodynamic hypothesis enunciated by Anfinsen [1], namely, that the native protein adopts the thermodynamically most stable conformation in *a given solvent* at a particular temperature. Consequently, much experimental and theoretical work has been expended to try to determine how the physical interactions between the amino acid residues determine the native structure and the folding pathways. This paper is a summary of the methodology and results therefrom, from our laboratory, for solving these two protein-folding problems.

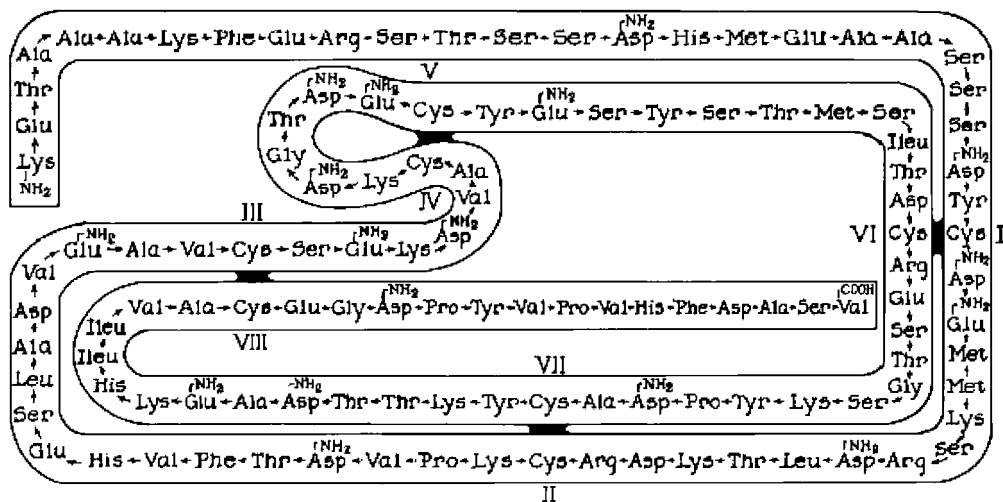


Fig. 1. – Amino acid sequence of RNase A.

2. – Experimental physical-chemical determination of structure

Before the successful application of X-ray and NMR methods to determine protein structure, and before the advent of recombinant DNA technology, the structure problem was attacked experimentally by using physical-chemical methods applied to the protein bovine pancreatic ribonuclease A (RNase A) [2], whose amino acid sequence is shown in fig. 1. The goal was to obtain distance constraints between amino acid residues that would limit the folding pattern of the native structure. These constraints were, first of all, the four disulfide bonds of RNase A (see fig. 1) and, secondly, non-covalent interactions that could be identified by spectroscopic, potentiometric titration, and chemical modification methods. With this approach, three specific non-covalent interactions were identified between tyrosyl and aspartyl residues, namely Tyr25-Asp14, Tyr 92-Asp38, and Tyr97-Asp83 of RNase A [2]. While these seven distance constraints (four disulfide bonds and three Tyr-Asp interactions) clearly place a restriction on how the backbone chain can fold, they are too few in number to determine the complete three-dimensional structure. A much larger number of distance constraints, and their proper distribution within the structure, are needed, the required number and distribution depending on the desired root-mean-square deviation (r.m.s.d.) between the experimentally determined structure and the actual one [3]; *i.e.* the larger the number of distance constraints, and the appropriateness of their distribution within the three-dimensional structure, the smaller will be the rmsd. With this recognition of the limitation of the use of only distance constraints (together with the associated polypeptide chain geometry) to determine three-dimensional structure, a theoretical approach with an empirical potential energy function (making use of only a few distance constraints) was developed [4-8] to compute

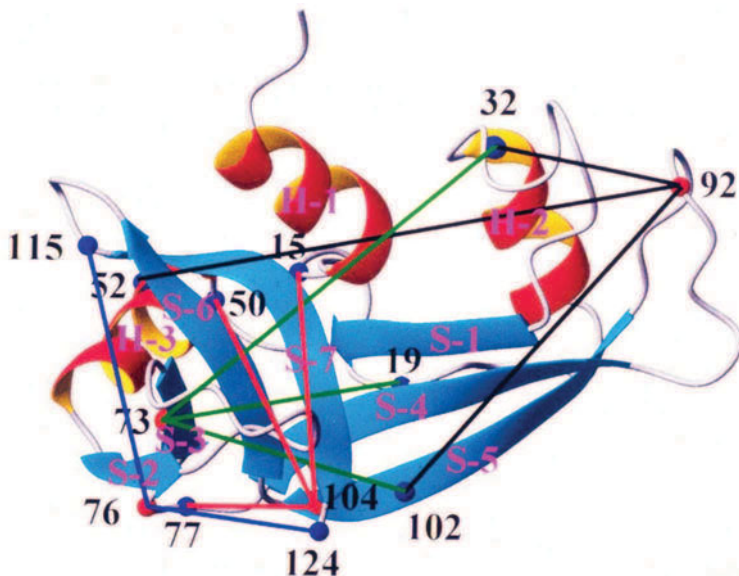


Fig. 2. – View of labeled mutants and corresponding distances (12) shown on the crystal structure of RNase A (11).

protein structure. Subsequently, the emphasis was shifted to computing the structure from the amino acid sequence, using only an empirical potential energy function without having to rely on distance constraints or on any other experimental data [9, 10]. This theoretical approach is discussed below, beginning with sect. 5.

It is of interest to point out that the unfolded state of RNase A, under folding conditions and with its disulfide bonds reduced, is not a statistical coil, but instead reflects the native structure [11] to some extent. This was demonstrated by FRET (Fluorescence Resonance Energy Transfer) experiments [12]. Several mutants with a fluorescence donor, Trp, at positions 73, 76, 92 and 104, respectively, and Cys (with the fluorescence acceptor, coumarin, subsequently linked to it) at positions 19, 32, 52, 77, 102, 115 and 124, respectively, were used to study the conformations of wild-type RNase A in the reduced state under folding conditions, with steady-state and time-resolved FRET measurements. These measurements provided distributions of intramolecular distances in the reduced state in nine mutants shown in fig. 2 [12]. The computed distribution of distances is quite wide but narrows down when the disulfide bonds are introduced. Further, the mean values of these distributions in the unfolded state are far from those expected for a statistical coil from polymer theory, but resemble those of the native structure. Therefore, the unfolded state is poised to fold to the native structure.

3. – Experimental physical-chemical determination of oxidative folding pathways of RNase A

The folding pathways of RNase A were determined both with and without its disulfide bonds intact. With the disulfide bonds reduced, folding can be induced by adding an oxidizing agent to re-form the disulfide bonds. Initially, these studies were carried out with a redox reagent consisting of oxidized and reduced glutathione, GSSG and GSH, respectively [13-17]. The kinetics of folding were complicated because of the presence of a large number of folding intermediates to which glutathione was covalently linked, but nevertheless were interpretable. This problem of bound glutathione was subsequently avoided by using the cyclic reagent dithiothreitol (DTT) instead of the linear glutathione because the folding intermediates did not contain bound DTT; thus, oxidative folding and reductive unfolding were studied with a redox system consisting of DTT^{ox} and DTT^{red} without linking DTT to the intermediates [12, 18-47]. Therefore, the reduced number of folding intermediate species, using DTT, made it much easier to interpret the kinetic data. Results of studies of oxidative folding and reductive unfolding pathways included the order of formation and reduction of disulfide bonds, their distribution at various stages of oxidative folding, identification of the rate-determining steps, and the intermediate structures formed in the rate-determining steps. In addition, multiple folding and unfolding pathways, some of which involved traps, were identified.

Oxidative folding of RNase A was studied by reducing the four disulfides, 26-84, 40-95, 58-110, and 65-72, shown in fig. 1. The oxidation reaction with DTT^{ox} was quenched at various times by blocking the un-reacted sulfhydryl groups with 2-aminoethyl methane thiosulfonate (AEMTS), and the products were separated chromatographically (see fig. 3A), and their concentrations were determined, thereby providing quantitative kinetic data on the folding process [18, 20-23, 28, 29, 33], where R is reduced RNase A; 1S, 2S, 3S, and 4S are folding intermediate ensembles of one-, two-, three-, and four-disulfide species, N is the native form of the wild-type protein, or of the mutant missing either the 40-95 or 65-72 disulfide bond, and des[40-95]_N and des[65-72]_N are three-disulfide species containing three native disulfides but missing the 40-95 or 65-72 disulfide bond, respectively. By desalting and removing DTT from an intermediate-stage oxidation mixture, and allowing for SH/S-S shuffling interchange, the existence of des[40-95]_N and des[65-72]_N was enhanced (see fig. 3B). The mechanism [28, 29, 35, 37, 38, 40, 41] that accounts for the kinetic data at 25°C is shown in fig. 4. The species in the steady-state pre-equilibrium are unstructured (indicated by the subscript U). In the rate-determining steps for the wild-type protein, the ensemble of 3S_U species reshuffles by SH/S-S interchange along two separate pathways (shown in fig. 4) to des[40-95]_N and des[65-72]_N, respectively, which have acquired the native structure, as determined by 3D NMR analysis [26, 27]. In two mutants, [C40A, C95A] and [C65A, C72A], N (equivalent to des[40-95] or des[65-72]) is formed in the rate-determining step by oxidation of 2S_U precursors rather than by SH/S-S reshuffling. The dominant pathways to des[40-95] and des[65-72] (to the extent of ~ 80% and ~ 20%, respectively) were observed in experiments on the wild-type protein [35, 41]. The minor pathways from 2S_U to des[40-95] and des[65-72] (to the extent

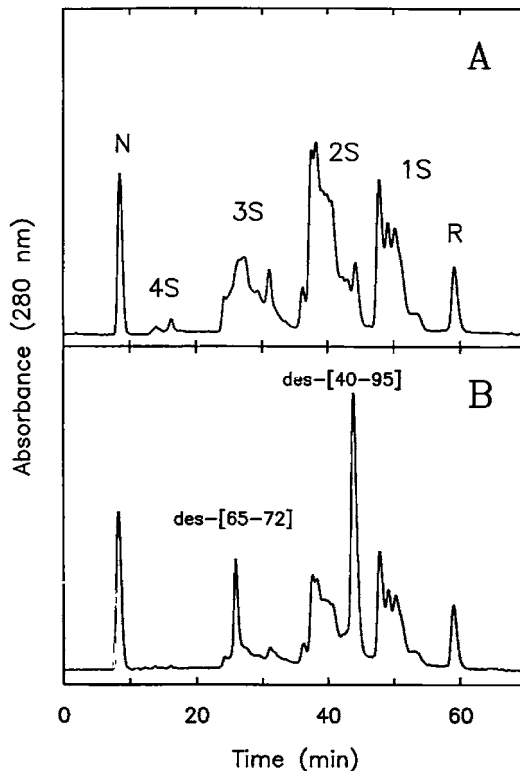


Fig. 3. – Chromatograms showing (A) a typical interrupted oxidative regeneration mixture, and (B) the same mixture that had been desalted to remove DTT, and allowed to reshuffle to enhance the formation of des[40-95]_N and des[65-72]_N (28).

of $\sim 5\%$ each) could not be observed in experiments with the wild-type protein. The mutants had to be used to detect these minor pathways.

We also found two other three-disulfide intermediates, des[26-84] and des[58-110], that are populated in the oxidative folding pathway of RNase A at a low temperature (15°C) even though they are not stable enough to be detected at 25°C [33]. These intermediates possess all but one of the four native disulfide bonds and have a stable tertiary structure at 15°C , similar to the two other observed intermediates, des[65-72] and des[40-95]. While the latter two des species lack one relatively *surface-exposed* disulfide bond, the other two des species each lack one *buried* disulfide bond (fig. 5). The des[26-84] and des[58-110] species are involved as kinetic traps in the rate-determining steps during the low-temperature oxidative folding from 3S_U of RNase A. By developing a new method [48], we were able to characterize the formation and consumption of des[26-84] and des[58-110] in oxidative folding.

A key problem in experimental protein folding is that of characterizing the distribution of disulfide bonds in each of the folding-intermediate ensembles 1S, 2S and 3S [25, 43].

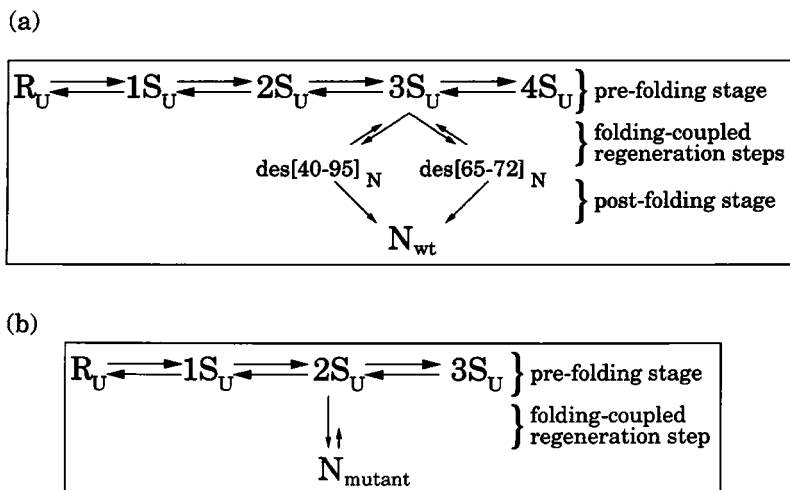


Fig. 4. – Oxidative folding mechanism of (a) wild-type RNase A, and (b) a three-disulfide mutant of RNase A.

In each of these ensembles, the 65-72 disulfide bond was found to be the dominant one. For example, in the 1S ensemble [25], there are 28 possible one-disulfide species in rapid SH/S-S reshuffling with each other. Approximately 50% of this ensemble is accounted for by only two species, the native 65-72 and the non-native 58-65 in the ratio of 4:1. None of the remaining 26 species exceeds 10% in the ensemble. Since the 65-72 and 58-65 S-S bond

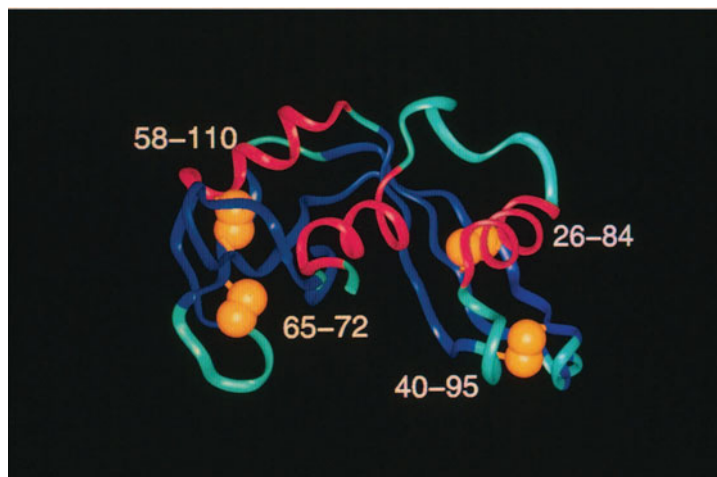


Fig. 5. – Schematic representation of the structure of RNase A, showing that disulfide bonds 40-95 and 65-72 are relatively exposed whereas 26-84 and 58-110 are buried.

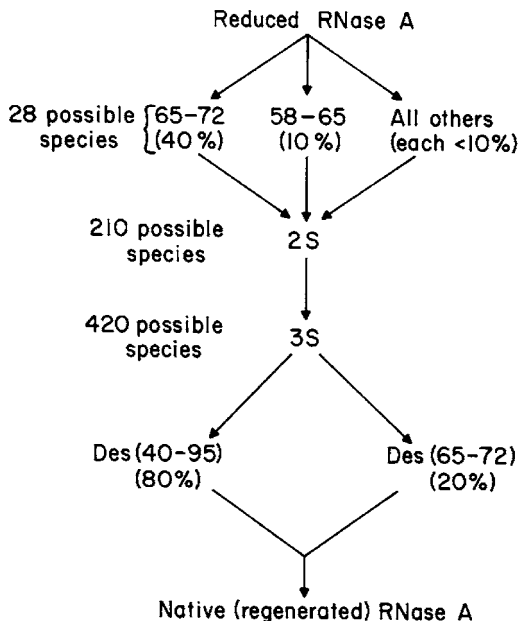


Fig. 6. – Summary of oxidative folding pathways of RNase A.

loops are the same size, their entropy of formation is the same. Therefore, the preference for the native 65-72 loop must be due to specific interactions that preferentially favor formation of this loop. In fact, conformational energy calculations carried out on a 58-72 peptide fragment in water [49] suggest that the observed preferred native pairing of the 65 and 72 cysteines arises from energetic determinants (adoption of the left-handed single-residue conformation by Gly 68, and side-chain interactions involving Gln69) contained within this peptide sequence. Since the ratio of the populations of the correct 65-72 disulfide bond to the incorrect 58-65 disulfide bond is observed experimentally to be the same (4:1) in both the 58-72 fragment [50] and in whole wild-type RNase A [25], it is these identified *local* interactions that lead to the preference for the 65-72 disulfide bond in the wild-type protein. *These are the types of interactions that are sought in our goal to determine the underlying inter-residue interactions that govern protein folding*, as stated in sect. 1.

The overall picture of the oxidative folding in RNase A, showing the possible sizes of the ensembles of the disulfide-bonded intermediates, is illustrated in fig. 6. As pointed out above, the 65-72 and 58-65 one-disulfide species, formed in the ratio of 4:1, constitute about 50% of the 1S ensemble. The 65-72 disulfide bond dominates in all successive ensembles and winds up, in the rate-determining step, in the dominant des[40-95] species. The sparser species, des[65-72], arises from all the other precursor species. These des species then go rapidly to N.

Mechanism of reductive protein unfolding

Yue-Jin Li, David M. Rothwarf and Harold A. Scheraga

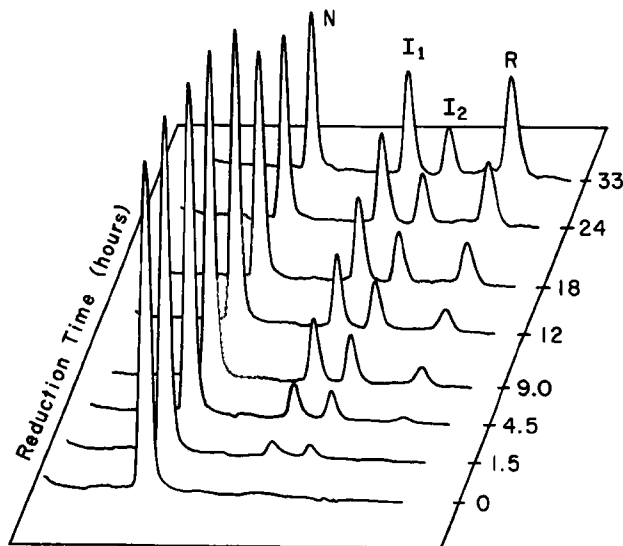


Fig. 7. – Chromatograms indicating the stages of reductive unfolding of RNase A. I_1 and I_2 correspond to des[65-72] and des[40-95], respectively.

The oxidative folding of several proteins that are homologous to RNase A, namely RNase B, onconase, and angiogenin have also been studied. RNase B is identical in amino acid sequence, and very similar in 3D structure to RNase A but has a carbohydrate attached to Asn34 [51]. The carbohydrate is actually a mixture of five isoforms which were successfully separated by mass spectrometry [47]. The carbohydrate does not influence the rate of reduction of RNase B, which is the same as that of RNase A [47], but enhances the oxidative folding rate of RNase B over that of RNase A without changing the folding pathway [52].

The reductive unfolding of RNase A was examined by similar AEMTS-blocking and chromatographic analysis. The several stages of unfolding are shown in fig. 7. The intermediate species I_1 and I_2 correspond to des[65-72] and des[40-95], respectively, indicating that the initial stages of reduction take place in two independent steps, reflecting the relative exposure of the 65-72 and 40-95 disulfides, as indicated in fig. 5. From an analysis of the reductive unfolding kinetics, the mechanism of fig. 8 was deduced [24].

Onconase (ONC) lacks the 65-72 disulfide bond of RNase A and, hence, follows a different folding pathway (unpublished results). In the reductive unfolding pathway, the protection of the 40-95 disulfide bond of RNase A is lacking from the homologous 30-75 disulfide bond of ONC. This local interaction in RNase A, which is absent in ONC, accounts for an observed difference in the reduction pathway and dynamics of the two proteins [53].

Mechanism of reductive protein unfolding

Yue-Jin Li, David M. Rothwarf and Harold A. Scheraga

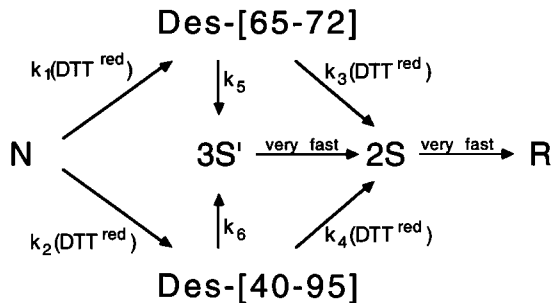


Fig. 8. – Parallel reductive unfolding pathways of RNase A at 15°C and 25°C, pH 8.0.

4. – Experimental physical-chemical determination of folding pathways of disulfide-intact RNase A

With the disulfide bonds intact, unfolding is induced by addition of guanidine hydrochloride (Gdn), and refolding is accomplished by diluting out the Gdn to a very low concentration. Disulfide-intact folding reactions occur very rapidly and, therefore, stopped-flow techniques are required to study these processes. The main results from such studies are the role that proline cis-trans isomerization plays in the various folding/unfolding stages. A particularly interesting result was the identification of an intermediate which folded very fast because its refolding was initiated from an unfolded state *before* its two cis and two trans prolines (that are present in the native state) had time to isomerize during the addition of Gdn and its subsequent rapid dilution [54]. This result was obtained by a double-jump stopped-flow technique in which the protein was first unfolded and then re-folded. This unfolded very-fast-refolding species is designated as U_{vf} . Other slower folding species, namely, a fast-folding species (U_f), and other species designated as a medium-folding one (U_m), a major slow-folding one U_s^{II} , and a minor slow-folding one (U_s^I) [55], were identified [56-63]. Figure 9 illustrates the various folding/unfolding stages, of RNase A with intact disulfide bonds; only three prolines are considered here because the fourth one, at position 42, plays no role in refolding kinetics [57], possibly because the Lys 41-Pro 42 peptide bond is overwhelmingly trans under both folding and unfolding conditions [64, 65] and, thus, the relative percentage of cis Pro 42 may be too small to have a discernible effect on the conformational folding.

Double-jump kinetic experiments on a P93A mutant [57] indicated that the Tyr92-Ala 93 peptide group (like the Tyr92-Pro93 peptide group in wild-type RNase A) is cis in the native state, a finding confirmed by NMR spectroscopy [66]. This is very unusual because the trans:cis ratio for non-N-substituted amino acids is typically greater than 100:1 [67]. The local interactions around Ala93 seem to provide sufficient energy to

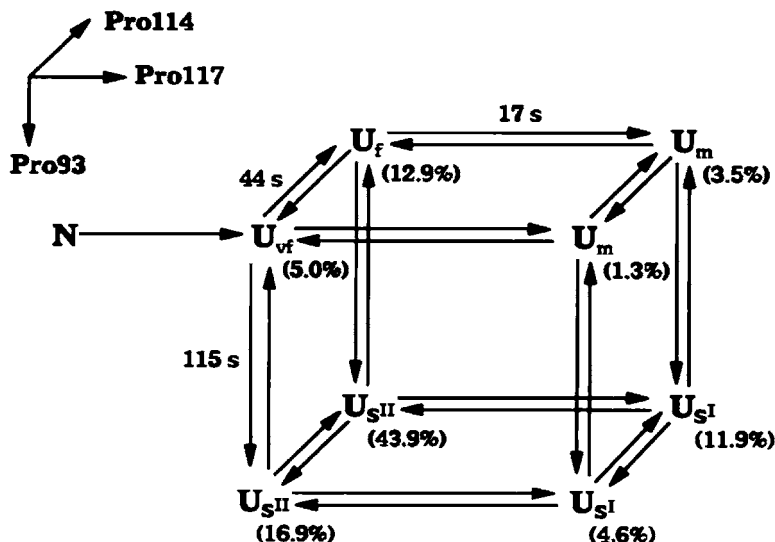


Fig. 9. – Kinetic and thermodynamic parameters for various steps in the folding/unfolding of disulfide-intact RNase A (57).

overcome its preference for trans, thereby enabling it to exist in the cis conformation. Presumably, these same interactions enable Pro93 in the wild-type protein to adopt the cis conformation. In an isolated X-Pro peptide group, the cis and trans conformations are almost equally energetic [67]. Therefore, it is not surprising to find both cis and trans prolines in proteins, with the local interactions around each specific proline providing the energy to force this residue to adopt a particular cis or trans conformation.

Native RNase A contains no tryptophan. A mutant, Y92W, provided information around Tyr92 [68] that supported the conclusions from earlier studies of proline-to-alanine mutants [57]. Continuous-flow measurements provided information about structural events in the $70 \mu\text{s}$ folding time range in which isomerization of proline peptide bonds was shown to have a major impact on the structural events during these early folding stages [69].

The purpose of the above experiments was to learn about the stability and structural changes in various parts of the RNase A molecule. A structurally plausible model accounts for the proline isomerization results in folding studies of RNase A [63]. In particular, the model shows that folding conditions strongly accelerate the cis/trans isomerization of proline peptide groups 93 and 114 to their native cis conformation, suggesting the presence of flickering local structure in their β -hairpins. The model, discussed in detail on page 5725 of ref. [63], also accounts for an unusual temperature dependence of Pro114 isomerization under folding conditions.

Finally, a series of mutants, in which tryptophan was incorporated individually at various sites, provided information about conformational stability in each of these ar-

eas [70]. Results of experiments on the local thermally-induced unfolding of each of these mutants shed new light on the thermal unfolding transitions in different parts of the RNase A molecule, generally supporting an earlier unfolding hypothesis of Burgess and Scheraga [71], as modified by Matheson and Scheraga [72].

5. – Theoretical determination of structure with an all-atom approach

Motivated by the experimental results mentioned in sect. 2, a theoretical approach was developed to compute the structures of native proteins [4], initially using only a hard-sphere potential for all atom pairs in the chain. Later, a more realistic pairwise potential energy function, with the bond lengths and bond angles kept fixed, was developed [5-10]. The potential energy of the whole molecule was treated as a sum over all the pairwise potentials, and subjected to minimization with respect to the torsional angles of the backbone and side chains to locate the global minimum. Alternatively, a Monte Carlo-plus-minimization (MCM) procedure was used to search the conformational space for the global minimum [73, 74]. Computational results obtained for the pentapeptide enkephalin [74], the cyclic decapeptide gramicidin S [75], and models of the fibrous protein collagen [76-87] were found to be in agreement with experiment. The MCM procedure was subsequently enhanced with an electrostatic component to produce an electrostatically-driven Monte Carlo (EDMC) procedure [88, 89], which greatly improved the efficiency of MCM. EDMC, together with two continuum hydration models [90, 91], led to excellent results for larger polypeptides, the 36-residue villin headpiece [92] and the 46-residue protein A [93]. (See fig. 10.) A summary of the all-atom approach is provided in ref. [94].

An important consideration for understanding the interactions involved in protein folding is the hydrophobic phenomenon. Therefore, the nature of hydrophobic interactions is discussed in the Appendix.

6. – Theoretical determination of structure with a United-Residue (UNRES) Model

Recognizing the difficulty in extending the all-atom approach to larger molecules (those containing of the order of 100 to 200 amino acids residues), a hierarchical procedure was developed [95] in which the initial search of conformational space was carried out with a united-residue (UNRES) model to reach only the *region* in which the global minimum would lie [94, 96-109]. The search is carried out with a Conformational Space Annealing (CSA) procedure [110-114], which includes a genetic-type algorithm. Then the UNRES models are converted to all-atom models [115, 116], and the search for the global minimum is continued with the EDMC procedure [88] with inclusion of solvent by a continuum hydration model [91]. The UNRES model is illustrated in fig. 11, and successive stages of the CSA procedure (in which a selected number of minima coalesce to small groups, one of which should contain the global minimum of the UNRES potential) are illustrated in figs. 12-14.

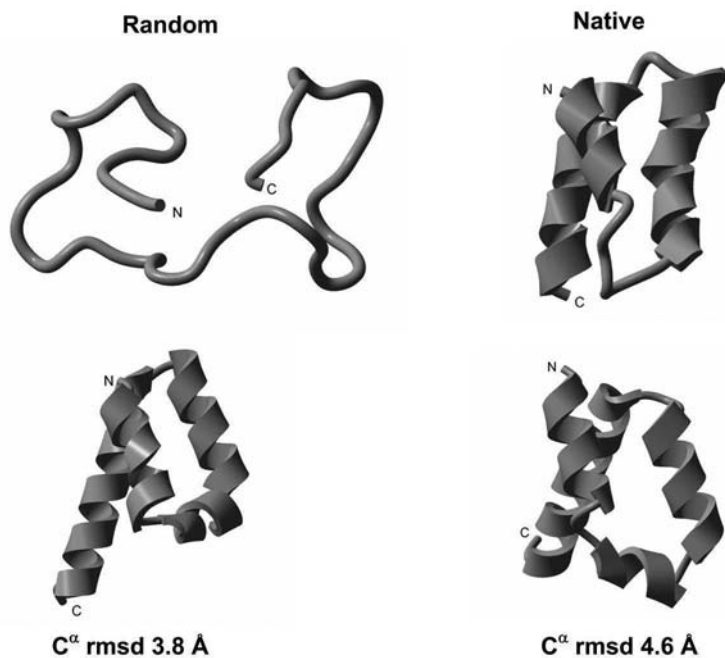


Fig. 10. – Illustration of random starting conformation of the 46-residue protein A, whose native structure is shown at the upper right. Structures computed with two different continuum hydration models are shown at the bottom, left and right, respectively (93).

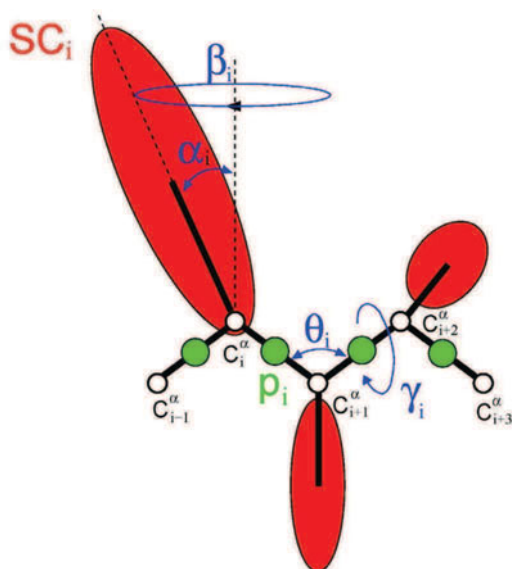


Fig. 11. – The UNRES model with four degrees of freedom ($\theta, \gamma, \alpha, \beta$). Virtual $C^\alpha \dots C^\alpha$ and $C^\alpha \dots SC$ bonds connect the backbone and side-chain groups, respectively.

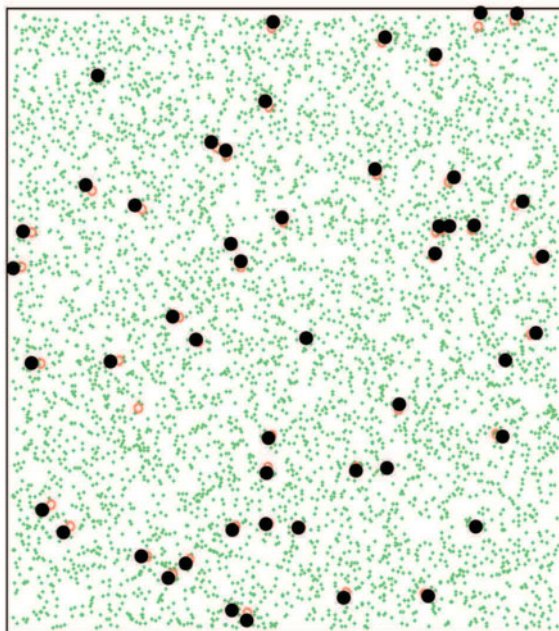


Fig. 12. – Schematic starting energy landscape for CSA, showing a finite number of selected minima (large circles) among many other local minima (dots).

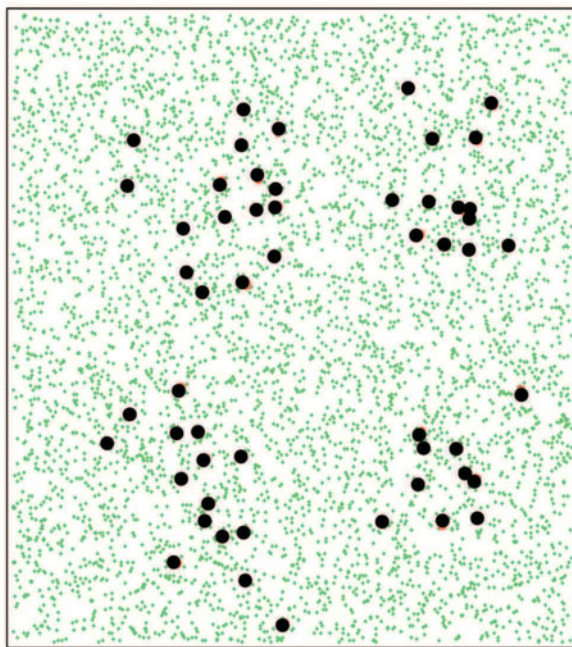


Fig. 13. – Schematic intermediate energy landscape for CSA.

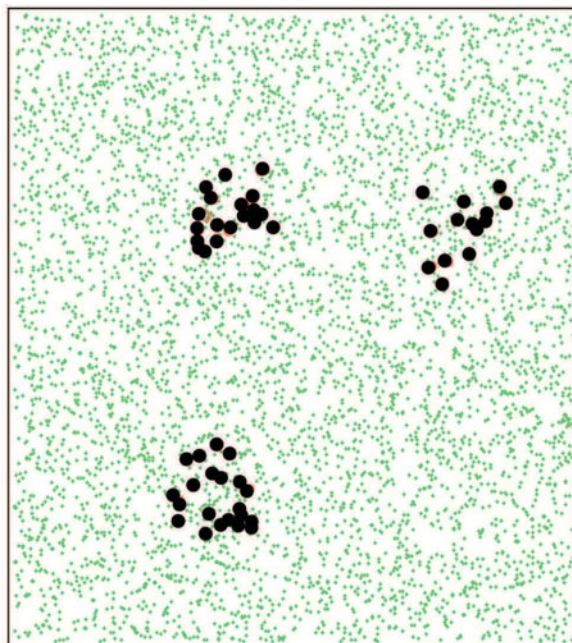


Fig. 14. – Schematic final energy landscape for CSA.

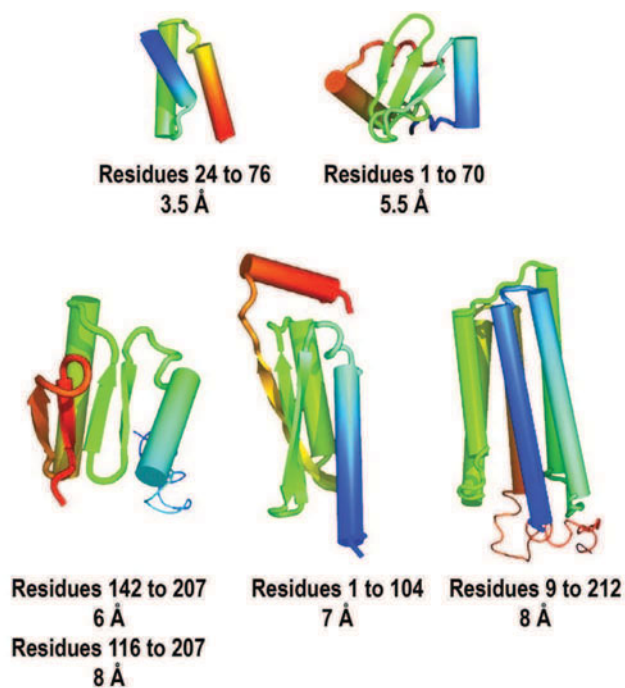


Fig. 15. – Results from CASP6.

The UNRES/CSA method was applied to several systems [115-119], and subjected to blind tests in several CASP (Critical Assessments of Protein Structure Prediction) exercises [120,121]. The most recent results, from CASP6, are illustrated in fig. 15. Good predictions, within a 6 Å rmsd, were obtained for large portions (up to 60–70 residues) of target proteins. The UNRES force field is presently being improved by implementing methods described in sect. 7, and the search procedure is being enhanced by inclusion of replica-exchange methodology [122,123].

7. – Molecular dynamics with the UNRES force field

Given the prediction success with UNRES in CASP exercises, this force field was adapted for molecular dynamics (MD). Since the fast degrees of freedom are averaged out in developing UNRES [124], an UNRES/MD procedure can partially overcome the time barrier in all-atom MD, *i.e.* it can produce trajectories that are further along, in time, compared to all-atom trajectories. The method is based on Langevin dynamics [125,126], with a time-dependent Langevin equation based in terms of generalized coordinates (q) and momenta ($\dot{q}$) related to the virtual bonds shown in fig. 11. With UNRES/MD, it has been possible to fold a protein containing as many as 75 amino acid residues [127] (see fig. 16). Figure 17 illustrates the folding of an $(\alpha + \beta)$ protein [127]. By generating 400 folding trajectories of protein A with UNRES/MD, it has been possible to obtain reliable statistics with which to compute folding kinetics [128]. The kinetic characteristics depend on the quantity being computed or measured (see fig. 18). With all-atom MD, it has been possible to assess the effect of temperature, friction, and random forces on the folding trajectories [129], and thereby resolve discrepancies in the literature among several theoretical and experimental results. Finally, the UNRES/MD method has been extended to treat multiple-chain proteins [130]; an example of the folding of a two-chain protein is shown in fig. 19.

Since MD provides free-energy information, UNRES/MD is currently being used to improve the UNRES force field by, among other enhancements, the inclusion of entropy [124].

8. – Conclusions

Experimental studies of protein structure have provided distance constraints which fostered the development of theoretical all-atom methods to compute protein structure from amino acid sequence. Similarly, experimental studies of protein folding pathways have provided information about inter-residue interactions that determine folding pathways and the final folded structure. A more recent development of an UNRES model of a polypeptide chain has provided an alternative to the all-atom approach that has led to successful calculations of large portions of large protein molecules and to UNRES/MD calculations of protein-folding pathways. Current efforts to improve both the UNRES and all-atom force fields are expected to lead to even more accurate predictions of protein structure and, thereby, of the interactions that lead to them. Finally, a brief discussion of hydrophobic interactions is provided in an Appendix.

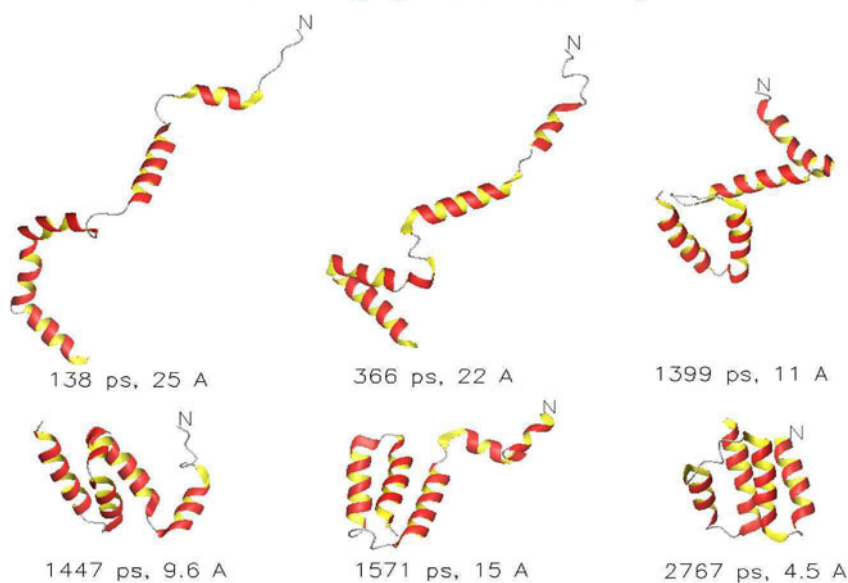
1CLB (α , 75 residues)

Fig. 16. – Folding trajectory of a 75-residue α -helical protein, starting from a fully-extended structure. (127).

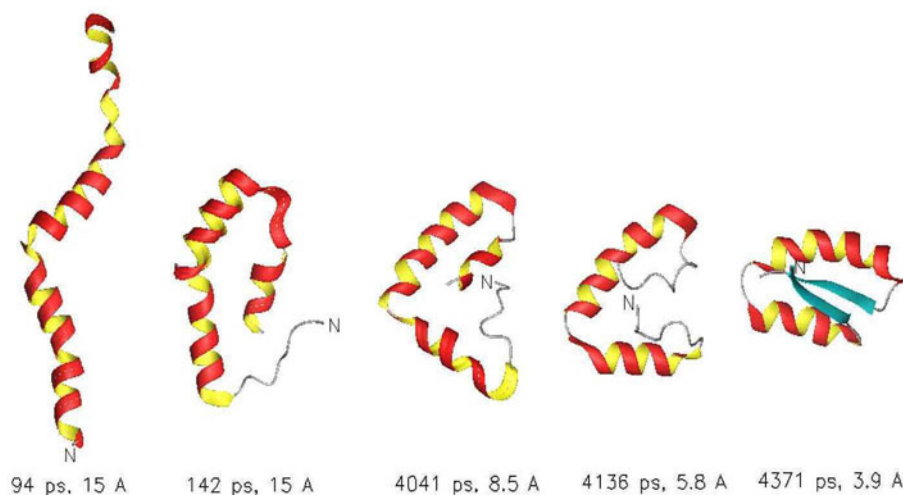
1E0G ($\alpha+\beta$, 48 residues)

Fig. 17. – Folding trajectory of an ($\alpha+\beta$) protein, starting from a fully-extended structure (127).

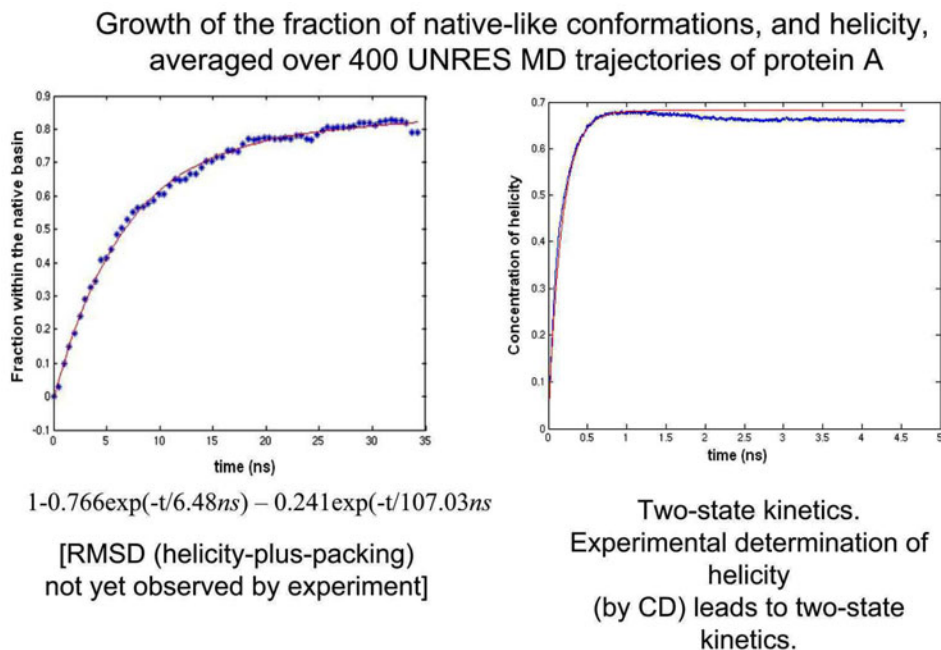


Fig. 18. – Illustration of how the folding trajectory depends on the quantity being computed (or measured). Left: a two-exponential curve based on rmsd; right: a single exponential curve based on helix content.

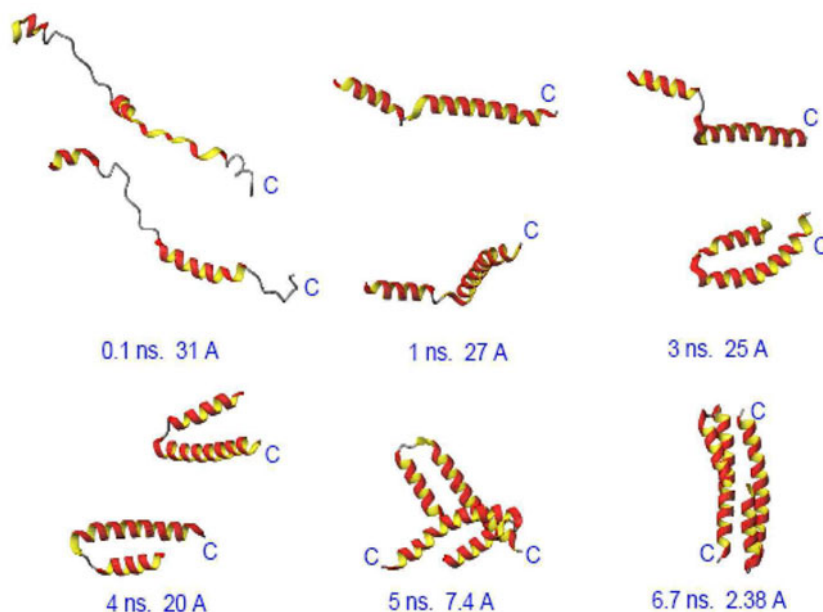


Fig. 19. – Folding of a two-chain protein by UNRES/MD (130).

APPENDIX

Hydrophobic interactions

In discussing the hydrophobic phenomenon, it is important to distinguish between two terms, hydrophobic hydration and hydrophobic interaction. Hydrophobic hydration refers to the interaction of a nonpolar particle with the surrounding water medium, while hydrophobic interaction refers to the interaction of two such hydrated nonpolar particles in water.

This subject was treated extensively by Némethy and Scheraga within the context of the statistical mechanics of liquid water [131], of liquid D₂O [132], and of the solubility of hydrocarbons in water, involving hydrophobic hydration [133]. An extension of the treatment of aqueous hydrocarbons led to a theory for hydrophobic interactions [134]. The theory was later improved by Griffith and Scheraga [135], who also extended the treatment to the hydration of the alkali halides at infinite dilution [136].

The original theory of liquid water [131] was based on a model that took account of the cooperativity of hydrogen-bond formation between water molecules. As a result of this cooperativity, water was regarded as a mixture of hydrogen-bonded clusters embedded among nonbonded water molecules. The water molecules were distributed among five energy levels corresponding to molecules with four, three, two, one, and zero hydrogen bonds, according to the Boltzmann principle. After evaluation of the partition function for this system, excellent agreement was obtained between experimental and computed values of the temperature-dependent thermodynamic properties of liquid water, including the minimum at 4° in the molar volume of liquid water, and the minimum at higher temperature for liquid D₂O because of the difference in the zero-point energies of hydrogen-bonded H₂O and D₂O. The cluster size was larger for D₂O than for H₂O because of this difference in zero-point energies and the cluster size in both liquids decreased with increasing temperature. In the later improvement of the theory [135], the mole fraction of nonbonded water molecules was found to be very small, leading to a picture of hydrogen-bonded clusters with distorted tetrahedrally-coordinated hydrogen bonds, similar to a model of Stillinger [137].

Aqueous hydrocarbon solutions were treated [133] by considering the thermodynamics for the transfer of a hydrocarbon molecule from liquid hydrocarbon to water at infinite dilution. If the resulting solution were ideal, the thermodynamic parameters should be those shown in the left-hand column of table I. However, real aqueous hydrocarbon solutions are not ideal, and exhibit the parameters shown in the right-hand column of table 1. This behavior is attributed to shifts in the energy levels of liquid water due to the hydrocarbon-water interactions [133]. As a consequence of these interactions, described more fully in ref. [133], the energy level corresponding to four-bonded water molecules is lowered, and the other energy levels are raised, resulting in a higher population of the four-bonded level in aqueous hydrocarbon solutions compared to this level in pure water, *i.e.* nonpolar solutes are considered as structure makers in that they lead to a higher degree of hydrogen bonding of the surrounding water, similar to clathrate-like structures in which these cage-like flickering structures form in the neighborhood of a nonpolar particle. Hence, $\Delta H < 0$ and $\Delta S < 0$, with the entropy change dominating to make $\Delta G > 0$, *i.e.* hydrocarbons have low solubility in water. The temperature dependence of these quantities, *e.g.*, the existence of a non-zero heat capacity, is due to the decrease in the size of the hydrogen-bonded water clusters with increasing temperature.

TABLE I. – *Thermodynamic parameters for formation of aqueous hydrocarbon solutions.*

Ideal solution	Hydrocarbon solution
$\Delta H = 0$	$\Delta H < 0^{(b)}$
$\Delta S = 0^{(a)}$	$\Delta S < 0^{(a,b)}$
$\Delta G = 0^{(a)}$	$\Delta G > 0^{(a,b)}$
$\Delta V = 0$	$\Delta V < 0^{(b)}$

^(a) After subtracting the ideal entropy of mixing.

^(b) These quantities are temperature dependent.

The numerical values of the experimental and calculated temperature-dependent thermodynamic properties of aqueous solutions of aliphatic and aromatic hydrocarbons were in excellent agreement with each other. Subsequent simulations of aqueous hydrocarbon solutions with molecular mechanics led to radial distribution functions for liquid water and aqueous hydrocarbon solutions [138, 139] that validated the models that had been used for the earlier statistical mechanical treatments [131, 133] of these systems.

Based on the treatment of aqueous hydrocarbon solutions, the hydrophobic interaction was treated by the model of fig. 20, which illustrates the formation of a hydrophobic interaction between the nonpolar side chains of alanine and leucine [134]. As the nonpolar groups approach each other, there is a small van der Waals attraction between

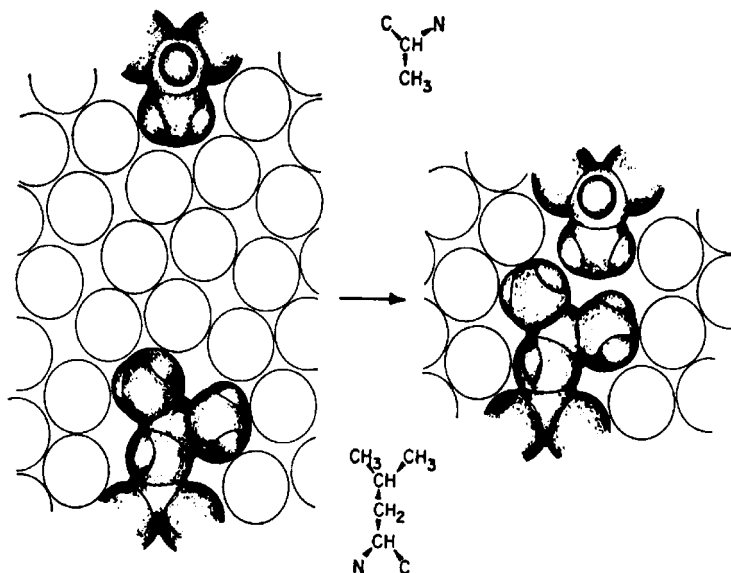


Fig. 20. – Example of formation of a hydrophobic interaction between the side chains of alanine and leucine (134). The open circles are only schematic representations of water molecules.

TABLE II. – *Thermodynamic parameters^(a) for hydrophobic hydration and hydrophobic interactions.*

Formation of hydrocarbon solutions	Formation of hydrophobic interactions
$\Delta H < 0$	$\Delta H > 0$
$\Delta S < 0$	$\Delta S > 0$
$\Delta G > 0$	$\Delta G < 0$
$\Delta V < 0$	$\Delta V > 0$

^(a) All these quantities are temperature dependent.

the nonpolar groups. But, by far, the largest contribution to the favorable free energy of hydrophobic interaction arises from the decreased number of water molecules around the nonpolar complex, compared to the number around the initially-separated partners. This situation is equivalent to the removal of some hydrocarbon from solution. Hence, the thermodynamic parameters for such a hydrophobic interaction should be opposite in sign compared to those of hydrophobic hydration, as illustrated in table II. Numerical values for such pairwise interactions, and their temperature dependence were presented for all pairs of nonpolar side chains of the naturally occurring amino acids [134]. Clearly, with $\Delta G < 0$, formation of a hydrophobic interaction is a favorable process and, as with hydrophobic hydration, the entropy dominates over the enthalpy. The positive value of ΔH indicates that hydrophobic interactions become stronger as the temperature increases in the range of positive ΔH . Various experimental tests were subsequently carried out [140], and validate the computed thermodynamic parameters for hydrophobic interactions. Finally, recent experiments [141] have provided evidence to support two related models [142,143] for the role of hydrophobic interactions in initiating and propagating protein folding.

* * *

This research was supported for many years by grants from NIH and NSF.

REFERENCES

- [1] ANFENSEN C. B., *Science*, **181** (1973) 223.
- [2] SCHERAGA H. A., *Fed. Proc.*, **26** (1967) 1380.
- [3] WAKO H. and SCHERAGA H. A., *Macromolecules*, **14** (1981) 961.
- [4] NÉMETHY G. and SCHERAGA H. A., *Biopolymers*, **3** (1965) 155.
- [5] SCHERAGA H. A., *Adv. Phys. Org. Chem.*, **6** (1968) 103.
- [6] MOMANY F. A., MCGUIRE R. F., BURGESS A. W. and SCHERAGA H. A., *J. Phys. Chem.*, **79** (1975) 2361.
- [7] NÉMETHY G., POTTLE M. S. and SCHERAGA H. A., *J. Phys. Chem.*, **87** (1983) 1883.

- [8] NÉMETHY G., GIBSON K. D., PALMER K. A., YOON C. N., PATERLINI G., ZAGARI A., RUMSEY S. and SCHERAGA H. A., *J. Phys. Chem.*, **96** (1992) 6472.
- [9] SCHERAGA H. A., *Nobel Symposium 11, on "Symmetry and Function of Biological Systems at the Macromolecular Level"*, edited by ENGSTROM A. and STRANDBERG B. (Almqvist and Wiksell, Stockholm) 1969, p. 43.
- [10] SCHERAGA H. A., *Chem. Rev.*, **71** (1971) 195.
- [11] WLODAWER A., SVENSSON L. A., SJÖLIN L. and GILLILAND G. L., *Biochemistry*, **27** (1988) 2705.
- [12] NAVON A., ITTAH V., LANDSMAN P., SCHERAGA H. A. and HAAS E., *Biochemistry*, **40** (2001) 105.
- [13] KONISHI Y., OOI T. and SCHERAGA H. A., *Biochemistry*, **20** (1981) 3945.
- [14] KONISHI Y., OOI T. and SCHERAGA H. A., *Biochemistry*, **21** (1982) 4734.
- [15] KONISHI Y., OOI T. and SCHERAGA H. A., *Biochemistry*, **21** (1982) 4741.
- [16] KONISHI Y., OOI T. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **79** (1982) 5734.
- [17] SCHERAGA H. A., KONISHI Y. and OOI T., *Adv. Biophys.*, **18** (1984) 21.
- [18] ROTHWART D. M. and SCHERAGA H. A., *J. Am. Chem. Soc.*, **113** (1991) 6293.
- [19] ROTHWART D. M. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **89** (1992) 7944.
- [20] ROTHWART D. M. and SCHERAGA H. A., *Biochemistry*, **32** (1993) 2671.
- [21] ROTHWART D. M. and SCHERAGA H. A., *Biochemistry*, **32** (1993) 2680.
- [22] ROTHWART D. M. and SCHERAGA H. A., *Biochemistry*, **32** (1993) 2690; 7064 (Erratum).
- [23] ROTHWART D. M. and SCHERAGA H. A., *Biochemistry*, **32** (1993) 2698.
- [24] LI Y.-J., ROTHWART D. M. and SCHERAGA H. A., *Nature Struct. Biol.*, **2** (1995) 489.
- [25] XU X., ROTHWART D. M. and SCHERAGA H. A., *Biochemistry*, **35** (1996) 6406.
- [26] SHIMOTAKAHARA S., RIOS C. B., LAITY J. H., ZIMMERMAN D. E., SCHERAGA H. A. and MONTELIONE G. T., *Biochemistry*, **36** (1997) 6915.
- [27] LAITY J. H., LESTER C. C., SHIMOTAKAHARA S., ZIMMERMAN D. E., MONTELIONE G. T. and SCHERAGA H. A., *Biochemistry*, **36** (1997) 12683.
- [28] ROTHWART D. M., LI Y.-J. and SCHERAGA H. A., *Biochemistry*, **37** (1998) 3760.
- [29] ROTHWART D. M., LI Y.-J. and SCHERAGA H. A., *Biochemistry*, **37** (1998) 3767.
- [30] IWAOKA M., JUMINAGA D. and SCHERAGA H. A., *Biochemistry*, **37** (1998) 4490.
- [31] XU X. and SCHERAGA H. A., *Biochemistry*, **37** (1998) 7561.
- [32] IWAOKA M., WEDEMEYER W. J. and SCHERAGA H. A., *Biochemistry*, **38** (1999) 2805.
- [33] WELKER E., NARAYAN M., VOLLES M. J. and SCHERAGA H. A., *FEBS Lett.*, **460** (1999) 477.
- [34] LAITY J. H., MONTELIONE G. T. and SCHERAGA H. A., *Biochemistry*, **38** (1999) 16432.
- [35] WEDEMEYER W. J., WELKER E., NARAYAN M. and SCHERAGA H. A., *Biochemistry*, **39** (2000) 4207; 7032 (Erratum).
- [36] SHIN H.-C. and SCHERAGA H. A., *J. Mol. Biol.*, **300** (2000) 995.
- [37] NARAYAN M., WELKER E., WEDEMEYER W. J. and SCHERAGA H. A., *Acc. Chem. Res.*, **33** (2000) 805.
- [38] WELKER E., NARAYAN M., WEDEMEYER W. J. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **98** (2001) 2312.
- [39] CAO A., WELKER E. and SCHERAGA H. A., *Biochemistry*, **40** (2001) 8536.
- [40] WELKER E., WEDEMEYER W. J., NARAYAN M. and SCHERAGA H. A., *Biochemistry*, **40** (2001) 9059.
- [41] SCHERAGA H. A., WEDEMEYER W. J. and WELKER E., *Methods in Enzymology*, edited by NICHOLSON A. W., Vol. **341** (Academic Press, San Diego) 2001, p. 189.
- [42] SAITO K., WELKER E. and SCHERAGA H. A., *Biochemistry*, **40** (2001) 15002.

- [43] WEDEMEYER W. J., XU X., WELKER E. and SCHERAGA H. A., *Biochemistry*, **41** (2002) 1483.
- [44] NARAYAN M., WELKER E. and SCHERAGA H. A., *Biochemistry*, **42** (2003) 6947.
- [45] NARAYAN M., WELKER E., WANJALLA C., XU G. and SCHERAGA H. A., *Biochemistry*, **42** (2003) 10783.
- [46] SHIN H.-C., NARAYAN M., SONG M.-C. and SCHERAGA H. A., *Biochemistry*, **42** (2003) 11514.
- [47] XU G., ZHAI H., NARAYAN M., MCLAFFERTY F. W. and SCHERAGA H. A., *Chem. Biol.*, **11** (2004) 517.
- [48] NARAYAN M., WELKER E. and SCHERAGA H. A., *J. Am. Chem. Soc.*, **123** (2001) 2909.
- [49] CARTY R. P., PINCUS M. R. and SCHERAGA H. A., *Biochemistry*, **41** (2002) 14815.
- [50] MILBURN P. J. and SCHERAGA H. A., *J. Protein Chem.*, **7** (1988) 377.
- [51] WILLIAMS R. L., GREENE S. M. and MCPHERSON A., *J. Biol. Chem.*, **262** (1987) 16020.
- [52] XU G., NARAYAN M. and SCHERAGA H. A., *Biochemistry*, **44** (2005) 9817.
- [53] XU G., NARAYAN M., KURINOV I., RIPOLL D. R., WELKER E., KHALILI M., EALICK S. E. and SCHERAGA H. A., *J. Am. Chem. Soc.*, **128** (2006) 1204.
- [54] HOURLY W. A., ROTHWARF D. M. and SCHERAGA H. A., *Biochemistry*, **33** (1994) 2516.
- [55] SCHMID F. X., *Methods Enzymol.*, **131** (1986) 70.
- [56] HOURLY W. A., ROTHWARF D. M. and SCHERAGA H. A., *Nature Struct. Biol.*, **2** (1995) 495.
- [57] DODGE R. W. and SCHERAGA H. A., *Biochemistry*, **35** (1996) 1548.
- [58] HOURLY W. A., ROTHWARF D. M. and SCHERAGA H. A., *Biochemistry*, **35** (1996) 10125.
- [59] HOURLY W. A. and SCHERAGA H. A., *Biochemistry*, **35** (1996) 11719.
- [60] HOURLY W. A. and SCHERAGA H. A., *Biochemistry*, **35** (1996) 11734.
- [61] JUMINAGA D., WEDEMEYER W. J., GARDUÑO-JÚAREZ R., McDONALD M. A. and SCHERAGA H. A., *Biochemistry*, **36** (1997) 10131.
- [62] JUMINAGA D., WEDEMEYER W. J. and SCHERAGA H. A., *Biochemistry*, **37** (1998) 11614.
- [63] BHAT R., WEDEMEYER W. J. and SCHERAGA H. A., *Biochemistry*, **42** (2003) 5722.
- [64] DYSON H. J., RANCE. M., HOUGHTEN R. A., LERNER R. A. and WRIGHT P. E., *J. Mol. Biol.*, **201** (1988) 161.
- [65] RALEIGH D. A., EVANS P. A., PITKEATHLY M. and DOBSON C. M., *J. Mol. Biol.*, **228** (1992) 338.
- [66] XIONG Y., JUMINAGA D., SWAPNA G. V. T., WEDEMEYER W. J., SCHERAGA H. A. and MONTELIONE G. T., *Protein Sci.*, **9** (2000) 421.
- [67] ZIMMERMAN S. S. and SCHERAGA H. A., *Macromolecules*, **9** (1976) 408.
- [68] SENDAK R. A., ROTHWARF D. M., WEDEMEYER W. J., HOURLY W. A. and SCHERAGA H. A., *Biochemistry*, **35** (1996) 12978.
- [69] WELKER E., MAKI K. RAMACHANDRA SHASTRY M. C., JUMINAGA D., BHAT R., SCHERAGA H. A. and RODER H., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 17681.
- [70] NAVON A., ITTAH V., LAITY J. H., SCHERAGA H. A., HAAS E. and GUSSAKOVSKY E. E., *Biochemistry*, **40** (2001) 93.
- [71] BURGESS A. W. and SCHERAGA H. A., *J. Theor. Biol.*, **53** (1975) 403.
- [72] MATHESON JR., R. R. and SCHERAGA H. A., *Biochemistry*, **12** (1979) 2437.
- [73] LI Z. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **84** (1987) 6611.
- [74] LI Z. and SCHERAGA H. A., *J. Mol. Struct. (Theochem)*, **179** (1988) 333.
- [75] DYGERT M., GO N. and SCHERAGA H. A., *Macromolecules*, **8** (1975) 750.
- [76] MILLER M. H. and SCHERAGA H. A., *J. Polym. Sci., Polym. Symp.*, **54** (1976) 171.
- [77] MILLER M. H., NÉMETHY G. and SCHERAGA H. A., *Macromolecules*, **13** (1980) 470.
- [78] MILLER M. H., NÉMETHY G. and SCHERAGA H. A., *Macromolecules*, **13** (1980) 910.
- [79] NÉMETHY G., MILLER M. H. and SCHERAGA H. A., *Macromolecules*, **13** (1980) 914.

- [80] NÉMETHY G. and SCHERAGA H. A., *Biopolymers*, **23** (1984) 2781; 581 (Erratum).
- [81] NÉMETHY G. and SCHERAGA H. A., *Biochemistry*, **25** (1986) 3184.
- [82] NÉMETHY G. and SCHERAGA H. A., *Biopolymers*, **28** (1989) 1573.
- [83] VITAGLIANO L., NÉMETHY G., ZAGARI A. and SCHERAGA H. A., *Biochemistry*, **32** (1993) 7354.
- [84] ZAGARI A., PALMER K., GIBSON K. D., NÉMETHY G. and SCHERAGA H. A., *Biopolymers*, **34** (1994) 51.
- [85] VITAGLIANO L., NÉMETHY G., ZAGARI A. and SCHERAGA H. A., *J. Mol. Biol.*, **247** (1995) 69.
- [86] PATERLINI M. G., NÉMETHY G. and SCHERAGA H. A., *Biopolymers*, **35** (1995) 607.
- [87] LEE J., SCHERAGA H. A. and RACKOVSKY S., *Biopolymers*, **40** (1996) 595.
- [88] RIPOLL D. R. and SCHERAGA H. A., *Biopolymers*, **27** (1988) 1283.
- [89] RIPOLL D. R. and SCHERAGA H. A., *J. Protein Chem.*, **8** (1989) 263.
- [90] OOI T., OOBATAKE M., NÉMETHY G. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **84** (1987) 3086; 6015 (Erratum).
- [91] VILA J., WILLIAMS R. L., VASQUEZ M. and SCHERAGA H. A., *Proteins: Struct., Function, and Genetics*, **10** (1991) 199.
- [92] RIPOLL D. R., VILA J. A. and SCHERAGA H. A., *J. Mol. Biol.*, **339** (2004) 915.
- [93] VILA J. A., RIPOLL D. R. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 14812.
- [94] SCHERAGA H. A., LIWO A., OLDZIEJ S., CZAPLEWSKI C., PILLARDY J., RIPOLL D. R., VILA J. A., KAZMIERKIEWICZ R., SAUNDERS J. A., ARNAUTOVA Y. A., JAGIELSKI A., CHINCHIO M. and NANIAS M., *Front. Biosci.*, **9** (2004) 3296.
- [95] SCHERAGA H. A., PILLARDY J., LIWO A., LEE J., CZAPLEWSKI C., RIPOLL D. R., WEDEMEYER W. J. and ARNAUTOVA Y. A., *J. Comput. Chem.*, **23** (2002) 28.
- [96] LIWO A., OLDZIEJ S., PINCUS M. R., WAWAK R. J., RACKOVSKY S. and SCHERAGA H. A., *J. Comput. Chem.*, **18** (1997) 849.
- [97] LIWO A., PINCUS M. R., WAWAK R. J., RACKOVSKY S., OLDZIEJ S. and SCHERAGA H. A., *J. Comput. Chem.*, **18** (1997) 874.
- [98] LIWO A., KAZMIERKIEWICZ R., CZAPLEWSKI C., GROTH. M., OLDZIEJ S., WAWAK R. J., RACKOVSKY S., PINCUS M. R. and SCHERAGA H. A., *J. Comput. Chem.*, **19** (1998) 259.
- [99] PILLARDY J., CZAPLEWSKI C., LIWO A., LEE J., RIPOLL D. R., KAZMIERKIEWICZ R., OLDZIEJ S., WEDEMEYER W. J., GIBSON K. D., ARNAUTOVA Y. A., SAUNDERS J., YE Y.-J. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **98** (2001) 2329.
- [100] LIWO A., CZAPLEWSKI C., PILLARDY J. and SCHERAGA H. A., *J. Chem. Phys.*, **115** (2001) 2323.
- [101] PILLARDY J., CZAPLEWSKI C., LIWO A., WEDEMEYER W. J., LEE J., RIPOLL D. R., ARLUKOWICZ P., OLDZIEJ S., ARNAUTOVA Y. A. and SCHERAGA H. A., *J. Phys. Chem. B*, **105** (2001) 7299.
- [102] LIWO A., ARLUKOWICZ P., CZAPLEWSKI C., OLDZIEJ S., PILLARDY J. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 1937.
- [103] OLDZIEJ S., KOZŁOWSKA U., LIWO A. and SCHERAGA H. A., *J. Phys. Chem. A*, **107** (2003) 8035.
- [104] MAKSIMIAK K., RODZIEWICZ-MOTOWIDŁO S., CZAPLEWSKI C., LIWO A. and SCHERAGA H. A., *J. Phys. Chem. B*, **107** (2003) 13496.
- [105] CZAPLEWSKI C., OLDZIEJ S., LIWO A. and SCHERAGA H. A., *Protein Eng., Design & Selection*, **17** (2004) 29.
- [106] LIWO A., OLDZIEJ S., CZAPLEWSKI C., KOZŁOWSKA U. and SCHERAGA H. A., *J. Phys. Chem. B*, **108** (2004) 9421.

- [107] LIWO A., ARŁUKOWICZ P., OLDZIEJ S., CZAPLEWSKI C., MAKOWSKI M. and SCHERAGA H. A., *J. Phys. Chem. B*, **108** (2004) 16918.
- [108] OLDZIEJ S., LIWO A., CZAPLEWSKI C., PILLARDY J. and SCHERAGA H. A., *J. Phys. Chem. B*, **108** (2004) 16934.
- [109] OLDZIEJ S., LAGIEWKA J., LIWO A., CZAPLEWSKI C., CHINCHIO M., NANIAS M. and SCHERAGA H. A., *J. Phys. Chem. B*, **108** (2004) 16950.
- [110] LEE J., SCHERAGA H. A. and RACKOVSKY S., *J. Comput. Chem.*, **18** (1997) 1222.
- [111] LEE J., SCHERAGA H. A. and RACKOVSKY S., *Biopolymers*, **46** (1998) 103.
- [112] LEE J. and SCHERAGA H. A., *Int. J. Quantum Chem.*, **75** (1999) 255.
- [113] LEE J., RIPOLL D. R., CZAPLEWSKI C., PILLARDY J., WEDEMEYER W. J. and SCHERAGA H. A., *J. Phys. Chem. B*, **105** (2001) 7291.
- [114] CZAPLEWSKI C., LIWO A., PILLARDY J., OLDZIEJ S. and SCHERAGA H. A., *Polymer*, **45** (2004) 677.
- [115] KAZMIERKIEWICZ R., LIWO A. and SCHERAGA H. A., *J. Comput. Chem.*, **23** (2002) 715.
- [116] KAZMIERKIEWICZ R., LIWO A. and SCHERAGA H. A., *Biophys. Chem.*, **100** (2003) 261; 91 (Erratum).
- [117] LEE J., LIWO A. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **96** (1999) 2025.
- [118] LIWO A., LEE J., RIPOLL D. R., PILLARDY J. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **96** (1999) 5482.
- [119] LEE J., LIWO A., RIPOLL D. R., PILLARDY J. and SCHERAGA H. A., *Proteins: Struct., Function Genetics, Suppl.*, **3** (1999) 204.
- [120] LEE J., LIWO A., RIPOLL D. R., PILLARDY J., SAUNDERS J. A., GIBSON K. D. and SCHERAGA H. A., *Int. J. Quantum Chem.*, **71** (2000) 90.
- [121] OLDZIEJ S., CZAPLEWSKI C., LIWO A., CHINCHIO M., NANIAS M., VILA J. A., KHALILI M., ARNAUTOVA Y. A., JAGIELSKA A., MAKOWSKI M., SCHAFROTH H. D., KAZMIERKIEWICZ R., RIPOLL D. R., PILLARDY J., SAUNDERS J. A., KANG Y. K., GIBSON K. D. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 7547.
- [122] NANIAS M., CHINCHIO M., OLDZIEJ S., CZAPLEWSKI C. and SCHERAGA H. A., *J. Comput. Chem.*, **26** (2005) 1472.
- [123] NANIAS M., CZAPLEWSKI C. and SCHERAGA H. A., *J. Chem. Theory Comput.*, **2** (2006) 513.
- [124] LIWO A., KHALILI M., CZAPLEWSKI C., KALINOWSKI S., OLDZIEJ S., WACHUCIK K. and SCHERAGA H. A., *J. Phys. Chem. B*, **111** (2007) 260.
- [125] KHALILI M., LIWO A., RAKOWSKI F., GROCHOWSKI P. and SCHERAGA H. A., *J. Phys. Chem. B*, **109** (2005) 13785.
- [126] KHALILI M., LIWO A., JAGIELSKA A. and SCHERAGA H. A., *J. Phys. Chem. B*, **109** (2005) 3798.
- [127] LIWO A., KHALILI M. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 2362.
- [128] KHALILI M., LIWO A. and SCHERAGA H. A., *J. Mol. Biol.*, **355** (2006) 536.
- [129] JAGIELSKA A. and SCHERAGA H. A., *J. Comput. Chem.*, **28** (2007) 1068.
- [130] ROJAS A. V., LIWO A. and SCHERAGA H. A., *J. Phys. Chem. B*, **111** (2007) 293.
- [131] NÉMETHY G. and SCHERAGA H. A., *J. Chem. Phys.*, **36** (1962) 3382.
- [132] NÉMETHY G. and SCHERAGA H. A., *J. Chem. Phys.*, **41** (1964) 680.
- [133] NÉMETHY G. and SCHERAGA H. A., *J. Chem. Phys.*, **36** (1962) 3401.
- [134] NÉMETHY G. and SCHERAGA H. A., *J. Phys. Chem.*, **66** (1962) 1773; 2888 (Erratum).
- [135] GRIFFITH J. H. and SCHERAGA H. A., *J. Mol. Struct.*, **682** (2004) 97.
- [136] GRIFFITH J. H. and SCHERAGA H. A., *J. Mol. Struct.*, **711** (2004) 33.
- [137] STILLINGER F. H., *Science*, **209** (1980) 451.
- [138] OWICKI J. C. and SCHERAGA H. A., *J. Am. Chem. Soc.*, **99** (1977) 7403.

- [139] OWICKI J. C. and SCHERAGA H. A., *J. Am. Chem. Soc.*, **99** (1977) 7413.
- [140] SCHERAGA H. A., *J. Biomol. Struct. Dyn.*, **16** (1998) 447.
- [141] DYSON H. J., WRIGHT P. E. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 13057.
- [142] TANAKA S. and SCHERAGA H. A., *Macromolecules*, **10** (1977) 291.
- [143] MATHESON JR., R. R. and SCHERAGA H. A., *Macromolecules*, **11** (1978) 819.

This page intentionally left blank

Lectures on biomolecular energy landscapes

P. G. WOLYNES

*Department of Chemistry and Biochemistry and Center for Theoretical Biological Physics
University of California, San Diego, La Jolla, CA, USA*

Introduction

Biomolecules are complex and heterogeneous, being the outcome of aeons of evolution. Evolution appears to be opportunistic and to take advantage of many happenstance events. At the same time selection is not perfect and biomolecules retain many features that are characteristic of molecules with random sequence of similar composition. By merging these two views a powerful framework for understanding the thermodynamics, and dynamics and function of biomolecules over the last decade has emerged. The first key problem in biomolecular science is to understand how a protein molecule can self-organize into a shape capable of specificity in chemistry and in its interaction with partners. It is this protein folding problem that is the main subject of these lectures.

Many of the features of protein folding dynamics are like those of any complex disordered polymeric system. On the other hand, proteins behave like engineered or selected systems in two ways: their components, the amino acids, have been chosen from a family that generically allows certain local structural organizations to occur. Thus most naturally occurring amino acids are capable of forming helices with repeating symmetry under the right condition [1]. Also, the side chains of the amino acids are used in proteins are similar enough in size that, at low resolution, one can imagine packing them easily together in many combinations [2]. Thus many polymers made of these particular subunits could have interesting forms of local organization. On the other hand, forming a folded protein molecule with a three-dimensional arrangement precise enough to carry out selective catalytic functions is far from trivial. Chemists and biologists have already been partially successful in making such detailed designs [3]. But natural evolution has

been more successful [4] yielding polymers with the capability of self-organization. To understand protein folding then, we must understand the properties of proteins as disordered systems. Then we must try to uncover those nonrandom features of proteins that are essential for their folding.

In my lectures here I will cover both the basics of this problem and give some hints of emphasizing the lessons that can be passed on to those more generally interested in biophysics.

I will then describe some recent case studies that reveal the current thinking in the field.

1. – The basis of folding landscapes

Folding can be understood using the idea of energy landscapes, an approach with its roots in the statistical mechanics of glasses and phase transitions [1-4]. Early work in biopolymer energy landscape theory showed that while polymers made with randomly chosen sequences of amino acids should exhibit the very complex multi-exponential kinetics associated with glasses, the phenomenology of natural proteins suggested that they in fact have evolved to avoid this kinetic complexity. Experiment and theory show folding proceeds fairly directly to the native structure which is energetically very stable. Down-hill trajectories to the folded state are opposed primarily by chain entropy. While the landscapes for polymers with a randomly chosen order of amino acids are predicted to be rugged, the landscapes of natural proteins have been smoothed to resemble a funnel [5]. This funnel topology makes predicting the mechanism of folding easy once the structure is known. On the other hand, building an energy landscape from first principles to have these characteristics is far from trivial, but has actually been achieved to some extent in the laboratory. The mathematical basis for understanding folding focuses on the statistics of the energy landscapes for *finite-size systems*. I first review the primary elements of this energy landscape theory of folding, then, how it can be used to understand folding kinetics in the laboratory and to predict protein structure from sequence, a key issue in practical molecular biology.

2. – Random sequences

A random heteropolymer would be able to bind to itself in many complex ways. Thus a completely random heteropolymer would be found as a random coil making only very sparse three-dimensional contacts or would be found condensed into a globule in which many contacts are made. The globule would have a fluctuating set of adventitious contacts that would largely be hydrophobic. This globular form of condensed protein molecule would, however, remain fluid. The collapsed molecule would move between many distinct conformations taking advantage of stabilizing contacts in a variety of ways. On a nanoscale, the protein would be a malleable gel. Macroscale gels have complex kinetics. The subtleties of fine cooking derive primarily from manipulating the mechanical properties of the protein solutions of food. Sequence analysis of some

proteins shows that they may be intrinsically disordered. These intrinsically disordered sequences may actually be gossamer molecules, possessing transient and fragile structures like those gels found in cooking. In contrast, most working proteins, enzymes, etc. are more compact than this and are fairly rigid. Once sufficient stabilizing contacts are made in a gel-like self-interacting biopolymer, one would expect the time scales of rearrangement would continue to lengthen. A strongly cross-linked molecule would exhibit the characteristics of artificial rubbers or polymer glasses. A protein like this would be a nanoscale *glass*. It would thermally occupy only a small number of possible conformations that can best take advantage of the adventitious contacts to form low-energy structures. But even this is not like a typical enzyme because the low-energy configurations will be structurally quite distinct and quite various in structure. Again transitions between such structures of a compact collapsed *random* heteropolymer would yield complex kinetics displayed by glasses. In addition the kinetics change rapidly with temperature or environmental conditions. Such a species would not be robust during molecular evolution or even to environmental changes.

3. – The statistical energy landscape

Few studies of purely random heteropolymers with defined sequences have been made in the laboratory. The difficulty of studying them arises because they are not soluble and will “crash” out of the solution. On the other hand, the sketch of the behavior of individual random heteropolymers narrated above is well established by computer simulations that have confirmed the basic qualitative ideas of energy landscape theory. Many of these simulations do not accurately model the stereochemistry intrinsic to peptides. Yet these lattice or “minimalist” models do capture the essence of a polymer as merely a necklace of beads [6-10]. Analyzing lattice polymers is much simpler than analyzing more realistic models because each bead is positioned on a specific location of a crystal lattice allowing an exact enumeration of configurations. *Much as chess is simpler to analyze than real warfare, lattice proteins are easier to study than more realistic models.* Ultimately the discreteness of possibilities available to a lattice protein allows specific counting of states and thus makes the direct evaluation of entropies and a more-or-less complete survey of the energy landscapes possible. To mimic a protein the beads must have differing interactions that can model the heterogeneity of real protein chains. Generally a random lattice heteropolymer will be “frustrated.” This term, borrowed from the field of spin glasses, describes the conflicting ways residues may interact so the system does not “know” how to order [11] giving rise to the diversity of states mentioned above. For most randomly chosen sequences it is impossible to simultaneously satisfy the desire of each residue to be surrounded by its most stabilizing partners because the chain cannot be broken. The unbreakable chain connections generally prevent locally optimal arrangements of the three-dimensional partners.

The energy landscape of a random heteropolymer, like the landscape of structural glasses, ultimately resembles the most extreme case of energetic ruggedness, the so-called random energy model introduced by Derrida to model spin glasses [12]. The idea is that

due to the many conflicting intersections, each protein has a seemingly unpredictably varying random energy when its conformation is globally modified. In comparison to the pairwise additive lattice models, the actual nonpairwise additive nature of the solvent-averaged forces exaggerates this trend toward uncorrelated randomness. Mean-field calculations suggest the random energy model actually does capture the statistics of the low energy states of the random heteropolymer.

Since there are no correlations in the model, the microcanonical description of the model becomes quite straightforward. The total number of configurations, Ω_0 is related to the configurational entropy through the equation

$$(1) \quad \Omega_0 = e^{S_0/k_B}.$$

The only other parameter characterizing the REM landscape is the mean-square fluctuation in energy states ΔE^2 . This quantity scales with the protein's length if the protein is compact.

The energy of any individual state is chosen from a Gaussian distribution because the energy comes from a sum of many conflicting terms,

$$(2) \quad P(E) = \frac{1}{\sqrt{2\pi\Delta E^2}} e^{-E^2/2\Delta E^2}.$$

At equilibrium the thermally averaged probability of finding a stated energy E is given by the Boltzmann factor further weighting this Gaussian

$$(3) \quad P = \frac{1}{Z} e^{-E/k_B T_e} = \frac{1}{\sqrt{2\pi\Delta E^2}} e^{-(E-\bar{E})^2/2\Delta E^2}.$$

This distribution is a shifted Gaussian centered on the averaged thermal energy

$$(4) \quad \bar{E} = \frac{-\Delta E^2}{2k_B T}.$$

From this distribution it follows that ever deeper states in the landscape farther and farther out in the Gaussian distribution are sampled as the temperature is decreased. But eventually, a problem arises. The thermally sampled states rapidly drop in number. The entropy drops with temperature as does the energy. The entropy falls so rapidly without any knowledge of an impending catastrophe—a collision with the third law of thermodynamics. The entropy is the logarithm of the number of configurations with the thermal averaged energy.

$$(5) \quad S = k_B \log \Omega P(E) = S_0 - \frac{\Delta E^2}{2k_B T}.$$

How far can the drop in entropy predicted by this result (exact for the random energy model) go? A problem must clearly occur at the temperature, T_0 , given by

$$(6) \quad T_0 = \sqrt{\frac{\Delta E^2}{2S_0}}.$$

At this temperature the entropy vanishes. The formula would appear to actually give a negative entropy below T_0 ! This possibility would violate the Third Law of Thermodynamics and is avoided by the system undergoing a thermodynamic transition at T_0 . The polymer becomes trapped in at most a few states. The real histogram of states must show its “graininess” at this temperature. Only a few states (a number polynomial rather than in the length of the chain) are to be found in the tail of the distribution. The actual energy distribution of a random heteropolymer can resemble a Gaussian until we sample its very low energy states. For a fraction of random sequences, in fact, only a single thermally occupied state might be found below T_0 . Would this highly frustrated molecule have organized itself merely by obeying statistical mechanics? It is a question of time [13, 14]. The problem with the random landscape becomes apparent when we examine how long a molecule would take to fold or, more important, when we think about molecular evolution! The thermodynamics of the REM actually describes quite well liquids as they become glasses! [15] The apparent vanishing of configurational entropy in supercooled liquids has been called the Kauzmann paradox [16]. Laboratory experiments show the impending vanishing of the entropy causes slowing dynamics of a viscous liquid when it is strongly cooled.

One can analyze the kinetic consequences of the entropy crisis for REM directly. To do this one must impose a set of reaction rules that describe the way the configurational states of the heteropolymers can inter-convert by elementary chain motions [17-19]. The slow dynamics of the random energy landscape comes about because those specific configurations stable enough to act as traps will be surrounded by more typical states. As the temperature is lowered, the system will be found in one of these lower energy states, but these configurations have very few structural elements in common. The different low-energy structures are all actually compromises—they satisfy the frustrating conflicts as best as can be done but in very different ways. There is no “plan” to the path to lowest energy. Thus if at one moment the protein is found in a satisfactorily low energy state, there is no way that the molecule can tell that there is not a slightly deeper, still more stable state which it would rather be in, but which is rather far off in the landscape. Through undirected Brownian motion the molecule continuously tries to find lower energy states, largely unsuccessfully, but eventually a still lower state may be found. Diffusion on the most rugged landscape is just like diffusion on an absolutely flat but high-dimensional energy landscape with a single deep minimum.

The detailed kinetic paths of a globally random energy landscape are complex to describe but the typical rates can be described in a statistically simple way when the density of states is high enough. When the polymer is in such a particularly low energy state, to escape from it the molecule must jump to other states which are more typical

in their energy. Therefore the activation barrier for trap escape increases upon cooling because the average energy $\overline{E}$ goes down rapidly with temperature. Super-Arrhenius temperature dependence average rate is given by the equation

$$(7) \quad R = R_0 e^{-\Delta E^2/2(k_B T)^2}.$$

The super-Arrhenius decrease of the rate plateaus once the bottom of the landscape is at T_0 . At the transition temperature, however, escape takes the same amount of time as the rate reflects the time to search through all the states

$$(8) \quad R = R_0 e^{-S_0/k_B}.$$

The REM is really a worst-case analysis. For a *correlated* landscape the rate depends on the number of basins, not the number of configurations but Plotkin *et al.* have shown it is still an exponentially small rate of exploration [20,21]. Thus even though thermodynamics indicates that a typical random heteropolymer will ultimately occupy only a small part of its configuration space at a low enough temperature, this part of phase space becomes quite kinetically difficult to reach at those temperatures by means of unguided Brownian motion.

There are ways to “nickel-and-dime” the huge exponential search time. The random energy model is a caricature because most energy landscapes are correlated. Polymer configurations with only few body interactions will have fairly similar values of the energy if they look similar. This degree of correlation is actually not even totally relevant for polymers because changing a single dihedral angle can bring together very distant and large parts of the sequence that were not in contact before giving kinetic topological constraints making the problem worse. Studying simply correlated energy landscapes in which pairs of energy levels are correlated shows the search that is easier because there are large basins of attraction [20,21] but the rate still scales exponentially with N . If topological constraints are ignored the locality of interaction further correlates the landscape [22,23] and does lower the search time so that it scales exponentially not with N but with a fractional power of N .

Evolution also argues against the long-term viability of a random energy landscape for proteins. Suppose a protein had, despite the typical kinetic difficulties, nevertheless efficiently evolved to fold and function but still possessed a random, rugged landscape. This could happen by choosing sequences that have traps to escape from that happen to have escape routes with barriers lower than expected. The statistics of local landscapes shows such trap selection to be a possibility [24]. Such landscapes are called “buffed” since their specific irregularities have been removed (by evolution!). However, a mutation in such a protein could stabilize a trap giving a protein with new ground state structure. The new structure would be unlikely to function as the old one did. By calculating their probability of occurrence, Plotkin and Wolynes showed that such “buffed” energy landscapes while possible are a less likely evolutionary outcome than the scenario to be

discussed below—that proteins have evolved to be minimally frustrated and simply avoid the trap problem.

4. – The energy landscape of long evolved proteins

Rather than using the buffering strategy to evolve sophisticated folding pathways that avoid becoming trapped as is typical for frustrated systems, Nature has taken a different route: make sure self-organization can be thermodynamically carried out without worrying about being trapped at all. Natural proteins simply do not seem to be as highly frustrated as typical random heteropolymers would be. Proteins do not fold by gradual loss of entropy until you run out of states, as they do in the REM. When looked at carefully native structures of well-folded protein do not actually have the obvious energetic conflicts that we would expect in random systems. Major compromises of the rules of structural chemistry are generally absent in X-ray structures. Evolved protein structures are not so highly frustrated as the ground states of random sequences structures should be. Instead there are many local themes of consistency and symmetry between a given sequence and the structure it adopts. One route to consistency was highlighted early in the work of Go [25]. He postulated that local secondary structure forces were consistent with packing acting between distant residues [26].

Bryngelson and Wolynes generalized this specific mechanism to allow more general sorts of self-consistency (not just involving consistency between secondary and tertiary structure) [1]. This more general idea is called the “minimum frustration principle.” The use of a modified REM also allows a quantitative formulation of the principle; in fact, it is clear the mechanism of consistency does not apply to secondary structure specifically. Tertiary interactions by themselves must be consistent with each other if they are sufficiently numerous. Most obviously the core of a protein is largely hydrophobic and when hydrophilic residues are found in the core, these strains are usually compensated in some kind of salt bridge. Tertiary structure consistency alone guides the folding route though secondary structure biases do play a role. Using a model with both secondary and tertiary biases Saven and Wolynes have suggested that local signals give about one-third of the bias needed for kinetically reliable folding [27]. To quantify the notion of consistency, the key used by Bryngelson and Wolynes is the idea of a “stratified” statistical energy landscape [2]. In contrast to a glass, the energy landscape of a minimally frustrated system can be naturally divided into layers with common energetic properties. The states within each layer having a similar low energy also have a specified degree of geometrical similarity to the ground state. There are many ways to measure similarity. The similarity measure might be chosen to quantify how many dihedral angles are in the correct configuration. If pairing forces are dominant it is better to stratify by how many pair contacts are correctly made, that is, how many are the same as in the native structure. In a minimally frustrated system as the energy decreases, *i.e.* the deeper the layer is, the more the structure of a configuration in that layer resembles the native structure. The energy decreases faster as the global minimum is approached than would be expected for a random heteropolymer. Brownian motion thus preferentially yields

the native state. Quantitatively, the average slope of the landscape towards the native structure is larger than the energetic slope found approaching other minima.

Non-native contacts still make random contributions to the energy but not enough for the polymer to get trapped at thermal energies. The size of typical non-native energies can be predicted from the random energy model. But even with non-native (partial) frustrating interactions the extra slope of the landscape towards the native structure leads to rapid flow, much as in a funnel. For a real protein, this folding funnel is doubtless pockmarked with many overlapping mini-funnels reflecting small traps but these traps are shallow and can be easily escaped by thermal motion. It is not necessary for evolution to have completely eliminated these traps. Landscape ruggedness need not be zero for the protein to fold. The traps just must be shallow on the thermal scale. As we have seen the entropy and the ruggedness energy parameter suffice to describe the landscape of a random heteropolymer. A landscape for a *foldable* protein must additionally be characterized by a third parameter, the size of the stability or better, the “specificity” gap which measures the difference in energy of the global minimum from the typical random states. Because this extra stability guides the Brownian motion to the ground state can be used as a collective coordinate to describe the most important aspects of the protein folding reaction. This is shown in fig. 1.

Bryngelson and Wolynes treated kinetics of minimally frustrated proteins in a simple way. They grouped together states which have a common value of the collective folding coordinate. The average flow of probability between different strata in the folding funnel is then determined not just by the energy of that layer but by the gradient of a total free energy as a function of the collective coordinate measuring the depth of the layer. The total free energy contains a key entropy term and averages over all the states of a structure. In the BW analysis each stratum is approximated as a separate random energy model to account for the non-native interactions that can be formed at any level. The free-energy profile as a function of the collective coordinates can be estimated by using the random energy model results to average over the states in a layer. The resulting free-energy profile is

$$(9) \quad F(Q) = \overline{E}(Q) - \frac{\Delta E^2(Q)}{2k_{\text{B}}T} - TS_0(Q),$$

where Q measures the depth of the layer, *i.e.* similarity to the native structure. As for the completely random heteropolymer the configurational entropy at a specific depth of the funnel depends on both the steric constraints in S_0 and the remaining ruggedness $\Delta E^2(Q)$. By having a glass transition at a certain level in the funnel, unusual traps may appear *but* now they have significant structural similarity to the native state.

As the protein organizes the entropy loss is offset by the average energy gradient of the funnel $E(Q)$. By trading entropy for energy, the protein’s Brownian motion will lead it to the global minimum if the temperature is low enough. If there were no energy gradient or if the energy gradient remained near zero until a final stabilizing crash nearly at the folded structure (as some suggest still in the area of structure prediction [28]) the

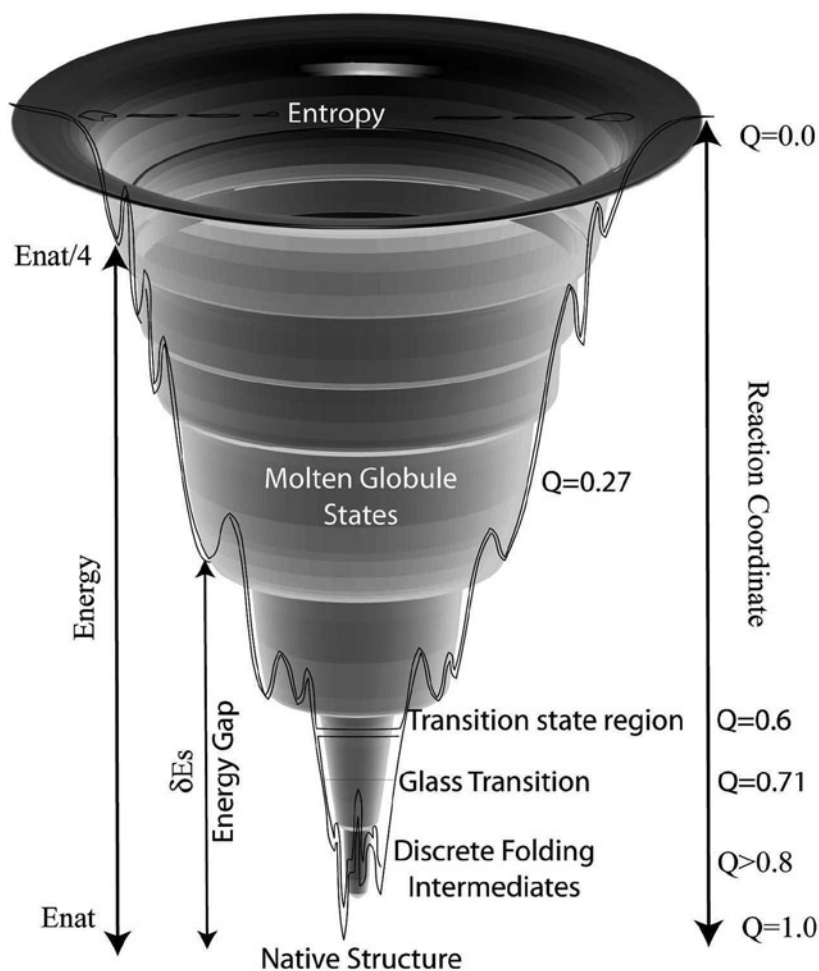


Fig. 1. – A schematic illustration of a typical small protein's folding funnel with its major landmarks. The radial coordinates express the entropy of an ensemble of states with a fixed value of the fraction of native structure Q , which is correlated with the energy E . The latter variables are expressed in the vertical axis. The near linear relation between energy and entropy means that free energy barriers in the profile are small on the scale of total energy. Notice that the transition state ensemble at $Q = 0.6$ occurs at a lower Q and therefore higher entropy than the glass transition at $Q = 0.71$, in this rough caricature. The transition state ensemble then is not configurationally unique in this case. Nevertheless the closeness of the two values of Q suggests that sometimes a unique transition state will be found when the Q of the transition state is higher. Discrete conformational substrates at high Q may sometimes be functional and can be thought of as conformational substates of the native protein or may represent kinetic traps.

resulting free-energy profile would exhibit a high entropic barrier giving again very slow search rates. For foldable heteropolymers the free-energy gradient provided by the funnel offsets the bulk of the huge entropy barrier allowing very rapid folding. Nowadays we know some proteins can fold in a few microseconds, while twenty years ago the seconds-to-hours timescale was usually quoted as a folding time. Thus in models, and perhaps in the laboratory only a small residual activation barrier actually remains due to the incomplete cancellation of entropy and energy through the folding process. While on the average the protein proceeds downward in specific energy towards the global minimum below T_F , the average flow may again be limited by trapping in the mini-funnels to structures having stable but incorrect non-native contacts whose local coordinates are orthogonal to the folding coordinate. These trapping events are a potential source of friction on the folding motions. Escaping from traps involves disentangling chains much as the slow motions in gels. The energy landscape theory quantitatively estimates the slowing by these transient traps by assuming each trap escape has its own collection reaction coordinate for unfolding from a collapsed structure. Since trap escape is an unfolding process energy landscape theory predicts the prefactor effect has an opposite temperature dependence to the funnel assisted folding itself. The total folding time which reflects both the mobility effects which get worse at low temperature and the thermodynamic effects which generally favor organization at low temperatures is therefore dramatically non-monotonic

$$(10) \quad \tau_F = D(T)^{-1} e^{\Delta F^\ddagger / K_B T}$$

because the diffusion constant has a temperature dependence given by eq. (7). Without trapping friction the thermodynamic effects alone would give rise to a folding time *vs.* temperature resembling a rectifier's response to a voltage [29] (see fig. 2). At high temperature the entropy barrier slows folding but folding becomes more downhill in a thermodynamic sense and faster at low temperature, finally being limited by the frictional slowing due to trapping. This characteristic non-monotonic behavior was first found in the analysis of Bryngelson and Wolynes. It has also been quantified carefully in many simulations. Socci, Onuchic and Wolynes showed these model simulations of protein folding could be analyzed using funnel reaction coordinates as Bryngelson and Wolynes suggested [30]. Because of the competition between funnel-guided folding trap escape on a rough potential, the temperature at which the energy gradient for fast folding can overcome the entropy gradient must exceed the glass transition temperature of collapsed states. Effective tunneling can be characterized then by the folding temperature being greater than the glass transition temperature of a random heteropolymer of the same composition. This limits the space of foldable sequences significantly. The landscapes of frustrated and funneled polymers are sketched in fig. 3. The statistical mechanical analysis of the energy landscape resolves the conceptual question of the quantitative meaning of "consistency," "harmony," etc. which figure in describing proteins from a structural viewpoint. The quantitative form of the minimal frustration principle can simply be stated " T_F exceeds T_G ." By evaluating these characteristic temperatures for realistic energy functions the hypothesis of minimal frustration (which is ultimately a biological

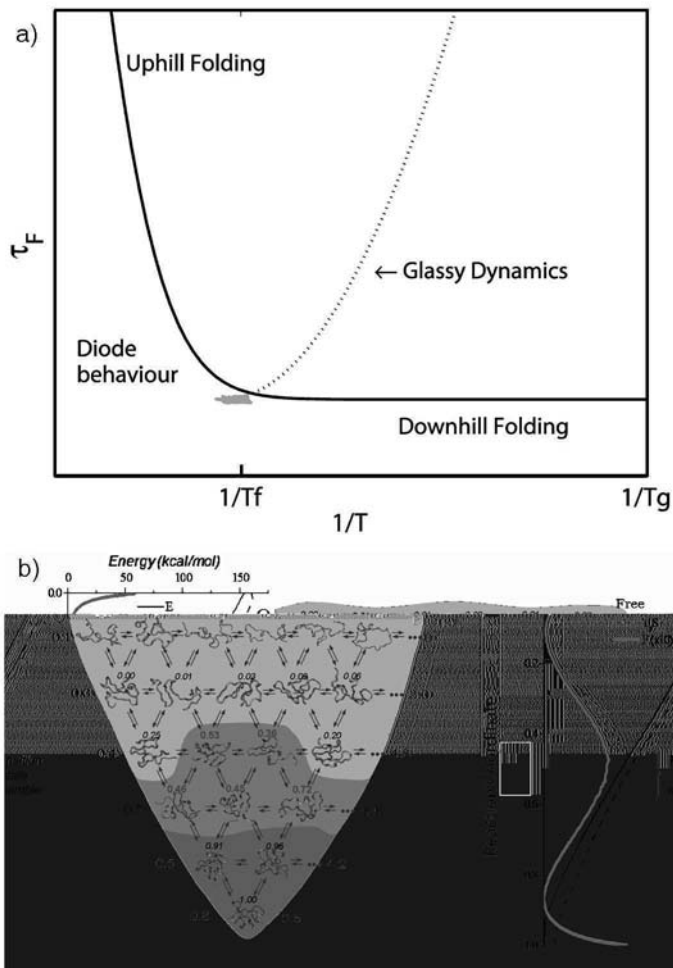


Fig. 2. – a) The folding time is a non-monotonic function of temperature. At high temperature folding is “uphill” due to entropy. If there is only a little ruggedness or nonadditivity in the energy law the folding may proceed downhill at high enough stability. Usually nonadditive forces or frustration contribute to the presence of an energy barrier and a slowing at low temperature. b) A variety of structures found in simulations of folding a small protein are shown superimposed on their locations in the folding funnel. The transition state structures (shown as a shaded region) are found in the intermediate Q range, as expected.

question) can be directly tested. Onuchic *et al.* estimated these temperatures from measurements of residual structure and dynamics of collapsed, but denatured protein [31]. On this basis the ratio of temperatures was $T_F/T_G = 1.6$. Chan has recently argued the ratio must actually be still larger than this estimate [32] if one is to describe the highly cooperative denaturation curves usually found for the most well-studied proteins.

One value of having derived this mathematical form of the “principle of minimal frustration” is that the quantitative formulation allows one to find in a systematic fashion

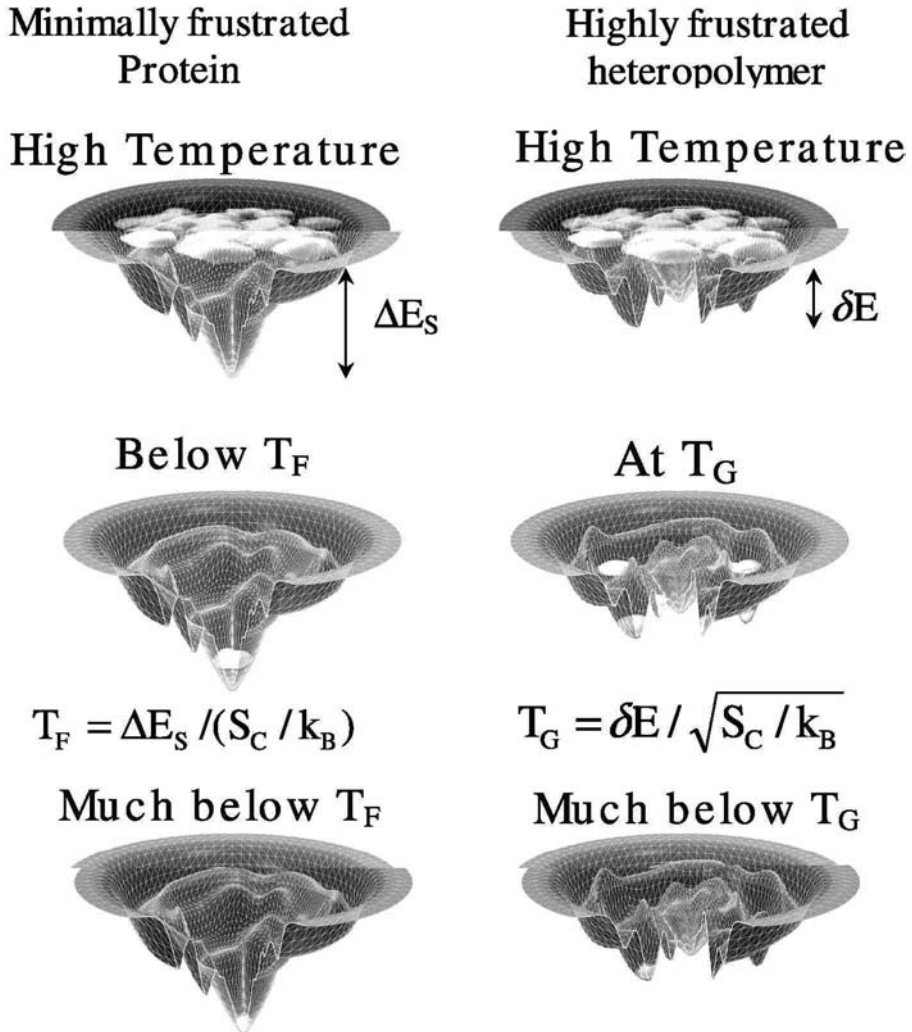


Fig. 3. – On the left is shown a series of funnels for a minimally frustrated protein with the thermally occupied states at different levels indicated by clouds. On the right the populated ensembles are superimposed on the rugged landscape of a typical random heteropolymer. The minimally frustrated system undergoes a first-order-like abrupt transition in the populations at a folding temperature T_F , while the random system undergoes a glass transition at T_G and configurational entropy vanishes leaving a small number of distinct conformations.

molecular potentials for predicting protein structure [33-36]. Also the T_F/T_G criterion can be used to design proteins that have been shown to successfully fold in the laboratory [35]. Theorists generally agree that minimizing frustration by considering as protein-like only those sequences and energy functions which yield T_F over T_G ratios bigger than 1 defines one of the main special features of proteins as distinguished from the majority of heteropolymers chosen at random.

The dominance of the funneling forces which follows from the minimum frustration principle suggests that the ensembles of structures characterizing partially folded natural proteins are best described by specifying which parts of the native structure are formed and which parts are not. Partially folded molecules that have some fraction of their native interactions formed have fewer opportunities to form frustrated non-native interactions. The weak cross linking caused by non-native contacts may slow folding, but only by a small factor. When the denatured state becomes strongly collapsed the friction effect increases. Such an effect has been observed and is called the “salt-induced detour” [37]. Even when specific non-native traps occur, the formation of the native contacts still must fight entropy, but because of minimal frustration this fight can go on locally and quasi-independently in many parts of the molecule. On a minimally frustrated funnel-like landscape the possible folding routes are numerous but these routes are not equally likely. Those routes that gain free energy of stabilization quickly but pay a low entropy cost are most important. Sometimes this greedy algorithm will necessitate backtracking, but this is a small effect. The dominant folding routes will often converge to a common region of phase space that will cause a bottleneck for reaching the folded state. This bottleneck in flow through configuration space is called the Transition State Ensemble (TSE). The free energy of this ensemble determines the folding rate [38].

Sometimes multiple bottlenecks occur as the protein organizes itself. There will then be low free-energy ensembles in between. These ensembles of structures would be described in chemical language as kinetic intermediates.

Since trap escape is predicted to be of secondary importance, the minimal frustration principle explains why the main features of folding kinetics can usually be predicted by knowing the stabilization energies of individual elements of the native structure and the entropic costs of bringing together parts of the scaffold. While stabilization is not uniform, heterogeneity of stability must be limited in magnitude or the protein would fray a great deal. Such fraying would allow aggregates to form easily. Thus on a minimally frustrated landscape to a first approximation a single energy scale and an entropy scale suffice to determine the dominant route of folding. These ingredients then must be functions of the contact pattern of the native protein alone. On funneled energy landscape only the “topology” of a protein in its native state determines the mechanism of its folding because these two energy scales must be balanced.

Different natural sequences having the same endpoint native structure, because of the funneled landscape, should exhibit only a handful of mechanisms for folding. Some of these will be more prominent in some thermodynamic conditions than others. This finite malleability of mechanisms describes many experiments [39] as we shall see below.

5. – Local *vs.* global descriptions of the folding landscape

The global energy landscape analysis tells us there is little energetic frustration (precisely $T_F > T_G$) so trapped states (prominent in the topomer search model) can be neglected.

But these globally consistent, minimally frustrated interactions are of finite range and act in real three-dimensional space. A local description of the energetics must be used to supplement the global funnel view. The mesoscale nature of proteins requires us to use both pictures at once. If the forces are short range, some parts of the protein need to locally assemble before the rest will inevitably fall into place. This is much like a crystallite nucleating a sample of supercooled liquid. The capillarity picture—a local version of the energy landscape theory—gives added detail to the description found using global collective coordinates. Even with short-range forces one can roughly find such global reaction coordinates by implementing capillarity theory in a structure-based model. The landscape paradigm remains intact in this local formulation because the free-energy profile can be found as the number of residues folded which neighbor each other, N_f . The free-energy profile contains a linear “bulk” term and an interfacial term scaling like $N_f^{2/3}$.

$$(11) \quad F(N_f) = (f_F - f_u)N_f + \gamma N_f^{2/3}.$$

From the macroscopic perspective the bulk term that scales with N_f reflects the free-energy difference per particle Δf between folded (f_F) and unfolded (f_u) protein; the small value of $\Delta f = f_F - f_u$ under folding conditions reflects the near cancellation of the entropy of unfolded state and the stabilizing energy of the native structure that is so familiar. In an early paper the “interface” term γ was taken by Bryngelson and Wolynes to be largely energetic and independent of the stability [40]. Putting in the typical stabilization of proteins under physiological conditions ($10k_B T$), BW incorrectly concluded that of the order ~ 100 residues would always need to be ordered at the folding transition state, a number comparable to a protein domain size. The error of their analysis was rather basic and apparent to others more skilled in thermodynamics of small systems. The correction to the error can be traced back to Kelvin’s work on the evaporation of small drops. The bulk transition temperature at which Δf vanishes does not coincide with the transition temperature of the cluster T_F unlike what Bryngelson and Wolynes thought. The freezing temperature is depressed by the surface contribution, as was known even to Lord Kelvin. We know that the initial and final full free energies must balance in folding—not their hypothetical bulk values of the free energies per particle! This was pointed out in the folding context by Finkelstein and Badredtinov [41, 42]. Taking account of this fact to rewrite eq. (1) one can show explicitly the free-energy profile referenced near the folding temperature again has few independent parameters,

$$(12) \quad F_{id}(\tilde{N}_f) = (-\tilde{\gamma} + \Delta H(T - T_F)/T_F)\tilde{N}_f + \tilde{\gamma}\tilde{N}_f^{2/3};$$

we find at T_F a crudely universal form having only one parameter essentially specifying

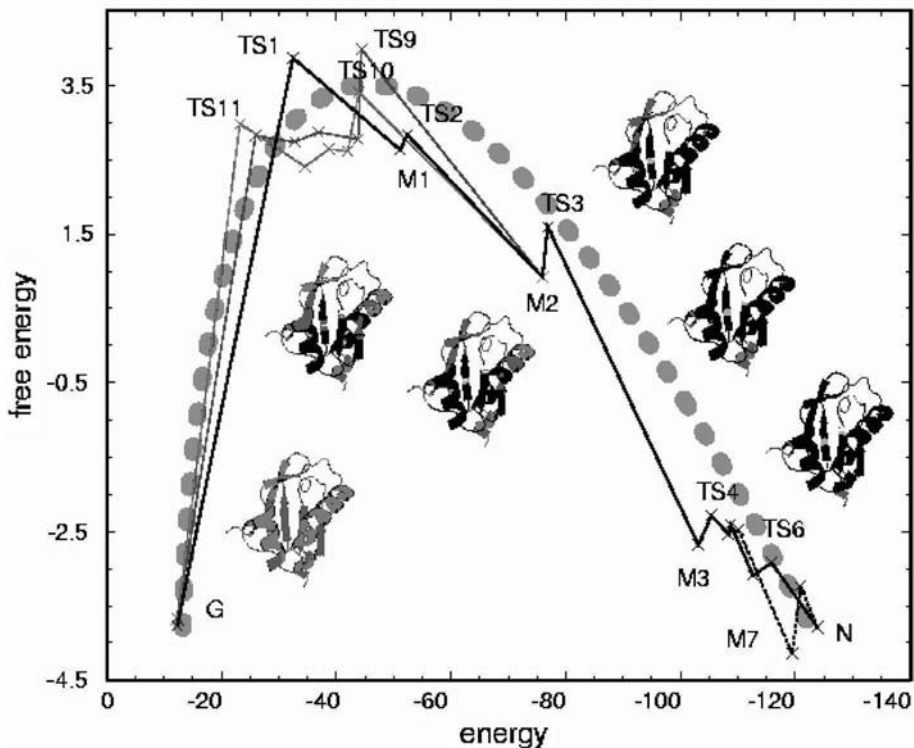


Fig. 4. – The free energy profile of UIA computed by a variational method from a minimally frustrated model is superimposed on the functional form expected from the crude capillarity picture. Notice that folding can be seen as the propagation of a front of ordered structure passing across the molecule. The motions within $k_B T$ at the barrier top roughly correspond to laying down a layer of ordered structures, usually. These will usually be diffusion influenced.

the barrier height. (This result is consistent with the homogeneous funnel model where energy scale should enter.) The temperature dependence of the stability enters through the enthalpy of unfolding ΔH . Despite the local interaction picture, the folded fraction $N_f = N_f/N$ emerges as a collective global coordinate for the folding process just as in the Bryngelson-Wolynes mean field approach. The specific numeric coefficients in ideal profile apply when the protein is nearly spherical and would differ for helical bundles or other nearly one-dimensional structures such as modular repeat proteins.

The free energy profile obtained by the “spherical cow model” universally implies a rather broad folding barrier. In fig. 4 we superimpose the spherical cow result on a detailed variational calculation for specific topology.

The transition state ensemble represents a broad ensemble. Structures within $k_B T$ of the barrier top should be included in the TSE. The breadth of the free-energy profile δN_{TST} is defined as the range over which the free energy changes by

$k_B T$. Thus one has $\delta \tilde{N}_{\text{TST}} = \sqrt{k_B T / (\delta^2 F / \delta^2 \tilde{N}^2)}$. At T_F this suggests the net number of residues needed to be moved to transit the scales like $N^{2/3}$ transition region $\delta N_{\text{TST}} \cong 2 \cdot \sqrt{2} / 3 N^{2/3} \sqrt{k_B T / \gamma}$. The folding profile is broader for longer chains, thus justifying a collective diffusive treatment of the chain motions for larger proteins, since elementary moves must typically displace a loop whose length will scale like $N^{1/3}$. We expect the dynamics of the front to be at the boarder of diffusion for proteins of the size of 100 residues, essentially a layer or so of protein must finally be laid down to commit the molecule to fold.

The interface between the “folded” and “unfolded” regions may have a complex structure. Thermal fluctuations and heterogeneity of native contact energy and entropy losses broaden the diffusing interface. Partially ordered regions may also wet the interface such as the liquid crystalline state in which helices are aligned but out of register [43] or where properly oriented structures existent in which there are plastic or rotating side chains. Such intermediately ordered phases will partially wet the interface between completely folded and unfolded states. Wetting reduces the interface energy γ . The resulting thermodynamic activation barrier may thus be smaller than anticipated. On the other hand, the expulsion of water may in some cases lead to a “drying” layer that could increase the interfacial energy [23, 44].

These wetting effects make it hard to predict the absolute magnitude of the folding barrier but the location of the interface can be well predicted because it depends mostly on balancing the large interactions against large entropies. The location of the transition state ensemble depends on the compromise between making strong local contacts and the pattern of entropy loss required by the topology of the final native structure. Theory can predict with some accuracy where the nucleation front will reach the critical size necessary to allow folding to continue downhill. The corresponding structural information can be found experimentally by protein engineering and kinetic φ value analysis [45, 46].

We see that energy landscape theory based on either local capillarity or the mean-field global pictures leads to a quasi-thermodynamic theory for folding based on a few reaction coordinates. These coordinates may be locally or globally defined. The resulting low-dimensional picture works because most of the elementary barrier-crossing steps describing the individual conformational transition of the chain come to a steady state before the folding bottleneck region is crossed. Due to this quasi-equilibrium, on an overall funnel-like surface, changes in protein stability become directly reflected in the observed kinetics [43, 45-47]. In small molecule chemistry, such extra-thermodynamic relations only hold for small perturbations. For folding, the collective nature of the reaction coordinate smoothes out the relationship between kinetics and thermodynamics and thus allows this simple relation to apply to much bigger changes in stability than in $k_e T$. In this way the energy landscape theory justifies in most cases the powerful ϕ value analysis approach pioneered by Fersht [48]. In contrast to minimally frustrated systems, if the landscape of proteins were very rugged, specific intermediates would form having specific, as opposed to fluctuating, non-native interactions. These intermediates would be found in the tail of the Gaussian energy distribution expected for a correlated random

energy model [20]. Changing any of the non-native contacts in these unpredictably structured intermediates would stabilize or destabilize specific non-native minima at random and qualitatively change the folding mechanism. If proteins were sufficiently frustrated, making kinetic predictions would be nearly impossible. In contrast, the funnel-like nature of the folding landscape allows a fairly direct structural interpretation of the φ values. Simulations confirm that this structural interpretation of ϕ values would break down dramatically for frustrated rough landscapes [49] because mutations would change the stability of idiosyncratically specific [50,51]. The fact that perfect funnel models based on the native structure of the protein are quite adequate at predicting the location of the key residues in the transition state for folding and the location of the capillary interface between folded and unfolded regions suggests the validity of the notion ruggedness is a small effect that can be roughly captured by the Bryngelson-Wolynes averaging into a dynamical prefactor.

Folding is not always a two-macrostate process. Nevertheless the entropy/energy imbalances intrinsic to any given protein topology often can provide all the information needed to determine the nature of folding intermediates and to predict which regions of the protein transition state ensembles are natively ordered. Energy functions based on perfect funnels do a good job of predicting the presence and structures of partially folded intermediates when they are observed [52].

Several seeming exceptions to the simple funnel picture have actually served to confirm our confidence in the energy landscape paradigm by predicting the presence of intermediates. The folding of cytochrome C [53] had sometimes been considered an exception to the funnel paradigm because of its fine structure in unfolding which was explored by Englanders' group using H/D exchange [54]. The location of the subunits properly folded in the intermediates is currently predicted, however, in detail by a perfect funnel model [55]. The quasi-independence of the different "foldons" found in this protein comes about because different fragments of the chain weakly contact each other. Instead, the mutual contacts between foldons and the heme provide the first main elements of the nucleation. The cooperation between different parts of the polypeptide chain needed in most proteins that do not have big cofactors is unnecessary in cytochrome because of the heme-side chain interactions.

There are exceptions to the naive single funnel paradigm—a notable example is the ROP dimer [56,57]. Kinetics and stability seem unrelated for ROP dimer folding. Some engineered mutations in this dimer speed up folding and unfolding. This observation cannot be reconciled with a quasi-equilibrium theory dominated by native-like interactions. Non-native interactions are needed. But it has been shown via simulations that this violation of the expected funnel behavior is a special (but by no means unique) case arising from the symmetry of the dimer that allows two near-degenerate, topologically distinct structures to compete on the energy landscape [58].

We learn something about the biology of this symmetry-induced frustration. All the engineered mutants bound by their target RNA *in vitro* yet they do not all function *in vivo*. The higher concentrations of RNA in the test tube allow binding to pull the equilibrium between the two degenerate structures over to the one that can bind. The

in vivo studies suggest that frustrated mutants, if they had arisen naturally, would likely lose out in natural selection. The ROP system may well be the felicitous “exception that proves the rule”—proteins need not have simple funnel landscapes—if they do not need to function.

Most biological functions, however, probably exploit frustration: multiple states of a protein provide switching capability (allostery) and may be needed sometimes to accommodate the multiple substrate conformations for processive enzyme transformations. Violations of the minimal frustration principle can often signal such functional constraints. Folding of proteins with complex functions will bring new challenges to energy landscape theory.

REFERENCES

- [1] BRYNGELSON J. D. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **84** (1987) 7524.
- [2] BRYNGELSON J. D. and WOLYNES P. G., *J. Phys. Chem.*, **93** (1989) 6902.
- [3] BRYNGELSON J. D., ONUCHIC J. N., SOCCI N. D. and WOLYNES P. G., *Proteins: Struct. Funct. Genet.*, **21** (1995) 167.
- [4] WOLYNES P. G., *Philos. Trans. R. Soc. London, Ser. A*, **363** (2005) 453.
- [5] LEOPOLD P. E., MONTAL M. and ONUCHIC J. N., *Proc. Natl. Acad. Sci. U.S.A.*, **89** (1992) 8721.
- [6] WOLYNES P. G., ONUCHIC J. N. and THIRUMALAI D., *Science*, **267** (1995) 1619.
- [7] LEVITT M. and WARSHEL A., *Nature*, **253** (1975) 694.
- [8] UEDA Y., TAKETOMI H. and GO N., *Biopolymers*, **17** (1978) 1531.
- [9] SHAKHNOVICH E. I. and GUTIN A. M., *Proc. Natl. Acad. Sci. U.S.A.*, **90** (1993) 7195.
- [10] DILL K. A., BROMBERG S., YUE K. Z., FIEBIG K. M., YEE D. P., THOMAS P. D. and CHAN H. S., *Protein Sci.*, **4** (1995) 561.
- [11] TOULOUSE, *Helv. Phys. Acta*, **57** (1984) 459.
- [12] DERRIDA B., *Phys. Rev. Lett.*, **45** (1980) 79.
- [13] SHAKHNOVICH E., FARZTDINOV G., GUTIN A. M. and KARPLUS M., *Phys. Rev. Lett.*, **67** (1991) 1665.
- [14] GUTIN A., SALI A., ABKEVICH V., KARPLUS M. and SHAKHNOVICH E. I., *J. Chem. Phys.*, **108** (1998) 6466.
- [15] WOLYNES P. G., *J. Res. Natl. Inst. Stand. Technol.*, **102** (1997) 187.
- [16] KAUZMANN W., *Chem. Rev.*, **43** (1948) 219.
- [17] WANG J., SAVEN J. G. and WOLYNES P. G., *J. Chem. Phys.*, **105** (1996) 11276.
- [18] WANG J., PLOTKIN S. S. and WOLYNES P. G., *J. Phys. I*, **7** (1997) 395.
- [19] SAVEN J. G., WANG J. and WOLYNES P. G., *J. Chem. Phys.*, **101** (1994) 11037.
- [20] PLOTKIN S. S., WANG J. and WOLYNES P. G., *Phys. Rev. E*, **53** (1996) 6271.
- [21] PLOTKIN S. S., WANG J. and WOLYNES P. G., *J. Chem. Phys.*, **106** (1997) 2932.
- [22] THIRUMALAI D., *J. Phys. I*, **5** (1995) 1457.
- [23] WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **94** (1997) 6170.
- [24] PLOTKIN S. S. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 4417.
- [25] GO N., *Annu. Rev. Biophys. Bioeng.*, **12** (1983) 183.
- [26] TAKETOMI H., UEDA Y. and GO N., *Int. J. Pept. Protein Res.*, **7** (1975) 445.
- [27] SAVEN J. G. and WOLYNES P. G., *J. Mol. Biol.*, **257** (1996) 199.
- [28] SCHUELER-FURMAN O., WANG C., BRADLEY P., MISURA K. and BAKER D., *Science*, **310** (2005) 638.

- [29] FEYNMAN R. P. and LEIGHTON R. B., *The Feynman Lectures in Physics*, Vol. **II** (Addison-Wesley, Reading) 1964.
- [30] SOCCI N. D., ONUCHIC J. N. and WOLYNES P. G., *J. Chem. Phys.*, **104** (1996) 5860.
- [31] ONUCHIC J. N., WOLYNES P. G., LUTHEY-SCHULTEN Z. and SOCCI N. D., *Proc. Natl. Acad. Sci. U.S.A.*, **92** (1995) 3626.
- [32] CHAN H. S., SHIMIZU S. and KAYA H., *Energy of Biol. Macromolecules E*, **380** (2004) 350.
- [33] GOLDSTEIN R. A., LUTHEY-SCHULTEN Z. A. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **89** (1992) 4918.
- [34] FUJITSUKA Y., TAKADA S., LUTHEY-SCHULTEN Z. A. and WOLYNES P. G., *Proteins: Struct. Funct. Genet.*, **54** (2004) 88.
- [35] JIN W. Z., KAMBARA O., SASAKAWA H., TAMURA A. and TAKADA S., *Structure*, **11** (2003) 581.
- [36] PAPOIAN G. A., ULANDER J., EASTWOOD M. P., LUTHEY-SCHULTEN Z. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U. S. A.*, **101** (2004) 3352.
- [37] OTZEN D. E. and OLIVEBERG M., *Proc. Natl. Acad. Sci. U. S. A.*, **96** (1999) 11746.
- [38] ONUCHIC J. N., SOCCI N. D., LUTHEY-SCHULTEN Z. and WOLYNES P. G., *Folding Des.*, **1** (1996) 441.
- [39] GRANTCHAROVA V., ALM E. J., BAKER D. and HORWICH A. L., *Curr. Opin. Struct. Biol.*, **11** (2001) 70.
- [40] BRYNGELSON J. D. and WOLYNES P. G., *Biopolymers*, **30** (1990) 177.
- [41] FINKELSTEIN A. V. and BADRETDINOV A. Y., *Mol. Biol.*, **31** (1997) 391.
- [42] FINKELSTEIN A. V. and BADRETDINOV A. Y., *Fold. Des.*, **3** (1998) 67.
- [43] LUTHEY-SCHULTEN Z., RAMIREZ B. E. and WOLYNES P. G., *J. Phys. Chem.*, **99** (1995) 2177.
- [44] LEVY Y. and ONUCHIC J. N., *Acc. Chem. Res.*, **39** (2006) 135.
- [45] FERSHT A. R., *Curr. Opin. Struct. Biol.*, **7** (1997) 3.
- [46] DAGGETT V. and FERSHT A., *Nat. Rev. Mol. Cell Biol.*, **4** (2003) 497.
- [47] OLIVEBERG M. and WOLYNES P. G., *Q. Rev. Biophys.*, **38** (2005) 245.
- [48] ITZHAKI L. S., OTZEN D. E. and FERSHT A. R., *J. Mol. Biol.*, **254** (1995) 260.
- [49] CLEMENTI C., CHEUNG M., NYMEYER H. and ONUCHIC J. N., *Biophys. J.*, **78** (2000) 46a.
- [50] CHAHINE J., NYMEYER H., LEITE V. B. P., SOCCI N. D. and ONUCHIC J. N., *Phys. Rev. Lett.*, **88** (2002).
- [51] NYMEYER H., SOCCI N. D. and ONUCHIC J. N., *Proc. Natl. Acad. Sci. U.S.A.*, **97** (2000) 634.
- [52] CLEMENTI C., JENNINGS P. A. and ONUCHIC J. N., *Proc. Natl. Acad. Sci. U.S.A.*, **97** (2000) 5871.
- [53] BAI Y. W., SOSNICK T. R., MAYNE L. and ENGLANDER S. W., *Science*, **269** (1995) 192.
- [54] MAITY H., MAITY M. and ENGLANDER S. W., *J. Mol. Biol.*, **343** (2004) 223.
- [55] WEINKAM P., ZONG C. H. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 12401.
- [56] WILLIS M. A., BISHOP B., REGAN L. and BRUNGER A. T., *Structure*, **8** (2000) 1319.
- [57] MUNSON M., ANDERSON K. S. and REGAN L., *Fold. Des.*, **2** (1997) 77.
- [58] LEVY Y., CHO S. S., SHEN T., ONUCHIC J. N. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 2373.

This page intentionally left blank

How organisms adapt to the environment: Sequence determinants of the habitat temperature and its physical rationale

K. B. ZELDOVICH, I. N. BEREZOVSKY and E. I. SHAKHNOVICH^(*)

*Department of Chemistry and Chemical Biology, Harvard University
12 Oxford St, Cambridge, MA 02138, USA*

There have been considerable attempts in the past to relate phenotypic trait —habitat temperature of organisms— to their genotypes, most importantly compositions of their genomes and proteomes. However, despite accumulation of anecdotal evidence, an exact and conclusive relationship between the former and the latter have been elusive.

We present an exhaustive study of the relationship between amino acid composition of proteomes, nucleotide composition of DNA, and optimal growth temperature of prokaryotes. Based on 204 complete proteomes of archaea and bacteria spanning the temperature range from -10 C to $+110\text{ C}$, we performed an exhaustive enumeration of all possible sets of amino acids and found a set of amino acids whose total fraction in a proteome is correlated, to a remarkable extent, with the optimal growth temperature. The universal set is Ile, Val, Tyr, Trp, Arg, Glu, Leu (IVYWREL), and the correlation coefficient is as high as 0.93. We also found that the G+C content in 204 complete genomes does not exhibit a significant correlation with optimal growth temperature ($R = -0.10$). On the other hand, the fraction of A+G in coding DNA is correlated with temperature, to a considerable extent, due to codon patterns of IVYWREL amino acids. Further, we found strong and independent correlation between OGT and frequency with which pairs of A and G nucleotides appear as nearest neighbors in genome sequences. This adaptation is achieved via codon bias. Further we analyze the physical reason for the

^(*) E-mail: eugene@belok.harvard.edu

observed amino acid composition bias and determine that this is due to positive design—that seeks to lower native state of proteins—and negative design that increases the energy of misfolded conformations. Together these two factors work to increase energy gap in proteins and therefore increase its stability. These findings present a direct link between principles of proteins structure and stability and evolutionary mechanisms of thermophylic adaptation.

1. – Introduction

As proteins and nucleic acids must remain in their native conformations at physiologically relevant temperatures, thermal adaptation requires adjustment of interactions within these biopolymers. Given the limited alphabet of amino acid residues, an apparent way to control protein stability is to properly choose the fractions of different residue types and then to arrange them in sequences that fold into and stable in unique native structures [1-4] at physiological conditions of a given organism. Various mechanisms of thermostability were discussed in the literature and many authors pointed out to changes in amino acid composition as one of the clearest manifestations of thermal adaptation [2, 4-8].

Indeed, it is well known that enhanced thermostability is reflected in specific trends in amino acid composition [9-11]. The most pronounced ones are in elevated fraction of charged residues [12-16], and/or increased amount of hydrophobic residues in (hyper)thermophilic organisms as compared to mesophilic ones. An early attempt of a systematic search for amino acids that are most significant for protein thermostability was made by Ponnuswamy *et al.* [17], who considered a set of 30 proteins and about 65000 combinations of different amino acids to find the amino acid sets serving as best predictors of denaturation temperature.

Furthermore, a number of authors explored possible relationship between an Optimal Growth Temperature (OGT) of organisms and nucleotide content of their genomes [16, 18, 19]. An increase in purine (A+G) load of bacterial genomes of some thermophiles was noted recently [20] as a possible primary adaptation mechanism.

However, despite the significant efforts, a clear and comprehensive picture of genomic and proteomic signatures of thermal adaptation remains elusive. First, it is not fully established which compositional biases in genomes and proteomes represent the most definitive signatures of thermophilic adaptation. On the genomic level is it G+C as widely believed by many after classical experimental study by Marmur and Doty [21] or alternative signatures such as purine loading index as suggested by others [19]? For proteomic compositions is it excess of charge residues or hydrophobic or both and which amino acids specifically are most sensitive to thermal adaptation? On a more fundamental level the key question is which factor—amino acid or nucleotide composition—is primary in thermal adaptation and which one is derivative? This issue is most vivid in prokaryotes whose genomes consist mostly of coding DNA, leading to a well-defined relationship between nucleotide and amino acid compositions. However this relationship may be not rigid due to degeneracy of genetic code. In principle amino acid and

nucleotide compositions can adapt, to a certain degree, independently due to possible application of codon bias.

Definite answers to these questions can be obtained only via a comprehensive study that considers, in a systematic way, all known prokaryotic genomes and all possible composition factors. Here, we carry out a comprehensive investigation of the relationship between the optimal growth temperature (OGT, T_{opt}) of prokaryotes and the compositions of their complete proteomes and compositions and pairwise nearest-neighbor correlations in their genomes. First, our goal is to find the sets of amino acids and nucleotides whose total content in a proteome (genome) serves as the best predictor of the OGT of an organism. We perform an exhaustive enumeration of all possible combinations of amino acid residue types, without making any *a priori* assumptions of the relevance of a particular combination for thermostability. Our analysis is based on 204 complete genomes of bacteria and archaea thriving under temperatures from -10°C to 110°C , comprising psychro-, meso-, thermo- and hyperthermophilic organisms. It reveals a particular combination of amino acids whose cumulative concentration in proteomes is remarkably well correlated with OGT. This result holds mainly for globular proteins comprising most parts of prokaryotic proteomes. For comparison, we apply this approach also to membrane proteins, having α -helical bundles [22] and β -barrels [23] structures, and show that the amino acids predictors of OGT found for proteomes of globular proteins does not work for membrane proteins. This finding clearly indicates that mechanisms of thermal stabilization of membrane proteins are different from those of globular proteins.

Next we turn to genomes and carry out the same comprehensive analysis to determine compositional genomic determinants of OGT. We find that purine load index, *i.e.* concentration of A+G exhibits highest correlation with OGT, consistent with an earlier observation made on several individual genomes [20]. Having found both proteomic and genomic compositional characteristics that correlate with OGT we then turn to the key question of whether genomic and proteomic determinants are independent ones or one is a derivative of the other. By running a set of reshuffling controls as described below we show that the primary factor is adaptation on the level of amino acid composition and the variation in DNA composition is largely a derivative of amino acid adaptation. While the nucleotide composition biases appear to be largely due to adaptation at the level of amino acid compositions, an additional DNA sequence adaptation that is independent of amino acid composition adaptation is still possible — at the level of higher-order correlation in nucleotide sequences. A possible mechanism for that in prokaryotes is through codon bias. Indeed we find a clear evidence for independent adaptation in DNA at the level of nearest-neighbor correlation of nucleotides (possibly due to strengthening of stacking interactions). We show that this adaptation of DNA sequences does indeed occur via codon bias.

Our next goal is to get insight into how observed patterns of change in amino acid compositions in response to extreme conditions of the environment are related to physical principles that govern stability and folding of globular proteins.

In terms of statistical physics the stability of the native state of a protein is determined by the Boltzmann factor $\exp[-\Delta E/k_{\text{B}}T]$, where ΔE is the energy gap between the native state and lowest-energy completely misfolded structures [24-28]. This factor imposes a

requirement that the energy gap must increase with temperature in order to preserve the uniqueness of the native state from destruction by thermal fluctuations. In principle, the growth of energy gap can be achieved by lowering energy of the native state (positive design), raising energy of misfolds (negative design), or both.

What mechanism does Nature use in its quest for thermophilic proteins? Our strategy to attack this important problem is as follows. In order to understand better the implication of the fundamental physical requirement of stability for protein sequences that evolve at elevated environmental temperature, we use the 27-mer lattice model of proteins. In this case, all compact conformations were enumerated [29] and, therefore, exact statistical-mechanical analysis is possible. Previously, protein thermodynamics [30, 31], folding [32, 33], and evolution [25, 34, 35] have been extensively studied by using this model. We simulate the process of thermal adaptation by the design of 27-mer sequences with selected (at a given environmental temperature T_{env}) thermal properties [27, 28]. The algorithm of design (see sect. 4) makes simultaneous unrestricted search in both conformational, sequence, and amino acid composition spaces. In our analysis we will focus on the amino acid composition of designed sequences as a function of the environmental temperature and we will compare the model findings with amino acid trends in real proteomes. Our main result is that thermal adaptation utilizes both positive design and negative design and we argue that adding amino acids from both extremes of hydrophobicity scale achieves exactly that goal —hydrophobic residues help with positive design while elevated concentration of charged residues helps to achieve stronger negative design. Further, we find an interesting and potentially important aspect of negative design: similar to positive design that strengthens certain native interactions, one manifestation of negative design is that it makes *specific non-native interactions* strongly repulsive. This in turn may lead to emergence of correlated mutations between amino acids that are not in contact in native structure.

2. – Results

2.1. Sequence determinants of thermal adaptation of soluble proteins. – We computed the $2^{20}-2 = 1,048,574$ values of the correlation coefficient R of the dependence between OGT $T_{\text{opt}}^{(j)}$ and the fractions $F^{(j)}$ for all possible sets of amino acids $\{a_i\}$, and found the set yielding the highest R . For the 86 proteomes with different OGT (see sect. 4) studied, the best set of amino acids is Ile, Val, Tyr, Trp, Arg, Glu, Leu (IVYWREL), giving the correlation coefficient between F_{IVYWREL} of a proteome and optimal growth temperature of the organism $R = 0.930$ (fig. 1). The quantitative relationship between the optimal growth temperature T_{opt} (in degree Celsius) and fraction F of IVYWREL amino acids reads

$$(1) \quad T_{\text{opt}} = 937F - 335.$$

The root-mean-square deviation between the actual and predicted OGTs for 86 genomes is 8.9°C. The same analysis performed on the complete set of 204 genomes yields KVV-

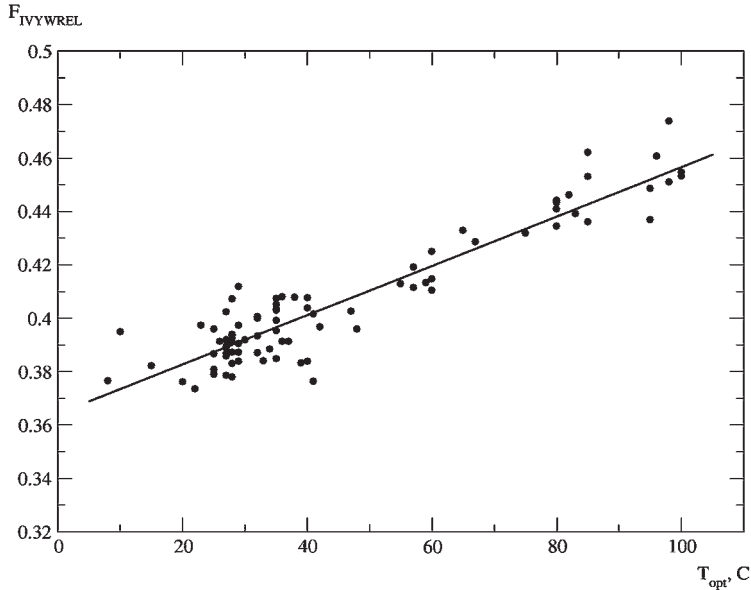


Fig. 1. – Correlation between the sum F of fractions of Ile, Val, Tyr, Trp, Arg, Glu, and Leu (IVYWREL) amino acids in 86 proteomes and the optimal growth temperature of organisms T_{opt} . The linear regression (red line) corresponds to the correlation coefficient $R = 0.93$. The optimal growth temperature T_{opt} (in degree Celsius) can be calculated from the total fraction F of IVYWREL in the proteome according to $T_{\text{opt}} = 937F - 335$. By construction, the IVYWREL set is the most precise predictor of OGT among all possible combinations of amino acids; other combinations statistically yield a larger error of prediction of OGT.

WREP as the best OGT predictor, with the correlation coefficient $R = 0.84$ and root-mean-square error of 9.2°C .

2.2. Statistical tests and controls. – We performed several tests to substantiate our procedure and to prove that the IVYWREL set is not a numerical aberration in a fit of noisy data, but a robust predictor of habitat temperature of prokaryotes.

First, as a control that the data are not overfitted, we randomly reshuffled the values of $T_{\text{opt}}^{(j)}$ between organisms, and, using the same exhaustive enumeration of all amino acid combinations, found the best composition-temperature correlation in the resulting artificially randomized proteome-temperature combinations. The distribution of the maximum correlation coefficient R_{max} obtained in 1000 such reshufflings follows a Gaussian shape centered around $\langle R_{\text{max}} \rangle = 0.344$, $\sigma = 0.060$. Therefore, the probability to find the observed correlation of $R = 0.93$ as a result of overfitting in a random combination of proteomes and temperatures is less than $p = 10^{-20}$, which proves an extremely high statistical significance of both the IVYWREL set and its correlation coefficient $R = 0.93$.

To demonstrate the robustness and stability of the IVYWREL predictor set, in fig. 2 we plot the distribution of correlation coefficients of the $F(T_{\text{opt}})$ dependence for different variations of the IVYWREL set. Figure 2 shows that the IVYWREL set is very tolerant

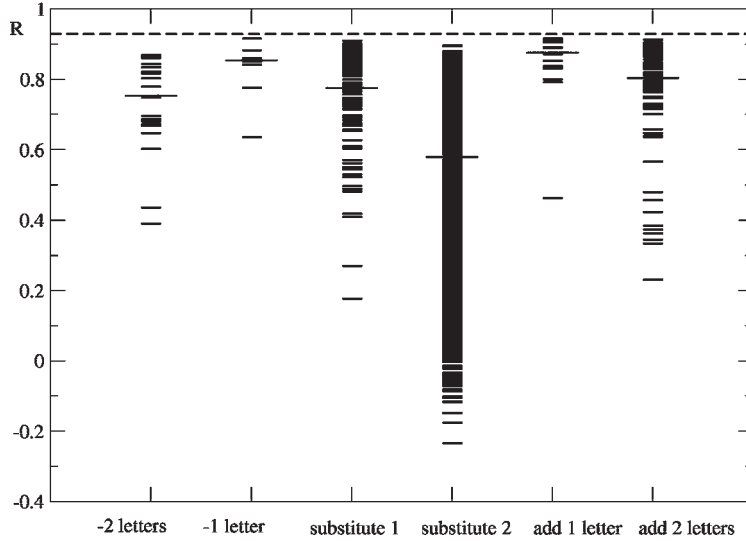


Fig. 2. – Distribution of the correlation coefficient R between OGT and fractions of amino acids in a proteome for different variation of the IVYWREL set, additions or deletions of one or two amino acids to/from the set, or substitution of one or two amino acids from the set by one or two amino acids not from the set. The dashed red line at $R = 0.93$ corresponds to the unperturbed IVYWREL. The horizontal red lines indicate the median values of the correlation coefficient for the given type of change of the predictor set.

to addition or removal of one amino acid, with the median value of R in excess of 0.8. Addition or deletion of two letters, or a substitution of one letter have a more significant effect on the accuracy of OGT prediction, with median R on the order of 0.75. Table I shows the examples of additions, deletions and substitutions in the IVYWREL set having the strongest or weakest effect on the correlation coefficient R .

To check the power of our method in the prediction of optimal growth temperature of prokaryotes, we randomly split the set of 86 proteomes into training sets of 43 and a test set of 43 species. We used the training set to determine the parameters of the best linear relationship between the fractions of all $2^{20} - 1$ amino acid sets and OGT, and employed this regression to predict OGTs of organisms in the test set. The accuracy of this prediction can be characterized by the root-mean-square difference $\sigma_{\Delta T}$ between predicted and actual OGT of the 43 organisms in the test set,

$$(2) \quad \sigma_{\Delta T} = \sqrt{\langle (\Delta T_j - \langle \Delta T \rangle)^2 \rangle}, \quad \Delta T_j = T_{\text{opt,predicted}}^{(j)} - T_{\text{opt}}^{(j)}.$$

Figure 3 presents the histogram of the values of $\sigma_{\Delta T}$ in 1000 test/train splits, with average $\sigma_{\Delta T}$ being equal to 12 °C for 43 organisms. This level of precision allows us to reliably discriminate between psychro-, meso-, thermo-, and hyperthermophilic organisms [2] knowing solely the amino acid composition of their proteomes.

TABLE I. – *Effects of the changes in the IVYWREL predictor combination on its predictive power. The plus sign means that an amino acid is added to the set, the minus sign indicates removal of an amino acid, “subst” indicates substitution.*

Type of change	R_{median}	Worst case		Best case	
		R_{min}	Change	R_{max}	Change
IVYWREL + 1	0.877	0.47	+A	0.921	+M
IVYWREL + 2	0.804	0.24	+AQ	0.917	+FP
IVYWREL – 1	0.855	0.64	–1	0.921	–W
IVYWREL – 2	0.754	0.40	–IE	0.874	–WE
IVYWREL subst. 1	0.776	0.18	E → A	0.914	W → H
IVYWREL subst. 2	0.580	–0.23	VE → AQ	0.902	WR → GP

Earlier, it has been suggested that the fraction of charged amino acid residues (Asp, Glu, Lys, Arg, DEKR) can be a significant predictor of living temperature [9,12-15]. We found that the fraction of these residues predicts the OGT with an accuracy (root-mean-square deviation) of 21 °C for 86 proteomes or 14 °C for 204 species. Similarly, a set of hydrophobic residues (Ile, Val, Trp, Leu, IVWL) predicts the OGT with an accuracy of

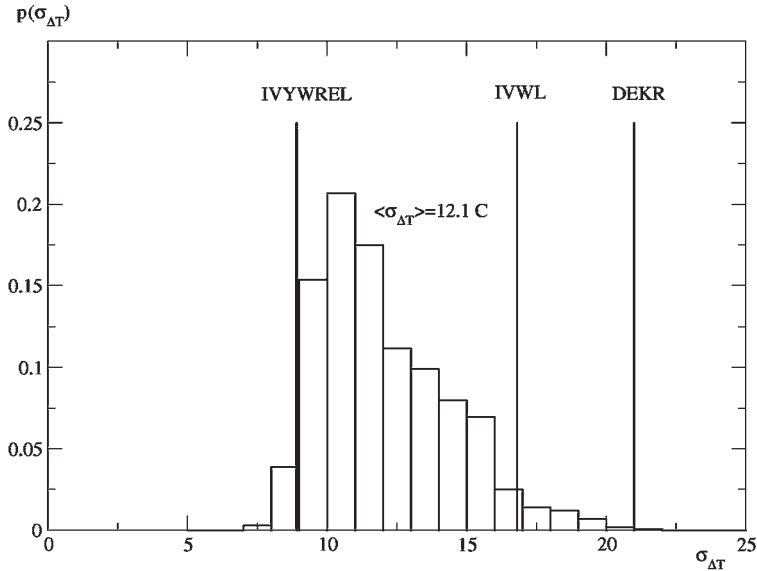


Fig. 3. – The distribution of the root-mean-square error $\sigma_{\Delta T}$ of the prediction of the OGT in the 1000 43-species test sets (black), and the prediction errors for 86 organisms using the IVYWREL set or sets of charged or hydrophobic residues. Individually, sets of hydrophobic or charged residues provide a much lower precision than their proper combination, IVYWREL.

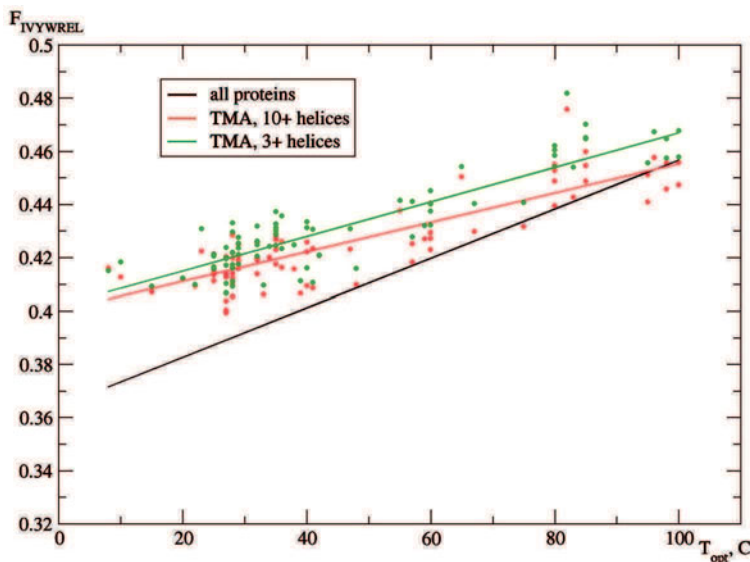


Fig. 4. – Total fraction of IVYWREL amino acids in alpha-helical membrane proteins containing three or more (green) or ten or more (red) helices, plotted against the OGT. Solid lines are linear regressions. The black line is the linear regression of the fraction of IVYWREL amino acids in all proteins in a proteome, same as in fig. 1. The fraction of IVYWREL in membrane proteins of thermophilic organisms is about the same as in all of their proteins. In mesophiles, membrane proteins are enriched with hydrophobic residues. For membrane proteins with ten or more helices, the five best predictors are ILVWYGERKP, FILVWYATSERKP, FILVWYATSEHRKP, IVYWREL, MFILVWYATSEHRKP, $0.85 < R < 0.86$.

16.8 °C for 86 proteomes. The IVYWREL predictor we discovered is accurate up to 8.9 °C for 86 proteomes. The result clearly demonstrates that consideration of both charged and hydrophobic residues is crucially important for predicting the living temperature and, thus, thermostability.

2'3. Membrane proteins and specific folds. – The calculation above made use of the complete proteomes, considering soluble and membrane proteins together. It is well known, however, that membrane proteins are markedly different from soluble ones, especially in terms of their stability and folding mechanism [36]. Interactions with the lipid bilayer presumably result in the mechanisms of thermostabilization of membrane proteins different from those of soluble ones. We used the TMHMM hidden Markov model [22] to identify α -helical membrane proteins in the proteomes of the 83 organisms, and performed the same analysis as for the complete proteomes. The IVYWREL combination is the best predictor of thermostability of membrane proteins with three or more transmembrane helices, $R = 0.89$, and is among the 5 highest- R predictors for membrane proteins with ten or more helices, $R = 0.85$ (see legend to fig. 4). In the latter case, the best predictor is IVYWRELGKP, $R = 0.86$. Figure 4 shows that the fraction of IVY-

TABLE II. – *Thermostability predictors for the major protein folds. Column 1, fold; column 2, number of species (out of 83) where proteins with that fold have been detected; column 3, average number of detected proteins per genome; column 4, maximum correlation coefficient between sum of fractions of amino acids and the corresponding amino acid combination; column 5, correlation coefficient between fraction of IVYWREL and OGT for the proteins with a given fold.*

Fold	Organisms	Proteins/Org	$R_{\max}$, predictor	R_{IVYWREL}
Beta barrel	83	79.2	0.91 IVWYAQERKP	0.87
Beta helix	59	20.3	0.90 ILVWYGERKP	0.81
Bundle	83	197	0.88 MILVWYANE	0.82
Globin	78	7.2	0.78 VWYGERKP	0.53
Cytochrome C	55	5.7	0.72 ILVWYGNQERKP	0.44
Ferredoxin	83	99	0.83 CMFILVWYGEHRKP	0.45
Lysozyme	50	4.2	0.72 CFILVWYGNDEP	0.50
Rossmann fold	83	292	0.90 ILWVERKP	0.86
Sandwich	82	27	0.85 FILVWYGNERKP	0.74
TIM-barrel	82	48	0.87 ILVWYGERKP	0.87

WREL amino acids in membrane α -helical proteins of mesophilic organisms is always higher than the overall IVYWREL fraction, whereas in hyperthermophilic organisms the fractions of IVYWREL in soluble proteins and in membrane α -helical bundles practically coincide. The weaker dependence of IVYWREL content on temperature suggests that the mechanism of thermostabilization of α -helical membrane proteins is different from that of soluble proteins (note however that IVYWREL compositions for soluble and α -helical membrane proteins appear to converge in hyperthermophiles). Next, we used the ProFTMB [23] server to identify transmembrane β -barrel proteins in the proteomes of 29 organisms uniformly covering the temperature scale. The most predictive amino acid combination for thermostability of transmembrane beta-barrels is CVYP, $R = 0.72$, while IVYWREL is a very poor predictor, $R = 0.35$. The low slope of IVYWREL correlation with OGT in α -membrane proteins and the CVYP predictor in TM β -barrels suggest that thermal adaptation in membrane proteins is governed by different rules than in globular ones (and probably different between different types of membrane proteins).

Table II shows that thermostability predictors of most of the folds are very similar to universal combination IVYWREL. The most abundant protein folds, such as α/β - and β -Barrel, Rossmann fold, and bundle, reveal a high-correlation coefficient between IVYWREL content in their sequences and OGT. Importantly, protein folds that invoke complementary mechanisms of stability, involving heme- and metal-binding (globin, cytochrome C, ferredoxin) or S-S bridges (lysozyme), show a significantly lower correlation between IVYWREL content and OGT.

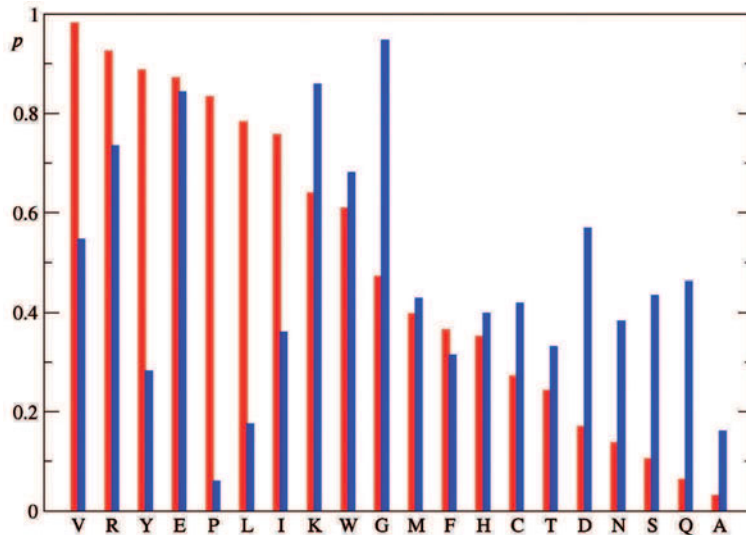


Fig. 5. – Histogram of the probability to find an amino acid among 1000 combinations of amino acids which are most correlated with OGT for real proteomes (red) and for artificial proteomes created from reshuffled DNA sequences (blue). The histogram for real proteomes supports the stability of IVYWREL predictor, while the difference between the two histograms suggests that amino acid biases upon thermophilic adaptation are not a consequence of the trends in nucleotide composition.

2.4. IVYWREL is not a consequence of nucleotide composition bias. – Proteins are encoded in the nucleotide sequences of their genes, and thermal adaptation presumably leads to increased stability of both proteins and DNA. Therefore, signatures of thermal adaptation in protein sequences can be due to the specific biases in nucleotide sequences and vice versa. In other words, one has to explore whether a specific composition of nucleotide (amino acid) sequences shapes the content of amino acid (nucleotide) ones, or thermal adaptation of proteins and DNA (at the level of sequence compositions) are independent processes.

In order to resolve this crucial issue, we applied the following logic. If amino acid biases are a consequence of just nucleotide biases and not protein adaptation, then proteomes translated from randomly reshuffled genomes will feature similar “thermal adaptation” trends in amino acid composition as observed in real proteomes. In contrast, if amino acid compositions are selected independently, then such control calculation will result in apparently different amino acid “trends” in randomly reshuffled genomes than observed in reality. We carried out this computational experiment by randomly reshuffling nucleotides in the coding sequences of 83 genomes and translating the reshuffled genomes into protein sequences. Then, we calculated the amino acid compositions of the resulting control proteomes, and found (using the same exhaustive enumeration of all amino acid combinations) the amino acid combinations that are best correlated with temperature. In fig. 5 we present the probability to find an amino acid in 1000 highest- R

combinations for real proteomes and control proteomes created from reshuffled DNA. The striking difference between the two plots suggests that the increase of IVYWREL with environmental temperature is not a consequence of the nucleotide composition bias of coding DNA. This result proves the existence of a specific adaptation pressure acting on the amino acid composition of proteins in the process of thermal adaptation.

2'5. Thermal adaptation of DNA. – In fig. 6a, we plotted the fraction of the proteomic thermostability predictor, IVYWREL, against the G+C content, a major contributor to pairing interactions in DNA, and a presumable indicator of DNA thermostability. A very weak negative correlation ($R = -0.14$) between the fractions of IVYWREL in the proteome and that of G+C in the coding DNA suggests that protein thermostability and its IVYWREL predictor are not consequences of enhanced GC content.

We used a complete enumeration of the $2^4 - 2 = 14$ possible sets of nucleotides to look for possible composition-OGT relationships in the coding DNA of prokaryotes (on average, coding DNA constitutes 85% of prokaryotic genomes). We found that the fraction of G+C in the complete genomes of 83 (fig. 6b) or 204 species does not show any significant correlation with optimal growth temperature. The same holds true for complete genome sequences, including noncoding DNA (data not shown).

The only combination of nucleotides whose fraction is statistically significantly correlated with temperature is purine composition, A+G, $R = 0.60$ (fig. 7a), in agreement with earlier observations made for several genomes [19]. As this correlation is very significant, there is a possibility that the trends in amino acid and nucleotide composition are tightly linked: either the proteomic predictor, IVYWREL, is a direct consequence of the A+G nucleotide bias (the possibility that we already discarded, see above), or vice versa, the A+G nucleotide bias automatically follows from the prevalence of IVYWREL in the proteomes of thermophilic organisms.

2'6. Purine loading bias is mainly due to IVYWREL. – To what extent can the increase of A+G with temperature be explained by the trends in amino acid composition? Obviously, the only way to adjust nucleotide sequences without affecting the encoded proteins is through the use of codon bias. A particularly important question is whether a specific codon bias is required to reproduce the trends in DNA composition.

To answer this question, we reverse-translated the protein sequences of the 83 organisms into DNA sequences without codon bias, *i.e.* by using synonymous codons with equal probabilities. As shown in fig. 7b, the fraction of A+G in the resulting nucleotide sequences is very significantly correlated with the optimal growth temperature, $R = 0.48$, which is very close to the corresponding value in actual genomes, $R = 0.60$ (fig. 7a). In other words, the amount of variation of A+G explained by the peculiarities of amino acid composition is almost the same with and without codon bias. (We note however that the slopes of dependences in figs. 7a,b are somewhat different suggesting that codon bias may be partly responsible for the overall purine composition of DNA.) Therefore, we conclude that the fraction of A+G in the coding DNA is largely defined by the composition of proteins. Indeed, by imposing the correct amino acid composition and choosing

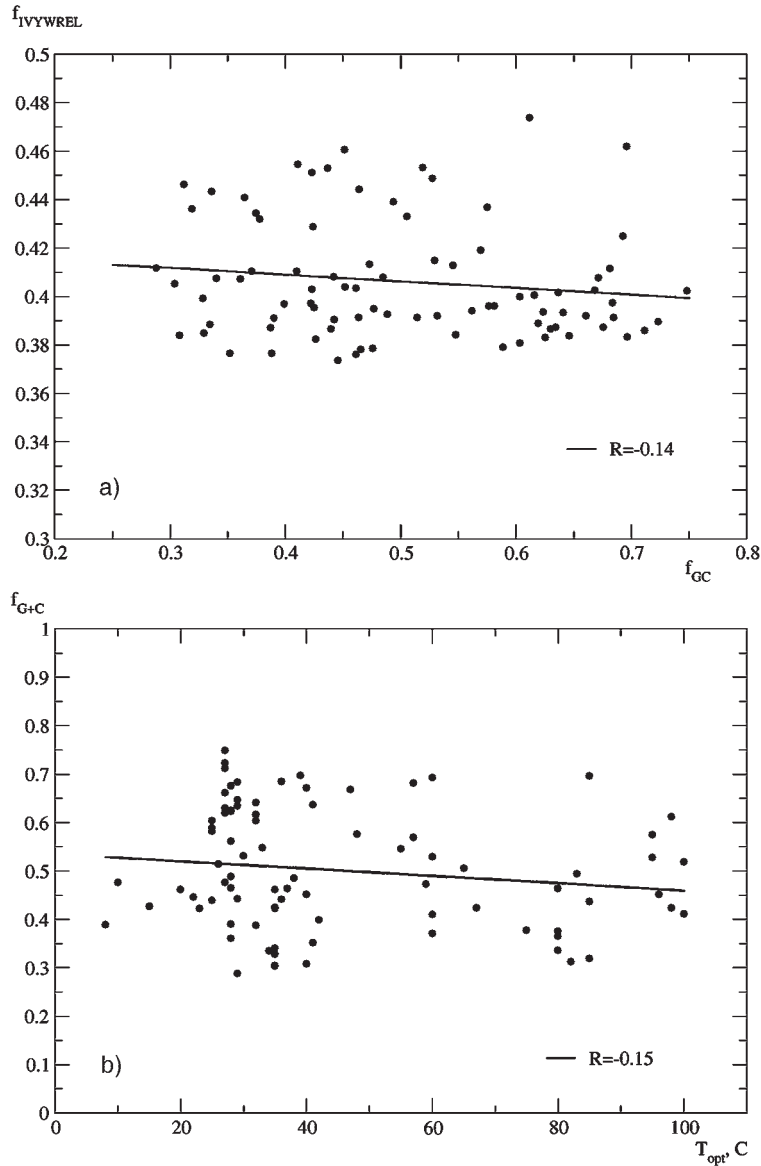


Fig. 6. – (a) Dependence of the fraction of IVYWREL amino acids in 83 proteomes (protein thermostability predictor) on the fraction of G+C in the coding DNA in the corresponding genomes. (b) Dependence of the G+C content in the coding DNA of the 83 complete genomes on the optimal growth temperature of the organisms. The correlation coefficient is $R = -0.15$, indicating that G+C content of the coding DNA is not related to the optimal growth temperature.

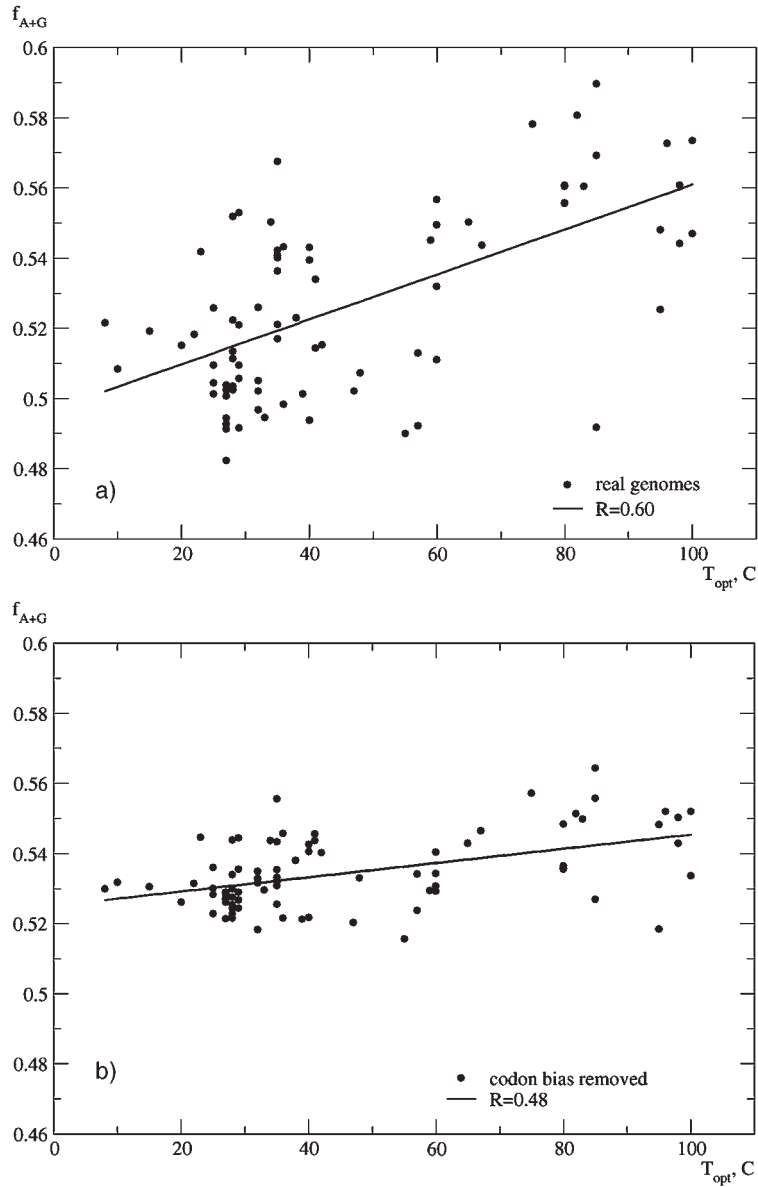


Fig. 7. – (a) The fraction of A+G in the coding DNA of the 83 complete genomes is highly correlated with the optimal growth temperature, $R = 0.60$. (b) When protein sequences of the 83 organisms are reverse translated into DNA without codon bias, the fraction of A+G remains correlated with OGT, $R = 0.48$.

the available synonymous codons with equal probabilities, one arrives at the correct prediction of the trend in DNA composition, increase of A+G with temperature. Together with the apparent irrelevance of G+C content for thermal adaptation on the organism scale, this finding suggests that on the level of nucleotide composition, direct selection pressure on DNA sequence composition appears to be weak.

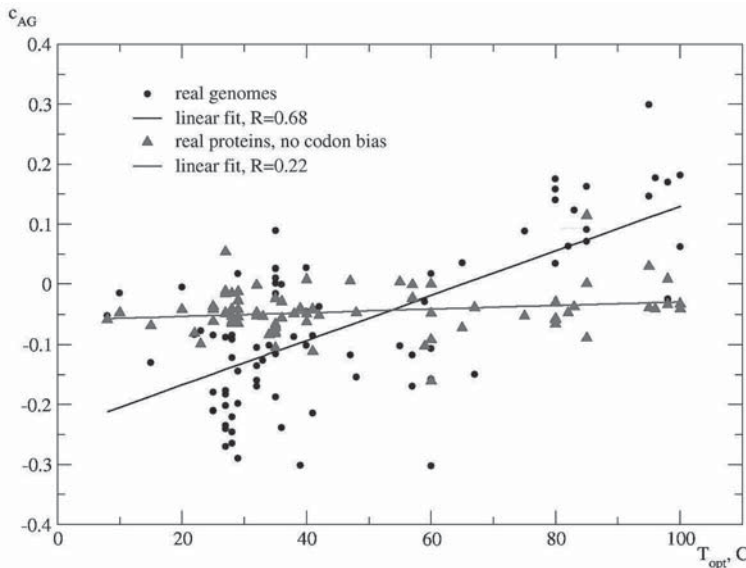


Fig. 8. – Dependence of the pairwise nearest-neighbor correlation function for A, G nucleotides c_{AG} on the optimal growth temperature in the genomes of 83 species (black) and in the DNA sequences obtained from the proteomes of 83 species without codon bias. While codon bias is not essential for reproducing the trends in nucleotide composition (fig. 6), nucleotide correlations are entirely dependent on the proper choice of codon bias. An increase of the number of ApG pairs enhances the stacking interactions in DNA, stabilizing it at elevated environmental temperatures.

2.7. Nearest-neighbor correlation in DNA sequences. – After nucleotide composition, the next level of description of DNA sequences is the pairwise nearest-neighbor correlation function, or normalized probability to find successive pairs of specific nucleotides. Although sequence correlations have a minor effect on base pairing, they determine the strength of the stacking interactions, where successive A and G nucleotides (ApG pairs) have a low energy, stabilizing the DNA [37, 38].

For each of the 83 genomes, we calculated the correlation function c_{ij} (see sect. 4) all 16 possible successive pairs of nucleotides in the coding strand. For each pair, we plotted its c_{ij} value in the 83 genomes against the OGT, and calculated the correlation coefficient, see table III, column 1. It turns out that the correlation function c_{AG} , or the excess probability to find successive pairs of A and G nucleotides (ApG pairs) in the coding DNA is significantly increasing with OGT, $R = 0.68$. The increase of c_{AG} with temperature is observed for both coding DNA and the complete genome sequence.

TABLE III. – *Correlation coefficients R between OGT and DNA sequence correlation function c_{ij} for the 16 combinations of nucleotides i, j in 83 genomes. Column 1, actual DNA sequences; column 2, DNA sequences obtained from reshuffled protein sequences retaining the actual codons used for every amino acid; column 3, DNA sequences obtained from reshuffled proteins without codon bias. ApG pairs are the most correlated with OGT in real genomes, but this property vanishes if codon bias is removed. Each of the columns is ordered by the value of R .*

Real DNA	Reshuffled proteins, real codon bias	No codon bias, reshuffled proteins
−0.589 TG	−0.597 TG	−0.825 CA
−0.515 CA	−0.459 AA	−0.698 TG
−0.458 GC	−0.458 CG	−0.563 GC
−0.443 AT	−0.456 GC	−0.437 AC
−0.432 CG	−0.452 AT	−0.355 AT
−0.302 AA	−0.388 CA	−0.127 TT
−0.202 GT	−0.244 GT	0.082 GG
−0.076 AC	0.010 TC	0.177 AG
0.092 TC	0.097 AC	0.216 CT
0.167 TT	0.187 TT	0.275 TC
0.200 GA	0.208 GA	0.334 GA
0.392 TA	0.446 GG	0.343 TA
0.479 GG	0.456 TA	0.396 AA
0.558 CC	0.567 CC	0.417 GT
0.601 CT	0.574 CT	0.443 CG
0.680 AG	0.736 AG	0.868 CC

We have shown above that codon bias does not define the dominant trends in nucleotide composition of coding DNA. Is codon bias necessary to explain the observed sequence correlations in coding parts of DNA, given that amino acid bias is established independently? To answer this question, we reverse-translated the actual protein sequences into nucleotide sequences, using synonymous codons with equal probabilities.

In principle, there are two possible sources of the correlations in DNA sequences, one stemming from the neighboring nucleotides within a codon, and another one stemming from the combination of nucleotides at the interface of successive codons. The latter possibility implies the dependence of correlations in DNA on the sequence correlations in proteins. Reshuffling of protein sequences while retaining the actual codons used for each amino acid removes the effect of codon interface (table III, column 2). These data

show that ApG pairs are still highly correlated with OGT, and correlations in protein sequences have a small effect on the correlations in DNA.

After removal of codon bias, ApG pairs are no longer correlated with OGT, $R = 0.20$ see table III, column 3 and fig. 8. Therefore, the correlation between the ApG pairs and temperature is entirely due to the evolved codon bias. Importantly, C/T pairs in coding sequences of DNA are also highly correlated with OGT. As we analyze correlations in the sense strand of DNA molecule, abundance of C/T pairs points out to the enrichment of anti-sense strand with ApG ones. Therefore, double-stranded DNA is stabilized by stacking interactions provided by ApG pairs that are spread in different locations of both sense and anti-sense strands. This conclusion holds for the whole DNA when both coding and non-coding parts are considered. Therefore, it appears that the crucial role of codon bias is to increase the number of ApG pairs in coding DNA in response to elevated environmental temperatures, enhancing the stacking interactions in DNA. We also note that the trend to increase the frequency of ApG pairs via codon adaptation may be another factor (besides amino acid adaptation, see above) that gives rise to increased overall composition of purine nucleotides A+G in hyperthermophiles, as can be seen from comparison of slopes of scatter plots shown in figs. 7a,b.

3. – Physical principles of protein design and the origin of IVYWREL

Now we turn to the analysis as to why the particular combination of amino acids —IVYWREL— serves as a predictor of OGT and why higher thermostability requires greater proportion of both charged and hydrophobic groups.

As a first step in this direction we design proteins with selected thermostability in a simple exactly solvable lattice model where proteins are presented as 27-mer sequences placed on cubic lattice. Such design is a first step towards modeling thermal adaptation of organisms, as there is a direct connection between OGT (environmental temperature, or T_{env}) of an organism and the melting temperature of its proteins [39,40].

We used the P-design procedure to create model 27-mer sequences that are stable at selected T_{env} (see sect. 4). We designed sets of 5000 model proteins for each T_{env} in the range $0.3 < T_{\text{env}} < 0.8$ in Miyazawa-Jernigan dimensionless units. The average melting temperature $\langle T_{\text{melt}} \rangle$ of lattice proteins is strongly correlated with T_{env} suggesting that the P-design procedure does work. It provides model proteins with enhanced stability in response to the increase of environmental temperature. The dependence of $\langle T_{\text{melt}} \rangle$ on T_{env} is close to linear and qualitatively matches the empirical linear relationship, $T_{\text{melt}} = 24.4 + 0.93 T_{\text{env}}$, between the average living temperature of the organism and melting temperature of its proteins [39].

As expected, the amino acid composition of designed proteins does depend on T_{env} for which they were designed. To quantify the differences between “low-temperature” and “high-temperature” amino acid compositions, we plotted a temperature dependences of the fractions of hydrophobic (LVWIFMPC), weak hydrophobic and polar (AGN-QSTHY), and charged (DEKR) amino acids for designed lattice proteins (fig. 9(a)) and natural (fig. 9(b)) proteomes. Figure 9(a) shows a significant increase of the amount of

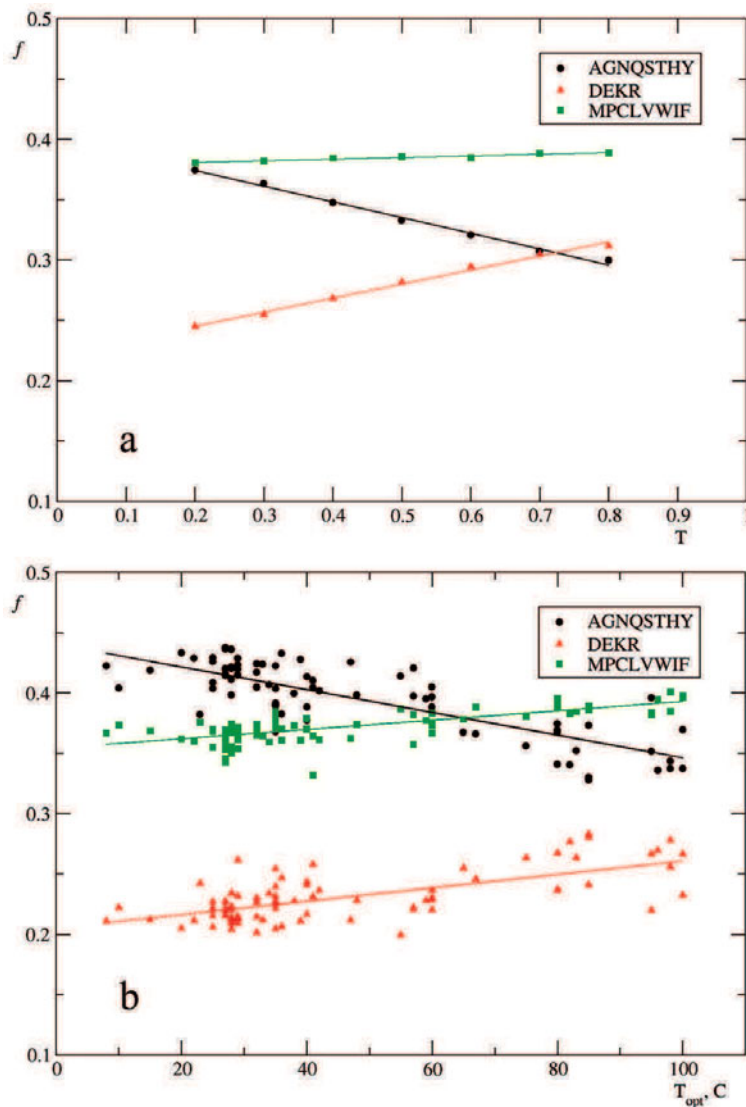


Fig. 9. – (a) Temperature dependences of the fractions of hydrophobic (LVWIFMPC), weak hydrophobic and polar (AGNQSTHY), and charged (DEKR) amino acids plotted against temperature T_{env} in the design experiment. The data received in P-design procedure applied to sets of 5000 27-mer sequences with random amino acid composition, using a different value of T_{env} for each set, $0.3 < T_{env} < 0.8$ (T_{env} is measured Miyazawa-Jernigan dimensionless units). (b) Temperature dependences of the fractions of amino acids in natural prokaryotic proteomes plotted against optimal growth temperature (OGT) of the organism. There are total 83 natural proteomes with optimal growth temperatures spanning interval from -10 to $+110^{\circ}\text{C}$.

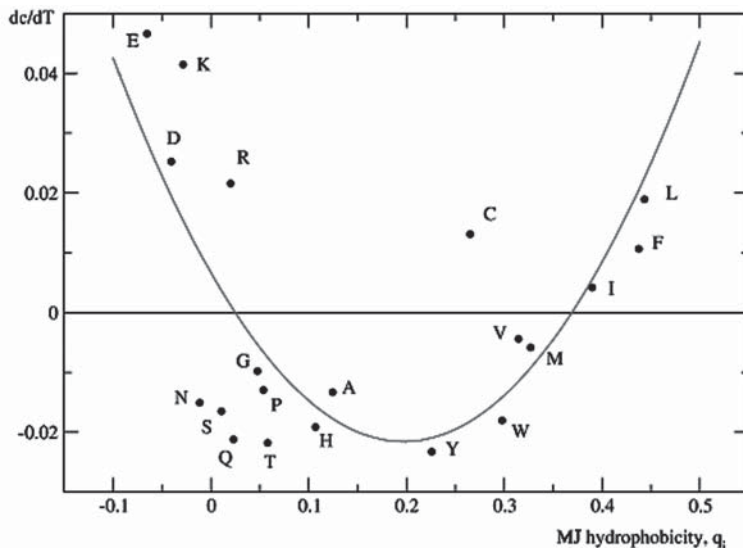


Fig. 10. – Temperature derivatives of the fractions of amino acids plotted against their hydrophobicity q (see sect. 4 for definition of hydrophobicity parameter q in Miyazawa-Jernigan parameter set). The temperature derivatives were obtained as slopes from linear regression between the frequency of every of the 20 amino acids in the designed proteome and T_{env} for which these sequences were designed. The parabolic fit is to guide the eye.

charged residues (red triangles) and slight increase of hydrophobic amino acids (green squares) at the expense of polar (black dots) ones. Remarkably the results shown in fig. 9 suggest that increase of thermostability is accompanied by growth of amino acid content from extremes of hydrophobicity scale — adding both charged and hydrophobic residues. This observation is further highlighted in fig. 2 that shows the changes of amino acid compositions of model proteins with T_{env} in designed proteomes for all 20 amino acids, ranked by their hydrophobicity according to the Miyazawa-Jernigan set of interaction parameters (see sect. 4). Figure 10 clearly shows that addition of amino acids to thermophilic model proteomes occurs from the extremes of hydrophobicity scale while the middle is depressed. The content of charged (Asp, Glu, Lys, Arg; DEKR) and four of the hydrophobic (Ile, Leu, Phe, Cys; ILFC) residues is increased with temperature at the expense of other residues, mostly polar ones. This observation shows that combining amino acids with maximum variance in their hydrophobicity is crucial for creating hyperthermostable model proteins. We refer to this effect as to “from both ends of hydrophobicity scale” trend.

To investigate whether the “from both ends” trend is observed in natural proteins, we analyzed the variation of amino acid composition in fully sequenced bacterial proteomes (83 species in total) of psycho-, meso-, thermo-, and hyper-thermophilic prokaryotes (habitat temperatures from -10 to $+110^\circ\text{C}$). Importantly, amino acid composition of 83 natural prokaryotic proteomes reveal similar trends, an increase of the contents of hydrophobic and charged residues and a decrease of the content of polar ones (fig. 9(b)).

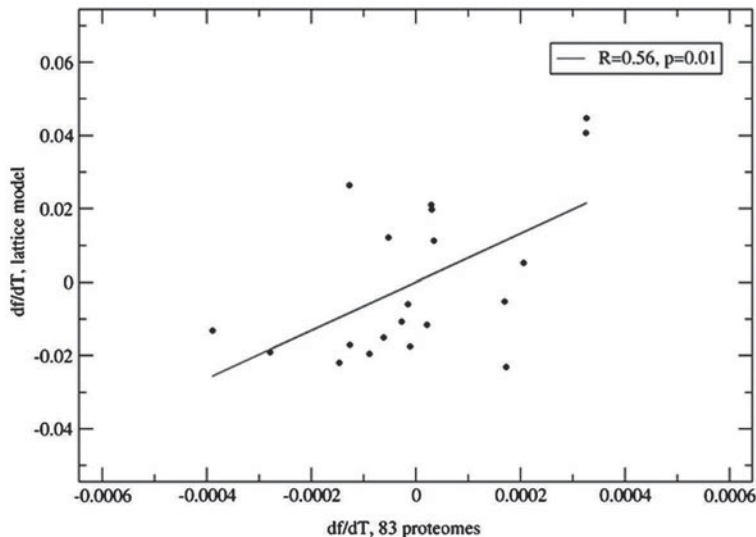


Fig. 11. – The scatter plot between the temperature derivatives of the fraction of each of 20 amino acids in designed lattice proteins (y -axis) against the corresponding temperature derivative calculated over the 83 natural proteomes (x -axis). The correlation coefficient is $R = 0.56$.

For a more direct comparison of the predictions of our model with the properties of natural proteomes, in fig. 11 we plotted the temperature derivative of the fraction of each of the amino acids in designed lattice proteins against the corresponding temperature derivative calculated over the 83 natural proteomes. The observed positive significant correlation ($R = 0.56$, $p = 0.01$) suggests that generic physical factors captured by this simple statistical-mechanical model played a major role in shaping the amino acid composition patterns across a wide range of habitat temperatures.

We hypothesize that the generic character of the “from both ends” trend that we universally observe in the model and natural proteins is related to the positive and negative elements of design. In this case, one (hydrophobic) end of the scale is responsible for positive design while another (hydrophilic) end provides negative design. In order to test this hypothesis, we first studied how the energy gap between the energy of the native state and that of misfolded conformations for the designed model proteins depends on T_{env} (fig. 12). Positive design is the major contributor to the effect (the slope of the temperature-dependent energy decrease of the native state with growth of T_{env} is -5.22), while the increase of the average energy of decoys with T_{env} (slope $+1.64$) is pronounced, but less significant. Nevertheless the results presented in fig. 12 provide a clear evidence that negative design works, along with positive design, in the selection of thermostable model proteins.

The data presented so far provides insight into averaged (over many model proteins) contributions to the energies of native conformation and decoys. However a question arises whether negative design works by increasing “average” non-native interactions of

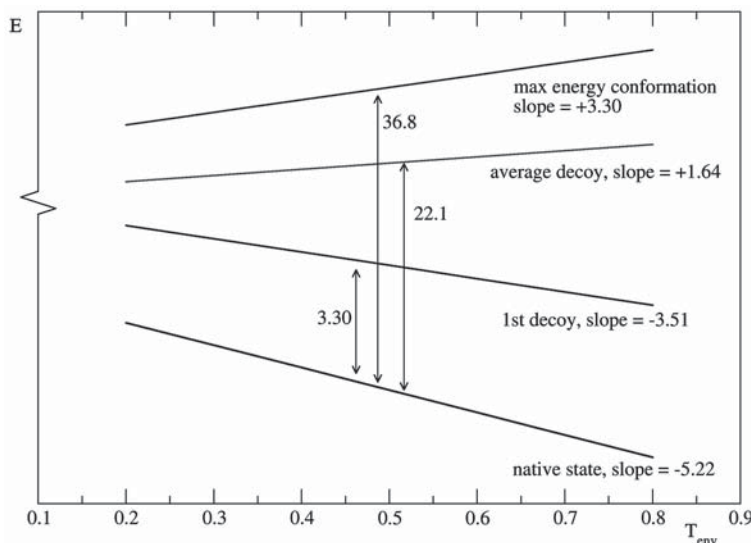


Fig. 12. – Temperature dependence of the contributions to the gap in P-design simulations of 5000 sequences. Positive design results in significant lowering of the native state energy (energy decrease in the interval of temperatures 0.2–0.8 has a slope equal to -5.22). Temperature-dependent increase of the decoys’ energy (averaged over all decoy structures) is less pronounced (slope is 1.64), pointing to the specific nature of the negative design.

by strengthening certain specific repulsive non-native interactions. Indeed, positive design results in many stabilizing contacts inside the hydrophobic core and in the placement of polar/charged residues mainly on the surface. In a similar vein, negative design may be based on introducing a few energetically disadvantageous non-native contacts that are persistent in many decoy structures increasing their energy [41, 42]. Therefore, non-native contacts responsible for negative design may well be specific for each sequence, making this effect more detectable if individual proteins are considered.

The exact nature of the lattice model makes a detailed residue-by-residue analysis of the action of both positive and negative design possible. To this end it is instructive to identify interactions, native and non-native contacts, between residues that play especially important role in stabilization of the native state and destabilization of decoys. The key idea here is that such important interactions should be conserved in all sequences that fold into a given structure. While identities of amino acids that form such a contact may vary from sequence to sequence, the strength (or energy) of key native or non-native contacts will be preserved: it will be either strongly repulsive or strongly attractive for all sequences that fold into a given structure [43]. Therefore in order to identify such key contacts, it is instructive to consider the distribution of energies of native and non-native residue contacts in multiple sequences that fold into the same native structure. Such analysis can reveal not only conserved strong native contacts but also possible conserved strong repulsive non-native contacts. To investigate such possibility, we designed 5000

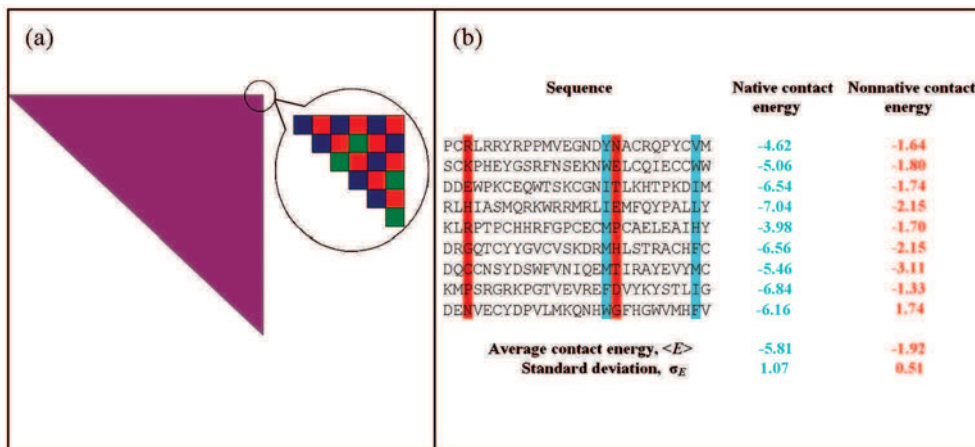


Fig. 13. – (a) Zoom-in into contact matrix of one of the 103346 compact lattice conformations, the cartoon. Even/even, odd/odd, diagonal, $(i, i + 1)$, $(i, i + 2)$ contacts do not exist in a $3 \times 3 \times 3$ lattice (red). There are 28 native (green) and 128 non-native (blue) contact in every compact conformation. (b) Illustration of the calculation of average energies and dispersion of all native and non-native contacts. 10 aligned sequences all folding into the same shown structure are presented. An example of residues making native (cyan) and non-native (red) contacts is shown. As an illustration energy of the selected native and non-native contacts are shown for each sequence (this energy is calculated according to the identities of amino acids forming a contact using Miyazawa-Jernigan knowledge-based potential). Average energy over all 10 aligned sequences of shown native and non-native contacts and its standard deviation are calculated for illustration here.

lattice proteins that all fold into the same (randomly chosen) native structure. To achieve that we used the design algorithm similar to P-design (see sect. 4), but for a fixed native structure, and checked *a posteriori* that the target structure is indeed the native state for all 5000 sequences. We designed a set of 5000 “mesophilic” sequences at $T_{\text{env}} = 0.2$ and 5000 “hyperthermophilic” sequences that fold to the same structure but are much more stable ($T_{\text{env}} = 0.8$).

The concept of native and non-native contacts for our lattice model is illustrated in fig. 13a. It is a cartoon with a zoom-in into the contact matrix of the lattice structure used in simulations. The contact matrix of any compact lattice conformation contains native (green, total 28 in any structure) and possible non-native (blue, total 128) contacts. All other contacts (red) are prohibited according to the properties of the cubic lattice. In order to identify important native and non-native contacts whose energies are conserved we applied the following procedure. First, for each of the 5000 sequences that fold into selected structure we calculated energies of all 28 native and all 128 non-native contacts (using the identities of residues and Miyazawa-Jernigan potentials that were employed to design sequences). Next, for each contact we calculated average energy and its standard deviation over all 5000 designed sequences (see fig. 13b for illustration of this calculation).

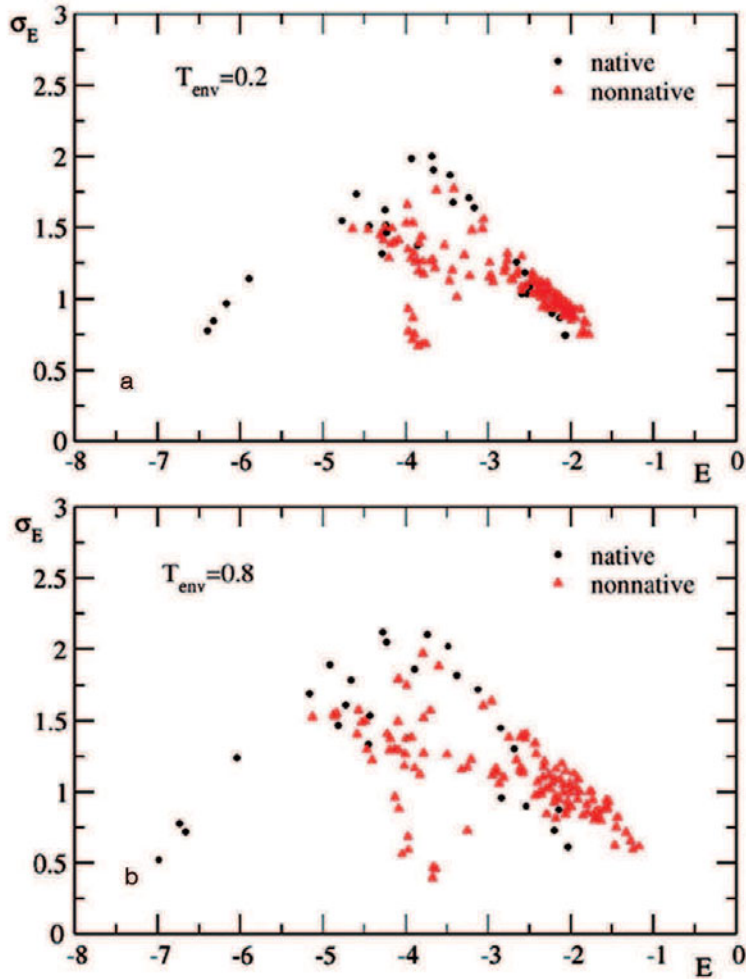


Fig. 14. – Average interaction energies and their standard deviations for all native (black dots) and non-native (red triangles) contacts calculated over 5000 sequences having the same native state. (a) Design of “mesophilic” lattice proteins, $T_{\text{env}} = 0.2$; (b) Design of “hyperthermophilic” lattice proteins, $T_{\text{env}} = 0.8$.

Contacts whose energy shows a very low standard deviation over all designed sequences are apparently the ones that are most important for stability. This procedure was carried out both for mesophilic sequences ($T_{\text{env}} = 0.2$) and for thermophilic sequences ($T_{\text{env}} = 0.8$) and the results are shown in fig. 14, which presents standard deviation of interaction energies of each native and non-native contact over all 5000 designed mesophilic sequences (a) and hyperthermophilic sequences (b), plotted against the average (over 5000 designed sequences) energy of that contact. The plot consists of $28 + 128 = 156$ datapoints,

covering all native and all possible non-native interactions. The native state clearly defines conserved low- and high-energy native contacts (shown in black) in most of the sequences, as the standard deviation is the lowest at the extreme values of the energy. Conserved attractive interactions are in the protein interior, corresponding to the lattice analog of the hydrophobic core; apparently, they emerge due to the action of positive design. The non-native contacts (red dots) follow a different pattern, with only a few conserved attractive interactions, suggesting the diversity of decoy structures. What is surprising to see, however, is that energies of certain *most repulsive* (high-energy) non-native contacts show a very low standard deviation —indicating that such contacts may be as important for protein stability as conserved native ones. Comparison of mesophilic and hyperthermophilic sequences shows clearly that emergence of strong and conserved attractive and repulsive interactions in key native and non-native contacts is directly related to sequence design that generates stable sequences: design of hyperthermostable sequences (fig. 14b) results in stronger and more conserved (lower dispersion of energy) attractive and repulsive specific native and non-native interactions. The only reason that repulsive energies of non-native contacts are conserved is that such contacts persist in certain frequent decoy structures and contribute to the widening of the gap between the native state and decoys. Such repulsive contacts are indirectly (via the sequence) related to a particular native state, are not numerous, and their role may be completely obscured in a “high-throughput” analysis where sequences with different native states are considered together, as in fig. 12. Therefore, we conclude that negative design involves a very specific strategic placement of repulsive contacts in certain decoy structures.

The results and analysis presented in fig. 14 have very important implication for real proteins. The requirement to conserve energy of key contacts in multiple sequences that fold into the same structure implies that amino acids forming such contacts can mutate in correlated way, for example by swaps. The observation that mutations may often occur as swaps to preserve specific attractive native and specific repulsive non-native interactions leads to a prediction of a peculiar dependence between frequency of amino acid substitutions (as in, *e.g.*, in BLOSUM matrices [44]) and interaction energy between amino acids. Indeed, as illustrated in fig. 15, a correlated mutation in the form of a swap can manifest itself *in sequence alignment* as a substitution between amino acids that are making the swap. The implication is that frequent substitutions will be observed between amino acids that strongly attract each other (to preserve specific stabilizing native contacts). More interestingly and perhaps more surprisingly, frequent substitutions are also predicted between amino acids that strongly repel each other (to preserve specific non-native repulsive contacts). In other words, we predict that the scatter plot between elements of amino acid interaction energy matrix and substitution matrix will be non-monotonic.

We tested this prediction by plotting the dependence of elements of substitution matrix BLOSUM62 [44] for 190 pairs of amino acids (synonymous substitutions are excluded) *vs.* their interaction energy as approximated by the knowledge-based Miyazawa-Jernigan potential [45] (fig. 16). This analysis indeed reveals a non-monotonic shape: The parabolic fit in fig. 16a highlights highly significant non-monotonic nature of the

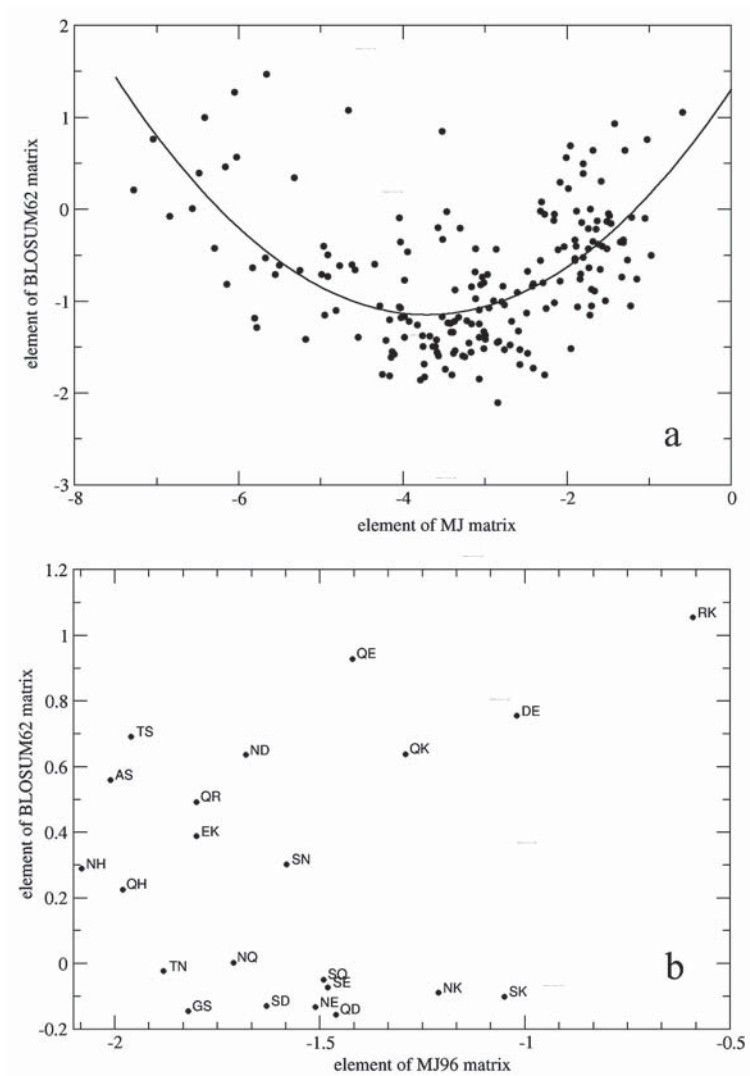


Fig. 16. – Scatter plots showing the dependence of the elements of the substitution matrices of interaction energies between amino acid residues. Only non-synonymous substitutions are presented. (a) BLOSUM62 substitution matrix derived on sequences of natural proteins (ordinate) *vs.* Miyazawa-Jernigan interaction energies between amino acids (abscissa). Red lines represent parabolic fit to highlight non-monotonic nature of the plot. $R^2 = 0.36$, $p < 10^{-4}$ for the fit and the coefficient at X^2 is 0.18 ± 0.02 . An alternative linear fit (not shown) is highly insignificant: $R^2 = 0.01$, $p = 0.89$. (b) Blow-up of the right upper corner of (a) with amino acid pairs labeled.

to Asparagine, Glutamic Acid to Arginine, etc. Several highly non-conservative substitutions show about “random” frequency (element of BLOSUM matrix close to zero, *e.g.*, for Asn to Lys) but this may be due to compensation of two opposite effects: suppression of highly non-conservative substitutions (*e.g.*, that change charge) and facilitation of correlated substitutions like the ones in the form of swaps as illustrated here.

Use of correlated mutations as predictors of spatial proximity of amino acids in the native structure has been proposed by many authors [46-49]. Indeed, statistical analysis shows that overall correlation between distance between amino acids and degree of correlation in multiple sequence alignments does exist [50]. However it has been noted that sometimes correlated mutations are observed between amino acids that are distant in native structure [46,51]. While sometimes such observations are discarded as false positives in the prediction algorithm [46], our analysis predicts that indeed residues that are distant in structure but may form important repulsive contacts in misfolded conformations may exhibit correlated mutations as illustrated in figs. 15 and 16.

3.1. Discussion. – Earlier works [39, 40, 52] have established an empirical correlation between OGT of an organism and the melting temperature of its proteins. Here we found the pronounced amino acid biases and related proteomic determinant of thermal adaptation: IVYWREL combination of amino acid residues, which are highly correlated with the OGT of prokaryotes. In order to better understand possible molecular mechanisms responsible for the observed highly significant proteomic trends, one must establish a connection between thermal adaptation at organismal and molecular levels.

Remarkably, all aminoacids from the IVYWREL predictor set are attached to tRNA by class I aminoacyl-tRNA synthetases [53]. Other amino acids belonging to the group of class I synthetases are Cys (prone to form stabilizing S-S bridges), Lys (important in the generic mechanism of thermostability, see below and [54]), and Met that are often placed in N-termini of proteins. Thus, class I amino acids, contrary to those of class II [55], may constitute a group of amino acids sufficient for the synthesis of thermostable proteins. Besides, all class I synthetases have the Rossmann fold structure, one of ancient LUCA domains [56] with high compactness and, therefore, stability [57]. These observations suggest a possible connection between thermal adaptation and evolution of protein biosynthesis.

As is known from statistical mechanics and the theory of protein folding [58], the stability of proteins is largely determined by the Boltzmann weight $\exp[-\Delta E/k_B T]$, where ΔE is the energy gap between the native state and the nearest in energy misfolded decoy structure, T is environmental temperature, and k_B is the Boltzmann constant. As IVYWREL amino acids are important for thermostability, it is natural to assume that the energy gap ΔE is established mainly by interactions between these types of residues. Thus, we suggest that natural selection adjusted the content of IVYWREL in the proteomes to maintain $\Delta E/k_B T$ at a nearly constant level irrespective of the environmental temperature. Interestingly, the total fraction of IVYWREL residues in the proteomes changes from 0.37 to 0.48 over the accessible temperature range. This relatively small yet significant change highlights a very delicate balance between hydrophobic, van der

Waals, ionic, and hydrogen-bond interactions in correctly folded proteins. Indeed, the IVYWREL set contains residues of all major types, aliphatic and aromatic hydrophobic (Ile, Val, Trp, Leu), polar (Tyr), and charged (Arg, Glu), both basic and acidic. Recently, we have shown, using exact statistical mechanical models of protein stability, that the increase of the content of hydrophobic and charged amino acids can be quantitatively explained as a physical response to the requirement of enhanced thermostability (“from both ends of the hydrophobicity scale” mechanism [59]), reflecting the positive and negative components of protein design.

The basic mechanism of protein thermal stability can be further augmented or modified by other biological, evolutionary, and environmental inputs. Availability of different amino acids in the environment, protein-protein interactions, and protein function requirements are just a few of those other major factors affecting the amino acid composition. Membrane proteins deliver an example where the mechanism of thermal adaptation is adjusted according to the specificity of the structure and interactions with its hydrophobic environment. In α -helical bundles, the temperature dependence of IVYWREL deviates from that of soluble proteins, whereas in β -barrels even the determinant of thermostability, CVYP, is completely different. These observations are suggestive of the differences between the folding processes of globular and membrane proteins. Thermostability predictors derived for individual folds confirm a generic nature of IVYWREL predictor, which dominates in the majority of protein folds. Specific mechanisms of structure stability present in metal-binding proteins (globin, cytochrome C, ferredoxin) or those stabilized by S-S bridges (lysozyme) explain lesser importance of IVYWREL predictor for these folds.

Thermostability of proteins is, however, only a partial prerequisite of thermal adaptation of an organism, as DNA molecules must also remain stable at elevated environmental temperatures. Numerous experimental and theoretical works established two fundamental interactions in double-stranded DNA, base pairing and base stacking as major determinants of DNA stability *in vitro*. GC pairs in the double helix have stronger base-pairing interactions than AT pairs [21], while purines, A and G, yield a low energy of stacking [37, 38, 60].

The role of G+C-content in establishing certain biases in the amino acid composition has been widely discussed [10, 18, 61-63]. Our high-throughput analysis of nucleotide and amino acid compositions has confirmed that the G+C content of DNA is not correlated with the optimal growth temperature of prokaryotes. The only signature of thermal adaptation in DNA composition that we found, increase of A+G content in the coding DNA, is to a considerable extent a consequence of the thermal adaptation of protein sequences. Elimination of codon bias and reverse translation of the protein sequences into nucleotide sequences does not change the major trends in DNA composition (fig. 7). This result suggests that composition-dependent (pairing) interactions in double-stranded DNA are not the bottleneck of thermal adaptation in prokaryotes, as even without codon bias the variation in composition of DNA follows the variation of composition of proteins across the whole range of environmental temperatures.

On the contrary, pairwise nearest-neighbor correlations observed in DNA sequences are largely independent from amino acid trends as they are entirely determined by the codon bias developed as a form of thermal adaptation in prokaryotes. We found that natural selection tailored the codon bias to increase the fraction of ApG pairs in both strands of DNA molecules of thermophilic organisms. Indeed, an increased number of ApG pairs leads to a lower stacking energy, and, thus, stabilization of DNA. We also demonstrated that the trend of higher frequency of ApG pairs in thermophiles persists for the whole DNA, including its coding and non-coding parts. To our knowledge, this consideration is one of the first physical models that quantitatively explains the necessity and one of possible reasons for codon bias in prokaryotes.

An important finding presented in this work is that amino acid composition adaptation is a primary factor while certain signatures of thermal adaptation at the level of nucleotide composition such as purine loading index may be partly derivative of amino acid adaptation requirement. This is perhaps not surprising given that many (but not all) proteins are present in cytoplasm in monomeric form and must be stabilized on their own —by interactions between their own amino acids (and also destabilize misfolded conformations, performing negative design as well). On the other hand, many proteins bind to DNA in prokaryotic cells and some of them may provide additional stabilization to DNA in hyperthermophiles making thermal adaptation at the level of DNA sequences a “collective enterprise” of the cell, relieving direct pressure on DNA sequence to adapt to high temperature. Positive superhelicity provided by gyrases may be one such mechanism and perhaps several other mechanisms of DNA stabilization by proteins could be discovered in future. In particular the observed purine-purine AG correlation as an “independent” (from amino acid composition) adaptation mechanism may be a signature of the important role of stacking interaction in DNA thermal stabilization. Alternatively, the AG correlations may be a signature of an additional thermal adaptation of DNA via modulation of local mechanical properties or facilitation of interaction with proteins. Elucidation of the exact mechanism(s) by which DNA in prokaryotic genomes adapt to high temperatures is a matter of future research.

To summarize, we have shown that thermal adaptation of proteins in prokaryotes is strongly manifest at the level of amino acid composition. Among all possible combinations of amino acids, the fraction of amino acids IVYWREL in a proteome is the most precise predictor of optimal growth temperature. We also observe a difference in the signatures of thermal adaptation of soluble and membrane proteins. We did not find a strong evidence for thermal adaptation of DNA at the level of nucleotide composition. The G+C content of the genomes is not correlated with environmental temperature, while A+G content increases with environmental temperature, to a considerable extent, as a consequence of thermal adaptation of proteins. At the same time, our analysis of sequence correlations in DNA shows that ApG nearest-neighbor pairs are overrepresented in both strands of dsDNA of thermophilic organisms. The abundance of ApG pairs is a direct consequence of the codon bias developed in thermophilic prokaryotes. These findings provide a definitive answer to the long-standing problem of which genomic and proteomic compositional trends reflect thermal adaptation.

The emergence of “from both ends” amino acid composition trend in thermal adaptation of proteins is a direct consequence of statistical mechanical principles that govern protein stability. According to the statistical-mechanical theory of protein folding, stability and uniqueness of the native structure is directly related to ΔE , the energy gap between native state and first (closest in energy) misfolded structure that is not structurally similar to the native conformation (see fig. 1 of ref. [26]). Therefore, increase of the environmental temperature demands higher melting temperature of proteins in an organism which in turn requires that proteins of thermophilic organisms have a larger energy gap ΔE in order to remain stable at elevated temperature. We found in this study that increase of the energy gap for thermophilic model proteins is achieved by decreasing the energy of the native state and simultaneously increasing the average energy of misfolded conformations, suggesting that both positive and negative design work to increase thermostability of proteins.

Remarkably, six of the seven residues of the IVYWREL set represent the two extremes of hydrophobicity scale, hydrophobic (IVWL) and charged (RE) types. Discrepancies between different hydrophobicity scales [64], the statistical nature of knowledge-based Miyazawa-Jernigan potential [45], and limitations of lattice model make it impossible to quantitatively compare the content of individual amino acids in lattice and natural proteomes or exactly predict the amino acid composition of thermophilic proteomes with very high accuracy (though our simple lattice calculations are in semi-quantitative agreement with data on natural proteomes, see, *e.g.*, fig. 9b). However, the increase of the content of hydrophobic and charged residues, is very robust in model calculations and shows the generic character of the “from both ends” trend.

In general, any protein design procedure is aimed at finding the unique lowest-energy sequence/structure combination that is separated from the competing misfolded structures by a large energy gap [41]. While positive design [65] is universally used in experiment, role and omnipresence of negative design are still under discussion [66]. The main challenge in the study of negative design stems from the difficulties in the modeling of relevant decoys and energetic effects of mutations that destabilize them [66]. Earlier, it has been shown that charged residues can be effectively used in negative design [42]. Another indirect evidence of the contribution of charged residues to negative design emerges from site-directed mutagenesis, where mutation of polar groups to charged ones on the surface of a protein leads to stabilization even in the absence of salt-bridge partners of the mutated group [67-69]. In a series of experiments [67, 68, 70], it has been shown that surface electrostatic interactions provide a marginal contribution to stability of the native structure, hence their possible importance for making unfavorable high-energy contact in decoys. An alternative view, proposed recently by Makhatadze and coauthors, suggests that long-range electrostatic interactions may contribute to stability of the native state [71]. Our simulations and proteomic analysis point to the positive design as the major contributor to widening of the energy gap between the native state and decoy structures, and corroborate the specific nature of negative design (fig. 12).

Specificity of negative design [41, 42, 65, 66] prompted us to create a large set of identically designed lattice proteins with the same native conformation. The advantage

of lattice model is that all native and non-native contacts are known, giving an exact description of positive and negative elements of design for each structure. We found that positive design is based on the strong attractive interactions between hydrophobic amino acid residues, spatially segregated from the charged ones. Negative design is mainly provided by the charged amino acids that form repulsive non-native contacts and thus increase energy of the decoys. The exhaustive analysis of the variance of energies of native and non-native interactions provides a detailed description of energetic contributions due to the negative design (fig. 14). In the case of thermophilic adaptation, both negative and positive elements of design become stronger because of the demand for an increased energy gap (fig. 14b). Therefore, thermophilic adaptation leads to the increase of the content of hydrophobic and hydrophilic amino acids, directly reflected in the “both ends mechanism”

Finally, we illuminate the generic nature of positive and negative elements of design and unravel its manifestation in evolution of protein sequences. A comparison of substitution rates in sequences of natural proteins (BLOSUM62 substitution matrix) and interaction energies according to knowledge-based Miyazawa-Jernigan potential reveals a specific conservation in the groups of hydrophobic and charged amino acids (fig. 16a). As a result, an interesting phenomenon emerges: residues making strong interactions (either attractive or repulsive) interchange in order to preserve energy of the contact.

Our study deepens the understanding of correlated mutations in proteins. With regards to native contacts, the fact that amino acids making strongly attractive native interactions should exhibit correlated mutations had been realized long ago. Several authors proposed to search for correlated mutations as a tool to determine possible native contacts from multiple sequence alignment [46-49]. Remarkably, our analysis of conservation of contact energies in sequence alignment of model proteins (see fig. 13b) suggests that residues that are not in contact in the native structure and repel each other may also exhibit correlated mutations. In this case, they preserve strong repulsive non-native interactions that keep energy of misfolded conformations sufficiently high compared to native states. As illustrated in fig. 15, the most straightforward type of a correlated mutation that conserves energy of a contact is a swap of amino acids. Such swaps can occur between mutually attractive amino acid residues or between strongly repulsive amino acid residues, to preserve strong attractive native contact (illustrated in blue in fig. 15) and a specific non-native contact in a misfolded conformation (in red in fig. 15), respectively.

Correlated mutations between residues that are distant in structure had been found earlier: for example in original paper of Gobel *et al.* [46] highly correlated mutations in bovine pancreatic trypsin inhibitor were found between residues that are as far as 20 Å apart. Vivid examples of correlated mutations between distant residues are provided by Larson and coauthors [72] for SH3 domain. These authors found several correlated mutations between distant surface residues that involve amino acids types (such as 11D, 52S in 1shf structure) that repel each other. These correlated mutations are between residues whose C_β atoms are 20 Å apart in native structure and which do not play a known functional role. Using double mutant technique, Horovitz and coauthors [47] suggested a relation between correlated mutations and energetic connectivity (*i.e.* nonadditivity of

stability effects in double mutation cycles) between corresponding amino acids. Further, analysis of energetic nonadditivity in double mutants in staphylococcal nuclease by Green and Shortle [73] showed that amino acids that are distant in structure may be “energetically coupled”, suggesting some form of indirect interaction between them. Green and Shortle attributed this effect to influence of mutations on unfolded state of proteins. More recently Lockless and Ranganathan [51], using the same technique of double mutants and a somewhat different algorithm to detect correlated mutations, found energetically coupled correlated mutations in PDZ domain. Based on the evolutionary consideration of multiple sequence alignments of PDZ domain, Lockless and Ranganathan suggested the existence of a “pathway of energetic connectivity” as a general property of many proteins. Fodor and Aldrich [50], however, examined different proteins and conclusively argued against the Lockless and Ranganathan’s “general principle of isolated pathways of evolutionary conserved energetic connectivity in proteins”.

How can one then explain the observation that distant residues whose mutations are highly correlated appear to be sometimes (but not always) energetically coupled? Experimental data [72, 73] reveal multiple examples when correlated mutations do occur to preserve strong repulsive interactions that are not present in the native structure but may appear in misfolded conformations. Our analysis suggests that energetic coupling between distant residues may be due to their specific interactions in certain misfolded conformations (fig. 14).

In this work we developed a simple exact model of thermophilic adaptation and discovered a fundamental, “from the both ends of hydrophobicity scale”, amino acid composition trend in Nature’s quest to enhance protein stability. We presented theoretical and proteomic evidence that this trend is due to positive and negative components of design that provide stability of protein structures. Both factors work in protein evolution and are manifested in peculiar patterns of multiple sequence alignments. While negative design is certainly not the only reason for recruitment of charged residues to the sequences of hyperthermophilic proteins, it appears to contribute significantly to a delicate balance of factors that determine protein stability. Further, it explains an energetic coupling in highly correlated mutations of distant residues and reveals the “swap” mechanism that preserve an interaction energy in correlated mutations. To conclude, the major lesson of this work is the discovery of physical and evolutionary mechanisms that underlie positive and negative design of natural proteins. Better understanding of fundamental natural mechanisms of protein stabilization can help in future engineering and design efforts.

4. – Methods

4.1. *Methods.* – The complete genomes were downloaded from NCBI Genome database and the optimal growth temperatures were collected from original publications, and from the web sites of the American Type Culture Collection, <http://www.atcc.org>, the German Collection of Microorganisms and Cell Cultures, <http://www.dsmz.de>, and PGTD database [74], <http://pgtdb.csie.ncu.edu.tw/>. The distribution of OGTs of prokaryotes with completely sequenced genomes shows that 75 species out of 204 have

an OGT of 37 °C (mostly human and animal pathogens), 27 species have an OGT of 30 °C, and 19 species grow best at 26 °C. Therefore, 121 genomes, or 59% of the data, correspond to just three values of OGT. To overcome this obvious bias in the data, we averaged the amino acid compositions of the organisms with OGTs of 26, 30, and 37 °C, respectively. After this procedure, our data set consists of 86 proteomes, 83 real proteomes and 3 averaged proteomes at 26, 30, and 37 °C. To validate the averaging of the overrepresented amino acid compositions we repeated the complete analysis for the 204 genomes, and did not find any major differences.

We represent the different sets of amino acids by vectors $\{a_i\}$ of 20 binary digits, where each digit a_i takes the value of 1 if the amino acid of type i is present in the set, and 0 otherwise. For the set to be nontrivial, at least one of the components of a_i must be 1 and at least another one must be zero. There are $2^{20} - 2 = 1,048,574$ such vectors, enumerating all possible subsets of 20 amino acids. Let $f_i^{(j)}$ be the fraction of amino acid of type i in proteome j . Then, for each of the 86 proteomes and each vector $\{a_i\}$ we calculate the total fraction $F^{(j)}$ of the amino acids from a particular subset,

$$F^{(j)} = \sum_{i=1}^{20} a_i f_i^{(j)}$$

and perform the linear regression between the values of $F^{(j)}$ and organism living temperatures $T_{\text{opt}}^{(j)}$. The higher the correlation coefficient R of this regression, the more important is a given subset of amino acids as a predictor of OGT. Since for each proteome $\sum_{i=1}^{20} f_i = 1$, there are in fact only $2^{19} - 1$ linearly independent combinations of fractions of amino acids; if the total fraction of N amino acids increases with OGT and the correlation coefficient is R , the total fraction of the complementary set of $20 - N$ amino acids is decreasing with OGT with the correlation coefficient $-R$. Therefore, the predictive powers of the complementary subsets of amino acids are equal. To resolve this ambiguity, we focus on the amino acid combinations which are positively correlated with OGT.

For 86 proteomes, the best predictor consists of amino acids IVYWREL, correlation coefficient $R = 0.930$, accuracy of prediction 8.9 °C (see sect. 2).

We also performed an exhaustive enumeration of $3^{20} - 1$ combinations of amino acids in a ternary model, where the coefficients a_i can take the values of -1 , 0 , and 1 . We expected that this method could be advantageous, as it allows to enrich the predictor combination by the amino acids whose fraction is decreasing with OGT (such amino acids would have a weight of -1). The best predictor combination in the ternary model consists of amino acids VYWP with weight $+1$ and CFAGTSNQDH with the weight -1 , correlation coefficient $R = 0.948$, accuracy of prediction of 8.1 °C for 86 proteomes. While this result is generally consistent with the IVYWREL predictor, the ternary model does not offer a significant increase of accuracy of the prediction of OGT as compared to a much simpler binary model. It is also possible to perform a full linear regression between the amino acid fractions and OGT, effectively allowing for non-integer values of $\{a_i\}$. This procedure involves a great risk of overfitting, and results only in a marginal improvement of accuracy of OGT prediction as compared to both binary and ternary models.

TMHMM prediction server [22] at <http://www.cbs.dtu.dk/services/TMHMM/> has been used to identify the sequences of alpha-helical membrane proteins in the 83 genomes. The number of proteins predicted in a genome did not correlate with OGT, so the prediction algorithm is sufficiently robust with respect to the changes of average amino acid composition. The PROFTmb server [23] at <http://cubic.bioc.columbia.edu/services/proftmb/> has been used to identify the transmembrane beta-barrels; proteins with 8 to 22 transmembrane beta-strands have been selected for further study.

We applied the same technique to enumerate all possible combinations of nucleotides in the coding DNA of prokaryotes and find the correlation coefficients between the total fractions of sets of nucleotides and OGT. To quantify the pairwise nearest-neighbor correlations in DNA, for each of the 16 possible pairs of nucleotide types i and j and for each of the 83 genomes we calculated the correlation function $c_{ij} = \frac{Nn_{ij}}{n_i n_j} - 1$, where N is the total number of nucleotides in the DNA sequence, n_i , n_j are the numbers of nucleotides of types i and j , and n_{ij} is the number of pairs where nucleotide j immediately follows nucleotide i in the coding strand. In random DNA without sequence correlations, $c_{ij} = 0$ for any i, j and for any nucleotide composition of the sequence.

Reshuffling of DNA and protein sequences has been performed by swapping two randomly chosen letters in the sequence, and repeating this procedure $2N$ times, where N is the length of the sequence.

Sequences of 10 protein folds (according to SCOP description [75]) with less than 50% of sequence identity were extracted from PDB database. Given a fold, representatives of its sequences were found in 83 complete proteomes by using blastp program with BLOSUM62 substitution matrix, e-value equals to 0.005. For every analyzed protein fold, representatives of its sequences were detected in organisms with OGTs covering whole interval +10 to 110 °C. Extracted proteomic sequences were used in calculations of thermostability predictor if the length of the alignment was more than 70% of the size of fold's sequence.

We use the standard lattice model of proteins as compact 27-unit polymers on a $3 \times 3 \times 3$ lattice [29]. The residues interact with each other via the Miyazawa-Jernigan pairwise contact potential [45]. It is possible to calculate the energy of a sequence in each of the 103346 compact conformations allowed by the $3 \times 3 \times 3$ lattice, and the Boltzmann probability of being in the lowest energy (native) conformation,

$$P_{\text{nat}}(T_{\text{env}}) = \frac{e^{-E_0/T_{\text{env}}}}{\sum_{i=0}^{103345} e^{-E_i/T_{\text{env}}}},$$

where E_0 is the lowest energy among the 103346 conformations, and T_{env} is the environmental temperature. The melting temperature T_{melt} is found numerically from the condition $P_{\text{nat}}(T_{\text{melt}}) = 0.5$. Note that if the energy spectrum E_i is sparse enough at low energies, the value of P_{nat} is determined chiefly by the energy gap $E_1 - E_0$ between the native state and the closest decoy structure that has no structural relation to the native state.

To design lattice proteins, we use here a Monte Carlo procedure (P-design [27, 28]) that maximizes the Boltzmann probability P_{nat} of the native state by introducing mutations in the amino acid sequence and accepting or rejecting them according to the Metropolis criterion. As this procedure takes the environmental temperature T_{env} as an input physical parameter, and generates amino acid sequences designed to be stable at T_{env} , it is an obvious choice for modeling the thermophilic adaptation.

Initially, the sequence is chosen at random; the frequencies of all amino acid residues in the initial sequences are equal to 5 percent. At each Monte Carlo step, a random mutation of one amino acid in a sequence is attempted, and P_{nat} of the mutated protein is determined. The native structure is determined at every step of the simulation; generally the native state changes upon mutation of the sequence. If the value of P_{nat} increased, the mutation is always accepted; if P_{nat} decreased, the mutation is accepted with the probability $\exp[-(P_{\text{nat}}(\text{old}) - P_{\text{nat}}(\text{new}))/p]$, with $p = 0.05$ (a Metropolis-like criterion). We chose $p = 0.05$ so that the average melting temperature of designed proteins is higher than the environmental temperature (see fig. S2 in Supporting Information), in agreement with experimental observations [40, 52]. The design procedure is stopped after 2000 Monte Carlo iterations. Such length of design runs is sufficient to overcome any possible effects of the initial composition of the sequences, so the amino acid composition of the designed sequences depends only on the environmental temperature T_{env} .

To relate the trends in amino acid composition with the physical properties and interaction energies of individual amino acids, we use hydrophobicity as a generic parameter characterizing an amino acid [64]. To characterize the hydrophobicity of amino acids in the simulations, we make use of the fact that the Miyazawa-Jernigan interaction energy matrix is very well approximated by its spectral decomposition [76]. Interestingly, it is sufficient to use only one eigenvector $\mathbf{q}$, corresponding to the largest eigenvalue, so the interaction (contact) energy E_{ij} between amino acids of types i and j reads $E_{ij} \approx E_0 + \lambda q_i q_j$ [76]. In this representation, hydrophobic residues have the largest values of q , while hydrophilic (charged) residues correspond to small q .

Sequence of the triosephosphate isomerase (7tim.pdb, chain a) was aligned against sequences of 39 other representatives of TIM barrel folds [75] with sizes less than 300 amino acid residues. Kalign web-server for multiple alignment of protein sequences was used (<http://msa.cgb.ki.se/cgi-bin/msa.cgi> [77]).

Coordinated substitutions in the multiple alignments were determined by using CRASP program (<http://wwwmgs.bionet.nsc.ru/mgs/programs/crasp> [78]). Only significant correlations, with correlation coefficient higher than critical threshold (0.311) were considered.

The complete genomes were downloaded from the NCBI Genome database at <http://www.ncbi.nih.gov/entrez/query.fcgi?db=Genome>.

* * *

We are grateful to M. D. FRANK-KAMENETSKII for illuminating discussions and M. KLOSTER for useful comments on initial version of this paper. We would like to thank an anonymous referee for the suggestion to analyze membrane proteins and to

H. BIGELOW (Columbia University) for the help with prediction of TM beta-barrels. This work was supported by the NIH. INB was supported by the Merck fellowship for genome-related research.

REFERENCES

- [1] SHAKHNOVICH E. I., *Phys. Rev. Lett.*, **72** (1994) 3907.
- [2] JAENICKE R. and BOHM G., *Curr. Opin. Struct. Biol.*, **8** (1998) 738.
- [3] JAENICKE R., *J. Biotechnol.*, **79** (2000) 193.
- [4] BEREZOVSKY I. N. and SHAKHNOVICH E. I., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 12742.
- [5] THOMPSON M. J. and EISENBERG D., *J. Mol. Biol.*, **290** (1999) 595.
- [6] QUEROL E., PEREZ-PONS J. A. and MOZO-VILLARIAS A., *Protein Eng.*, **9** (1996) 265.
- [7] VETRIANI C., MAEDER D. L., TOLLIDAY N., YIP K. S., STILLMAN T. J., BRITTON K. L., RICE D. W., KLUMP H. H. and ROBB F. T., *Proc. Natl. Acad. Sci. U.S.A.*, **95** (1998) 12300.
- [8] ANDERSON D. E., HURLEY J. H., NICHOLSON H., BAASE W. A. and MATTHEWS B. W., *Protein Sci.*, **2** (1993) 1285.
- [9] SZILAGYI A. and ZAVODSZKY P., *Structure*, **8** (2000) 493.
- [10] KREIL D. P. and OUZOUNIS C. A., *Nucleic Acids Res.*, **29** (2001) 1608.
- [11] PE'ER I., FELDER C. E., MAN O., SILMAN I., SUSSMAN J. L. and BECKMANN J. S., *Proteins*, **54** (2004) 20.
- [12] CAMBILLAU C. and CLAVERIE J. M., *J. Biol. Chem.*, **275** (2000) 32383.
- [13] DAS R. and GERSTEIN M., *Funct. Integr. Genomics*, **1** (2000) 76.
- [14] CHAKRAVARTY S. and VARADARAJAN R., *Biochemistry*, **41** (2002) 8152.
- [15] TEKAIA F., YERAMIAN E. and DUJON B., *Gene*, **297** (2002) 51.
- [16] SINGER G. A. and HICKEY D. A., *Gene*, **317** (2003) 39.
- [17] PONNUSWAMY P., MUTHUSAMY R. and MANAVALAN P., *Int. J. Biol. Macromol.*, **4** (1982) 186.
- [18] MUSTO H., NAYA H., ZAVALA A., ROMERO H., ALVAREZ-VALIN F. and BERNARDI G., *FEBS Lett.*, **573** (2004) 73.
- [19] LAMBROS R. J., MORTIMER J. R. and FORSDYKE D. R., *Extremophiles*, **7** (2003) 443.
- [20] LIN F. H. and FORSDYKE D. R., *Extremophiles*, 2006.
- [21] MARMUR J. and DOTY P., *J. Mol. Biol.*, **5** (1962) 109.
- [22] SONNHAMMER E. L. L., VON HELJNE G. and KROGH A., in *Proceedings of the Sixth International Conference on Intelligent Systems for Molecular Biology*, edited by J. GLASGOW T. L., F. MAJOR, R. LATHROP, D. SANKOFF and C. SENSEN (AAAI Press, Menlo Park, CA) 1998, pp. 175.
- [23] BIGELOW H. R., PETREY D. S., LIU J., PRZYBYLSKI D. and ROST B., *Nucleic Acids Res.*, **32** (2004) 2566.
- [24] GOLDSTEIN R. A., LUTHEY-SCHULTEN Z. A. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **89** (1992) 4918.
- [25] SHAKHNOVICH E. I., *Chem. Rev.*, **106** (2006) 1559.
- [26] SHAKHNOVICH E. I. and GUTIN A. M., *Proc. Natl. Acad. Sci. U.S.A.*, **90** (1993) 7195.
- [27] MORRISSEY M. P. and SHAKHNOVICH E. I., *Fold. Des.*, **1** (1996) 391.
- [28] HOANG T. X., TROVATO A., SENO F., BANAVAR J. R. and MARITAN A., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 7960.
- [29] SHAKHNOVICH E. and GUTIN A., *J. Chem. Phys.*, **93** (1990) 5967.

- [30] PANDE V. S., GROSBERG A. Y. and TANAKA T., *Proc. Natl. Acad. Sci. U.S.A.*, **91** (1994) 12976.
- [31] SHAKHNOVICH E. I. and GUTIN A. M., *Nature*, **346** (1990) 773.
- [32] KLIMOV D. K. and THIRUMALAI D., *Phys. Rev. Lett.*, **76** (1996) 4070.
- [33] SHAKHNOVICH E. I., *Phys. Rev. Lett.*, **72** (1994) 3907.
- [34] GOVINDARAJAN S. and GOLDSTEIN R. A., *Proc. Natl. Acad. Sci. U.S.A.*, **93** (1996) 3341.
- [35] TAVERNA D. M. and GOLDSTEIN R. A., *Biopolymers*, **53** (2000) 1.
- [36] BOWIE J. U., *Nature*, **438** (2005) 581.
- [37] FRIEDMAN R. A. and HONIG B., *Biophys. J.*, **69** (1995) 1528.
- [38] SAENGER W., *Principles of Nucleic Acid Structure* (Springer-Verlag, New York) 1984.
- [39] GROMIHA M. M., OOBATAKE M. and SARAI A., *Biophys. Chem.*, **82** (1999) 51.
- [40] MCFALL-NGAI M. J. and HORWITZ J., *Exp. Eye Res.*, **50** (1990) 703.
- [41] BUTTERFOSS G. L. and KUHLMAN B., *Annu. Rev. Biophys. Biomol. Struct.*, **35** (2006) 49.
- [42] SUMMA C. M., ROSENBLATT M. M., HONG J. K., LEAR J. D. and DEGRADO W. F., *J. Mol. Biol.*, **321** (2002) 923.
- [43] MIRNY L. A. and SHAKHNOVICH E. I., *J. Mol. Biol.*, **291** (1999) 177.
- [44] HENIKOFF S. and HENIKOFF J. G., *Proteins*, **17** (1993) 49.
- [45] MIYAZAWA S. and JERNIGAN R. L., *J. Mol. Biol.*, **256** (1996) 623.
- [46] GOBEL U., SANDER C., SCHNEIDER R. and VALENCIA A., *Proteins*, **18** (1994) 309.
- [47] HOROVITZ A., BOCHKAREVA E. S., YIFRACH O. and GIRSHOVICH A. S., *J. Mol. Biol.*, **238** (1994) 133.
- [48] NOIVIRT O., EISENSTEIN M. and HOROVITZ A., *Protein Eng. Des. Sel.*, **18** (2005) 247.
- [49] OLMEA O. and VALENCIA A., *Fold. Des.*, **2** (1997) S25.
- [50] FODOR A. A. and ALDRICH R. W., *J. Biol. Chem.*, **279** (2004) 19046.
- [51] LOCKLESS S. W. and RANGANATHAN R., *Science*, **286** (1999) 295.
- [52] HOCHACHKA P. and SOMERO G., *Biochemical Adaptation. Mechanism and Process in Physiological Evolution* (Oxford University Press) 2002.
- [53] ERIANI G., DELARUE M., POCH O., GANGLOFF J. and MORAS D., *Nature*, **347** (1990) 203.
- [54] BEREZOVSKY I. N., CHEN W. W., CHOI P. J. and SHAKHNOVICH E. I., *PLoS Comput. Biol.*, **1** (2005) e47.
- [55] KLIPCAN L., SAFRO I., TEMKIN B. and SAFRO M., *FEBS Lett.*, **580** (2006) 1672.
- [56] SHAKHNOVICH B. E., DEEDS E., DELISI C. and SHAKHNOVICH E., *Genome Res.*, **15** (2005) 385.
- [57] ENGLAND J. L., SHAKHNOVICH B. E. and SHAKHNOVICH E. I., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 8727.
- [58] SHAKHNOVICH E. I., *Curr. Opin. Struct. Biol.*, **7** (1997) 29.
- [59] BEREZOVSKY I. N., ZELDOVICH K. B. and SHAKHNOVICH E. I., *PLoS Comput. Biol.*, **3** (2007) 0498.
- [60] YAKOVCHUK P., PROTOZANOVA E. and FRANK-KAMENETSKII M. D., *Nucleic Acids Res.*, **34** (2006) 564.
- [61] HURST L. D. and MERCHANT A. R., *Proc. Biol. Sci.*, **268** (2001) 493.
- [62] SUHRE K. and CLAVERIE J. M., *J. Biol. Chem.*, **278** (2003) 17198.
- [63] WANG H. C., SUSKO E. and ROGER A. J., *Biochem. Biophys. Res. Commun.*, **342** (2006) 681.
- [64] BASTOLLA U., PORTO M., ROMAN H. E. and VENDRUSCOLO M., *Gene*, **347** (2005) 219.
- [65] SHIFMAN J. M. and MAYO S. L., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 13274.
- [66] BOLON D. N., GRANT R. A., BAKER T. A. and SAUER R. T., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 12724.

- [67] PJURA P., MATSUMURA M., BAASE W. A. and MATTHEWS B. W., *Protein Sci.*, **2** (1993) 2217.
- [68] SALI D., BYCROFT M. and FERSHT A. R., *J. Mol. Biol.*, **220** (1991) 779.
- [69] ZHANG X. J., BAASE W. A., SHOICHET B. K., WILSON K. P. and MATTHEWS B. W., *Protein Eng.*, **8** (1995) 1017.
- [70] PEREZ-JIMENEZ R., GODOY-RUIZ R., IBARRA-MOLERO B. and SANCHEZ-RUIZ J. M., *Biophys. Chem.*, **115** (2005) 105.
- [71] STRICKLER S. S., GRIBENKO A. V., GRIBENKO A. V., KEIFFER T. R., TOMLINSON J., REIHLE T., LOLADZE V. V. and MAKHATADZE G. I., *Biochemistry*, **45** (2006) 2761.
- [72] LARSON S. M., RUCZINSKI I., DAVIDSON A. R., BAKER D. and PLAXCO K. W., *J. Mol. Biol.*, **316** (2002) 225.
- [73] GREEN S. M. and SHORTLE D., *Biochemistry*, **32** (1993) 10131.
- [74] WEI L., HUANG E. S. and ALTMAN R. B., *Structure Fold. Des.*, **7** (1999) 643.
- [75] MURZIN A. G., BRENNER S. E., HUBBARD T. and CHOTHIA C., *J. Mol. Biol.*, **247** (1995) 536.
- [76] LI H., TANG C. and WINGREEN N. S., *Phys. Rev. Lett.*, **79** (1997) 765.
- [77] LASSMANN T. and SONNHAMMER E. L., *BMC Bioinformatics*, **6** (2005) 298.
- [78] AFONNIKOV D. A. and KOLCHANOV N. A., *Nucleic Acids Res.*, **32** (2004) W64.

This page intentionally left blank

Molecular recognition through local elementary structures: Transferability of simple models to real proteins

R. A. BROGLIA

Dipartimento di Fisica, Università di Milano - Milano, Italy
INFN, Sezione di Milano - via C. Celoria 16, 20133 Milano, Italy
The Niels Bohr Institut, University of Copenhagen
Blegdamsvej 17, DK 2100 Copenhagen Ø, Denmark

G. TIANA, D. PROVASI and L. SUTTO

Dipartimento di Fisica, Università di Milano - Milano, Italy
INFN, Sezione di Milano - via C. Celoria 16, 20133 Milano, Italy

1. – Introduction

The functional properties of proteins depend upon their three-dimensional structures. This native structure arises because particular sequences of amino acids (see fig. 1) in polypeptide chains fold to generate, from linear chains, compact domains with specific three-dimensional structures (see fig. 2).

Although all the information required for the folding of a protein chain is contained in its amino acid sequence, one has not yet learned to read this information so as to be able to predict the detailed three-dimensional structure of a real protein whose sequence is known. This is the protein folding problem, one of the great unsolved problems of science (cf. ref. [1], p. 508 see also [2]). Great in two senses: 1) important as well as 2) formidable. This second meaning can be directly appreciated looking at fig. 3, where an all atom representation of the native conformation of the HIV-1-PR, an enzyme which plays a central role in the life cycle of the HIV virus is displayed. To make any progress

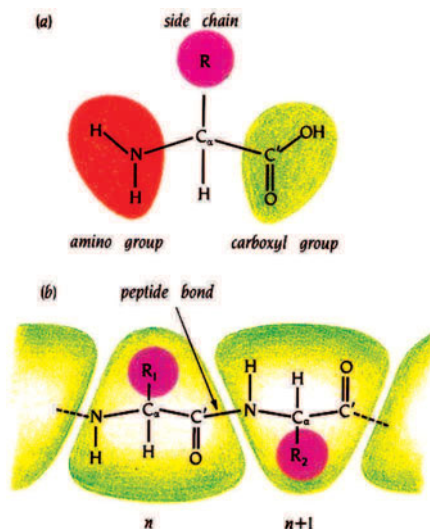


Fig. 1. – Proteins are built up by amino acids that are linked by peptide bonds into a polypeptide chain. (a) Schematic diagram of an amino acid. A central carbon atom C_{α} is attached to an amino group, NH_2 , a carboxyl group $C'OOH$, a hydrogen bond, H, and a side chain, R. (b) In a polypeptide chain the carboxyl group of amino acid n has formed a peptide bond, $C-N$, to the amino group of amino acid $n + 1$. One water molecule is eliminated in this process. The repeating units, which are called residues, are divided into main-chain atoms and side chains. The main-chain part, which is identical in all residues, contains a central C_{α} atom attached to an NH group, a $C' = O$ group, and an H atom. The side-chain R, which is different residues, is bound to the C_{α} atom (see [3]).

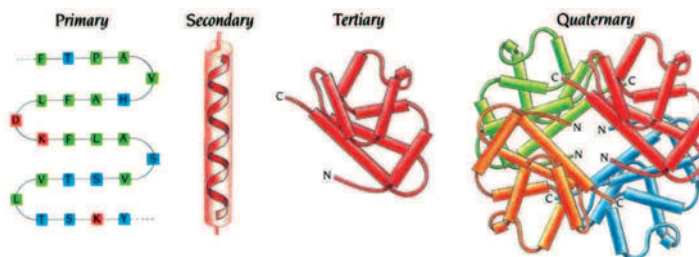


Fig. 2. – The amino acid sequence of a protein's polypeptide chain is called its *primary* structure. Different regions of the sequence form local regular *secondary* structure, such as alpha (α) helices or beta (β) strands. The *tertiary* structure is formed by packing such structural elements into one or several compact globular units called *domains*. The final protein may contain several polypeptide chains arranged in a *quaternary* structure. By formation of such tertiary and quaternary structure amino acids far apart in the sequence are brought close together in three dimensions to form a functional region, an *active site* (see [3]).

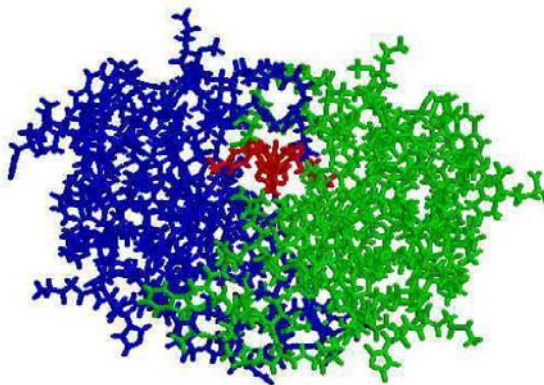


Fig. 3. – All atoms representation of the homodimeric protein HIV-1-PR. Also shown in red, is a drug designed to block the active site of the protease.

in solving the protein folding problem one needs: a) to develop a model of proteins which contains the essentials of the system but which, at the same time, is sufficiently simple so as to allow the calculation of the variety of dynamic and static properties characterizing the system; b) to recognize the fact that one is simply forced to simplify the forces acting among the amino acids. The basis for fulfilling requirements a) and b) are displayed in figs. 4 and 5 and is the theme of sect. 2.

Concerning the importance of the protein folding problem (see meaning 1) above), it can easily be appreciated through the fact that if one knew the details of the folding process, one could design inhibitors of this process and thus of the biological activity of any chosen target protein. In particular of proteins which play a central role in the life cycle of bacteria and viruses. For example of the HIV (Human Immunodeficiency

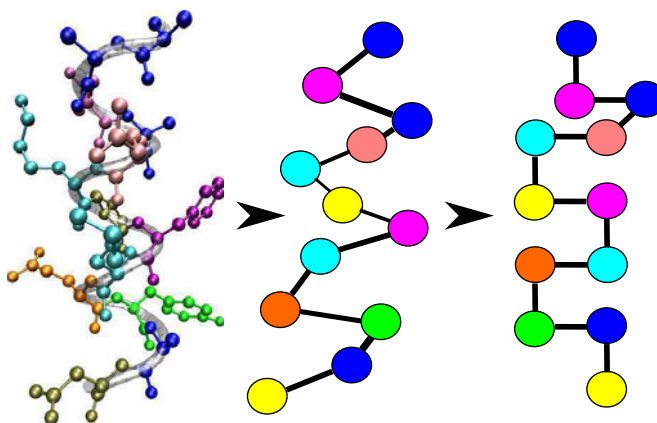


Fig. 4. – Different levels of simplification in the description of the polypeptide chain.

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	-0,13	0	0,12	0,26	0,03	-0,07	0,34	-0,22	0,14	-0,01	0,25	0,28	0,1	0,08	0,43	-0,06	-0,09	-0,1	-0,09	0,09
C	0	-1,06	0,03	0,69	-0,23	-0,08	-0,19	0,16	0,71	-0,08	0,19	0,13	0	0,05	0,24	-0,02	0,19	0,06	0,08	0,04
D	0,12	0,03	0,04	-0,15	0,39	-0,22	-0,39	0,59	-0,76	0,67	0,65	-0,3	0,04	-0,17	-0,72	-0,31	-0,29	0,58	0,24	0
E	0,26	0,69	-0,15	-0,03	0,27	0,25	-0,45	0,35	-0,97	0,43	0,44	-0,32	-0,1	-0,17	-0,74	-0,26	0	0,34	0,29	-0,1
F	0,03	-0,23	0,39	0,27	-0,44	-0,38	-0,16	-0,19	0,44	-0,3	-0,42	0,18	0,2	-0,29	0,41	0,29	0,31	-0,22	-0,16	0
G	-0,07	-0,08	-0,22	0,25	-0,38	-0,38	0,2	0,25	0,11	0,23	0,19	-0,14	-0,11	-0,06	-0,04	-0,16	-0,26	0,16	0,18	0,14
H	0,34	-0,19	-0,39	-0,45	-0,16	0,2	-0,29	0,49	0,22	0,16	0,99	-0,24	-0,21	-0,02	-0,12	-0,05	-0,19	0,19	-0,12	-0,34
I	-0,22	0,16	0,59	0,35	-0,19	0,25	0,49	-0,22	0,36	-0,41	-0,28	0,53	0,25	0,36	0,42	0,21	0,14	-0,25	0,02	0,11
K	0,14	0,71	-0,76	-0,97	0,44	0,11	0,22	0,36	0,25	0,19	0	-0,33	0,11	-0,38	0,75	-0,13	-0,09	0,44	0,22	-0,21
L	-0,01	-0,08	0,67	0,43	-0,3	0,23	0,16	-0,41	0,19	-0,27	-0,2	0,3	0,42	0,26	0,35	0,25	0,2	-0,29	-0,09	0,24
M	0,25	0,19	0,65	0,44	-0,42	0,19	0,99	-0,28	0	-0,2	0,04	0,08	-0,34	0,46	0,31	0,14	0,19	-0,14	-0,67	-0,13
N	0,28	0,13	-0,3	-0,32	0,18	-0,14	-0,24	0,53	-0,33	0,3	0,08	-0,53	-0,18	-0,25	-0,14	-0,14	-0,11	0,5	0,06	-0,2
P	0,1	0	0,04	-0,1	0,2	-0,11	-0,21	0,25	0,11	0,42	-0,34	-0,18	0,26	-0,42	-0,38	0,01	-0,07	0,09	-0,28	-0,33
Q	0,08	0,05	-0,17	-0,17	-0,29	-0,06	-0,02	0,36	-0,38	0,26	0,46	-0,25	-0,42	0,29	-0,52	-0,14	-0,14	0,24	0,08	-0,2
R	0,43	0,24	-0,72	-0,74	0,41	-0,04	-0,12	0,42	0,75	0,35	0,31	-0,14	-0,38	-0,52	0,11	0,17	-0,35	0,3	-0,16	-0,25
S	-0,06	-0,02	-0,31	-0,26	0,29	-0,16	-0,05	0,21	-0,13	0,25	0,14	-0,14	0,01	-0,14	0,17	-0,2	-0,08	0,18	0,34	0,09
T	-0,09	0,19	-0,29	0	0,31	-0,26	-0,19	0,14	-0,09	0,2	0,19	-0,11	-0,07	-0,14	-0,35	-0,08	0,03	0,25	0,22	0,13
V	-0,1	0,06	0,58	0,34	-0,22	0,16	0,19	-0,25	0,44	-0,29	-0,14	0,5	0,09	0,24	0,3	0,18	0,25	-0,29	-0,07	0,02
W	-0,09	0,08	0,24	0,29	-0,16	0,18	-0,12	0,02	0,22	-0,09	-0,67	0,06	-0,28	0,08	-0,16	0,34	0,22	-0,07	-0,12	-0,04
Y	0,09	0,04	0	-0,1	0	0,14	-0,34	0,11	-0,21	0,24	-0,13	-0,2	-0,33	-0,2	-0,25	0,09	0,13	0,02	-0,04	0,06

Fig. 5. – The interaction matrix derived by Miyazawa and Jernigan (see [4] table VI).

Virus) responsible for AIDS (Acquired Immuno Deficiency Syndrome). In this case, a particularly suited target is the HIV-1-Protease.

In the present contribution we shall aim at linking the physical case of the (lattice) model solution of the protein folding problem (see contribution of Tiana *et al.* for the extension to the continuum), to the design of a folding inhibitor of the HIV-1-PR which is likely not to create resistance. The assesment of the possibilities that this inhibitor can be used as a lead in the design of a (non-conventional) drug to fight AIDS is made in the contribution of D. Provasi *et al.* (*in vitro* experiments) and of S. Rusconi *et al.* (experiments in infected cells).

2. – The lattice model

A useful theoretical approach to study protein folding is a simplified lattice model, where the protein is represented by a string of beads that is arranged on a cubic lattice. The configurational energy of a chain of N monomers is given by

$$(1) \quad E = \frac{1}{2} \sum_{i,j}^N U_{m(i),m(j)} \Delta(|\vec{r}_i - \vec{r}_j|),$$

where $U_{m(i),m(j)}$ is the effective interaction potential between monomers $m(i)$ and $m(j)$, $\vec{r}_i$ and $\vec{r}_j$ denote their lattice positions and $\Delta(x)$ is the contact function. In eq. (1) the pairwise interaction is different from zero when i and j occupy nearest-neighbour sites, *i.e.* $\Delta(a) = 1$ and $\Delta(na) = 0$ for $n \geq 2$, where a indicates the step length of the lattice. In addition to these interactions, it is assumed that on-site repulsive forces prevent two amino acids to occupy the same site simultaneously, so that $\Delta(0) = \infty$ (excluded volume ansatz). The folding of the chain is simulated by Monte Carlo (MC) methods. As far as

the interaction among amino acids is concerned, we shall consider a 20-letters representation of protein sequence where U is a 20×20 matrix. A possible realization of this matrix (cf. fig. 5) is given in ref. [4] (table VI), where it was derived from frequencies of contacts between different amino acids in protein structures. The model we study here is a generic heteropolymer model which has been shown to reproduce important generic features of protein folding thermodynamics and kinetics, independent of the particular potential chosen [5,6]. This is achieved by using the same potential to design sequences and to simulate folding. However, in using such an approach, one should keep in mind that the labelling of amino acids (spherical beads all of the same size and with no side chain) is generic too and may be no obvious relation between those labels and labels for real amino acids.

2'1. Inverse folding problem: the design of good folders. – Although simple, the lattice model is still too complicated to be used in trying to solve the protein folding problem (see however sect. 4). Previous to this we shall discuss a related, although simpler problem, that is, the inverse folding problem. This problem can be economically stated in terms of the question: which are the sequences of amino acids which fold on short call to a selected native conformation?

This problem has, at least for small, single domain proteins, a simple solution: good folder sequences are characterized by a large gap $\delta = E_C - E_N$ (compared to the standard deviation σ of the contact energies) between the energy E_N of the sequence in the native conformation and the lowest energy E_C of the conformations structurally dissimilar to the native conformation [6-8], E_C being a quantity which is solely determined by the composition of the protein. Operatively, protein sequences are designed by minimizing, for fixed amino acid composition, the energy of the native conformation with respect to the amino acid sequence. In other words, good folders are associated with a normalized gap $\xi = \delta/\sigma \gg 1$, quantity closely related to the z -score [9]. Furthermore, starting from a designed sequence which displays a large gap, all mutated sequences which preserve (to some extent) the gap fold into the native conformation [10].

2'2. Role of the different amino acids in the folding process. – We shall now discuss some of the results of a Monte Carlo (MC) simulation study of the dynamics of a 36-mer chain characterized by a polymer sequence, denoted S_{36} (cf. fig. 6(b)), designed by minimizing the energy in the target (native) conformation shown in fig. 6(a). In the units we are considering [4] ($RT_{\text{room}} = 0.6 \text{ kcal/mol}$), the energy of S_{36} is $E_{\text{nat}} = -16.5$. While this is not the sequence of lowest energy, in particular the sequence displayed in fig. 6(c) has an energy in the native conformation of -17.13 , S_{36} has a sufficiently low energy and a large value of ξ ($= 8.33$) so that it can encode for a “wild-type” protein.

Monte Carlo simulations of folding performed on S_{36} at $T = 0.20$ (in our units) and using a standard algorithm described extensively in the literature [11, 12], in which, at each MC step, a monomer is picked up at random and end and crankshaft moves as well as corner flips are considered, indicate that this designed chain folds in a rather short “time”, of the order of $8 \cdot 10^6$ MC steps. At $T = 0.28$ the folding time is even shorter, $6.5 \cdot 10^5$ MC steps ($1 \text{ MC step} \approx 10^{-13} \text{ s}$). The fractional population of the native

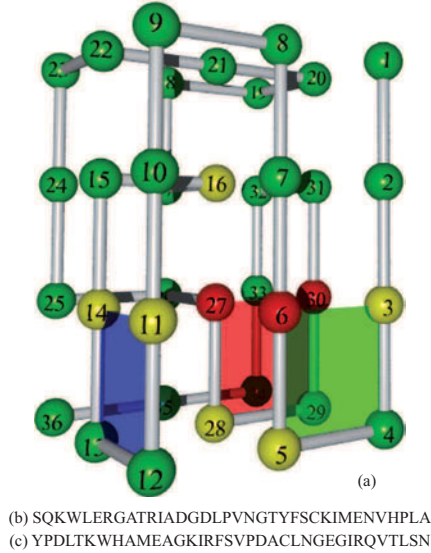


Fig. 6. – (a) The conformation of the 36-mer chosen as the native state in the design procedure. Each amino-acid residue is represented as a bead occupying a lattice site. The design tends to place the most strongly interacting amino acids in the interior of the protein where they can form most contacts. In real proteins these amino acids are highly hydrophobic. The strongest interactions are between groups *D*, *E* and *K* (cf. (b)), the last one being buried deep in the protein (amino acid in site 27). Amino acids occupying “hot” sites (sites 6, 27, 30) have been represented by red beads, those occupying “warm” sites (sites 3, 5, 11, 14, 16 and 28) by yellow and those occupying cold sites by green beads. The local elementary structures (LES) formed by the amino acid sequences $S_4^1 \equiv (3, 4, 5, 6)$, $S_4^2 \equiv (27, 28, 29, 30)$ and $S_4^3 \equiv (11, 12, 13, 14)$ and stabilized by the contacts 3-6, 27-30 and 11-14 are explicitly shown by shadowed areas. (b) Designed amino acid sequence S_{36} . (c) Designed sequence S'_{36} corresponding to $E_n = -17.13$.

state corresponding to these two temperatures ($T = 0.20$ and 0.28) is 91% and 10%, respectively, to be compared with a population of 0.5 and of 10^{-5} for the heteropolymer *folding temperatures* of $T = 0.25$ (temperature at which the probability for folding as well as for unfolding is $1/2$) and $T = 0.40$ (temperature at which native contacts break essentially as fast as they are formed due to thermal fluctuations) respectively. All the calculations discussed below were carried out (unless otherwise stated) at the temperature $T = 0.28$, optimal from the point of view of allowing for the accumulation of representative samples of the different simulations, and at the same time leading to a consistent population of the native conformation.

Mutations in the designed sequence are introduced by replacing a single monomer in S_{36} by a monomer of a different type. In our case, there exist 19 such possible substitutions for each monomer in S_{36} , implying $36 \times 19 = 684$ different sequences to study. Actually, a systematic behavior of folding can be deduced by analyzing a relatively small number of such altered chains. To show this, we start our analysis by choosing a

residue of S_{36} arbitrarily, and study the folding dynamics of the corresponding 19 altered chains. The MC simulations were performed up to a maximum number of $15 \cdot 10^6$ steps, averaging over 12 different starting random coil configurations for each altered sequence. By repeating the same procedure for other few selected sites along the original chain, it is possible to conclude that the behavior of altered chains can be generally classified into three categories:

- (1) chains which still fold to the native structure,
- (2) chains which fold to a unique compact structure, but different than the native one, and
- (3) chains which, although becoming compact, do not fold to a unique structure at all during simulations.

To characterize quantitatively the above three different behaviors, we find that the quantity ΔE_{loc} defined as the difference between the energies of the altered and the intact (“wild-type”) chain, both calculated in the native configuration, plays a key role. More precisely, such local energy difference, $\Delta E_{\text{loc}}[m'(i) \rightarrow m(i)]$, for a mutation at site i , where the monomer $m(i)$ in S_{36} is replaced by a monomer $m'(i) \neq m(i)$ is given by

$$(2) \quad \Delta E_{\text{loc}}[m'(i) \rightarrow m(i)] = \sum_{j \neq i} (U_{m'(i), m(j)} - U_{m(i), m(j)}) \Delta(|\mathbf{r}_i - \mathbf{r}_j|).$$

We have calculated all the 684 values of $\Delta E_{\text{loc}}[m' \rightarrow m]$, and found they fall in the range $0 \leq \Delta E_{\text{loc}}[m' \rightarrow m] < 5.66$, a number to be compared with the energy gap $\delta = 2.5$ associated with S_{36} in the native conformation. We classify the impact of mutation by the ability of the mutated sequence to fold into or close to the native conformation. We define the degree of folding (similarity parameter) q [13] as the fraction of correctly formed contacts ($q = 1$ corresponds to the native state and $q \ll 1$ corresponds to misfolded states).

The following rules are obtained from the results of the calculations:

- (1) $\Delta E_{\text{loc}}[m' \rightarrow m] < 1$: the altered chain always folds to the native structure ($q = 1$).
- (2) $1 < \Delta E_{\text{loc}}[m' \rightarrow m] < 2.5$: the altered chain folds to a unique structure, sometimes different than the native one, with q being smaller but close to one. In some cases, however, folding to the native structure may still occur ($q = 1$).
- (3) $2.5 < \Delta E_{\text{loc}}[m' \rightarrow m] < 4$: twilight zone: for some mutations, chains fold into near native structure with $q \sim 1$, other mutations lead to misfolding with $q \ll 1$.
- (4) $\Delta E_{\text{loc}}[m' \rightarrow m] > 4$: the altered chain does not fold to a unique structure at all during the simulation time, and now $q \ll 1$.

For a given site i , we can therefore classify the 19 possible mutations according to the rules (1–4). For rule (2), additional information about the dynamical behavior of the chain is required. We find that 73.675% mutations fall into the first class $\Delta E_{\text{loc}}[m' \rightarrow m] < 1$, 26.3% into the second class $1 < \Delta E_{\text{loc}}[m' \rightarrow m] < 4$, and the rest 0.015% into the class (4). Thus, a relatively large fraction of mutations yield altered chains still folding into the native structure, some mutations lead to limited misfolding and only a small fraction of mutations leads to complete misfolding.

Mutations at a given site may yield values $\Delta E_{\text{loc}}[m' \rightarrow m]$ which do not correspond to a single class, *i.e.* sometimes values smaller and larger than one occur at the same site. However, assuming that mutations are not selective, *i.e.* that they all occur with equal probability at a single site, an approximate scheme can be envisaged to classify the different sites according to the average magnitude of the damage (impact energy) caused by mutations. This is done by calculating the average value of $\Delta E_{\text{loc}}[m' \rightarrow m]$ for each site i as

$$(3) \quad \Delta \bar{E}_{\text{loc}}(i) = \frac{1}{19} \sum_{m'} \Delta E_{\text{loc}}[m'(i) \rightarrow m(i)].$$

In this way, to each site i is associated a mean value $\Delta \bar{E}_{\text{loc}}(i)$ and the following simple scheme emerges (remember that $\delta = 2.5$ for S_{36}):

- (1) when $\Delta \bar{E}_{\text{loc}}(i) < 1$, chains having mutations at site i are likely to belong to the first category, *i.e.* on average they fold to the native structure and we denote i as a “cold” site,
- (2) when $1 < \Delta \bar{E}_{\text{loc}}(i) < 2$ they behave on average as in the second category mentioned above, and i is denoted as a “warm” site,
- (3) when $\Delta \bar{E}_{\text{loc}}(i) > 2$ the resulting altered chains are likely to yield unfolded structures, and the site is denoted as a “hot” site.

We have classified the 36 monomers of S_{36} according to this scheme as shown schematically in fig. 6(a). We find that 27 sites can be considered as cold sites (green beads), 6 as warm sites (*i.e.* $\approx 17\%$ of the amino acids represented by yellow beads) and only 3 (*i.e.* $\approx 8\%$ of the amino acids represented by red beads) as hot ones (see table I(a)). Thus, 75% of the heteropolymer chain admits single error substitutions in the correct amino acid sequence yielding altered chains still folding to the native structure. Only in a relatively small fraction of the chain (about 10–15%, *i.e.* on all the hot sites and some of the warm sites), mutations have catastrophic effects leading to complete misfolding. Additional simulations have confirmed the general trend predicted by this empirical scheme.

2'3. Extension of the inverse folding strategy. – The strategy of minimizing the energy of a sequence in the native conformation with respect to amino acid sequence for fixed composition (relative frequency of the different types of amino acids) leads to good folders. This is one of the characteristics displayed by proteins, but says nothing concerning active site activity, nor aggregation (see, *e.g.*, [14]). In fact, a better notional protein could, in principle, be obtained by minimizing energy imposing at the same time maximization of activity (for example cleavage) and minimization of aggregation (see figs. 7 and 8). This procedure could shed light into the question of how and if activity affects folding and vice versa, a question which could be of relevance in the design of conventional, active site centered drugs. To gain a better understanding concerning the relation between folding and reduced aggregation may be of importance to learn, for example, how to fight fibril connected diseases (like Alzheimer, Creutzfeld-Jakob, type-2 diabetes, Parkinson, etc.).

TABLE I. – (a) The average values of $\Delta\bar{E}_{\text{loc}}$ for each site of S_{36} (sequence (b) in fig. 6 equal to sequence (a) of fig. 12 belonging to the supercluster with FN_1 , see fig. 12(g)). Indicating the hot (red) and warm (yellow) sites; (b) same as in (a) but for the sequence S_{36}^a (see fig. 12(b)) belonging to a different supercluster than S_{36} (supercluster with FN_2 , see fig. 12(e) and (q)).

(a)	Site	$\Delta\bar{E}_{\text{loc}}$	Site	$\Delta\bar{E}_{\text{loc}}$	Site	$\Delta\bar{E}_{\text{loc}}$	Site	$\Delta\bar{E}_{\text{loc}}$
	1	0.24	10	0.64	19	0.38	28	1.31
	2	0.73	11	1.46	20	0.63	29	0.95
	3	1.94	12	0.50	21	0.65	30	2.48
	4	0.75	13	0.55	22	0.40	31	0.99
	5	1.09	14	1.85	23	0.28	32	0.78
	6	2.79	15	0.88	24	0.77	33	0.68
	7	0.77	16	1.79	25	0.44	34	0.42
	8	0.27	17	0.80	26	0.54	35	0.66
(b)	9	0.09	18	0.38	27	3.46	36	0.26
	Site	$\Delta\bar{E}_{\text{loc}}$	Site	$\Delta\bar{E}_{\text{loc}}$	Site	$\Delta\bar{E}_{\text{loc}}$	Site	$\Delta\bar{E}_{\text{loc}}$
	1	0.61	10	1.46	19	0.05	28	1.54
	2	1.94	11	0.52	20	0.75	29	0.56
	3	0.62	12	0.51	21	0.95	30	0.69
	4	0.39	13	0.83	22	0.45	31	2.49
	5	1.09	14	1.21	23	0.21	32	0.14
	6	0.92	15	1.30	24	0.54	33	0.74
	7	2.79	16	3.04	25	0.92	34	0.20
	8	0.66	17	0.75	26	1.14	35	0.83
	9	0.09	18	0.30	27	2.01	36	0.97

2.4. *How many mutations can a designed protein tolerate?* – A complete enumeration of all sequences S_{36} obtained by introducing mutations in S_{36} have been carried out up to seven mutations keeping fixed the amino acid composition of the chain (swappings), and up to four mutations without this constraint (pointlike). Carrying out Monte Carlo simulations of the dynamics of the mutated sequences with the same composition of S_{36} and $\Delta E < \delta$, one finds that, in 98% of the cases they can reach the native conformation in a time comparable to the folding time of S_{36} . Repeating the same analysis on 10^3 sequences with up to four pointlike mutations, we observed that only in 57 cases did the chain find conformations dissimilar from the native one, with lower energy, and it was not able to find the native conformation within the simulation time. In table II it is shown

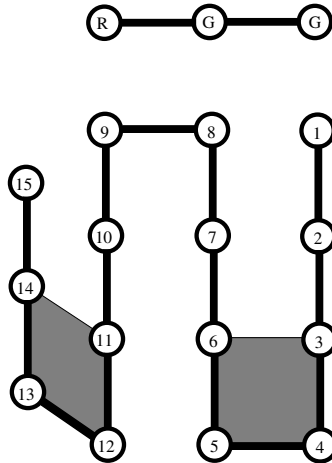


Fig. 7. – Possible extension of the inverse folding approach for the design of good folders. In this case swappings among the 36-mer amino acids are accepted if they reduce the energy of the sequence in the native state and at the same time, increase the interaction with the amino acids located in sites 1, 8, 9 with the substrate RGG (*i.e.* R interact with 9, G with 8 and G with 1).

that $2.78 \cdot 10^8$ sequences, *i.e.* 1.9% of all the sequences generated by seven composition conserving mutations which preserve (to any extent) the gap, provide a lower limit to the total number of sequences folding into the same structure. In particular, it is estimated that there exist about 10^{30} sequences which fold to the native conformation shown in fig. 6(a). These results demonstrate that good folders are highly designable. The lesson emerging from the extreme resilience demonstrated by the 36-mer to mutations will be particularly useful in discussing the stability of the HIV-1-PR to point mutations and the eventual manifestation of drug resistance.

The degeneracy of the protein folding code is well-documented: many sequences exist that can fold to similar native conformations [3, 15]. Besides homologs, *i.e.* proteins that have a clear evolutionary connection, homologous sequences (differing by a few neutral mutations) which are often (but not always) functionally related, there exist analogs, *i.e.* structurally similar proteins that have non-homologous sequences (with sequence identity

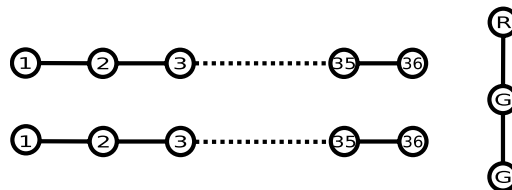


Fig. 8. – The same as in fig. 7 but now to the two requirements set above one adds that of making a minimum the probability that the two chains aggregate.

TABLE II. – *The number of mutated sequences S'_{36} which fold into the native conformation shown in fig. 6(a). In column one the number of mutations m is shown. Columns 2 and 3 are associated with composition conserving results (c.), while columns 4 and 5 correspond to pointlike mutations (n. c.). Columns 2 and 4 display the number of sequences associated with a change in energy ΔE smaller than the gap δ , while columns 3 and 5 display the total number of sequences associated with the number of mutations m (for details see ref. [10]).*

m	$\Delta E < \delta$ (c.)	Tot (c.)	$\Delta E < \delta$ (n.c.)	Tot (n.c.)
1			613	684
2	447	630	$1.59 \cdot 10^5$	$2.27 \cdot 10^5$
3	6518	14280	$2.30 \cdot 10^7$	$4.89 \cdot 10^7$
4	$1.37 \cdot 10^5$	$5.30 \cdot 10^5$	$1.99 \cdot 10^9$	$7.68 \cdot 10^9$
5	$2.10 \cdot 10^6$	$1.66 \cdot 10^7$		
6	$2.53 \cdot 10^7$	$5.16 \cdot 10^8$		
7	$2.78 \cdot 10^8$	$1.55 \cdot 10^{10}$		

less than 20%), unrelated functions and no evident evolutionary relation [15, 16]. *The analysis of the origin of analogs emphasizes the physical aspect of molecular evolution because they share common fold but not function.*

An important question is whether present analogs emerged as a result of a long divergent evolution or they originated from dissimilar sequences/structures and converged to structurally homologous folds? A physical approach to address this question is to study the “topography” of the space of sequences that fold into the same target conformation. In particular, the connectivity of the space of sequences via neutral nets (*i.e.* mutations that preserve the foldability into this structure [17]) may be a good evidence for divergent evolution as an origin of analogs, while the presence of disconnected “isles” in sequence space would be an argument in favor of convergent evolution origin of analogs. This question can be addressed rigorously by analyzing the properties of the sequence space within the framework of lattice model.

Within the present context, the lattice model has two clear advantages. First, folding properties of each sequence can be examined directly from folding dynamics simulations which are quite feasible for such models. Second, all sequences designed to fold into a given conformation have the same length. This factor simplifies comparison between sequences eliminating the need to introduce arbitrary gap/insertion penalties.

In particular, a single parameter, defined as the ratio of non-matching residues to the total number of residues (an overlap parameter) in a straightforward alignment of two sequences can serve as a natural measure of similarity between them. Using this parameter as a metric in sequence space we show that sequences folding to the same native conformation cluster together in a rather articulated, hierarchical fashion, in the space of sequences. Such clustering has a direct evolutionary implications, and, more relevant in connection with the main subject of the present review, implications concerning the

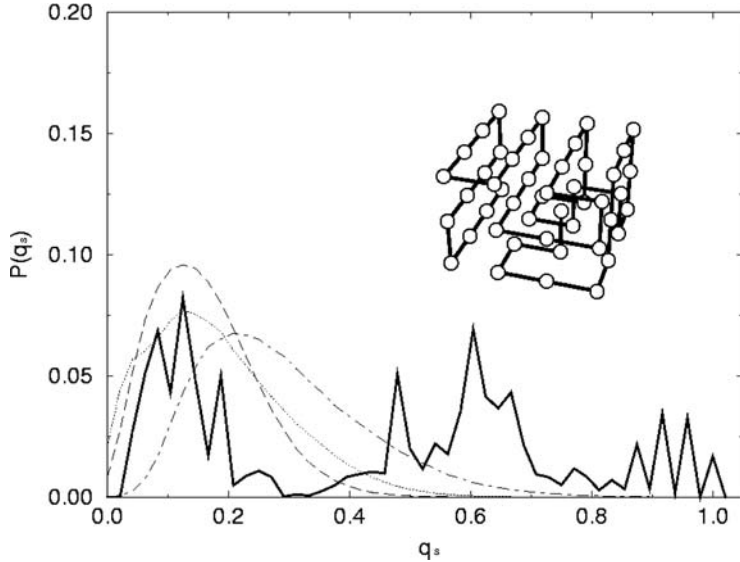


Fig. 9. – The probability distribution for $q_s^{\alpha\beta}$ over 1000 sequences with energy below -25.50 found in MC minimizations at temperature $T = 0.01$ (solid line). The target conformation is displayed in the inset. The other curves correspond to $T = 0.05$ (dash-dotted), $T = 0.08$ (dotted) and $T = 0.10$ (dashed).

insurgence or less of non-conventional drug resistance to folding inhibitors (see sect. 6).

As previously stated, the similarity parameter $q_s^{\alpha\beta}$ between sequence α and sequence β , defined as

$$(4) \quad q_s^{\alpha\beta} = \frac{1}{N} \sum_{i=1}^N \delta(\sigma_i^\alpha, \sigma_i^\beta),$$

serves as a metric in sequence space. Here $\delta(\sigma_i^\alpha, \sigma_i^\beta)$ is 1 if the i -th residue of sequence α is equal to the i -th residue of sequence β and zero otherwise. A natural parameter to describe the energy surface in sequence space is the distribution

$$(5) \quad P(q_s) = \langle \delta(q_s - q_s^{\alpha\beta}) \rangle,$$

where $\langle \rangle$ indicates a proper (“thermal”) average over all possible pairs of sequences α and β .

To study the properties of the distribution $P(q_s)$ we have chosen the 48-mer target structure (native conformation) shown in the inset to fig. 9. Using standard Monte-Carlo optimization in sequence space [18-20] we designed many sequences that have low energy in the target structure. The optimization in sequence space was carried out with constant aminoacid composition corresponding to “average” composition in natural proteins [2]. This condition implies optimization of the sequence fitness parameter ξ via optimization of the gap δ keeping variance of interaction energies σ (that depends on aminoacid

composition only [20]) fixed. The elementary MC move in sequence space that preserves aminoacid composition is swap of two randomly chosen residues. The acceptance of a “move” is controlled by applying a standard Metropolis criterion with “selective temperature” T that serves as a measure of a degree of evolutionary pressure towards sequences having larger energy gaps. For this reason we shall set a label “s” to this “selective temperature” (*i.e.* T_s) to clearly distinguish from the “normal temperature”, used for example, in studying the dynamics of the folding process (*cf.*, *e.g.*, fig. 13 below).

The results are shown in fig. 9 for several simulation temperatures ranging from $T = 0.01$ to $T = 0.08$. The $P(q_s)$ plot recorded at $T = 0.01$ features a number of peaks most pronounced centered at $q_s = 0.1$, $q_s = 0.55$ and $q_s = 0.95$ ⁽¹⁾. The pronounced peak at $q \approx 0.55$ disappears in the $P(q_s)$ plots taken at higher temperatures. For example, at $T = 0.08$ only two peaks are apparent, one at $q = 0.1$ and a set of smaller peaks in the high- q region.

Since our MC design sampling probes different regions of sequence space at low-temperature simulations and at high-temperature ones, the comparison of the results allows to get insight into the nature of energy landscape in sequence space in a range of energies. Apparently the low-temperature simulation at $T = 0.01$ samples the bottom part of the energy spectrum in sequence space. The peaks of $P(q_s)$ at $q_s \approx 1$ and at $q_s = 0.55$ indicate that the low-energy part of space of designed sequences can be divided into clusters, each cluster containing sequences that differ only by a few mutations. *Furthermore, within each cluster, there are 6–8 sites in which the associated residues are conserved. These sites correspond to all of the hot sites and some of the warm sites.* The size of these clusters varies to a considerable extent but no sequence is isolated. This means that, for each sequence, it is possible to introduce a swapping of amino acids between most of the cold sites and still obtain a good folding sequence belonging to the same cluster. The maximum number of swappings that preserve a sequence within a cluster is around 15% of the length of the chain. If this limit is exceeded, a wider cooperative rearrangement of residues (between 40% and 60%), corresponding to the second peak in fig. 9 is needed in order to produce a sequence which again displays an energy within the gap δ in the native conformation⁽²⁾. Such greater sequence rearrangements correspond to moves in the space of sequences, from one cluster to the next one.

⁽¹⁾ The calculation performed at $T = 0.01$, keeping only those sequences with $E < -25.50$ has the meaning of simulating the behaviour at $T \approx 0$, where the MC algorithm fails. For this case we found a dependence of the position of the peaks on the sampling rate, if the latter is $\gg 10^3$. This is a consequence of the entropic term in the free energy which would disappear if $T \approx 0$, so we discarded the results at large sampling rates as an artifact of the calculation.

⁽²⁾ This phenomenon seems, to some extent, inverse to what is observed in the mutation pattern of the HIV-1-PR subject to conventional drug pressure (Saquinavir, Indinavir, etc.). In this case *primary mutations* are observed which modify the active site strongly reducing the effectiveness of the drug followed by “*secondary mutations*” boosting the enzymatic activity somewhat impaired by the primary mutations. The resistance induced mechanism seems to be typical of active site centered site drugs. We shall see below that non-conventional inhibitors, that is inhibitor of the folding of the target protein may lead to non-conventional resistance, that is the evolution of the protein from one cluster to another and eventually from one supercluster to another.

Since all sequences belonging to a cluster are equivalent, *i.e.* they display $> 85\%$ overlap among them, it is possible to compare different clusters by comparing sequences chosen at random from them. In other words, each cluster behaves like a class of equivalence. From a detailed analysis of the sequences belonging to different clusters, it turns out that clusters group into “super-clusters”. Clusters within each super-cluster are associated with a similarity parameter $q_s \approx 0.55$, while clusters belonging to different super-clusters have $q_s \approx 0.10$ – 0.15 .

Within each super-cluster, most of the “hot” and a fraction of the “warm” sites are in the same (or very similar) position and the type of residues occupying each of these sites is conserved in all of the sequences. In other words, all the sequences belonging to a supercluster have the same or very similar FN. This fact emphasizes once more the importance of “hot” and “warm” sites to govern the low-energy properties of protein sequences. Furthermore, half of the “hot” plus “warm” sites of sequences belonging to the first super-cluster are in the same positions as the “hot” plus “warm” sites of sequences of the second super-cluster and the other half move to neighboring sites (see fig. 10A). The common “hot”-“warm” sites are different for residues which are very similar (meaning that the columns of MJ matrix indicating their interaction with all the other monomers, are very similar, *i.e.* D with E and Q, K with R, see fig. 11). This suggests that the presence of the two super-clusters may be due to the quasi-degeneracy of these matrix elements⁽³⁾. The substitution of a residue in a “hot” site with a very similar one (eventually a rather warm yellow-redish amino acid) causes, to a small extent, a rearrangement of the other “hot” sites, and to a much larger degree the rearrangement of the other (“warm” and “cold”) residues. To substantiate this scenario, we have repeated the calculation with a random-generated matrix having the same distribution of elements of the MJ matrix. Because in this case all the columns of the matrix are different, we expect no super-clusters. In fact, the distribution $P(q_s)$ for this case is very similar to the distribution shown in fig. 9, except for the absence of the peak at $q_s \approx 0.1$ (data not shown).

The schematic illustration of the low energy part of the sequence space is sketched in fig. 10B(a). For each good folder sequence it is possible to mutate a few “cold” sites, up to 15% of them and still obtain a good folding sequence. This produces a cluster. If more “cold” sites are changed, there is a partial rearrangement of many other “cold” sites to keep the energy low, leading to different clusters. When a “hot” site is mutated with a similarly strongly interacting residue (if the interaction matrix allows for such a possibility), the other “hot” sites become slightly rearranged, while the “cold” sites are

⁽³⁾ It is an open question how these results will find their counter part in the case of real proteins. The complexity of the (all atom) amino acid-amino acid interaction may open the possibility for high degeneracy. On the other hand, side chains confer the different amino acids very selective partnerships thus indicating an eventually low degeneracy. The validity of the second picture will made non-conventional drugs particularly attractive, as non-conventional drug induced resistance will be easy to fight through a cocktail of a reduced number of peptides representative of the LES associated with the (few) different FN. On the other hand, the first scenario (high degeneracy) will make non-conventional drug resistance a serious problem.

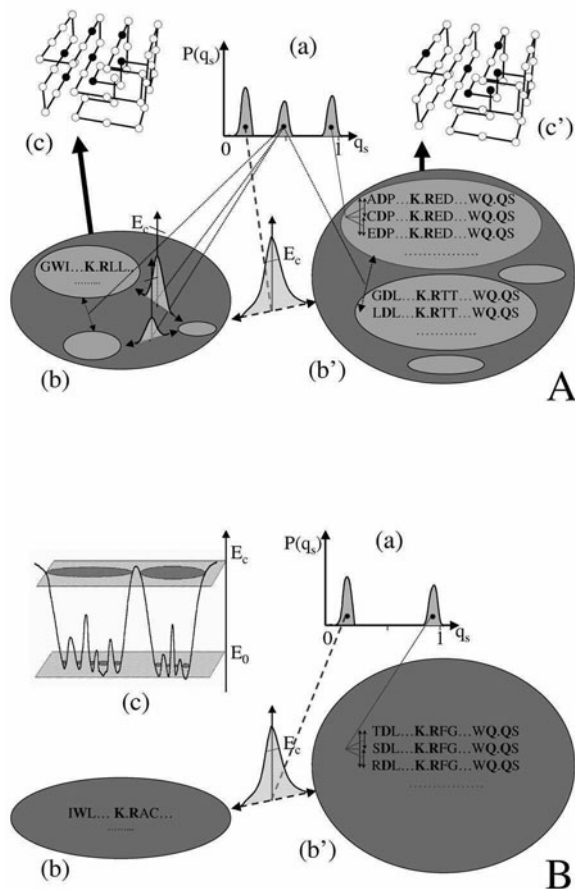


Fig. 10. – Schematic representation of the structure of the space of sequences folding into the same conformation. The calculations were carried out for a 48-mers chain making use of Miyazawa-Jernigan contact energies among the amino acids. (A) The space of 48-mers sequences, all folding to the same native conformation (cf. (c) or (c')) as derived from an analysis performed at very low evolutionary temperature ($T = 0.01$), resulting in the distribution $P(q_s)$ shown in (a). The main structure observed is that of superclusters (as shown in (b) and (b')) in dark grey) which are separated by potential barriers higher than E_c . Within each supercluster “hot” residues (black dots in (c) and (c')) are highly conserved. Between two superclusters the number, type and position of the “hot” residues can change to some extent, the associated change in contact energy being responsible for the large barrier shown between supercluster (b) and (b'). A supercluster is built out of a large number of clusters which are shown in light grey. Potential barriers among clusters display energies all smaller than E_c (cf. energy plots between clusters inside the supercluster (b)). (B) The space of 48-mers sequences all folding to the same native conformation (cf. (c) and (c')) in plot (A) of the figure), designed at a high evolutionary temperature ($T = 0.08$). In this case the function $P(q_s)$ shown in (a) displays only two peaks ($q_s \approx 1$ and $q_s \approx 0$), so that the space of sequences is divided only in superclusters ((b) and (b')). In (c) we display a schematic representation of the evolution of the space of designed sequences, making use of low- and high-evolutionary temperature.

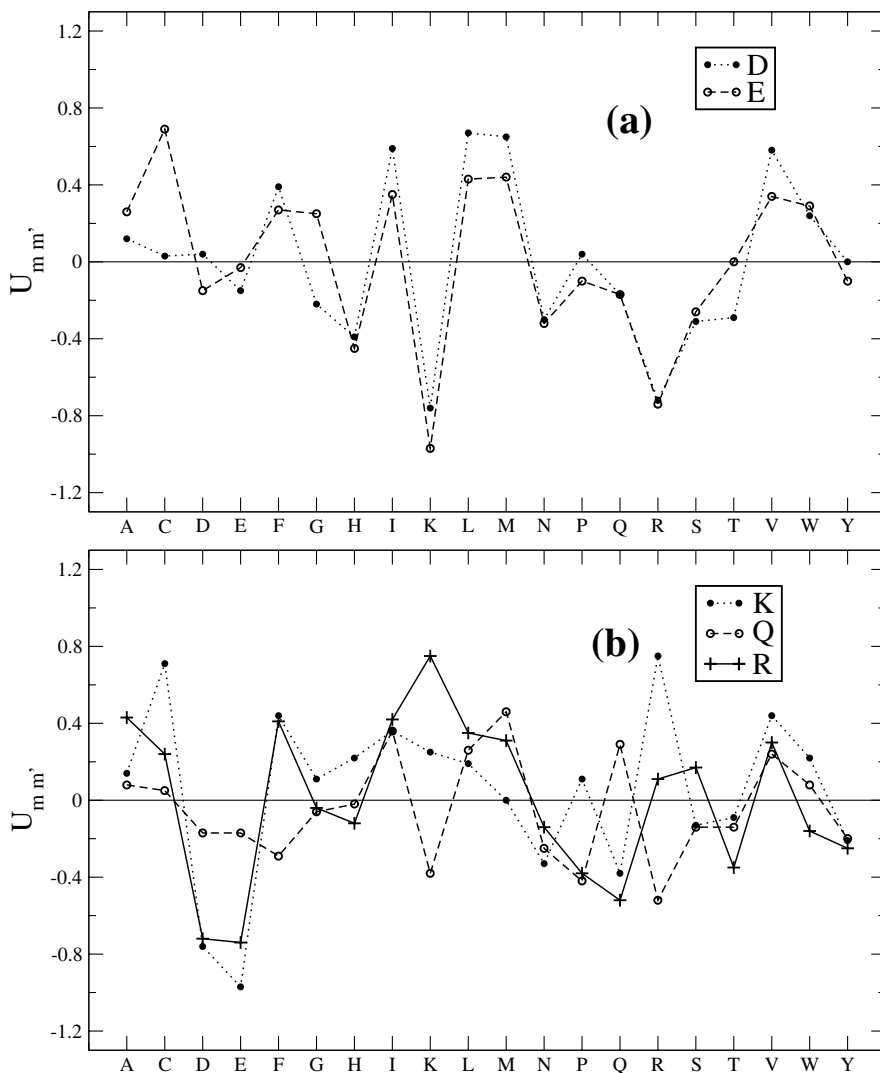


Fig. 11. – Miyazawa-Jernigan matrix elements associated with: (a) amino acids D and E (third and fourth lines of fig. 5), (b) amino acids K, Q and R (lines ninth, fourteenth and fifteenth of fig. 5).

heavily rearranged, the net result being a new super-cluster (see figs. 10B and 12).

An important aspect of this analysis concerns the height of the “barriers” in sequence space. This question is closely related to the issue of connectivity of clusters via neutral network, *i.e.* set of mutations that preserve the ability of all intermediate sequences to fold into the same structure. While the question of what constitutes a viable folding sequence is a complicated one having a number of aspects (thermodynamic, kinetic and functional aspects [17, 21, 22]) a necessary condition for a sequence to fold into a given structure is

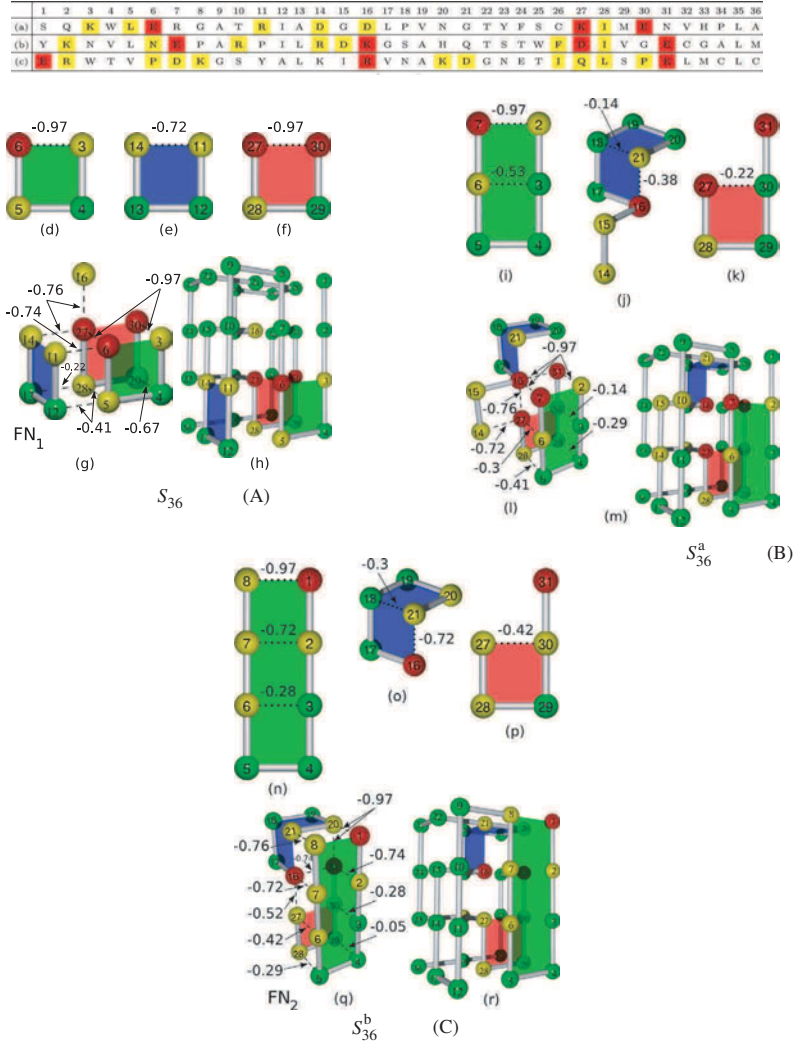


Fig. 12. – (A) The folding of the sequence S_{36} (see (a), as well as fig. 6(b)) into the native structure (h) (see also fig. 6(a)) is controlled by the three LES 3-6, 11-14, 27-30 shown in (d)-(f) (in terms of green, blue and red coloured surfaces). These LES build, in their relative native conformation, the folding nucleus $(FN)_1$ displayed in (g). The interactions stabilizing the LES are shown as dotted lines while those representing the interaction between LES are given in term of dashed lines. The corresponding contact energies (see fig. 5) are also shown. Hot, warm and cold amino acids are represented in terms of red, yellow and green beads, respectively. The sequence S_{36} belongs to one of the two superclusters associated with the degeneracy in contact interactions shown in fig. 11. The other two sequences labeled S_{36}^a and S_{36}^b (see (b) and (c)) belong to the second supercluster (see (B),(C)) associated with the $(FN)_2$ (see (q)), this in spite of the fact the folding nucleus shown in (l) and the LES shown in (i)-(k) correspond to a small variations of $(FN)_2$ and of the set of LES shown in (n)-(p). In fact, the sequences S_{36}^a and S_{36}^b are inhibited by the same p-LES, namely (16-21).

TABLE III. – The average values of $\Delta\bar{E}_{\text{loc}}$ for each site of S_{36}^b (sequence displayed in fig. 12(c)) corresponding to a second cluster within the supercluster associated with S_{36}^a (see fig. 12(e) and (q)).

Site	$\Delta\bar{E}_{\text{loc}}$	Site	$\Delta\bar{E}_{\text{loc}}$	Site	$\Delta\bar{E}_{\text{loc}}$	Site	$\Delta\bar{E}_{\text{loc}}$
1	2.05	10	0.46	19	0.11	28	1.31
2	1.46	11	0.61	20	1.94	29	0.15
3	0.47	12	0.20	21	1.80	30	1.08
4	0.09	13	0.80	22	0.74	31	2.51
5	0.56	14	0.46	23	0.49	32	0.47
6	0.94	15	0.90	24	0.69	33	0.71
7	1.76	16	2.66	25	0.34	34	0.03
8	1.73	17	0.98	26	1.23	35	0.92
9	0.37	18	0.79	27	1.28	36	0.44

that its energy in this structure is lower than E_C [7, 18, 20]. From this “minimalistic” point of view the clusters are connected by neutral networks (see fig. 12(B) and (C)) while superclusters are not—the top of the barrier in sequence space between them lies above the level of E_C , *i.e.* the mutation “pathway” between superclusters passes through the “sea” of non-folding sequences (see fig. 10).

The schematic representation of the structure and connectivity of the part of the sequence space sampled at higher temperatures (*i.e.* that fits the native structure with higher energy but still lower than E_C) is shown in fig. 10B. The main difference between $T = 0.01$ and higher temperature simulations is that the peak of $P(q_s)$ at $q = 0.55$ does not exist at higher temperature sampling. This indicates that the structure of the sequence space sampled at $T = 0.08$ does not include apparent clusters. However the superclusters remain the same with the same “hot” sites conserved in each of them. Figure 10B(c) schematically represents the energy landscape in the sequence space obtained both from low temperature sampling and high temperature sampling. It clearly shows that there are two characteristic barriers—between clusters and between superclusters. When the sampling temperature exceeds the former the apparent picture of clusters disappears and the simulations samples broad basins each corresponding to a supercluster, in our terminology.

In the case of the 36-mer discussed in connection with fig. 6 and tables I and II, we find that there are two superclusters (see discussion carried out above in connection with fig. 11 as well as fig. 19). The first one displaying the distribution of hot and warm sites shown in fig. 6 (see also fig. 12(A)). The second one is shown in fig. 12(C) (see also tables I(b) and III).

Summing up, stability of a designed sequence to mutations, and thus the number of

TABLE IV. – *Evolution of the folding nucleus of a 36-mer under the effect of mutations which brings the good folder from one supercluster to another. Starting from the sequence labelled $S_{36}(FN_1)$ (see figs. 6(b) and 12(a)) associated with the folding nucleus FN_1 (see fig. 12(g)), the system evolves into the mutated sequence $(S_{36})_{\text{mut}} = S_{36}^a$ (see fig. 12(b)) associated with the **intermediate** folding nucleus (FN_{mut}) , see figs. 12(l) to the sequence $S_{36}^b(FN_2)$ (see fig. 12(c)) associated with the supercluster FN_2 (see figs. 12(q)). From fig. 19 it is seen that $(S_{36})_{\text{mut}}$ which we call S_{36}^a belongs in fact to the folding nucleus FN_2 . Thus $(FN)_{\text{mut}} \in FN_2$.*

	LES			FN	E_N	δ/σ
$S_{36}(FN_1)$	3-6 (−0.97)	11-14 (−0.72)	27-30 (−0.97)	−8.57	−16.5	8.3
$S_{36}^a(FN_2)$	2-7 (−1.5)	14-21 (−0.72)	27-31 (−0.22)	−9.03	−16.34	7.8
$S_{36}^b(FN_2)$	1-8 (−1.97)	16-21 (−1.02)	27-31 (−0.42)	−8.39	−15.83	6.1

mutations which preserve folding to the same native structure, is controlled by the energy gap δ (see table II). Neutral mutations are those which lead to sequences which preserve (to any extent) the gap, a result which is tantamount to saying that the amino acids occupying most of the hot sites and some of the warm sites of the native conformation are conserved. Correlated simultaneous mutations on a number of cold and warm sites, and some hot sites, can change the folding nucleus (and thus the LES), bringing the sequence from one supercluster to another supercluster (see table IV).

3. – Hierarchical folding of a model protein

Let us study the time evolution of the different native contacts of the designed sequence S_{36} . A surprisingly simple movie runs under our eyes [23]. Starting from a random coil (fig. 13(a)), few local elementary structures (LES) built out of the monomer sequences $S_1 \equiv (3, 4, 5, 6)$, $S_2 \equiv (27, 28, 29, 30)$ and $S_3 \equiv (11, 12, 13, 14)$ are formed only after $\approx 10^2$ steps of the MC simulation (fig. 13(b)). They are controlled by the local contacts 3-6, 27-30 and 11-14, involving some of the most strongly interacting amino acids, in any case all those occupying the hot sites and some of those occupying warm sites [24] in the native conformation. These contacts achieve stability ($> 80\%$) after 10^5 MC steps and when they assemble together after $\approx 10^6$ MC steps (fig. 13(c)) they form the (post-critical) folding nucleus [25, 26] of the notional protein, from which it reaches the native conformation (fig. 13(d)) in a short time ($\approx 10^3$ MC steps), provided the energy of the system is lower than E_C , where it has not to compete with the bulk of misfolded conformations. The local elementary structures S_1 , S_2 and S_3 , and not the individual monomers, thus take care, through local guidance and non-local long-range correlations⁽⁴⁾ (bonding

⁽⁴⁾ Within the present scenario of protein folding the question of the relative importance of local- *vs.* non-local contacts is somewhat transcended. In fact, they become complementary expressions of the same reality and, in the same way that nothing is gained in expressing a quantum phenomenon in terms of particles, that cannot equally well be described in terms of waves, the paradigm of non-local contact description of protein folding, the formation of the

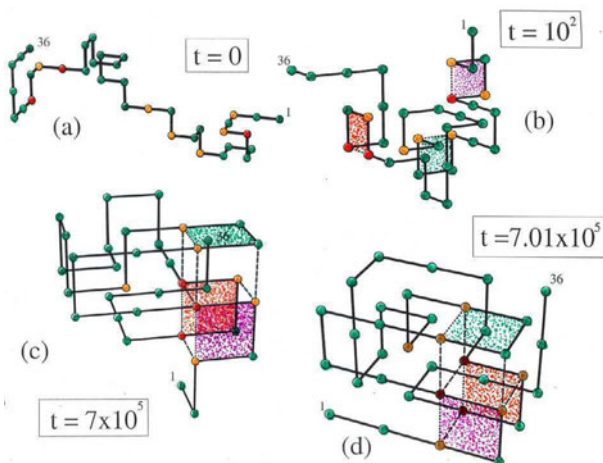


Fig. 13. – Snapshots of the folding of the sequence S'_{36} (cf. fig. 2(c)), whose energy in the native conformation is $E_n = -17.13$. Starting from a random conformation (a), the system forms after $\approx 10^2$ MC steps partially folded intermediates (b), involving three sets of four amino acids (3-6, 11-14, 17-30), whose stability is provided by the bonding indicated by dotted lines. When the partially folded intermediates come together to form the folding core (indicated by dotted and dashed lines) after $7 \cdot 10^5$ MC steps (c), the system folds to the native conformation after only 10^3 MC steps (d). The amino acids participating in the bonding of the partially folded intermediates (dotted lines) are among some of the most strongly interacting amino acids, which occupy, in the native conformation (d), “hot” and “warm” sites [24] indicated by red and yellow beads, respectively. The monomers number 36 of the sequence S'_{36} are indicated for each conformation.

between local elementary structures), of the process of protein folding⁽⁵⁾, and can be viewed as the local “bricks” of a dynamical LEGO kit to model protein folding.

If instead of S_{36} , one uses a sequence corresponding to the second supercluster, *e.g.*, S_{36}^b (see fig. 12(c)) one observes essentially the same hierarchical evolution, where in this case the LES seem to be $S_1'' = 1-8$, $S_2'' = 16-21$ and $S_3^b = 27-31$ (see fig. 12 (n)–(p)).

4. – Solving the protein folding problem in the case of a notional protein (three-step-strategy (3SS))

With the help of the results discussed above, we have developed a strategy which allows to predict the three-dimensional native conformation of a model protein from its amino acid sequence [27], provided the contact energies acting among the amino acids are known. The algorithm consists of three steps, namely 1) finding good candidates for

folding core, can be described in terms of local elementary structures, entities which epitomize the local-contact picture but at the same time also the phenomenon of molecular recognition, paradigm of non-local contacts.

⁽⁵⁾ This result is also supported by the disruptive effect mutations which affect the stability of these structures have on the folding ability of the design sequence.

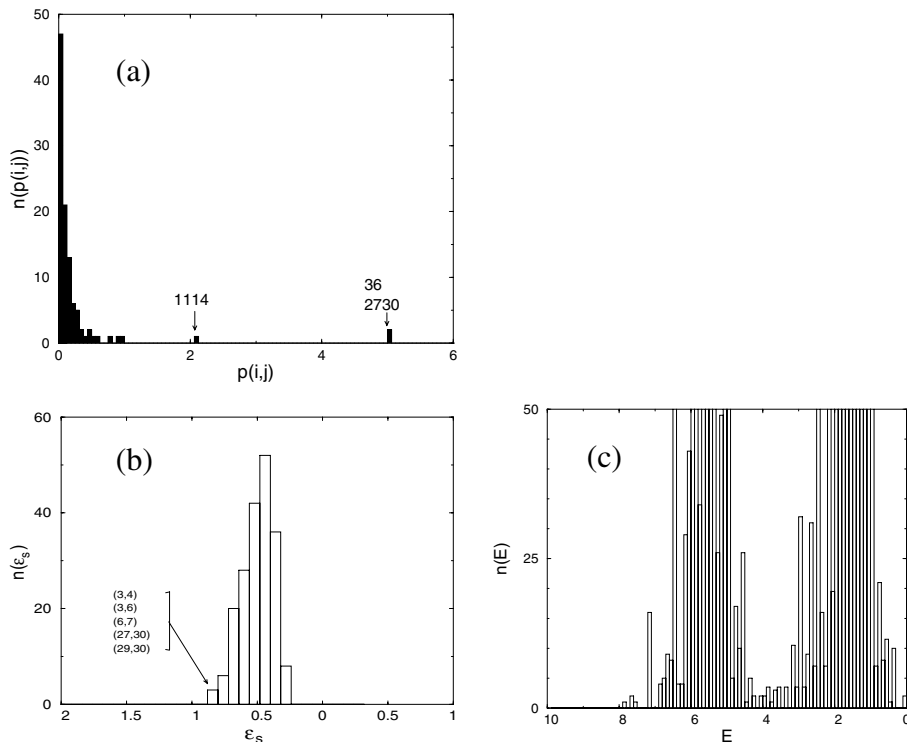


Fig. 14. – (a) The distribution of the parameter $p(i,j)$, whose maximization allows to find the closed elementary structures. (b) The distribution of the energy density ϵ_s , employed to find open elementary structures. (c) The distribution of the energies associated with the possible folding nuclei of sequence S'_{36} , build of the elementary structures 3-4-5-6, 11-12-13-14 and 27-28-29-30 (for details cf. ref. [27]).

the role of local elementary structures, 2) finding the folding nucleus and 3) finding the native conformation relaxing the residues not participating in the folding nucleus. This algorithm is based on the hierarchical sequence of events that allows the chain to fold fast and works because at each step only a limited portion of the configurational space has to be searched through. Examples of the three steps needed to identify the LES and the native structure of S_{36} are shown in fig. 14.

Another example is provided in fig. 15, where the results of the 3SS applied to a 36-mer optimized on a conformation of poor local contacts [28], which therefore provides an important test to the algorithm. In this case one finds that of the three LES, two of them are open (no internal contacts but the peptidic ones) while one is closed (internal contacts between amino acids 1 and 6, see fig. 15(b)), the folding nucleus being shown in fig. 15(c). These elements allow, also in this probing case, to determine the native structure shown in fig. 15(d). The time evolution of a number of important contacts is shown in fig. 15(e).

The correctness of the hierarchical view of protein folding based on LES and on the (post critical) folding nucleus is testified by the fact that, making use of these elements,

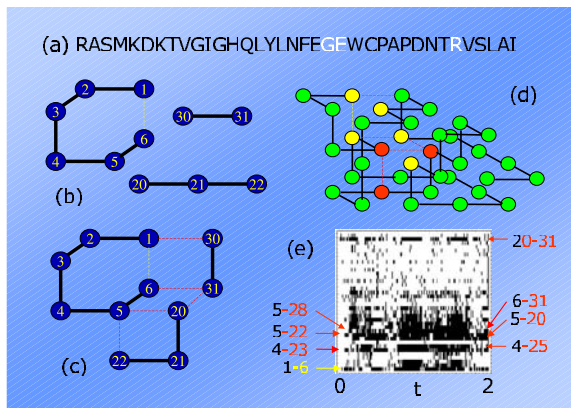


Fig. 15. – Prediction scheme for the sequence (a). In (b) are shown the resulting local elementary structures, two of which are open. All the amino acids participating in them are displayed in (a) in terms of white symbols. The same colour is used for the amino acid participating in the (internal) contacts of the closed structure. In (c) are shown the disposition of these structure to form the (unique) folding core with which is associated the (single) completely compact conformation (d) with energy lower then E_C . This conformation coincides with the one used in the literature [28] to design the sequence shown in (a), the associated (native) energy being $E_n = -15.99$. Furthermore, Monte Carlo simulations testify to the correctness of the predicted native conformation (d). This can be seen in (e), where the time evolution of the native contacts for a particular run is displayed as a function of the MC steps of the folding simulations. The sequence in (a) folds into the structure shown in (d) in approximately $2 \cdot 10^6$ MC steps. The normalized gap associated with sequence (a) is $\xi = 6.3$. The hot, warm and cold sites of the protein in its native conformation calculated according to [24] are displayed in terms of red, yellow and green beads, respectively.

and of the contact energies used to design the protein, one can determine the native structure of the protein from its amino acid sequence.

We conclude this Section noting that while the inverse folding model provides much insight into the folding of proteins, this insight is based on the knowledge of the native conformation. Consequently, the (3SS) is, to our knowledge, the only *bona fide* (model) solution of the protein folding problem.

5. – Lattice model design of resistance proof, folding-inhibitor peptides

In the previous section it has been shown that single domain proteins fold according to a hierarchical mechanism [23, 29]. Starting from an elongated, denaturated⁽⁶⁾, conformation, it is found that highly conserved and strongly interacting amino acids lying close along the designed chain form small local elementary structures (LES). Due to the

⁽⁶⁾ The unfolded state is the most difficult state to be characterized. One thing in any case seems clear: the unfolded state is formed by an ensemble of random conformation rich in some (local) native contacts associated with the stretches of the protein which eventually lead to LES.

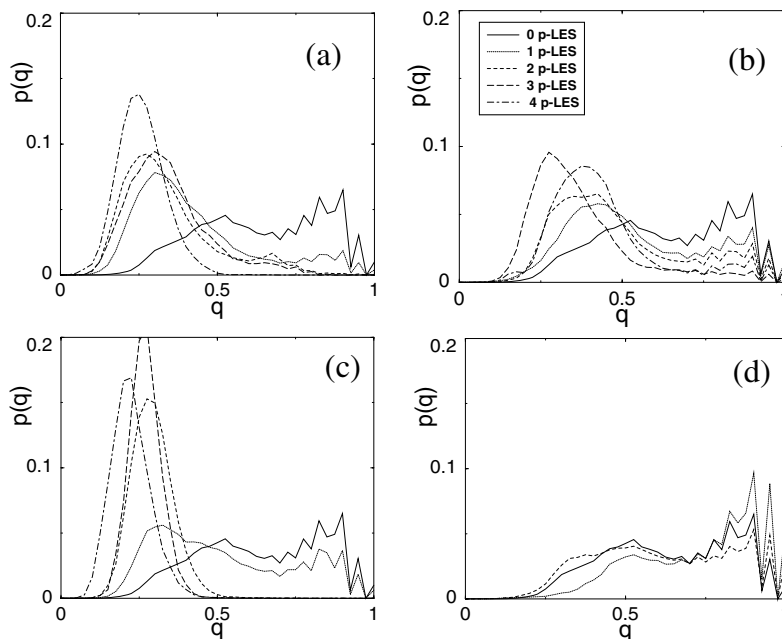


Fig. 16. – The equilibrium distribution of the order parameter q of the S_{36} sequence (see fig. 6(b)) in presence of n_p p-LES of kind 3'-6' (a), 11'-14' (b) and 27'-30' (c), calculated at temperature $T = 0.24$ in the units chosen ($RT_{\text{room}} = 0.6$ kcal/mol). As a control, a string corresponding to the residues 8-11 of the protein was also used (d).

small conformational space available and to their large attractive propensity, these LES are formed at a very early stage in the folding process and are very stable. The rate limiting step of this process corresponds to the assembly of the LES to build their native, nonlocal contacts (folding nucleus). This nucleation can be done in a relatively short time, because LES, moving as almost rigid entities, not only reduce the conformational space available to the protein but also display low probability of forming non-native interactions and thus aggregation (see [14]). Furthermore they interact with each other more strongly than single amino acids belonging to these structures do [23]. *In other words, LES, intermediates between single amino acids and the folding nucleus, provide the basis of molecular recognition directing the folding of a protein. The nucleation event corresponds to the overcoming of the major free energy barrier found in the whole folding process [26]. After this is accomplished the remaining conformational space available to the protein is so small that the system reaches the native state almost immediately.*

In keeping with these results it is not surprising that it is possible to destabilize the native conformation of a protein making use of peptides whose sequence is identical to that of the protein LES. Such peptides (p-LES) interact with the protein (in particular with the complementary fragments (LES) in the (post-critical) FN) with the same energy that stabilizes the nucleus, thus competing with its formation (see ref. [30]).

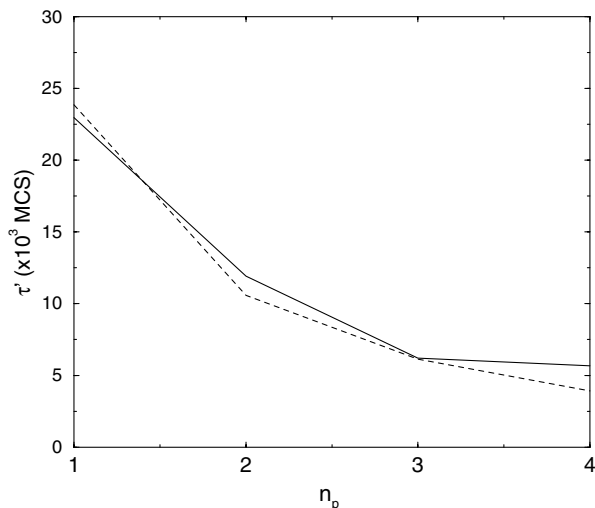


Fig. 17. – The mean binding time τ' between the residue 30 of the protein and 3' of the p-LES, as a function of the number n_p of p-LES (solid line). The result of MC simulations is compared with the random search time (dashed line).

This can be seen from the results reported in fig. 16, results obtained carrying out Monte Carlo simulations of the system composed of the target protein S_{36} and a number n_p of p-LES (at temperature $T = 0.24$) collecting the histogram of the order parameter q , defined as the relative number of native contacts, parameter which measures the extent to which the equilibrium state reached by the protein is similar or less to the native conformation.

In fig. 16(a) is displayed the equilibrium distribution of q , for the system composed of sequence S_{36} and a number of p-LES 3-6 as a function of n_p (concentration). While the distribution of q -values in the absence of the p-LES (solid line) shows a two-peak shape, reflecting an all-none transition between the native ($q > 0.7$) and the unfolded ($q < 0.6$) state, the presence of p-LES reduces markedly the stability of the protein. The effect of the other p-LES, *i.e.* 11-14 and 27-30, is similar to that found for the peptide 3-6 and are displayed in figs. 16(b) and 16(c).

The thermodynamics which is at the basis of the disruptive mechanism of p-LES is quite simple. Because the interaction between complementary LES and between p-LES and its partner is essentially the same, the second mechanism is preferred to the first. In this way the target protein increases its internal entropy (the unfolded state is much less organized than the native one) by an amount larger than that lost in translational entropy (of the peptide).

It has also been found that the dynamical properties of p-LES make them suitable to be used as drugs. This can be seen, for example, by calculating the probability $P(t)$ that the bond between the residue 30 of the protein and 3 of any of the p-LES 3-6 is

formed as a function of time starting from a random conformation of the protein and of the peptides. This bond is chosen as representative of the interaction between the whole LES 27-30 and the p-LES 3-6, the dynamics of the other bonds associated with the same LES being quite similar. The slope of the calculated probability function is well fitted by a single exponential $P(t) (1 - \exp[t/\tau'])$, where τ' is the characteristic time of bond formation. The dependence of τ' on the number of n_p of p-LES is shown in fig. 17 as a solid curve, where it is compared to the average time needed for the p-LES to build the bond 30-3 with the protein after a random search in the volume of the cell where the Monte Carlo simulations are carried out. The agreement between the two calculations indicates that the random search is the actual mechanism which leads to the binding of the p-LES to the complementary LES 27-30 (see also sect. 8). The fact that $P(t)$ is well reproduced by a single exponential indicates furthermore that this is the only mechanism operative. In particular, this result excludes the possibility that the p-LES binds tightly to some other part of the protein. Such a scenario would produce a double or more fold exponential shape of $P(t)$. We note that the binding time τ' of p-LES 3-6 to the protein is much shorter than the docking time between the complementary LES 3-6 and 27-30. The reason for this result is to be associated with the fact that, unlike LES, p-LES are not slowed down by the polymeric connection with the rest of the protein.

A consequence of this fact is the ability of p-LES to bind the LES of the protein even if this is in its equilibrium native state. The p-LES can take advantage of the thermal fluctuations of the protein and make use of the fact that these fluctuations display a recursion time (which, assuming that the system is ergodic, is equal to the mean first passage time) much longer than the time p-LES need to bind the LES, to enter and disrupt the protein. As a matter of fact, calculating the distribution of q starting from the protein in the native state, one finds the same distribution displayed in fig. 16 (equilibrium distribution)⁽⁷⁾.

To further test the validity of these results, we have repeated the above calculations making use of peptides corresponding to segments of the protein sequence other than those corresponding to LES. In fig. 16 the effect of peptides corresponding to residues 8-11 is shown. One can notice that the protein is not destabilized to any significant extent.

Calculations as those described earlier and leading to the results displayed in figs. 16 and 17 have also been carried out for another 36-mer model protein (see fig. 18). The outcome of these calculations are, as a rule, in agreement with those found in connection with sequence S_{36} . To be noted, however, an important difference found in connection

⁽⁷⁾ Note that in the case of molecular dynamic simulations as well as target protein tests of the inhibition ability of a peptide, one can alter either temperature and/or pH (in particular in the second case), so as to tune the probability that the native protein is in the native conformation. This is not the case in test on virus in cell cultures. In fact, if the target protein is very stable (*e.g.*, like the lysozyme and likely the haemagglutinin, both displaying a number of disulfide bonds), the only way the peptide can dock its complementary LES is just after (or during) the protein is expressed by the ribosome and before it folds. Fortunately, the HIV-1-PR is marginally stable under biological conditions. It can thus be targeted both before and after it has folded.

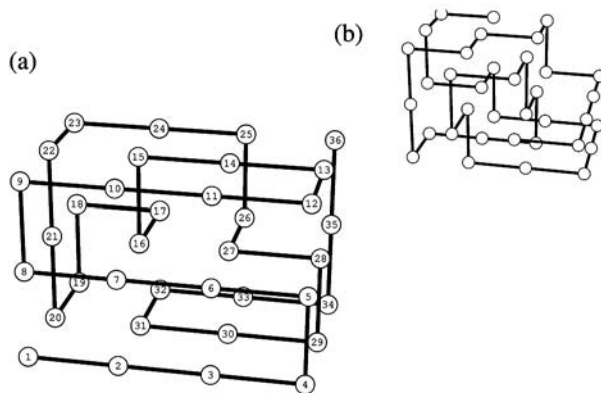


Fig. 18. – The two 36-mer native structures used in the calculations of the inhibitory properties of p-LES. The associated designed sequences [30] are (a) $S_{36} \equiv \text{SQKWLERGATRIADGDLPVNGTYFSCKIMENVHPLA}$, (b) $\text{RASMKDKTVGIGHQLYLNFEGEWCAPDNTRVSLAI}$.

with sequence (b) shown in the caption to fig. 18. This designed protein displays, in the folding process, three LES of length 2, 3 and 6, respectively. While the p-LES built of six residues (closed LES) inhibits folding as those described earlier, the other two p-LES do not (open LES). This is connected with the small size of these p-LES, which makes them quite unspecific and thus unable of molecular recognition. In fact, the probability that a p-LES binds to some part of the protein other than the target LES decreases exponentially with the number of residues involved. This result will prove of importance in the design of an inhibitor for the HIV-1-PR (see contribution by D. Provasi).

6. – Drug resistance

In trying to mimic the development of drug resistance of a viral protein one has repeated Monte Carlo simulations of the system formed by the target protein (S_{36}) and three p-LES 3-6, but this time introducing single-random point mutations in S_{36} , a situation typical of conventional drug resistance. Two types of results have been observed: (I) if the proteins is (upon mutation) still able to fold the scenario corresponding to figs. 16(a)-(c) still applies, (II) in those cases in which the p-LES does not bind to the protein, the corresponding sequence is not able to fold.

The situation is however quite different if one introduces a conspicuous number of simultaneous swappings such that the 36-mer moves from the original supercluster to a second supercluster (see figs. 10 and 12). In this case, the resulting sequence can fold as the original p-LES (3-6, 11-14 and 27-30) of S_{36} do not display essentially any folding inhibition ability (see fig. 19). Thus, we are in presence of a new type of drug resistance which we shall call non-conventional drug resistance, a situation which could be found in the case in which the target protein is associated with a virus displaying a high rate of reproduction and a low level of genetic accuracy, typical of enzymes associated with retroviruses

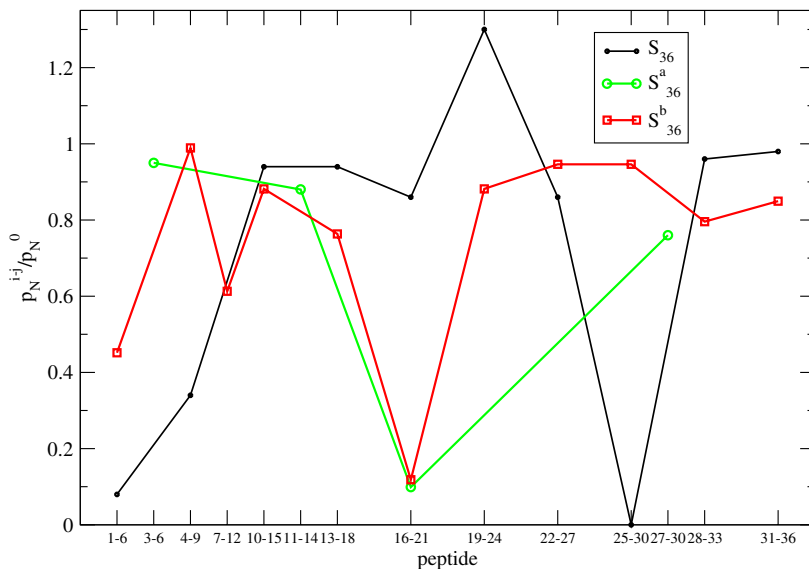


Fig. 19. – Ratio of the probability p_N^{i-j} that the protein is in the native state in the presence of three peptides starting at amino acid $\#i$ and ending at amino acid $\#j$, and of the probability that the protein by itself is in the native conformation. This ratio is instrumental for the search of the LES of the lattice model notional protein S_{36} , S_{36}^a and S_{36}^b (see fig. 12). For this search the sequence of peptides used display a sequence identical to the stretch of the target protein in question. On the other hand, if one were looking for a lead to an inhibitor which eventually could act equally well on a variety of mutants, one would then use a single set of peptides derived for example from the wild-type sequence (S_{36}). The population of the native state at $T = 0.24$ for a system composed of S_{36} (S_{36}^a , S_{36}^b) in the presence of three peptides displaying a sequence identical with that of the target protein (S_{36} continuous, S_{36}^a dotted, S_{36}^b dashed), normalized with respect to the population p_N^0 of the protein by itself ($= 0.78, 0.81, 0.67$ in the case of S_{36} , S_{36}^a and S_{36}^b , respectively).

(like, *e.g.*, the HIV-1-PR). This resistance could be, in principle, overcome making use of cocktail of p-LES representative of the different superclusters (two in the present case, corresponding *e.g.*, to peptide p(3-6) (sequence S_{36}) and peptide p(25-30) (sequence S_{36}^b))

7. – Design and folding of dimeric proteins

A large number of proteins perform their biological activity under the shape of dimers (or oligomers). A dimer is a protein whose native conformation is a globule build out of two disjoint chains. Depending whether the two chains have the same sequence or not, they are referred to as homodimers or heterodimers. Notwithstanding this difference, they have been observed to fold through three major mechanism. Some of the known dimers fold according to a three-state mechanism ($D \rightarrow I \rightarrow N$), where first the denaturated chains (D) assume the native conformations of each chains independent of each other (I: folding intermediate), and subsequently the two parts come together to

form the dimer (N: native). A different behaviour is displayed by other proteins, whose chains dimerize without populating any monomeric native-like intermediate (two-state process, $D \rightarrow N$). In this case, one can only identify the unfolded monomer (D) and the native dimers (N). A third mechanism, known as “domain-swapping”, implies the interaction of domains of, as a rule, large structures of any chain with similar structure of the other chain. Such a mechanism seems not to be of relevance within the subject of the present review.

It emerges from the above classification and from lattice model calculations [31], that the folding and thus the biological activity of three-state dimers, like for example that of the HIV-1-PR, can be inhibited following the same strategy as that discussed in connection with single-domain, globular protein (see the contribution of Provasi *et al.*).

8. – Conclusions

Within lattice model, the folding of proteins is a hierarchical process governed by LES, formed very early in the folding process and stabilized through local contacts among highly conserved amino acids. The docking of the LES gives rise to the (post) critical FN and shortly after, the native state is reached. Short peptides named p-LES displaying a sequence identical to that of LES are found to inhibit folding and thus biological activity. Furthermore, it is unlikely that point mutations can lead to escape mutants, in that the interaction between LES as well as between p-LES and LES is controlled by amino acids which play an essential role in the folding process. It is remarkable that these (model) results are highly transferable to real proteins, in particular to the inhibition of the HIV-1-PR [32,33] (see contribution of Provasi *et al.* and Rusconi *et al.* to this volume).

REFERENCES

- [1] FEHRST A., *Structure and Mechanism in Protein Science* (Garland Publishing Inc., New York) 1999.
- [2] CREIGHTON T., *Proteins. Structure and Molecular Properties* (W.H. Freeman & Co, New York) 1992.
- [3] BRANDEN C. and TOOZE J., *Introduction to Protein Structure* (Garland Publishing, Inc., New York) 1998.
- [4] MIYAZAWA S. and JERNIGAN R., *Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation*, *Macromolecules*, **18** (1985) 534.
- [5] SHAKHNOVICH E. I., *Theoretical studies of protein-folding thermodynamics and kinetics*, *Curr. Opin. Struct. Biol.*, **7** (1997) 29.
- [6] SHAKHNOVICH E. I., *Modeling protein folding: The beauty and power of simplicity*, *Folding & Design*, **1** (1996) R50.
- [7] GOLDSTEIN R., LUTHEY-SCHULTEN Z. and WOLYNES P., *Optimal folding codes from spin-glass theory*, *Proc. Natl. Acad. Sci. U.S.A.*, **89** (1992) 4918.
- [8] SALI A., SHAKHNOVICH E. I. and KARPLUS M., *Kinetics of protein folding: a lattice model study for the requirements for folding to the native state*, *J. Mol. Biol.*, **235** (1994) 1614.
- [9] BOWIE J., LUTHEY-SHULTEN R. and EISEMBERG D., *A method to identify protein sequences that fold into a known three-dimensional structure*, *Science*, **253** (1991) 164.

- [10] BROGLIA R. A., TIANA G., ROMAN H. E., VIGEZI E. and SHAKHNOVICH E. I., *Stability of designed proteins against mutations*, *Phys. Rev. Lett.*, **82** (1999) 4727.
- [11] VERDIER P. H., *Monte Carlo studies of lattice-model polymer chains. III. Relaxation of Rouse coordinates*, *J. Chem. Phys.*, **59** (1973) 6119.
- [12] HILHORST H. J. and DEUTCH J. M., *Analysis of Monte Carlo results on the kinetics of lattice polymer chains with excluded volume*, *J. Chem. Phys.*, **63** (1975) 5153.
- [13] SHAKHNOVICH E. I., FARZTDIMOV G., GUTIN A. M. and KARPLUS M., *Protein folding bottlenecks: A lattice Monte Carlo simulation*, *Phys. Rev. Lett.*, **67** (1991) 1665.
- [14] BROGLIA R. A., TIANA G., PASQUALI S., ROMAN H. E. and VIGEZI E., *Folding and Aggregation of Designed Protein Chains*, *Proc. Natl. Acad. Sci. U.S.A.*, **95** (1998) 12930.
- [15] BRENNER S. E., CHOTHIA C. and HUBBARD T. J., *Population statistics of protein structures: lessons from structural classifications.*, *Curr. Opin. Struct. Biol.*, **7** (1997) 369.
- [16] ROST. B., *Protein structures sustain evolutionary drift*, *Folding & Design*, **2** (1997) S19.
- [17] GOVINDARAJAN S. and GOLDSTEIN R., *Evolution of Model Proteins on a Foldability Landscape*, *Proteins*, **29** (1997) 461.
- [18] SHAKHNOVICH E. I. and GUTIN A., *Engineering of stable and fast-folding sequences of model proteins*, *Proc. Natl. Acad. Sci. U.S.A.*, **90** (1993) 7195.
- [19] SHAKHNOVICH E. I. and GUTIN A., *A new approach to the design of stable proteins*, *Prot. Eng.*, **6** (1993) 793.
- [20] SHAKHNOVICH E. I., *Proteins with selected sequences fold into unique native conformation*, *Phys. Rev. Lett.*, **72** (1994) 3907.
- [21] MIRNY L., ABKEVICH V. and SHAKHNOVICH E., *How evolution makes proteins fold quickly*, *Proc. Natl. Acad. Sci. U.S.A.*, **95** (1998) 4976.
- [22] GOVINDARAJAN S. and GOLDSTEIN R. A., *On the Thermodynamic Hypothesis of Protein Folding*, *Proc. Natl. Acad. Sci. U.S.A.*, **95** (1998) 5545.
- [23] BROGLIA R. A. and TIANA G., *Hierarchy of events in the folding of model protein*, *J. Chem. Phys.*, **114** (2001) 7267.
- [24] TIANA G., BROGLIA R. A., ROMAN H. E., VIGEZI E. and SHAKHNOVICH E. I., *Folding and misfolding of designed heteropolymeric chains with mutations*, *J. Chem. Phys.*, **108** (1998) 757.
- [25] SHAKHNOVICH E. I., ABKEVICH V. and PTITSYN O., *Conserved residues and the mechanism of protein folding*, *Nature*, **379** (1996) 96.
- [26] ABKEVICH V. I., GUTIN A. M. and SHAKHNOVICH E. I., *Specific nucleus as the transition-state for protein-folding – Evidence from the lattice model*, *Biochemistry*, **33** (1994) 10026.
- [27] BROGLIA R. A. and TIANA G., *Reading the three-dimensional structure of a protein from its amino acid sequence*, *Prot. Struct. Funct. Design.*, **45** (2001) 421.
- [28] ABKEVICH V. I., GUTIN A. M. and SHAKHNOVICH E. I., *Impact of local and nonlocal interactions on thermodynamics and kinetics of protein-folding*, *J. Mol. Biol.*, **252** (1998) 460.
- [29] TIANA G. and BROGLIA R. A., *Statistical analysis of native contact formation in the folding of design proteins*, *J. Chem. Phys.*, **114** (2001) 2503.
- [30] BROGLIA R. A., TIANA G. and BERERA R., *Resistance proof, folding-inhibitor drugs*, *J. Chem. Phys.*, **118** (2003) 4754.
- [31] TIANA G. and BROGLIA R. A., *Folding and design of dimeric proteins*, *Proteins*, **49** (2002) 82.
- [32] BROGLIA R. A., TIANA G., SUTTO L., PROVASI D. and SIMONA F., *Design of HIV-1-PR inhibitors which do not create resistance: blocking the folding of single monomers*, *Protein Sci.*, **14** (2005) 2668.
- [33] BROGLIA R. A., PROVASI D., VASILE F., OTTOLINA G., LONGHI R. and TIANA G., *A folding inhibitor of the HIV-1 Protease*, *Proteins*, **62** (2006) 928.

This page intentionally left blank

Exploring mechanisms of protein folding and binding in signal transduction networks

G. M. VERKHIVKER

*Department of Pharmacology, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093-0392, USA*

1. – Introduction

Proteins form complex interaction networks that determine much of the physiology and function of the cell. The architecture of protein-protein interaction networks [1-4] was recently proposed to be scale-free with most of the proteins having only few connections but with relatively fewer protein hubs establishing a considerably large number of links [5-14]. The presence of scale-free networks and protein hubs in biological systems has indicated that evolutionary processes may exhibit mechanisms of preferential attachment based on duplication and divergence of genes [15-22]. The organization of protein interaction networks may be imprinted in structural properties of the binding sites which enable network hubs to interact with diverse range of protein systems by incorporating conformational flexibility and even intrinsic disorder in one or both binding partners. Functional promiscuity is often linked with protein conformational diversity, which can be considered as evolvability traits that enable existing enzymes to rapidly develop new activities. When combined with the classical mechanisms of gene duplication mutation and selection and models of divergent evolution [23-29], conformational diversity provides a powerful mechanism to facilitate the evolution of new functions. There is a growing evidence from direct evolution experiments that proteins with promiscuous functions may have divergently evolved to acquire higher specificity and activity based on their ability to alter functions using sequence plasticity of a relatively small number of amino acid substitutions [23-29]. Numerous biochemical and genomic analysis have suggested that proteins may have developed the ability to improvise novel or alter existing functions

using plasticity of amino acids residing inside or near active sites [30-32]. Direct evolution studies have also indicated that substantial changes in the promiscuous functions of a protein need not come at the expense of its native function as evolution of a new function may be driven by mutations that have a minor effect on the native function but larger effects on the promiscuous functions [30-32].

Protein binding sites have evolved not only to be structurally flexible, but also functionally adaptive, capable of binding with high affinity to a diverse range of proteins and ligands which may be different from their natural binding partners in composition, size, and shape [33-35]. Alanine scanning mutagenesis combined with structural and thermodynamic studies, have shown that despite a typical large size of intermolecular interfaces, binding affinity and specificity may be determined by a relatively small fraction of the interface, forming energetically important hot spots at protein binding sites [36-43]. Evolution can often find convergent solutions to stable intermolecular interfaces and accommodate flexible and even intrinsically unstructured binding partners by using structural plasticity of the binding site and stabilizing contributions of the hot spot residues [44]. The intrinsically unstructured [45-47], intrinsically disordered [48-51] or natively unfolded [52-54] are frequently involved in regulatory functions in the cell where structural disorder can be relieved upon binding of the protein to its target molecule. The intrinsic lack of structure and functional conformational transitions occurring upon binding may offer important functional advantages, including the ability of binding to several different targets [45,46,48-51], high specificity coupled with low affinity [48,55], the precise control and simple regulation of the binding thermodynamics [45,46,48-51,56] and the increased rates of specific macromolecular association [48-51].

The regulation of the cyclin-dependent kinase (Cdk) activity is essential for the proper timing and coordination of numerous nuclear processes [57,58]. Cdks are regulated by binding to their cyclin protein partners, which play a key role in the cell division cycle through their ability to activate and enhance the substrate affinity of their associated Cdks and forming active heterodimeric complexes [59-61]. Structural studies have provided a detailed view of the activation mechanism of Cdk2 by cyclin A and the mode of recognition of substrates [62-66]. The activity of Cdks throughout the cell cycle is directed by a combination of several mechanisms, including coordination of Cdk phosphorylation and dephosphorylation, which is required for controlled activation and deactivation of Cdks, and variations in the levels of the Cdk inhibitor proteins, CKIs, responsible for the deactivation of the Cdk cyclin complexes [67-73]. The inhibitors of the Kip/Cip family (CKI) $p21^{Waf1}$ and $p27^{Kip1}$ regulate cell proliferation through binding and inactivating Cdk2 and Cdk4 cyclin binary complexes and are highly dynamic polypeptides that exhibit limited, transient secondary structure in solution, and lack globular tertiary structure [67-70]. The $p21^{Waf1}$ and $p27^{Kip1}$ proteins can inhibit Cdkc-cyclin activity by blocking formation of active Cdk-cyclin complexes via binding to inactive Cdk or by binding to the active complex. The intrinsically disordered nature of $p21^{Waf1}$ and $p27^{Kip1}$ may be associated with binding diversity and the ability to bind and inhibit a diverse family of Cdk-cyclin assemblies, including Cdk2-cyclin-A, Cdk2-cyclin-E, and Cdk4-cyclin-D complexes [67-70].

Coupling of folding and binding accompanied by a disorder-order transition has been experimentally detected for $p27^{Kip1}$ [67-70]. The crystal structure of the 69-amino acid N-terminal inhibitory domain of $p27^{Kip1}$ bound to the phosphorylated cyclin-A-Cdk2 complex at 2.3 Å resolution [68] has revealed an ordered conformation of $p27^{Kip1}$ protein comprising of a rigid coil, α -helix, β -hairpin, β -strand, and 3_{10} -helix (fig. 1). Instead of contributing directly to folding of $p27^{Kip1}$, the hydrophobic residues of coil, β -hairpin, and β -strand form intermolecular contacts in the tertiary complex, that are further consolidated with six backbone-backbone hydrogen bonds formed by the β -strand and stabilizing interactions of the 3_{10} -helix in the catalytic cleft of Cdk2 (fig. 1). The crystal structure of the tertiary complex has revealed that the 3_{10} -helix region of $p27^{Kip1}$ is deeply inserted within the catalytic cleft of Cdk2, causing significant conformational changes in the Cdk2 portion of the interface and inhibiting kinase activity [68]. While the presence of a conserved 3_{10} -helix region among the inhibitors of the Kip/Cip family raised the possibility that all members of this family inhibit the cyclin-A-Cdk2 complex activity in the same manner, it was discovered that mutations within the 3_{10} -helix of $p21^{Kip1/Cip1/Waf1}$ (L76A/Y77A) and $p27^{Kip1}$ (F87A/Y88A/Y89A) did not affect the inhibitory activity toward cyclin-A-Cdk2 and cyclin-E-Cdk2. Consequently, the 3_{10} -helix motif may be specifically utilized in selectively inhibiting the Cdk2 activity primarily by the $p57^{Kip2}$, but not by the $p27^{Kip1}$ and $p21^{Kip1/Cip1/Waf1}$ members of the Kip/Cip family [71, 72].

Circular-dichroism spectroscopy experiments have shown that the unbound form of the $p27^{Kip1}$ inhibitory domain is intrinsically disordered, but not entirely unfolded and contains some marginally stable helical structure in the α -helix region [70]. Reducing or increasing the stability of the α -helix in the unbound $p27^{Kip1}$, respectively, with proline or alanine substitutions has an only marginal effect on the binding thermodynamics. However, while the disruption of the α -helix in the unbound form by proline mutations does not affect the rate of $p27^{Kip1}$ binding, a marginal stabilization of the α -helix upon alanine substitutions in this region slows down the rate of formation for the inhibited complex by approximately 3 times. Consequently, it has been suggested that the intrinsic disorder of the α -helix in the unbound form of $p27^{Kip1}$ can result in a kinetic advantage during $p27^{Kip1}$ folding transitions coupled to the binding process [70]. Even in a crowded, thermodynamically nonideal environment, where intrinsically disordered proteins may adopt a stable structure, the $p27^{Kip1}$ protein does not experience any significant conformational change as detected by circular-dichroism and fluorescence spectra [73]. Hence, molecular crowding effects are not necessarily sufficient to induce ordered structure in the intrinsically disordered $p27^{Kip1}$ protein. According to the originally proposed $p27^{Kip1}$ binding mechanism [68-70], substrate recruitment to Cdk2 is initially carried out by the hydrophobic docking site on cyclin A that forms the local contacts with the RNLFG moiety of the $p27^{Kip1}$ coil, serving as an initial anchor in the tertiary complex formation. Importantly, these interactions do not require structural rearrangements in cyclin A that provides the rigid portion of the intermolecular interface. The importance of the LFG motif is manifested in the crystal structure of $p27^{Kip1}$ complexed with cyclin-A-Cdk2 in which the N-terminal half of the 68-residue $p27^{Kip1}$ fragment containing the RNLFG moiety associates with cyclin-A-Cdk2 through binding

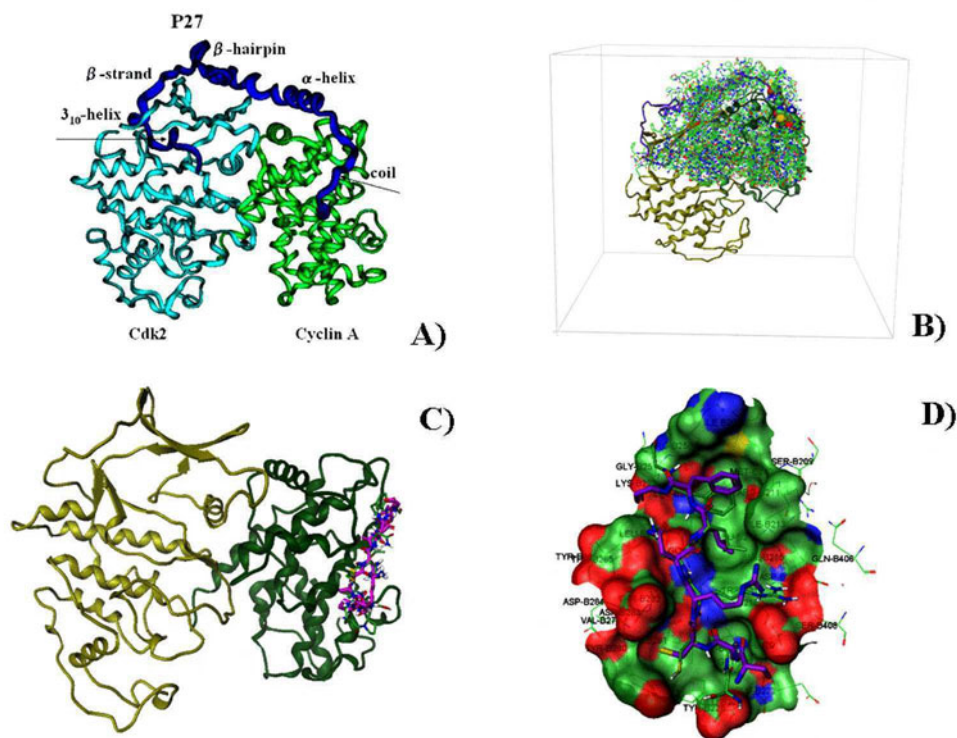


Fig. 1. – (A) The crystal structure of the tertiary complex with the bound 69-amino acid $p27$ protein (residues 25-93) comprising sequentially the coil (residues from 25 to 34), α -helix (residues from 35 to 60), β -hairpin (residues from 61 to 71), β -strand (residues 75 to 81), and 3_{10} -helix (residues from 85 to 90). Schematic drawing shows $p27$ in blue, cyclin A in green, and Cdk2 in light blue. (B) The crystal structure of the tertiary complex with the bound $p27$ protein in the search box that encompasses the $p27^{Kip1}$ tertiary complex with a large 10.0 Å cushion added to each side which guarantees a sufficiently unbiased conformational search. The collection of random starting conformations of the $p27^{Kip1}$ coil used in simulated tempering dynamics is presented for illustration. (C) Convergent solutions for the low-energy conformations of the $p27^{Kip1}$ coil superimposed with the crystallographic coil conformation in the docking recognition site of cyclin A (shown in blue). (D) A close-up of the hot spot in the docking recognition site of cyclin A: the residues lining the recognition groove form several subsites, including the primary hydrophobic pocket: Met-210, Ile-213, Leu-214, Trp-217, Leu-253; the arginine site and secondary lipophilic pocket formed by Val-221, Ile-281. In addition, binding is facilitated by a number of hydrogen bond interactions with the peptide backbone, including the side-chain of Gln-254. The predicted conformations of the $p27^{Kip1}$ coil (shown in blue) superimposed with the crystallographic coil conformation in the docking recognition site of cyclin A (shown in atom-based default colors).

to a shallow hydrophobic groove on the surface of cyclin A [70].

Structural and kinetic experiments have demonstrated that $p27^{Kip1}$ can interact with cyclin-A-Cdk2 through a mechanism involving binding-coupled, protein folding (folding-on-binding) consisting of two distinct steps [74-77]. In the first step of the kinetic mechanism a highly dynamic coil domain of $p27^{Kip1}$ interacts with the cyclin A subunit of the complex, followed by a much slower binding to Cdk2 domain. Rapid binding of $p27^{Kip1}$ coil to cyclin A tethers the inhibitor to the binary cyclin-A-Cdk2 complex, which may reduce the entropy barrier associated with slow binding to the catalytic subunit. These results combined with sequence and structural analysis of multiple Cdks and cyclins have suggested that the binding specificity of the CKIs, including $p21^{Waf1}$ and $p27^{Kip1}$ proteins, may arise from the sequential folding-upon-binding mechanism of CKI's, determined by initial selective recognition by the cyclin regulatory subunits [74-77]. The concept of thermodynamic tethering implies that assemblies formed through the binding of unstructured proteins to pre-existing protein assemblies lead to complexes with exceptionally high thermodynamic stability, a mechanism that has been observed for $p27^{Kip1}$ binding to the cyclin-A-Cdk2 complex [74-77]. The discovered effect of thermodynamic tethering of $p27^{Kip1}$ onto cyclin-A-Cdk2 complexes has been proposed as a general phenomenon associated with unstructured proteins, which can function by binding independently to multiple subunits within multi-protein assemblies [77].

Structural basis for specificity of substrates and recruitment proteins by the cyclin-A-Cdk2 complex, determined by a hydrophobic binding site on the cyclin-A subunit, is mediated through sequences which share a consensus RXL motif found in binding sequences in a variety of cyclin-A-Cdk2 partner proteins, including $p21^{Waf1}$, $p27^{Kip1}$, and $p57^{Kip2}$ inhibitors of the Kip1/Cip1 family [78,79]. A number of cyclin-A-Cdk2 complexes crystal structures with recruitment proteins have been recently reported, including a peptide with the sequence RRLFGEDPPKE from $p107$, the peptides STSRHKKLMFK from $p53$, KPSACRNLFGP from $p27^{Kip1}$, PVKRRLDLE from E2F, and AGSAKRRLFGE from $p107$ [80]. The crystallographic studies combined with isothermal calorimetry measurements on cyclin-A-Cdk2 complexes with these peptides have shown that the cyclin recruitment site can recognize diverse, conformationally constrained target sequences and can favorably accommodate a second hydrophobic residue C-terminal or next adjacent to the Leu residue of the "RXL" motif making important contributions to the recruitment peptide recognition [80,81]. The structure-activity analysis for the series of $p21^{Kip1/Cip1/Waf1}$ -derived peptides [82-85] has indicated that these peptides bind with the cyclin-A-Cdk2 complex via a common mechanism occurring through the initial recruitment by the binding groove on the cyclin subunit, whereas uncomplexed Cdks could not be recognized and blocking of this recruitment site prevents recognition and subsequent phosphorylation of Cdk substrates.

The emerging realization that protein folding and molecular recognition phenomena share a number of universal aspects, including the complex nature of interactions on the underlying energy landscape [86-90] is manifested by a balance between opposing thermodynamic forces during the process, the loss of conformational entropy and the energy gain upon the native structure formation, that ultimately determines the ther-

modynamic free-energy barrier of the reaction. These physical arguments, that have been fruitful in understanding protein folding mechanisms, have also led to an elegant, fly-casting mechanism derived from the energy landscape theory and proposed to explain kinetic advantages of unstructured proteins in binding [86, 87]. A microscopic study of the $p27^{Kip1}$ conformational dynamics and the energy landscape analysis of protein folding transitions coupled to binding was recently undertaken by simulating hierarchy of $p27^{Kip1}$ structural loss in the course of multiple and independent temperature-induced Monte Carlo simulations from the crystal structure of the $p27^{Kip1}$ -phosphorylated cyclin-A-Cdk2 tertiary complex [91-93]. We have characterized the transition state ensemble (TSE) of the $p27^{Kip1}$ folding reaction coupled to binding as an expanded form of the native bound structure with the native-like topology established in the β -hairpin and the β -strand regions of the intermolecular interface, and a considerable flexibility in the coil, α -helix, and 3_{10} -helix regions. These structural elements of $p27^{Kip1}$, that are largely formed in the TSE, are also consistently observed to disintegrate last in the ensemble of unfolding-unbinding simulations. Furthermore, the ensemble of so-called “post-critical” conformations, that appear immediately after the TS and lead to a rapid descent to the native structure is characterized by a considerable structural consolidation of the β -hairpin and the β -strand intermolecular contacts, in particular, by strengthening in six backbone-backbone hydrogen bonds formed by the β -strand in the tertiary complex. Simulations of $p27^{Kip1}$ unfolding-unbinding at high temperatures and subsequent determination of the TSE have shown that the topological requirements of the native intermolecular interface to order the β -hairpin and the β -strand structural elements are critical for nucleating a rapid kinetic transition to the native complex and ultimately dictate the $p27^{Kip1}$ folding mechanism [91-93].

In this work, we analyze role of major structural elements of $p27^{Kip1}$ in determining the binding mechanism with Cdk2-cyclin-A binary complex. We determine a structural basis of specificity for recruitment proteins by the cyclin-A-Cdk2 complex which may be primarily determined by the hydrophobic docking site on the cyclin A and unfrustrated binding with the consensus RNLFG motif of the $p27^{Kip1}$ coil during the initiation binding event. We show that the proposed atomic picture of the $p27^{Kip1}$ molecular recognition with the Cdk2-cyclin-A complexes is consistent with the nucleation-condensation mechanism of protein folding [94]. The results of this study provide important insights in understanding molecular recognition in protein interaction networks which may have evolved common mechanisms to ensure functional adaptability in binding with conformationally diverse binding partners.

2. – Simulation methods

2.1. Monte Carlo simulations: the energy model. – We have analyzed the hierarchy of the $p27^{Kip1}$ structural loss, which is monitored by using temperature-induced Monte Carlo unfolding-unbinding simulations initiated from the crystal structure of the tertiary complexes at a range of temperatures. The molecular recognition energetic model used in this study includes intramolecular energy terms, given by torsional and non-

bonded contributions of the DREIDING force field [95], and the intermolecular energy contributions calculated using the AMBER force field [96] to describe protein-protein interactions combined with an implicit solvation model [97]. The dispersion-repulsion and electrostatic terms have been modified and include a soft core component that was originally developed in free energy simulations to remove the singularity in the potentials and improve numerical stability of the simulations [98]. The protein-protein interaction energy between atom i of the ligand and atom j of the protein from the binary complex is written as

$$(1) \quad E_{ij}^{inter} = E_{ij}^{dr} + E_{ij}^{es} + E_{ij}^{sol} ,$$

where E_{ij}^{dr} is the dispersion-repulsion interaction energy, E_{ij}^{es} is the electrostatic interaction energy, and E_{ij}^{sol} is the change in solvation energy of atom i in the presence of atom j plus the change in solvation energy of atom j in the presence of atom i . In this study, the formula for calculating the dispersion-repulsion interaction between the two atoms is given by the soft-core form

$$(2) \quad E_{ij}^{dr} = \frac{A_{ij}}{(r_{ij}^6 + \delta_{LJ}^6)^2} - \frac{B_{ij}}{(r_{ij}^6 + \delta_{LJ}^6)} ,$$

where r_{ij} is the distance between atoms i , A_{ij} and B_{ij} are calculated using geometric combination rules:

$$(3) \quad A_{ij} = 2^{12} (\epsilon_i \epsilon_j)^{1/2} (r_i^* r_j^*)^6 ,$$

$$(4) \quad B_{ij} = 2^7 (\epsilon_i \epsilon_j)^{1/2} (r_i^* r_j^*)^3 ,$$

and ϵ and r^* are parameters of the original AMBER force field. Note that the original AMBER force field uses a different set of combination rules. The quantity, δ_{LJ} , is the soft-core value for the dispersion-repulsion potential. Currently we use $\delta_{LJ} = 2.75 \text{ \AA}$. The formula used for the Columbic interaction between two atoms with partial charges q_i and q_j is

$$(5) \quad E_{ij}^{es} = \frac{q_i q_j}{4\pi\epsilon_0 D (r_{ij}^6 + \delta_{ES}^6)^{1/3}} ,$$

where the soft-core value for the electrostatics is $\delta_{ES} = 1.75 \text{ \AA}$ and the relative dielectric constant is $D = 2.0$. A solvation term was added to the interaction potential to account for the free energy of interactions between the explicitly modelled atoms of the protein-protein system and the implicitly modelled solvent. The term was derived by considering the transfer of atom, from an environment where it is completely surrounded by solvent, to an environment in which it has explicit atomic neighbors. The solvent-accessible surface is created by rolling a spherical water probe with a radius of $1.4 \text{ \AA}$ over the Van der Waals surface of the molecule. The center of the probe traces the solvent-accessible surface.

2'2. Monte Carlo simulations: simulated tempering dynamics. – We have carried out equilibrium simulations with the ensembles of protein kinase conformations using parallel simulated tempering dynamics [99] with 100 replicas of the ligand-protein system attributed respectively to 100 different temperature levels that are uniformly distributed in the range between 5300 K and 300 K. The Cdk2-cyclin-A portion of the tertiary complex is held fixed in its minimized crystallographic conformation, while the rigid-body degrees of freedom and rotatable angles of the $p27^{Kip1}$ protein are treated as independent variables during simulations. $p27^{Kip1}$ conformations and orientations are sampled in a parallelepiped that encompasses the crystallographic structure of the $p27^{Kip1}$ tertiary complex with a large 10.0 Å cushion added to every side of this box surrounding the interface which guarantees a sufficiently unbiased conformational search. Monte Carlo moves are performed simultaneously and independently for each replica at the corresponding temperature level. After each simulation cycle, that is completed for all replicas, exchange of configurations for every pair of adjacent replicas at neighboring temperatures is introduced. The m -th and n -th replicas, described by a common Hamiltonian $H(X_1, \dots, X_m, \dots, X_n, \dots, X_N)$, are associated with the inverse temperatures β_m and β_n , and the corresponding conformations X_m and X_n . The exchange of conformations between adjacent replicas m and n is accepted or rejected according to Metropolis criterion with the probability

$$(6) \quad p = \min(1, \exp[-\delta]),$$

$$(7) \quad \delta = [\beta_n - \beta_m][H(X_m) - H(X_n)].$$

2'3. Binding free energy calculations. – Binding free energies are computed using the molecular mechanics AMBER force field and the solvation energy term based on continuum generalized Born and solvent accessible surface area (GB/SA) solvation model [100]. The binding free energy of the ligand-protein complex can be written as follows:

$$(8) \quad G_{bind} = G_{complex} - G_{protein} - G_{ligand},$$

where the average total free energy of the molecule G is evaluated as follows:

$$(9) \quad G_{molecule} = G_{solvation} + E_{MM} - TS_{solute},$$

$$(10) \quad G_{solvation} = G_{cavity} + G_{vdw} + G_{pol}.$$

In the GB/SA model, the G_{cavity} and G_{vdw} contributions are combined together via evaluating solvent-accessible surface areas:

$$(11) \quad G_{SA} = G_{cavity} + G_{vdw} = \sum_i \sigma_i SA_i,$$

G_{SA} is the nonpolar solvation term derived from the solvent-accessible surface area (SA).

$$(12) \quad G_{pol} = -166.0 \left(1 - \frac{1}{\epsilon}\right) \sum_i \sum_j \frac{q_i q_j}{(r_{ij}^2 + \alpha_{ij}^2 \exp(-D_{ij}))^{0.5}}.$$

G_{pol} is the polar solvation energy which is computed using the GB/SA solvation model. S_{solute} is the vibrational entropy of the molecule. E_{MM} is the molecular mechanical energy of the molecule summing up the electrostatic E_{es} interactions, van der Waals contributions E_{vdw} , and the internal strain energy E_{int} . The structures of ligand-protein complexes are subjected to conjugate gradient minimization as implemented in the version 7.0 of the MacroModel molecular modelling software package [100]. The ensembles of structures for the uncomplexed protein and ligand are generated by using all ligand-protein conformations. Separation into the protein and ligand structures is followed by minimization of the complex, the unbound protein and the unbound ligand conformation. A residue-based cutoff of 8 Å is set for computing nonbonded van der Waals interactions and 20 Å residue-based cutoff is used for computing electrostatic interactions. The solute entropy contribution is not included in the MM/GBSA model.

3. – Results and discussion

The most stable elements of $p27^{Kip1}$ during unfolding and unbinding dynamics at temperatures $T = 300\text{ K}–380\text{ K}$ are the β -hairpin and the β -strand, while the coil portion, α -helix and 3_{10} -helix fluctuate in a considerably broader range. As temperature increases to $T = 400\text{ K}$, temperature-induced motions produce coupled unfolding and unbinding on the simulation time scale, evident from dramatic and progressively increasing root mean square deviation (RMSD) from the native structure (figs. 2, 3). Despite considerable differences among individual trajectories, the analysis of multiple independent simulations at high temperatures results in a systematic structural polarization in the hierarchy of structural loss during $p27^{Kip1}$ unfolding and unbinding. The most persistent interactions of $p27^{Kip1}$ at the intermolecular interface are formed by the β -hairpin and the β -strand elements that maintain their structural integrity considerably longer during unbinding/unfolding than other $p27^{Kip1}$ elements (fig. 2). The shape of the distribution for the $p27^{Kip1}$, and its β -hairpin and β -strand components deviates considerably from the concerted picture. We have determined that as the number of intermolecular hydrogen bonds progressively decreases and $p27^{Kip1}$ gradually dissociates from the bound complex, these structural elements continue to maintain their native bound conformation intact until the late stage of the process. Despite large fluctuations of the $p27^{Kip1}$ coil, there is a visible lack of symmetry in the hierarchy of its structural loss, reflected in the occupancy distribution which is shifted towards moderate RMSD from the native bound structure and accompanied by an appreciable average number of the intermolecular hydrogen bonds (fig. 3). We have found a predominant tendency in the hierarchy of structural loss that is preserved on average in the ensemble of multiple

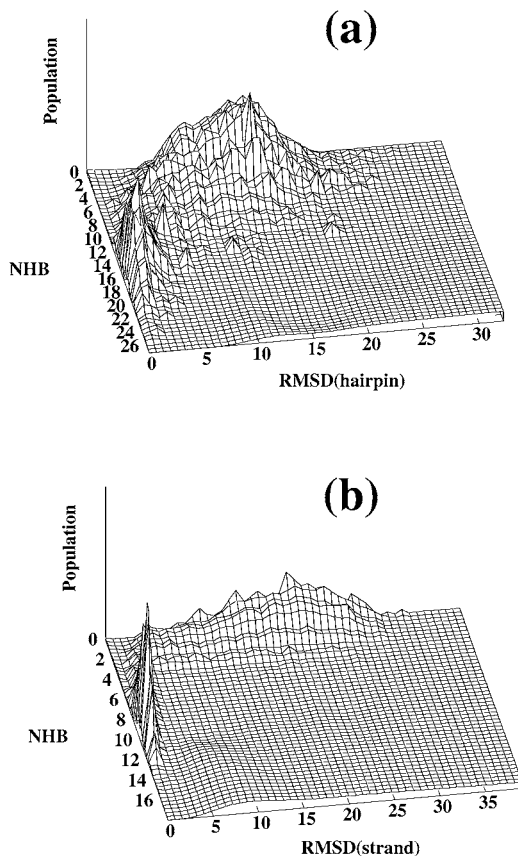


Fig. 2. – Three-dimensional population histograms generated from 1000 independent unfolding/unbinding simulations at $T = 400$ K initiated from the crystal structure of the complex to monitor hierarchy of structural loss for $p27^{Kip1}$ structural elements: β -hairpin (a), and β -strand (b). The histograms are built as a function of the total number of the native intermolecular hydrogen bonds (NHB) formed by the entire $p27^{Kip1}$ in the complex and the RMSD from the crystal structure conformation for each of the individual $p27^{Kip1}$ segments.

temperature-induced trajectories. According to this trend, the ensemble of largely unbound and unfolded $p27^{Kip1}$ conformations can be characterized by the entire loss of all intermolecular contacts with the Cdk2 portion of the interface, whereas these conformations tend to fluctuate along cyclin A frequently encountering weak interactions with the recognition groove.

To characterize the ensemble of largely unstructured $p27^{Kip1}$ conformations, which dominate at the end of temperature-induced unfolding/unbinding simulations, we performed a more detailed statistical analysis of the ensemble of unfolded $p27^{Kip1}$ conformations by initiating independent simulations from 1000 of these structures at lower temperatures. Strikingly, these largely unstructured $p27^{Kip1}$ conformations are char-

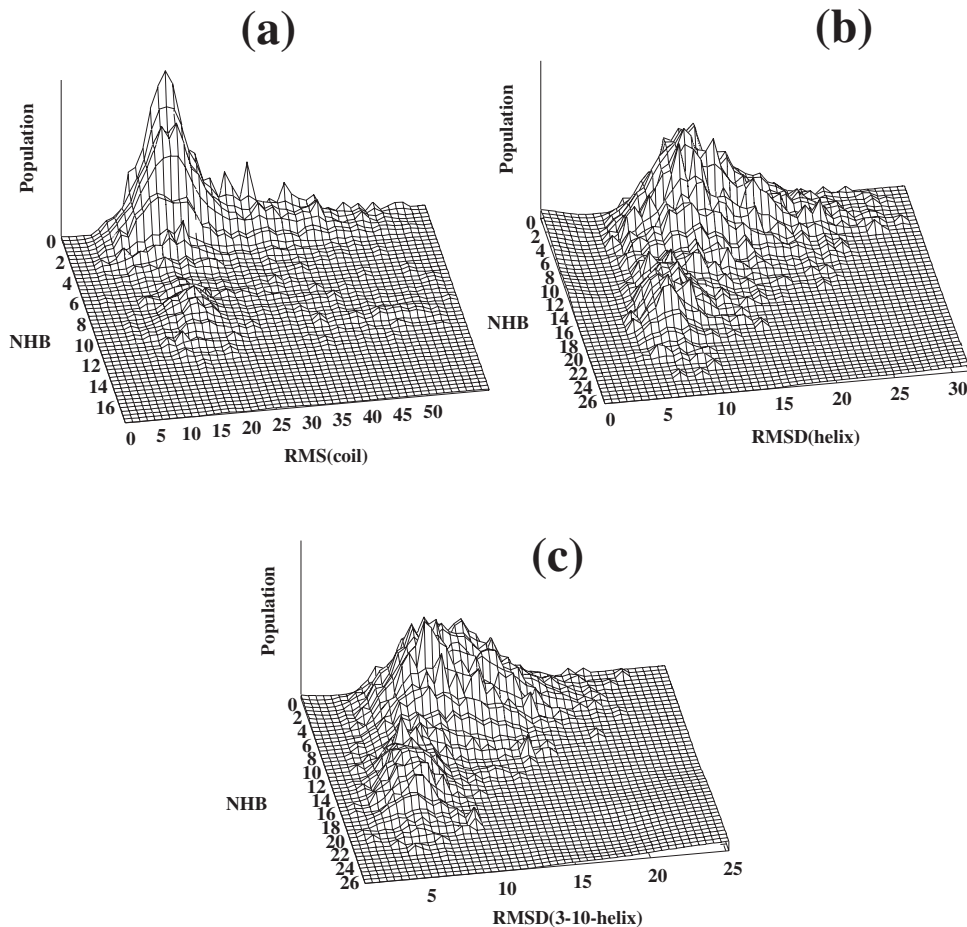


Fig. 3. – Three-dimensional population histograms generated from 1000 independent unfolding/unbinding simulations at $T = 400$ K initiated from the crystal structure of the complex to monitor hierarchy of structural loss for $p27^{Kip1}$ structural elements: coil (a), α -helix (b), and 3_{10} -helix (c). The histograms are built as a function of the total number of the native intermolecular hydrogen bonds (NHB) formed by the entire $p27^{Kip1}$ in the complex and the RMSD from the crystal structure conformation for each of the individual $p27^{Kip1}$ segments.

acterized by transitions between mostly unbound, collapsed conformations and weakly bound $p27^{Kip1}$ conformations, fluctuating in a non-specific manner along cyclin A side of the interface. Analysis of the recurrent features observed at late stages of coupled unfolding and unbinding yields some interesting insights into the initial stages of the $p27^{Kip1}$ binding mechanism. To provide insight into thermodynamics and dynamics of the $p27^{Kip1}$ coil intermolecular contacts, we have compared RMSD probability distributions for the entire coil segment of $p27^{Kip1}$ and individual residues (Leu32, Phe33, and

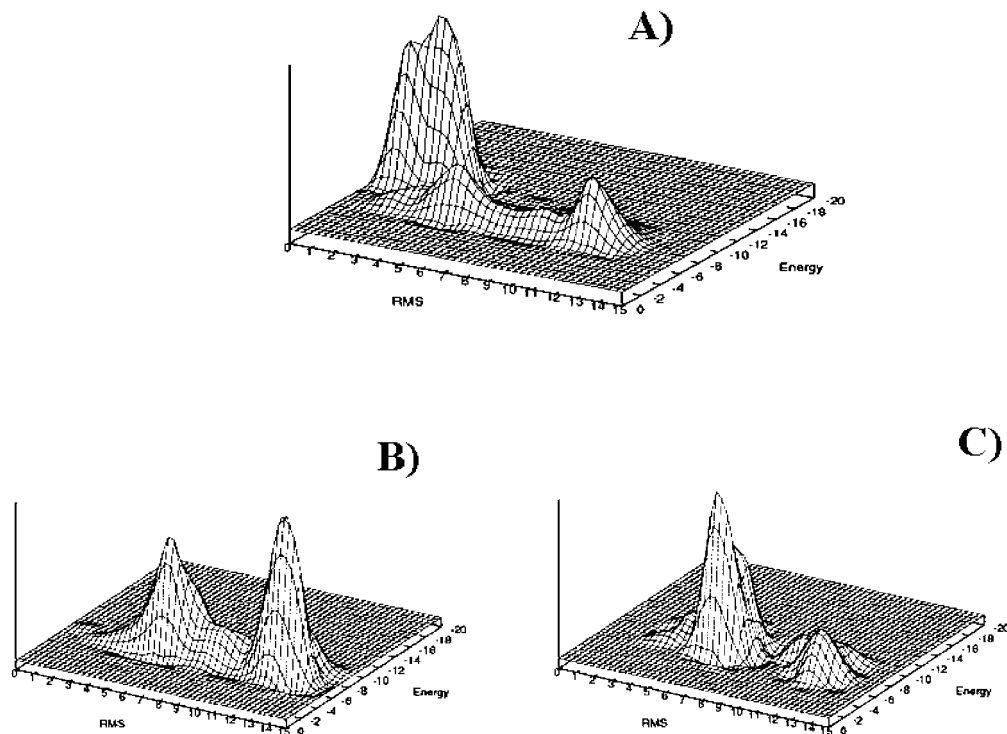


Fig. 4. – Three-dimensional histograms generated from 1000 independent structure predictions simulated tempering simulations of the $p27^{Kip1}$ coil (A) β -strand (B) and 3_{10} -helix (C) by considering these $p27^{Kip1}$ components as fully flexible individual peptides. The histograms are built as a function of binding free energy and the RMSD from the crystal structure conformation for each of the individual $p27^{Kip1}$ segments.

Gly34) from the RNLFQ motif which are believed to be critical in anchoring the $p27^{Kip1}$ protein to cyclin A. As a result, with a overall highly flexible $p27^{Kip1}$ coil, the residues from the RNLFQ motif have a significant population of native-like states. Hence, the formation of entropically favorable and weakly bound $p27^{Kip1}$ conformations fluctuating along the cyclin A portion of the interface may be beneficial at the initial stage of the reaction. This is consistent with the experimental evidence pointing to this region of the intermolecular interface as a potential initiation docking site during binding reaction.

To analyze the role of the major structural elements of $p27^{Kip1}$ in determining the recognition mechanism of binding with the cyclin-A-Cdk2 binary complex, we have also conducted simulated tempering structure predictions of the SACRNLFQ fragment of the $p27^{Kip1}$ coil, β -hairpin, β -strand and 3_{10} -helix by considering these $p27^{Kip1}$ components as fully flexible individual peptides, each with a large number of independent degrees of freedom, ranging from 30 to over 40 rotatable bonds (fig. 1, 4). According to the energy landscape theory of biomolecular binding, topological features of the native complexes

that are critical for robust structure prediction are often driven by consistent ordering of the recognition ligand motif in its native conformation. Strikingly, the results of simulations demonstrate that despite a larger number of rotatable degrees of freedom and the disordered nature of the initial starting conformations of the $p27^{Kip1}$ coil, distributed in a completely unbiased fashion over an enormous amount of conformational space (fig. 1A), an appreciable level of structural consistency in multiple simulated tempering runs can be detected (fig. 1B, 4A). The results of simulations reveal a remarkable structural similarity between the dominant binding mode of the RNLFG moiety, which associates with cyclin-A-Cdk2 through binding to a shallow hydrophobic groove on the surface of the cyclin A, and its native conformation in the crystal structure of the $p27^{Kip1}$ complex with cyclin-A-Cdk2 (fig. 1C, 4A). Hence, the RNLFG peptide can present a recognition anchor of the $p27^{Kip1}$ protein by exhibiting structural consistency with the crystal structure of the entire $p27^{Kip1}$ complex, a property that we termed structural harmony [91-93]. We find that for the $p27^{Kip1}$ coil interactions with the recognition groove on cyclin A represent unfrustrated binding, when the native interactions of the $p27^{Kip1}$ coil are consistent with the $p27^{Kip1}$ interactions. This results in gradual energy decrease as the native interactions are progressively formed and a dominant, conformational funnel leading to the native structure emerges. Furthermore, the energetics of the SACRNLFG peptide coil interactions with the recognition groove of cyclin A is also more favorable than that of 3_{10} -helix and β -strand peptide motifs (fig. 4). In contrast, the interactions of other structural elements of $p27^{Kip1}$ are considerably more frustrated, particularly β -strand (fig. 4B) and 3_{10} -helix motifs (fig. 4C) cannot recognize uniquely either Cdk2 or cyclin-A proteins in the binary complex and as such are likely incapable of providing initial recognition during the binding reaction.

The hot spot of cyclin A is comprised of the residues lining the recognition groove forming several subsites, including the primary hydrophobic pocket: Met-210, Ile-213, Leu-214, Trp-217, Leu-253; the arginine site and secondary lipophilic pocket formed by Val-221, Ile-281. In addition, binding is facilitated by a number of hydrogen bond interactions with the peptide backbone. Interestingly, the relatively low hydrogen bonding ability in the binding region in the hot spots of cyclin A may indicate that protein hubs would place fewer geometric constraints in the recognition hot spots to facilitate diversity of the binding partners, as fewer and less complex complementary polar interactions would be necessary for initial binding. Interestingly, the topography of the cyclin A recognition groove is similar in sequence composition and structural topography of the recognition hot spot in the hinge region of the constant fragment (Fc) of human immunoglobulin G (Ig) protein, which represents the consensus binding site for natural proteins and synthetic peptides [44]. A considerable similarity between structural topography of the protein hubs and hot spots present in different protein interaction networks may suggest that molecular recognition of the protein hubs hot spots may have evolved to ensure rather common mechanisms of functional adaptability in binding with conformationally diverse binding partners.

To understand the effect of structural stability of different $p27^{Kip1}$ secondary elements on the molecular recognition mechanism, we investigate the effect of forming specific in-

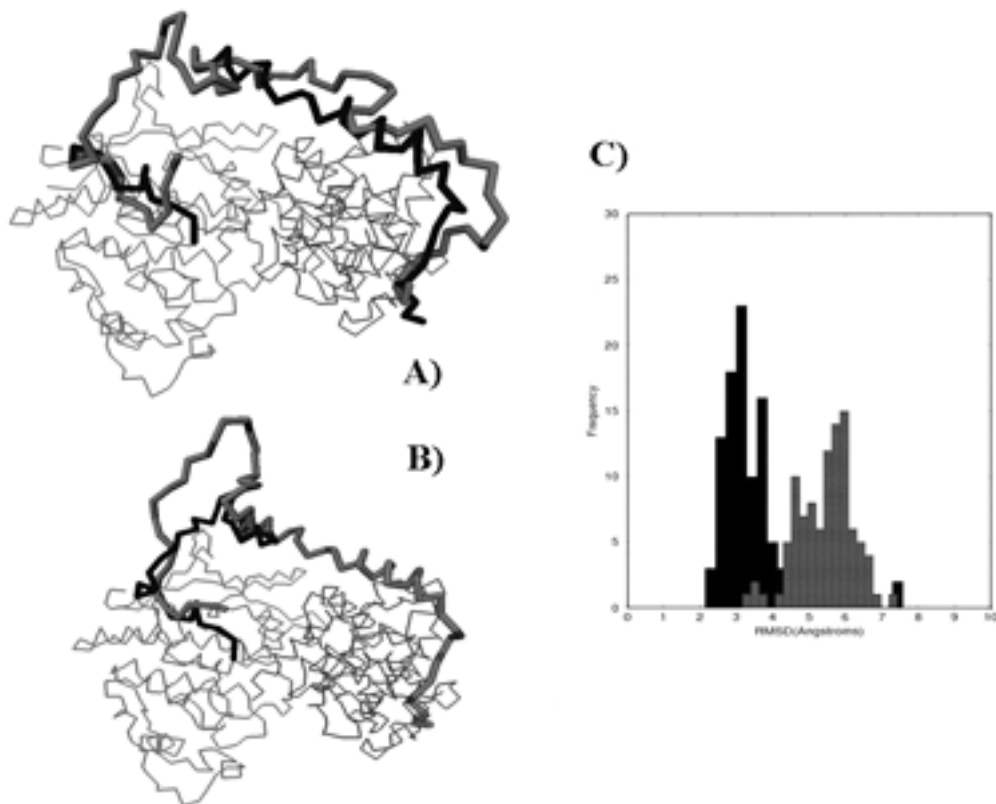


Fig. 5. – Superposition of the crystal structure of $p27^{Kip1}$ in the tertiary complex with typical final $p27^{Kip1}$ protein conformations at the end of a simulated annealing run when both the β -hairpin and the β -strand elements are constrained in their native bound conformations (A) and when both the coil and α -helix elements are constrained in their native bound conformations (B). (C) RMSD distribution frequency from the results of independent simulated annealing runs: when the β -hairpin and the β -strand are kept in their native bound conformations (black filled graph); the coil and α -helix are kept in their native bound conformations (grey shadow graph).

termolecular interactions on the structure prediction of the $p27^{Kip1}$ tertiary complex. By constraining different secondary structure elements of $p27^{Kip1}$ in their native bound conformations and conducting multiple simulated tempering simulations, we analyze differences in the success rate of predicting the native structure of $p27^{Kip1}$ in the tertiary complex (fig. 5). The important role of the $p27^{Kip1}$ coil interactions with cyclin A in the initial recognition event and binding thermodynamics may suggest that further consolidation of the adjacent $p27^{Kip1}$ helix in its native bound conformation would enhance the structural stability and the intermolecular interface of the $p27^{Kip1}$ complex with cyclin-A-Cdk2 binary complex. However, when both the $p27^{Kip1}$ coil and the α -helix elements

are held fixed in their native bound conformations during simulated tempering runs, the convergence to the native bound structure of $p27^{Kip1}$ could not be achieved on the same simulation time scale, even though the total number of degrees of freedom is considerably reduced from 169 rotatable bonds for the fully flexible $p27^{Kip1}$ to 82 rotatable bonds in these partially constrained structure prediction simulations (fig. 5A). In contrast, when both the β -hairpin and the β -strand elements are constrained in their native bound conformations, a better convergence to the native $p27^{Kip1}$ bound structure is observed in the course of simulated tempering simulations using two very distinct energy models (fig. 5B). It is important to emphasize that the imposed constraints do not drastically reduce the total number of degrees of freedom in the latter case as 116 rotatable bonds are treated as independent degrees of freedom when the β -hairpin and the β -strand elements are constrained in their native bound conformations. These results can be reconciled if we assume that molecular recognition of the $p27^{Kip1}$ protein is described within the framework of the nucleation-condensation mechanism in which a minimal accumulation of stable intermediates may be favored in the denatured state and stabilization of stable secondary structure elements is not necessary before the TSE is formed [94]. Furthermore, any potential nucleation site should be weakly populated in the denatured state and become structured and important only in the TSE rather than in the denatured state. According to this hypothesis, we may suggest that structure prediction simulation runs initiated from the post-critical structures, which appear immediately after the TS, and are characterized by structural consolidation of the β -hairpin and the β -strand intermolecular contacts, should consistently become more native-like after quenching at lower temperatures. According to this conjecture, protein conformations where these local secondary structure elements, that are rarely detected in the TSE, are artificially stabilized should encounter a significant kinetic barrier and exhibit a lack of convergence to the native state on a relatively short time scale of structure prediction simulations. It may be therefore argued that consolidation of specific native intermolecular interactions between the β -hairpin and β -strand segments of $p27^{Kip1}$ and the cyclin-A-Cdk2 binary complex, that are rarely populated in the unbound state, but are present in the TSE, should not only accelerate the folding reaction, but also result in the improved convergence in structure prediction of the $p27^{Kip1}$ tertiary complex. Conversely, the excessive stabilization of the unbound $p27^{Kip1}$ states with the intermolecular interactions between $p27^{Kip1}$ and the cyclin-A-Cdk2 binary complex that are barely represented in the TSE, such as the coil and the α -helix, may lead to the increased roughness of the underlying energy landscape and indeed has the adverse effect on the success rate in predicting the native bound protein structure.

These results suggest an atomic picture of the $p27^{Kip1}$ binding mechanism which is consistent with the experimental evidence pointing to the importance of the RNLFG recruitment motif in recognizing an initiation docking site on cyclin A during the tertiary complex formation. A relatively large entropy cost of forming nonlocal interactions during the initiation binding event encourages probing the local intermolecular contacts between the conserved LFG motif of the $p27^{Kip1}$ coil and the recognition groove of cyclin. By analogy with the fly-casting mechanism, the unfolded $p27^{Kip1}$ may anchor weakly in

this region to favorably position the remainder of the protein for coupled folding and binding event. However, further stabilization of the intermolecular contacts for the rigid coil and the α -helix elements of $p27^{Kip1}$, that are rarely present in the TSE, may be disfavored and $p27^{Kip1}$ could exploit its intrinsic disorder in the α -helix until thermal fluctuations allow the system to wrap around the binding partner, with the β -hairpin starting to form hydrophobic interactions and the β -strand adopting the native structure topology at the intermolecular interface. The entropy loss caused by retaining non-local, intermolecular contacts simultaneously at the remote regions of the interface may not be fully compensated by the lack of sufficient strength in these interactions which are yet to be consolidated. Consequently, $p27^{Kip1}$ may find more advantageous at this stage to partially release contacts with the cyclin in order to optimize the balance between the loss in conformational entropy and the energy gain upon formation of the native interactions for the β -strand and β -hairpin elements of $p27^{Kip1}$. At this stage of the process, before the TSE is reached, the conformational entropy dominates and the formation of new native interactions leads to a greater loss in the entropy than the gain in the energy. *i.e.* the coupled folding and binding reaction proceeds up-hill in free energy terms. Only after structural consolidation of the β -hairpin and the β -strand intermolecular contacts is complete and the major free-energy barrier is passed, the coil, α -helix, and 3_{10} -helix could steadily consolidate their native conformations concomitantly with the formation of the native intermolecular contacts. These results reconcile experimental data on the recognition of substrate recruitment motifs, suggesting that $p27^{Kip1}$ may favorably utilize interactions between the $p27^{Kip1}$ coil and the hydrophobic patch on cyclin A to reduce the entropy loss at the initial stage of the reaction. However, these interactions provide only loose geometrical requirements and may not be essential to precisely position the substrate with respect to the active site of Cdk2. This is also consistent with the proposal that $p27^{Kip1}$ may be ubiquitinated through C-terminus region of the protein which can remain flexible in the tertiary complex and thereby accessible for ubiquitination and interaction with the degradation machinery [74-77].

REFERENCES

- [1] GOH K. I., OH E., JEONG H., KAHNG B. and KIM D., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 12583.
- [2] WATTS D. J. and STROGATZ S. H., *Nature*, **393** (1998) 440.
- [3] SOLE R. V., PASTOR-SATORRAS R., SMITH E. and KEPLER T. B., *Adv. Complex Syst.*, **5** (2002) 43.
- [4] WAGNER A., *Proc. R. Soc. London, ser. B*, **270** (2003) 457.
- [5] BARABASI A. L. and ALBERT R., *Science*, **286** (1999) 509.
- [6] JEONG H., TOMBOR B., ALBERT R., OLTVAI Z. N. and BARABASI A. L., *Nature*, **407** (2000) 651.
- [7] ALBERT R., JEONG H. and BARABASI A. L., *Nature*, **406** (2000) 378.
- [8] JEONG H., MASON S. P., BARABASI A. L. and OLTVAI Z. N., *Nature*, **411** (2001) 41.
- [9] ALBERT R. and BARABASI A. L., *Rev. Mod. Phys.*, **74** (2002) 47.

- [10] RAVASZ E., SOMERA A. L., MONGRU D. A., OLTVAI Z. N. and BARABASI A. L., *Science*, **297** (2002) 1551.
- [11] BARABASI A. L. and OLTVAI Z. N., *Nat. Rev. Genet.*, **5** (2004) 101.
- [12] YOOK S. H., OLTVAI Z. N. and BARABASI A. L., *Proteomics*, **4** (2004) 928.
- [13] ALBERT R., *J. Cell Sci.*, **118** (2005) 4947.
- [14] DUNKER A. K., CORTESE M. S., ROMERO P., IAKOUCHEVA L. M. and UVERSKY V. N., *FEBS J.*, **272** (2005) 5129.
- [15] QIAN J., LUSCOMBE N. M. and GERSTEIN M., *J. Mol. Biol.*, **313** (2001) 673.
- [16] KAREV G. P., WOLF Y. I., RZHETSKY A. Y., BEREZOVSKAYA F. S. and KOONIN E. V., *BMC Evol. Biol.*, **2** (2002) 18.
- [17] KOONIN E. V., WOLF Y. I. and KAREV G. P., *Nature*, **420** (2002) 218.
- [18] JORDAN I. K., MARINO-RAMIREZ L., WOLF Y. I. and KOONIN E. V., *Mol. Biol. Evol.*, **21** (2004) 2058.
- [19] ZHANG Z., LUO Z. W., KISHINO H. and KEARSEY M. J., *Mol. Biol. Evol.*, **22** (2005) 501.
- [20] DOKHOLYAN N. V., SHAKHNOVICH B. and SHAKHNOVICH E. I., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 14132.
- [21] TIANA G., SHAKHNOVICH B. E., DOKHOLYAN N. V. and SHAKHNOVICH E. I., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 2846.
- [22] DEEDS E. J., ASHENBERG O. and SHAKHNOVICH E. I., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 311.
- [23] JAMES L. C., ROVERSI P. and TAWFIK D. S., *Science*, **299** (2003) 1362.
- [24] JAMES L. C. and TAWFIK D. S., *Trends Biochem. Sci.*, **28** (2003) 361.
- [25] ROODVELDT C., AHARONI A. and TAWFIK D. S., *Curr. Opin. Struct. Biol.*, **15** (2005) 50.
- [26] EARL D. J. and DEEM M. W., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 11531.
- [27] BOLON D. N. and MAYO S. L., *Proc. Natl. Acad. Sci. U.S.A.*, **98** (2001) 14274.
- [28] BLOOM J. D., WILKE C. O., ARNOLD F. H. and ADAMI C., *Biophys. J.*, **86** (2004) 2758.
- [29] BLOOM J. D., LABTHAVIKUL S. T., OTEY C. R. and ARNOLD F. H., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 5869.
- [30] AHARONI A., GAIDUKOV L. KHERSONSKY O., MCQ GOULD S., ROODVELDT C. and TAWFIK D. S., *Nat. Genet.*, **37** (2005) 73.
- [31] GERLT J. A., BABBITT P. C. and RAYMENT I., *Arch. Biochem. Biophys.*, **433** (2005) 59.
- [32] YOSHIKUNI Y., FERRIN T. E. and KEASLING J. D., *Nature*, **440** (2006) 1078.
- [33] ATWELL S., ULTSCH M., DE VOS A. M. and WELLS J. A., *Science*, **278** (1997) 1125.
- [34] SUNDBERG E. J. and MARIUZZA R. A., *Structure Fold. Des.*, **8** (2000) R137.
- [35] DEMCHENKO A. P., *J. Mol. Recognit.*, **14** (2001) 42.
- [36] CLACKSON T. and WELLS J., *Science*, **267** (1995) 383.
- [37] DELANO W. L., *Curr. Opin. Struct. Biol.*, **12** (2002) 14.
- [38] JONES S. and THORNTON J. M., *Proc. Natl. Acad. Sci. U.S.A.*, **93** (1996) 13.
- [39] LO CONTE L., CHOTHIA C. and JANIN J., *J. Mol. Biol.*, **285** (1999) 2177.
- [40] BOGAN A. A. and THORN K. S., *J. Mol. Biol.*, **280** (1998) 1.
- [41] HU Z., MA B., WOLFSON H. and NUSSINOV R., *Proteins*, **39** (2000) 331.
- [42] MA B., WOLFSON H. J. and NUSSINOV R., *Curr. Opin. Struct. Biol.*, **11** (2001) 364.
- [43] JANIN J., *Structure*, **7** (1999) R277.
- [44] DELANO W. L., ULTSCH M. H., DE VOS A. M. and WELLS J. A., *Science*, **287** (2000) 1279.
- [45] WRIGHT P. E. and DYSON H. J., *J. Mol. Biol.*, **293** (1999) 321.
- [46] DYSON H. J. and WRIGHT P. E., *Curr. Opin. Struct. Biol.*, **12** (2002) 54.

- [47] DYSON H. J. and WRIGHT P. E., *Adv. Protein Chem.*, **62** (2002) 311.
- [48] DUNKER A. K., LAWSON J. D., BROWN C. J., WILLIAMS R. M., ROMERO P., OH J. S., OLDFIELD C. J., CAMPEN A. M., RATLIFF C. M., HIPPS K. W., AUSIO J., NISSEN M. S., REEVES R., KANG C., KISSINGER C. R., BAILEY R. W., GRISWOLD M. D., CHIU W., GARNER E. C. and OBRADOVIC Z., *J. Mol. Graph. Model.*, **19** (2001) 26.
- [49] DUNKER A. K. and OBRADOVIC Z., *Nat. Biotechnol.*, **19** (2001) 805.
- [50] DUNKER A. K., BROWN C. J. and OBRADOVIC Z., *Adv. Protein. Chem.*, **62** (2002) 25.
- [51] DUNKER A. K., BROWN C. J., LAWSON J. D., IAKOUCHEVA L. M. and OBRADOVIC Z., *Biochemistry*, **41** (2002) 6573.
- [52] UVERSKY V. N., *Protein Sci.*, **11** (2002) 739.
- [53] UVERSKY V. N., GILLESPIE J. R. and FINK A. L., *Proteins: Struct. Funct. Genet.*, **41** (2000) 415.
- [54] UVERSKY V. N., *Eur. J. Biochem.*, **269** (2002) 2.
- [55] PLAXCO K. W. and GROSS M., *Nature*, **386** (1997) 657.
- [56] SPOLAR R. S. and RECORD M. T., *Science*, **263** (1994) 777.
- [57] MORGAN D. O., *Nature*, **374** (1995) 131.
- [58] MORGAN D. O., *Annu. Rev. Cell. Dev. Biol.*, **13** (1997) 261.
- [59] DE BOND T. H. L., ROSENBLATT J., JANCARIK J., JONES H. D., MORGAN D. O. and KIM S. H., *Nature*, **363** (1993) 595.
- [60] BROWN N. R., NOBLE M. E., ENDICOTT J. A., GARMAN E. F., WAKATSUKI S., MITCHELL E., RASMUSSEN B., HUNT T. and JOHNSON L. N., *Structure*, **3** (1995) 1235.
- [61] JEFFREY P. D., RUSSO A. A., POLYAK K., GIBBS E., HURWITZ J., MASSAGUE J. and PAVLETICH N. P., *Nature*, **376** (1995) 313.
- [62] RUSSO A. A., JEFFREY P. D. and PAVLETICH N. P., *Nat. Struct. Biol.*, **3** (1996) 696.
- [63] PAVLETICH N. P., *J. Mol. Biol.*, **287** (1999) 821.
- [64] BROWN N. R., NOBLE M. E., ENDICOTT J. A. and JOHNSON L. N., *Nat. Cell. Biol.*, **1** (1999) 438.
- [65] BROWN N. R., NOBLE M. E., LAWRIE A. M., MORRIS M. C., TUNNAH P., DIVITA G., JOHNSON L. N. and ENDICOTT J. A., *J. Biol. Chem.*, **274** (1999) 8746.
- [66] HONDA R., LOWE E. D., DUBININA E., SKAMNAKI V., COOK A., BROWN N. R. and JOHNSON L. N., *EMBO J.*, **24** (2005) 452.
- [67] KRIWACKI R. W., HENGST L., TENNANT L., REED S. I. and WRIGHT P. E., *Proc. Natl. Acad. Sci. U.S.A.*, **93** (1996) 11504.
- [68] RUSSO A. A., JEFFREY P. D., PATTEN A. K., MASSAGUE J. and PAVLETICH N. P., *Nature*, **382** (1996) 325.
- [69] SCHULMAN B. A., LINDSTROM D. L. and HARLOW E., *Proc. Natl. Acad. Sci. U.S.A.*, **95** (1998) 10453.
- [70] BIENKIEWICZ E. A., ADKINS J. N. and LUMB K., *Biochemistry*, **41** (2001) 752.
- [71] HASHIMOTO Y., KOHRI K., KANEKO Y., MORISAKI H., KATO T., IKEDA K. and NAKANISHI M., *J. Biol. Chem.*, **273** (1998) 16544.
- [72] ADKINS J. N. and LUMB K., *Proteins: Struct. Funct. Genet.*, **46** (2002) 1.
- [73] FLAUGH S. L. and LUMB K. J., *Biomacromolecules*, **2** (2001) 538.
- [74] LACY E. R., FILIPPOV I., LEWIS W. S., OTIENO S., XIAO L., WEISS S., HENGST L. and KRIWACKI R. W., *Nat. Struct. Mol. Biol.*, **11** (2004) 358.
- [75] LACY E. R., WANG Y., POST J., NOURSE A., WEBB W., MAPELLI M., MUSACCHIO A., SIUZDAK G. and KRIWACKI R. W., *J. Mol. Biol.*, **349** (2005) 764.
- [76] SIVAKOLUNDU S. G., BASHFORD D. and KRIWACKI R. W., *J. Mol. Biol.*, **353** (2005) 1118.
- [77] BOWMAN P., GALEA C. A., LACY E. and KRIWACKI R. W., *Biochim. Biophys. Acta.*, **1764** (2006) 182.

- [78] ZHU L., HARLOW E. and DYNLACHT B. D., *Genes Dev.*, **9** (1995) 1740.
- [79] CHEN J., SAHA P., KORNBLUTH S., DYNLACHT B. D. and DUTTA A., *Mol. Cell. Biol.*, **16** (1996) 4673.
- [80] LOWE E. D., TEWS I., CHENG K. Y., BROWN N. R., GUL S., NOBLE M. E., GAMBLIN S. J. and JOHNSON L. N., *Biochemistry*, **41** (2002) 15625.
- [81] ZHELEVA D. I., MCINNES C., GAVINE A. L., ZHELEV N. Z., FISCHER P. M. and LANE D. P., *J. Pept. Res.*, **60** (2002) 257.
- [82] KONTOPIDIS G., ANDREWS M. J., MCINNES C., COWAN A., POWERS H., INNES L., PLATER A., GRIFFHS G., PATERSON D., ZHELEVA D. I., LANE D. P., GREEN S., WALKINSHAW M. D. and FISCHER P. M., *Structure*, **11** (2003) 1537.
- [83] ATKINSON G. E., COWAN A., MCINNES C., ZHELEVA D. I., FISCHER P. M. and CHAN W. C., *Bioorg. Med. Chem. Lett.*, **12** (2002) 2501.
- [84] ANDREWS M. J., MCINNES C., KONTOPIDIS G., INNES L., COWAN A., PLATER A. and FISCHER P. M., *Org. Biomol. Chem.*, **2** (2004) 2735.
- [85] CASTANEDO G., CLARK K., WANG S., TSUI V., WONG M., NICHOLAS J., WICKRAMASINGHE D., MARSTERS J. C. JR. and SUTHERLIN D., *Bioorg. Med. Chem. Lett.*, **16** (2006) 1716.
- [86] SHOEMAKER B. A., PORTMAN J. J. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **97** (2000) 8868.
- [87] LEVY Y., WOLYNES P. G. and ONUCHIC J. N., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 511.
- [88] VERKHIVKER G. M., BOUZIDA D., GEHLHAAR D. K., REJTO P. A., FREER S. T. and ROSE P. W., *Curr. Opin. Struct. Biol.*, **12** (2002) 197.
- [89] TSAI C. J., KUMAR S., MA B. and NUSSINOV R., *Protein Sci.*, **8** (1999) 1181.
- [90] KUMAR S., MA B., TSAI C. J., SINHA N. and NUSSINOV R., *Protein Sci.*, **9** (2000) 10.
- [91] VERKHIVKER G. M., BOUZIDA D., GEHLHAAR D. K., REJTO P. A., FREER S. T. and ROSE P. W., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 5148.
- [92] VERKHIVKER G. M., *Biopolymers*, **75** (2004) 420.
- [93] VERKHIVKER G. M., *Proteins*, **58** (2005) 706.
- [94] FERSHT A. R., *Proc. Natl. Acad. Sci. U.S.A.*, **97** (2000) 1525.
- [95] MAYO S. L., OLAFSON B. D. and GODDARD W. A. III., *J. Phys. Chem.*, **94** (1990) 8897.
- [96] CORNELL W. D., CIEPLAK P., BAYLY C. L., GOULD I. R., MERZ K. M., FERGUSON D. M., SPELLMEYER D. C., FOX T., CALDWELL J. W. and KOLLMAN P. A., *J. Am. Chem. Soc.*, **117** (1995) 5179.
- [97] STOUTEN P. F. W., FRÖMMEL C., NAKAMURA H. and SANDER C., *Mol. Simul.*, **10** (1993) 97.
- [98] BEUTLER T. C., MARK. A. E., VAN SCHAIK R. C., GERBER P. R. and VAN GUNSTEREN W., *Chem. Phys. Lett.*, **22** (1994) 529.
- [99] SUGITA Y. and OKAMOTO Y., *Chem. Phys. Lett.*, **314** (1999) 141.
- [100] MOHAMADI F., RICHARDS N. G. J., GUIDA W. C., LISKAMP R., LIPTON M., CAUFIELD C., CHANG G., HENDRICKSON T. and STILL W. C., *J. Comput. Chem.*, **11** (1990) 440.

This page intentionally left blank

Spin glasses, tubes and proteins

J. R. BANAVAR

*Department of Physics, 104 Davey Lab, The Pennsylvania State University
University Park PA 16802, USA*

T. XUAN HOANG

Institute of Physics and Electronics, VAST - 10 Dao Tan, Hanoi, Vietnam

A. MARITAN(*)

*Dipartimento di Fisica “G. Galilei”, Università di Padova
Via Marzolo 8, 35131 Padova, Italy*

CNISM, Unità di Padova - Via Marzolo 8, 35131 Padova, Italy

INFN, Sezione di Padova - Via Marzolo 8, 35131 Padova, Italy

A. TROVATO

CNISM, Unità di Padova - Via Marzolo 8, 35131 Padova, Italy

*Dipartimento di Fisica “G. Galilei”, Università di Padova
Via Marzolo 8, 35131 Padova, Italy*

Proteins are molecular machines which play a vital role in life. The study of proteins has proved to be a daunting problem because of its sheer complexity. We discuss how two physics-based ideas, spin glasses and the phase behavior of a compact flexible tube, are useful for the development of a framework for understanding proteins. The tube picture provides a simple explanation for how geometry and symmetry determine the menu of possible native state folds of proteins, whereas the spin glass paradigm is useful for determining how one of the folds from this predetermined menu is selected as the native-state structure of a given protein sequence.

(*) Lectures delivered by Amos Maritan

Physics has provided an understanding of the essential features underlying the phases of inanimate matter in terms of the principles of geometry and symmetry. The fixed and finite number of crystal structures are determined by the requirements of periodicity and space-filling. Living matter is also governed by physical law. The hallmarks of life are information, replication, natural selection, and functionality. The DNA molecule is the repository of information. Base-pairing and the double-helix structure [1] of the DNA molecule provide a logical mechanism for replication. A similar simple understanding of the physics of protein molecules has been lacking.

We conjecture that the common attributes of proteins, and not the details of the amino acid sequence, determine the fixed menu of putative native-state folds. Our hypothesis, supported by the results of protein research over the last several decades, provides a simple explanation for the astonishing similarities between proteins and yields predictions that differ sharply from current beliefs. Our framework is useful for understanding protein folding, amyloid formation and protein interactions and has important implications for natural selection.

Proteins [2, 3], the workhorse molecules of life, play a central role in all living organisms—for example, enzymes catalyze biochemical reactions, proteins such as hemoglobin transport small molecules such as oxygen, hormonal proteins transfer signals and gene regulatory proteins bind to the DNA and switch the genes on or off. Under physiological conditions, globular proteins fold rapidly and reproducibly into their native-state structures [4]. The folding is predominantly driven by the aversion to water of some amino acid side chains leading to the creation of a hydrophobic core in the compact folded state [5]. In addition, other interactions such as hydrogen bonds, Van der Waals interactions and electrostatic effects play a role in the folding process.

The paradigm of spin glasses [6] has proved to be useful for understanding proteins. A spin glass is characterized by random frustrated interactions that one cannot satisfy simultaneously even in its ground state. Spin glass ideas [7] have been applied to understand the fit of a protein sequence into its native-state structure. The basic notion is that a random sequence of amino acids would necessarily be governed by conflicting interactions in any compact conformation. In order for a protein-like sequence to be able to fold rapidly and reproducibly into its native state, it must be a good fit to its native state and be minimally frustrated [8]. This picture has led to the notion of a folding funnel-like free-energy landscape [9]. Spin glass techniques [10] have been used for the evaluation of approximate interaction energies for protein structure recognition and for the design of optimal sequences characterized by a pronounced funnel.

In order to apply spin glass ideas to proteins, one begins with a sequence of amino acids and determines the nature of the resulting free-energy landscape [11]. This procedure practically necessitates a one-sequence-at-a-time description of proteins [12] and does not account for the fact that, in spite of significant differences in their sequence of amino acids, proteins behave similarly to an extraordinary degree. Furthermore, the spin glass framework does not provide any hints on how one might address the key problem of what selects the class of native folds adopted by proteins.

An example of a stunning similarity in the behavior of globular proteins is the observation that their native states are housed in modular structures commonly built of helices and sheets, which are repetitive motifs for which hydrogen bonds provide the scaffolding [13, 14]. Steric effects [15] also encourage the polypeptide chain to adopt either the helical or the extended conformation. Interestingly, the mechanisms of hydrogen bonds and sterics are unrelated to each other and yet lead to the same consequences. This suggests that the phase of physical matter housing protein native structures may have these motifs as the preferred ground-state structures. Our aim, in this paper, is to understand the physics underlying this previously unstudied phase of matter.

Nature provides us a strong hint on how one might proceed: DNA and proteins are chain molecules and, quite generally, biomolecular structures are characterized by the common occurrence of the helix [1, 13]. From everyday experience, the helix is a natural, compact conformation of a short, flexible tube [16]⁽¹⁾ and the pitch-to-radius ratio of a tube curled up into a space-filling helix coincides nicely with that of protein helices [17]. This suggests that one ought to investigate the phase behavior of a flexible tube subject to compaction in order to investigate whether it is related to and can explain protein behavior.

The coarse-grained flexible tube model [18] captures two essential ingredients of proteins. First, the space within a tube roughly allows for the packing of the main chain protein atoms and local steric effects are encapsulated by constraints on the local radius of curvature. Second, the effects of the geometrical constraints imposed by the chemistry of backbone hydrogen bonds is represented by the inherent anisotropy of a tube, which encourage nearly parallel placement of nearby tube segments. This anisotropy is an inherent feature of *all* chain molecules because there is a special direction associated with an object that is part of a chain defined by the tangent to the chain or equivalently by the locations of the adjoining objects along the chain.

At zero temperature, a short tube, subject to a self-attraction promoting compaction, can exist in either a compact phase or a swollen phase as one varies the ratio of the range of attractive interaction to the tube size. When the ratio is large compared to 1, there are many choices in the relative positioning of the tube segments. In the other limiting case of small ratio, the tube cannot avail itself of the attraction and the conformations are merely self-avoiding. Both these phases are characterized by a high degeneracy. Surprisingly, the energy landscape becomes vastly simpler in a marginally compact phase in which the tube is barely able to avail itself of the self-attraction. The number of options is greatly reduced and tube segments subject to attraction need to be placed alongside and

⁽¹⁾ One ensures self-avoidance in a system of hard spheres by considering all pairs of spheres and ensuring that, for each pair, the mutual distance between the sphere centers exceeds the sphere diameter. The generalization of this for a flexible tube of non-zero thickness in the continuum entails the use of a many-body interaction and removes the singularity in the conventional pairwise description. The recipe considers all triplets of points along the axis of a flexible tube. Draw circles through each of the triplets and require that none of the radii is smaller than the tube radius.

parallel to each other. The proximity of this phase of matter to the phase transition to the swollen phase leads to sensitivity of the structures to perturbations. Furthermore, helices and sheets are the preferred structures in this marginally compact phase [18]. One may understand the emergence of the second building block, the almost planar β -sheets, from the *zig-zag* pattern adopted by a pair of *discrete* tube segments, barely subject to an attractive interaction, placed alongside each other [18]. The inherent biaxial symmetry of the resulting structure promotes planarity.

Recent work [19-21] has probed the phase diagram of a short discrete tube, made up of just one type of amino acid, characterized by generic geometrical constraints arising from sterics and explicit backbone hydrogen bonding. On allowing for mild variations in the overall hydrophobicity and local curvature energy penalty parameters, one obtains a free-energy landscape, that is determined by geometry and symmetry, with multiple minima corresponding to a menu of folds which are modular structures built of helices and sheets as in proteins.

Armed with the insights obtained from these studies of a short tube, we conjecture that the intrinsic geometry and the physical and chemical properties of a polypeptide chain lead to protein-native state structures being housed in a novel phase of matter whose characteristics are determined by the common attributes of all proteins and not by the sequence of amino acids of any specific protein. Such behavior is reminiscent [22] of the selection principles underlying the fixed and finite number of crystal structures. Unlike crystals, protein are neither infinite nor periodic. Yet, our hypothesis is that the menu of protein native-state structures is determined by considerations of geometry and symmetry.

Our hypothesis is supported by numerous observations. Anfinsen [4] wrote, “Biological function appears to be more a correlate of macromolecular geometry than of chemical detail.” The propensity to form helices and sheets [13-15] arises mainly from consideration of the chain backbone and not the details of the amino acid side chains. Quite remarkably, the total number of distinct folds is only of the order of a few thousand [23], a fact exploited in structure prediction techniques [24]. Proteins are relatively short-chain molecules and longer globular proteins form domains which fold autonomously [25]. The fold topology of single domains is conserved by natural evolution [26] and the duplication and recombination of single domains is thought to be a major determinant of the evolution of multi-domains proteins in higher organisms [27]. Many proteins share the same native-state fold [28] and the mutation of one amino acid into another only rarely leads to radical changes in the native-state structure. Indeed, multiple protein functionalities can arise within the context of a single fold [29].

Recent experiments [2, 30] have been successful in mapping out the nature of the folding transition state in several cases. Interestingly, the rate of protein folding is not vastly different [24, 31] even when one has large changes in the amino acid sequence [31, 32], as long as the overall topology of the folded structure is the same. Furthermore, mutational studies [24, 30] have shown that, in the simplest cases, the structures of the transition states are also similar in proteins sharing the same native-state topology.

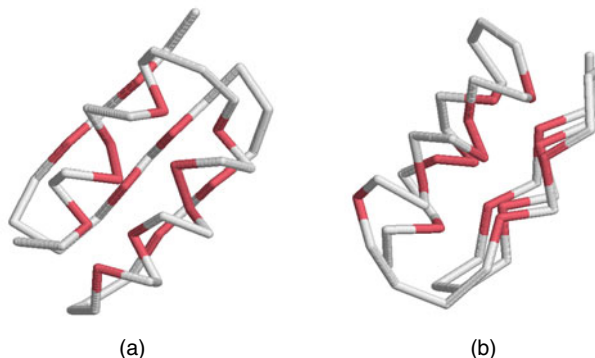


Fig. 1. – Ground state, deduced in computer simulations, of a designed hydrophobic-polar sequence of length 48. The conformations shown in (a) and (b) are the same but viewed from a different angle. The hydrophobic (H) and polar (P) amino acids are shown in red and grey colors, respectively. The details of the original homopolymer model are described in ref. [19-21]. The hydrophobicity energy parameter is set to be -0.4 , if a contact is H-H and 0 , otherwise. The local bending energy penalty is chosen uniformly to be equal to 0.2 . The design procedure is based on a simple recipe of helices having a hydrophobic pattern of $(i, i + 3)$ or $(i, i + 4)$ and strands having a pattern of $(i, i + 2)$, where the two variables denote the position of hydrophobic amino acids in the sequence.

Anfinsen [4] wrote that “the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment”. The results described above strongly suggest instead that the key role of the sequence is not in determining the native conformation but rather in *choosing* the native conformation from a *fixed, predetermined menu of folds*. The spin glass paradigm, while useful for understanding the quality of the sequence-structure fit, is unable to address the issue of what selects the candidate structures that proteins are housed in. The tube paradigm, on the other hand, provides an explanation of this vital issue.

We turn now to an exploration of the consequences of our hypothesis.

The pre-sculpted free-energy landscape allows one to easily design sequences that fold to one of the predetermined structures [21]. Figure 1 shows a protein-like structure made up of a two-helix bundle and a three-stranded β -sheet that is the ground state of a designed sequence of length 48 in the two-letter (hydrophobic-polar) code.

Evolution can be thought of as a “random walk” that forms a connected network [33] in sequence space—there is no similar continuous variation in structure space. The characteristics required for protein native-state structures to be targets of an evolutionary process are stability and diversity [34]. Stability is needed because one would not want to mutate away a DNA molecule able to encode a useful protein and diversity, in order to allow evolution to build complex and versatile forms. The mechanism for natural selection arises naturally within the framework of a predetermined menu of conserved folds. DNA molecules that code for amino acid sequences that fit well into one of these putative native-state structures and have useful functionalities thrive at the expense of

molecules that create sequences that are not useful. Stability would be lost were each sequence to determine its own distinct native-state structure leading to a continuum of putative native-state structures. Substantial neutrality in the evolution of sequences or genotypes and a predetermined discrete menu of phenotypes or native-state structures are necessary for proteins to act as the targets of natural selection.

Protein interactions are at the heart of the network of life. As sequences undergo evolution and natural selection, there are two distinct scenarios for the behavior of the corresponding native-state structures. In order to maintain and especially enhance the interactions among proteins, the corresponding structures must either co-evolve in a coherent manner to retain the classic lock-and-key mechanism or, more simply, take advantage of the existence of the menu of folds determined, not by the sequence, but by considerations common to all proteins. Such a menu would provide a fixed backdrop [35] for the action of natural selection. Optimal sequences would not only fold into the desired native-state structure (fig. 1) but would also be fit in the environment of other proteins. Indeed, in this picture, sequences and functionalities evolve within the constraint of these folds, which, in turn, are immutable and determined by physical law.

In several instances, protein interactions are carried out by natively or intrinsically unstructured proteins (IUPs), which under physiological conditions do not exhibit extensive structural order in solution [36, 37]. A key feature of IUPs is their high conformational flexibility, that allows them to interact with different molecular partners and adopt relatively rigid conformations in the presence of natural ligands, thus undergoing a loss of conformational entropy upon binding [38]. The conformation in the bound state is determined not only by the amino acid sequence but also by the structure of the interacting partner [39]. This behaviour nicely fits within the pre-sculpted free-energy landscape framework. Structures in the predetermined menu available for short homopolymeric tubes can indeed be selected by a proper choice of the geometry of the interacting ligand/substrate [21].

The optimal packing of many short chains [20] is most readily carried out by forming cross-linked β -architectures where two or more β -sheets are layered in a “sandwich” structure (fig. 2), as commonly observed in the amyloid phase [40, 41]. Indeed, one measure of the fitness of a sequence in its environment is its ability to remain soluble and not aggregate with other protein chains. A range of human diseases such as Alzheimer’s, spongiform encephalopathies, Type-II diabetes and light-chain amyloidosis lead to degenerative conditions and involve the deposition of plaque-like material in tissue arising from the aggregation of proteins [42]. A variety of proteins not involved in these diseases also form *in vitro* aggregates very similar to those implicated in the diseased state showing that, as predicted by the tube model, this aggregational tendency is a generic property of polypeptide chains with the specific sequence of amino acids again playing at best a secondary role. The spin glass paradigm, which is a sequence-based, cannot address the issue of why such aggregation ought to occur generically. Sequence heterogeneity would play a role in determining the nature of the β -structure formed by the assembly of the aggregating chains, *e.g.* the cross-linked “triangular” β -helix arrangement [21].

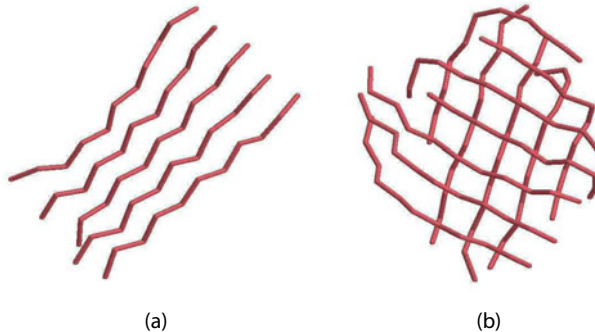


Fig. 2. – Aggregated structures formed by five and ten homopolymer chains of length 11 in the model described in ref. [19, 20]. The hydrophobicity energy parameter is chosen to be -0.08 and the local bending energy penalty 0.3 which leads to a single chain having a helical ground state. No hydrogen bonding is allowed for the amino acids at the ends of the chains. The lowest-energy conformations obtained in long simulations at $T = 0.19$ for five chains (a), and for ten chains (b) are shown. The systems are confined within cubic boxes of edge $L = 80 \text{ \AA}$ and $L = 100 \text{ \AA}$, respectively.

The spin glass [6] with random, frustrated interactions has multiple ground states and is characterized by stability and diversity [34]. Strikingly, the phase of matter housing the ground states of short tubes has an energy landscape, even in the absence of heterogeneity or randomness, characterized not only by stability and diversity but also the resulting modular structures possess exquisite sensitivity to the right types of perturbations. Furthermore, unlike the spin degrees of freedom in a spin glass, the rich variety of structures are all geometry based and are exploited by nature in the context of the proteins, the molecules of life, using the classic lock-and-key mechanism.

In summary, the experimental data accumulated over the years suggest a considerable simplicity to the protein problem. Just as the class of cross-linked β -structures are determined from geometrical considerations, the menu of protein native-state structures is also determined by the common attributes of globular proteins. The allowed folds, made up of helices and sheets connected together by turns, are structures, which are stable, yet characterized by flexibility and sensitivity. The rich repertoire of amino acids allows for sequence design—proteins are those sequences, which not only fit harmoniously into one of the structures in the predetermined menu but are also able to interact synergistically with other cell products. Strikingly, these interactions are facilitated by the native-state structures themselves. The menu of folds act as a fixed backdrop for the evolution of sequences and functionalities. Recent work [43] has provided further confirmation of the validity of the framework proposed here by demonstrating a one-to-one correspondence between native folds of single-domain proteins and compact conformations of homopolypeptides generated by randomly assembling hydrogen-bonded secondary-structure elements. Furthermore, active-site geometries are reproduced as well by simple packing of compact hydrogen-bonded structures [43].

Rose *et al.* [44] have presented a chemistry-based approach complementary to the tube model in which “the sidechain/backbone paradigm is inverted. A folding protein selects its fold from a limited repertoire of stable scaffolds, each built from a composite of hydrogen-bonded α -helices and/or β -strands. Folding is an inherently digital process in which the formative interactions are among backbone elements.”

* * *

We are indebted to G. ROSE for helpful discussions. This work was supported by PRIN 2003, no. 2003025755, and 2005, no. 2005027330, INFN, NASA, NSF IGERT grant DGE-9987589, and the NSC of Vietnam.

REFERENCES

- [1] WATSON J. D. and CRICK F. H. C., *Nature*, **171** (1953) 737.
- [2] FERSHT A. R., *Structure and Mechanism in Protein Science: A Guide to Enzyme Catalysis and Protein Folding* (W.H. Freeman and Company, New York) 1999.
- [3] FINKELSTEIN A. V. and PTISTYN O., *Protein Physics: A Course of Lectures* (Academic Press, New York) 2002.
- [4] ANFINSEN C. B., *Science*, **181** (1973) 223.
- [5] KAUZMANN W., *Adv. Protein Chem.*, **14** (1959) 1.
- [6] BINDER K. and YOUNG A. P., *Rev. Mod. Phys.*, **58** (1986) 801; MEZARD M., PARISI G. and VIRASORO M., *Spin Glass Theory and Beyond* (World Scientific, Singapore) 1987; ANDERSON P. W., *Phys. Today*, **43**, issue no. 3 (1990) 9.
- [7] STEIN D. L., *Proc. Natl. Acad. Sci. U.S.A.*, **82** (1985) 3670; FRAUENFELDER H., SLIGAR S. G. and WOLYNES P. G., *Science*, **254** (1991) 1598.
- [8] BRYNGELSON J. D. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **84** (1987) 7524.
- [9] LEOPOLD P. E., MONTAL M. and ONUCHIC J. N., *Proc. Natl. Acad. Sci. U.S.A.*, **89** (1992) 8271; WOLYNES P. G., ONUCHIC J. N. and THIRUMALAI D., *Science*, **267** (1995) 1619; DILL K. A. and CHAN H. S., *Nat. Struct. Biol.*, **4** (1997) 10.
- [10] GOLDSTEIN R. A., LUTHEY-SCHULTEN Z. A. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **89** (1992) 9029; BRYNGELSON J. D., ONUCHIC J. N., SOCCI N. D. and WOLYNES P. G., *Proteins*, **21** (1995) 167; KLIMOV D. K. and THIRUMALAI D., *Phys. Rev. Lett.*, **76** (1996) 4070; CHAN H. S. and DILL K. A., *Proteins*, **30** (1998) 2.
- [11] LI Z. Q. and SCHERAGA H. A., *Proc. Natl. Acad. Sci. U.S.A.*, **84** (1987) 6611.
- [12] The theoretical procedure for studying spin glasses entails averaging over the quenched disorder as exemplified in RAMANATHAN S. and SHAKHNOVICH E. I., *Phys. Rev. E*, **50** (1994) 1303.
- [13] PAULING L., COREY R. B. and BRANSON H. R., *Proc. Natl. Acad. Sci. U.S.A.*, **37** (1951) 205.
- [14] PAULING L. and COREY R. B., *Proc. Natl. Acad. Sci. U.S.A.*, **37** (1951) 729.
- [15] RAMACHANDRAN G. N. and SASISEKHARAN V., *Adv. Protein Chem.*, **23** (1968) 283; SRINIVASAN R. and ROSE G. D., *Proc. Natl. Acad. Sci. U.S.A.*, **96** (1999) 14258.
- [16] BANAVAR J. R., GONZALEZ O., MADDOCKS J. H. and MARITAN A., *J. Stat. Phys.*, **110** (2003) 35.
- [17] MARITAN A., MICHELETTI C., TROVATO A. and BANAVAR J. R., *Nature*, **406** (2000) 287.
- [18] BANAVAR J. R. and MARITAN A., *Rev. Mod. Phys.*, **75** (2003) 23.

- [19] HOANG T. X., TROVATO A., SENO F., BANAVAR J. R. and MARITAN A., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 7960.
- [20] BANAVAR J. R., HOANG T. X., MARITAN A., SENO F. and TROVATO A., *Phys. Rev. E*, **70** (2004) 041905.
- [21] HOANG T. X., MARSELLA L., TROVATO A., SENO F., BANAVAR J. R. and MARITAN A., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 6883.
- [22] SCHRODINGER E., *What is life?* (Doubleday, Garden City, N.Y.) 1956.
- [23] CHOTHIA C., *Nature*, **357** (1992) 543.
- [24] BAKER D., *Nature*, **405** (2002) 39.
- [25] PRIVALOV P. L., *Adv. Protein Chem.*, **35** (1982) 1.
- [26] PONTING C. P. and RUSSELL R. R., *Annu. Rev. Biophys. Biomol. Struct.*, **31** (2002) 45.
- [27] CHOTHIA C., GOUGH J., VOGEL C. and TEICHMANN S. A., *Science*, **300** (2003) 1701.
- [28] MATTHEWS B. W., *Annu. Rev. Biochem.*, **62** (1993) 139.
- [29] HOLM L. and SANDER C., *Proteins*, **28** (1997) 78.
- [30] CHITI F., TADDEI N., WHITE P. M., BUCCIANINI M., MAGHERINI F., STEFANI M. and DOBSON C. M., *Nat. Struct. Biol.*, **6** (1999) 1005; MARTINEZ J. C. and SERRANO L., *Nat. Struct. Biol.*, **6** (1999) 1010.
- [31] PERL D., WELKER C., SCHINDLER T., SCHRODER K., MARAHIEL M. A., JAENICKE R. and SCHMID F. X., *Nat. Struct. Biol.*, **5** (1998) 229.
- [32] KIM D. E., GU H. and BAKER D., *Proc. Natl. Acad. Sci. U.S.A.*, **95** (1998) 4982.
- [33] MAYNARD SMITH J., *Nature*, **225** (1970) 563.
- [34] ANDERSON P. W., *Proc. Natl. Acad. Sci. U.S.A.*, **80** (1983) 3386.
- [35] DENTON M. and MARSHALL C., *Nature*, **410** (2001) 417.
- [36] WRIGHT P. E. and DYSON H. J., *J. Mol. Biol.*, **293** (1999) 7960.
- [37] DUNKER A. K. and OBRADOVIC Z., *Nat. Biotechnol.*, **19** (2001) 805.
- [38] DYSON H. J. and WRIGHT P. E., *Curr. Opin. Struct. Biol.*, **12** (2002) 54.
- [39] TOMPA P., *Trends Biochem. Sci.*, **27** (2002) 527.
- [40] PETKOVA A. T., ISHII Y., BALBACH J. J., ANTZUTKIN O. N., LEAPMAN R. D., DELAGLIO F. and TYCKO R., *Proc. Natl. Acad. Sci. U.S.A.*, **99** (2002) 16742.
- [41] NELSON R., SAWAYA M. R., BALBIRNIE M., MADSEN A. O., RIEKEL C., GROTHE R. and EISENBERG D., *Nature*, **435** (2005) 773.
- [42] KELLY J. W., *Curr. Opin. Struct. Biol.*, **8** (1998) 101; DOBSON C. M., *Nat. Rev. Drug Discov.*, **2** (2003) 154.
- [43] ZHANG Y., HUBNER I. A., ARAKAKI A. K., SHAKHNOVICH E. and SKOLNICK J., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 2605.
- [44] ROSE G. D., FLEMING P. J., BANAVAR J. R. and MARITAN A., *Proc. Natl. Acad. Sci. U.S.A.*, **103** (2006) 16623.

This page intentionally left blank

Studying protein folding with laser tweezers

C. CECCONI

Dipartimento di Fisica, Università di Modena e Reggio Emilia - Modena, Italy

E. A. SHANK

*Department of Microbiology and Molecular Genetics, Harvard Medical School
Boston, MA, USA*

S. MARQUSEE

*Department of Molecular & Cell Biology and the Institute for Quantitative Biology
University of California, Berkeley, Berkeley CA 94720-3206, USA*

C. J. BUSTAMANTE

Howard Hughes Medical Institute, University of California, Berkeley, Berkeley, CA, USA

1. – Introduction

Proteins must fold into compact and unique three-dimensional structures to carry out their specific functions. If folding goes wrong, proteins become useless and often toxic molecules for living cells. Millions of people around the world suffer from diseases caused by protein misfolding, such as Gaucher's disease, Alzheimer's disease and Parkinson's disease [1, 2]. Despite its importance, our understanding of the basic rules that govern how a protein attains its native structure is still incomplete. This lack of information is partly due to the inadequacy of conventional bulk methods to study a process that is highly heterogeneous. During folding, individual molecules are thought to follow different pathways and populate different intermediate structures on their journey to the native state [3]. Such a diversity of behaviors is often blurred in the ensemble average measured

by bulk methods. In bulk experiments, the signal is always an ensemble- and time-average of the contributions of a large, de-phased population of molecules, in which the many different folding routes and non-cumulative intermediate states are often not clearly resolved. Moreover, the unfolding potential derived from these experiments is often expressed in terms of a reaction coordinate, typically the β Tanford (β_T) value, which is not easy to interpret [4].

Recent advances in single-molecule manipulation techniques, such as atomic force microscopy (AFM) and laser tweezers [5], have made it possible to revisit protein folding with a new approach. In these experiments single molecules are directly manipulated and their behavior under tension is described in terms of a well-defined reaction coordinate, namely their molecular end-to-end distance. These methods present a number of advantages over more traditional bulk techniques, allowing us: i) to monitor in real time the fluctuations between different molecular conformations and characterize directly the thermodynamics and kinetics of these processes, ii) to measure directly the potential of mean force of a molecule as a function of its extension, iii) to probe the deformability of the folded structure, iv) to characterize the entropic elasticity of the unfolded state, and v) to measure the magnitude of the forces that hold together the protein's tertiary structure. Moreover, because of their single-molecule nature, these experiments permit the investigation of alternative, less probable folding trajectories.

The large majority of mechanical manipulation studies on protein molecules published to date have been carried out using AFM. In these studies the mechanical unfolding of long polymeric proteins—either naturally occurring [6] or biochemically synthesized [7,8]—has been characterized in terms of a variety of parameters, such as: i) unfolding forces, ii) rates of unfolding, iii) distances from the folded structure to the transition state, iv) unfolding intermediates, and v) anisotropy of the energy landscape [9-11]. These studies have provided a wealth of new information about the mechanisms by which proteins unfold under tension and allowed us to explore phenomena previously inaccessible to experimental investigation. However, because of the stiffness of the employed AFM probes, the range of forces accessible in these experiments has been limited. A typical AFM cantilever has a spring constant of about 50 pN/nm, and thus it has a root-mean square (RMS) force noise of ~ 15 pN. Such a cantilever therefore cannot be used to monitor events that typically occur at very low forces, such as refolding transitions and fluctuations between different molecular conformations [12,13]. Because of this, single-molecule mechanical manipulation experiments have so far been restricted primarily to the study of the high-unfolding behavior of proteins. These experimental limitations have recently been overcome in the work published in *Science* by Cecconi *et al.* [14]. In this work, the authors present a method to manipulate individual globular proteins in the low-force regime of the laser tweezers, where optically trapped beads have a RMS force noise of less than 1 pN [14,15]. Using this experimental approach, Cecconi and co-authors directly monitored the unfolding and *refolding* trajectories of *E. coli* ribonuclease HI (RNase H) molecules and uncovered information inaccessible to more traditional bulk techniques.

This paper provides a brief overview of the basic principles of the laser tweezers, details the experimental procedure that we devised to make globular proteins amenable

to mechanical manipulation, and describes and discusses the major results and findings of Cecconi *et al.* regarding the folding mechanism of RNase H obtained through single-molecule manipulation studies.

2. – Laser tweezers

Laser tweezers were invented by Arthur Ashkin and colleagues in 1986 when they discovered that by focusing a laser beam through a microscope objective they could trap and manipulate small particles of high refractive index, such as plastic beads and oil droplets [16]. Since its invention, laser tweezers have become a versatile tool to trap and manipulate a diverse range of small objects, from living cells down to single atoms, with many applications in biophysics and other disciplines [17, 18]. The mechanism of optical trapping for objects that are much larger than the wavelength of light can be explained through a simple ray optics model. Light photons carry momentum equal to h/λ , where h is Planck's constant and λ is the wavelength of light. When a stream of photons interacts with an object and it is either absorbed or scattered the momentum carried by light changes. Since momentum is conserved, the rate of change in momentum of the light must be equal and opposite to the rate of change in momentum of the scattering object. Because of this, a particle that acts as a positive lens and refracts light towards its center can be entrained in a light beam.

Consider a collimated beam of light with a Gaussian intensity profile hitting a refractive sphere that is situated off-axis in the beam, as shown in fig. 1a). For each ray in the beam the components of momentum flux parallel and perpendicular to the optical axis are

$$(1) \quad (dp/dt)_{\text{parallel}} = n_B(W/c) \cos \theta$$

and

$$(2) \quad (dp/dt)_{\text{transverse}} = n_B(W/c) \sin \theta,$$

respectively, where W is the power of the light ray, θ is the angle of the ray to the optic axis, c is the speed of light, and n_B is the refractive index of the liquid surrounding the sphere [19]. The reaction impulse (F) felt by the bead is equal but opposite to the change in the light momentum flux summed over all rays passing through the bead:

$$(3) \quad F_{\text{bead}} = - \sum_{\text{rays}} (dp/dt)_{\text{in}} - (dp/dt)_{\text{out}}.$$

If the bead hit by the beam of light acts as a converging lens, as shown in fig. 1a), then the strong central rays are refracted away from the beam axis, while the weaker peripheral rays are refracted towards it. This results in a net downward change of light momentum flux that causes the bead to feel a force that draws it towards the center of the beam. However, since the parallel component of the ray momentum is reduced

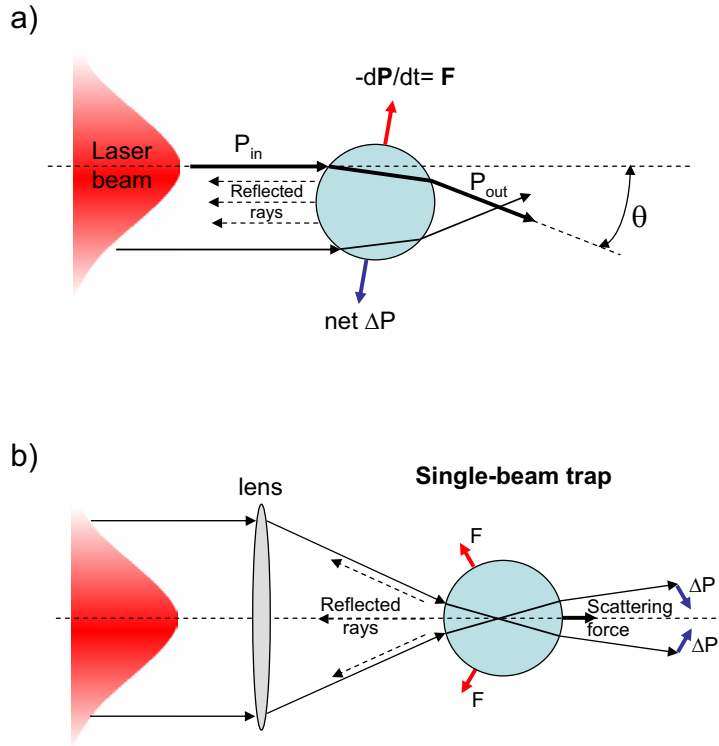


Fig. 1. – Light momentum effects on a refractive sphere. a) A collimated laser beam with a Gaussian intensity profile entrains a bead, but it also pushes it away from the light source. b) A focused laser beam with a Gaussian intensity profile traps a bead at a particular point in space. The scattering force caused by reflected rays is balanced by the refraction of the high angle rays and the particle is trapped slightly off center.

after refraction, the bead feels an additional scattering force that pushes it downstream. Additionally, the bead is scattered forward by reflected rays. Thus, when a particle is hit by a collimated beam of light with an intensity profile peaked in its center, the particle light source [20].

To trap a bead at a particular point in space, the beam of light must be focused (fig. 1b)). In this situation, a bead at the focus is pushed forward by the reflected rays until the scattering force is compensated by a backward force caused by the increase in the forward momentum flux of the rays refracted by the bead. The bead is thus trapped slightly downstream of the light focus. The larger the focal cone angle of the beam is, the stronger the backward force that opposes the axial escape of the bead is. Thus, to efficiently trap a bead with a single-beam trap, the back aperture of a high numerical aperture objective must be completely filled by the trapping beam.

In order to trap the bead at the light focus and avoid the use of large focal cone angles, dual counter-propagating laser beams with a common focal point must be used, as shown

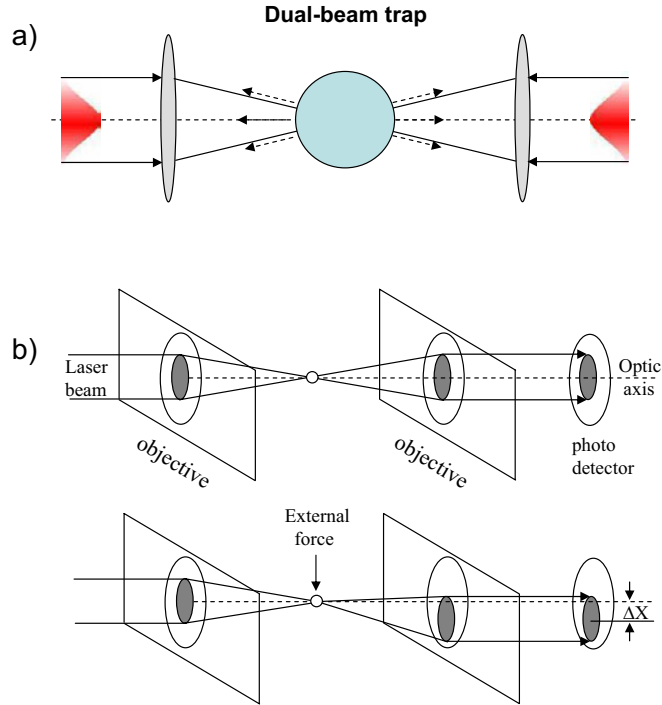


Fig. 2. – Dual counter-propagating laser beams. a) In a dual-beam trap, the scattering forces caused by reflected rays are canceled, while the gradient forces are doubled; the bead is trapped at the light focus. b) Light-momentum force sensor, based on fig. 1 of Smith *et al.* [15]. Two low numerical aperture (NA) beams are used with two high-NA objectives. Each objective is used twice: to focus one beam and collect the other beam for analysis. The narrow cones of light used in this set-up allow the collection of nearly all the light leaving the trap, even after significant deflection of the beam. When an external force is applied on the bead, the bead moves slightly downward and the light is refracted asymmetrically. By collecting all the light momentum leaving the trap it is possible to measure the force acting on the bead as $F = (W/c)(x/R_L)$, where W is the light intensity, c is the speed of light, x is the offset measured by the photodetector and R_L is the focal length of the lens [15].

in fig. 2a). In this experimental set-up, the scattering forces generated by reflected rays are canceled, while the transverse gradient forces are doubled. Using narrow cones of light with this laser tweezer configuration allowed Smith *et al.* to devise a new method for measuring optical trap forces that relies on the direct measurement of the transverse change in momentum flux of the trapping beam [15] (fig. 2b)). When a force F is applied on the trapped particle, the particle moves downward until the light pressure from the deflected beam exactly balances the external force. If narrow cones of light are used and the back aperture of the laser trap objectives is underfilled, then all the light momentum leaving the trap can be collected and directed to position-sensitive photodetectors. The offset distance measured by the photodetectors can then be transformed into the angular

deflection of light to calculate the change in the momentum flux of the trapping beam that, according to momentum conservation, equals the force applied on the particle. This force-measurement method presents several advantages over more traditional ones, being independent of the particle's size or shape, the distance beyond the cover glass, and the viscosity and refractive index of the buffer [15]. The experimental design depicted in fig. 2b) cannot be used to measure forces in a single-beam trap because a single beam of light with such small cone angle does not trap a bead efficiently.

The data shown in this paper were collected with a momentum-flux force sensor dual beam laser trap built by Steve B. Smith and detailed in Smith *et al.* [15].

3. – Synthesis of molecular constructs for use in mechanical manipulation studies

A major issue in single-molecule manipulation studies is to find conditions that facilitate the attachment of the molecule under study to movable substrates, while keeping the strength of the interactions between the tethering surfaces to a minimum. This is a difficult task to fulfill when trying to manipulate nanometer-sized globular proteins with micrometer-sized beads, such those used in laser tweezer experiments. A direct attachment of the protein molecule to the beads would require the large tethering surfaces to come so close to each other they would interact, thus compromising the measurements. To overcome this problem, we developed a new experimental procedure to make globular proteins amenable to mechanical manipulation in laser tweezer experiments [21]. This approach relies on the use of molecular handles, ~ 500 bp DNA molecules, to specifically connect the protein to polystyrene beads and keep the attachment points at a distance at which unspecific interactions between the tethering surfaces are negligible (fig. 3). One end of each DNA molecule is covalently attached to a cysteine residue of the protein through a disulfide bond, while the other end is bound to a bead through either streptavidin-biotin or digoxigenin-antibodies interactions. The thiol-thiol chemistry between the handles and the protein is mediated by 2,2'-dithiodipyridine (DTDP) [22-24]. An excess of DTDP is first used to activate cysteine-bearing proteins, which are then reacted with a mixture of the two handles to produce the DNA-protein chimeras used in laser tweezer experiments. Only 50% of the chimeras synthesized through this method have the required configuration of one biotin- and one digoxigenin-labeled handles; the rest of the molecules carry identical handles and therefore do not function in laser tweezer experiments. The time course of the disulfide bond formation can be followed spectrophotometrically at 343 nm via the release of the leaving group pyridine-2-thione. The activation of cysteine residues with DTDP is typically complete within a few minutes, whereas the attachment of the DNA molecules to the activated proteins follows slower kinetics and can take hours, or even days, to reach completion (data not shown) [21]. The success and extent of the protein/DNA coupling reaction can be assessed both by gel electrophoresis and AFM. Indeed, DNA-protein complexes run as distinct retarded bands in polyacrylamide gels, and can be easily visualized by AFM imaging, (data not shown) [21].

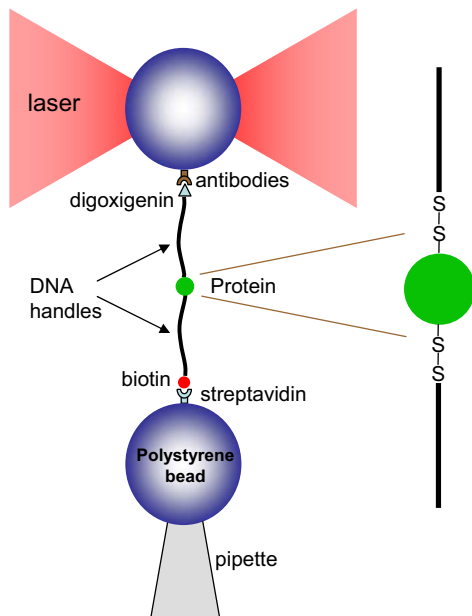


Fig. 3. – Schematic representation of the experimental set-up used to mechanically manipulate individual globular proteins by laser tweezers (not to scale). During the experiment the molecule is stretched and relaxed several times by moving the pipette relative to the optical trap and the force applied on the molecule is determined by measuring the change in momentum flux of the light beams leaving the trap. The extension of the molecule is determined using a light “lever system” [31].

To ensure that meaningful information could be obtained from the manipulation of our molecular constructs, we assessed the effect of the handles on the structure and enzymatic activity of the protein through bulk experiments. We showed that DNA-modified RNase H retains its overall fold and its enzymatic activity (fig. 4a) and b), respectively) an indication that in our laser tweezer samples the protein preserves its three-dimensional conformation [14]. This experimental procedure devised to manipulate individual RNase H molecules should be general and adaptable to the study of most proteins.

4. – Mechanical manipulation of single RNase H molecules by laser tweezers

E. coli ribonuclease HI (RNase H) is an enzyme that hydrolytically cleaves the ribonucleotide backbone of RNA-DNA hybrids in a divalent cation (Mg^{2+} or Mn^{2+}) dependent manner [25]. It is a small single-domain protein of 155 amino acids, consisting of both α -helices and β -sheets, and lacking disulfide bonds and prosthetic groups [26]. The structure, stability and folding of RNase H have been intensively studied with bulk techniques by several groups [26-28]. Despite this effort, however, crucial aspects of the folding mechanism of this protein have proven difficult to investigate with traditional en-

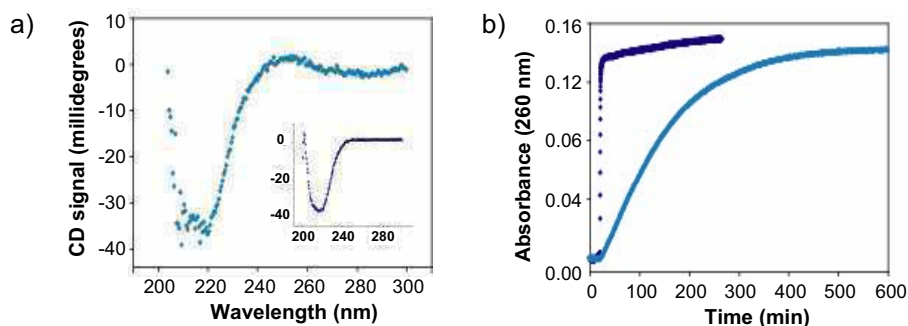


Fig. 4. – Effect of DNA handles on protein's structure and enzymatic activity. a) Circular dichroism spectra of RNase H alone (inset) and with 40 bp dsDNA molecules attached (light blue dots), adapted from Cecconi *et al.* [14]. The two spectra are very similar, indicating that the DNA-modified protein retains its overall fold. These experiments were carried out with 40 bp dsDNA because it was technically difficult to obtain enough sample with long handles (558 bp) for circular dichroism studies. b) Spectrophotometric activity assay of the protein alone (dark blue dots), and with 558 bp dsDNA attached (light blue dots). The DNA-modified protein retains its enzymatic activity; adapted from Cecconi *et al.* [14].

semble methods. For example, in bulk studies, RNase H has been observed to populate a moderately stable, partially structured kinetic intermediate, postulated to be a molten globule [29] (fig. 5). However, because this partially folded conformation forms within the dead time of the measuring instrument (12 ms), it has been difficult to establish directly

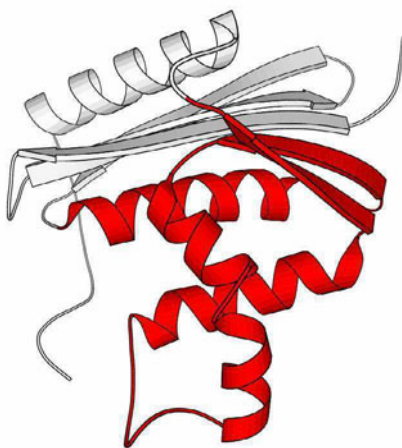


Fig. 5. – Ribbon structure of RNase H. The red region is the part of the molecule that is structured in the burst-phase intermediate as determined by pulse labeling hydrogen-deuterium exchange experiments [29].

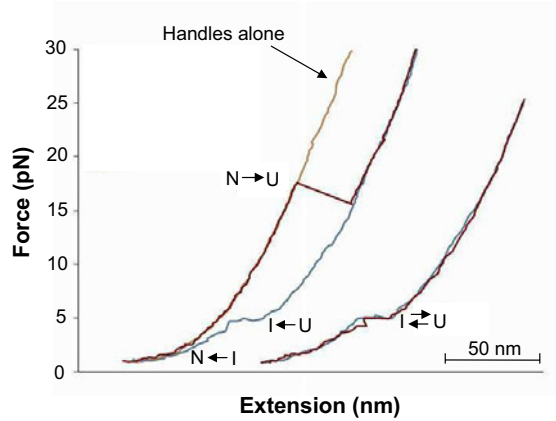


Fig. 6. – Mechanical manipulation of individual RNase H molecules. Two successive force-extension cycles obtained by pulling and relaxing the same RNase H molecule are shown; one cycle is displaced laterally for clarity, adapted from Cecconi *et al.* [14]. The stretching (red) and relaxation (blue) traces display sudden changes in extension corresponding to unfolding and refolding events of the protein. The size of the high-force stretching transition at ~ 19 pN is consistent with the complete unfolding of the native state. The size of the low-force stretching and relaxation transitions at ~ 5.5 pN is instead consistent with the unfolding and refolding of an intermediate, partially folded, structure. In fact, the refolding of the protein into this intermediate state does not restore the original length of the molecule, leaving a gap between the stretching and the relaxation traces at ~ 5.5 pN (left curve). If the intermediate state is not given enough time at low forces to fold into the native structure, it unfolds again in next pulling cycle at about the same force at which it folded, indicating that the unfolding of I is reversible (right curve). When handles alone with no protein were pulled no transitions were observed (yellow trace).

whether this intermediate is on- or off-pathway and whether it is an obligatory step in the folding trajectory to the native state. Moreover, it has not been clear whether this intermediate is a distinct thermodynamic state or simply a redistribution of the unfolded ensemble under folding conditions [30].

With the intent of shedding light on some of these unresolved issues, we decided to reinvestigate the folding of RNase H with a new approach: using laser tweezers. Individual RNase H molecules were tethered to polystyrene beads by means of molecular handles, as depicted in fig. 3 and described above. One bead was held in a force-measuring optical trap, while the other was held by suction at the end of a micropipette connected to a piezo-electric actuator. During the experiment, the molecule was stretched and relaxed multiple times by moving the micropipette relative to the optical trap to generate force-extension curves like the one shown in fig. 6. When mechanically manipulated, RNase H unfolds in a two-state manner (native state $N \rightarrow U$ unfolded state) at ~ 19 pN, and refolds through an intermediate structure (I) at ~ 5.5 pN, which then folds into the native state at lower forces (~ 1 pN). If the $I \rightarrow N$ transition fails, the intermediate is stretched in the next pulling cycle and observed to unfold at ~ 5.5 pN (fig. 6, right

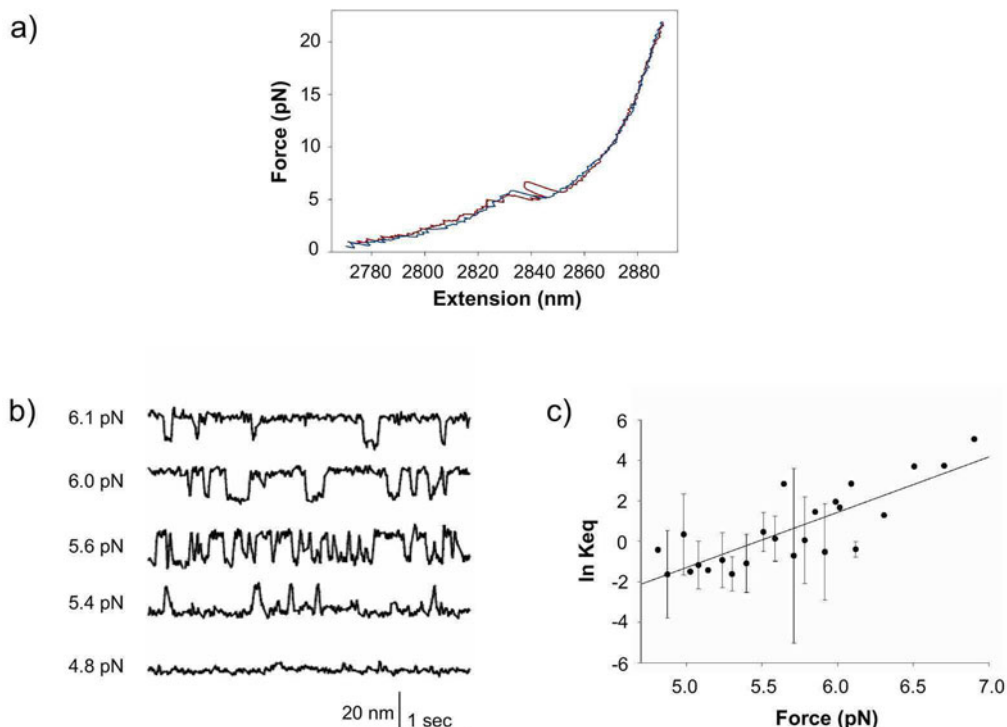


Fig. 7. – Fluctuations between the intermediate and the unfolded states. a) Force extension-curve showing interconversion between the U and I states, adapted from Cecconi *et al.* [14]. b) Extension *vs.* time traces of RNase H at various constant forces, adapted from Cecconi *et al.* [14]. When held at a force near 5.5 pN, RNase H shows bistability and hops between the intermediate and the denaturated form of the protein. By changing the preset force the $U \rightleftharpoons I$ equilibrium can be shifted. c) Equilibrium constant (K_{eq}) of the $U \rightleftharpoons I$ transition at any given force calculated as the ratio of the lifetimes of the U and I states, adapted from Cecconi *et al.* [14].

panel). The fact that the unfolding and refolding traces of I nearly coincide, indicates that the $U \rightleftharpoons I$ transition occurs reversibly under our experimental conditions. In fact, consistent with this observation, when the molecule was manipulated slowly enough (at a speed less than 300 nm/s), the protein was sometimes observed to fluctuate between the unfolded and intermediate states during force-extension cycles (fig. 7a)). In order to investigate these fluctuations in more detail, we fixed the applied force on the molecule to a predetermined value near 5.5 pN using the force-feedback mode of the instrument, and monitored the fluctuations in extension of the molecule as a function of time. Under these experimental conditions, the molecule shows bistability, hopping between the U and I states in a force-dependent manner (fig. 7b)). From the force-dependent rates of unfolding and refolding of the intermediate—calculated as the inverse of the average life times of the I and U states—we estimated the position of the transition state between

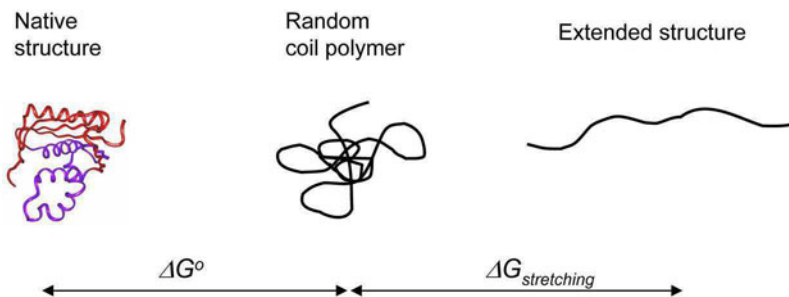
U and I by fitting these rates with the Arrhenius-like equation:

$$(4) \quad k_{I \rightarrow U} = k_m k_{I \rightarrow U}^0 \exp[F \Delta x_{I \rightarrow U}^\ddagger / k_B T],$$

where k_m represents the contribution of the instrument to the absolute rates, $k_{I \rightarrow U}^0$ is the unfolding rate at zero force, F is the force applied on the molecule, $k_B T$ is the thermal energy, and $\Delta x_{I \rightarrow U}^\ddagger$ is the distance from I to the transition state along the reaction coordinate [14, 31]. A similar expression holds true for the reverse $U \rightarrow I$ transition. From the slope of $\ln k$ vs. F plots we estimated $\Delta x_{I \rightarrow U}^\ddagger$ to be 5 ± 1 nm, and $\Delta x_{U \rightarrow I}^\ddagger$ to be 6 ± 1 nm. The distance from I to the transition state is much larger than that found for the mechanical unfolding of any protein native state studied so far ($\Delta x_{N \rightarrow U}^\ddagger$ is typically less than 1 nm), and it is similar to that found for the unfolding of biomolecules whose structure is mostly stabilized by secondary interactions [31]. The large $\Delta x_{I \rightarrow U}^\ddagger$ value, and thus the large mechanical compliance and “stretchiness” of I , could therefore reflect the intermediate state’s lack of those stereospecific tertiary contacts that stabilize native structures and make them brittle under force.

From the lifetimes of the unfolded and intermediate state in constant force-experiments we could also calculate the equilibrium constant of the $U \rightleftharpoons I$ transition at difference forces and then extrapolate to zero force to estimate K_{eq} ($F = 0$) (fig. 7c)). This analysis yielded a K_{eq} ($F = 0$) that corresponds to a $\Delta G_{(UI)}$ of 9 ± 1.5 kcal/mol. The change in free energy associated with the mechanical denaturation of a molecule comprises the free-energy difference between the folded and unfolded states at zero force (ΔG^0) and the free-energy change to stretch the unfolded state (ΔG_{str}) (fig. 8). Assuming the response of a polypeptide chain to force to be well described by the worm-like chain model (WLC) [32], ΔG_{str} can be calculated as the area under the WLC force-extension curve integrated from zero to the extension of the unfolded polypeptide chain; for our system ΔG_{str} was calculated to be 5 ± 1 kcal/mol [14]. After correction for ΔG_{str} , the free-energy change of the $U \rightleftharpoons I$ transition measured in our experiments corresponds well with that measured in bulk studies ($\Delta G_{(UI)(\text{bulk})} = 3.6 \pm 0.1$ kcal/mol).

Two other independent methods were used to evaluate the thermodynamics of the $U \rightleftharpoons I$ transition. In one approach, we determined the potential of mean force of the reversible unfolding of the intermediate as the average area under the unfolding/refolding plateau; this analysis yielded $\Delta G_{(UI)} = 9 \pm 2$ kcal/mol. In the second method, we analyzed the probability of U refolding to I as a function of force using the statistics of a two-state system and obtained a $\Delta G_{(UI)}$ of 8.9 ± 0.25 kcal/mol [14]. These free-energy values, after correction for ΔG_{str} , also correlate well with $\Delta G_{(UI)(\text{bulk})}$ (table I). The remarkable similarity between the free-energy value measured in ensemble experiments and those acquired at the single molecule level suggest that there is a resemblance between the intermediate structures observed using these two techniques. To further examine this similarity, we mechanically manipulated a variant of RNase H (I53D RNase H). I53D RNase H carries a point mutation in the core region of the protein that destabilizes the bulk kinetic intermediate, making the protein fold in a two-state manner [33]. In agreement with what is observed in solution, in our single molecule experiments the I53D



Worm-like chain interpolation formula

$$F = \frac{\kappa_B T}{P} \left[\frac{1}{4} \left(1 - \frac{x}{L_c} \right)^{-2} - \frac{1}{4} + \frac{x}{L_c} \right]$$

Fig. 8. – Correcting for the stretching of the unfolded state. The Gibbs free-energy difference (ΔG) is a thermodynamic function and it thus depends only on the initial and final state of a process. The chemical and mechanical denaturation of a protein are processes that share the same initial state, namely the native structure of the molecule, but have quite different final states. In chemical denaturation experiments the final state is a random coil polymer, whereas in mechanical manipulation studies the final state is a stretched molecule. To compare free energies measured in solution with those obtained in our experiments, the work done to extend the polypeptide chain, $W = \Delta G_{\text{str}}$, must be taken into account. This work can be calculated using the worm-like chain (WLC) interpolation formula: $F = (k_B T/P)(1/(4(1 - x/L)^2) + x/L - 1/4)$, where P , x and L are the persistence length, end-to-end distance and contour length of the molecule, respectively [32]. This formula has been shown to describe well the behavior of a polypeptide chain under tension [34].

variant refolds through a gradual compaction rather than a sharp transition, showing no evidence of an intermediate state (fig. 9). This data further support the idea of identity between the mechanical and bulk intermediates and make our findings relevant to the folding process of RNase H in solution. We could therefore use laser tweezers to answer questions that more traditional techniques had failed to address. The interconversion between the intermediate and the unfolded states during constant force measurements appears to be a first-order process, as indicated by the distributions of the dwell times of the I and U states (fig. 10). These results, together with the sharpness of the $U \rightleftharpoons I$ transitions and the narrowness of the unfolding and refolding force distributions of I (data not shown) suggest that the intermediate that populates the refolding trajectories of RNase H is a well-defined, compact, thermodynamically distinct molecular structure, probably held together mostly by local interactions.

TABLE I. – Comparison between the change in free energies for the $U \rightleftharpoons I$ transition measured in bulk and in single-molecule mechanical studies.

	Bulk experiments	Laser tweezer experiments
$\Delta G_{(UI)}$	3.6 ± 0.1 kcal/mol	3.8 ± 0.82 kcal/mol 4 ± 2.6 kcal/mol 4 ± 2 kcal/mol

The interconversion between the intermediate state and the unfolded state was sometimes observed to stop during constant force measurements (fig. 11). This arrest was always associated with a downward step in the extension *vs.* time trace corresponding to a decrease in the end-to-end distance of the molecule. This molecular compaction well corresponded to that expected for the I to N transition at the given forces. In fact, when the molecule was stretched after hopping had ceased, we always observed a high force event at ~ 19 pN, corresponding to the $N \rightarrow U$ transition, and not the unfolding of the intermediate state at ~ 5.5 pN. Thus, during constant force experiments, we were sometimes able to directly observe the molecule cross the folding barrier to the native state. In about 80% of these cases, the jump into the native structure clearly took place from the intermediate state, as in fig. 11. In the rest of the cases, I might have been too short-lived to be resolved in our measurements. These data indicate that the refolding intermediate observed in our experiments is on-pathway. Moreover, since all our force-extension curves display the low-force $U \rightarrow I$ transition, the same intermediate also appears to be an obligatory step in the refolding trajectories of RNase H.

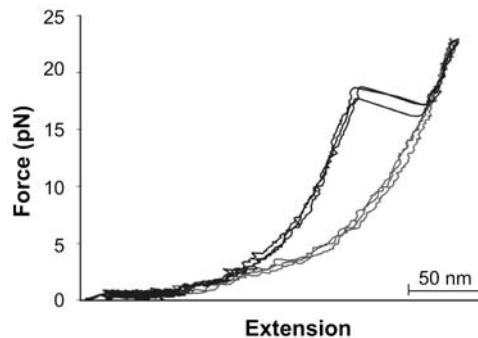


Fig. 9. – Force-extension curves for the I53D RNase H variant, from Cecconi *et al.* [14]. This mutant refolds gradually, showing no low force refolding or unfolding transitions (compare behavior at ~ 5.5 pN to that shown in fig. 6).

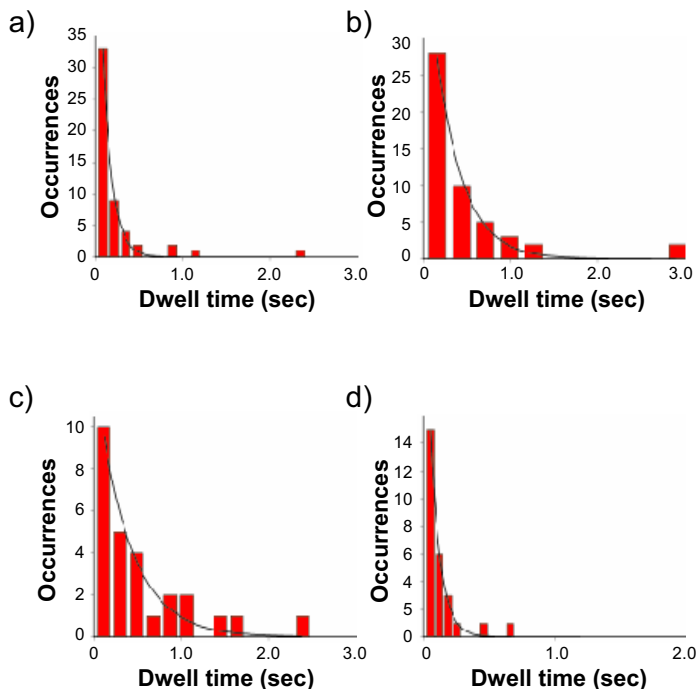


Fig. 10. – Dwell time distributions of the unfolded and intermediate states of RNase H during constant force measurements, adapted from Cecconi *et al.* [14]. The distributions are well fit by single exponentials, revealing the first-order nature of the $U \rightleftharpoons I$ transition. From these fits we estimated the unfolding and refolding rate constants of the intermediate. At force equal to 5.1 pN, the $U \rightleftharpoons I$ equilibrium is shifted toward I with $k_f = 9.1 \pm 0.5 \text{ s}^{-1}$ and $k_u = 3.27 \pm 0.24 \text{ s}^{-1}$, panel a) and b), respectively. Instead, at 5.9 pN the unfolded state predominates and $k_f = 2.6 \pm 0.35 \text{ s}^{-1}$ and $k_u = 13.1 \pm 0.68 \text{ s}^{-1}$, panel c) and d), respectively.

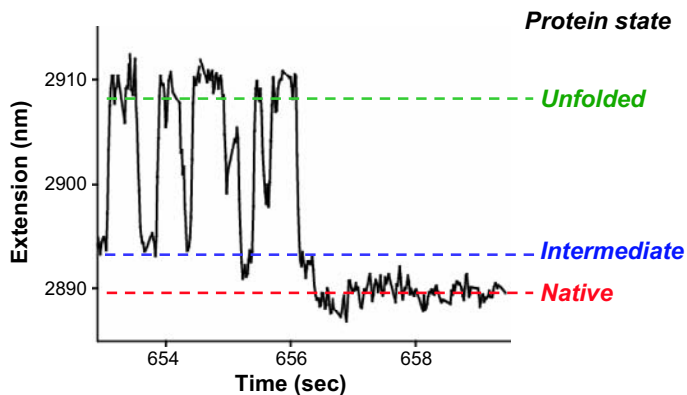


Fig. 11. – Extension *vs.* time traces of RNase H, adapted from Cecconi *et al.* [14]. The molecule is directly observed to fold into the native structure from the intermediate state.

5. – Conclusion

Protein folding is thought to be a very heterogeneous process where individual molecules fold into their native state through many different pathways on a multidimensional funnel-shaped energy landscape [3]. The experimental characterization of such a multitude of folding routes requires measurements at the single-molecule level, since ensemble measurements provide only average information about the folding process. In this paper we describe an experimental approach to study protein folding at single-molecule level. This method relies on the use of molecular handles to mechanically manipulate individual globular proteins in the low force regime of optical tweezers. In these experiments, single molecules are stretched and relaxed multiple times and their behavior under tension is described as a function of their molecular end-to-end distance. The low spring constant of the laser trap allows the manipulation of the molecules at close to equilibrium conditions and the direct monitoring of fluctuations between different molecular conformations. We applied this experimental approach to study the folding mechanism of RNase H. We characterized the complete mechanical unfolding and refolding trajectories of individual RNase H molecules and identified a refolding intermediate (*I*) that correlates with the conformation sampled in ensemble studies, which has been postulated to be a molten globule. Unlike in bulk experiments where *I* forms within the dead time of the measuring instruments, in our measurements the interconversion between the intermediate and the unfolded states could be monitored and visualized in real time, allowing us to characterize directly the kinetic, thermodynamic and mechanical properties of this molten globule. We demonstrated that the intermediate that populates the folding trajectories of RNase H is a compact, thermodynamically distinct state, which is both obligatory and on-pathway to the native structure. Moreover, we showed that under tension the intermediate exhibits unusual mechanical properties unfolding at low forces (only a few piconewtons) and displaying great compliance, deforming by 5 nm before unfolding, a property typical of structures stabilized by weak, non stereospecific tertiary contacts [31].

The data presented in Cecconi *et al.* and discussed in this paper demonstrate the power of direct mechanical manipulation to access processes that are difficult or impossible to examine with more traditional experimental approaches. By slowing down the folding/unfolding process of RNase H with force we were able to monitor and characterize an intermediate whose signal was concealed by time- and ensemble-averaging, and so could not to be directly observed in bulk experiments. Most significantly, we were able to characterize the mechanical properties of a molten globule structure. The experimental method presented here represents a completely new approach to single molecule manipulation studies and sets the stage for many new future experiments.

REFERENCES

- [1] COHEN F. E. and KELLY J. W., *Nature*, **426** (2003) 905.
- [2] DOBSON C. M., *Nature*, **426** (2003) 884.

- [3] BALDWIN R. L., *J. Biomolec. NMR*, **5** (1995) 103.
- [4] FERSHT A. R., *Structure and Mechanism in Protein Science: a Guide to Enzyme Catalysis and Protein Folding* (W.H. Freeman, New York) 1999.
- [5] STRICK T. *et al.*, *Phys. Today* **54**, issue no. 10 (2001) 46.
- [6] RIEF M. *et al.*, *Science*, **276** (1997) 1109.
- [7] ROUNSEVELL R., FORMAN J. R. and CLARKE J., *Methods*, **34** (2004) 100.
- [8] YANG G. L. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **97** (2000) 139.
- [9] BROCKWELL D. J. *et al.*, *Nat. Struct. Biol.*, **10** (2003) 731.
- [10] CARRION-VAZQUEZ M. *et al.*, *Nat. Struct. Biol.*, **10** (2003) 738.
- [11] CARRION-VAZQUEZ M. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **96** (1999) 3694.
- [12] FERNANDEZ J. M. and LI H. B., *Science*, **303** (2004) 1674.
- [13] LEE G. *et al.*, *Nature*, **440** (2006) 246.
- [14] CECCONI C. *et al.*, *Science*, **309** (2005) 2057.
- [15] SMITH S. B., CUI Y. J. and BUSTAMANTE C., Optical-trap force transducer that operates by direct measurement of light momentum, *Methods in Enzymology*, vol. **361** (2003) pp. 134-162.
- [16] ASHKIN A. *et al.*, *Opt. Lett.*, **11** (1986) 288.
- [17] GRIER D. G., *Nature*, **424** (2003) 810.
- [18] SVOBODA K. and BLOCK S. M., *Annu. Rev. Biophys. Biomol. Struct.*, **23** (1994) 247.
- [19] GORDON J. P., *Phys. Rev. A*, **8** (1973) 14.
- [20] SMITH S. B., *Stretch Transitions Observed in Single Biopolymer Molecules (DNA and Protein) using Laser Tweezers*, Doctoral Thesis, University of Twente, The Netherlands, (1998).
- [21] CECCONI C. *et al.*, Protein-DNA chimeras for single molecule mechanical folding studies with the optical tweezers, in preparation.
- [22] GRASSETTI D. R. and MURRAY J. F. JR., *Arch. Biochem. Biophys.*, **119** (1967) 41.
- [23] PEDERSEN A. O. and JACOBSEN J., *Eur. J. Biochem.*, **106** (1980) 291.
- [24] RIENER C. K., KADA G. and GRUBER H. J., *Anal. Bioanal. Chem.*, **373** (2002) 266.
- [25] BLACK C. B. and COWAN J. A., *Inorg. Chem.*, **33** (1994) 5805.
- [26] YANG W. *et al.*, *Science*, **249** (1990) 1398.
- [27] DABORA J. M. and MARQUESEE S., *Protein Sci.*, **3** (1994) 1401.
- [28] RASCHKE T. M., KHO J. and MARQUESEE S., *Nature Struct. Biol.*, **6** (1999) 825.
- [29] RASCHKE T. M. and MARQUESEE S., *Nature Struct. Biol.*, **4** (1997) 298.
- [30] PARKER M. J. and MARQUESEE S., *J. Mol. Biol.*, **293** (1999) 1195.
- [31] LIPHARDT J. *et al.*, *Science*, **292** (2001) 733.
- [32] BUSTAMANTE C. *et al.*, *Science*, **265** (1994) 1599.
- [33] SPUDICH G. M., MILLER E. J. and MARQUESEE S., *J. Mol. Biol.*, **335** (2004) 609.
- [34] KELLERMAYER M. S. Z. *et al.*, *Science*, **276** (1997) 1112.

Oligarchy in protein folding: The upper and lower classes in protein chains

G. TIANA, D. PROVASI, A. AMATORI and L. SUTTO

Dipartimento di Fisica, Università di Milano and INFN, Sezione di Milano - Milano, Italy

R. A. BROGLIA

Dipartimento di Fisica, Università di Milano and INFN, Sezione di Milano - Milano, Italy

The Niels Bohr Institute, University of Copenhagen - Copenhagen, Denmark

1. – History of the hierarchical view of folding

Proteins are polymers composed of amino acids displaying complex and heterogeneous interactions. Amino acids tend to form the best possible network of hydrogen bonds, and at the same time locate hydrophobic residues in the interior, align their dipole moments, prevent residues with the same charge to come together, optimize Van der Waals interactions, etc. Most of the time these are contradictory requirements and no sequence of amino acids can optimize all of the interactions at the same time. In other words, even in their lowest energy states, proteins are not able to optimize locally all the interactions.

A protein is then a *frustrated* system [1]. As a consequence, one would expect it to display the typical features of such systems: an equilibrium state populated by many dissimilar conformations (at temperatures comparable with the typical interaction energy between amino acids, *i.e.* $kT = 2.4\text{ kJ/mol}$), rugged energy landscapes, strong dependence on the initial conditions of its folding, lengthy equilibration times, and so on. Actually, proteins do exactly the opposite. They display a low-entropy, essentially unique equilibrium conformation in a wide range of temperatures (typically from 0 to 60°C), separated by a sharp, first-order-like transition from high-energy, disordered conformations. They are able to reach the native state in times which go from microseconds to seconds, displaying usually a Poissonian kinetics.

The reason why a protein does not display the features of frustrated system has long been clarified: evolution selects sequences to minimize the overall frustration of the protein [2]. In spite of this, explaining why proteins display antithetic properties to those displayed usually by frustrated systems is not straightforward. In what follows we shall provide evidence that the peculiar properties displayed by proteins are connected with the fact that the stabilization energy of proteins is not distributed uniformly among the various amino acids, but is concentrated in few, mutually interacting fragments. Anyway, this idea has roots far in the past.

It was Cyrus Levinthal who, in 1969, first introduced the concept of nucleation in connection with protein folding. Explaining how proteins can avoid a random, time-consuming search in conformational space, he concluded that *protein folding is speeded and guided by the rapid formation of local interactions which then determine the further folding of the peptide. This suggests local amino acid sequences which form stable interactions and serve as nucleation points in the folding process* [3].

In 1973 Wetlaufer started to talk about “structural regions”, that is *sections of peptide chain that can be enclosed in a compact volume* [4]. Analyzing ten proteins of known structure, he was able to divide them in structural regions of 40–150 residues, mostly continuous along the peptidic chain. He suggested that *chain continuity is an important feature because it is just what one would expect of a process of nucleation followed by rapid growth*. In his view, the folding starts from a marginally stable nucleus of 8–18 consecutive residues and rapidly extends to the chain-wise nearby structural regions.

In 1981 Rose and Lesk extend Wetlaufer’s idea, introducing the concept of hierarchical folding [5]. They identify in myoglobin and ribonuclease a hierarchy where structural regions of larger length scale include structural regions of smaller length scale, and suggest that this structural hierarchy reflects the hierarchy of events during the folding process. The hierarchical organization is what makes the folding process fast. It is interesting from the historical point of view to note that all these seminal works are sons of their time, that is the early years of protein structure investigations. Thus, they are strictly connected with the structural properties of proteins.

In the nineties physicists started to cope with the protein folding problem. In 1994 Shakhnovich used a lattice model to show that the formation of a nucleus built out of less than 20% of the residues is a necessary and sufficient condition to reach the native conformation starting from an elongated chain [6]. This postcritical nucleus is defined as *the minimal sized fragment of the new phase (i.e., the native state) which inevitably grows further to the new state*. It is called postcritical because the term “critical nucleus” is reserved for the structural features which characterize the top of the main free energy barrier which separate the native from the denatured state, from where the protein has 1/2 probability to fold and 1/2 probability to fall back into the denatured state. Consequently, the postcritical nucleus has a lower free energy than the critical nucleus, on the side of the native state. To be precise, a structural definition of postcritical nucleus is inconsistent: as emphasized by Lifshits and Pitaevskii, in the kinetics of first-order transitions there is not a critical point, but a critical region of size $\delta = (T/\alpha)^{1/2}$, where α is the work associated with the formation of the surface of the nucleus [7].

Again departing from a structural view towards an energy-landscape theory of protein folding, Wolynes defined “foldons” as those fragments of the protein whose folding temperature is much higher than their glass transition temperature, that is whose native state energy lies well below that of the competing conformations [8]. He divided thirty proteins into foldons of average size 38 residues which, interestingly, well overlap with exon regions of DNA. Although the definition of foldons is based on equilibrium thermodynamics, the consequences on kinetics are straightforward: *different parts of proteins may fold at different times quasi-independently, provided that the interactions guiding the protein to its folded structure rather than to other misfolded configurations are largely contained in discrete contiguous parts of the sequences (i.e., foldons).*

The article “Is protein folding hierarchic?” by Baldwin and Rose summarizes the experimental evidences in favour of a picture where *folding begins with structures that are local in sequence and marginal in stability; these local structures interact to produce intermediates of ever-increasing complexity and grows, ultimately, into the native conformation* [9]. The most evident argument for a hierarchical scenario are folding intermediates, where the structural features of some of the steps of the hierarchy can be detected experimentally. But not all proteins display intermediates. Nonetheless, a number of experiments described by Baldwin and Rose suggest that the hierarchic mechanism acts also for these proteins, just the intermediates are less stable. In other words, the formation of local structure does not correspond to a metastable state in the overall free-energy landscape, and consequently it can hardly be detected by standard experimental techniques.

From a physical point of view, the hierarchy of events in folding seems to reflect the hierarchy in which stabilization energy is distributed among the different amino acids of proteins. A quick look at the effect point mutations have on the stability of the native state of proteins suffices to realize that only few amino acids are critical to stabilize the protein [10-12]. To go more in depth into the problem, one can use semiempirical force fields to obtain the contribution of the amino acids to the overall stability. In fig. 1 the stabilization energies per residue for three small proteins are displayed, namely Protein G, the SH3 domain of Src and CI2. Since available force fields are usually not completely reliable, we have calculated such energies making use of three independent methods. First, the interaction energy of each amino acid is calculated starting from the experimental destabilization effects $\Delta\Delta G_{UN}$ of pointlike mutations, solving a coupled set of linear equations (solid curve in eq. (1)), as described in ref. [13]. Another method is to use directly the FOLD-X force field [14] and the associated results are displayed as a dashed curve. A third method is to average the atomic contributions of the Gromos force field (dotted curve). Following this method, which has provided good results in comparing with the experimental data [15], one obtains in such a way a residue-based energy but disregards the effective contribution of the solvent. Although the three curves are not identical for each protein, all of them identify some regions where a large amount of stabilization energy is consistently concentrated (*e.g.*, residues 3-8 and 50-55 in Protein G; 6-8, 18-28, 37-42 and 47-51 in SH3; 47-52 in CI2).

Calculations performed with minimal lattice models confirm that evolution concentrates the stabilization energy in a small fraction of amino acids in order to produce

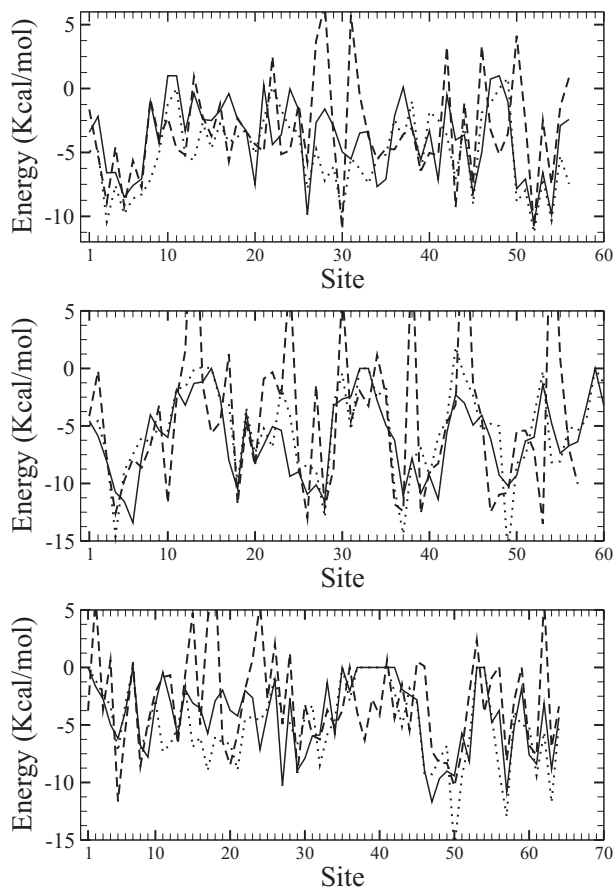


Fig. 1. – The stabilization energy per residue for Protein G (upper panel), Src-SH3 (middle panel) and CI2 (lower panel). The solid curves indicate the energy calculated from the experimental values of $\Delta\Delta G_{UN}$, the dotted curves are obtained averaging the Gromos force field, while the dashed curves show the FOLD-X energy contributions.

stable and fast folding proteins [16-18] and that this is responsible for the hierarchical mechanism which allows the protein to fold fast [19, 20] (see also the contribution of Broglia *et al.* to this volume). In brief, the fragments of the protein built out of consecutive, strongly interacting residues are formed very early in the folding process, as a result of the search on only a limited volume of the conformational space of the protein, search which is consistently biased towards the native state by the attractive interactions between these amino acids. Such fragments are called local elementary structures (LES) [21]. The rate-limiting step in the folding process is the assembly of LES into the folding nucleus, a step where most of the conformational entropy of the chain is squeezed out. Just after this, the entropy of the protein is so reduced that the remaining amino acids can found their native position essentially instantly (for a review see ref. [20]).

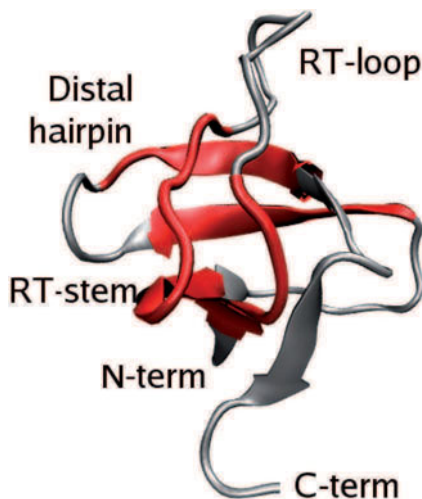


Fig. 2. – The native structure of the SH3 domain, composed of 60 amino acids. In the figure the regions 6–8, 18–28, 37–42 and 47–51, where most of the stabilization energy is concentrated, are highlighted.

2. – The folding of the SH3 domain: A computational model

In order to elucidate in detail the hierarchical nature of the folding process, one needs a test protein and a model to describe it. The test protein we have chosen is the SH3 domain of Src (displayed in fig. 2), because it is one of the smallest stable proteins (60 residues) and because it has been widely characterized experimentally [10].

The model needed to investigate the folding of SH3 should be a trade-off between an appropriate description of the protein and its computational lightness. The most realistic models are those which describe explicitly the electronic density, meaning that one has to solve, under some approximations, Schrödinger's equation (for example, within Density Functional Theory). This is indeed the most detailed description one can give of a protein, but it is also the most expensive from a computational point of view (the associated computational algorithms scale as the cube of the number of electrons). This approach can be used to study small parts of the protein, such as the catalytic site of an enzyme, which involves few tens of atoms, but is inappropriate for the overall mechanism of protein folding (see, for example, the contribution by Paolo Carloni to this volume). Simpler models make use of classical force fields which depend only on the spatial coordinates of the atoms which build out the protein and of the surrounding solvent (typically, tens of thousands of atoms). These models can be used to study processes which take place on time scales of up to microseconds, such as unfolding of small proteins, partial conformational changes of larger proteins, but not protein folding (see, for example, the contribution of Wilfred van Gunsteren *et al.*). The most extensive effort done in this respect has been a single folding trajectory of the 36-residues villin

headpiece [22], corresponding to a calculation of several months on a massive parallel computer. Note however that a single folding trajectory is not statistically representative of the folding process of a protein. Consequently, with only this knowledge it is not possible to say something of relevance about the underlying kinetics. The situation is even worse concerning equilibrium thermodynamics. To describe protein equilibrium, the simulation has to fold and unfold the protein several times, in which case equilibrium quantities can be calculated by means of the ergodic theorem. Although non-Newtonian dynamic algorithms will be useful in this respect, the possibility of describing protein equilibrium is still out of reach.

An alternative approach, which we have used in the following, is to employ simplified models, which retain only the basic ingredients which are thought to allow proteins to fold fast to a unique and stable equilibrium state. These ingredients are the heterogeneity of the interactions among amino acids and the polymeric character of the protein [23]. Of course there exists no model for all seasons, and consequently this will only be useful to describe the general mechanism of protein folding, not other aspects of proteins such as enzymatic activity, post-translational modifications, etc.

The model describes a protein as an inextensible chain, where amino acids are represented by spherical beads centered around the C_α atom, thus allowing a realistic accounting for the protein backbone. The potential energy of the protein depends on the positions $\{r_i\}$ and sequence $\{\sigma_i\}$ of amino acids according to

$$(1) \quad U(\{r_i\}, \{\sigma_i\}) = \sum_{i < j} B(\sigma_i, \sigma_j) \theta(R(\sigma_i, \sigma_j) - |r_i - r_j|),$$

where θ is Heaviside's step function, $R(\sigma_i, \sigma_j)$ is the range of interaction which depends on the kind of amino acids, and $B(\sigma_i, \sigma_j)$ is the interaction energy between an amino acid of kind σ_i and one of kind σ_j . There is also a hard-core repulsion which depends on the kind of residue (for details see ref. [24, 25]).

Although a number of realizations of the matrix B are available in the literature, none of them have provided satisfactory results in folding proteins. To overcome this problem, we have adopted the maximum degree of generality: since we think that it is the heterogeneity of the interaction and not its detailed features to make proteins fold, we have chosen a random-generated interaction matrix B . Of course, evolution has used a different potential function to select proteins to fold to the structures we know by X-ray crystallography or by nuclear magnetic resonance. Consequently, in order to study the folding of model proteins, we have also to find good folder sequences within the framework of the model, by reproducing evolution in the computer.

To reproduce the evolution of SH3, we assume that the fitness of a sequence is simply its ability to fold and be stable on the SH3 conformation. The elementary move of evolution is a swap between two amino acids, keeping fixed the conformation of SH3, and the move is accepted or not based on a Metropolis energy criterion [26, 17]. Folding sequences are those which display an overall low energy, in particular lower than a sequence-independent threshold E_c . Examples of sequences displaying different native

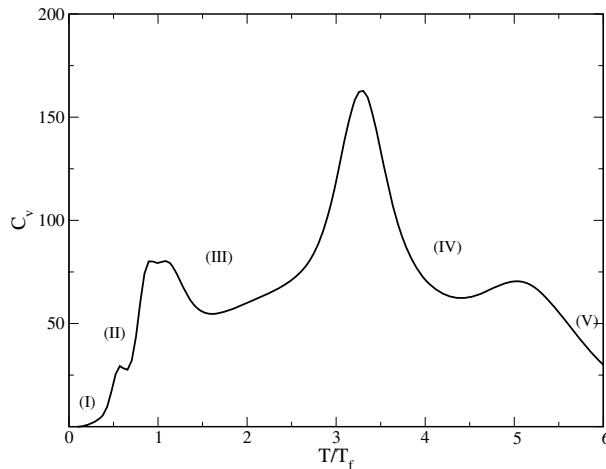


Fig. 3. – The specific heat C_v of the best sequence designed to fold to the SH3 conformation.

energies are given in table I of ref. [24]. The quantity E_c , expressed in units of the standard deviation σ_B of the interaction matrix, results in the value $-65\sigma_B$. The lowest-energy sequence found has energy $-73\sigma_B$ and folds within an dRMSD of $3.1\text{ \AA}$ to the crystallographic conformation. The quality of the folding worsen as the energy increases: sequences at E_c always display dRMSD above $6\text{ \AA}$ and are thus bad folders.

3. – Equilibrium thermodynamics

The main quality of a simplified protein model as that described above is that it is relatively easy to sample its conformational space and, by means of multicanonical algorithms [27], obtain its equilibrium properties. The specific heat of the best folding sequence is displayed in fig. 3.

The peaks in the specific heat mark sharp, first-order-like⁽¹⁾ transitions between different phases. Starting from high temperature, the phases labelled as V and IV in fig. 3 correspond to unfolded, coil conformations (the radius of gyration is larger than $15\text{ \AA}$, whereas that of the native state is $10\text{ \AA}$). These temperatures are too high to be observed experimentally. The highest C_v peak marks the transition to an unfolded globular conformation (labelled as III; the radius of gyration is about $12\text{ \AA}$, while the dRMSD is about $6\text{ \AA}$). The peak between states III and II corresponds to the folding transition (T_f) observed experimentally. The state labelled as II (dRMSD about $3\text{ \AA}$)

⁽¹⁾ A phase transition is properly defined only in an infinite system. A protein is a finite and also quite small system, and consequently fluctuations are not negligible and broaden consistently the peaks in C_v . Nonetheless, in the following we will use the term “phase transition” in an extended way also concerning finite systems.

TABLE I. – *The equilibrium population of two native structures of SH3 (cf. fig. 2) for the different thermodynamic states of SH3.*

State	RT stem (6–10)–(18–23)	distal hairpin (38–50)
native (II)	1	1
unfolded globule (III)	0.8	1
unfolded coil (IV)	0.25	0.25
unfolded coil (V)	0.08	0.04

is the equilibrium native state, while state I is a frozen native state, lying at too low temperatures to be observed in solution experiments.

One of the most interesting features of the thermodynamic states described above is that the presence of native structures is not limited to the native state (II). In table I are listed the equilibrium formation probabilities of the RT stem, that is the structure built out of two strands which connects the RT loop to the rest of the protein (cf. fig. 2), and of the distal hairpin (cf. fig. 2). These native structures are not formed only in the native state, as expected, but are consistently detectable also in the unfolded states (III) and (IV). As a matter of fact, these structures are consistent with Wolynes’ definition of foldons although in our framework foldons do not exhaustively cover the entire protein sequence. Moreover, the RT stem and the distal hairpin are in contact in the native conformation. The formation of these contacts happens to take place exactly at the folding temperature. Consequently, the interaction within and across these two structures form the core which stabilizes the native state of the protein.

Another interesting feature which emerges from this protein model is that the remarkable stability of RT stem and distal hairpin is not present only in the best folding sequence designed on the SH3 conformation, but in all folding sequences (the formation probability of the distal hairpin ranges between 0.93 and 1, while that of the RT stem between 0.80 and 1). Even more interestingly, they are populated also in the bad folding sequences whose ground-state energy is comparable with E_c . For example, in the case of sequence called s_7 in ref. [24] the formation probability of the distal hairpin is 0.38 and that of the RT stem is 0.32. Summing up, the protein displays few structures which are stabilized by local contacts and which are conserved not only at high temperatures, but also over sequences selected (even mildly) by evolution.

4. – Folding kinetics

With the simplified model it is possible also to repeat a large number of folding kinetics trajectories, starting from a random-generated conformation, and following the typical sequence of events to the native state. We have repeated 25 folding runs at the temperature $T = 0.8T_f$, below the folding temperature. The average dRMSD as a function of time is displayed in fig. 4 and displays a double exponential behaviour, with characteristic times $4.8 \cdot 10^{-9}$ s and $3.3 \cdot 10^{-6}$ s, respectively. The former corresponds

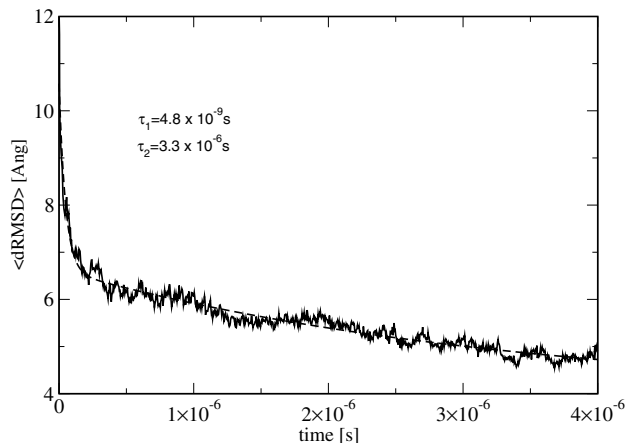


Fig. 4. – The mean dRMSD as a function of time.

to the compaction of the chain from the elongated initial conformation to an unfolded globule, a process too fast to be detected experimentally. The slower process corresponds to the folding to the native state and matches with the single exponential usually detected experimentally.

The model also allows to study the formation probability of single contacts as a function of time, something which is quite difficult to be done experimentally. Figure 5 shows examples of this quantity for some contacts of SH3. There are three typical kinds of behaviour for the formation of native contacts: 1) fast (compared to the overall folding time, *i.e.* μ s), exponential and stable, 2) slower, exponential and stable, 3) even slower and non-exponential. The formation probability of contacts between residues 40–44 and

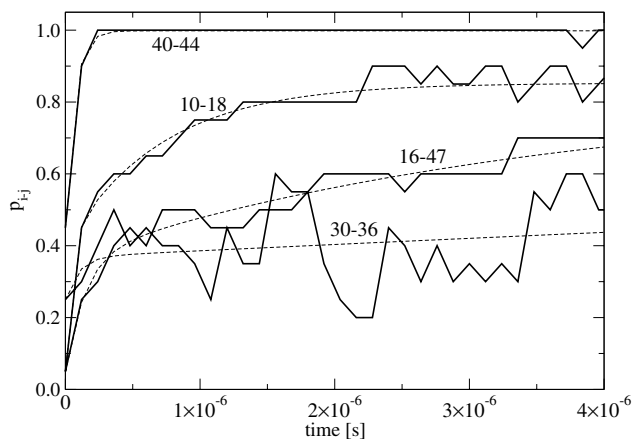


Fig. 5. – The formation probability of some selected native contacts as a function of time.

TABLE II. – *The characteristic times τ_i associated with single- or double-exponential fit of the formation probability of selected native contacts. The last column indicate the correlation coefficient r of the fit.*

Contact	τ_1	τ_2	r
40–44	$6.7 \cdot 10^{-8}$ s		0.992
10–18	$7.0 \cdot 10^{-7}$ s	$3.7 \cdot 10^{-7}$ s	0.966
16–47	$3.6 \cdot 10^{-6}$ s	$1.4 \cdot 10^{-7}$ s	0.969
30–36	$4.1 \cdot 10^{-5}$ s	$9.8 \cdot 10^{-8}$ s	0.320

10–18 is fast and stable (cf. fig. 5), and can be fitted with a single exponential and with a double exponential, respectively (the associated characteristic times and correlation coefficients are listed in table II). A single exponential for the formation probability p_{ij} of the native contact between residues i and j can arise for a number of mechanisms. The simplest is a transition between two states, namely from not bound to bound, following the master equation

$$(2) \quad \frac{\partial p_{ij}}{\partial t} = k_{\text{in}}(1 - p_{ij}) - k_{\text{out}}p_{ij},$$

provided that the rates k_{in} and k_{out} are independent of time (or depend on time on a time scale which is completely different from the time scale of the process under consideration). A double exponential arises if there are three possible states. Comparison with fig. 4 suggests that these states are: an extended non-bound state, a compact non-bound state, a bound state. The time independence of the formation and disruption rate indicates that the formation of these contacts is spontaneous and does not depend on any other process. The structures built out of residues 6–10 and 18–23, where all contacts behave as 40–44 and 10–18, are then “local elementary structures” along the folding pathway (cf. refs. [20] and [19]). In this simple model the stability is given by $k_{\text{in}}/(k_{\text{in}} + k_{\text{out}})$, while the characteristic time is $\tau = (k_{\text{in}} + k_{\text{out}})^{-1}$. For contacts of the type 40–44 or 10–18, k_{in} is large because the two residues are close along the sequence, and thus they do not have to explore a large conformational space to find a bound state. On the other hand, k_{out} is small because the two residues are strongly attractive, resulting in a small overall formation time τ and high stability.

Another class of contacts is represented by that between residues 16 and 47, which, by the way, belong to the local elementary structures discussed above. The formation of this contact is again exponential, but it is formed on a longer time scale, comparable with the overall folding time. The fact that the dynamics is exponential implies that it involves again few states, being insensitive to kinetic traps. Its large stability cannot be explained by a large value of k_{in} , because this time the two monomers are far along the chain, and then need to explore a larger fraction of conformational space than before. This fraction is anyway smaller than that associated with a homopolymeric chain because

the early formation of stable local elementary structures reduces effectively the length of the chain. The stability of the contact is then assured by the very small value of k_{out} , due to the fact that the interaction between residues 16 and 47 is enhanced by the favourable interaction between the whole local elementary structures they belong to. These contacts build the folding nucleus.

Finally, the dynamics of residues which do not belong to local elementary structures is non-exponential, because their formation strongly depends on the state of the folding nucleus, and consequently k_{in} and k_{out} depend explicitly on time. The interaction between these residues is too weak, sometimes even repulsive, to form stable structures on their own. Only the formation of the folding nucleus reduces their conformational entropy to such an extent that they can be stably in their native position.

Summing up, there is a three-level hierarchy in the folding of SH3. The first level involves some fragments of strongly attractive residues which get structured very early. These fragments are quite apart along the chain. The second level corresponds to the assembly together of these fragments into the postcritical folding nucleus, on a time which almost coincides with the folding time. The third level is the reaching by the protein of the native state, when the rest of the residues fall into their native position.

5. – The uneven distribution of energy

We have seen that the stabilization core of the protein, involving a limited subset of residues, is responsible for its remarkable stability. On the other hand, the postcritical folding nucleus, again composed of few residues, is responsible for the fast folding kinetics. In the case of SH3 the stabilization core and the postcritical folding nucleus almost coincide (keep also in mind the non-exactness of the definition of postcritical folding nucleus). In principle, they have not to coincide pointwise, as highlighted by Dokholyan and coworkers studying Protein A [28]. However, one can identify in the distribution of interaction energy among the amino acids in the native conformation a common factor which controls the position of both the stabilization core and the postcritical folding nucleus. In fact, it is straightforward that the residues where most of the interaction energy is concentrated are critical for the stability of the protein, and thus serve as stabilization core. But they are important also from the kinetical point of view. It is unavoidable that some of these residues lie close along the chain, and this gives rise to the local elementary structures, as described in the previous section. Thus, the stabilization core and the postcritical folding nucleus, although being different entities, are tightly connected.

In fig. 6 the contribution of each residue to the total energy of the native conformation is shown, that is

$$(3) \quad E_i \equiv \sum_j B(\sigma_i, \sigma_j) \theta(R(\sigma_i, \sigma_j) - |r_i - r_j|).$$

It is clear that the stabilization core and the postcritical folding nucleus involve the mutually interacting regions where the stabilization energy is concentrated. There are

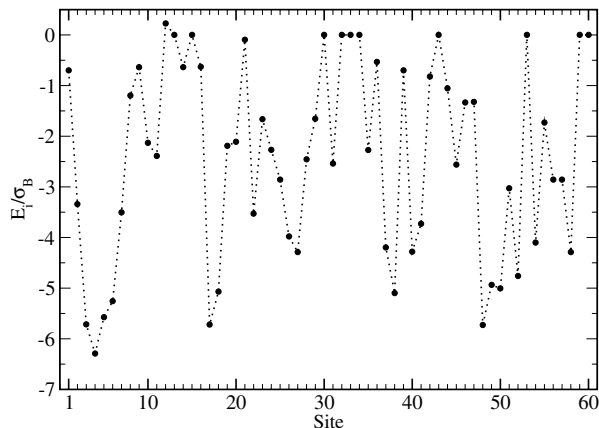


Fig. 6. – The energy per residue in the native state of SH3.

certainly strongly interacting residues which have little importance for the stability or for the folding kinetics, but these are the isolated ones, put aside from the network of interactions which controls the protein folding.

6. – Inhibition of the folding of SH3

It was shown [29-31] that a strategy to inhibit the biological activity of proteins is to avoid that they populate their native state. This can be done with peptides displaying the same sequence as fragments of the protein itself and thus acting as trojan horses. At the light of the model results discussed above, this can be done following two strategies which, in principle, are different:

- 1) Using as inhibitors peptides with the same sequence as foldons, in order to shift the thermodynamic equilibrium towards the unfolded state. The peptide binds to the rest of the stabilization core mimicking the same interactions which stabilize the whole protein. If the concentration of peptides is large enough, the entropy loss due to the binding is smaller than that associated with that necessary to populate the native state, and the equilibrium is shifted towards a state where the protein is partially unfolded and the peptide bound to it. This can be done notwithstanding the initial state of the protein, because the peptide takes advantage of the thermal fluctuation of the protein.
- 2) Using peptides mimicking local elementary structures, thus interfering with the folding kinetics. When the protein exits the ribosome and is still in its unfolded state, the peptide binds to its complementary local elementary structures preventing the nucleus from being formed and then the protein from reaching its native state. This unfolded state is necessarily long lasting, because the peptide has a strong affinity for the protein. Whether or not the unfolded state is also the long-time equilibrium state depends on the concentration of peptide.

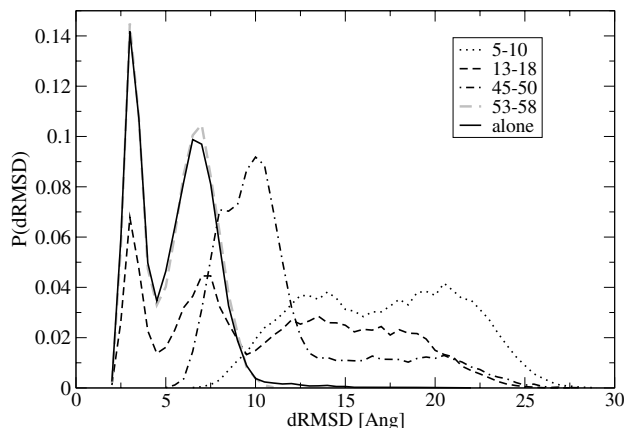


Fig. 7. – The equilibrium distribution of the dRMSD of SH3 in the presence of peptides with sequence identical to that of fragments of the protein itself.

Of course, since the stabilization core is very similar to the postcritical folding nucleus and the foldons forming the stabilization core are essentially identical to the local elementary structures, both mechanisms will act simultaneously. The main caveat is the stability of the protein. If the protein displays strong thermal fluctuations, interfering with its kinetics is identical to interfering with its equilibrium. But if the protein is very stable, the time needed by the peptide to enter the native conformation and shift the equilibrium to the unfolded state can be very long. In this case the only mechanism which could inhibit the protein on a reasonable time scale is to block it while it is folding.

In fig. 7 the equilibrium distribution of the dRMSD is displayed for the protein alone and for a system composed of the protein and three peptides with the same sequence of some fragments of the protein⁽²⁾. These calculations have been performed at $T = 1.05T_f$, that is at a temperature which is quite high with respect to the biological temperature, starting from the native conformation. As expected, the peptides corresponding to the foldons which build out the stabilization core are able to shift the thermodynamic equilibrium to the unfolded state. The population of the native state in the presence of peptides of kind $i - j$, defined as

$$(4) \quad p_N^{i-j} = \int_0^6 d(\text{dRMSD}) p(\text{dRMSD}),$$

and normalized to the population of the native state p_N^0 of the protein alone is displayed in fig. 8.

⁽²⁾ For computational convenience, in these calculations we have neglected the non-native interactions, studying thus a Gō-model.

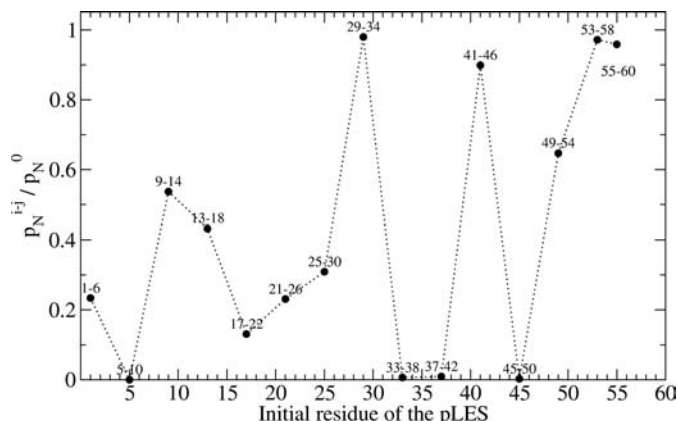


Fig. 8. – The normalized native state population for peptides whose sequence cover exhaustively the protein.

If we repeat the calculations done, for example, with peptide 45–50 starting from the unfolded state at $T = 1.05T_f$, the resulting relative population of the native state is identical to that obtained from the simulation started from the native state (*i.e.* 0.04). This means that the inhibition mechanism is the former of the two described above, that is that under equilibrium conditions. Decreasing the temperature to $T = 0.8T_f$ the simulation starting from the unfolded state still provides a high degree of inhibition (*i.e.* $p_N^{45-50}/p_N^0 = 0.14$), while that starting from the native conformation and lasting for $4 \cdot 10^{-4}$ s is weakly affected by the peptide (*i.e.* $p_N^{45-50}/p_N^0 = 0.89$). Consequently, at this temperature the leading mechanism is the latter, that is the kinetic one. Of course, from a practical point of view, it is immaterial which mechanism is acting, because the peptides which act following the former act also following the latter.

7. – Conclusions

The uneven distribution of energy among the amino acids is a key feature given by evolution to SH3. The fragments of the protein which contain most of this energy are responsible for both its thermodynamic stability and its fast folding kinetics. It is possible to exploit these features to inhibit the folding, and thus the activity, of SH3. Peptides with the same sequence of one of these fragments interfere with the folding kinetics (mostly at low temperature) or with the thermodynamic stability of the native state (mostly at high temperature). Since they are consequence of some general physical features of proteins, it is likely that the results obtained for SH3 can be reproduced for most single-domain proteins, as well as for each of the domains of multidomain proteins.

REFERENCES

- [1] TOULOUSE G., *Commun. Phys.*, **2** (1977) 115.
- [2] FRAUENFELDER H. and WOLYNES P. G., *Phys. Today*, **47** (1994) 58.
- [3] LEVINTHAL C., *Mossbauer Spectroscopy in Biological Systems*, in *Proceedings of a meeting held at Allerton House, Monticello, Illinois*, edited by DEBRUNNER J. T. P. and MUNCK E. (University of Illinois Press) 1969, pp. 22-24.
- [4] WETLAUFER D. B., *Proc. Natl. Acad. Sci. U.S.A.*, **70** (1973) 697.
- [5] LESK A. M. and ROSE G. D., *Proc. Natl. Acad. Sci. U.S.A.*, **78** (1981) 4304.
- [6] ABKEVICH V. I., GUTIN A. M. and SHAKHNOVICH E. I., *Biochemistry*, **33** (1994) 10026.
- [7] LIFSHITS E. M. and PITAEVSKII L., *Physical Kinetics* (Pergamon, Oxford) 1981.
- [8] PANCHENKO A. R., LYTHEY-SCHULTEN Z. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **93** (1996) 2008.
- [9] BALDWIN R. L. and ROSE G. D., *TIBS*, **24** (1999) 26; 77.
- [10] YI Q., BYSTROFF C., RAJAGOPAL P., KLEVIT R. E. and BAKER D., *J. Mol. Biol.*, **283** (1998) 293.
- [11] ITZHAKI L. S., OTZEN D. E. and FERSHT A. R., *J. Mol. Biol.*, **254** (1995) 260.
- [12] MCALLISTER E., ALM E. and BAKER D., *Nature Struct. Biol.*, **7** (2000) 669.
- [13] SUTTO L., TIANA G. and BROGLIA R. A., *Protein Sci.*, **15** (2006) 1638.
- [14] GUEROIS R. and SERRANO L., *J. Mol. Biol.*, **304** (2000) 967.
- [15] TIANA G., SIMONA F., DE MORI G. M. S., BROGLIA R. A. and COLOMBO G., *Protein Sci.*, **13** (2004) 113.
- [16] TIANA G., BROGLIA R. A., ROMAN H. E., VIGEZZI E. and SHAKHNOVICH E., *J. Chem. Phys.*, **108** (1998) 757.
- [17] TIANA G., BROGLIA R. A. and SHAKHNOVICH E., *Proteins*, **39** (2000) 244.
- [18] TIANA G., DOKHOLYAN N. V., BROGLIA R. A. and SHAKHNOVICH E., *J. Chem. Phys.*, **121** (2004) 2381.
- [19] TIANA G. and BROGLIA R. A., *J. Chem. Phys.*, **114** (2001) 2503.
- [20] BROGLIA R. A., TIANA G. and PROVASI D., *J. Phys. Condensed Matter*, **R16** (2004) 111.
- [21] BROGLIA R. A., TIANA G., PASQUALI S., ROMAN H. E. and VIGEZZI E., *Proc. Natl. Acad. Sci. U.S.A.*, **95** (1998) 12930.
- [22] DUAN Y. and KOLLMAN P. A., *Science*, **282** (1998) 740.
- [23] SHAKHNOVICH E. I. and GUTIN A. M., *Biophys. Chem.*, **34** (1989) 187.
- [24] AMATORI A., TIANA G., SUTTO L., FERKINGHOFF-BORG J., TROVATO A. and BROGLIA R. A., *J. Chem. Phys.*, **123** (2005) 54904.
- [25] AMATORI A., FERKINGHOFF-BORG J., TIANA G. and BROGLIA R. A., *Phys. Rev. E*, **73** (2006) 61905.
- [26] SHAKHNOVICH E. I. and GUTIN A. M., *Nature*, **346** (1990) 773.
- [27] FERKINGHOFF-BORG J., *Eur. Phys. J. B*, **29** (2002) 481.
- [28] DOKHOLYAN N. V., BULDRYEV S. V., STANLEY H. E. and SHAKHNOVICH E. I., *Folding and Design*, **3** (1998) 577.
- [29] BROGLIA R. A., TIANA G. and BERERA R., *J. Chem. Phys.*, **118** (2003) 4754.
- [30] BROGLIA R. A., TIANA G., SUTTO L., PROVASI D. and SIMONA F., *Protein Sci.*, **14** (2005) 2668.
- [31] BROGLIA R. A., PROVASI D., VASILE F., OTTOLINA G., LONGHI R. and TIANA G., *Proteins*, **62** (2006) 928.

This page intentionally left blank

Analysis of the driving forces for biomolecular solvation and association

W. F. VAN GUNSTEREN, D. P. GEERKE and D. TRZESNIAK

*Laboratory of Physical Chemistry, ETH, Swiss Federal Institute of Technology
 CH-8093 Zürich, Switzerland*

C. OOSTENBRINK

*Computational Medicinal Chemistry and Toxicology, Department of Pharmacochemistry
 Vrije Universiteit - De Boelelaan 1083 P 262, NL-1081 HV Amsterdam, The Netherlands*

N. F. A. VAN DER VEGT

Max-Planck Institute for Polymer Research - Ackermannweg 10, D-55128 Mainz, Germany

1. – Introduction

Many processes occurring in a molecular system will be determined by the free-energy difference between two or more states, because the free energy is a measure of the probability of finding the system in a given state. The free energy F of a system of N particles in a volume V at temperature T is a $6N$ -dimensional integral over all particle coordinates $\vec{r}$ and momenta $\vec{p}$ of the Boltzmann factor of the system Hamiltonian $H(\vec{p}, \vec{r})$ (kinetic plus potential energy)

$$(1) \quad F_{NVT} = -k_B T \ln \left([N! h^{3N}]^{-1} \int \int \exp[-H(\vec{p}, \vec{r})/k_B T] d\vec{p} d\vec{r} \right),$$

where k_B is Boltzmann’s constant and h is Planck’s constant. The factor $N!$ must be omitted when dealing with distinguishable particles. Since the integrand, the Boltzmann

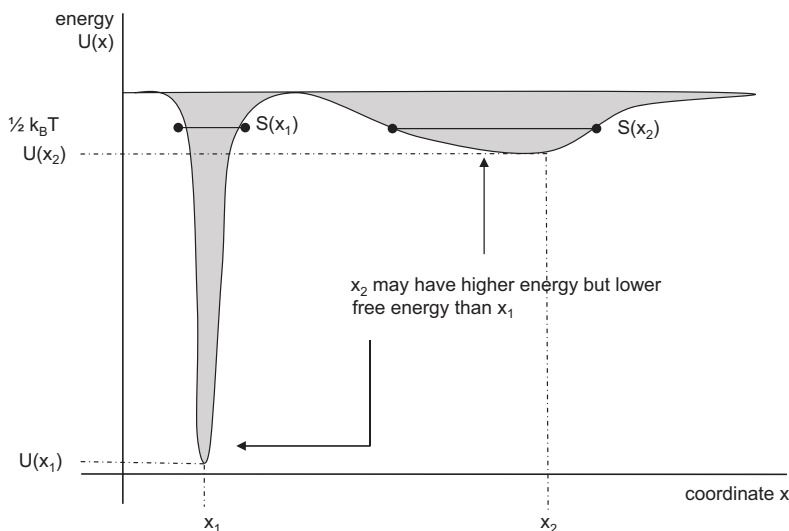


Fig. 1. – Energy(U)-entropy(S) compensation at finite temperature T for a one-dimensional (x) system.

factor, is everywhere positive, the omission of configurations in the integrals leads to systematic (not cancelling) errors. Not only will low-energy structures or configurations determine the free energy, but also configurations of higher energy and greater abundance will do so. In other words, both energy U and entropy S contribute to the free energy at finite temperatures:

$$(2) \quad F = U - TS.$$

It is well known that entropy plays an essential role in many biomolecular processes, in particular in molecular solvation and in association in the condensed phase. Changes in free energy that drive these processes may result from changes in internal energy U or in entropy S , which may work together or against each other depending on the relative strengths of the nonbonded interactions between the various components (atoms, molecules) of the system. Figure 1 illustrates the phenomenon of energy-entropy compensation: two conformations x_1 and x_2 of a molecule may have $U(x_1) \ll U(x_2)$, while $F(x_1) > F(x_2)$ if at a given temperature $S(x_1) \ll S(x_2)$. The entropy is a measure of the extent of configuration space (x) accessible to the molecular system at a given temperature T .

Since the free energy change ΔF of a given process can be written as $\Delta F = \Delta U - T\Delta S$, different combinations of energy change ΔU and entropy change ΔS may result in the same ΔF . For solvation and condensed-phase association processes at room temperatures, the ΔU and $T\Delta S$ values are generally larger than the ΔF value. In other words, ΔU and ΔS do partially compensate each other, which we will call “physico-chemical”

energy-entropy compensation. When considering the process of solvation, *i.e.* bringing a solute molecule (indicated by the index u) into a solvent (indicated by the index v), we may separate the solute-solute (uu), solute-solvent (uv), and solvent-solvent (vv) contributions to the energy and entropy:

$$(3) \quad \Delta F = \Delta U_{uu} + \Delta U_{uv} + \Delta U_{vv} - T\Delta S_{uu} - T\Delta S_{uv} - T\Delta S_{vv}.$$

The process of solvation can be described as taking a solute molecule and turning on the nonbonded interaction between the solute and the solvent. In other words, only the solute-solvent term $H_{uv}(\vec{p}, \vec{r})$ in the Hamiltonian will contribute directly to ΔF . It has been shown [1-3] that this leads to exact statistical-mechanical compensation of the solvent-solvent energy and entropy in (3),

$$(4) \quad \Delta U_{vv} = T\Delta S_{vv}.$$

This implies that the solvent rearrangement energy ΔU_{vv} and entropy ΔS_{vv} do not contribute to ΔF . Following the same line of argument, one has

$$(5) \quad \Delta U_{uu} = T\Delta S_{uu}$$

or

$$(6) \quad \Delta F = \Delta U_{uv} - T\Delta S_{uv}.$$

Only the solute-solvent forces drive the solvation process. The same type of reasoning can be carried through for the process of molecular association of two solutes in a solvent. Unfortunately, the changes ΔU_{uv} and ΔS_{uv} that drive the process cannot be measured experimentally, they can only be calculated from molecular simulation trajectories or ensembles. On the other hand, ΔU and ΔS values for solvation or association can be measured and calculated, but they may be not representative for the driving forces of the process in case the exactly compensating solvent-solvent terms ΔU_{vv} and ΔS_{vv} are much larger than the essential solute-solvent terms ΔU_{uv} and ΔS_{uv} . Due to the generally much larger number of solvent-solvent interactions compared to solute-solvent interactions, it is much harder to compute precise ΔU and ΔS values from a molecular simulation than it is to calculate ΔU_{uv} and ΔS_{uv} values.

In this article we illustrate the effects of statistical-mechanical and physico-chemical energy-entropy compensation for the processes of solvation and association of small (polar and apolar) solute molecules in pure and mixed solvents. These basic processes are determining many biomolecular processes such as polypeptide or protein folding, protein-ligand, protein-protein or protein-DNA association, and partitioning of solute molecules between different solvents (water, membranes).

We do not wish to review the literature concerning solvation or association processes, but only demonstrate the above-mentioned compensation effects using examples from our own work [4-9].

2. – Methods

The process of solvation can be described using a Hamiltonian in which the interaction between the solute and solvent molecules V_{uv} is made dependent on a coupling parameter λ , such that for $\lambda = 0$ there is no solute-solvent interaction and for $\lambda = 1$ the full solute-solvent interaction is present. Then, the free energy of solvation can be written as [2]

$$(7) \quad \Delta F^{\text{solv}} = \int_0^1 \left\langle \frac{\partial V_{uv}(\lambda)}{\partial \lambda} \right\rangle_{\lambda} d\lambda,$$

in which $\langle \dots \rangle_{\lambda}$ is an average over an ensemble generated using $V(\vec{r}; \lambda)$. For the energy of solvation one has

$$(8) \quad \begin{aligned} \Delta U^{\text{solv}} &= \langle V_{uv}(\lambda = 1) \rangle_{\lambda=1} - \langle V_{uv}(\lambda = 0) \rangle_{\lambda=0} \\ &\quad - (k_B T)^{-1} \int_0^1 \left\{ \left\langle V_{vv} \frac{\partial V_{uv}(\lambda)}{\partial \lambda} \right\rangle_{\lambda} - \langle V_{vv} \rangle_{\lambda} \left\langle \frac{\partial V_{uv}(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \right\} d\lambda \\ &= \Delta U_{uv}^{\text{solv}} + \Delta U_{vv}^{\text{solv}}, \end{aligned}$$

where we have, without loss of generality, assumed that $V_{uu} = 0$, or alternatively that the λ -independent V_{uu} term is included in the λ -independent V_{vv} term. For the entropy one has

$$(9) \quad \begin{aligned} T \Delta S^{\text{solv}} &= -(k_B T)^{-1} \int_0^1 \left\{ \left\langle V_{uv}(\lambda) \frac{\partial V_{uv}(\lambda)}{\partial \lambda} \right\rangle_{\lambda} - \langle V_{uv}(\lambda) \rangle_{\lambda} \left\langle \frac{\partial V_{uv}(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \right\} d\lambda \\ &\quad - (k_B T)^{-1} \int_0^1 \left\{ \left\langle V_{vv} \frac{\partial V_{uv}(\lambda)}{\partial \lambda} \right\rangle_{\lambda} - \langle V_{vv} \rangle_{\lambda} \left\langle \frac{\partial V_{uv}(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \right\} d\lambda \\ &= T \Delta S_{uv}^{\text{solv}} + T \Delta S_{vv}^{\text{solv}}. \end{aligned}$$

From these formulae the exact statistical-mechanical compensation of energy and entropy,

$$(10) \quad \Delta U_{vv}^{\text{solv}} = T \Delta S_{vv}^{\text{solv}},$$

is evident. These two terms contribute exactly oppositely to ΔF^{solv} , so

$$(11) \quad \Delta F^{\text{solv}} = \Delta U^{\text{solv}} - T \Delta S^{\text{solv}}$$

can be simplified to

$$(12) \quad \Delta F^{\text{solv}} = \Delta U_{uv}^{\text{solv}} - T \Delta S_{uv}^{\text{solv}}.$$

These formulae (7)-(9) are the thermodynamic integration (TI) formulae [10] for the free energy, energy and entropy. For small solutes, which do not perturb the solution too much, even simpler particle-insertion (PI) formulae [11] and procedures can be used. If

the molecular system is simulated at constant pressure instead of constant volume, the (Helmholtz) free energy F becomes the (Gibbs) free enthalpy G ($G = F + pV$), and the energy U becomes the enthalpy H ($H = U + pV$).

When considering solute *versus* solvent contributions to the free energy of association, similar expressions can be derived. In this case, the role of the λ -parameter that regulates the interaction $V_{uv}(\lambda)$ between solute and solvent is played by the reaction coordinate r that describes the distance between the two solute molecules for which the association free energy or potential of mean force is to be calculated [7]. In eqs. (7)-(12) λ is to be replaced by r (with appropriate integration boundaries) and the superscript *solv* by *assoc*.

3. – Results

In the following we consider a variety of molecular systems: polar as well as apolar solutes and solvents, pure solvents as well as solvent mixtures of varying composition. We consider the processes of solvation and of association of two or more molecules in the condensed phase. For the different systems and processes we calculate ΔF , ΔU , ΔS , ΔU_{uv} and ΔS_{uv} values or their constant pressure counterparts.

3.1. Solvation.

3.1.1. Solvation of polar and apolar solutes in polar and apolar pure solvents. Table I lists free enthalpy of solvation for a number of apolar and polar compounds in solvents of varying polarity, from water to cyclohexane [4]. The root-mean-square (rms) deviation between simulated and experimental values is of the order of $k_B T$ (2.5 kJ mol⁻¹ at 300 K, 1 atm) or lower. This shows that the GROMOS 53A6 force field [12] adequately describes the partitioning of the various solutes between the different solvents. This indicates the level of accuracy of the energetic quantities that can be obtained from simulations using a thermodynamically calibrated force field. This level of accuracy should be kept in mind when interpreting the calculated solute-solvent energy and entropy values ΔU_{uv} and ΔS_{uv} , which cannot be compared to experimental values.

3.1.2. Solvation of methane in aqueous solutions of different co-solvents. Figures 2-4 show the free enthalpy of solvation for a methane solute in aqueous solutions of NaCl, acetone and DMSO as a function of NaCl concentration or co-solvent mole fraction [5]. The lower panels show the energetic and entropic solute-solvent contributions to the solvation process relative to those for solvation in pure water ($\Delta\Delta U_{uv}$ and $\Delta\Delta S_{uv}$). One observes that entropy disfavors solvation increasingly with salt concentration, while the energy contribution is rather insensitive to the salt concentration (fig. 2). Solvation of methane in acetone aqueous solution shows a quite different picture (fig. 3). The entropy favours solvation, while the energy is lightly favourable, not changing beyond 0.4 mole fraction acetone. Solvating methane in DMSO aqueous solution offers yet another picture (fig. 4). Although the free enthalpy of solvation decreases with mole fraction, *i.e.* favours

TABLE I. – Free enthalpy of solvation [kJ mol^{-1}] of neutral polar and apolar compounds in different solvents at 298.15 K and 1 atm: values from simulation (ΔG_{sim}) and experiment (ΔG_{exp}), and root-mean-square (rms), average and average absolute deviations of the calculated values from experiment per class (polar, apolar) of compounds. In the simulations, all solutes were described by the GROMOS 53A6 force field parameter set, as were liquid cyclohexane, chloroform, methanol, and DMSO. Water was described by the SPC model, acetonitrile by the model of Gee and van Gunsteren. In parentheses: naturally occurring amino acid side chain for which the compound is an analog. Data from ref. [4].

Solute	Water		Cyclohexane/chloroform		Methanol		DMSO		Acetonitrile	
	ΔG_{sim}	ΔG_{exp}	ΔG_{sim}	ΔG_{exp}	ΔG_{sim}	ΔG_{exp}	ΔG_{sim}	ΔG_{exp}	ΔG_{sim}	ΔG_{exp}
apolar										
methane (Ala)	8.7	8.4	0.6	0.8	3.2	1.6	4.3	5.0		
ethane	7.4	7.5					0.0	1.2		
propane (Val)	8.6	8.2	-8.5	-6.7	-4.2	-4.8	-2.6	-1.5		
butane (Ile)	8.7	8.8	-11.4	-13.0	-6.2	-7.8				
isobutane (Leu)	10.3	9.6	-10.9	-9.8	-4.9	-6.7			-4.2	-5.9
<i>n</i> -pentane	10.2	9.8			-8.4	-9.5	-4.2	-6.2	-7.1	-8.7
<i>n</i> -hexane	11.5	10.5			-10.4	-11.7	-7.0	-7.2	-9.0	-11.2
<i>n</i> -nonane	12.6	12.7			-16.9	-18.4	-16.0	-13.4	-16.1	-17.1
cyclohexane	2.9	5.2	-21.0	-18.5	-15.5	-13.8	-16.1	-10.1		
toluene (Phe)	0.0	-3.1	-25.2	-20.5	-16.5	-18.2	-18.7	-18.9	-16.2	-19.8
rms deviation	1.3		2.4		1.5		2.5		2.2	
average deviation	0.3		-1.5		1.1		-1.2		2.0	
average absolute deviation	0.9		2.0		1.4		1.8		2.0	
polar (neutral)										
methanol (Ser)	-22.1	-21.2	-5.3	-6.9	-25.0	-20.5	-23.5	-21.2	-16.1	-16.7
ethanol (Thr)	-20.0	-20.5	-8.3	-9.5	-24.4	-22.2	-23.3	-22.3	-17.7	-18.1
1-propanol	-19.9	-21.2			-25.2	-24.9	-24.7	-23.8		
acetic acid (Asp)	-30.6	-28.0	-15.2	-9.1						
propioic acid (Glu)	-27.2	-27.0	-18.4	-15.8						
urea	-60.9	-58.5					-54.3	-56.5		
NH ₃	-18.6	-18.0	-2.5	-3.3	-16.9	-11.8				
MMA	-20.1	-19.1	-2.0	-2.1						
DMA	-18.7	-18.0	-3.4	-0.8	-12.9	-12.3				
TMA	-13.9	-13.6	-8.8	-6.2						
acetamide (Asn)	-40.6	-40.6	-12.8	-12.6						
propamide (Gln)	-38.5	-39.4	-14.8	-13.9						
rms deviation	1.2		2.5		3.2		1.7		0.5	
average deviation	-0.5		-1.1		-2.5		-0.5		0.5	
average absolute deviation	1.0		1.9		2.5		1.6		0.5	

solvation, as was observed for the acetone solution, the driving forces of the solvation process are different. The energy strongly favours solvation, while the entropy is lightly unfavourable, again not changing beyond 0.4 mole fraction. These figures illustrate that similar free energy changes may be caused by different driving forces, even for rather comparable molecular systems.

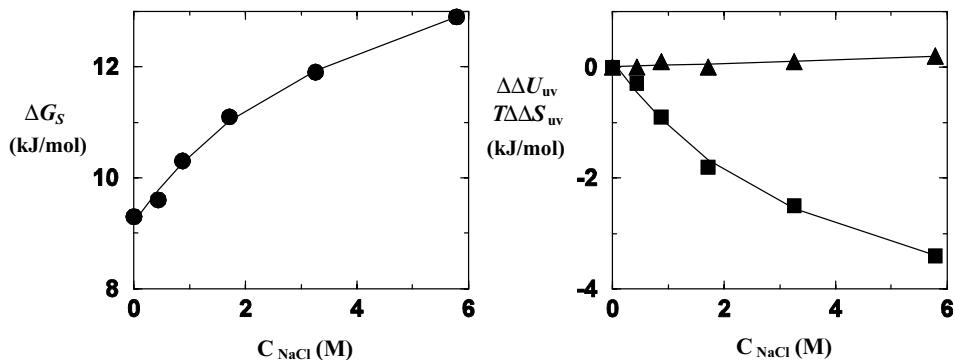


Fig. 2. – Methane solvation free enthalpy ΔG^{solv} (left panel) and solute-solvent energy $\Delta\Delta U_{uv}^{\text{solv}}$ (triangles) and entropy $T\Delta\Delta S_{uv}^{\text{solv}}$ (squares) relative to neat water (right panel) as a function of Na^+Cl^- concentration of the aqueous solution. Data from ref. [5].

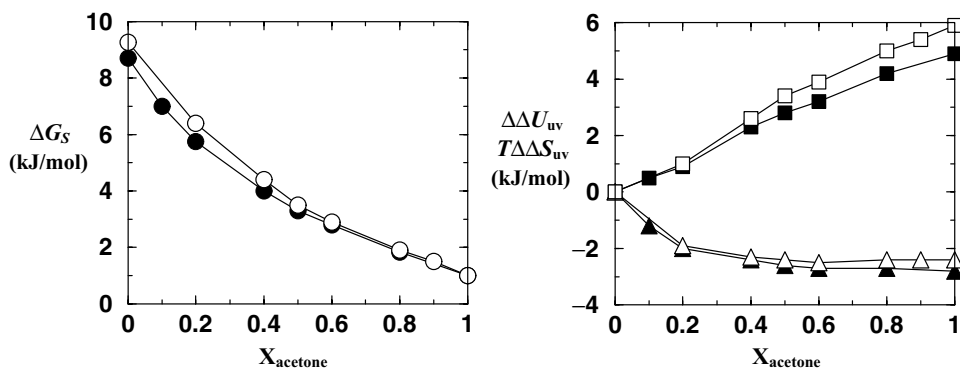


Fig. 3. – Methane solvation free enthalpy ΔG^{solv} (left panel) and solute-solvent energy $\Delta\Delta U_{uv}^{\text{solv}}$ (triangles) and entropy $T\Delta\Delta S_{uv}^{\text{solv}}$ (squares) relative to neat water (right panel) as a function of acetone mole fraction of the aqueous solution. Solvent mixtures with SPC water (closed symbols) and SPC/E water (open symbols) are shown. Data from ref. [5].

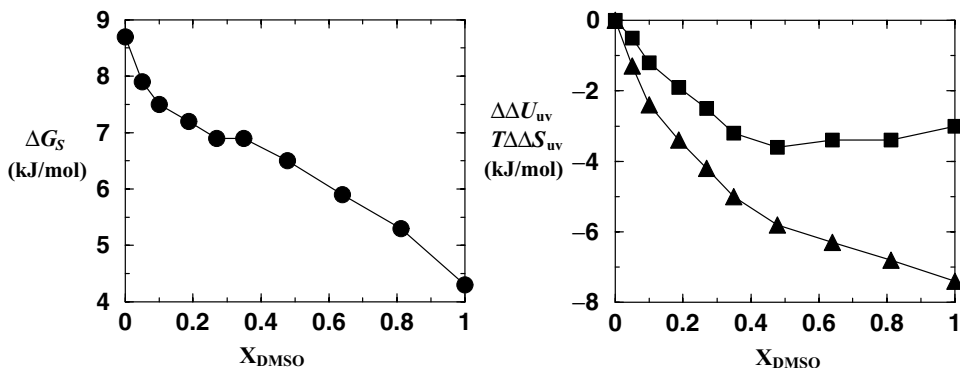


Fig. 4. – Methane solvation free enthalpy ΔG^{solv} (left panel) and solute-solvent energy $\Delta\Delta U_{uv}^{\text{solv}}$ (triangles) and entropy $T\Delta\Delta S_{uv}^{\text{solv}}$ (squares) relative to neat water (right panel) as function of DMSO mole fraction of the aqueous solution. Data from ref. [5].

3'1.3. Solvation of apolar solutes in aqueous solutions of different co-solvents. Table II shows the solvation free enthalpies, energies and entropies relative to solvation in pure water for different small apolar solutes [6]. The relative and absolute contributions of energy and entropy do vary considerably between the solutes. Comparing the different co-solvents, one observes cases of energy and entropy co-acting as well as counteracting, for example depending on which acetone model was used. Similar relative free enthalpy values may have different energetic versus entropic components.

3'2. Association.

3'2.1. Association of two hydrophobic or hydrophilic solutes in pure water. Figure 5 shows the free energy of association of two (united atom) methane molecules in pure water relative to the gas phase, *i.e.* excluding the direct methane-methane interaction $V_{uu}(r)$ [7]. The effect of the degree of hydrophobicity or hydrophilicity of the solute pair is investigated by varying the strength of the methane-water interaction in V_{uw} . This is done by changing the attractive van der Waals interaction parameter C_6 or charging the two molecules of the solute pair with equal but opposite charges. Using the standard GROMOS interaction the relative ΔF^{assoc} (solute-solute interaction excluded) behaves as expected for two hydrophobic particles in a polar solvent. The most favourable configuration would be achieved if the two methane solutes were on top of each other. Halving the C_6 (CH_4 -OW water) van der Waals parameter weakens the attraction between solutes and solvent (second panel fig. 5). This leads to lower ΔF^{assoc} values at short separation distances. This is expected because the system wants to promote solute association. The opposite behaviour is observed when decreasing the hydrophobicity by doubling the C_6 (CH_4 -OW) van der Waals parameter (third panel fig. 5). In this case, the solvent separated minimum shifts to shorter distance (0.65 nm) and becomes deeper. The bottom panel of fig. 5 shows the result of completely annihilating the hydrophobicity by adding opposite partial charges (+0.5 and -0.5 e) to the methane solute pair. It is now extremely unfavourable for the solutes to be close together, because water would like to interact with each of them and therefore the relative ΔF^{assoc} is purely repulsive. When adding the vacuum methane-methane potential energy to the solid curves in fig. 5, we recover the absolute ΔF^{assoc} as a function of distance r (dashed curves).

Figure 6 shows the relative entropy of association ΔS^{assoc} as a function of the methane-methane distance r obtained using the TI formula eq. (9) for association and using temperature differentiation of the free energy, together with the solute-solvent contribution $\Delta S_{uv}^{\text{assoc}}$ [7]. The first three panels show the entropy change upon association that is expected for hydrophobic particles in water [6]. The ΔS^{assoc} curves are difficult to converge within 2 ns per methane-methane distance, because of the large solvent-solvent contribution. The solute-solvent part of the entropy, $\Delta S_{uv}^{\text{assoc}}$, which represents the entropic driving force for the association, is much less noisy. For hydrophobic solutes in aqueous solution, the minima in $\Delta F^{\text{assoc}}(r)$ are generally determined by the solute-solvent entropy. For the dipolar (charged) methane pair, the

TABLE II. – *Thermodynamic data of solute transfer from pure water to cosolvent/water mixtures at 298 K and 1 atm. Units are kJ mol^{-1} . $\Delta\Delta G_s$, solute transfer Gibbs free energies; $\Delta\Delta U_{uv}$, solute-solvent transfer energies; $T\Delta\Delta S_{uv}$, solute-solvent transfer entropies. The data for helium, neon, argon, krypton, xenon, and methane were calculated by PI using at least 2.5×10^8 particle insertions in 1000-2000- p simulation trajectories of the cosolvent/water mixture; the data for ethane, propane, n -butane, iso -butane, and neo -pentane were calculated by TI. At least 21 λ -point simulations were performed under constant pressure and temperature conditions, requiring a total simulation time of 20 ns per solute for each cosolvent/water mixture. Data from ref. [6]*

Solute	NaCl (11%)			Urea (15%)			DMSO (10%)			Acetone(I) (10%)			Acetone(I) (50%)			Acetone(II) (10%)		
	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$	$\Delta\Delta G_s$	$\Delta\Delta U_{uv}$	$T\Delta\Delta S_{uv}$
helium	2.0	-0.4	-2.4	1.3	-0.1	-1.4	0.0	-0.2	-0.2	-0.4	-0.2	0.2	-1.7	-0.6	1.1	-0.1	-0.2	-0.1
neon	2.8	0.2	-2.6	1.1	-0.7	-1.8	-0.3	-0.7	-0.4	-0.7	-0.4	0.3	-2.4	-1.0	1.4	-0.4	-0.6	-0.2
argon	3.7	-0.1	-3.8	0.8	-2.3	-3.1	-0.9	-1.8	-0.9	-1.3	-1.0	0.3	-4.4	-2.1	2.3	-1.1	-1.7	-0.6
krypton	4.0	-0.2	-4.2	0.4	-3.2	-3.6	-1.2	-2.3	-1.1	-1.6	-1.2	0.4	-5.2	-2.6	2.6	-1.5	-2.3	-0.8
xenon	4.4	-1.2	-5.6	-0.5	-4.7	-4.2	-1.5	-2.7	-1.2	-2.2	-1.7	0.5	-6.8	-3.6	3.2	-2.1	-3.4	-1.3
methane	4.2	-0.2	-4.4	0.5	-3.1	-3.6	-1.1	-2.1	-1.0	-1.7	-1.2	0.5	-5.4	-2.6	2.8	-1.4	-2.2	-0.8
ethane	5.6	-0.8	-6.4	-0.9	-6.0	-5.1	-2.8	-4.4	-1.6	-2.9	-2.4	0.5	-8.4	-4.6	3.8	-2.6	-4.3	-1.7
propane	5.4	-0.3	-5.7	-2.2	-7.3	-5.1	-4.8	-5.8	-1.0	-5.8	-3.9	1.9	-13.0	-6.5	6.5	-4.8	-6.2	-1.4
n -butane	7.4	-1.2	-8.6	-2.7	-10.6	-7.9	-5.8	-7.9	-2.1	-6.4	-5.6	0.8	-15.7	-9.3	6.4	-4.3	-6.0	-1.7
iso -butane	5.7	-1.2	-6.9	-2.3	-10.2	-7.9	-6.5	-6.7	-0.2	-6.7	-4.9	1.8	-15.2	-7.7	7.5	-5.9	-7.7	-1.8
neo -pentane	8.2	-0.6	-8.8	-2.2	-10.9	-8.7	-6.0	-8.3	-2.3	-6.7	-4.3	2.4	-15.8	-7.1	8.7	-5.3	-9.7	-4.4

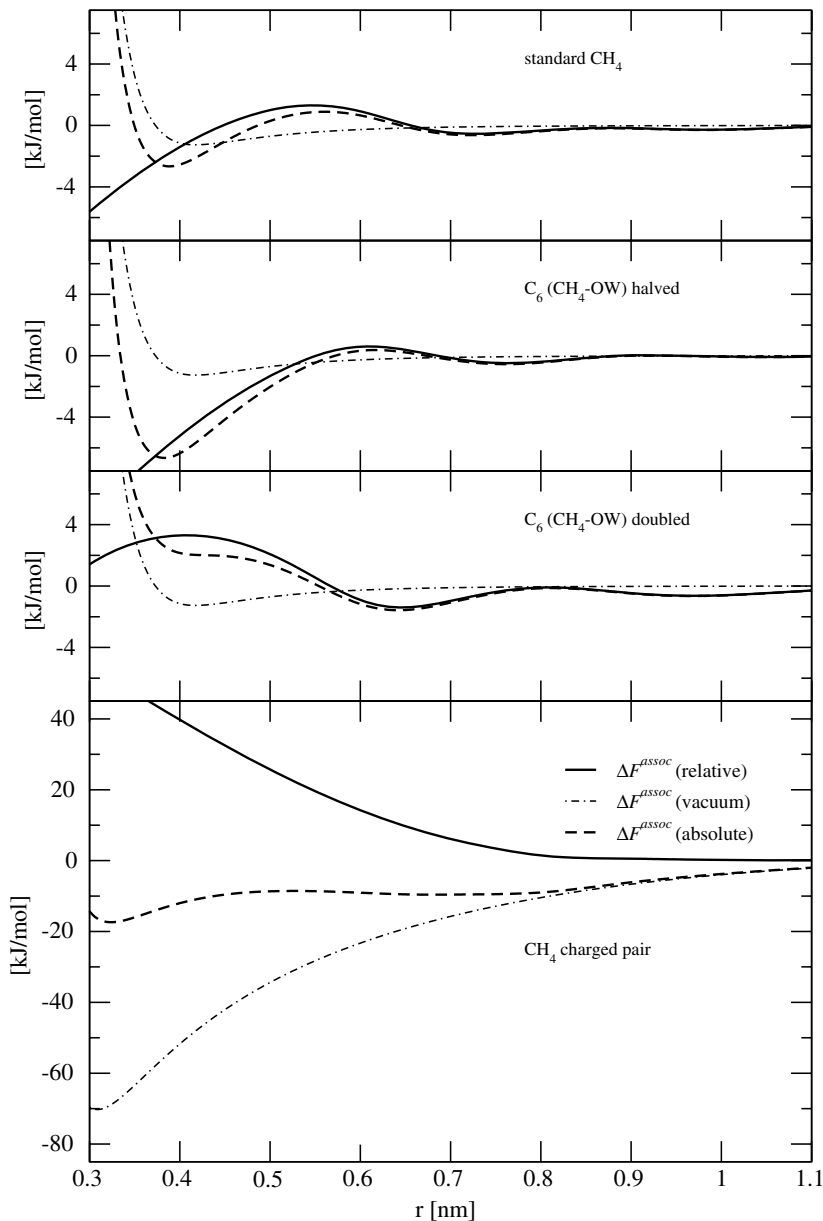


Fig. 5. – Relative free energy of association (ΔF^{assoc}) as a function of the distance between two methane molecules (solute-solute interaction excluded) in aqueous solution. From top to bottom, the panels show the results for the standard methane van der Waals and Coulomb interaction with water, with the attractive van der Waals $\text{C}_6 (\text{CH}_4\text{-OW})$ parameter halved, with this parameter doubled, and with the methane pair charged ($+0.5\text{ e}$, -0.5 e). The dot-dashed curves show the vacuum methane-methane interaction and the dashed curves show the absolute free energy (solute-solute interaction included). Data from ref. [7].

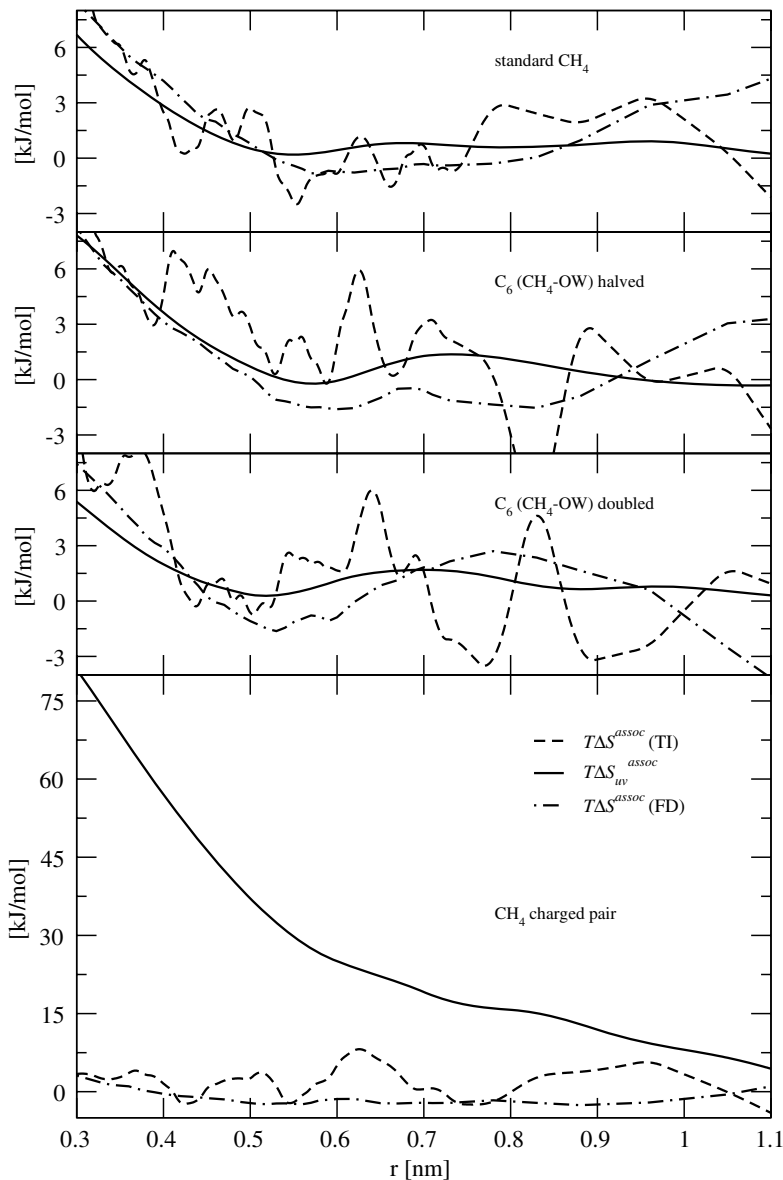


Fig. 6. – Entropy as a function of the distance between two methane molecules (solute-solute interaction excluded) in aqueous solution. Entropy of association $T\Delta S^{\text{assoc}}$ as obtained from thermodynamic integration (dashed lines) or temperature differentiation of the free energy (dot-dashed lines), and the solute-solvent contribution $T\Delta S_{\text{uv}}^{\text{assoc}}$ (solid lines). From top to bottom, the panels show the results for the standard methane van der Waals and Coulomb interaction with water, with the attractive van der Waals C_6 ($\text{CH}_4\text{-OW}$) parameter halved, with this parameter doubled, and with the methane pair charged (+0.5 e, −0.5 e). Data from ref. [7].

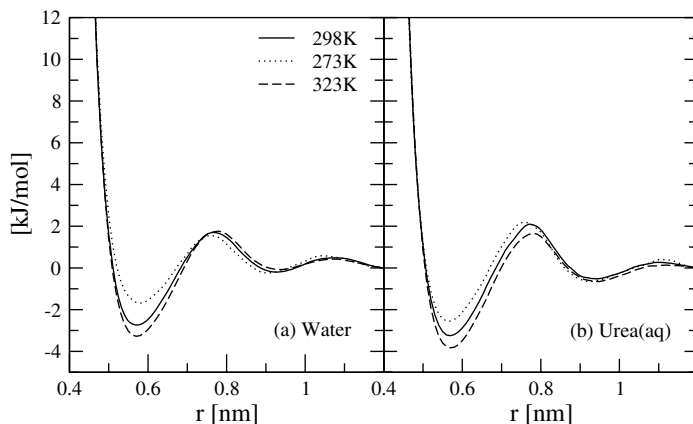


Fig. 7. – Free enthalpy of association (ΔG^{assoc}) as function of the distance between two *neo*-pentane molecules in pure water (left panel) and in 6.9 mol l^{-1} urea solution (right panel) at 1 atm and 273 K (dotted lines), at 298 K (solid lines), and at 323 K (dashed lines). Data from ref. [8].

solute-solvent entropy and the (relative) ΔF^{assoc} behave similarly: both decrease with increasing solute separation. As the ions separate, stronger ordering of water is induced due to the larger solute dipole moment.

3.2.2. Association of two hydrophobic solutes in water and in aqueous urea. Figure 7 shows the free enthalpy of association $\Delta G^{\text{assoc}}(r)$ for a pair of *neo*-pentane molecules in water (left panel) and in 6.9 mol l^{-1} urea solution (right panel) for three temperatures [8]. The curves for urea and water are rather similar and the *neo*-pentane pair contact minimum at about 0.57 nm becomes deeper with increasing temperature, indicating that *neo*-pentane association is entropically driven in both solutions. In fig. 8, the enthalpic and entropic contributions are shown together with their solvent-solvent ($\Delta H_{\text{vv}}^{\text{assoc}} = T\Delta S_{\text{vv}}^{\text{assoc}}$) and solute-solvent ($\Delta H_{\text{uv}}^{\text{assoc}}, T\Delta S_{\text{uv}}^{\text{assoc}}$) counterparts. In water, the entropy stabilises the contact minimum, while the enthalpy destabilises it. These contributions largely compensate each other in the free enthalpy. The solvent separated minimum around 0.92 nm is enthalpically favourable, while entropically being unfavourable. The middle panels show the contribution $\Delta H_{\text{vv}}^{\text{assoc}} = T\Delta S_{\text{vv}}^{\text{assoc}}$ of the solvent reorganisation upon *neo*-pentane association. The values are larger and are oscillating differently between water and urea. Yet, because of exact statistical-mechanical energy-entropy compensation, these oscillations do not show up in the ΔG^{assoc} curves. However, they do mask the solute-solvent contributions $\Delta H_{\text{uv}}^{\text{assoc}}$ and $T\Delta S_{\text{uv}}^{\text{assoc}}$ (lower panel) in the total ΔH^{assoc} and $T\Delta S^{\text{assoc}}$ curves (upper panels). So, interpreting the ΔH^{assoc} and $T\Delta S^{\text{assoc}}$ curves in terms of enthalpic and entropic driving forces may lead to spurious conclusions, because the real driving forces for the association are contained in $\Delta H_{\text{uv}}^{\text{assoc}}$ and $T\Delta S_{\text{uv}}^{\text{assoc}}$. Comparing these curves between water and urea, the differences are less pronounced. Both in water and in urea, the *neo*-pentane contact minimum in

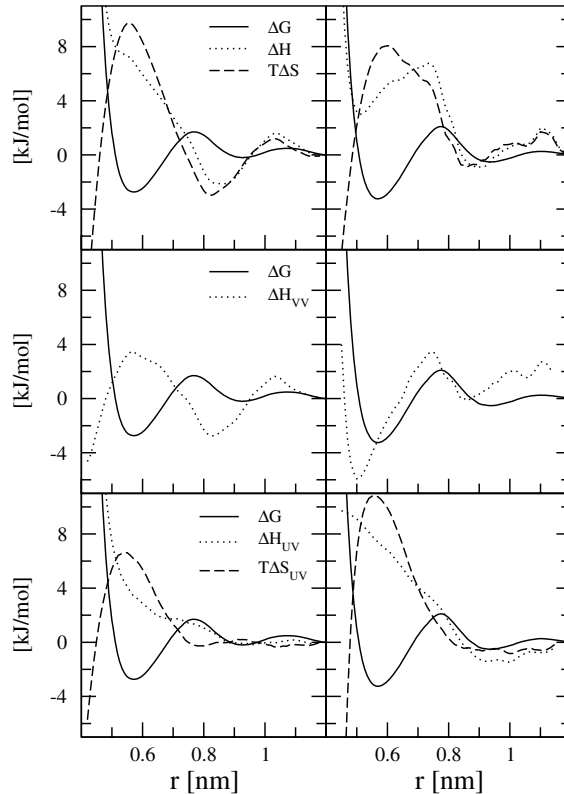


Fig. 8. – Free enthalpy of association (ΔG^{assoc}) and its enthalpic (ΔH^{assoc}) and entropic ($T\Delta S^{\text{assoc}}$) contributions (upper panels), its solvent-solvent ($\Delta H_{\text{vv}}^{\text{assoc}} = T\Delta S_{\text{vv}}^{\text{assoc}}$) contributions (middle panels), and its solute-solvent ($\Delta H_{\text{uv}}^{\text{assoc}}$, $T\Delta S_{\text{uv}}^{\text{assoc}}$) contributions (lower panels) as function of the distance between two *neo*-pentane molecules in pure water (left panels) and in 6.9 mol l^{-1} urea solution (right panels) at 1 atm and 298 K. Data from ref. [8].

ΔG^{assoc} is stabilised by the solute-solvent entropy and destabilised by the solute-solvent enthalpy. The solvent separated minimum at 0.92 nm in ΔG^{assoc} is in water stabilised by the solute-solvent entropy, whereas in urea by the solute-solvent enthalpy.

3.2.3. Association of many hydrophobic solutes in water and in urea solutions. The thermodynamic-integration and particle-insertion methods that were used to obtain the free energies of solvation and association presented in the previous sections cannot so easily be used to obtain these quantities for the formation of larger clusters of solutes in solution. An alternative is to simulate a solute-solvent mixture and to analyse the solute cluster sizes at equilibrium. The free energy of cluster growth can be approximated using the relation [9, 13]

$$(13) \quad \Delta G_{\text{growth}} = -k_{\text{B}}T \ln \left(\frac{[m_{i+1}]C_0}{[m_i][m_1]} \right),$$

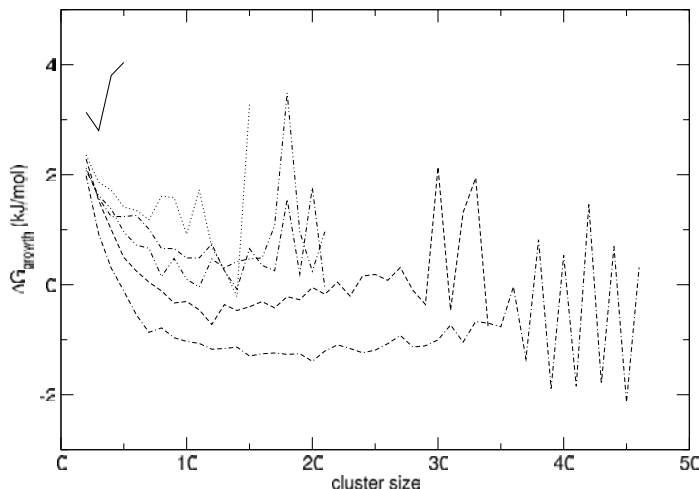


Fig. 9. – Free enthalpy of methane cluster growth as a function of the cluster size for different methane-urea-water mixtures: $C_{5,0}$ (solid line), $C_{20,0}$ (dotted line), $C_{25,0}$ (dashed line), $C_{20,6}$ (dot-dashed line), $C_{20,10}$ (dot-dot-dashed line), and $C_{25,10}$ (dash-dash-dotted line). The first index on $C_{i,j}$ denotes the methane concentration in 10^{-10} M and the second index the urea concentration in M. Large fluctuations at the end of the curves stem from poor statistics for large cluster sizes. Data from ref. [9].

in which $[m_i]$ is the concentration of a cluster of i solutes and C_0 is a unit of concentration. In fig. 9 this free energy is shown as a function of cluster size for various solutions of methane in water or aqueous urea. The composition of a system is roughly indicated by the indices i (concentration of methane in 10^{-10} M) and j (urea concentration in M) of the system code $C_{i,j}$. It is observed that methane cluster formation in water only becomes favourable for clusters of more than five methanes, and ΔG_{growth} becomes lower for higher methane concentrations, as expected. Addition of urea to the solution does favour methane cluster formation. So, urea would enhance the hydrophobic effect at higher concentrations. We note that for larger cluster sizes the results become unreliable due to insufficient statistics or sampling.

4. – Conclusions

The examples shown in this study illustrate how molecular simulation can be used to interpret experimental thermodynamic data on solvation and association of solutes in pure solvents or solvent mixtures. Due to the sizeable entropic effects, interpretation in terms of single solute-solvent configurations is not meaningful. Due to exact statistical-mechanical energy-entropy compensation of solvent-solvent contributions, $\Delta U_{\text{vv}} = T\Delta S_{\text{vv}}$, in the free energy, these contributions may mask the real driving forces contained in the solute-solvent terms ΔU_{uv} and $T\Delta S_{\text{uv}}$. The latter terms should be used to interpret the solvation or association mechanism, not the measurable but less

informative ΔU and $T\Delta S$ terms. The physico-chemical energy-entropy compensation that is also observed for many systems implies that biomolecular force fields should not only be calibrated to reproduce experimental free energies of solvation, but also energies and entropies of solvation.

* * *

Financial support by the National Center of Competence in Research (NCCR) Structural Biology of the Swiss National Science Foundation (SNSF) is gratefully acknowledged.

REFERENCES

- [1] YU H. A. and KARPLUS M., *J. Chem. Phys.*, **89** (1988) 2366.
- [2] PETER C., OOSTENBRINK C., VAN DORP A. and VAN GUNSTEREN W. F., *J. Chem. Phys.*, **120** (2004) 2652.
- [3] BEN-AMOTZ D., RAINERI F. O. and STELL G., *J. Phys. Chem. B*, **102** (2005) 6866.
- [4] GEERKE D. P. and VAN GUNSTEREN W. F., *Chem. Phys. Chem.*, **7** (2006) 671.
- [5] VAN DER VEGT N. F. A. and VAN GUNSTEREN W. F., *J. Phys. Chem. B*, **108** (2004) 1056.
- [6] VAN DER VEGT N. F. A., TRZESNIAK D., KASUMAJ B. and VAN GUNSTEREN W. F., *Chem. Phys. Chem.*, **5** (2004) 144.
- [7] TRZESNIAK D. and VAN GUNSTEREN W. F., *Chem. Phys.*, **330** (2006) 410.
- [8] VAN DER VEGT N. F. A., LEE M.-E., TRZESNIAK D. and VAN GUNSTEREN W. F., *J. Phys. Chem. B*, **110** (2006) 12852.
- [9] OOSTENBRINK C. and VAN GUNSTEREN W. F., *Phys. Chem. Chem. Phys.*, **7** (2005) 53.
- [10] KIRKWOOD J. G., *J. Chem. Phys.*, **3** (1935) 300.
- [11] WIDOM B., *J. Chem. Phys.*, **39** (1963) 2808.
- [12] OOSTENBRINK C., VILLA A., MARK A. E. and VAN GUNSTEREN W. F., *J. Comp. Chem.*, **25** (2004) 1656.
- [13] RASCHKE T. M., TSAI J. and LEVITT M., *Proc. Natl. Acad. Sci. U.S.A.*, **98** (2001) 5965.

This page intentionally left blank

Structural basis of dielectric permittivity of proteins: Insights from quantum mechanics

K. RAHA(*) and K. M. MERZ jr.(**)

*104 Chemistry Building, The Pennsylvania State University
University Park, PA 16802, USA*

In this report, we demonstrate the use of semiempirical quantum mechanics (QM) and molecular dynamics simulations (MD) in conjunction with the Frohlich-Kirkwood theory to calculate the dielectric permittivity of proteins. The proteins staphylococcus nuclease and T4 lysozyme were examined in order to investigate the structural basis of the macroscopic dielectric permittivity from microscopic simulations. Use of QM allowed a realistic representation of electronic polarizability of the proteins, which is otherwise inaccessible because of the use of fixed point charge models in the classical force fields which are typically used to study proteins. The key findings of this study were: dielectric permittivity is not a constant but varies with region of the protein, and its structural and electronic features. It is the highest in the surface and boundary regions and drops off sharply towards the interior of the protein. Electronic polarization, whether due to solvent, or to the protein environment significantly influences permittivity.

(*) Current Address: Department of Pharmaceutical Chemistry, University of California, San Francisco, QB3 Building, Room 501C 1700 4th Street, CA 94143, USA.

(**) Current Address: Department of Chemistry, Quantum Theory Project, University of Florida, 2328 New Physics Building, PO Box 118435, Gainesville, FL 32611-8435, USA.
E-mail: merz@qtp.ufl.edu

1. – Introduction

Dielectric permittivity of proteins is a macroscopic constant that plays an important role in diverse phenomena including enzyme catalysis [1], ion conductance within channels, signal transduction [2,3], electron transfer [4,5], and in ligand binding or molecular recognition [6,7]. The polarizability of the protein interior in response to a solvent reaction field, impacts the strength of the electrostatic interactions between the various interacting functional groups that are responsible for the physico-chemical properties of a protein or an enzyme [8-11]. The physical picture of dielectric permittivity in homogeneous solvents like highly polar water or non-polar chloroform is well understood [12,13]. Proteins, on the other hand, are a complex milieu of charged and neutral groups at physiological pH, and direct methods to measure their dielectric permittivities in the medium they exist in (usually aqueous solution) is nontrivial [14].

Measurements on several dry protein powders have estimated the permittivity of proteins to be in the range of 2-5 [15], but experimental pK_a measurements and mutagenesis studies have hinted at much higher dielectric permittivity in the condensed phase [14,16,17]. Popular continuum models of solvation, that use the Poisson-Boltzmann (PB) equation often treat dielectric permittivities as empirical constants that are set at fixed values for both the solvent (*e.g.*, 78.0 for water) and the solute [18-20]. The internal dielectric of the protein is treated as an adjustable parameter to reproduce an experimental observation, for example, perturbations of the pK_a of buried ionizable residues, or electron transfer reorganization energies [21,22]. A dielectric constant of 4, often used to model the interior of a protein, has been shown by Garcia-Moreno and coworkers [14,16] to underestimate the polarizability experienced by buried groups in proteins. In an illuminating study, Schutz and Warshel used the perturbation of the pK_a of ionizable residues in a set of proteins to underscore the model-dependence of protein dielectric permittivity [21]. Their calculations demonstrate the variation of protein dielectric constants needed to reproduce observed apparent pK_a using microscopic, semi-macroscopic and macroscopic models. The authors have shown that macroscopic models often require a large range from 20 to as much as -60 for the protein dielectric to explain observed apparent pK_a 's of ionizable residues in some proteins. Semi-macroscopic methods, on the other hand, appear to be more consistent in their estimation of protein dielectric constants (which is in the range of 4 to 6) in models that reliably reproduce observed apparent pK_a 's. These apparent ambiguities have spawned considerable interest in the physical origin of the dielectric permittivity of proteins.

Computer simulations have been employed by various groups to gain a better understanding of factors that determine dielectric permittivity [23-26]. All atom molecular dynamics (MD) simulations in conjunction with the Frohlich-Kirkwood (FK) theory have been used to calculate the dielectric permittivity of the solute from the probability distribution of the dipole moment. King *et al.* have addressed the theoretical considerations that relate the electric field to polarizability of the medium within the scope of microscopic simulations, and the FK theory has been successfully used to gain insights into the phenomenon [24,25,27]. Pitera *et al.* have dissected contributions from charged and

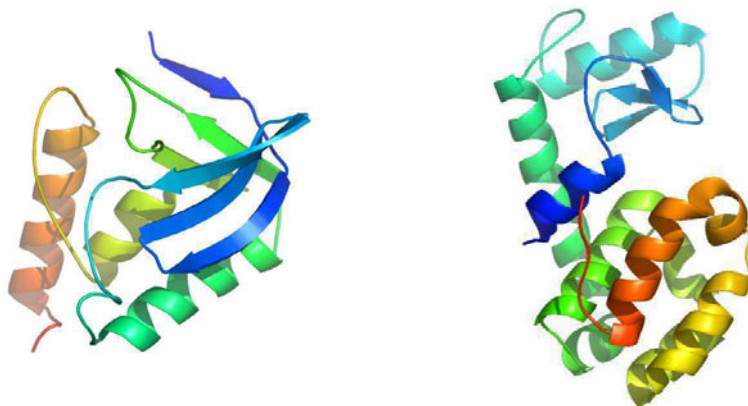


Fig. 1. – Ribbon trace of the backbone of staphylococcal nuclease and T4-lysozyme. The sequence and the secondary structure are also shown in the figure.

neutral groups, ligand binding, and studied the effects of environment such as pH, salt concentration, temperature and non-polar solvents on dielectric permittivity [24].

In this report, we describe impact of introducing the electronic polarizability on the calculation of the dielectric permittivity of a protein using FK theory and MD simulations. An accurate description of electronic polarizability of a multielectron system such as a protein in an MD simulation is a daunting undertaking and can only be achieved by solving the Schrödinger equation for the entire system (protein and solvent) at every time step of the simulation when the forces are calculated. This however, is computationally intractable at the present time. Hence, the electronic polarizability in this study was approximated from atomic charges of the protein calculated using semiempirical electronic structure theory and our version of the linear scaling divide and conquer (D&C) method [28]. We studied the proteins staphylococcal nuclease and T4-lysozyme to get physically based insights into the structural basis of dielectric permittivity (fig. 1). We used gas-phase and solvated partial charges to calculate permittivities that were compared with fixed partial charges. The contribution of charged, uncharged, sidechain, and backbone groups, to dielectric permittivity with respect to polarization was also examined.

2. – Results

In figs. 2a and b, dielectric permittivities during the course of the MD simulations for staphylococcal nuclease, T4 lysozyme are plotted. Staphylococcal nuclease had a high permittivity of ~ 50 (averaged over final 1 ns of the simulation) when the gas-phase CM2 charge model was used. T4 lysozyme had a permittivity of ~ 22 . When a non-polarizable fixed charge model was used, it led to higher dielectric permittivities when compared to QM charges. In the case of staphylococcal nuclease, a dielectric permittivity of ~ 65

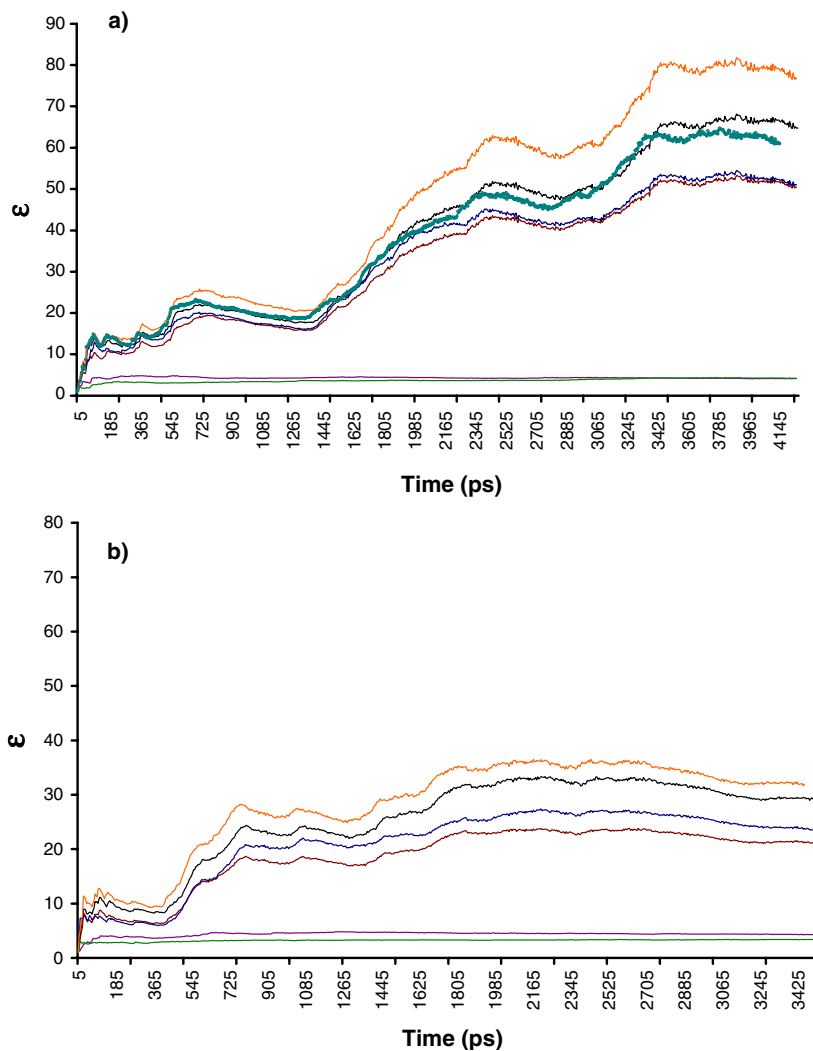


Fig. 2. – a) Dielectric permittivity of staphylococcal nuclease calculated from MD simulation and QM calculations. Color code: AMBER charges (black); gas phase CM2 charges (red); sidechain residues (blue); uncharged residues (purple); backbone residues (green); charge transfer (turquoise); solvated CM2 charges (orange). b) Dielectric permittivity of T4 lysozyme calculated from MD simulation and QM calculations. Color code: AMBER charges (black); gas phase CM2 charges (red); sidechain residues (blue); uncharged residues (purple); backbone residues (green); solvated CM2 charges (orange).

TABLE I. – *Dielectric permittivity averaged over final nanosecond of MD simulation calculated using gas phase CM2 charges and solvated CM2 charges (in parenthesis) in concentric spheres of radius 5 Å, 10 Å, 15 Å, and 20 Å for proteins staphylococcus nuclease and T4 lysozyme.*

Protein	20 Å	15 Å	10 Å	5 Å
Stn. nuclease	30 (45)	10 (15)	3 (3)	1 (1)
T4 Lysozyme	25 (34)	7 (9)	2 (3)	1.5 (1.5)

was obtained using AMBER partial charges. For T4 lysozyme, permittivity of 31 was obtained using AMBER partial charges, which agreed well with the permittivity of 30 previously calculated by Smith *et al.* [26]. This similarity is in spite of differences in the force fields used (GROMOS *vs.* AMBER), simulation protocol, model representation (united non-polar atom model *vs.* all-atom model) and water models (SPC *vs.* TIP3P) between the two studies.

In figs. 2a and b, the effects of various protein groups on the dielectric permittivity during the simulations are also shown. The contribution from groups such as sidechain, backbone, and uncharged residues of the proteins was calculated by zeroing the charges on all other groups. We found that dipole moment fluctuation due to sidechain groups was the single largest contributor to the dielectric permittivities of the proteins. In the case of staphylococcal nuclease the contribution from the sidechain groups led to dielectric permittivity of greater than 50. In T4 lysozyme, including only the sidechain groups led to a permittivity of 25. On the other hand, the permittivities calculated from uncharged and backbone groups of all proteins were very low (~ 4). Similar dependence of dielectric permittivity on charged and sidechain groups has been previously observed [24, 26].

The dielectric permittivity in different protein regions was investigated by dividing the proteins into concentric spheres of radii 5 Å, 10 Å, 15 Å and 20 Å from the center of mass. The partial charges on atoms of residues within the cutoff radius were considered and partial charges on atoms outside the cutoff radius were set to zero. The dipole moment fluctuations during MD simulations were calculated within the spheres using the CM2 charge model. In table I the dielectric permittivities (averaged over final 1 ns of simulation) in these concentric spheres are reported. The common feature was that in the core of the protein, the dielectric permittivity was in the range of 1–5, increasing to higher values in spheres of larger radii. This observation sheds light on the polarizability of the protein cores. Permittivities of 1–5 imply that the core of the protein, even in the case of the staphylococcal nuclease with its buried polar groups, is not as polarizable as the surface regions of the proteins.

In order to account for polarization of electron density in the presence of solvent reaction field, the solvated CM2 (CM2_{sol}) partial charges on the atoms of the protein were calculated from the trajectory during the simulation using the PB/SCRF method [29]. It was found that solute polarization in response to solvent reaction field, increases the dielectric permittivity of the proteins. For staphylococcal nuclease, the permittivity

increased to ~ 80 for the entire protein due to polarization (fig. 1a). Interestingly, the permittivity calculated from a fixed AMBER charge model was 65, and that calculated from the gas phase CM2 (CM2_{gp}) charges was 50. For 20 Å sphere Staphylococcus nuclease permittivity reduced to a value of ~ 45 using CM2_{sol}, which was a significant ($\sim 50\%$) reduction in the permittivity. The corresponding reduction using CM2_{gp} is from 50 to 30 in the 20 Å sphere. For the 15 Å sphere, the permittivity was 15 for CM2_{sol}, compared to 10 for the CM2_{gp} and for the 10 Å sphere; the permittivity was 3 both for CM2_{sol} and CM2_{gp}. The permittivities for the 5 Å spheres are similar at 1 for both CM2_{sol} and CM2_{gp}.

In the case of T4 lysozyme (fig. 2b), the permittivity increased to 33 due to solute polarization, although not as much as in the case of staphylococcal nuclease. It was comparable to permittivity calculated using the AMBER charge model [30]. For the 20 Å sphere, CM2_{sol} calculated permittivity was 34 compared to 25 calculated using CM2_{gp}. Interestingly the reduction in permittivity in the 20 Å sphere was not as drastic as observed in Staphylococcus nuclease. For the 15 Å sphere, the permittivity calculated using CM2_{sol} was 9 compared to 7 calculated with CM2_{gp}. For the 10 Å sphere and 5 Å sphere it was ~ 3 and 1.5, respectively, compared to 2 and 1.5 calculated using GP-CM2.

The dielectric properties of staphylococcal nuclease are of special interest because of the experimental studies that have attempted to elucidate the dielectric constant in the protein core by pK_a -based approaches. Notably, the work done by Garcia-Moreno *et al.* [14] provided guidance for our computer experiments. Garcia-Moreno *et al.* have calculated the dielectric phenomenologically, as the dielectric constant in the Born equation that reproduces the observed pK_a shift of a buried ionizable residue such as glutamate (Glu) or lysine (Lys) [14, 17]. These residues were engineered as glutamate or lysine point mutations at position 66 in the hydrophobic core of the protein, replacing the existing valine residue. The pK_a shift measured by equilibrium thermodynamics was 4.3 units for Glu in water and 4.9 units for Lys in water. The authors used the Born formalism to obtain a dielectric constant of 10–12 to explain the pK_a shifts from experimental measurements.

In subsequent studies based on experimental thermodynamic measurements and X-ray crystallographic analysis of wild-type and mutant structures, the authors argued that high dielectric constants in the hydrophobic core reflect penetration of water into the core of the protein [17]. In our study on staphylococcal nuclease, it was found that calculations on the entire protein led to very high dielectric permittivities in the range of 50–80. It is well known that high dielectric constants of polar solvents such as water are due to the presence of permanent dipoles that can reorient quickly in response to an applied electric field. Proteins, on the other hand, are conformationally constrained, so high dielectric permittivity of staphylococcal nuclease is surprising.

The dipole moment fluctuation, or $\langle M^2 \rangle - \langle M \rangle^2$, is a slowly converging quantity, as also noted in previous studies [24, 27], and it is possible that in spite of more than 4 ns of equilibrium MD simulation the dipole moment fluctuation had not converged. However in spheres of 20 Å, 15 Å, and 10 Å the dielectric permittivity reduced from a value of 80 to 50, and converged to values of 15 and 3, respectively. This implies that the magnitude

of dielectric permittivity of the entire protein is dependent on the contributions to the dipole moment fluctuation from groups on the surface. In order to trace the origins of high dielectric permittivity of staphylococcal nuclease, the structural aspects of the protein was analyzed further. The total charge on the protein is +10 due to 27 positively charged residues and 17 negatively charged residues. The positively charged residues included 22 Lys and 5 Arg residues and the negatively charged residues consisted of 11 Glu and 7 Asp residues. It was clear from the 3-dimensional structure that large and flexible residues such as lysine and glutamate dominated the charged residues, both positive and negative. Visual inspection and solvent accessible surface area calculations on the protein revealed that lysine and glutamate residues were found on the surface of the protein that allowed these sidechains maximum flexibility.

The secondary structure of staphylococcal nuclease includes 3 α helices, 3 β sheets and 10 turns. It appears that there are unusually large numbers of turns in this protein compared to other proteins. For example, T4 lysozyme has 10 α helices, 4 β sheets and 4 turns. Helices and sheets are structured regions that are conformationally constrained and stabilize a protein by packing the hydrophobic core [31]. On the other hand, loops and turns are conformationally flexible regions in a protein fold that are responsible for domain movement during protein folding and molecular recognition. Hence, the presence of a large number of turns in staphylococcal nuclease would increase the conformationally flexible of staphylococcal nuclease relative to other proteins in general. Interestingly, the preponderance of charged residues on the turns of this protein is higher than in T4 lysozyme. For example, 13% of all charged residues in staphylococcal nuclease are present on the turns and 19% of them are found in α helices and β sheets. By comparison, only 2% of the charged residues are present in the turn regions of T4 lysozyme compared to 23% that are present in helix and sheet regions.

This is further underscored in figs. 3a and b. The residues present beyond the sphere of radius 20 Å (red in fig. 3a) are the charged residues Lys and Glu found in turns (fig. 3b). The residues present beyond the sphere of radius 15 Å (orange in fig. 3a) include 12 Lys and 1 Glu, mostly on loops and turns (fig. 3b). We note that excluding residues beyond the 20 Å sphere in these calculations leads to a reduction in the dielectric permittivity of the protein from 80 to 50, and exclusion of residues beyond 15 Å leads to permittivity of 15. These observations shed light on the origins of high dielectric permittivity of staphylococcal nuclease in these calculations. The presence of charged residues in highly flexible regions of the protein leads to tremendous dipole moment fluctuations and is also responsible for the somewhat non-convergent behavior of dielectric permittivity. In general, this suggests that the presence of charged residues on beta-turns, loops, and other conformationally flexible regions lead a protein to have high dielectric permittivity compared to other proteins in which charged residues are present in the more structurally constrained regions of the protein. In the case of T4 lysozyme, the protein has high α -helical content and the bulk of the positively charged residues are found in α helices and β sheets. The permittivity calculated for T4 lysozyme is significantly lower at 35.

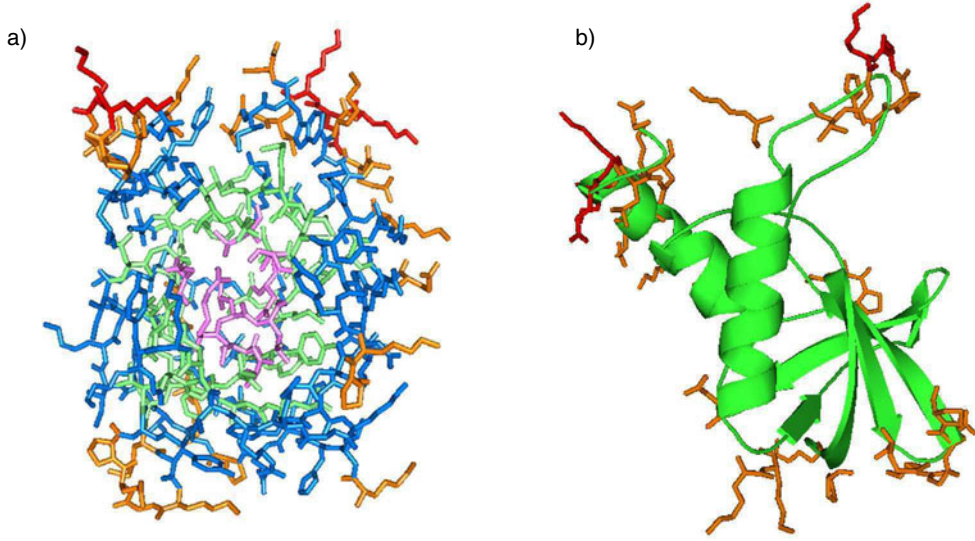


Fig. 3. – a) Residues in concentric spheres of radius 5 Å (purple), 10 Å (green), 15 Å (blue) and 20 Å (orange) in staphylococcus nuclease. Residues beyond 20 Å are colored red. b) Secondary structure of staphylococcus nuclease depicted in green. The residues in the 20 Å shell and beyond are retained in the figure.

3. – Materials and methods

The Fröhlich-Kirkwood theory, which has been presented and used numerous times in the literature was used to calculate effective dielectric permittivity of a solute from computer simulation [23-26]. We will briefly summarize the theory here. For a solute of volume V and permittivity ε embedded in a uniform continuum with permittivity ε_{RF} , the charge distribution can be represented by a point charge and point dipole at its center of mass. If the fluctuations of the solute dipole moment M are obtained from a simulation at temperature T , and volume V of the solute, then Frohlich-Kirkwood [32] theory can be used to calculate the dielectric permittivity in the solute cavity. The dielectric permittivity of a solute is a function of the probability distribution of its total dipole moment fluctuation $p(M)$; specifically the average fluctuation $\langle M^2 \rangle - \langle M \rangle^2$ and is calculated from the charge distribution of a molecular solute such as protein. Hence the fluctuations $\langle M^2 \rangle - \langle M \rangle^2$ are related to the permittivity ε through

$$(1) \quad \frac{\langle M^2 \rangle - \langle M \rangle^2}{3\varepsilon_0 V k_B T} = \frac{(2\varepsilon_{\text{RF}} + 1)(\varepsilon - 1)}{(2\varepsilon_{\text{RF}} + \varepsilon)}$$

which can be rearranged to

$$(2) \quad \varepsilon = \frac{1 + \frac{\langle M^2 \rangle - \langle M \rangle^2}{3\varepsilon_0 V k_B T} \frac{2\varepsilon_{\text{RF}}}{(2\varepsilon_{\text{RF}} + 1)}}{1 - \frac{\langle M^2 \rangle - \langle M \rangle^2}{3\varepsilon_0 V k_B T} \frac{1}{(2\varepsilon_{\text{RF}} + 1)}}.$$

Here ε_0 is the permittivity of vacuum and k_B is the Boltzmann constant. The dipole moment of the molecular system was calculated as

$$(3) \quad M = \sum_{i=1}^N q_i r_i,$$

where N is the number of atoms, q_i is the partial charge on the atom and r_i is the distance of the atom from a fixed origin, which in this case was chosen to be the center of mass. The dipole moment fluctuation can be calculated along each axis in order to eliminate the influence of the choice of the coordinate axis [24, 26]:

$$(4) \quad \langle M^2 \rangle - \langle M \rangle^2 = \langle M_x^2 \rangle - \langle M_x \rangle^2 + \langle M_y^2 \rangle - \langle M_y \rangle^2 + \langle M_z^2 \rangle - \langle M_z \rangle^2.$$

Multi-nanosecond equilibrium MD simulations were performed using the AMBER6 [33] suite of programs on the proteins staphylococcal nuclease, T4 lysozyme (see table I). The protein coordinates were obtained from the protein data bank (PDB) [34] and protons were added to the heavy atoms using the leap module of AMBER6. The sander module was used to minimize hydrogen positions by keeping the heavy atoms fixed using a force constant of 5000 kcal/mol/Å. The proteins were then solvated in a water box using the TIP3P water model and counter ions were added to neutralize the charge on the system. The water molecules and counter ions were then energy minimized by constraining the protein atoms with force constant of 25 kcal/mol/Å. The positional constraints on the protein were then released by doing energy minimization and releasing 5 kcal/mol/Å in constraints for every 600 cycles of minimization followed by 2500 cycles of unconstrained minimization. Then the temperature was raised from 0 K to 300 K and the system was equilibrated for another 20 ps with SHAKE to constrain the bonds involving hydrogen atoms and a time step of 2 fs was used. After the system had reached equilibrium, the simulation was continued for more than 3 nanoseconds and trajectories collected every 5 ps. The details of individual simulations for each protein are given in table II.

After collecting the trajectories from the MD simulation, the configuration at the beginning of the simulation was chosen as a reference state, and subsequent protein coordinates from the MD trajectory were fit to the reference using the least squares fit method of the *carnal* module of AMBER6. The charges on all protein atoms from the individual coordinates of the trajectory were then calculated using the linear scaling divide and conquer (D&C) semiempirical QM program DivCon [35]. AM1 Hamiltonian was used, and Mulliken, CM1 and CM2 charges were obtained for all atoms in a protein. For the D&C calculation, the protein was divided into subsystems corresponding to the

TABLE II. – *Comparison of properties of Staphylococcus nuclease and T4 Lysozyme and their MD simulation details.*

	Staphylococcus nuclease	T4 Lysozyme
PDB ID	5STN	1L54
Number of atoms	2226	2650
Total charge	10.0	9.0
Number of residues	136	164
Number of positive residues	27	27
Number of negative residues	17	18
Vol (nm ³)	19	23
Length of simulation (ps)	4265	3480
Temperature (K)	300	300
Number of waters	6601	8267
Charge ^(*) on 20 Å sphere	7 (3)	7 (2)
Charge ^(*) on 15 Å sphere	−1 (11)	3 (6)
Charge ^(*) on 10 Å sphere	−3 (13)	−1 (10)
Charge ^(*) on 5 Å sphere	0 (10)	2 (7)

(*) Charge in brackets is on residues excluded from the sphere.

constituent amino acid residues. Each residue comprised of one subsystem, and an inner buffer of 4.0 Å and an outer buffer of 2.0 Å were used in the calculations. The CM2 charges obtained from the QM calculations and the coordinates of the proteins corresponding to each snapshot from the trajectories were used to calculate the distribution of the dipole moment ($\langle M^2 \rangle - \langle M \rangle^2$) using eqs. (2) and (3). For the sake of comparison to a non-polarizable model, the dielectric permittivity resulting from fixed atomic charges throughout the simulation was also calculated. The atomic partial charges in this case were taken from the AMBER force field for all amino acid residues. The dielectric permittivity was calculated from the dipole moment fluctuation, as shown in eq. (3). Although the experimental dielectric constant of water at 300 K is 78.5, $\epsilon_{\text{RF}} = 68.0$ was chosen based on previous simple point charge (SPC) simulation studies of water at 300 K and due to the similarity between SPC and TIP3P water models [23, 26, 36]. The only unknown in eq. (2) then was the volume of the protein. Based on the study done by Pitera *et al.* [24] the partial specific volume of proteins and the molecular weight was used calculate the molecular volume.

Multi-nanosecond equilibrium MD simulations were performed using the AMBER6 [33] suite of programs on the proteins under equilibrium conditions. For details of the simulations see materials and methods in the supporting material. We collected the trajectories from the MD simulations, and protein coordinates were fit to the initial reference configuration using the least-squares fit method of the *carnal* module of AMBER6. The partial charges on all protein atoms from the individual snapshots of the trajectory were then recalculated using the linear scaling divide and conquer (D&C) semiempirical QM program DivCon04 [35]. The AM1 Hamiltonian was used, and Mulliken, CM1 and CM2 [37] charges were obtained for all atoms in the protein. These charges were used to calculate the dielectric permittivity from the dipole moment fluctuation ($\langle M^2 \rangle - \langle M \rangle^2$) (see eqs. (1) and (2) in materials and methods). For the sake of comparison to a non-polarizable model, the dielectric permittivity calculated using fixed partial atomic charges from the AMBER force field [38] throughout the simulation is also reported.

4. – Discussion

In this study of dielectric permittivity of proteins, the issue of electronic polarizability and its impact on dielectric permittivity has been addressed using semiempirical QM methods. From these calculations it is clear that the polarization due to environment, whether it is the protein or the solvent, has an important role in determining macroscopic dielectric permittivity. In the deep core regions, dielectric permittivity reduces to values of 1–4, while on the surface it is as high as 80.

Admittedly, phenomenon such as dielectric screening at the surface of proteins ought to be much more pronounced. In contrast, permittivity reduces in boundary regions that are between the surface and the core regions of the protein. From the point of view of protein folding and protein-ligand interaction, burial of an ion pair in the core of a protein, or penetration of a ligand into the boundary and core regions of the protein would compensate for a desolvation penalty paid by charged and polar functional groups of the buried ion-pair or the ligand when they are sequestered from the solvent into a low dielectric medium, provided these groups are involved in specific electrostatic interactions. A dielectric permittivity of 4–15 in the core and boundary regions of the protein imply stronger Coulombic interaction due to reduced charge screening by the medium. However, the energy cost of burying charged groups is directly related to the dielectric permittivity of the medium and, as Garcia-Moreno *et al.* have pointed out, the cost of transferring a charged particle from an aqueous medium to a low dielectric medium of permittivity 2 is 38 kcal/mol, compared to 3 kcal/mol if the permittivity of the medium is 20 [14]. Hence, like other biophysical phenomenon, burial of charged groups in the core of a protein or at a protein ligand interface is possibly a delicate energetic balance influenced by dielectric permittivity.

The chemical and structural complexities of proteins are determinants of its dielectric permittivity. This does not come as a surprise. Ours and previous studies using microscopic simulations have established this. Therefore, use of a single internal dielectric constant for all regions and all proteins in continuum electrostatic models is a cause for

concern. The need for multiple dielectric constants in continuum calculations has been emphasized time and again [21,39-41] and recent reports in the literature have described the development and implementation of algorithms that use site-dependent dielectric constants in continuum electrostatic models [30,42,43]. In this study we demonstrate the use of semiempirical QM calculations with MD simulations and the Fröhlich-Kirkwood theory to understand crucial aspects of permittivity in proteins. In future efforts, we will address the important role of discreteness of interaction of water with proteins in modulating their dielectric permittivity.

REFERENCES

- [1] WARSHEL A., *Annu. Rev. Biophys. Biomol. Struct.*, **32** (2003) 425.
- [2] VARMA S. and JAKOBSSON E., *Biophys. J.*, **86** (2004) 690.
- [3] BASTUG T. and KUYUCAK S., *Biophys. J.*, **84** (2003) 2871.
- [4] BASHFORD D., KARPLUS M. and CANTERS G. W., *J. Mol. Biol.*, **203** (1988) 507.
- [5] CUKIER R. I. and NOCERA D. G., *Annu. Rev. Phys. Chem.*, **49** (1998) 337.
- [6] BIOT C., BUISINE E. and ROOMAN M., *J. Am. Chem. Soc.*, **125** (2003) 13988.
- [7] SHARP K. A., FRIEDMAN R. A., MISRA V., HECHT J. and HONIG B., *Biopolymers*, **36** (1995) 245.
- [8] WARSHEL A. and AQVIST J., *Annu. Rev. Biophys. Biophys. Chem.*, **20** (1991) 267.
- [9] PAUL C. H., *J. Mol. Biol.*, **155** (1982) 53.
- [10] VARDARAJAN R., ZEWEIT T. E., GRAY H. B. and BOXER S. G., *Science*, **243** (1989) 69.
- [11] KRISHTALIK L. I. and TOPOLEV V. V., *Biochim. Biophys. Acta.*, **1459** (2000) 88.
- [12] ONSAGER L., *J. Am. Chem. Soc.*, **58** (1936) 1486.
- [13] KIRKWOOD J. G., *J. Chem. Phys.*, **7** (1939) 911.
- [14] GARCIA-MORENO B., DWYER J. J., LATTMAN E. E., SPENCER D. S. and STITES W. E., *Biophys. Chem.*, **64** (1997) 211.
- [15] PETHIG R., *Dielectric and Electronic Properties of Biological Materials* (Wiley, New York) 1979.
- [16] FITCH C. A., KARP D. A., LEE K. K., SITES W. E., LATTMAN E. E. and GARCIA-MORENO B., *Biophys. J.*, **82** (2002) 3289.
- [17] DWYER J. J., GITTS A. G., KARP D. A., LATTMAN E. E., SPENCER D. S., STITES W. E. and GARCIA-MORENO B., *Biophys. J.*, **79** (2000) 1610.
- [18] ANTOSIEWICZ J., MCCAMMON J. A. and GILSON M. K., *J. Mol. Biol.*, **238** (1994) 415.
- [19] GILSON M. K. and HONIG B. H., *Nature*, **330** (1987) 84.
- [20] HONIG B. and NICHOLLS A., *Science*, **268** (1995) 1144.
- [21] SCHUTZ C. N. and WARSHEL A., *Proteins: Struct. Funct. Genet.*, **44** (2001) 400.
- [22] SHARP K., *Biophys. J.*, **74** (1998) 1241.
- [23] ALPER H. E. and LEVY R. M., *J. Chem. Phys.*, **91** (1989) 1242.
- [24] PITERA J. W., FALTA M. and VAN GUNSTEREN W. F., *Biophys. J.*, **80** (2001) 2546.
- [25] SIMONSON T. and PERHIA D., *Proc. Natl. Acad. Sci. U.S.A.*, **92** (1995) 1082.
- [26] SMITH P. E., BRUNNE R. M., MARK A. E. and VAN GUNSTEREN W. F., *J. Phys. Chem.*, **97** (1993) 2009.
- [27] KING G., LEE F. S. and WARSHEL A., *J. Chem. Phys.*, **95** (1991) 4366.
- [28] DIXON S. L. and MERZ JR. K. M., *J. Chem. Phys.*, **107** (1997) 879.
- [29] GOGONEA V. and MERZ JR. K. M., *J. Phys. Chem. A*, **103** (1999) 5171.
- [30] ROCCHIA W., ALEXOV E. and HONIG B., *J. Phys. Chem. B*, **105** (2001) 6507.

- [31] WONG K. B., CLARKE J., BOND C. J., NEIRA J. L., FREUND S. M., FERSHT A. R. and DAGGETT V., *J. Mol. Biol.*, **296** (2000) 1257.
- [32] FROHLICH H., *Theory of Dielectrics* (Clarendon, Oxford, UK) 1958.
- [33] CASE D. A., CALDWELL J. W., CHEATHAM II T. E., ROSS W. S., SIMMERLING C. L., DARDEN T. A., MERZ K. M., STANTON R. V., CHENG A. L., VINCENT J. J., CROWLEY M., FERGUSON D. M., RADMER R. J., SEIBEL G. L., SINGH U. C. and KOLLMAN P. A. (University of California, San Francisco) 1997.
- [34] BERMAN H. M., WESTBROOK J., FENG Z., GILLILAND G., BHAT T. N., WEISSIG H., SHINDYALOV I. N. and BOURNE P. E., *Nucl. Acid. Res.*, **28** (2000) 235.
- [35] DIXON S. L., VAN DER VAART A., GOGONEA V., VINCENT J. J., BROTHERS E. N., SUÁREZ D., WESTERHOFF L. M. and MERZ JR. K. M., *DivCon Lite* (The Pennsylvania State University) 1999.
- [36] SHEN M. and FREED K. F., *Biophys. J.*, **82** (2002) 1791.
- [37] LI J. B., ZHU T. H., CRAMER C. J. and TRUHLAR D. G., *J. Phys. Chem.*, **102** (1998) 1820.
- [38] CORNELL W. D., CIEPLAK P., BAYLAY C. I., GOULD I. R., MERZ K. M., FERGUSON D. M., SPELLMEYER D. C., FOX T., CALDWELL J. W. and KOLLMAN P. A., *J. Am. Chem. Soc.*, **117** (1995) 5179.
- [39] WARSHEL A., RUSSEL S. T. and CHURG A. K., *Proc. Natl. Acad. Sci. U.S.A.*, **81** (1984) 4785.
- [40] WARSHEL A. and PAPAZYAN A., *Curr. Opin. Struct. Biol.*, **8** (1998) 211.
- [41] MEHLER E. L., FUXREITER M., SIMON I. and GARCIA-MORENO B., *Proteins: Struct. Funct. Genet.*, **48** (2002) 283.
- [42] SIMONSON T., *Int. J. Quantum Chem.*, **73** (1999) 45.
- [43] WISZ M. S. and HELLINGA H. W., *Proteins: Struct. Funct. Bioinfor.*, **51** (2003) 360.

This page intentionally left blank

Comparisons of protein electrostatic potentials as a means to understanding function

P. J. WINN and R. C. WADE

*EML-Research gGmbH - Villa Bosch, Schloss-Wolfsbrunnengasse 33
69118 Heidelberg, Germany*

M. ZAHRAN

*EML-Research gGmbH - Villa Bosch, Schloss-Wolfsbrunnengasse 33
69118 Heidelberg, Germany*

*Département de Bioinformatique Génomique et Moléculaire
Université Denis Diderot, Paris 7 - Paris, France*

1. – Introduction

Proteins perform diverse chemical and structural functions in biology, and these are dependent on the physico-chemical properties of the protein. The function of every protein is linked to the three-dimensional arrangement of the amino acids from which it is constructed. For each protein, the type and order of amino acids is read from the cellular DNA and this information is used to synthesise the protein. Mutations in DNA may change the amino acid sequence of a protein. This will lead to different physical properties. Mutations may lead to different mechanisms for folding a protein's amino acid chain and even to folds. Mutations may result in subtle changes in interaction properties, such as the rate of ligand binding, or the specificity for protein or small molecule binding partners. Catalytic mechanisms may be subtly changed, or catalytic function may be lost. Changes in protein properties/function may in turn affect the function of cells and thus of an organism, finally leading to adaptive advantage or disadvantage and evolutionary

selection. Evolutionary changes in the physico-chemical properties of a protein are thus the connection between protein amino acid sequence mutation and function.

Evolution can be studied by comparative genomics at the DNA level. Sequences carry snapshots of the historical record of molecular evolution. Statistical methods can be applied to develop phylogenies and understand gene duplication and transfer events. The same techniques may be applied to protein amino-acid sequences, since, for a protein, these are what the DNA sequence encodes. It is possible to rationalise functional variation by amino acid sequence analysis, the assumption being that similar sequences have related evolutionary history and thus function. However, such interpretations may be misleading [1], particularly in predicting protein-protein interactions [2]. Analysis of the three-dimensional structures of proteins, and their associated physico-chemical properties connects sequence to physics and chemistry, and thus to protein function. Such analyses can be time consuming but are often very informative [3-5]. Here we discuss some examples where the electrostatic potentials of proteins were found to be correlated with their function. We then highlight examples from ubiquitination and related pathways where electrostatic potential appears to be an important modulator of function. We review the high-throughput modeling of the ubiquitin conjugating enzymes and the systematic analysis of the resulting structures and their electrostatic potentials.

2. – Calculation of electrostatic potentials in biomolecular systems

If biomolecules are simulated using all-atom representations, then electrostatic interaction energies are usually evaluated with a Coulombic term. This is typically the case for molecular dynamics simulations. However, biological systems are large, and all-atom simulations are often too costly. For this reason, solvent effects are often treated in an average way by using dielectric constants suitable for solvent (~ 80 for water) and solute (typically 4 for protein). The Poisson equation is then appropriate. Implicit inclusion of ions can be achieved by adding a Boltzmann term dependent on ion concentration, ion charge and electrostatic potential in space. This is implemented in the Poisson-Boltzmann equation

$$(1) \quad -\nabla \cdot \varepsilon \nabla \phi = \rho + \sum_i q_i n_i e^{q_i \phi / kT},$$

where ε is the dielectric constant, varying over space, ϕ the electrostatic potential, ρ the fixed charge density, q_i the charge on an ionic species, n_i the concentration of the i -th ionic species, k the Boltzmann constant and T the temperature.

The Boltzmann exponential is often expanded as a Taylor series; the linearised Poisson-Boltzmann equation is then derived by truncation to the first term of the Taylor series. Numerically the Poisson-Boltzmann and linearised Poisson-Boltzmann equations are often solved via use of finite-difference methods on a grid on and around the system of interest. Such implementations are used in, *e.g.*, Brownian dynamics simulations and pK_a calculations. A broader discussion of the Poisson-Boltzmann equation, its numerical solution and its application can be found elsewhere [6].

3. – Examples of the use of electrostatic potentials in biomolecular systems

Complementary electrostatic potentials between interacting protein surfaces are to be expected; although ions or water molecules trapped at a protein interface can bridge between residues that would be otherwise electrostatically repulsive. A hydrophobic patch on a protein surface will tend to have a small electrostatic potential, compared to an acidic, basic or polar patch. Thus, the electrostatic potential can also capture aspects of hydrophobic interaction properties. A study of 12 protein-protein interfaces found anti-correlation (complementarity) between the electrostatic potentials of interacting proteins [7]. However, no correlation was found between the charges at the interfaces, implying that charges distant from the interface are important in establishing electrostatic complementarity.

Ullmann *et al.* [4] superposed the physiologically equivalent but structurally divergent proteins cytochrome c6 and plastocyanin on the basis of their electrostatic potentials. They found that the aromatic residues tryptophan 63 in cytochrome c6 and tyrosine 83 in plastocyanin likely have an equivalent function. A functional role for tryptophan 63 is supported by the conservation of this position as an aromatic residue. The same technique was used to compare ferredoxin and flavodoxin, which have very different structures. Under conditions of iron deficiency, flavodoxin fulfils the function of the iron-sulfur protein ferredoxin. Optimal alignment of their electrostatic potentials leads to an overlap of their active sites, rather than their centres of mass, and is in agreement with experimental data [3].

Acetylcholinesterase degrades acetylcholine that has been released at the synaptic junction, thus terminating signal transmission. Its electrostatic potential is important for this classical function [8]. Acetylcholinesterase (AChE) is thought to have a non-catalytic role in the developing nervous system that is also dependent on its electrostatic properties. An electrostatic motif is conserved in the cholinesterases and in structurally similar neural cell-adhesion proteins [9]. Recent work supports a role for this conserved electrostatic patch in interactions with the proteins laminin-1 and collagen IV [10].

Copper zinc superoxide dismutase (CuZnSOD) is a highly efficient enzyme. Its electrostatic potential is important in guiding the superoxide substrate to its active site; enhancing this electrostatic guidance has been shown to make the enzyme even more efficient [11]. The electrostatic potential of CuZnSOD is highly conserved across species [5, 12], although residues identified as being important for the electrostatic potential of CuZnSOD in one species are not strictly conserved in other species. Mutations are usually accompanied by compensatory mutations elsewhere in the enzyme, and the evolution of charged residues mirrors the evolution of different species [5]. For example the prokaryotic enzymes have an additional loop (KDxK), compared to the eukaryotic enzymes, whereas other charged residues deemed important for eukaryotic CuZnSOD function are not present [5, 12]. Livesay *et al.* [5] show that the spatial charge distribution within the CuZnSOD superfamily is conserved.

Electrostatic potentials have been used to estimate rates of protein association [13] and also enzymatic rate constants for use in simulations of enzyme kinetics in metabolic

pathways [14]. A well-studied case is barnase and its inhibitor barstar. These two proteins have highly complementary electrostatic potentials, and variation in the rates of association of different mutants can be correlated with the rate constants computed from Brownian dynamics simulations, using a linearised Poisson-Boltzmann representation of the electrostatic potential [15]. Elsewhere, it has been shown that the rate of electron transfer between plastocyanin mutants and cytochrome *f* can be related to the mutants' relative electrostatic similarities [16]. The above examples indicate the importance of electrostatic potentials to protein function. There are many more examples in the literature.

4. – Comparison of protein electrostatic potentials: Protein Interaction Property Similarity Analysis (PIPSA)

A systematic comparison of the electrostatic potentials of a family of proteins should allow quantitative and qualitative classification of proteins. Measures of similarity have long been used to compare and predict the properties of small organic compounds, *e.g.*, as an aid to drug discovery [17]. The similarity between the molecular interaction fields of proteins can be determined by PIPSA [18,19]. “Molecular interaction field” indicates some property that mediates an intermolecular interaction; this may be, *e.g.*, the hydrophobicity of the protein surface or the electrostatic potential. PIPSA is implemented with two measures of similarity, the Hodgkin index (SI_{Hodgkin}) and the Carbo index (SI_{Carbo}), eqs. (2) and (3). Both indices evaluate the overlap integral between the fields, Φ_a and Φ_b , of two molecules, *a* and *b*, but they apply different normalisation factors. SI_{Hodgkin} has a value of +1 when the potentials are identical, and −1 when they are the exact inverse of each other. SI_{Carbo} has a value of +1 as long as the two potentials are correlated, but not necessarily of the same magnitude, and −1 if they are anti-correlated.

$$(2) \quad SI_{\text{Hodgkin}} = \frac{2 \int \Phi_a \Phi_b \, dv}{\int \Phi_a \, dv + \int \Phi_b \, dv},$$

$$(3) \quad SI_{\text{Carbo}} = \frac{\int \Phi_a \Phi_b \, dv}{(\int \Phi_a \, dv)^{1/2} (\int \Phi_b \, dv)^{1/2}}.$$

PIPSA analysis is performed on superposed protein structures. Comparison of electrostatic potentials is usually via solution of the linearised Poisson-Boltzmann equation, which is performed on a grid. The potential of two proteins is then compared at the points on the grid that are within a skin around the surface of each protein, and the similarity index calculated. The skin is defined by a distance from the molecular surface (σ) and a thickness (δ), as shown in fig. 1. Further details of the method and its applications can be found elsewhere [18].

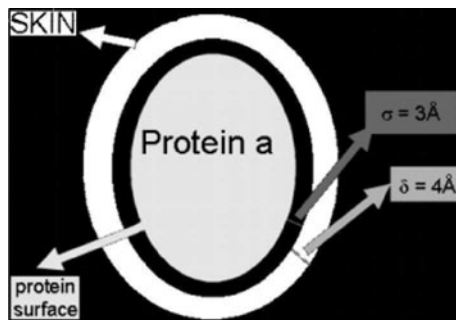


Fig. 1. – Defining a molecular skin for use in PIPSA. PIPSA analysis is performed in a skin around each of the proteins of interest. The skin is defined by a distance σ from the molecular surface, and a skin thickness of δ . These are typically 3 and 4 Å, respectively. The bounding surfaces of the skin are calculated as the position of the centres of spheres of radius σ and $\sigma + \delta$ rolled over the protein surface. The electrostatic potential is calculated at all points on a grid around each protein. All grid points within the skin of some protein, protein *a*, are then compared with those within the skin of some other protein, protein *b*, and the Hodgkin and Carbo similarity indices are calculated according to eqs. (2) and (3).

5. – Electrostatic potentials in the ubiquitin and ubiquitin like systems

Ubiquitin family protein (UFP) ligation pathways regulate many cellular events. Examples include degradation of defective proteins, cell cycle progression, endocytosis and DNA repair, but there are many other functions that are less well understood [20]. Ubiquitin is a 76 amino acid protein. It was the first UFP to be discovered, and its ligation pathway is the most studied. The other UFPs are known as ubiquitin-like proteins (UBL), and include the now well-studied SUMO and Nedd8 [20]. The C-terminus of a UFP is ligated, via an isopeptide linkage, to the $N(\epsilon)$ atom of a lysine residue of the target protein. The post-translational modification alters the protein interaction properties of the target protein and thus changes its cellular role. Dysfunction of ubiquitin ligation has been implicated in many diseases, including cancer and neurodegeneration [21]. As well as UFPs, other multidomain proteins have small folding domains that have the same topology as ubiquitin. These will be referred to here as ubiquitin fold domains. They appear to function as protein recognition modules, often in a UFP pathway [22].

The mechanism by which UFPs are ligated to their targets is thought to be similar. Three enzymatic stages are required. An ATP-dependent activating enzyme (E1) forms a thiol ester bond with the terminal glycine of the ubiquitin family protein (UFP). The UFP is transferred from the E1 to a UFP conjugating enzyme (Ubc or E2). This then transfers the UFP to a target substrate, directly or with the help of a ligase enzyme (E3). A further enzyme (E4) may also be required for the formation of multiubiquitin chains of length greater than four [23].

UFPs are present in all eukaryotes. The importance of the UFPs has led to much research activity over recent years, and this seems likely to continue. This makes the UFP

ligation pathways a very interesting system for understanding the relationship between electrostatic potential, protein function and protein interactions. Variations in electrostatic potential between functionally different UFPs can be compared, and examined in an evolutionary context. Isofunctional UFP pathways can be examined for electrostatic conservation or deviation. This should lead to insights into the evolutionary process, and also help predict protein function.

Analysis of the electrostatic potentials of all UFPs (57) and ubiquitin fold domains (58) available (7th August 2006) in the Protein data bank (www.rcsb.org) was performed with PIPSA. Each domain was fit onto the SUMO structure with pdb id 1A5R. 1A5R was truncated to Y21 to E93, and this was designated 1A5R_short. Residues beyond this range are variable in structure. All other UBL structures were truncated to match this length, *e.g.*, ubiquitin was considered from residue 1 to 72. N- and C-ter were treated as neutral, being neither blocked or charged. Missing side chain atoms were automatically added using the program SPDV. A dendrogram (Epogram) of the similarity of their potentials is shown in fig. 2. There are distinct clusters of proteins with similar electrostatic potentials. Ubiquitin and Nedd8 have very similar potentials, and cluster together on the epogram. Experimentally, it has been seen that the selectivity of Nedd8 and ubiquitin for their respective ligation pathways is dependent on one amino acid. Mutation of this amino acid (A72 in Nedd8, R72 in ubiquitin) to the corresponding amino acid in the other UFP results in a change of pathway specificity [24]. The different SUMO and ubiquitin orthologs have anti-correlated electrostatic potentials with an SI_{Hodgkin} of about -0.5 . The Epogram suggests which pathways the different ubiquitin fold domains interact with; for example, the domains with electrostatic potentials similar to SUMO likely interact with the SUMO pathway. The anti-correlation between the SUMO and ubiquitin potentials is also reflected in the E2s that are specific for SUMO. Figure 3 shows that the Ubc9 orthologs, responsible for the conjugation of SUMO, form a distinct cluster separate from the other E2s. This gives an indication of how electrostatic potential might be used to distinguish between different UFP ligation pathways. Since SUMO may be ligated to lysines to block their ubiquitination [25], a large difference between SUMO and ubiquitin may be advantageous.

Comparative sequence analysis of the E2s gives different results to those of the electrostatic potential analysis [26]. This can be seen in fig. 3, in the family of E2C proteins, which are all highly similar in amino acid sequence but have very different electrostatic potentials. Elsewhere, little correlation was found between pairwise amino acid sequence similarity and pairwise electrostatic potential similarity [19,27]. To understand this consider some hypothetical protein; the mutation of a few negatively charged residues to positively charged residues would create a large change in electrostatic potential but a small change in sequence similarity. Thus, it is easy to see why there is often no strong correlation between sequence identity and electrostatic potential.

Since interacting protein partners may co-evolve it is not clear when evolution will tend to conserve electrostatic potentials. Proteins that interact with cellular membranes, or DNA presumably have a constraint placed on them due to the conserved charged nature of these surfaces. Similarly certain physical processes such as electron transfer

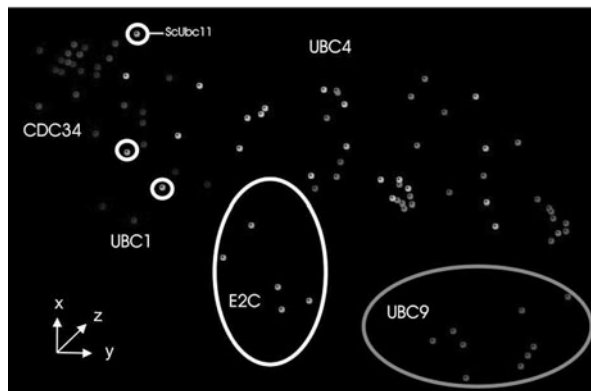


Fig. 3. – A representation of the E2 enzymes electrostatic similarity. The picture is a view of a three-dimensional representation of the pairwise electrostatic similarity matrix. Each point represents a protein. The sequences with high similarity to ScUbc9 form a cluster in the bottom right corner (marked by a grey ellipse), indicating that they all have similar electrostatic properties. The sequences similar to clam E2-C are circled in white (one large white ellipse, and 3 small circles). They do not form a clear cluster, notably ScUbc11 (top left) is a clear outlier. This is discussed in more detail in the main text. This representation is from an analysis of 211 E2 enzymes. The 211 by 211 matrix of pairwise comparisons was projected into a three-dimensional space. The volume of the space is defined by taking the two points furthest apart (*i.e.* the two proteins with least similarity); the point that defines the largest area is then added to the first two, and finally the point that defines the largest volume. All other points are projected into this volume. Thus, two points close in space have similar electrostatic properties, and those far apart should be very different. The plot here represents a subset of the 211 proteins, full results can be found elsewhere [26].

or enzymatic catalysis may also impose evolutionary constraints on the electrostatic potential of a protein. In the case of two or more interacting proteins, it is possible for correlated mutations to occur, and for their electrostatic potentials to mutate with respect to other species.

The E2s that are phylogenetically clustered with the clam protein E2-C are circled in fig. 3 and labelled E2C. They are thought to lead to the ubiquitination and thus degradation of mitotic cyclins [28–31]. Destruction of cyclin B is thought to be necessary for exit from mitosis, and thus continuation of the cell cycle [32]. However ScUbc11, the *S. cerevisiae* (baker's yeast) sequence with closest sequence identity to E2-C, is not essential and does not ubiquitinate cyclin B in a cell-free system [33]. ScUbc11 lacks a 22 amino acid N-terminal extension compared to the other E2-C orthologs. Whilst this is a possible reason for ScUbc11's different behaviour [33], this extension is not well conserved in the other E2-C orthologs. ScUbc11 also has an electrostatic potential that is divergent from all of its sequence orthologs (fig. 4 A–D) and is thus unlikely to interact with the same proteins. In clam cell-free extracts, the other orthologs ubiquitinate mitotic cyclins; thus, based on its electrostatic potential, we expect ScUbc11 not to. However, the machinery in yeast that ubiquitinates mitotic cyclins may have evolved a different

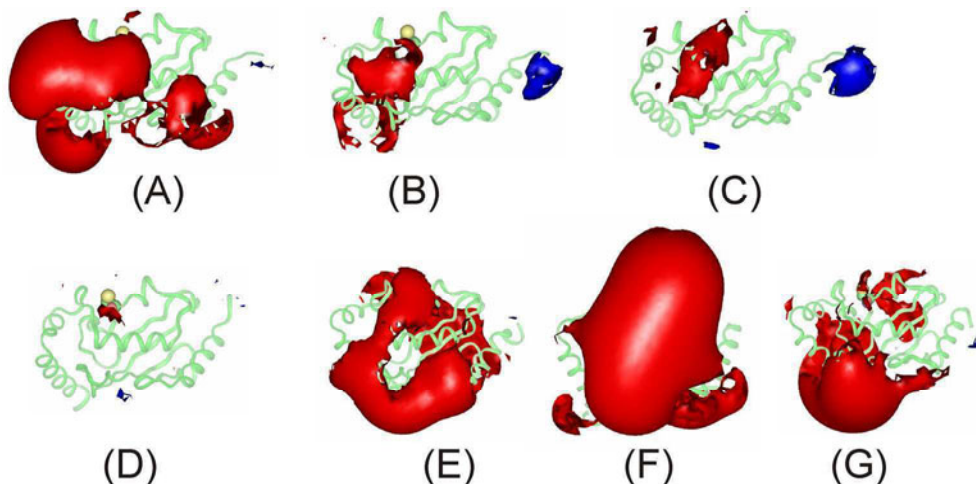


Fig. 4. – The isotopotential contours around the E2s most similar to clam E2-C (A-D) and around E2s known to ubiquitinate histones *in vitro*. The protein structures are shown as a backbone ribbon, with the atoms of the catalytic cysteine shown as spheres, the cysteine sulphur is coloured yellow. The electrostatic potential contours are at $\pm 2 \text{ Kcal Mol}^{-1} \text{ e}^{-1}$ (red negative, blue positive). The potentials were calculated at zero ionic strength to emphasise the differences/similarities in potential (solution of the Poisson equation). The same qualitative conclusions are evident even when calculations are performed at 100 mM ionic strength [26]. E2-C sequence orthologs: (A) ScUbc11, the sequence ortholog from *S-cerevisiae*, (B) Ubcp4, the ortholog from *S-pombe*, (C) E2-C from clam, (D) Ubh10, the human ortholog. Ubiquitinators of histones from *S. cerevisiae*: (E) ScUbc8 (F) ScUbc3 (G) ScUbc2.

electrostatic potential from that in clam, and thus ScUbc11 may still be able to interact with this machinery. ScUbc11 knockouts are viable, but no other yeast E2 has been found to be solely responsible for cyclin B ubiquitination, suggesting that another E2 may take over the ScUbc11 role once it has been knocked out. As this example illustrates, high-throughput comparison of protein electrostatic potentials is likely to be useful for experimental design and interpretation. They allow proteins that are likely to have similar behaviour to be selected for an assay. They also allow the selection of proteins that are likely to functionally differ from known proteins.

ScRad6p, ScCdc34p and the ScUbc8p have all been shown to ubiquitinate histones (H2A or H2B) *in vitro* without requiring an E3 [34-37], although the ScUbc8p result is unclear [38]. All three of these proteins have similar strongly negative electrostatic potentials (fig. 4, E-G) and acidic C-terminal extensions (not shown). Indeed, the ability of ScCdc34p and ScRad6p to ubiquitinate histones *in vitro* has been shown to be dependent on the presence of their C-terminal tail [35,39]. This suggests that E2s with a strong negative potential will ubiquitinate histones *in vitro*, presumably a consequence of the highly positive potential associated with histones. *In vivo* it appears that only Rad6 is responsible for histone ubiquitination [37].



Fig. 5. – A snapshot of a search for the five E2s with highest electrostatic potential similarity to a query, Uniprot code Q9ZU75. A cartoon of the related structures, their potential on the solvent accessible surface and their isopotentials are returned. Each entry also has links to other potentially relevant data such as protein interactions (DIP) [44] and mutation data (Prosa 2) [43]. The potentials were calculated by solution of the linearised Poisson-Boltzmann equation with ionic strength 50 mM. The isopotential contours are at ± 0.5 kT, the solvent accessible surface representation is coloured from red ($-kT$) to white (neutral) to blue ($+kT$).

6. – High-throughput modelling of protein electrostatic potentials

To comprehensively investigate the E2 proteins we have developed a computational pipeline to model their structures and electrostatic potentials. E2 sequences are taken from the SMART database [40] as it is updated. They are then aligned onto a set of sequence profiles and modelled by comparison with known structures. The known structures require some pre-refinement as discussed elsewhere [41]. The electrostatic potentials are calculated with the linearised Poisson Boltzmann equation, using the program UHBD [42]. The data are presented on the website www.ubiquitin-resource.org. This allows the user to query with Uniprot and Ensembl codes. The user may ask for the E2s that are electrostatically most similar to a query, or may ask for a comparison of the electrostatic potentials of a query list. The result is a webpage with pictures of the electrostatic potentials of the E2s and the similarity index of the protein pairs (fig. 5). The user may download the models of the E2 structures and the grid file of the electrostatic potential, for displaying locally. Amino-acid mutation data may be retrieved from databases such as Swissprot (www.expasy.org) and displayed on the modelled struc-

tures using Prosat2 [43]. There are also links to the “Database of Interacting Proteins” (DIP) [44], to the sequences’ Uniprot or Ensembl entry and to various data on the model quality. The website gives researchers in the field of ubiquitination easy access to the calculated electrostatic potentials. It allows them to compare potentials in the same way that they currently compare sequences. This will thus allow the community to have a better understanding of how electrostatic potentials are related to function and to evolution.

There are many areas where care is required when performing calculations for the comparison of electrostatic potentials and these are discussed in more detail elsewhere [14, 41]. Simple changes in side chain rotamers between different crystal structures may cause problems with quantitative predictions of enzyme kinetic parameters [14]. For the clustering analysis that we have performed for the E2s we have found that models have whole surface electrostatic potentials similar to their crystal structures [41]. This suggests that this is less of a problem in this case, although care is still need when considering, *e.g.*, histidine protonation [41]. Small misalignments during the homology modelling have also been shown to be tolerable [26]. Other possible problems are variations in loop conformations between experimental structures and, for comparative modelling, unexpected structural elements compared to the template.

Inconsistencies and errors in the models produced by the pipeline will make it difficult to meaningfully compare the electrostatic potentials of different structures. Checks are made on both the sequences and the templates used in the modelling process [14, 41]. For fully automated modelling further checks are required on the resulting models. The sequence alignment of model and template are compared to determine if their N- and C-terminii are not congruent, which might indicate an incomplete model sequence. An incomplete model will not give electrostatic potentials useful for comparison with other structures, although it may be useful for structural interpretation. Similarly, models with an insertion or deletion of more than five amino acids, ones that have unknown amino acid types (an X in the sequence entry), or ones that have abnormal structural characteristics [41, 45] are also not used for electrostatic calculations. Suitable care in the modelling and comparison process has allowed us to produce meaningful results, as discussed above and elsewhere [14, 26, 41].

7. – Conclusions

Electrostatic interactions are important mediators of protein interactions. There are now many cases where analysis of electrostatic potentials has been used to provide greater insight into biological systems. PIPSA allows high throughput analysis of the electrostatic similarity between different proteins. This is useful for quantitative predictions, *e.g.*, of enzyme catalytic constants, and for qualitative classification. This allows for a greater insight into the evolutionary constraints on protein electrostatic potentials and on their use for protein function prediction. Analysis of electrostatic potentials is more complicated than the analysis of sequence, because of problems handling structural variation and flexibility. Nonetheless good results have been obtained, and a long-term goal would be to compare electrostatic potentials as routinely as sequences currently are.

* * *

We thank AMIT BANERJEE for productive discussions. This work was supported by the Klaus Tschira Foundation and the National Institute of Health (USPHS) grant GM59467.

REFERENCES

- [1] GERLT J. A. and BABBITT P. C., *Genome Biol.*, **1** (2000) reviews0005.1.
- [2] MIKA S. and ROST B., *PLoS Comput. Biol.*, **2** (2006) e79.
- [3] ULLMANN G. M., HAUSWALD M., JENSEN A. and KNAPP E. W., *Proteins*, **38** (2000) 301.
- [4] ULLMANN G. M. *et al.*, *Biochemistry*, **36** (1997) 16187.
- [5] LIVESAY D. R., JAMBECK P., ROJNUCKARIN A. and SUBRAMANIAM S., *Biochemistry*, **42** (2003) 3464.
- [6] FOGOLARI F., BRIGO A. and MOLINARI H., *J. Mol. Recognit.*, **15** (2002) 377.
- [7] MCCOY A. J., EPA V. C. and COLMAN P. M., *J. Mol. Biol.*, **268** (1997) 570.
- [8] RADIC Z. *et al.*, *J. Biol. Chem.*, **272** (1997) 23265.
- [9] BOTTI S. A., FELDER C. E., SUSSMAN J. L. and SILMAN I., *Protein Eng.*, **11** (1998) 415.
- [10] JOHNSON G. and MOORE S. W., *Neurosci. Lett.*, **337** (2003) 37.
- [11] GETZOFF E. D. *et al.*, *Nature*, **358** (1992) 347.
- [12] WADE R. C., GABDOULLINE R. R., LUDEMANN S. K. and LOUNNAS V., *Proc. Natl. Acad. Sci. U.S.A.*, **95** (1998) 5942.
- [13] GABDOULLINE R. R. and WADE R. C., *Curr. Opin. Struct. Biol.*, **12** (2002) 204.
- [14] STEIN M., GABDOULLINE R. R., BESSON B. and WADE R. C., in *Experimental Standard Conditions of Enzyme Characterizations, Proceedings of the 2nd International Beilstein Symposium*, edited by M. G. HICKS and C. KETTNER (Beilstein Institute, Ruedesheim) 2006.
- [15] GABDOULLINE R. R. and WADE R. C., *Biophys. J.*, **72** (1997) 1917.
- [16] DE RIENZO F. *et al.*, *Biophys. J.*, **81** (2001) 3090.
- [17] GOOD A. C., in *Molecular Similarity in Drug Design*, edited by DEAN P. M., (Blackie Academic & Professional, London) 1995, pp. 1.
- [18] WADE R. C., GABDOULLINE R. R. and DE RIENZO F., *Int. J. Quantum Chem.*, **83** (2001) 122.
- [19] BLOMBERG N., GABDOULLINE R. R., NILGES M. and WADE R. C., *Proteins*, **37** (1999) 379.
- [20] WELCHMAN R. L., GORDON C. and MAYER R. J., *Nat. Rev. Mol. Cell. Biol.*, **6** (2005) 599.
- [21] REBOUD-RAVAUX M., *Protein Degradation in Health and Disease* (Springer Verlag, Heidelberg) 2002.
- [22] HUANG D. T. *et al.*, *Mol. Cell.*, **17** (2005) 341.
- [23] KOEGL M. *et al.*, *Cell*, **96** (1999) 635.
- [24] WALDEN H. *et al.*, *Mol. Cell.*, **12** (2003) 1427.
- [25] DESTERRO J. M., RODRIGUEZ M. S. and HAY R. T., *Mol. Cell.*, **2** (1998) 233.
- [26] WINN P. J. *et al.*, *Structure*, **12** (2004) 1563.
- [27] DE RIENZO F., GABDOULLINE R. R., MENZIANI M. C. and WADE R. C., *Protein Sci.*, **9** (2000) 1439.
- [28] TOWNSLEY F. M. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **94** (1997) 2362.
- [29] OSAKA F., SEINO H., SENO T. and YAMAO F., *Mol. Cell. Biol.*, **17** (1997) 3388.
- [30] TOKUMOTO M., NAGAHAMA Y. and TOKUMOTO T., *FEBS Lett.*, **458** (1999) 375.

- [31] YU H., KING R. W., PETERS J. M. and KIRSCHNER M. W., *Curr. Biol.*, **6** (1996) 455.
- [32] KING R. W., DESHAIES R. J., PETERS J. M. and KIRSCHNER M. W., *Science*, **274** (1996) 1652.
- [33] TOWNSLEY F. M. and RUDERMAN J. V., *Yeast*, **14** (1998) 747.
- [34] KAISER P. *et al.*, *J. Biol. Chem.*, **269** (1994) 8797.
- [35] GOEBL M. G. *et al.*, *Science*, **241** (1988) 1331.
- [36] JENTSCH S., MCGRATH J. P. and VARSHAVSKY A., *Nature*, **329** (1987) 131.
- [37] ROBZYK K., RECHT J. and OSLEY M. A., *Science*, **287** (2000) 501.
- [38] QIN S., NAKAJIMA B., NOMURA M. and ARFIN S. M., *J. Biol. Chem.*, **266** (1991) 15549.
- [39] SUNG P., PRAKASH S. and PRAKASH L., *Genes Dev.*, **2** (1988) 1476.
- [40] LETUNIC I. *et al.*, *Nucleic Acids Res.*, **34** (2006) D257.
- [41] WINN P. J. *et al.*, *Proteins*, **58** (2005) 367.
- [42] MADURA J. D. *et al.*, *Comput. Phys. Commun.*, **91** (1995) 57.
- [43] GABDOULLINE R. R., ULBRICH S., RICHTER S. and WADE R. C., *Nucleic Acids Res.*, **34** (2006) 79.
- [44] SALWINSKI L. *et al.*, *Nucleic Acids Res.*, **32** (2004) D449.
- [45] HOOFT R. W., VRIEND G., SANDER C. and ABOLA E. E., *Nature*, **381** (1996) 272.

This page intentionally left blank

Computational structural proteomics of the kinases binding specificity and drug resistance

G. M. VERKHIVKER

*Department of Pharmacology, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093-0392, USA*

1. – Introduction

Cataloging the protein kinase complement of the human genome (the “kinome”) has projected human genome analysis through a focus on one large gene superfamily, providing important insights into classification and evolution of human protein kinases [1-5]. The human genome analysis has also facilitated a detailed comparison of related kinases in human and model organisms and a comprehensive analysis of protein phosphorylation in normal and disease states [6,7]. Protein tyrosine kinases, which are responsible for the development of many cancers have emerged as a major class of drug targets for many different therapies [8,9]. These phosphorylated tyrosines serve as docking sites for recruiting downstream signaling molecules, including non-receptor tyrosine kinases that trigger a variety of cellular responses and cross-talk between signalling pathways, which is deregulated in various diseases. The highly conserved nature of the ATP binding site, present in more than 500 identified protein kinases, has appeared as the central concern in development of cancer therapeutics as selectivity mechanisms of these pharmacological agents against protein kinase targets remain elusive [10-12]. Evolutionary relationships between protein kinases are often employed to infer inhibitor binding profiles from sequence analysis [13,14]. Classical selectivity analysis based on sequence homology of protein kinases assumes that phylogenetically similar kinases should display similar inhibitor binding profiles. However, protein kinases binding with small molecules may display selectivity within a given kinase subfamily, while exhibiting cross-activity with kinases that are

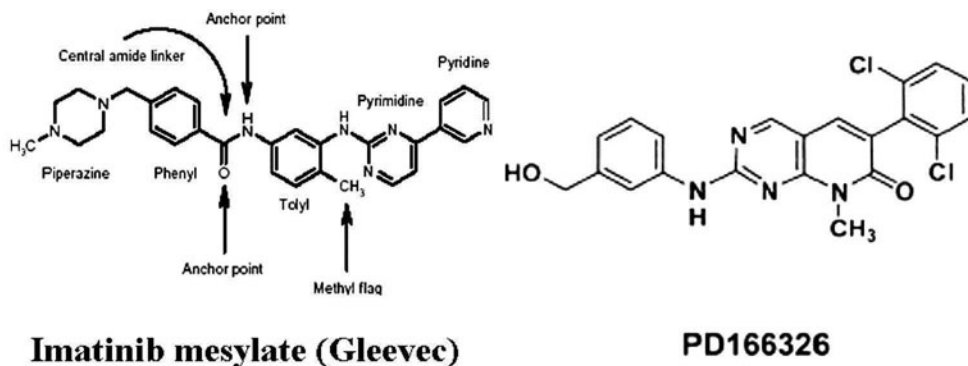


Fig. 1. – Chemical structures of STI-571 (Gleevec) and PD166326 pyrido[2,3-d]pyrimidine kinase inhibitors studied in this work.

phylogenetically remote from the prime target. The recently proposed chemogenomic classification of kinase space based on small molecule selectivity data generally differs from sequence-based dendrogram analysis, but remains comparable for closely homologous targets. The emerging insights into kinase function and evolution combined with a rapidly growing number of publically available crystal structures of protein kinases complexes have motivated computational analysis of sequence-structure relationships in determining the binding function of protein tyrosine kinases.

Imatinib mesylate (Gleevec, STI571, or CP57148B) is a prime example of inherent complexity in inferring binding specificity from sequence homology of kinase domains (fig. 1). This inhibitor is effective in treating early-stage chronic myelogenous leukemia acting as an effective direct inhibitor of ABL (ABL1), ARG (ABL2), KIT, and PDGFR tyrosine kinases [15-17]. The phylogenetic dendrograms based on primary sequence alignment of the protein tyrosine kinases assign the ABL kinase as more closely related to SRC tyrosine kinase family than to PDGFR family. However, Imatinib mesylate exhibits comparable efficiencies and high selectivity at inhibiting PDGFR, KIT, and ABL tyrosine kinases, while being largely ineffective in suppressing the tyrosine kinase activity of SRC family [15-17]. Phylogenetic analysis would not have predicted selective inhibition of PDGFR, KIT, and ABL tyrosine kinases by Imatinib mesylate without cross-referencing to a broader range of closely related tyrosine kinases. Protein kinases act as conformationally flexible signaling switches that have evolved elaborate mechanisms for maintaining a dynamic equilibrium between active and inactive states of the enzyme which plays a critical role in molecular recognition with the kinase inhibitors. The protein kinases crystal structures have revealed considerable structural differences between closely related active and highly specific inactive forms of the enzyme [18, 19]. A remarkable structural versatility of protein kinases in adopting distinct inactive conformations has also guided the development of specific kinase inhibitors as exemplified by the success of Imatinib in blocking the activity of the ABL kinase [15-17]. The high selectivity of Imatinib mesylate at inhibiting ABL, KIT, and PDGFR tyrosine kinases is arguably achieved by recogniz-

ing the specific inactive conformation of the ABL kinase [20-22]. However, the original concept that the inactive ABL conformation could present a structurally unique target has been recently questioned, as emerging proteomics approaches to kinase profiling [23] revealed somewhat unexpected broader spectrum of Imatinib activity, including binding to SYK kinase and some SRC tyrosine kinases, including LCK, FRK and FYN. A novel binding mode for Imatinib mesylate in the complex with the active SYK kinase has provided a plausible structural model for explaining Imatinib mesylate inhibition profiles with activated tyrosine kinases [24]. The conformational diversity of the ABL kinase crystal structures has been recently specified by classifying an array of alternative protein conformations into active, inactive, and intermediate states as well as the inactive-like conformations that closely resemble the inactive structure of SRC kinase [21-25]. This discovery has supported the initial notion that the structure of the ABL activation loop in the Imatinib complex is a natural protein conformation and the interconversion between distinctly different inactive conformations may be indeed a characteristic feature of the ABL kinase domain. Drug resistance toward Imatinib can frequently develop in advanced-stage of chronic myelogenous leukemia [26-28]. The major mechanism of resistance is the emergence of point mutations in the ABL kinase domain which interfere with drug binding. Structural studies have suggested that common mutations in the P-loop of ABL may cause resistance to Imatinib by impairing the flexibility of the kinase domain and restricting its ability to adopt the inactive conformation required for the productive inhibitor binding [26-28].

Pyrido-[2,3-d]pyrimidine drug molecules were originally synthesized as tyrosine kinase inhibitors of ATP-binding [29] and were identified as selective and potent inhibitors of SRC kinases [30,31]. Pyrido-[2,3-d]pyrimidine compounds are not only potent inhibitors of ABL kinase, but also can display a stronger activity against ABL compared to Imatinib. The most promising drug candidate in the class of pyrido [2,3-d]pyrimidines is PD166326 inhibitor (fig. 1), which is a highly potent dual specificity compound, active primarily against ABL, SRC and LCK kinases in a variety of *in vitro* binding assays with $IC_{50} = 1-5$ nM [32,33]. This compound has also demonstrated activity against FGFR, PDGFR and EGFR tyrosine kinases in a variety of *in vitro* assays with IC_{50} s in the range of 50–150 nM. Remarkably, pyrido-pyrimidine-type kinase inhibitors were found to be capable of suppressing some of the most frequently detected mutations of ABL in the P-loop that cause resistance towards Imatinib [34-36]. This class of inhibitors has exhibited a broad spectrum of activity and selectivity against protein kinases sharing a conserved Thr residue at the critical, gate-keeper position of the ATP binding site. Although the evolutionary profile of the Thr gate-keeper residue is conserved in all protein tyrosine kinases from the ABL and SRC families, the binding profile of Imatinib mesylate against ABL and SRC kinases is drastically different. The biochemical screening of existing clinical compounds and novel dual SRC and ABL kinase inhibitors has demonstrated feasibility of potent inhibition against Imatinib-resistant mutant variants of ABL. The activity of dual SRC-ABL kinase inhibitors was recently tested against a panel of 58 ABL kinase mutants revealing the varying pattern of PD166326 sensitivity and a compound-specific mutagenicity profile with previously undescribed mutants residues G250, E255,

T315, L248 and F317 [32-36]. Cataloging the identity of the pyrido-[2,3-d]pyrimidine resistant ABL mutants has revealed that mutations of the active site residues which are involved in the favorable intermolecular contacts (L248, F317, T315, Y253) can confer cross-resistance to the structurally distinct compounds, including Imatinib, Dasatinib, and PD166326. This study has emphasized potential pharmacological consequences of the binding promiscuity in drug resistance, suggesting that conformationally permissive binding may be impaired primarily via mutations of the kinase domain residues directly interacting with the inhibitor. Discovery of alternative BCR-ABL kinase inhibitors that have activity against Imatinib-resistant mutants is critically important for patients who relapse on Imatinib therapy.

The energy landscape view of molecular recognition [37-39], whereby a protein sequence can adopt diverse structures in binding with different partners, provides a model framework whereby sequence conservation and structural diversity can be interlinked in deciphering how proteins can rapidly evolve new binding functions. We employ ligand binding dynamics in sequence and structure space of protein tyrosine kinase conformations to unravel the imprint of evolutionary conservation and conformational selection on the binding function and specificity of tyrosine kinases. The employed model of biomolecular binding, based on the energy landscape theory implies presence of an ensemble of multiple conformational states for both interacting molecules. This model, often termed "conformational selection", assumes that conformational flexibility of the binding partners is determined by the equilibrium distribution of the protein and ligand conformational states, which is shifted upon binding towards the thermodynamically most stable complex [40,41]. It is worth noting that the analysis of the topography and connectivity in the space of protein tyrosine kinase sequences capable of high-affinity binding to diverse spectrum of kinase inhibitors may be useful in considering divergent evolution models of the protein binding function as the corresponding sequences are connected by a small number of mutations in the binding site residues and share the universally conserved amino acids in a few strategic hot spot positions which largely control the mode of recognition and specificity profile of tyrosine kinase family. In contrast, the presence of disconnected isles in sequence space may be an argument in favor of the convergent evolution for achieving binding function.

We propose *in silico* approach for a computational proteomics analysis of the tyrosine kinases binding in which evolutionary, structural and dynamical profiles of biomolecular interactions are probed to decipher the molecular basis of kinase selectivity. We apply this approach to elucidate the molecular basis of binding specificity and drug resistance for the clinically important class of inhibitors, including Imatinib and the pyrido-[2,3-d]pyrimidine inhibitors. We propose that structural origins of the pyrido-[2,3-d]pyrimidines activities against a specific spectrum of tyrosine kinases and Imatinib-resistant mutants may be largely determined by the underlying dynamics of the inhibitor-kinase interactions. This computational study allows to directly verify a long-standing hypothesis of whether the activity profile of multi-targeted kinase inhibitors may be related to variations in their sensitivity to the activation state of the enzyme and may result from a potent inhibition of both active and inactive tyrosine kinases crystal structures.

Examining dynamics and energetics of the inhibitor binding interactions on the kinome scale may have important implications in understanding the role of conformational adaptability of multi-targeted kinase inhibitors in developing mechanisms of drug resistance.

2. – Materials and methods

2.1. Structural classification of protein tyrosine kinases. – All publically available crystal structures of protein tyrosine kinases in the Protein Data Bank (PDB) [42] are used to categorize the structural space into panels of active and inactive kinase conformations. The following protein structures, employed in this study, provide the extensive coverage of tyrosine kinase family and include: 6 ABL (1fpu, 1iep, 1m52, 1opj, 1opk, 1opl); 1 ACK1 (1u4d); 3 KIT (1pkg, 1t45, 1t46); 1 ZAP70 (1u59); 1 BTK (1k2p); 1 ITK (1sm2); 1 CSK (1byg); 2 SYK (1xbb, 1xbc); 2 EGFR (1m14, 1m17); 4 IGF1R (1m7n, 1k3a, 1jqh, 1p4o); 2 EphA2 (1mqb, 1jpa); 1 FAK (1mp8); 4 FGFR1 (2fgi, 1fgk, 1fgi, 1agw); 2 HGFR (1r0p, 1r1w); 1 FLT3 (1rjb); 3 HCK (1ad5, 1qcf, 2hck); 4 INSR (1irk, 1ir3, 1i44, 1p14); 5 LCK (1qpc, 1qpd, 1qpe, 1qpj, 3lck); 1 MUSK (1luf); 4 SRC (1fmk, 1ksw, 2ptk, 2src); 1 TIE2 (1fvr); 1 KDR/VEGFR (1vr2). Among 51 protein tyrosine kinase crystal structures, the following 20 structures belong to the category of inactive kinase conformations: 3 ABL (1fpu, 1iep, 1opj); 2 KIT (1t45, 1t46); 1 BTK (1k2p); 1 CSK (1byg); 2 IGF1R (1m7n, 1k3a); 2 EphA2 (1mqb, 1jpa); 1 FAK (1mp8); 1 FGFR1 (1fgk); 1 HGFR (1r1w); 1 FLT3 (1rjb); 1 HCK (1ad5); 1 INSR (1irk); 1 MUSK (1luf); 1 SRC (1fmk); 1 TIE2 (1fvr).

The superposition of crystal structures from the protein tyrosine kinase family into a common reference frame is based on similarity of C_α atoms for a common set of residues defining the ATP binding site.

2.2. A hierarchical model of biomolecular recognition. – We employ a hierarchical model, which provides an opportunistic solution to achieve a synergy of an adequate conformational sampling and an accurate binding energetics. In this approach, equilibrium sampling of the inhibitor conformations with the multiple states of tyrosine kinases is first performed by replica-exchange Monte Carlo simulations using the soft-core AMBER force field. The molecular recognition energetic model used in dynamics simulations includes intramolecular energy terms, given by torsional and nonbonded contributions of the DREIDING force field [43], and the intermolecular energy contributions calculated using the AMBER force field [44] to describe protein-protein interactions combined with an implicit solvation model [45, 46]. We assume that the simplified soft-core force field follows the shape of the “true” potential and can detect the density of low-energy states in the broad regions surrounding favorable binding modes. However, the simplified soft-core energy model is less accurate in detecting the exact energetics of the inhibitor binding modes. We have applied a hierarchy of energy functions to utilize the robustness of the soft-core AMBER energy function in characterizing the multitude of the available low-energy binding modes and the topology of the binding energy landscape along with the accuracy of the MM/GBSA approach in quantifying the exact magnitude

of ligand-protein interactions. In the second stage of the protocol, the inhibitor binding free energies are computed for respective ligand-protein conformations generated in equilibrium simulations at $T = 300$ K, using the molecular mechanics AMBER force field and the solvation energy term based on continuum generalized Born and solvent accessible surface area solvation model [47,48]. This hierarchical approach allows to take into account equilibrium contributions of both active and inactive kinase structures to the computed binding free energies and thereby obtain a more accurate estimate of the inhibitor binding affinity with the panel of studied tyrosine kinases.

2'3. Monte Carlo binding simulations: simulated tempering dynamics. – We have carried out equilibrium simulations with the ensembles of protein kinase conformations using parallel simulated tempering dynamics [49,50] with 100 replicas of the ligand-protein system attributed respectively to 100 different temperature levels that are uniformly distributed in the range between 5300 K and 300 K. In simulations with ensembles of multiple protein conformations, protein conformations are linearly assigned to each temperature level, that implies a consecutive assignment of protein conformations starting from the highest temperature level and allows each protein conformation from the ensemble at least once be assigned to a certain temperature level. Monte Carlo moves are performed simultaneously and independently for each replica at the corresponding temperature level. After each simulation cycle, that is completed for all replicas, exchange of configurations for every pair of adjacent replicas at neighboring temperatures is introduced. The m -th and n -th replicas, described by a common Hamiltonian $H(X_1, \dots, X_m, \dots, X_n, \dots, X_N)$, are associated with the inverse temperatures β_m and β_n , and the corresponding conformations X_m and X_n . The exchange of conformations between adjacent replicas m and n is accepted or rejected according to Metropolis criterion with the probability

$$(1) \quad p = \min(1, \exp[-\delta]),$$

$$(2) \quad \delta = [\beta_n - \beta_m][H(X_m) - H(X_n)].$$

At equilibrium, the fraction of time that the ligand-protein system spends at a protein conformation $\lambda = i$ to time spent at a protein conformation $\lambda = j$ is determined by the Boltzmann distribution

$$(3) \quad \frac{P(\lambda_i = 1, \lambda_{m \neq i} = 0)}{P(\lambda_j = 1, \lambda_{l \neq j} = 0)}$$

and provides a measure for ordering protein conformations according to their interaction free energies with the inhibitor. The protein conformations that deliver the lowest interaction energy for the inhibitor during equilibrium simulation would dominate the distribution with the highest probability. By taking into account the inhomogeneity of the molecular system, it is possible to dynamically optimize the step sizes at each temperature during Monte Carlo simulations [51]. For the optimal performance of replica-exchange methods a number of free parameters such the appropriate temperature distribution,

range of temperatures, number of Monte Carlo sweeps at each temperature and number of swaps between different temperature levels after each cycle have to be chosen. We have adapted a protocol, where the highest temperature is defined as the “melting” temperature of the system. In this procedure, by starting at an arbitrary temperature, 100–500 Monte Carlo moves are attempted and the acceptance ratio is determined. If this parameter is less than 80%, the temperature is doubled. When the acceptance ratio exceeds this threshold, a sufficient number of moves to completely “melt” the system is applied and this temperature gets chosen as the highest temperature level. The efficient replica exchange simulation protocol typically requires a sufficient overlap between two adjacent energy probability distributions to ensure the convergence of the equilibrium ensemble. We have found that the considerable structural differences between multiple kinase conformational states, which “compete” for binding with the inhibitor required as many as 100 replicas of the system to guarantee a non-negligible probability of moving through the entire temperature range for all adjacent pairs of replicas. In this scenario, replica at low temperatures can escape from local minima which guarantee each replica of the system to be equilibrated in the canonical distribution with its own temperature. Replica-exchange Monte Carlo simulations of the inhibitor binding with protein tyrosine kinase crystal structures were repeated 100 times, each time allowing a different initial assignment of random starting ligand configurations and protein conformations to temperatures. We have found that the employed length of each simulation run (10000 samples at each temperature) and the use of soft-core potentials for sampling inhibitor conformations are sufficient attributes to ensure convergence in the equilibrium distribution of protein kinase conformations at each temperature. The inhibitor conformations and orientations are sampled in a parallelepiped that encompasses the crystallographic structures of all superimposed kinases complexes with a very large 10.0 Å cushion added to every side of this box surrounding the binding interface. This results in an unbiased set of random initial ligand conformations covering the entire kinase domain, without imposing any artificial constraints to prevent inhibitor from drifting away from the active site. The protein structure of each complex is held fixed in its minimized conformation, while rigid body degrees of freedom and the ligand rotatable angles are treated as independent variables. The dihedral angles allowed to rotate include those linking sp^3 hybridized atoms to either sp^3 or sp^2 hybridized atoms and single bonds linking two sp^2 hybridized atoms. The energy used to determine exchanges between temperature levels is the total energy, which includes not only the intermolecular energy of the ligand-protein complex, but also the internal energy of both the protein and the ligand. Since the protein structure of each complex is held fixed in its minimized conformation, the internal energy of each protein tyrosine kinase is recorded and fixed, but contributes to the total energy of each replica during the exchange of protein conformations.

3. – Results and discussion

We investigate how sequence conservation profiles and temperature-dependent selection for a specific protein and inhibitor conformational states are interlinked in enabling

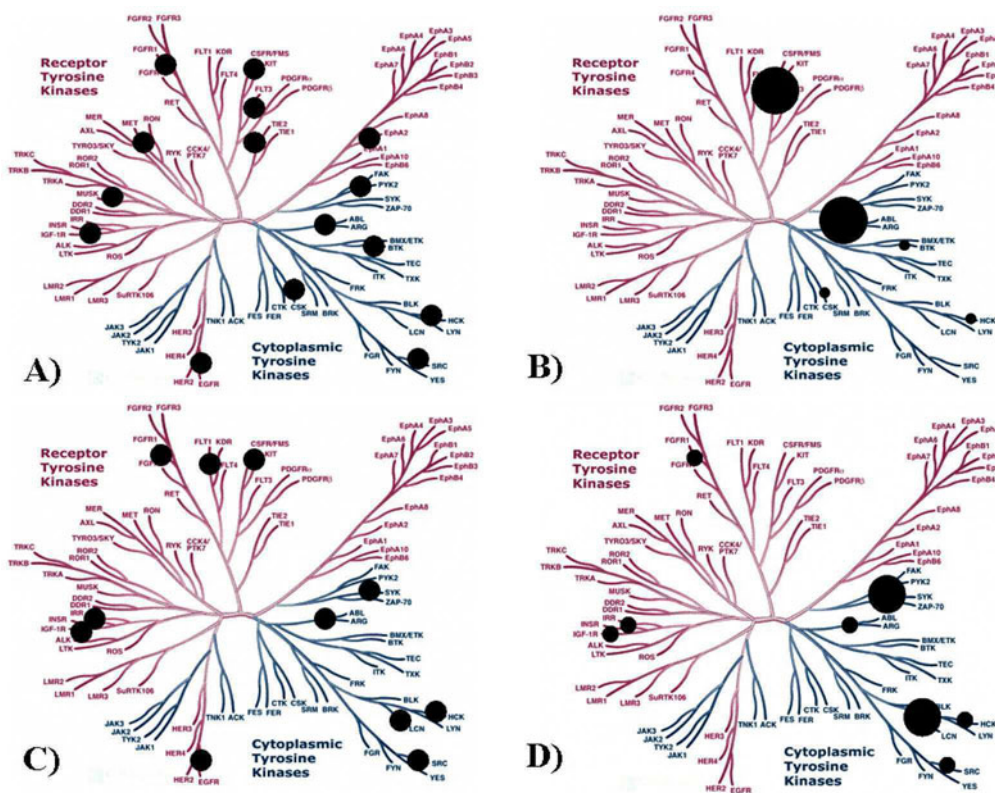


Fig. 2. – The kinase dendrogram was adapted from (Manning01) and is reproduced with the kind permission from Science (<http://www.sciencemag.org>) and Cell Signaling Technology, Inc. (<http://www.cellsignal.com>). The filled circles are used to map phylogenetic footprint of tyrosine kinases from the dendrogram with the structural space of inactive (A) and active (C) protein tyrosine kinase conformations. Graphical mapping of the phylogenetic footprint of the protein tyrosine kinases with the respective densities of protein tyrosine kinases in the inactive conformation (B) and active conformation (D) yielding the low-energy complexes with Imatinib mesylate. The size of the filled circles mapped onto phylogenetic dendrogram of protein tyrosine kinases is proportional to the frequency of recruiting the respective protein tyrosine kinase conformation to form a low-energy complex with the respective inhibitor in the course of simulations.

functional promiscuity and selectivity for three classes of clinically important classes of cancer drugs targeting protein tyrosine kinases. A kinome dendrogram representation of evolutionary relationships between protein tyrosine kinases is often employed to infer functional inhibitor profiles from sequence analysis. We have performed a simple mapping of the tyrosine kinases dendrogram with the structural space of available inactive (fig. 2A) and active tyrosine kinase conformations (fig. 2C). An extensive coverage of the tyrosine kinase phylogenetic footprint with the inactive kinases conformations would have suggested a much broader binding activity profile for Imatinib mesylate that has been actually determined. Furthermore, this analysis would not have predicted selective

inhibition of PDGFR, KIT, and ABL tyrosine kinases by Imatinib mesylate without suggesting cross-binding to a broader range of closely related tyrosine kinases. Hence, functional profile of Imatinib mesylate binding may not be directly linked with the position of protein kinases on evolutionary dendrogram. To provide a more robust functional mapping from sequence analysis to a structurally driven binding phenotype, we conduct *in silico* profiling of kinase inhibitors using Monte Carlo simulations of inhibitor binding with the ensemble of tyrosine kinase conformations that allows to implicitly incorporate protein flexibility and assess the role and contribution of the inhibitor-protein interactions on the kinase binding function and specificity. The results of simulations can be conveniently summarized by graphically mapping the phylogenetic footprint of tyrosine kinases with the respective densities of inactive (fig. 2B) and active (fig. 2D) protein tyrosine kinase structures which yield the low-energy complexes with Imatinib mesylate. In agreement with the experimental data, we have observed a high degree of structural similarity between the predicted inhibitor conformations and the crystal structure of Imatinib mesylate in the ABL kinase complex. The high selectivity towards the inactive conformations of ABL and KIT is achieved via convergence to a narrow spectrum of bound conformations featuring the unique binding mode of Imatinib mesylate in the crystal structure of the ABL kinase complex (fig. 3). The vast majority of Imatinib mesylate low-energy conformations reside within root mean square (RMSD) $\text{RMSD} = 2 \text{ \AA}$ from the crystallographic inhibitor conformation. There is an appreciable correlation between small deviations within Imatinib mesylate crystallographic binding mode and energetic variations. A smaller peak of predicted bound conformations lying within $\text{RMSD} = 0.5 \text{ \AA}$ from the crystal structure corresponds to the lowest binding free energies of -16 kcal/mol . Simulations with protein tyrosine kinases in the active form have revealed a significant structural departure of the low-energy inhibitor conformations from the Imatinib mesylate crystal structure featured in the ABL kinase complex (fig. 4). Furthermore, these bound conformations are distributed over a rather broad range of RMSD values, which may indicate a multitude of alternative binding conformations for Imatinib mesylate with the active tyrosine conformations. The distribution of binding free energies is shifted towards energy values with a peak nearing -6 kcal/mol , which are considerably weaker than the ones observed in simulations with the inactive tyrosine kinase structures (fig. 3). The dominant low-energy inhibitor binding mode in complexes with the active tyrosine kinases structures is similar to the recently determined crystal structure of Imatinib mesylate bound to the active conformation of SYK kinase (fig. 4). We have found that the active conformations of a very limited subset of active tyrosine kinases conformations, namely SYK and LCK structures are capable of accommodating fluctuations of Imatinib mesylate around a novel binding mode of Imatinib mesylate (figs. 3, 4). These data offer support to recent chemical proteomics discovery suggesting that in addition to inhibiting a small group of protein tyrosine inactive kinases structures with high efficiency, Imatinib mesylate can also bind with weaker affinity to LCK and SYK kinases [23, 24]. In the absence of an experimental structure of Imatinib mesylate bound to mono-phosphorylated ABL, the proposed binding mode may also provide a useful structural model for interpreting Imatinib mesylate inhibition data.

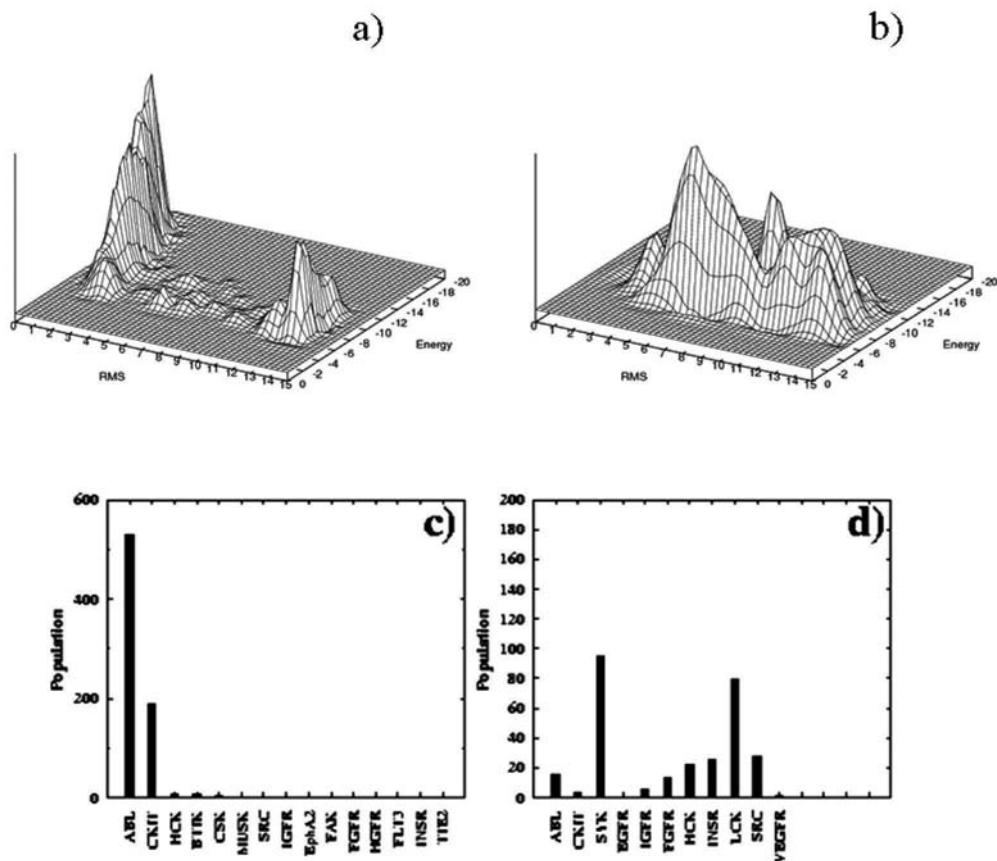


Fig. 3. – The distribution of protein tyrosine kinases for the predicted Imatinib mesylate complexes with the ensemble of INACTIVE (A) and ACTIVE (B) protein tyrosine kinase structures. Three-dimensional population histograms generated from equilibrium binding simulations of Imatinib mesylate inhibitor at $T = 300$ K as a function of root mean square (RMSD) values from the crystal structure conformation and the binding free energy of the respective inhibitor state in the complex with the INACTIVE (A) and ACTIVE (B) ABL kinase structures. Analysis of Imatinib mesylate binding energies is based on 1000 low-energy data points collected at $T = 300$ K.

The potent inhibition profile of most active pyrido-[2,3-d]pyrimidine PD166326 inhibitor (fig. 1) against a spectrum of tyrosine kinases, including ABL, SRC and LCK, has indicated that this inhibitor may bind to the ABL kinase domain independent of the activation loop conformation, existing in a dynamic equilibrium between the open and closed conformations [21,22]. The equilibrium distributions of the tyrosine kinase crystal structures interacting with the inhibitor PD166326 reveal a consistent convergence to the ABL and SRC families of tyrosine kinases (fig. 5). In accordance with *in vitro* data, the results of simulations reproduce a broad spectrum of experimental activities for the

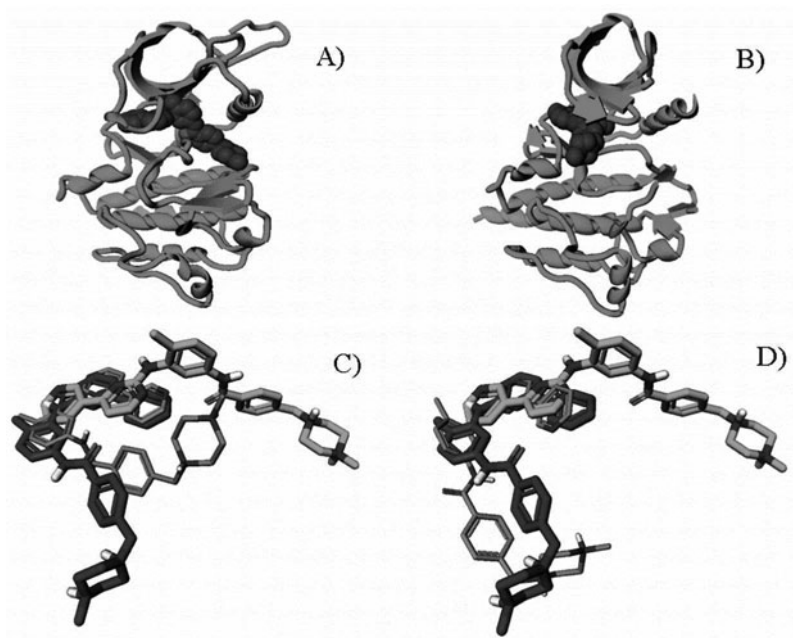


Fig. 4. – The crystal structure of Imatinib mesylate in the complex with the ABL inactive kinase structure (A) and SYK active kinase structure (B). The predicted conformation of Imatinib mesylate in the complex with the ABL kinase and the crystal structure conformation are virtually identical (C,D). The predicted conformation of Imatinib mesylate (displayed in light stick) in the complex with the LCK (C), and SYK (D) kinases superimposed with the crystal structure of Imatinib mesylate bound to the SYK active kinase conformation (displayed in dark stick). For illustration purpose, the crystal structure of Imatinib mesylate in the complex with the ABL kinase is also shown on panels (C) and (D).

pyrido-[2,3-d]pyrimidine inhibitors, including a potent inhibition of the ABL, SRC, HCK, LCK kinases and somewhat less favorable binding to the FGFR and PDGFR tyrosine kinases. The computational proteomics analysis agrees with the structural and biochemical data [21-25] and provides a dynamical view of the inhibitor-protein interactions. We have found that functional adaptability of the pyrido-[2,3-d]-pyrimidines inhibitor in binding with structurally diverse active and inactive forms of ABL may be achieved through moderate fluctuations around the crystallographic binding mode, without imposing stringent requirements for a specific protein conformation. Importantly, a strong preference in the binding signal towards ABL and SRC tyrosine kinases is exhibited irrespective of the activation state of the enzyme whereas the predicted bound conformation of PD166326 conforms to the crystallographic binding mode from the complex with the active ABL kinase (fig. 6). Indeed, the dynamics of the inhibitor-kinase binding interactions is relatively insensitive to the structure of the activation loop as both active and inactive ABL and SRC kinases can form the thermodynamically stable complexes with

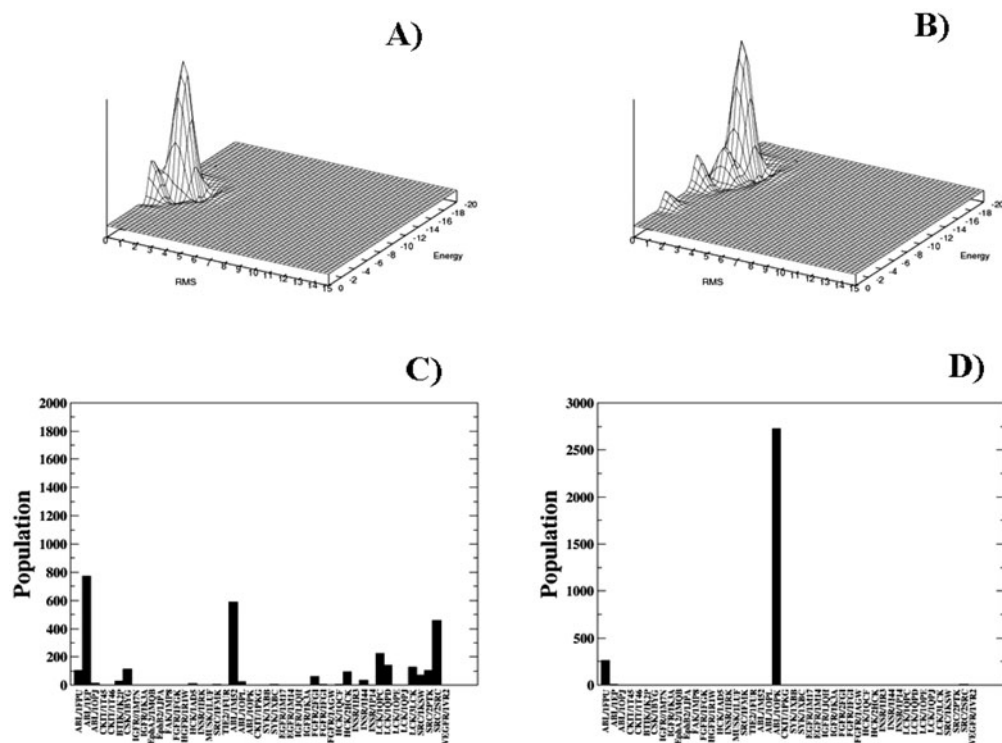


Fig. 5. – Three-dimensional population histograms generated from equilibrium binding simulations of PD166326 inhibitor at $T = 300$ K as a function of root mean square (RMSD) values from the lowest predicted bound conformation and the binding free energy of the respective inhibitor state in the complex with the INACTIVE (A) and ACTIVE (B) ABL kinase structures. Analysis of PD166326 binding energies is based on 1000 low-energy data points collected at $T = 300$ K. The distribution of protein tyrosine kinases states from the predicted PD166326 complexes in simulations at $T = 320$ K (C) and $T = 300$ K (D). X-axis labels denote the protein tyrosine kinase names and respective PDB codes. The PDB codes match the description of the crystal structures of protein tyrosine kinases in sect. 2 and point to the respective kinase conformations.

the inhibitors at lower temperature (fig. 5). However, the distribution of the tyrosine kinase crystal structures interacting with PD166326 at lower temperature is determined primarily by the active ABL kinase (fig. 5). The binding free-energy analysis of the PD1663236 bound conformations with the active and inactive ABL kinase structures manifests this salient structural feature of the inhibitor dynamics leading to a more favorable energetics with the ABL active state (fig. 5). This result indicates that although PD166326 can bind to both active and inactive ABL kinases, the respective variations in the P-loop conformations may be accommodated with different binding affinity.

It was initially proposed that the hydroxymethyl portion of PD166326 may form additional hydrogen bonds with the kinase domain, which may contribute to the increasing stability of interactions and partially explain the increased potency of this pyrido[2,3-

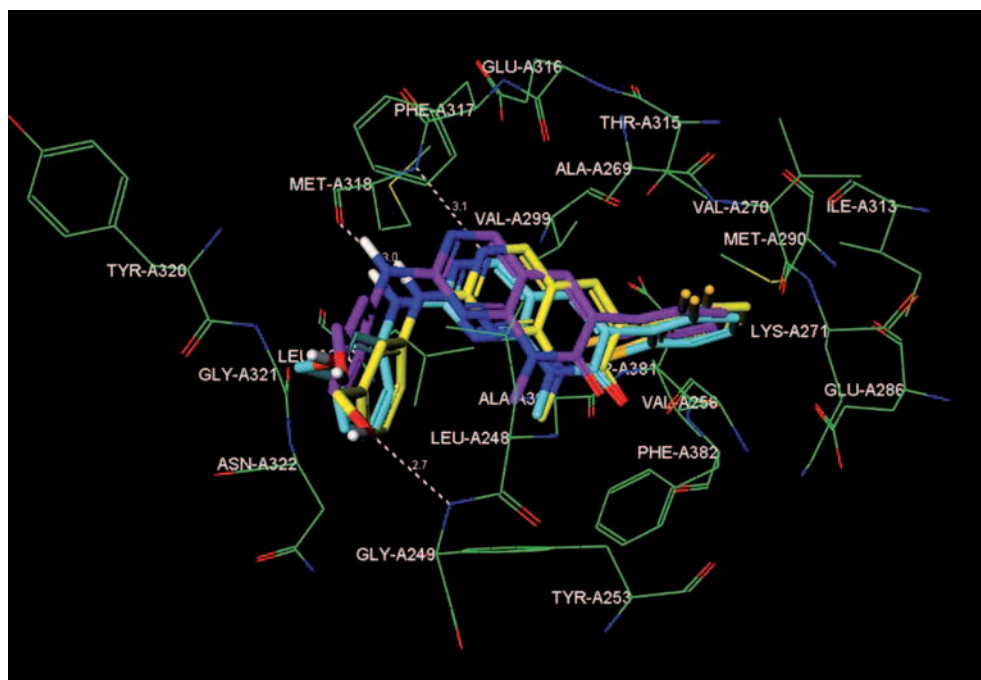


Fig. 6. – A view of the intermolecular network formed by the PD166326 inhibitor in the active site with the amide nitrogen and carboxyl oxygen of Met-318 is shown. The crystallographic conformation of PD166326 in the complex with the active ABL structure (pdb code 1OPK) (displayed with the stick model in light blue) is overlaid with the predicted binding mode of PD166326 bound to the active ABL structure (pdb code 1OPK) (with the stick model in yellow) and the predicted PD166326 conformation bound to the inactive ABL structure (pdb code 1IEP) (with the stick model in purple). Note that PD166326 can frequently form and break an additional hydrogen bond with the NH of Gly-249 from P-loop of the active ABL kinase.

d]pyrimidine derivative [21-25]. However, subsequent crystallography studies, which have determined the unique binding mode of PD166326 interacting with the active form of ABL kinase, have not indicated any specific interaction formed by the hydroxymethyl group of the inhibitor. We have found that during equilibrium simulations PD166326 can frequently form and break an additional hydrogen bond with the NH of Gly-249 from P-loop of the ABL kinase, capitalizing on the favorable interaction pattern provided by the conformation of the active conformation (fig. 6). Consequently, potential structural variations of the P-loop featured in the active and inactive ABL states may have an effect on the PD166326 binding energetics than on the inhibition profiles of more conformationally-tolerant pyrido[2,3-d]pyrimidines. The observed reduction in conformational tolerance of the ABL kinase and the P-loop region to interactions with PD166326 could impair the stability of the inhibitor-kinase interaction and partially explain the increased sensitivity of the PD166326 compound towards ABL mutations in the P-loop [32, 33].

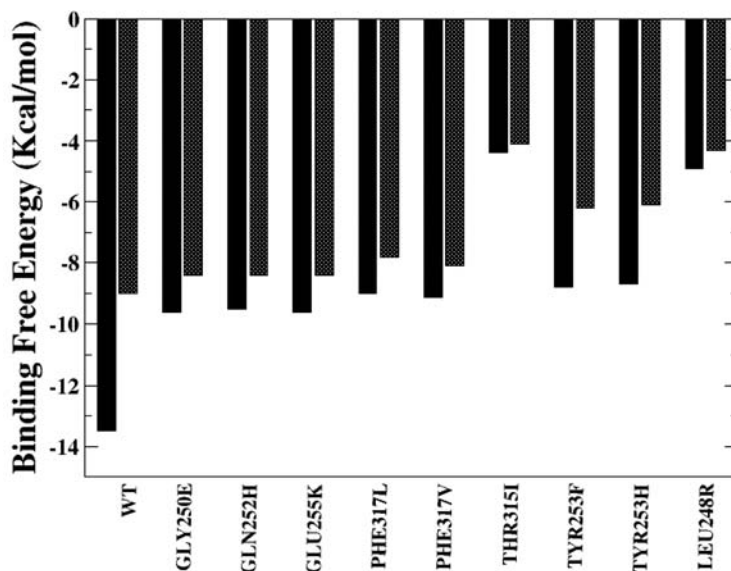


Fig. 7. – A comparison of the computed binding free energies for PD166326 with the wild-type ABL and a set of clinically relevant ABL kinase mutants: Gly-250E, Gln-252H, Glu-255K, Phe-317L, Phe-317V, Thr-315I, Tyr-253F, Tyr-253H, and Leu-248R. The binding free energy values shown in filled back bars are with the active wild-type ABL (pdb code 1M52) and the mutant forms of the enzyme assuming the active conformational state. The binding free energy values shown in filled dashed bars are with the inactive wild-type ABL (pdb code 1IEP) and the mutant forms of the enzyme assuming the inactive conformational state.

We have also elucidated the molecular basis of the PD166326 sensitivity against a number of clinically relevant ABL kinase mutants, including Gly-250E, Gln-252H, Glu-255K, Phe-317L, Phe-317V, Thr-315I, Tyr-253F, Tyr-253H, and Leu-248R (fig. 7). To understand the role of the ABL conformational states on the pyrido[2,3-d]pyrimidine patterns of sensitivity against clinically relevant ABL mutants, we have conducted binding free energy calculations in which the wild-type ABL and the studied mutant forms of the enzyme can assume either active or inactive conformational states. The computational procedure uses the wild-type ABL crystal structures as the templates to model the ABL mutant conformations with the penultimate rotamer library [52]. In this procedure, all rotamers for each mutated residue are generated and the best minimized rotamer replaces the wild-type residue. The energies of the modified proteins with multiple mutations are optimized using a simple self-consistent procedure, in which optimization of the rotamer position for a mutated residue is followed by the same procedure for the next mutated residue. This procedure is iterated in a self-consistent manner until convergence is achieved and the optimal rotamer positions of the mutated residues replace the corresponding wild-type residues.

Mutations in the mobile P-loop of the ABL kinase domain, including Gly-250E, Gln-252H, Glu-255K could arguably favor a shift in the equilibrium towards an active, Imatinib-resistant conformation and cause resistance to conformationally specific inhibitors. We have found that the differences in the binding free energies of PD166326 with the wild-type ABL and the P-loop mutants (Gly-250E, Gln-252H, Glu-255K) are quite moderate and are relatively insensitive to considerable structural variations featured in the active and inactive states of the ABL kinase (fig. 7). While the impact of Phe-317L, Phe-317V, Tyr-253F, and Tyr-253H mutations seems to be only marginally different, the computed free energies reveal a considerably greater effect of the conformational state of ABL on the resulting binding free-energy differences for the Tyr-253F and Tyr-253H mutants. These results are consistent with the outcome of a cell-based screening for resistance of ABL-positive leukemia, which identified a pattern of specific kinase mutations, including that cause resistance to PD166326 [32, 33]. In contrast, we have found that the identical substitutions of several active site residues which are directly involved in the favorable intermolecular contacts, including L248, F317, T315, Y253 can confer resistance to the pyrido-[2,3-d]pyrimidine inhibitors. Computational profiling of the inhibitor sensitivity with the ABL kinase mutants unambiguously detects Thr-315I and Leu-248R variants as causing a considerable reduction in the binding free energy with respect to the wild-type kinase and the respective decrease in the inhibitor sensitivity (fig. 7). This agrees with the cell-based screening experiments which uncovered a significant deleterious effect of these modifications on the inhibitory profiles of the pyrido-[2,3-d]pyrimidine compounds [32, 33]. The results of our study suggest that protein tyrosine kinases may have evolved an “escape mechanism of drug resistance” against conformationally adaptable inhibitors by generating a specific pattern of substitutions from the active site residues which are involved in the favorable intermolecular contacts and may directly impair the inhibitor-protein recognition.

4. – Conclusions

Understanding and predicting the molecular basis of protein kinases specificity against existing therapeutic agents remains highly challenging and deciphering this complexity presents an important problem in discovery and development of effective cancer drugs. Computational proteomics analysis of the ligand-protein interactions using parallel simulated tempering with an ensemble of the tyrosine kinases crystal structures agrees with the experimental activity profiles and reveals important molecular determinants of the kinase specificity. We have found that sequence plasticity of the binding site residues alone is not sufficient to enable protein tyrosine kinases to readily evolve binding activities with inhibitors. The proposed *in silico* proteomics analysis unravels mechanisms by which protein conformational diversity of tyrosine kinases is linked with sequence plasticity of the binding site residues and conformational ligand adaptability in enabling binding specificity of the tyrosine kinase complexes. Evolutionary conservation patterns in protein tyrosine kinases need to be considered in the context of structural similarity analysis of the inhibitor binding modes and classification of the respective protein

conformational changes which may lead to more robust metrics of functional annotation. Functionally important annotations of binding specificity can be attempted by combining multiple sequence alignments, evolutionary analysis of binding sites and *in silico* profiling of ligand-protein interactions on genomic scale. The presented computational approach may be useful in complementing proteomics technologies to characterize activity signatures of small molecules against a large number of potential kinase targets.

REFERENCES

- [1] HANKS S. K. and HUNTER T., *FASEB J.*, **9** (1995) 576.
- [2] KOSTICH M., ENGLISH J., MADISON V., GHEYAS F., WANG L., QIU P., GREENE J. and LAZ T.M., *Genome Biol.*, **3** (2002) research0043.1-0043.12.
- [3] KRUPA A. and SRINIVASAN N., *Genome Biol.*, **3** (2002) research0066.1-0066.14
- [4] MANNING G., WHYTE D. B., MARTINEZ R., HUNTER T. and SUDARSANAM S., *Science*, **298** (2002) 1912.
- [5] MANNING G., PLOWMAN G. D., HUNTER T. and SUDARSANAM S., *Trends Biochem. Sci.*, **10** (2002) 514.
- [6] HUBBARD S. R. and TILL J. H., *Annu. Rev. Biochem.*, **69** (2000) 373.
- [7] HUBBARD S. R., *Curr. Opin. Struct. Biol.*, **12** (2002) 735.
- [8] SRIDHAR R., HANSON-PAINTON O. and COOPER D. R., *Pharm. Res.*, **17** (2000) 1345.
- [9] MADHUSUDAN S. and GANESAN T. S., *Clin. Biochem.*, **37** (2004) 618.
- [10] SAWYERS C. L., *Genes and Dev.*, **17** (2003) 2998.
- [11] BAIN J., McLAUCHLAN H., ELLIOTT M. and COHEN P., *Biochem. J.*, **371** (2003) 199.
- [12] KNIGHT Z. A. and SHOKAT K. M., *Chem. Biol.*, **12** (2005) 621.
- [13] VULPETTI A. and BOSOTTI R., *Farmaco.*, **59** (2004) 759.
- [14] VIETH M., HIGGS R. E., ROBERTSON D. H., SHAPIRO M., GRAGG E. A. and HEMMERLE H., *Biochim. Biophys. Acta*, **1697** (2004) 243.
- [15] DRUKER B. J., *Adv. Cancer Res.*, **91** (2004) 1.
- [16] WONG S. and WITTE O. N., *Annu. Rev. Immunol.*, **22** (2004) 247.
- [17] DEININGER M., BUCHDUNGER E. and DRUKER B. J., *Blood*, **105** (2005) 2640.
- [18] HUSE M. and KURIYAN J., *Cell*, **109** (2002) 275.
- [19] NOBLE M. E., ENDICOTT J. A. and JOHNSON L. N., *Science*, **303** (2004) 1800.
- [20] SCHINDLER T., BORNMAN W., PELLICENA P., MILLER W. T., CLARKSON B. and KURIYAN J., *Science*, **289** (2000) 1938.
- [21] NAGAR B., BORNMAN W. G., PELLICENA P., SCHINDLER T., VEACH D. R., MILLER W. T., CLARKSON B. and KURIYAN J., *Cancer Res.*, **62** (2002) 4236.
- [22] NAGAR B., HANTSCH O., YOUNG M. A., SCHEFFZEK K., VEACH D., BORNMAN W., CLARKSON B., SUPERTI-FURGA G. and KURIYAN J., *Cell*, **112** (2003) 859.
- [23] FABIAN M. A., BIGGS W. H. 3RD, TREIBER D. K., ATTERIDGE C. E., AZIMIOARA M. D., BENEDETTI M. G., CARTER T. A., CICERI P., EDEEN P. T., FLOYD M., FORD J. M., GALVIN M., GERLACH J. L., GROTZFELD R. M., HERRGARD S., INSKO D. E., INSKO M. A., LAI A. G., LELIAS J. M., MEHTA S. A., MILANOV Z. V., VELASCO A. M., WODICKA L. M., PATEL H. K., ZARRINKAR P. P. and LOCKHART D. J., *Nat. Biotechnol.*, **23** (2005) 329.
- [24] ATWELL S., ADAMS J. M., BADGER J., BUCHANAN M. D., FEIL I. K., FRONING K. J., GAO X., HENDLE J., KEEGAN K., LEON B. C., MULLER-DIECKMANN H. J., NIENABER V. L., NOLAND B. W., POST K., RAJASHANKAR K. R., RAMOS A., RUSSELL M., BURLEY S. K. and BUCHANAN S. G., *J. Biol. Chem.*, **279** (2004) 55827.

- [25] LEVINSON N. M., KUCHMENT O., SHEN K., YOUNG M. A., KOLDOBSKIY M., KARPLUS M., COLE P. A. and KURIYAN J., *PLoS Biol.*, **4** (2006) 0753.
- [26] CORBIN A. S., LA ROSEE P., STOFFREGEN E. P., DRUKER B. J. and DEININGER M. W., *Blood*, **101** (2003) 4611.
- [27] MICHOR F., HUGHES T. P., IWASA Y., BRANFORD S., SHAH N. P., SAWYERS C. L. and NOWAK M. A., *Nature*, **435** (2005) 1267.
- [28] COWAN-JACOB S. W., GUEZ V., FENDRICH G., GRIFFIN J. D., FABBRO D., FURET P., LIEBETANZ J., MESTAN J. and MANLEY P. W., *Mini Rev. Med. Chem.*, **4** (2004) 285.
- [29] BOSCHELLI D. H., WU Z., KLUTCHKO S. R., SHOWALTER H. D., HAMBY J. M., LU G. H., MAJOR T. C., DAHRING T. K., BATLEY B., PANEK R. L., KEISER J., HARTL B. G., KRAKER A. J., KLOHS W. D., ROBERTS B. J., PATMORE S., ELLIOTT W. L., STEINKAMPF R., BRADFORD L. A., HALLAK H. and DOHERTY A. M., *J. Med. Chem.*, **41** (1998) 4365.
- [30] KRAKER A. J., HARTL B. G., AMAR A. M., BARVIAN M. R., SHOWALTER H. D. and MOORE C. W., *Biochem. Pharmacol.*, **60** (2000) 885.
- [31] MOASSER M. M., SRETHAPAKDI M., SACHAR K. S., KRAKER A. J. and ROSEN N., *Cancer Res.*, **59** (1999) 6145.
- [32] VON BUBNOFF N., VEACH D. R., VAN DER KUIP H., AULITZKY W. E., SANGER J., SEIPEL P., BORNMAN W. G., PESCHEL C., CLARKSON B. and DUYSER J., *Blood*, **105** (2005) 1652.
- [33] WOLFF N. C., VEACH D. R., TONG W. P., BORNMAN W. G., CLARKSON B. and ILARIA R. L. JR., *Blood*, **105** (2005) 3995.
- [34] DORSEY J. F., JOVE R., KRAKER A. J. and WU J., *Cancer Res.*, **60** (2000) 3127.
- [35] WISNIEWSKI D., LAMBEK C. L., LIU C., STRIFE A., VEACH D. R., NAGAR B., YOUNG M. A., SCHINDLER T., BORNMAN W. G., BERTINO J. R., KURIYAN J. and CLARKSON B., *Cancer Res.*, **62** (2002) 4244.
- [36] VON BUBNOFF N., VEACH D. R., MILLER W. T., LI W., SANGER J., PESCHEL C., BORNMAN W. G., CLARKSON B. and DUYSER J., *Cancer Res.*, **63** (2003) 6395.
- [37] SHOEMAKER B. A., PORTMAN J. J. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **97** (2000) 8868.
- [38] LEVY Y., WOLYNES P. G. and ONUCHIC J. N., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 511.
- [39] VERKHIVKER G. M., BOUZIDA D., GEHLHAAR D. K., REJTO P. A., FREER S. T. and ROSE P. W., *Curr. Opin. Struct. Biol.*, **12** (2002) 197.
- [40] KUMAR S., MA B., TSAI C. J., SINHA N. and NUSSINOV R., *Protein Sci.*, **9** (2000) 10.
- [41] TSAI C. J., MA B. and NUSSINOV R., *Proc. Natl. Acad. Sci. U.S.A.*, **96** (1999) 9970.
- [42] BERMAN H. M., WESTBROOK J., FENG Z., GILLILAND G., BHAT T. N., WEISSIG H., SHINDYALOV I. N. and BOURNE P. E., *Nucleic Acids Res.*, **28** (2000) 235.
- [43] MAYO S. L., OLAFSON B. D. and GODDARD W. A. III., *J. Phys. Chem.*, **94** (1990) 8897.
- [44] CORNELL W. D., CIEPLAK P., BAYLY C. L., GOULD I. R., MERZ K. M., FERGUSON D. M., SPELLMEYER D. C., FOX T., CALDWELL J. W. and KOLLMAN P. A., *J. Am. Chem. Soc.*, **117** (1995) 5179.
- [45] STOUTEN P. F. W., FRÖMMEL C., NAKAMURA H. and SANDER C., *Mol. Simul.*, **10** (1993) 97.
- [46] BEUTLER T. C., MARK. A. E., VAN SCHAIK R. C., GERBER P. R. and VAN GUNSTEREN W., *Chem. Phys. Lett.*, **222** (1994) 529.
- [47] MOHAMADI F., RICHARDS N. G. J., GUIDA W. C., LISKAMP R., LIPTON M., CAUFIELD C., CHANG G., HENDRICKSON T. and STILL W. C., *J. Comput. Chem.*, **11** (1990) 440.

- [48] KOLLMAN P. A., MASSOVA I., REYES C., KUHN B., HUO S., CHONG L., LEE M., LEE T., DUAN Y., WANG W., DONINI O., CIEPLAK P., SRINIVASAN J., CASE D. A. and CHEATHAM T. E. III., *Acc. Chem. Res.*, **33** (2000) 889.
- [49] SUGITA Y. and OKAMOTO Y., *Chem. Phys. Lett.*, **314** (1999) 141.
- [50] VERKHIVKER G. M., REJTO P. A., BOUZIDA D., ARTHURS S., COLSON A. B., FREER S. T., GEHLHAAR D. K., LARSON V., LUTY B. A., MARRONE T. and ROSE P. W., *Chem. Phys. Lett.*, **336** (2001) 495.
- [51] BOUZIDA D., KUMAR S. and SWENDSEN R. H., *Phys. Rev. A*, **45** (1992) 8894.
- [52] LOVELL S. C., WORD J. M., RICHARDSON J. S. and RICHARDSON D. C., *Proteins*, **40** (2000) 389.

Blocking the protein folding machinery. Rational design of inhibitors of the molecular chaperone Hsp90 as new anticancer agents

G. COLOMBO(*)

*Istituto di Chimica del Riconoscimento Molecolare del CNR
Via Mario Bianco 9, 20131 Milano, Italy*

Cancer therapy now aims at disabling oncogenic pathways that are selectively operative in tumor cells, so to spare normal tissues and limit side effects in humans. This “targeted therapy” relies on a better understanding of cancer genes, particularly those implicated in tumor cell proliferation and survival [1]. Accordingly, targeted inhibition of the Bcr-Abl kinase with small molecule antagonists has produced dramatic clinical responses in malignancies driven by this oncogene [2]. However, such approach may not be immediately available for the majority of tumors where multiple molecular abnormalities and genetic instabilities may elude the identification of one single, disease-driving oncogene [1]. Conversely, pathways that intersect multiple essential functions of tumor cells may provide wider therapeutic opportunities. A prime target for this strategy is Heat Shock Protein 90 (Hsp90), a molecular chaperone that oversees the correct conformational development of polypeptides and protein refolding through sequential ATPase cycles and stepwise recruitment of cochaperones. This adaptive pathway contributes to the cellular stress response to environmental threats, including heat, heavy metal poisoning, hypoxia, etc., and is dramatically exploited in cancer, where Hsp90 ATPase activity is upregulated by ~ 100 -fold [3]. The repertoire of Hsp90 client proteins is restricted mainly to growth-regulatory and signaling molecules, especially kinases and transcription factors, which may contribute to tumor cell maintenance [3]. Therefore, targeted suppression of Hsp90

(*) E-mail: g.colombo@icrm.cnr.it

ATPase activity with a small molecule inhibitor, the benzoquinone ansamycin antibiotic 17-allylamino-17-demethoxygeldanamycin (17-AAG) showed promising anticancer activity in preclinical models, and recently completed safety evaluation in humans [4]. One Hsp90 client protein with critical roles in tumor cell proliferation and cell viability is survivin, an Inhibitor of Apoptosis (IAP) protein selectively over-expressed in cancer [5, 6]. Accordingly, targeting the survivin-Hsp90 complex may provide a strategy to simultaneously disable multiple signaling pathways in tumors, and a peptidomimetic antagonist of this interaction structurally different from 17-AAG, Shepherdin, inhibited the chaperone activity and exhibited potent and selective anticancer activity in preclinical models [7].

In this paper, we report the computational/theoretical structure-based design and characterization of *shepherdin*, a novel peptidomimetic antagonist of the complex between Hsp90 and survivin [7]. For its potent and broad antitumor activity, selectivity of action in tumor cells *vs.* normal tissues, and inhibition of tumor growth *in vivo* without toxicity, shepherdin (K79-L87, KHSSGCAFL) and its retroinverso analog shepherdin-RV may offer a promising approach for rational cancer therapy. We will then show how we could identify the minimum sequence of shepherdin, labeled shepherdin-min (K79-G83, KHSSG) required for activity in acute leukemia cancer cells. The structures of these peptides are studied by means of long time-scale MD simulations in explicit water. Subsequently, the dominant structures are docked to Hsp90, and the resulting complexes are also relaxed by means of long time-scale MD simulations to identify at equilibrium the dominant interactions responsible for binding. Finally, we will describe the use of the information developed in this part to identify a new non-peptidic small molecule which represents the prototype for a new class of compounds which can selectively inhibit Hsp90's chaperone activity. Computational and theoretical results will be benchmarked by experimental validations *in vitro* and *in vivo*.

1. – Materials and methods

1.1. Peptide molecular-dynamics simulations. – The peptides that have been simulated comprised the survivin sequence K79-K90 (KHSSGCAFLIVK) [5, 6]. The survivin sequence K79-L87 (KHSSGCAFL, Shepherdin) and Shepherdin-RV (LFACGSSHK, all D-aminoacids), and the minimal peptide K79-G83 (KHSSG). The C- and N-termini of each peptide were capped to avoid electrostatic artifacts due to the attractions between free opposite charges at the C- and N-termini. The side chain of Lys (K79 in Shepherdin and shepherdin-min and K87 in Shepherdin-RV) was considered to be protonated, bearing a net charge of +1. The charge state of each peptide molecule is consistent with the solution conditions of the experiments. Each peptide was solvated with water in a periodic truncated octahedron, large enough to contain the peptide and 0.9 nm of solvent on all sides. All solvent molecules within 0.15 nm of any peptide atom were removed. Two Cl^- counterions were added to the K79-K90 system, while one Cl^- counterion was added to Shepherdin, Shepherdin-RV and Shepherdin-min to ensure electroneutrality of the solution. Different initial conditions and conformations were used for the different peptide systems studied. For peptide K79-K90, the starting structure was totally extended to

avoid biases in the conformational search. The simulation (production run) was of 100 ns. Shepherdin was subjected to two 100 ns long Molecular Dynamics (MD) simulations (for a total of 200 ns), starting from either the totally extended conformation (all backbone dihedrals set to 180 degrees) or the conformation the peptide has in the survivin crystal structure [8]. Shepherdin-RV, being the most active agent *in vivo*, was simulated for longer time-scales by combining five different simulations starting with different initial velocities on the atoms of the peptide and the solvent, obtained from a Maxwellian velocity distribution at the desired temperature of 300 K. The initial conformation is in all cases completely extended. Four simulations were 100 ns long, while one was 53 ns long for a total time of 453 ns. Shepherdin-min was simulated for 200 nanoseconds.

Each system was initially energy minimized with a steepest descent method for 1000 steps. In all simulations the temperature was maintained close to the intended value of 300 K by weak coupling to an external temperature bath [9] with a coupling constant of 0.1 ps. The peptide and the rest of the system were coupled separately to the temperature bath. The *GROMOS96* force field [10] was used. The simple point charge (SPC) [11] water model was used. The LINCS algorithm [12] was used to constrain all bond lengths. For the water molecules the SETTLE algorithm [13] was used. A dielectric permittivity $\epsilon = 1$ and a time step of 2 fs were used. A cut-off was used for the calculation of the non-bonded van der Waals interactions. The cut-off radius was set to 0.9 nm. The calculation of electrostatic forces utilized the PME implementation of the Ewald summation method. All atoms were given an initial velocity obtained from a Maxwellian distribution at the desired initial temperature of 300 K. In each simulation, the density of the system was adjusted performing the first equilibration runs at NPT condition by weak coupling to a bath of constant pressure ($P_0 = 1$ bar, coupling time $\tau_P = 0.5$ ps) [9]. All simulations, starting from the appropriate peptide geometry, were equilibrated by 50 ps of MD runs with position restraints on the peptide to allow relaxation of the solvent molecules. These first equilibration runs were followed by other 50 ps runs without position restraints on the peptide. The production runs using NVT conditions, after equilibration, were 50 ns long for all of the complexes. All the MD runs and the analysis of the trajectories were performed using the GROMACS software package [14]. Conformational cluster analysis of the combined trajectories for Shepherdin (200 ns) or Shepherdin-RV (453 ns) and Shepherdin-min was performed using the method described in Daura *et al.* [15]: count the number of neighbors using a cut-off of 0.15 nm Root Mean Square Deviation (RMSD) between the optimal backbone superposition of different structures, take the structure with the largest number of neighbors with all its neighbors as a cluster and eliminate it from the pool of clusters. This procedure is repeated for the remaining structures in the pool. The most populated clusters, corresponding to the most visited structures in the MD simulations, for Shepherdin and Shepherdin-RV comprised hairpin type of structures (representative structures are reported in fig. 1). The most populated structural cluster for Shepherdin[79-83] was in contrast characterized by an extended structure. The representative structures (dominant structures of the peptide in solution) of the most populated cluster were used for docking experiments on Hsp90.

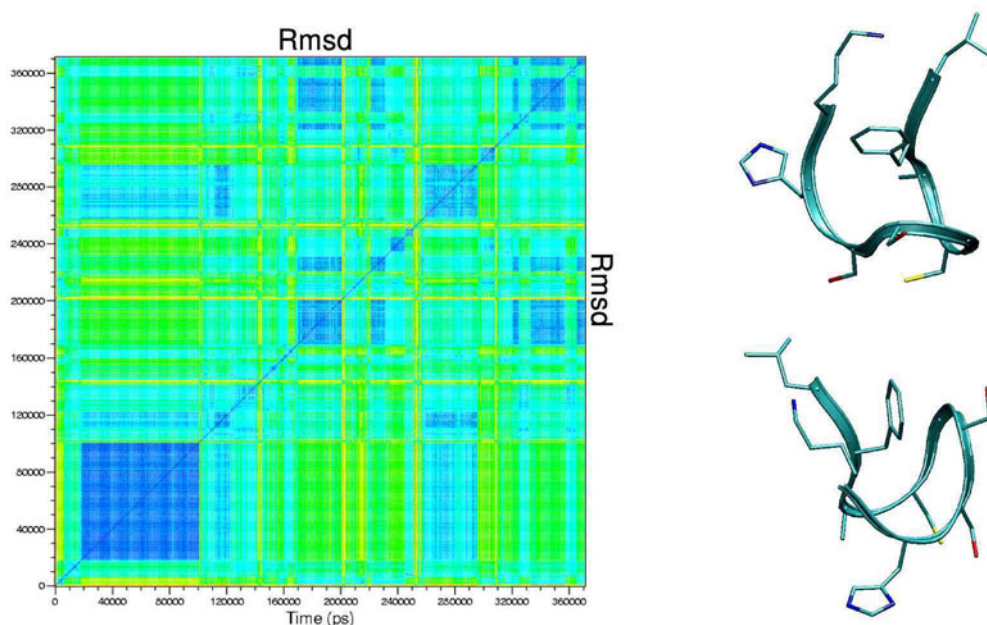


Fig. 1. – Main structures from extensive MD simulations of Shepherdin.

1'2. Docking procedure. – The β -hairpin structures of Shepherdin and Shepherdin-RV, and the extended structure of Shepherdin-min were subjected to blind docking experiments on the N-terminal domain of Hsp90 using the program Autodock [16]. The crystal structure of the protein was taken from the protein data bank (pdb code 1YET). The original X-ray structure contains the ligand Geldanamycin (GA), which was removed from the active site to yield the apo-open form of Hsp90. To test the viability of the docking procedure, and its ability to reproduce the experimental structure, a blind docking procedure was initially run on GA. The ligand was removed from Hsp90 and a docking experiment was run with no information on the position of the binding site. The minimum free energy structure determined in this experiment is exactly superimposable to the X-ray derived structure described [17].

The docking procedure used can be summarized as follows. Mass-centered grid maps were generated with 0.25 Å spacing by the program Autogrid for the whole Hsp90 protein target. Lennard-Jones parameters 12-10 and 12-6 (default parameters in the program package) were used for modeling hydrogen bonding and Van der Waals interactions, respectively. The distance-dependent dielectric permittivity of Mehler and Solmajer [18] was used for the calculation of the electrostatic grid maps. The Lamarckian genetic algorithm (LGA) and the pseudo-Solis and West methods were applied for minimization using default parameters. The number of generations was set to 25 million in all runs, and the stopping criterion was therefore defined by the total number of energy evalu-

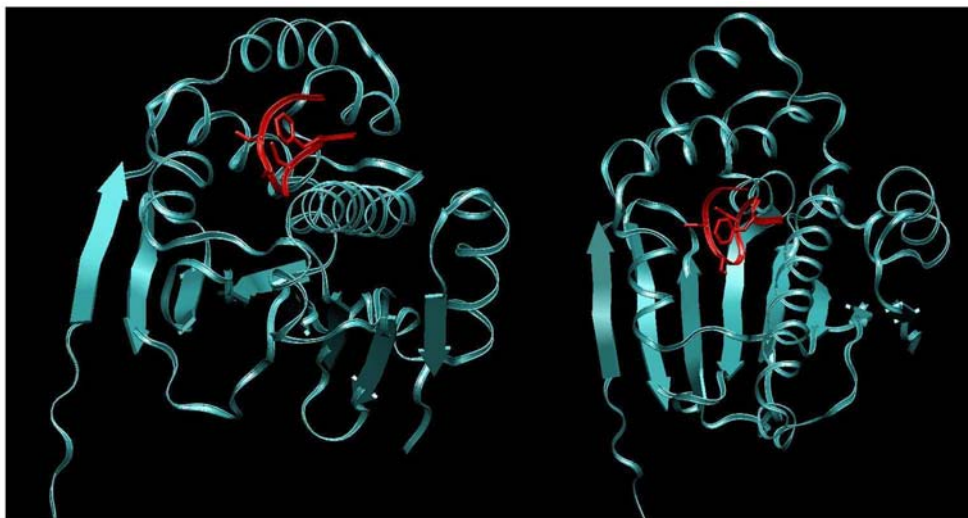


Fig. 2. – Structures of the docked complex between Hsp90 and Shepherdin.

ations. Random starting positions on the entire protein surface, random orientations, and torsions (flexible ligand only) were used for the ligand. For Shepherdin 100 different runs were performed with the parameters described above. For Shepherdin-RV, a total of 350 runs were performed. The results of docking runs were analyzed by the clustering procedure described by Hetenyi *et al.* [19] were classified by a two-step procedure. First, the docked conformations of the ligand peptides were listed in increasing energy order. The structure of the complex that corresponded to the global minimum energy was used as the starting point of the first refinement molecular-dynamics run. Second, the ligand conformation with the lowest energy was used as a reference, and all conformations with a center-of-mass-to-center-of-mass distance of less than 3 Å from the reference were taken to belong to the first class. After a ligand was assigned to a class, it was not used again for other classes. The process was then repeated for all hitherto unclassified conformations until all conformations were put in a class. The energy minima are represented in fig. 2. The representative structure of the most populated class was then used for the second molecular dynamics refinement run. The refinement molecular dynamics runs of the complexes obtained after the Autodock runs were each 70 nanoseconds long.

1'3. Pharmacophore generation. – Three different pharmacophore models were built and labeled PHARM1, PHARM2 and PHARM3 based on the results of MD simulations. The conformation of shepherdin and the orientations of its side-chain functional groups in the most populated structural cluster from MD trajectories of the complex were used as structural template (fig. 3). The distributions of dihedral values and distances among critical functionalities were used to define upper and lower boundaries for geometric constraints. The HypoGen Module of the Catalyst[®] program from Accelrys was used for this

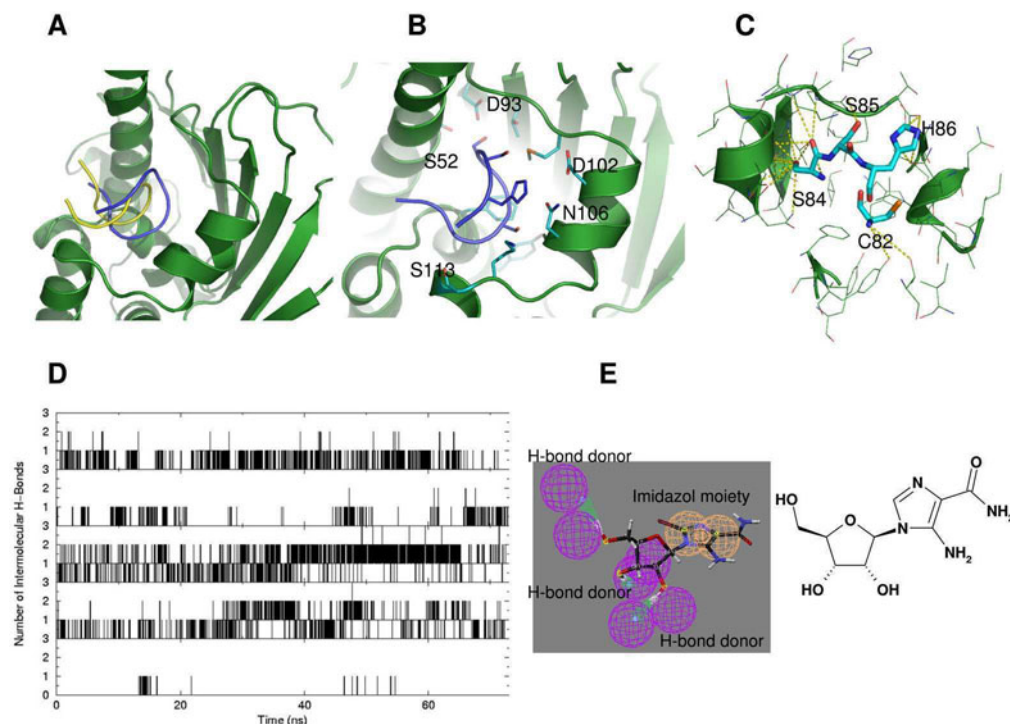


Fig. 3. – A) Structure of the MD-refined docked complex between shepherdin and the N-domain of Hsp90 (green). Initial (yellow) and final (blue) structure of shepherdin within the ATP-binding pocket of Hsp90 with two different orientations. B) Side-chain groups of Hsp90 responsible for the main interactions with shepherdin. C) Close-up of the groups of shepherdin (colored by atom) responsible for stabilizing interactions with Hsp90 (green). D) Time evolution of the H-bonding interactions between Hsp90 and shepherdin groups (from top to bottom) γ -OH of S84, γ -OH of S85, $N\epsilon$ of H86, $N\delta$ -H of H86, and SH of C82. E) The pharmacophoric hypothesis mapped on the 3D structure of AICAR and its molecular structure (5-aminoimidazole-4-carboxamide-1- β -D-ribofuranoside, AICAR).

purpose. For PHARM1 a four-point model was created by assembling 3 H-bond donors mapped over the two Ser γ -OH and the Cys SH groups, and one imidazole moiety mapped over the imidazole ring of His of shepherdin. The torsional and distance restraints were added to restrict the database search. Furthermore, we imposed location constraints (the volume in which the functions can reside), specifying the radius of the spheres according to Catalyst's defaults. PHARM2 and 3 were constructed in a similar way, augmenting PHARM1 with an aromatic function centered on the position of the Phe80 benzene ring and a hydrophobic function centered on the S atom of Cys82 (PHARM2), or by the presence of a positive charge mapped on the position of the ammonium group of Lys87 (PHARM3). These were used as queries for a search in the NCI3D database using the database search module of Catalyst.

2. – Results

2'1. *Shepherdin and Shepherdin RV.* – The peptides of the Shepherdin series were identified as described in the last paragraph of the introduction to this paper.

The modelling study of the peptide K79-L87 named shepherdin and its retroinverso version L87-K79 (all D aminoacid) shepherdin-RV start with a long molecular-dynamic simulation with the target to identify the characteristic conformations of these peptides in solution. Analysis of the trajectories predicts that both shepherdin and shepherdin-RV have a dominant configuration characterized by a turn involving S82-G83 in shepherdin and G83-S84 in shepherdin-RV, and overall β -hairpin geometry (fig. 1). The most populated conformation of shepherdin-RV shows a higher degree of compactness, with the aromatic ring of F80 packing on the turn region (fig. 1). The representative β -hairpin conformations of both peptides were subjected to multiple docking experiments on Hsp90 using the AutoDock program package. In all cases, the peptides were predicted to dock into the ATP binding site of Hsp90 (figs. 2). The geometry of the final complex is highly correlated with that of the complex between Hsp90 and GA [17], with the turn region of the peptides closely tracing the ansa ring backbone of GA. Shepherdin and Shepherdin-RV make 13 and 18 predicted hydrogen bonds with the ATP pocket of Hsp90, respectively, involving the side chains of H80, S81, S82, the carbonyl group of G83, and the side chains of K87 and C82 (shepherdin-RV). Except for D93, the complementary residues of Hsp90 predicted to make contact with shepherdin and/or shepherdin-RV largely overlap with amino acids implicated in GA binding [17], including S113, which has been recently shown to contribute to stepwise accessibility of the ATP pocket of Hsp90 to GA. Shepherdin and shepherdin-RV are predicted to assume more extended conformations than GA in the Hsp90 pocket, and bury a solvent accessible surface of 498 and 546 Å², respectively, as opposed to 402 Å² buried by GA.

To check these structural predictions, and validate experimentally that shepherdin engaged Hsp90 differently from GA, we introduced targeted mutations in the ATP pocket of Hsp90, and tested their effect on shepherdin binding. Individual substitution of N51, S52, D102, or S113 in the N domain of Hsp90 reduced binding to shepherdin by 20–60%, whereas mutagenesis of “GA-specific” D93 had no effect, and a scrambled peptide did not bind wild-type or mutant Hsp90 (fig. 4).

2'2. *Simulations of Shepherdin-RV mutants.* – To investigate the impact of single-point mutation on the structure-activity relationship properties of Shepherdin-RV, and to shed more light on the determinants of the interaction between the peptide and Hsp90, two mutants peptides (H80A and C84A) were simulated with long time scale all-atom MD simulations. A total of four simulations (two runs for each mutant) are calculated for the two peptide mutants. Two different initial conformations were used: one completely extended and the second one from the dominant Shepherdin-RV conformation found in the previous runs, and subjected to mutation. The purpose of the first simulation is to identify the characteristic conformations of these peptides in solution and the stability of the Shepherdin-RV β -hairpin conformation after the mutation.

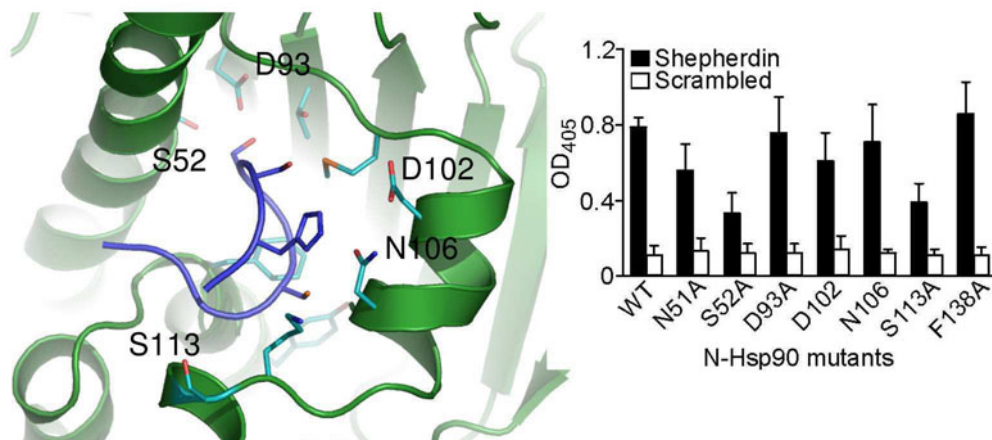


Fig. 4. – Predicted important Hsp90-Shepherdin Interactions (left) and their experimental verification (right) through mutagenesis experiments.

In the case of the C84A mutant, simulations suggest that the mutation dramatically decreases the tendency of the peptide to form a stable hairpin like structure. In the 100 ns time span from the completely extended conformation neither the analysis of the time evolution of secondary structure, nor the structural cluster analysis are able to identify a hairpin conformation similar to that observed for the original sequence. The second simulation, from a preformed hairpin structure, shows that the mutant peptide retain the hairpin conformation for about 10 ns and after that the turn geometry changes for a long period, before complete loss of the initial conformation.

An analogous behaviour is observed for H80A mutant peptide. Both the analysis of the time evolution of secondary structure and the structural clustering suggest that the hairpin is not the dominant conformation in solution, despite being present for a smaller percentage.

These results clearly suggest that the bent, hairpin like conformation of Shepherdin is a fundamental determinant for recognition with the active site of Hsp90. It is worth noting at this point, that this type of conformation for the same sequence is also present in the native structure of Shepherdin, suggesting a certain level of structural preorganization for this sequence, optimized for binding to Hsp90.

2'3. *Shepherdin*[79-83]. – The combination of theoretical analysis and experimental verification described in the previous subsections suggest that the minimal motif necessary for recognition should contain the HSSG sequence. Based on these considerations, a new short peptide, with sequence KHSSG spanning residues 79-83 of survivin was synthesized and named Shepherdin[79-83].

To understand possible structure-activity relationships of this peptide and of its interaction with Hsp90, Shepherdin[79-83] was simulated in isolation in explicit water. The peptide did not populate any preferred ordered secondary structure, and so the hydrophilic side chains and the backbone carboxyl and amino groups tended to maximize their interactions with the surrounding water solvent. Cluster analysis of the 200-nanosecond simulations determined that the main conformational family of shepherdin[79-83] was characterized by a slight bend geometry involving residues His-80, Ser-81, and Ser-82. Docking experiments on Hsp90 with this geometry predicted that shepherdin[79-83] bound to the ATP pocket of Hsp90. Two different orientations of shepherdin[79-83] were observed: one that corresponded to the global free-energy minimum structure of the shepherdin[79-83]-Hsp90 complex and one that represented the most frequently obtained structure after statistical clustering of all the structures studied during the Autodock simulations. The sites of contact between Hsp90 and shepherdin[79-83] in either configuration overlapped. In the global free-energy minimum configuration, the side chain of His-80 in shepherdin[79-83] made hydrophobic contacts with Ile-96 and hydrogen bonded with Gly-97 in Hsp90, the side chain of Ser-81 in shepherdin[79-83] hydrogen bonded with the side chains of Asp-102 and Asn-106 in Hsp90, and Ser-82 in shepherdin[79-83] hydrogen bonded with Asn-51 and Phe-138 in Hsp90. In the most frequently obtained shepherdin configuration, His-80 in shepherdin[79-83] formed a hydrophobic interaction with Ile-96 in Hsp90 but was also involved in a new hydrogen bonding interaction with the side chain of Asp-54 in Hsp90; Ser-81 interacted with Asp-93 and Asn-106 in Hsp90, and Ser-82 interacted with Asn-106 and Asp-102 in Hsp90. Consistent with these molecular dynamics predictions, in biochemical experiments, shepherdin[79-83] efficiently displaced ATP binding from the N-domain of recombinant Hsp90, whereas the scrambled peptide was ineffective.

2.4. Characterization of Hsp90/shepherdin binding interface. – The dominant conformations of shepherdin in solution were investigated through all-atom, explicit solvent MD simulations for a total time span of 400 ns. Statistical cluster analysis showed that shepherdin displays one main conformation, characterized by the presence of a turn involving residues G83-S84 and an overall hairpin geometry (see Plescia *et al.* [7] for details). The remaining clusters were mainly extended conformations, with the peptide backbone groups involved in hydrogen bonding with water.

The most populated conformation was subjected to multiple blind docking experiments on Hsp90 using the Autodock program. In all cases, the peptide was predicted to bind within the ATP binding site of Hsp90 (fig. 3A-C). Analysis of the blind docking results through the procedure described by Hetenyi *et al.* [19] showed that low-energy poses are all closely correlated with one another, with an RMSD from the global minimum structure lower than 2.5 Å. Control docking experiments were conducted with the extended structures representative of other clusters, but in those cases it was not possible to univocally identify any particular binding site on Hsp90.

The free-energy minimum structure of the Hsp90/shepherdin complex was then subjected to two long, 54 and 73 ns, all-atom MD simulations. Analysis of the statistical

and time-dependent distribution of the interactions between functional groups of the ligand and of the chaperone was carried out to develop pharmacophore models, keeping into account the motional and flexibility properties of both the ligand and the receptor. Shepherdin partially reoriented to increase the total number of stabilizing contacts with the ATP binding pocket of Hsp90 (fig. 3A,B). Attention was focused in particular on the analysis of hydrogen-bonding, hydrophobic/aromatic and charge-charge interactions, as these represent the most common types of intermolecular forces determining host/guest recognition in drugs.

The functional groups of shepherdin involved in direct or water-mediated hydrogen bonds with the binding pocket of Hsp90 included the γ -OH functions of Ser84, Ser85 and the imidazole ring of His86 (fig. 3B-E). The latter, in particular, could satisfy hydrogen-bonding conditions being involved in interactions both as an acceptor (N ϵ atom) and as a donor (N δ -H functional group). The remaining hydrogen-bonding group on shepherdin, Cys82, was involved to a lesser extent in intermolecular H-bonding interactions; however, its presence was shown experimentally to be necessary to ensure binding. Moreover, it displayed hydrophobic interactions with the side chains of Hsp90 Leu108 and the alkyl part of Asn109. Cys82 is also important for preserving the hairpin structure: mutations to Ala on the isolated peptide lead to loss of ordered hairpin structure.

To define the presence of hydrophobic/aromatic interactions, the contacts involving the side chains of Phe80 and His86 were monitored during simulation. Shepherdin Phe80 was found to be in contact mainly with the charged/polar side chains of Lys59, Asn52 and Asn55 on Hsp90, while shepherdin His86 was not significantly involved in hydrophobic/aromatic contacts with Hsp90 residues.

Finally, the role of the positively charged ammonium group on the side chain of shepherdin Lys87 was found to be only marginally involved in interactions with the backbone carbonyl oxygens of Hsp90 residues Phe135 and Gly136, being mostly exposed to the water solvent during MD simulations.

2.5. Pharmacophoric hypotheses and small molecule identification. – Three different pharmacophore models were built and labeled PHARM1, PHARM2 and PHARM3 based on the results of MD simulations. The conformation of shepherdin and the orientations of its side-chain functional groups in the most populated structural cluster from MD trajectories of the complex were used as structural template (fig. 3C,E). The distributions of dihedral values and distances among critical functionalities were used to define upper and lower boundaries for geometric constraints.

PHARM1 (fig. 3E) consisted of four pharmacophoric points: three H-bonding donor functionalities mapped over the side chain OH group of Ser84, Ser85 and the SH group of the Cys82, plus one imidazole ring moiety mapped on the position of the corresponding ring of His86 (fig. 3E). Each imidazole atom was allowed to bear any substituent or be a bridgehead in the presence of a fused ring. PHARM2 consisted of two H-bonding donor groups corresponding to the γ -OH of Ser84 and Ser85, one aromatic function centered on the position of the Phe80 benzene ring, and one hydrophobic function centered on the

S atom of Cys82. PHARM3 had the same properties as PHARM1, augmented by the presence of a positive charge mapped on the position of the ammonium group of Lys87.

The three models described above were used as queries for a search of the NCI3D database of molecules (containing approximately 160 000 compounds). The search with PHARM1 yielded 73 compounds, the search with PHARM2 yielded 42 compounds, while PHARM3 gave no hits. In experimental tests, the molecules corresponding to hits of PHARM2 proved to be extremely insoluble due to the presence of aromatic/hydrophobic groups and had thus no tumor-cell-killing effect. The search with PHARM1 yielded, among others, 20 hits reminiscent of the class of known purine-based inhibitors of the ATP-binding pocket of Hsp90. Interestingly, one of the non-purine-based hits that was found to be effective in experiments, AICAR (fig. 3E), was not previously known to interfere with Hsp90 chaperone functions and was characterized by a novel molecular structure among Hsp90 antagonists.

All of the data generated with simulations were punctually verified experimentally.

3. – Discussion

In this study, we used structure-based rational studies to identify and characterize shepherdin, a novel anticancer peptidomimetic modeled on the survivin-Hsp90 binding interface [6]. All theoretical predictions were subjected to experimental verification and the activities of the peptides were evaluated *both in vitro and in vivo*, in a large multidisciplinary effort. Shepherdin engages the ATP pocket of Hsp90 with unique binding characteristics, destabilizes survivin plus several additional client proteins, and causes massive killing of tumor cells by apoptotic and nonapoptotic mechanisms [7]. Shepherdin is selective in its antitumor activity, and does not affect the viability of normal cells or tissues, including human hematopoietic progenitors. When administered *in vivo*, shepherdin is safe and well tolerated, and inhibits growth of different tumor cell types without systemic or organ toxicity. Taken together, these features may make shepherdin an attractive lead prodrug for “targeted” cancer therapy [7].

Although initially designed as a high-affinity ($K_D \sim 80$ nM) inhibitor of the survivin-Hsp90 interaction, the data presented here suggest that shepherdin may function as a more global antagonist of Hsp90 chaperone activity. This conclusion is based on the structure-function analysis of shepherdin, and in particular its ability to expansively engage the chaperone ATP pocket, compete for the Hsp90-ATP complex, and destabilize multiple Hsp90 client proteins in addition to survivin, *in vivo*. Because of these features, shepherdin appears ideally suited to interfere with the periodicity of Hsp90 ATPase cycles, by directly preventing ATP binding [17], and/or by competing with cochaperone recruitment, especially that of p50^{cdc37}, which is required for ATPase activity and shares overlapping binding contacts with shepherdin [20]. In this context, the simultaneous destabilization of survivin levels [5], combined with the acute collapse of Hsp90 function [5], would be expected to cause a general breakdown of multiple cell proliferation and cell survival pathways in tumor cells, suitable for therapeutic exploitation [4].

When tested as an anticancer agent in tumor models, shepherdin was selective and well-tolerated, sparing normal cells, preserving colony-forming ability of purified human hematopoietic progenitors, and causing no organ or systemic toxicity after prolonged administration *in vivo*.

Experimental tests with the minimal sequence of five residues also confirm theoretical hypotheses. The results of MD simulations on the structure of Shepherdin[79-83] and of its complex with Hsp90 were challenged with competition experiments by use of enzyme-linked immunosorbent assay (ELISA). Apoptosis, Hsp90 client protein expression, and mitochondrial dysfunction were evaluated in Acute Myeloid Leukemia (AML) types (myeloblastic, monocytic, and chronic myelogenous leukemia in blast crisis), patient-derived blasts, and normal mononuclear cells. Effects of shepherdin[79-83] on tumor growth were evaluated in AML xenograft tumors in mice ($n = 6$). Organ tissues were examined histologically. Taken together, these results showed that Shepherdin[79-83] bound to Hsp90, inhibited formation of the survivin-Hsp90 complex, and competed with ATP binding to Hsp90.

Based on this knowledge, we used a structure- and dynamics-based rational design to identify the non-peptidic small molecule AICAR as a structurally novel inhibitor of Hsp90 (fig. 3E). The compound was selected to engage the ATP-binding pocket of the N-terminal domain of Hsp90, with binding and functional properties that mimic those of the peptidic antagonist of the survivin-Hsp90 complex, shepherdin. Accordingly, AICAR bound the Hsp90 N-domain, destabilized multiple Hsp90 client proteins *in vivo*, including survivin, and exhibited broad antiproliferative activity in multiple tumor cell lines, although at higher concentrations than those required to obtain the same growth inhibitory effect with 17-AAG, with induction of apoptosis and inhibition of cell proliferation. Reminiscent of the selective anticancer activity of shepherdin, AICAR did not affect proliferation of normal human fibroblasts.

In summary, shepherdin has the molecular features of both an inhibitor of a critical protein-protein interaction in tumor cells, *e.g.*, survivin-Hsp90, and an enzymatic antagonist of Hsp90 ATPase cycles. Because of these combined features, plus its considerably higher potency compared to other Hsp90 inhibitors, *e.g.*, 17-AAG, shepherdin may provide a potent and selective new anticancer agent in humans, consistent with the use of peptidomimetics in targeted cancer therapy. In addition, we narrowed the shepherdin binding interface to a short stretch of amino acids between H80 and C84 in the survivin sequence. Previously, mutagenesis of H80 or C84 resulted in dominant negative phenotypes with mitotic defects and induction of apoptosis in tumor cells, thus further underscoring their critical roles in survivin function. This small cluster of residues may thus provide a manageable platform for further derivatization of shepherdin, as well as for chemical screenings to identify shepherdin-like small molecules with enhanced, “targeted” anticancer activity in humans. The strategy presented here can be used in a general way to generate small molecule from peptide leads, keeping flexibility, solvation and dynamics into account.

REFERENCES

- [1] VOGELSTEIN B. and KINZLER K. W., *Nat. Med.*, **10** (2004) 789.
- [2] PAEZ J. G., JANNE P. A., LEE J. C., TRACY S., GREULICH H., GABRIEL S., HERMAN P., KAYE F. J., LINDEMAN N., BOGGON T. J., NAOKI K., SASAKI H., FUJII Y., ECK M. J., SELLERS W. R., JOHNSON B. E. and MEYERSON M., *Science*, **304** (2004) 1497.
- [3] WHITESELL L. and LINDQUIST S. L., *Nat. Rev. Cancer*, **5** (2005) 761.
- [4] NECKERS L. and IVY S. P., *Curr. Opin. Oncol.*, **15** (2003) 419.
- [5] ALTIERI D. C., *Nat. Rev. Cancer*, **3** (2003) 46.
- [6] FORTUGNO P., BELTRAMI E., PLESCIA J., FONTANA J., PRADHAN D., MARCHISIO P. C., SESSA W. C. and ALTIERI D. C., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 13791.
- [7] PLESCIA J., SALZ W., XIA F., PENNATI M., ZAFFARONI N., DAIDONE M. G., MELI M., DOHI T., FORTUGNO P., NEFEDOVA Y., GABRILOVICH D. I., COLOMBO G. and ALTIERI D. C., *Cancer Cell*, **7** (2005) 457.
- [8] VERDECIA M. A., HUANG H., DUTIL E., KAISER D. A., HUNTER T. and NOEL J. P., *Nat. Struct. Biol.*, **7** (2000) 602.
- [9] BERENDSEN H. J. C., POSTMA J. P. M., GUNSTEREN W. F. v., NOLA A. D. and HAAK J. R., *J. Chem. Phys.*, **81** (1984) 3684.
- [10] VAN GUNSTEREN W. F., BILLETER S. R., EISING A. A., HUNENBERGER P. H., KRUGER P., MARK A. E., SCOTT W. R. P. and TIRONI I. G. *Biomolecular Simulation: The GROMOS 96 manual and user guide* (vdf Hochschulverlag, ETH Zurich, Switzerland) 1996.
- [11] BERENDSEN H. J. C., GRIGERA J. R. and STRAATSMA T. P., *J. Phys. Chem.*, **91** (1987) 6269.
- [12] HESS B., BEKKER H., FRAAIJE J. G. E. M. and BERENDSEN H. J. C., *J. Comp. Chem.*, **18** (1997) 1463.
- [13] MIYAMOTO S. and KOLLMAN P. A., *J. Comp. Chem.*, **13** (1992) 952.
- [14] SPOEL D. V. D., LINDAHL E., HESS B., BUUREN A. R. v., APOL E., MEULENHOF P. J., TIELEMAN D. P., SIJBERS A. L. T. M., FEENSTRA K. A., DRUNEN R. v. and BERENDSEN H. J. C., *Gromacs User Manual version 3.2*. 2004, www.gromacs.org.
- [15] DAURA X., GADEMANN K., JAUN B., SEEBACH D., GUNSTEREN W. F. v., and MARK A. E., *Angew. Chem. Int. Ed.*, **38** (1999) 236.
- [16] MORRIS G. M., GOODSELL D. S., HALLIDAY R. S., HUEY R., HART W. E., BELEW R. K. and OLSON A. J., *J. Comp. Chem.*, **19** (1998) 1639.
- [17] STEBBINS C. E., RUSSO A. A., SCHNEIDER C., ROSEN N., HARTL F. U. and PAVLETICH N. P., *Cell*, **89** (1997) 239.
- [18] MEHLER E. L. and SOLMAJER T., *Protein Eng.*, **4** (1991) 903.
- [19] HETENYI C. and SPOEL D. v. d., *Protein Sci.*, **11** (2002) 1729.
- [20] ROE S. M., PRODROMOU C., O'BRIEN R., LADBURY J. E., PIPER P. W. and PEARL L. H., *J. Med. Chem.*, **42** (1999) 260.

This page intentionally left blank

Energy landscapes of bimolecular binding and molecular modulators of protein-protein interactions

G. M. VERKHIVKER

*Department of Pharmacology, University of California San Diego
9500 Gilman Drive, La Jolla, CA 92093-0392, USA*

1. – Energy landscapes and molecular recognition

Molecular recognition between proteins and flexible target molecules, including other proteins, nucleic acids and small molecules is often accompanied by a considerable flexibility of the protein binding sites and structural rearrangements upon binding between the associated partners [1-4]. Accessibility of alternative conformational states is important for protein function, including assembly, molecular recognition, regulation of biological activity, and enzymatic catalysis. It has been recognized that proteins are not adequately described by a single conformational state, but are better represented by a manifold of low-energy protein conformations, conformational substates, on a rugged energy landscape [4-9]. Current view of the protein energy landscape picture implies that the conformational substates, that represent local minima of the protein, are organized in hierarchical tiers that are separated by barriers which can be crossed by thermal activation. Within this hierarchy, alternative conformational states are defined by significant differences in protein conformation and large energy barriers, while modest coordinate changes, with concomitantly smaller energy barriers, characterize alternative protein conformational substates. Furthermore, the distribution of protein conformational substates is heterogeneous, and may depend on the topology of the protein structure where the proteins with similar architectures can exhibit similar large-scale dynamic behavior [10,11]. Analogous to a typical folded protein, ligand-protein complexes generally have a well-defined native structure, but on a microscopic level a ligand-protein system may exhibit

structural disorder that is revealed on different length and time scales: by rotation of a local protein side-chain, by conformational change of the ligand in the active site or by a collective conformational change associated with a movement of the protein backbone, side-chains and a change of the ligand binding mode. Local flexibility within a protein often changes upon target binding involving a significant change of conformational entropy, which is typically expressed by equilibrium fluctuations between conformational substates, while retaining a global topology of the native fold. The magnitude of conformational diversity observed in proteins ranges from fluctuation of side chains to the movement of loops and secondary structures, and even to global tertiary structure rearrangements. NMR, in particular, has proven particularly effective in revealing the true conformational diversity of proteins. Dynamics of the peptide-protein interfaces and protein flexibility can have a profound effect in determining binding thermodynamics and kinetics and modulating binding affinity and specificity [12-16].

Analogous to the protein folding phenomenon [17-19], the process of complex formation, when one or both binding partners are flexible, can evolve over a vast conformational space and under the influence of a multitude of conflicting interactions. A large number of available conformational states during molecular recognition of large flexible protein systems necessitates the use of a statistical characterization and the energy landscape analysis, originally introduced in protein folding and further developed in ligand-protein binding [20-25]. The energy landscape view emphasizes a description of a system in terms of statistical ensembles and as a result a complex interplay between the entropy and enthalpy contributions to the free energy during reaction. The emerging realization that protein folding and molecular recognition phenomena share a number of universal aspects, including the complex nature of interactions on the underlying energy landscape, is manifested by a balance between opposing thermodynamic forces during the process, the loss of conformational entropy and the energy gain upon the native structure formation, that ultimately determines the thermodynamic free energy barrier of the reaction. These physical arguments, that have been fruitful in understanding protein folding mechanisms, have also led to an elegant, fly-casting mechanism derived from the energy landscape theory and proposed to explain kinetic advantages of unstructured proteins in binding [26-28]. The energy landscape theory of molecular recognition has been exploited in a phenomenological mechanism of biomolecular binding which implies the presence of an ensemble of multiple conformational states for both interacting molecules and thereby generalizes the traditional concepts of "lock and key" and "induced fit" mechanisms. This model, often termed "conformational selection", (fig. 1) assumes that conformational flexibility of the binding partners is determined by the equilibrium distribution of the protein and ligand conformational states, which is shifted upon binding towards the thermodynamically most stable complex [29-33]. It has been suggested that variations in structural stability and flexibility during molecular recognition may be associated with the ruggedness at the bottom of the underlying binding energy landscape. The differences in the shape of the binding energy landscape can be related to various functions such as specificity or permissiveness in recognition. The less stable and less selective complexes may have a more rugged bottom of the energy landscape with low

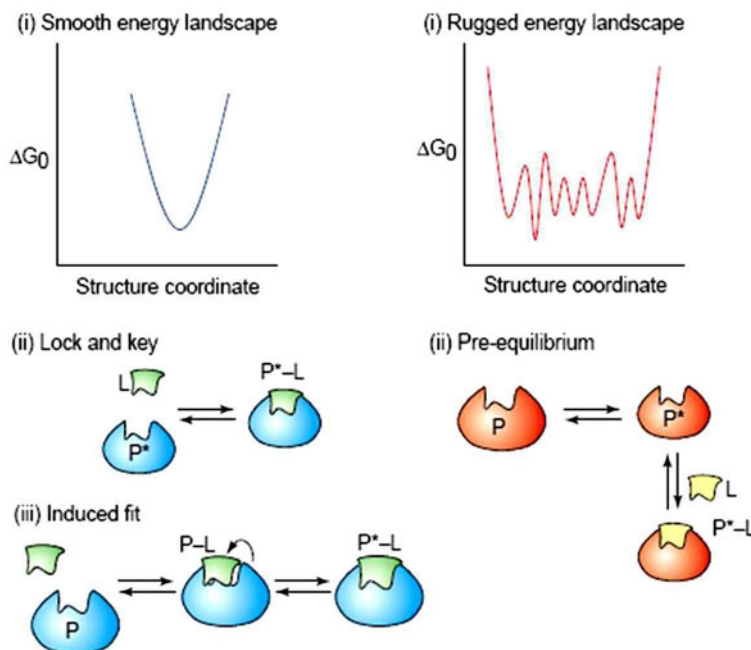


Fig. 1. – The single stable conformer model (i) corresponds to a functional mode of either lock and key (ii) or induced fit (iii). The energy landscape model of biomolecular binding: conformational flexibility of the binding partners is determined by the equilibrium distribution of the pre-existing protein and ligand conformational states, which is shifted upon binding towards functional state corresponding to the thermodynamically most stable complex (ii). Adopted from and reproduced with the kind permission from James LC, Tawfik DS (2003) Conformational diversity and protein evolution—a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361-368.

barriers between conformers of the complex. Highly specific peptide-protein complexes are likely to be relatively rigid, with a steep funnel of conformations leading to the native structure (fig. 1). In the “induced fit” model, the molecular association of flexible interfaces and specificity of biomolecular interactions can be described by conformational isomerization or folding of one or both partners to attain the native stable complex which follows the formation of an initial encounter complex. The sequential selection is reminiscent of the idea of kinetic proof-reading [34-36] and may have implications for the binding mechanism as the required binding specificity may be achieved in this case through formation of an intermediate metastable complex, paving the way for a slower, irreversible evolution towards a highly specific complex with a slow rate of dissociation. From a thermodynamic perspective, the conformational selection model and induced fit can be considered as complementary models, as due to thermal fluctuations there always exists some character of the induced fit on the atomic scale of molecular interactions. However, from a structural and dynamic perspective, ligands of similar binding affinity may bind

via different mechanisms and lead to distinct interactions with tertiary elements of a signaling pathway. The crystal structures of a monoclonal IgE antibody, SPE7, existing in two very different conformations each binding structurally distinct antigens, provided support to the pre-existing equilibrium hypothesis [37]. While the predominant unbound conformation has a flat binding site, which is reminiscent of antibodies that bind proteins or peptides, the alternative conformational state contains a deeper pocket, typical for binding haptens [37].

2. – Binding hot spots and convergent solutions at protein-protein interfaces

The recent evidence from direct evolution experiments has indicated that substantial changes in functional promiscuity of a protein are often linked with the inherent conformational diversity, which can be considered to be traits of the ability of a protein to evolve [37-39]. Protein binding sites have evolved not only to be conformationally flexible, but also functionally adaptive with a diverse repertoire of protein systems capable of binding with high affinity to different binding partners [40-42]. These studies have revealed that evolution can find convergent solutions to stable intermolecular interfaces by using structural plasticity of the binding site residues. Multiple binding partners can target the same hot spot residues in the protein binding site by mimicking specific interactions of the natural proteins while using the interacting groups from the different structural scaffolds. Alanine scanning mutagenesis of protein-protein interfacial residues, combined with structural and thermodynamic studies, have enabled discovery of energetically important regions at intermolecular interfaces that are critical in determining binding affinity [43,44]. Binding hot spots have been detected in various protein-protein interfaces where, despite a typically large size of an intermolecular interface, binding affinity and specificity can be determined by a localized functional epitope consisting of only a small fraction of the interfacial residues [45]. A comprehensive analysis of protein-protein interfaces and a survey of hot spots compiled from various protein binding sites have demonstrated a diversity of interaction patterns and a lack of general rules for hydrophobicity, polarity, or shape, that can be used to unambiguously predict hot spots at the intermolecular interfaces [46,47]. A recent analysis of conserved residues in 11 clustered interface families comprising a total of 97 crystal structures has shown that the composition of hot spots is typically enriched by certain residues, such as Trp, Tyr, Arg, His, Gln, Asn, and Pro, and can be surrounded by a shell of less important residues, shielding hot spot residues from solvent and, thereby, providing a requirement for their favorable interactions [47-49].

The relationship between conformational flexibility and convergent evolution has been demonstrated in a recent study of the recognition in the hinge region of the constant fragment (Fc) of human immunoglobulin G which represents the consensus recognition site for binding natural proteins and synthetic peptides [42]. The consensus binding region on the surface of the Fc protein, which is characterized by a high degree of solvent accessibility and a predominantly nonpolar character of the interface, undergoes a series of local conformational changes in order to complement the distinct surfaces of each binding

partner. Interestingly, the low hydrogen bonding ability in the binding region (19% of the surface is capable of hydrogen bonding compared with 37% on average) has indicated that this site places fewer geometric constraints on binding partners, because fewer complementary polar interactions are necessary for binding [42]. The crystal structures of the Ig-Fc protein in complexes with domain C2 of protein G, domain B1 of protein A, and 13-residue cyclic DCAWHLGELVWCT peptide have been determined at high resolution [42]. From the energy landscape perspective on molecular recognition, evolutionary conserved residues in the binding site of the Ig-Fc protein may have a strong effect on the energy gradients towards the native structure. In our recent energy landscape analysis of biomolecular binding, we have investigated the role of the hot spot residues at the Ig-Fc protein interface in fulfilling the thermodynamic and kinetic requirements for specific binding [50].

3. – Targeting P53-MDM2 interfaces with molecular modulators

An intriguing consequence of the energy landscape perspective on protein conformational diversity is that it may provide a model for analysis of functional adaptability and multi-specificity whereby a protein accommodating multiple conformational states may exert structural plasticity of the binding site and respective diversity of the binding functions. Structural plasticity and equilibrium switching between distinct conformational states of the MDM2 (murine double minute 2) protein may involve global and local rearrangements of both backbone and side-chains, which may lead to different active site configurations and different binding specificities. During the past decade, MDM2 has emerged as the principal cellular antagonist of p53 by limiting the p53 tumor suppressor function [51-60]. The p53 protein is inactivated in the presence of MDM2 and does not stimulate the expression of genes involved in apoptosis, cell cycle arrest, or DNA repair. In some tumors where MDM2 is overexpressed, p53 is constantly inhibited and tumor growth is favored. The regions corresponding to residues 17-125 of MDM2 have been crystallized in complex with the p53 peptide (15-29) and the biochemical basis of MDM2-mediated inhibition of p53 was further elucidated by the X-ray structure solved at 2.6 Å resolution [61,62]. The crystal structure has shown that MDM2 forms a deep hydrophobic cleft (about 25 Å long and 10 AA wide), lined with 14 hydrophobic and aromatic residues into which the transactivation domain of p53 binds, thereby concealing itself from the interaction with the transcriptional machinery. Site-directed mutagenesis has shown the importance of p53 residues Leu-14, Phe-19, Leu-22, Trp-23, and Leu-26, of which Phe-19, Trp-23, and Leu-26 are the most critical. Similarly, mutations of MDM2 at residues Gly-58, Glu-68, Val-75, or Cys-77 result in lack of p53 binding [63,64]. The two large hydrophobic amino acids p53 Phe-19 and Trp-23 are located face to face on the same side of the helix, and together with p53 Leu-26, they point toward the MDM2 protein, where they are surrounded by several hydrophobic residues (Leu-54, Leu-57, Ile-61, Met-62, Tyr-67, Val-75, Val-93, Phe-86, Ile-99, Phe-91, and Ile-103) from the MDM2 protein. p53 makes H-bonds between the backbone amide of Phe-19 and the O of MDM2 Gln-72, between the N of Trp-23 and the backbone carbonyl of MDM2 Leu-54,

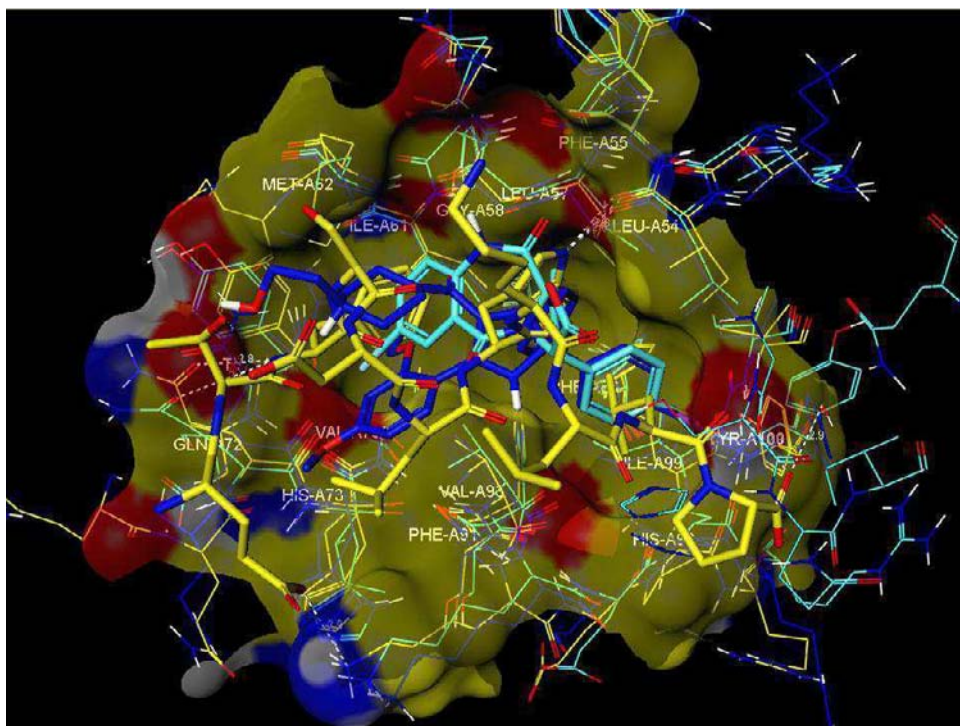


Fig. 2. – The crystal structures of p53 peptide (in yellow), Nutlin-2 (in blue) and 1,4-benzodiazapine-2,5-dione analog (in light blue) in the binding site of MDM2. A view of intermolecular network formed by the p53 peptide and the inhibitors in the MDM2 binding site. Conformational variations of the MDM2 active residues forming a hot spot at the intermolecular interface: the MDM2 binding site residues from the crystal structure of the XDM2 in the complex with the bound p53 peptide (PDB entry 1YCQ) are shown in yellow; the MDM2 binding site residues from the crystal structure of humanized MDM2 with Nutlin-2 (PDB entry 1RV1) shown in blue. the MDM2 binding site residues from the crystal structure of humanized MDM2 with a 1,4-benzodiazapine-2,5-dione analog (PDB entry 1T4E) are shown in light blue.

and between the backbone carbonyl of Asn29 and the hydroxyl of MDM2 Tyr-100. The interaction between p53 and MDM2 is essentially hydrophobic, and 70% of the atoms at the interface are nonpolar (fig. 2). Structural studies of the MDM2-p53 interactions and binding with the p53-derived peptides revealed global conformational changes of the overall structure of MDM2, stretching far beyond the binding cleft, indicating significant changes in the domain dynamics of MDM2 upon ligand binding [65-71]. Experimental measurements of the strength of the p53-MDM2 bond range from a K_d of 60 to 700 nM, depending on the length of the p53 peptide. The solution structure and dynamics of apo-MDM2 revealed the nature of the conformational changes in MDM2 that accompany binding of p53 [72]. Conformational plasticity of the apo-MDM2 could allow for the more open conformation of the binding cleft than the one observed in structures of

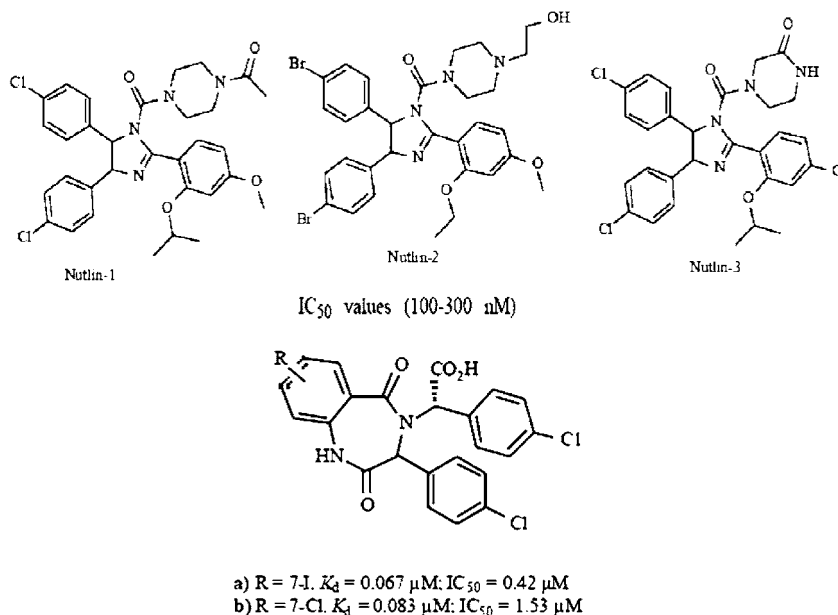


Fig. 3. – (A) Chemical structures of Nutlins, *cis*-imidazoline MDM2 binding inhibitors (B) Chemical structures of benzodiazepinediones, including TDP521252 and TDP665759 —1,4-benzodiazapine-2,5-dione MDM2 inhibitors.

complexes with small molecules and peptides [72]. As a result, the MDM2 protein may bind specifically to several different partners, although, to date, all the known liganded structures of MDM2N are highly similar to one another. Structural studies of MDM2 have also indicated that the protein region between residues 16 and 24, can form a lid that closes over the p53-binding site [73]. The binding of p53 to MDM2 induces conformational changes resulting in ultimate displacement of the lid which competes with p53 for the same binding site.

The current quest for small molecule MDM2 antagonists for cancer therapy is based on the rationale that relieving p53 repression by MDM2 will result in enhanced p53 activity in cancers with wild-type p53. The only inhibitors available for a long time were synthetic peptides, which have demonstrated to be very useful tools to show both *in vitro* and *in vivo*, inhibition of MDM2-p53 interaction and induce apoptosis in tumor cells, but they are far from becoming drugs due to their insolubility or impermeability [74-78]. Small molecule antagonists featured as effective anticancer for targeting MDM2 for cancer therapy have recently begun to emerge, although the repertoire of these inhibitors is rather limited and none of them has yet entered human clinical trials [79,80]. A class of synthetic low molecular weight compounds, *cis*-imidazoline analogs referred to as Nutlins has recently been discovered as potent inhibitors of MDM2-p53 interactions. The crystal structure of one of the compounds, Nutlin-2 (fig. 3) has been solved in the complex with

the humanized MDM2 [81-83] revealing that these inhibitors utilize the imidazole ring as a “holding” scaffold projecting three hydrophobic groups into the three hydrophobic pockets of MDM2 and mimicking critical MDM2-p53 interactions in the binding site (fig. 2). Recently, a library of 22 000 1,4-benzodiazepine-2,5-diones was screened using the high throughput direct binding assay for binding to the p53-binding domain of MDM2 [84, 85]. The most potent antagonists (fig. 3) that were co-crystallized with HDM2 have also mimicked p53 binding pattern, with two chlorophenyl substituents occupying two of the three hydrophobic pockets of the MDM2 cleft, while an iodophenyl or chlorophenyl group occupies the third hydrophobic pocket (fig. 3). An integrated, virtual database screening strategy has led to a series of potent quinolinol-based compounds bound to MDM2 with a K_i of 120 nM and representing a promising new class of non-peptide inhibitors of the MDM2-p53 interaction [86, 87]. The tertiary structures of small-molecule complexes which have been solved to date with the humanized *Xenopus* MDM2 complexes share their alpha-helix mimetic properties and are similar to that of the original complex p53-derived peptide with the human MDM2, suggesting that the *Xenopus* version of MDM2 may be less conformationally mobile and more preordered to binding.

Computational studies of MDM2-p53 binding were pioneered by Kollman and coworkers when they introduced a computational alanine scanning approach and evaluated the individual contributions of key p53 residues to the binding free energy [88]. This approach has been further developed to examine the binding MDM2-p53 interface and the effect of mutating key residues in the human p53-MDM2 complex to computationally design a potent beta-peptide mimic of p53 [89]. Structure-based computational design was successfully applied to design libraries of p53 mimics using a scaffold that projects side-chain functionalities using the pharmacophore requirements derived from the MDM2 interacting motif of p53 [90]. A computational approach to increase the affinity of a protein-peptide complex was recently introduced by designing extensions which interact with the protein outside the canonical peptide binding pocket. The approach was tested by designing extensions for p53 to bind with increased affinity to the MDM2 oncoprotein [91]. Conformational changes of the p53-binding cleft of MDM2 have been studied by molecular dynamics simulations, suggesting a wider and more stable topology of the binding cleft in the complexes, while apo-MDM2 exhibited a narrower and highly flexible cleft [92]. Conformational flexibility of MDM2 and the existence of multiple ligand-binding domains highlight the complex intradomain communication and considerable versatility of MDM2 in binding with small molecules and interacting proteins which may be largely determined by the underlying binding energy landscape.

4. – Materials and methods

4.1. *Structural analysis.* – The crystallographic conformations of the Ig-Fc protein used in simulations with an engineered 13-residue peptide include: 1DM2 (co-crystal of human IGG1 with the DCAWHLGELVWCT peptide); 1FCC (crystal structure of the C2 fragment of protein G in the complex with the Fc domain of human IG); and 1FC2 (crystal structure of a human Fc fragment of IG in the complex with the fragment B

of protein A). All publically available crystal structures of the MDM2 protein from the Protein Data Bank (PDB) [93] are used to categorize the conformational space of MDM2: 1TTV (NMR structure of a complex between MDM2 and a small molecule inhibitor); 1Z1M (24 NMR structures of unliganded MDM2); 2GV2 (MDM2 in the complex with an 8-mer p53 peptide analogue); 1RV1 (crystal structure of human MDM2 with an imidazole inhibitor Nutlin-2); 1T4F (structure of human MDM2 in the complex with an optimized p53 peptide); 1YCQ (crystal structure of xenopus MDM2 bound to the tansactivation domain of p53); 1YCR (crystal structure of human MDM2 bound to the tansactivation domain of p53); 1T4E (crystal structure of human MDM2 in complex with a benzodiazepine inhibitor); 2AXI (crystal structure of human HDM2 in complex with a beta-hairpin). We have also expanded a spectrum of low-energy conformational states which may mimic a range of protein equilibrium fluctuations near the native structures. The computational procedure uses the penultimate rotamer library [94] to build and optimize multiple rotamers for the critical MDM2 binding site residues, including Leu-54, Leu-57, Ile-61, Met-62, Tyr-67, Gln-72, Val-75, Phe-86, Phe-91, Val-93, Ile-99, Tyr-100, and Ile-103. The energies of the modified protein are optimized using a simple self-consistent procedure in which optimization of the rotamer position for a modified residue is followed by the same procedure for the next mutated residue until convergence is achieved.

4.2. Hierarchical models of biomolecular recognition. – We employ a hierarchical model, which provides an opportunistic solution to achieve a synergy of an adequate conformational sampling and an accurate binding energetics. Equilibrium sampling of the inhibitor conformations with the multiple conformational states is first performed by replica-exchange Monte Carlo simulations [95,96] using the soft-core AMBER force field. The molecular recognition energetic model used in dynamics simulations includes intramolecular energy terms, given by torsional and nonbonded contributions of the DREIDING force field [97], and the intermolecular energy contributions calculated using the AMBER force field [98] to describe protein-protein interactions combined with an implicit solvation model [99,100]. The simplified soft-core force field is assumed to reproduce adequately the shape of the “true” potential and detect the low-energy protein and inhibitor states. Hierarchical model of biomolecular recognition relies on the robustness of the soft-core AMBER energy function in characterizing the multitude of the low-energy binding modes combined with the accuracy of the MM/GBSA approach [101,102] which is applied to the generated bound states to quantify the binding free energy values.

5. – Results and discussion

5.1. Conformational landscape of MDM2 and specific binding with small molecular mimics. – Conformational differences in the MDM2 structures are reflected in significant variations of the p53-binding cleft (fig. 3) indicating significant changes in the MDM2 dynamics upon ligand binding. The binding site of the apo-structure of human MDM2 is quite shallow and is considerably less accessible for binding than the complexed forms of MDM2. Although the tertiary structures of the small-molecule complexes with the hu-

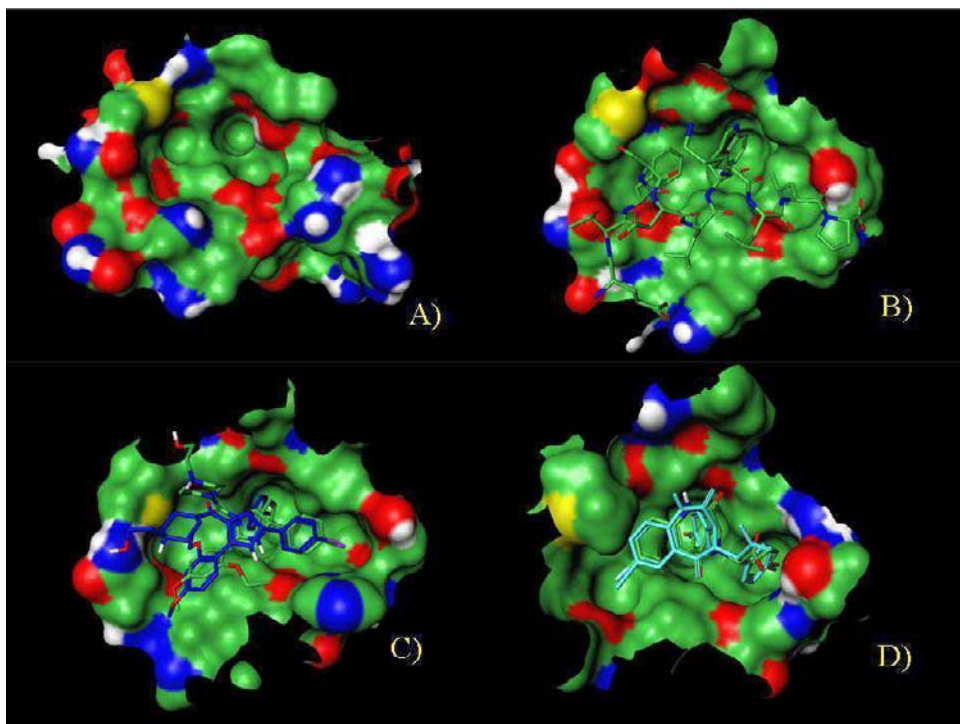


Fig. 4. – A) The Connolly surface of the MDM2 binding site of a representative solution structure of apo-MDM2 (one of 24 reported solution structures is used; PDB entry 1Z1M). B) The Connolly surface of the MDM2 binding site of the crystal structure of the XDM2 in the complex with the bound p53 peptide (PDB entry 1YCQ). The crystallographic conformation of p53 peptide is shown in default colors. C) The Connolly surface of the MDM2 binding site from the crystal structure of humanized MDM2 with Nutlin-2 shown in blue (PDB entry 1RV1) superimposed with the computationally predicted lowest energy structure of the inhibitor (shown in default colors). D) The Connolly surface of the MDM2 binding site from the crystal structure of humanized MDM2 with a 1,4-benzodiazapine-2,5-dione analog in light blue (PDB entry 1T4E) superimposed with the computationally predicted lowest energy structure of the inhibitor (shown in default colors).

manized *Xenopus* MDM2 (pdb entries 1RV1, 1T4E) are seemingly similar to that of the original p53 peptide complex with the human MDM2 (pdb entries 1YCQ, 1YCR), there are a number of important differences in the side-chain conformations of the binding site residues which are necessary for providing the high affinity inhibitor binding. Structural differences in the side-chain orientations which are influenced by the binding partner can be observed for Leu-57, Met-62 Gln-72, Leu-57, Lys-94, and Tyr-100 residues (fig. 2). These residues appear to assume multiple conformations resulting in a considerable structural plasticity of the MDM2 binding site and the ability to accommodate distinct binding partners (figs. 2, 4). Consequently, the organization of the MDM2 conformational land-

scape may be quite complex with the different tiers of protein fluctuations produced in response to binding large peptides and small molecules, which may be associated with collective movements of the protein backbone and side-chains and separated by appreciable energy barriers. This structural analysis calls for a study of the MDM2 conformational preferences in binding with small molecules which is important to understand the conformational landscape of MDM2 binding. We analyzed the role of MDM2 conformational flexibility in eliciting specific binding with small molecular mimics, specifically the highly potent imidazole-based Nutlin-2 and 1,4-benzodiazepine-2,5-dione analog with the known co-crystal structures. A recently introduced computational approach for examining protein binding specificities is used to model the inhibitor binding with the ensembles of all publically available MDM2 crystal structures and NMR structures, including the apo-MDM2 structure and the MDM2 bound conformations in complexes with the peptides and small molecular mimics. In these computational experiments, parallel simulated tempering with the multiple MDM2 protein conformations operates in the structural space of the known crystallographic conformations which always maintain their native fold, while the conformational space of the inhibitor and the dynamics of the inhibitor-MDM2 interactions are extensively sampled at a range of temperatures. The important result of this *in silico* experiment is that studied inhibitors elicit specific MDM2 conformational preferences that are almost exclusively determined by the MDM2 crystallographic conformations from the complexes with small molecular modulators (fig. 4). In contrast, the NMR structures of apo-MDM2 and MDM2 bound conformations from the complexes with the peptides are considerably less favorable during equilibrium fluctuations with the inhibitors. This result reinforces the earlier conjecture proposed in the structural study of apo-MDM2 that the binding cleft of the apo-form which is very different from that of the complexed form is shallow and not likely to provide much opportunity for ligand design. Furthermore, we have observed that the MDM2 crystal structure from the complex with the 1,4-benzodiazepine-2,5-dione contributes greatly to the binding equilibrium in simulations with both benzodiazepine-based analog and imidazole-based Nutlin-2 molecule. However, the MDM2 bound conformation from the crystal structure of the Nutlin-2 complex appears to be more specific and contributes greatly to equilibrium only in simulations with Nutlin-2 (fig. 5). The MDM2 bound conformation from the complex with the 1,4-benzodiazepine-2,5-dione may provide an entropically favorable, low-energy basin favorable in binding with a number of small molecules, which may be instructive in the design of new classes of MDM2 inhibitors to increase the chemical diversity of compounds targeting MDM2-p53 interactions. The computed binding free energies are in agreement with the experimental values and support initial insights from the equilibrium simulations revealing that high affinity inhibitors induce specific protein responses upon ligand binding (fig. 5). These results favor the hierarchical organization model of the MDM2 conformational landscape, suggesting that the range of favorable MDM2 conformational fluctuations in response to binding small molecules may be separated from a tier of the conformational states relevant for binding p53-derived peptides and peptidomimetics. A molecular level analysis of the MDM2 binding with existing small molecular mimics demonstrates that conformational flexibility may allow for

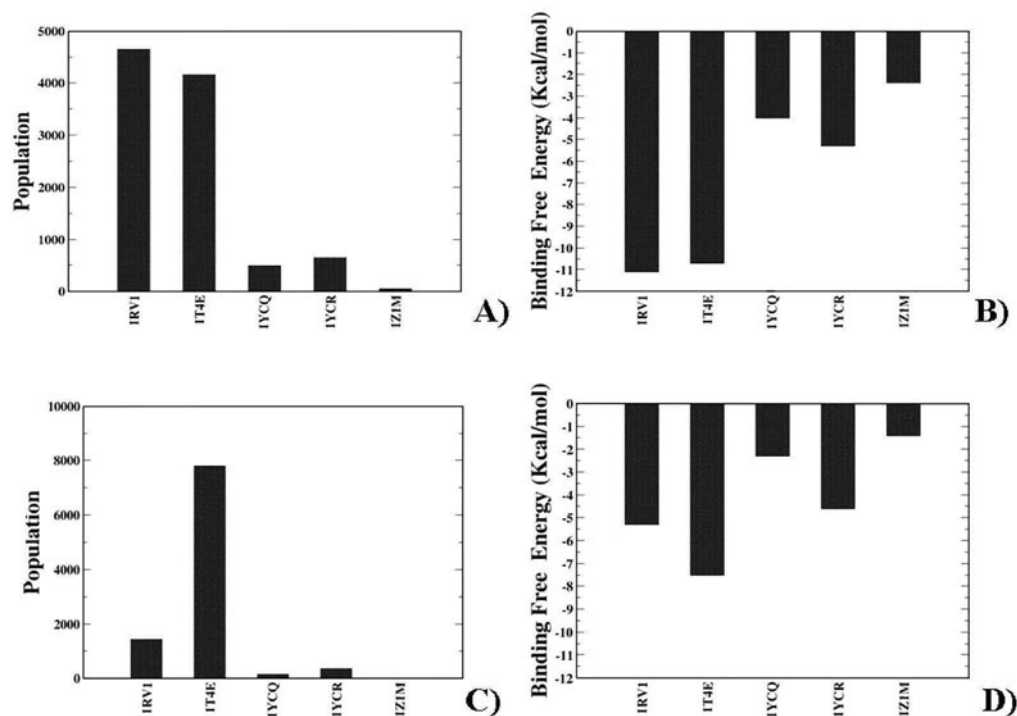


Fig. 5. – The equilibrium distribution of the MDM2 structures interacting with the Nutlin-2 inhibitor (A) and a 1,4-benzodiazapine-2,5-dione analog (C). X-axis labels denote the MDM2 pdb entry codes. Binding free energy analysis of the Nutlin-2 inhibitor (B) and a 1,4-benzodiazapine-2,5-dione analog (D) with the MDM2 structures.

functional diversity without depending on the evolution of sequence diversity, which can greatly facilitate the potential for rapidly evolving new functions and structures. The proposed analysis reveals important molecular determinants of the binding specificity and may be useful for better understanding of the biological implications of the MDM2 dynamics and future efforts in anticancer drug design.

5.2. The energy landscape analysis of a hot spot at the consensus binding site of the constant fragment (Fc) of human immunoglobulin G. – Conserved interactions that are present in all binding interfaces with the Fc fragment of Ig share a set of contacts with side-chains of key residues —His-433, His-435, Ser-254, Ile-253, Met-252, Met-428, Arg-255, Asn-434, and Tyr-436. The conserved peptide-protein interactions in the crystal structure of the native complex include hydrophobic packing onto Met-252 and His-435 residues, hydrogen bonding to main chain of Ile-253, and salt bridges with His-433, whereas specific interactions with the peptide are also formed via hydrogen bonding to Asn-434, Ser-254, and salt bridges with Arg-255 (fig. 6). We present a computational

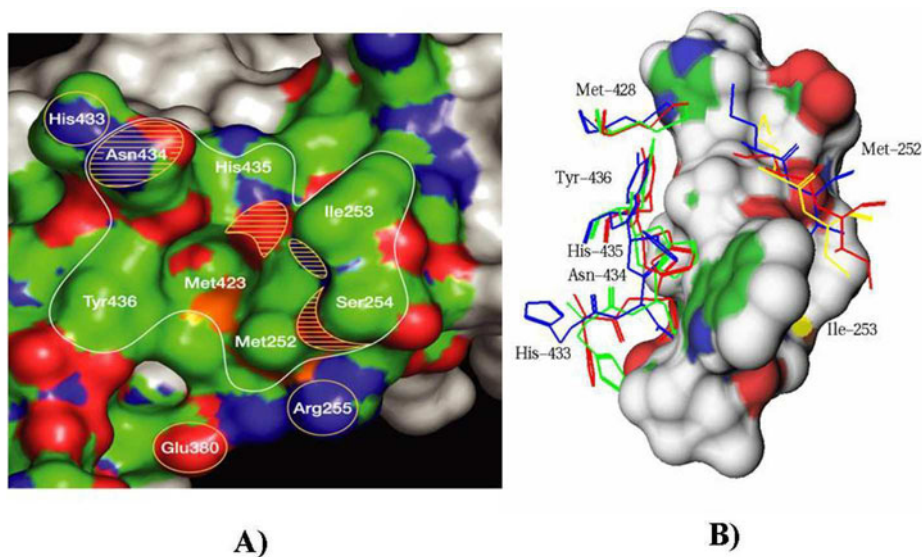


Fig. 6. – (A) The hinge region of the constant fragment (Fc) of human immunoglobulin G (Ig) protein, which represents the consensus binding site for natural proteins and synthetic peptides, including domain B1 of protein A, domain C2 of protein G, rheumatoid factor, neonatal Fc-receptor. Adopted from and reproduced with the kind permission from Arkin, M. R. and Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. (*Nature Rev. Drug Discov.*, **3** (2004) 301-317). (B) Conformational variations of the key protein active residues forming a hot spot at the intermolecular interface from the native complex (side-chains are shown in yellow), from the complexes with the protein G (side-chains in blue) and protein A (side-chains are shown in red). Connolly surface of the DCAWHLGELVWCT peptide color coded by atom type.

analysis of the thermodynamic and kinetic aspects of molecular recognition between a 13-residue cyclic peptide DCAWHLGELVWCT and the Fc fragment of the Ig protein studied using simulations of the peptide binding with an ensemble of multiple protein conformations. It has been proposed that different proteins may bind to the consensus binding site of the Fc fragment of Ig by redistributing the equilibrium population of the protein conformational substates depending upon the corresponding binding partner. The equilibrium fluctuations of the DCAWHLGELVWCT peptide at the temperature range between $T = 300$ K and $T = 400$ K (fig. 7) show that at higher temperatures the low-energy peptide bound conformations populate both the native-like structures and alternative peptide conformations, bound to alternative regions on the protein surface. As temperature lowers, the native-like conformations tend to dominate the thermodynamic equilibrium. The scatter graphs between the root mean square deviations (RMSD) from the crystal structure of the peptide and energies generated from equilibrium binding simulations at a range of temperatures between $T = 300$ K and $T = 400$ K reveal energy gradients, or binding funnels near the binding site, which are characterized by a

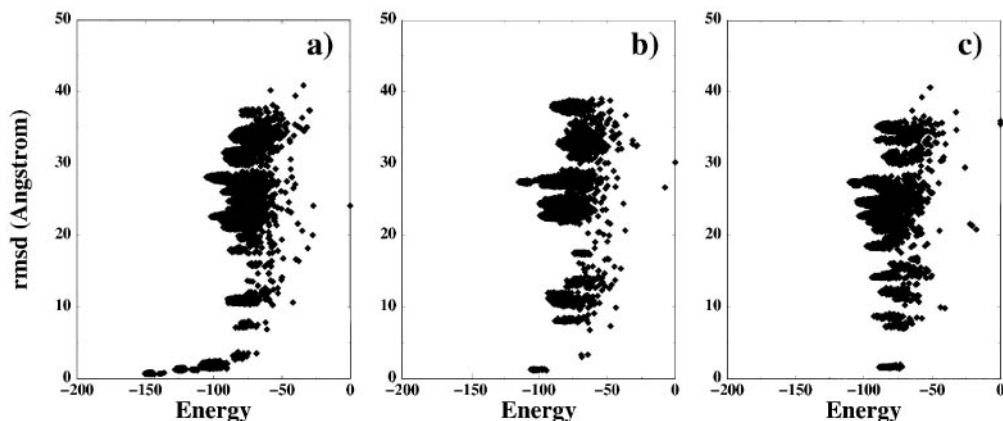


Fig. 7. – The scatter graphs between the RMSDs from the crystal structure of the peptide and energies generated from equilibrium binding simulations at a range of temperatures between $T = 300$ K and $T = 400$ K with the native protein conformation (a) and protein bound conformations from the complexes with the protein G (b) and protein A (c).

steady decrease in energy of the system as the degree of similarity between the native and docked peptide structures increases (fig. 7a). At lower temperatures, as the system begins to sample low-energy states, we observe a considerable stability gap between the crystal structure of the complex and alternative low-energy peptide conformations. This gap provides the thermodynamic stability of the native state and allows to locate unambiguously the crystal structure of the complex as the lowest energy conformation in equilibrium simulations. The binding energy landscape displays a funnel-like shape near the binding site of the native complex and a distinct energy gradient toward the native structure, which is not present when the peptide interacts with alternative protein conformations of Ig (figs. 7b, 7c). The presence of the dominant binding funnel near the active site of the native protein for the recognition peptide indicates that the shape of the binding energy landscape for the native complex may contribute to the mechanism of specific recognition. The consensus binding region of Fc undergoes local conformational changes to accommodate distinct binding partners. In particular, positional and side-chain adjustments of the hydrophobic residues Met-252 and Ile-253 provide a considerable extent of adaptability in the consensus binding site. We attempt to delineate the role that these residues could play in contributing to the funnel-like character of the energy landscape near the binding site and determining kinetic accessibility of the native complex. We have conducted equilibrium binding simulations of peptide binding with various single alanine mutants of the native protein structure. In these structures, one at a time, protein active-site residues His-433, His-435, Ile-253, and Met-252 were mutated to Ala. We find that deletions of hydrophobic side-chains for Met-252 (fig. 8a) and Ile-253 (fig. 8b) and the mutations in His-433 (fig. 8c) and His-435 residues (fig. 8d) have indeed a noticeable effect on the shape of the energy landscape. A significant en-

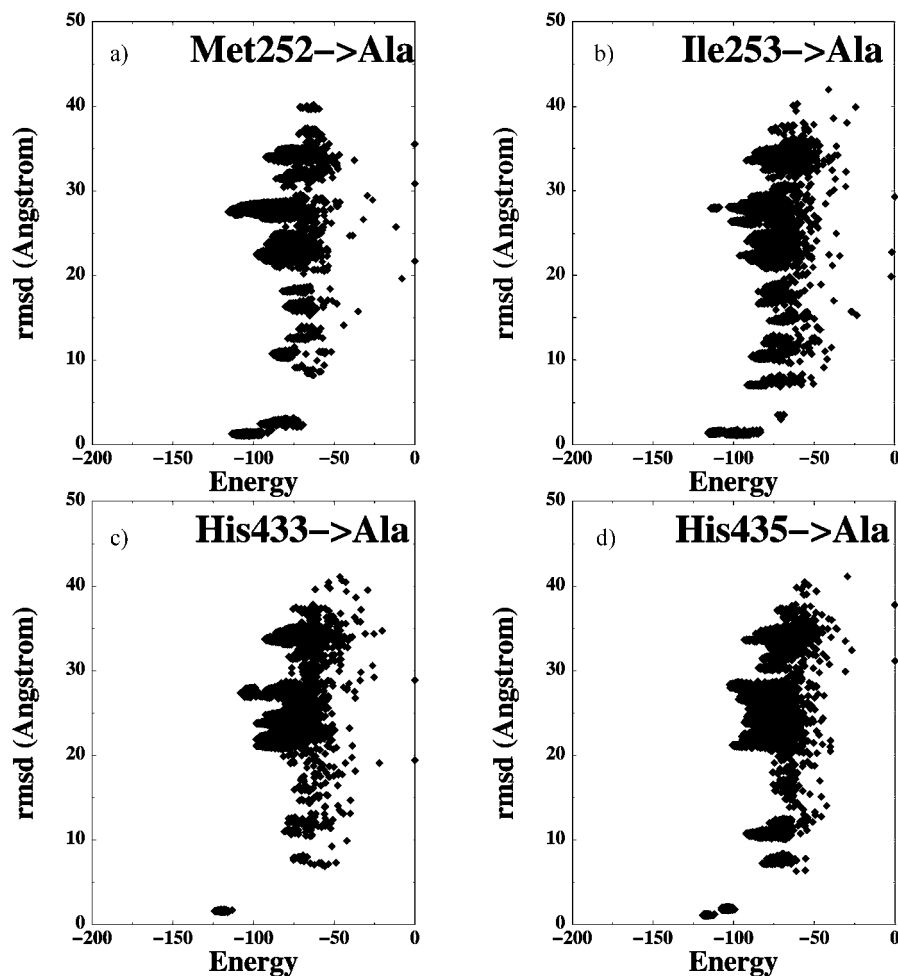


Fig. 8. – The scatter graphs between the RMSDs from the crystal structure of the peptide and energies generated from equilibrium binding simulations at a range of temperatures between $T = 300$ K and $T = 400$ K with the native protein conformation where Met-252 is mutated to Ala (a), Ile-253 is mutated to Ala (b), His-433 is mutated to Ala (c), and His-435 is mutated to Ala (d).

ergy gap, which is present in the wild-type complex, is reduced for these mutants and there is a pronounced overlap in the energies of native-like conformations and alternative binding modes of the peptide. Hence, these active-site residues, which form a majority of favorable hydrophobic contacts, may contribute to kinetic accessibility of the native complex by creating an energy gradient near the active site towards the crystal structure of the complex. The results indicate that a combination of hydrophobic contacts and specific interactions with Met-252, Ile-253, His-433, and His-435 residues may determine the shape of the energy landscape for the native complex and play an important role

in kinetic regulation of specific recognition with the peptide. The organization of protein interaction networks may be imprinted in structural properties of the binding sites which enable network hubs such as MDM2 and Ig proteins to ensure similar mechanisms of functional adaptability in binding with conformationally diverse binding partners.

REFERENCES

- [1] MCCAMMON J. A., *Curr. Opin. Struct. Biol.*, **8** (1998) 245.
- [2] CARLSON H. A. and MCCAMMON J. A., *Mol. Pharmacol.*, **57** (2000) 213.
- [3] VAN REGENMORTEL M. H., *J. Mol. Recognit.*, **12** (1999) 1.
- [4] KUMAR S., MA B., TSAI C. J., SINHA N. and NUSSINOV R., *Protein Sci.*, **9** (2000) 10.
- [5] FRAUENFELDER H., *Nat. Struct. Biol.*, **2** (1995) 821.
- [6] LEESON D. T. and WIERSMA D. A., *Nat. Struct. Biol.*, **2** (1995) 848.
- [7] FRAUENFELDER H. and LEESON D. T., *Nat. Struct. Biol.*, **5** (1998) 757.
- [8] REJTO P. A. and FREER S. T., *Prog. Biophys. Mol. Biol.*, **66** (1996) 167.
- [9] TSAI C. J., MA B. and NUSSINOV R., *Proc. Natl. Acad. Sci. U.S.A.*, **96** (1999) 9970.
- [10] SINHA N. and NUSSINOV R., *Proc. Natl. Acad. Sci. U.S.A.*, **98** (2001) 3139.
- [11] KESKIN O., JERNIGAN R. L. and BAHAR I., *Biophys. J.*, **78** (2000) 2093.
- [12] KAY L. E., MUHANDIRAM D. R., WOLF G., SHOELSON S. E. and FORMAN-KAY J. D., *Nat. Struct. Biol.*, **5** (1998) 156.
- [13] FORMAN-KAY J. D., *Nat. Struct. Biol.*, **6** (1999) 1086.
- [14] ZIDEK L., NOVOTNY M. V. and STONE M. J., *Nat. Struct. Biol.*, **6** (1999) 1118.
- [15] CAVANAGH J. and AKKE M., *Nat. Struct. Biol.*, **7** (2000) 11.
- [16] LEE A. L., KINNEAR S. A. and WAND A. J., *Nat. Struct. Biol.*, **7** (2000) 72.
- [17] MIRNY L. and SHAKHNOVICH E., *Annu. Rev. Biophys. Biomol. Struct.*, **30** (2001) 361.
- [18] ONUCHIC J. N. and WOLYNES P. G., *Curr. Opin. Struct. Biol.*, **14** (2004) 70.
- [19] SHAKHNOVICH E., *Chem. Rev.*, **106** (2006) 1559.
- [20] ONUCHIC J. N., NYMEYER H., GARCIA A. E., CHAHINE J. and SOCCI N. D., *Adv. Protein Chem.*, **53** (2000) 87.
- [21] PLOTKIN S. S. and ONUCHIC J. N., *Q. Rev. Biophys.*, **35** (2002) 111.
- [22] PLOTKIN S. S. and ONUCHIC J. N., *Q. Rev. Biophys.*, **35** (2002) 205.
- [23] WOLYNES P. G., *Philos. Transact. A Math. Phys. Eng. Sci.*, **363** (2005) 453.
- [24] VERKHIVKER G. M., BOUZIDA D., GEHLHAAR D. K., REJTO P. A., FREER S. T. and ROSE P. W., *Curr. Opin. Struct. Biol.*, **12** (2002) 197.
- [25] WANG J. and VERKHIVKER G. M., *Phys. Rev. Lett.*, **90** (2003) 188101.
- [26] SHOEMAKER B. A., PORTMAN J. J. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **97** (2000) 8868.
- [27] LEVY Y., WOLYNES P. G. and ONUCHIC J. N., *Proc. Natl. Acad. Sci. U.S.A.*, **101** (2004) 511.
- [28] VERKHIVKER G. M., BOUZIDA D., GEHLHAAR D. K., REJTO P. A., FREER S. T. and ROSE P. W., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 5148.
- [29] TSAI C. J., MA B. and NUSSINOV R., *Proc. Natl. Acad. Sci. U.S.A.*, **96** (1999) 9970.
- [30] TSAI C. J., KUMAR S., MA B. and NUSSINOV R., *Protein Sci.*, **8** (1999) 1181.
- [31] KUMAR S., MA B., TSAI C. J., SINHA N. and NUSSINOV R., *Protein Sci.*, **9** (2000) 10.
- [32] TSAI C. J., MA B., KUMAR S., WOLFSON H. and NUSSINOV R., *Crit. Rev. Biochem. Mol. Biol.*, **36** (2001) 399.
- [33] MA B., SHATSKY M., WOLFSON H. J. and NUSSINOV R., *Protein Sci.*, **11** (2002) 184.
- [34] HOPFIELD J. J., *Proc. Natl. Acad. Sci. U.S.A.*, **71** (1974) 4135.

- [35] GULUKOTA K. and WOLYNES P. G., *Proc. Natl. Acad. Sci. U.S.A.*, **91** (1994) 9292.
- [36] SLUTSKY M. and MIRNY L. A., *Biophys J.*, **87** (2004) 4021.
- [37] JAMES L. C., ROVERSI P. and TAWFIK D. S., *Science*, **299** (2003) 1362.
- [38] JAMES L. C. and TAWFIK D. S., *Trends Biochem. Sci.*, **28** (2003) 361.
- [39] ROODVELDT C., AHARONI A. and TAWFIK D. S., *Curr. Opin. Struct. Biol.*, **15** (2005) 50.
- [40] ATWELL S., ULTSCH M., DE VOS A. M. and WELLS J. A., *Science*, **278** (1997) 1125.
- [41] SUNDBERG E. J. and MARIUZZA R. A., *Struct. Fold. Des.*, **8** (2000) R137.
- [42] DELANO W. L., ULTSCH M. H., DE VOS A. M. and WELLS J. A., *Science*, **287** (2000) 1279.
- [43] CLACKSON T. and WELLS J., *Science*, **267** (1995) 383.
- [44] DELANO W. L., *Curr. Opin. Struct. Biol.*, **12** (2002) 14.
- [45] JONES S. and THORNTON J. M., *Proc. Natl. Acad. Sci. U.S.A.*, **93** (1996) 13.
- [46] BOGAN A. A. and THORN K. S., *J. Mol. Biol.*, **280** (1998) 1.
- [47] LO CONTE L., CHOTHIA C. and JANIN J., *J. Mol. Biol.*, **285** (1999) 2177.
- [48] HU Z., MA B., WOLFSON H. and NUSSINOV R., *Proteins*, **39** (2000) 331.
- [49] MA B., WOLFSON H. J. and NUSSINOV R., *Curr. Opin. Struct. Biol.*, **11** (2001) 364.
- [50] VERKHIVKER G. M., BOUZIDA D., GEHLHAAR D. K., REJTO P. A., FREER S. T. and ROSE P. W., *Proteins*, **48** (2002) 539.
- [51] MOMAND J., ZAMBETTI G. P., OLSON D. C., GEORGE D. and LEVINE A. J., *Cell*, **69** (1992) 1237.
- [52] FINLAY C. A., *Mol. Cell. Biol.*, **13** (1993) 301.
- [53] CHEN J., MARECHAL V. and LEVINE A. J., *Mol. Cell. Biol.*, **13** (1993) 4107.
- [54] CHEN J., WU X., LIN J. and LEVINE A. J., *Mol. Cell. Biol.*, **16** (1996) 2445.
- [55] LEVINE A. J., *Cell*, **88** (1997) 323.
- [56] VOUSDEN K. H. and LU X., *Nat. Rev. Cancer*, **2** (2002) 594.
- [57] KUBBUTAT M. H., JONES S. N. and VOUSDEN K. H., *Nature*, **387** (1997) 299.
- [58] BROOKS C. L. and GU W., *Mol. Cell.*, **21** (2006) 307.
- [59] BOND G. L., HU W. and LEVINE A. J., *Curr. Cancer Drug Targets.*, **5** (2005) 3.
- [60] POYUROVSKY M. V. and PRIVES C., *Genes Dev.*, **20** (2006) 125.
- [61] KUSSIE P. H., GORINA S., MARECHAL V., ELENBAAS B., MOREAU J., LEVINE A. J. and PAVLETICH N. P., *Science*, **274** (1996) 948.
- [62] BOTTGGER A., BOTTGGER V., GARCIA-ECHEVERRIA C., CHENE P., HOCHKEPPEL H. K., SAMPSON W., ANG K., HOWARD S. F., PICKSLEY S. M., and LANE D. P., *J. Mol. Biol.*, **269** (1997) 744.
- [63] BOTTGGER V., BOTTGGER A., HOWARD S. F., PICKSLEY S. M., CHENE P., GARCIA-ECHEVERRIA C., HOCHKEPPEL H. K. and LANE D. P., *Oncogene*, **13** (1996) 2141.
- [64] FREEDMAN D. A., EPSTEIN C. B., ROTH J. C. and LEVINE A. J., *Mol. Med.*, **3** (1997) 248.
- [65] SHIMIZU H., BURCH L. R., SMITH A. J., DORNAN D., WALLACE M., BALL K. L. and HUPP T. R., *J. Biol. Chem.*, **277** (2002) 28446.
- [66] SCHON O., FRIEDLER A., BYCROFT M., FREUND S. M. and FERSHT A. R., *J. Mol. Biol.*, **323** (2002) 491.
- [67] ISSAEVA N., FRIEDLER A., BOZKO P., WIMAN K. G., FERSHT A. R. and SELIVANOVA G., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 13303.
- [68] SCHON O., FRIEDLER A., FREUND S. and FERSHT A. R., *J. Mol. Biol.*, **336** (2004) 197.
- [69] BURCH L., SHIMIZU H., SMITH A., PATTERSON C. and HUPP T. R., *J. Mol. Biol.*, **337** (2004) 129.
- [70] CHI S. W., LEE S. H., KIM D. H., AHN M. J., KIM J. S., WOO J. Y., TORIZAWA T., KAINOSHO M. and HAN K. H., *J. Biol. Chem.*, **280** (2005) 38795.

- [71] SHENG Y., SARIDAKIS V., SARKARI F., DUAN S., WU T., ARROWSMITH C. H. and FRAPPIER L., *Nat. Struct. Mol. Biol.*, **13** (2006) 285.
- [72] UHRINOVA S., UHRIN D., POWERS H., WATT K., ZHELEVA D., FISCHER P., MCINNES C. and BARLOW P. N., *J. Mol. Biol.*, **350** (2005) 587.
- [73] MCCOY M. A., GESELL J. J., SENIOR M. M. and WYSS D. F., *Proc. Natl. Acad. Sci. U.S.A.*, **100** (2003) 1645.
- [74] CHENE P., *Nat. Rev. Cancer.*, **3** (2003) 102.
- [75] KLEIN C. and VASSILEV L. T., *Br. J. Cancer.*, **91** (2004) 1415.
- [76] CHENE P., *Mol. Cancer Res.*, **2** (2004) 20.
- [77] FISCHER P. M. and LANE D. P., *Trends Pharmacol. Sci.*, **25** (2004) 343.
- [78] ZHANG Z. and ZHANG R., *Curr. Cancer Drug Targets.*, **5** (2005) 9.
- [79] BUOLAMWINI J. K., ADDO J., KAMATH S., PATIL S., MASON D. and ORES M., *Curr. Cancer Drug Targets.*, **5** (2005) 57.
- [80] FRY D. C. and VASSILEV L. T., *J. Mol. Med.*, **83** (2005) 955.
- [81] VASSILEV L. T., VU B. T., GRAVES B., CARVAJAL D., PODLASKI F., FILIPOVIC Z., KONG N., KAMMLOTT U., LUKACS C., KLEIN C., FOTOUHI N. and LIU E. A., *Science*, **303** (2004) 844.
- [82] VASSILEV L. T., *Cell Cycle*, **3** (2004) 419.
- [83] TOVAR C., ROSINSKI J., FILIPOVIC Z., HIGGINS B., KOLINSKY K., HILTON H., ZHAO X., VU B. T., QING W., PACKMAN K., MYKLEBOST O., HEIMBROOK D. C. and VASSILEV L. T., *Proc. Natl. Acad. Sci. U.S.A.*, **1033** (2006) 1888.
- [84] GRASBERGER B. L., LU T., SCHUBERT C., PARKS D. J., CARVER T. E., KOBLISH H. K., CUMMINGS M. D., LAFRANCE L. V., MILKIEWICZ K. L., CALVO R. R., MAGUIRE D., LATTANZE J., FRANKS C. F., ZHAO S., RAMACHANDREN K., BYLEBYL G. R., ZHANG M., MANTHEY C. L., PETRELLA E. C., PANTOLIANO M. W., DECKMAN I. C., SPURLINO J. C., MARONEY A. C., TOMCZUK B. E., MOLLOY C. J. and BONE R. F., *J. Med. Chem.*, **48** (2005) 909.
- [85] KOBLISH H. K., ZHAO S., FRANKS C. F., DONATELLI R. R., TOMINOVICH R. M., LAFRANCE L. V., LEONARD K. A., GUSHUE J. M., PARKS D. J., CALVO R. R., MILKIEWICZ K. L., MARUGAN J. J., RABOISSON P., CUMMINGS M. D., GRASBERGER B. L., JOHNSON D. L., LU T., MOLLOY C. J. and MARONEY A. C., *Mol. Cancer Ther.*, **5** (2006) 160.
- [86] DING K., LU Y., NIKOLOVSKA-COLESKA Z., WANG G., QIU S., SHANGARY S., GAO W., QIN D., STUCKEY J., KRAJEWSKI K., ROLLER P. P. and WANG S., *J. Med. Chem.*, **49** (2006) 3432.
- [87] LU Y., NIKOLOVSKA-COLESKA Z., FANG X., GAO W., SHANGARY S., QIU S., QIN D. and WANG S., *J. Med. Chem.*, **49** (2006) 3759.
- [88] MASSOVA I. and KOLLMAN P. A., *J. Am. Chem. Soc.*, **121** (1999) 8133.
- [89] ZHONG H. and CARLSON H. A., *Proteins*, **158** (2005) 222.
- [90] LU F., CHI S. W., KIM D. H., HAN K. H., KUNTZ I. D. and GUY R. K., *J. Comb. Chem.*, **8** (2006) 315.
- [91] SOOD V. D. and BAKER D., *J. Mol. Biol.*, **357** (2006) 917.
- [92] ESPINOZA-FONSECA L. M. and TRUJILLO-FERRARA J. G., *Biopolymers*, **83** (2006) 365.
- [93] BERMAN H. M., WESTBROOK J., FENG Z., GILLILAND G., BHAT T. N., WEISSIG H., SHINDYALOV I. N. and BOURNE P. E., *Nucleic Acids Res.*, **28** (2000) 235.
- [94] LOVELL S. C., WORD J. M., RICHARDSON J. S. and RICHARDSON D. C., *Proteins*, **40** (2000) 389.
- [95] SUGITA Y. and OKAMOTO Y., *Chem. Phys. Lett.*, **314** (1999) 141.

- [96] VERKHIVKER G. M., REJTO P. A., BOUZIDA D., ARTHURS S., COLSON A. B., FREER S. T., GEHLHAAR D. K., LARSON V., LUTY B. A., MARRONE T. and ROSE P. W., *Chem. Phys. Lett.*, **336** (2001) 495.
- [97] MAYO S. L., OLAFSON B. D. and GODDARD W. A. III., *J. Phys. Chem.*, **94** (1990) 8897.
- [98] CORNELL W. D., CIEPLAK P., BAYLY C. L., GOULD I. R., MERZ K. M., FERGUSON D. M., SPELLMEYER D. C., FOX T., CALDWELL J. W. and KOLLMAN P. A., *J. Am. Chem. Soc.*, **117** (1995) 5179.
- [99] STOUTEN P. F. W., FRÖMMELE C., NAKAMURA H. and SANDER C., *Mol. Simul.*, **10** (1993) 97.
- [100] BEUTLER T. C., MARK. A. E., VAN SCHAIK R. C., GERBER P. R. and VAN GUNSTEREN W., *Chem. Phys. Lett.*, **222** (1994) 529.
- [101] MOHAMADI F., RICHARDS N. G. J., GUIDA W. C., LISKAMP R., LIPTON M., CAUFIELD C., CHANG G., HENDRICKSON T. and STILL W. C., *J. Comput. Chem.*, **11** (1990) 440.
- [102] KOLLMAN P. A., MASSOVA I., REYES C., KUHN B., HUO S., CHONG L., LEE M., LEE T., DUAN Y., WANG W., DONINI O., CIEPLAK P., SRINIVASAN J., CASE D. A. and CHEATHAM T. E. III., *Acc. Chem. Res.*, **33** (2000) 889.

This page intentionally left blank

Tetrameric voltage-gated ion channels investigated by molecular dynamics and bioinformatics

S. PANTANO

VIMM Venetian Institute of Molecular Medicine – Via Orus 2, 35129 – Padova, Italy

Institut Pasteur, Montevideo, Mataojo 2020, 11400, Uruguay

M. BERRERA

SISSA, Scuola Internazionale Superiore di Studi Avanzati, IIT, Istituto Italiano di Tecnologia-SISSA Unit, and INFM-DEMOCRITOS – Via Beirut 2, 34014 – Trieste, Italy

VIMM Venetian Institute of Molecular Medicine – Via Orus 2, 35129 – Padova, Italy

C. ANSELMi

SISSA, Scuola Internazionale Superiore di Studi Avanzati, IIT, Istituto Italiano di Tecnologia-SISSA Unit, and INFM-DEMOCRITOS – Via Beirut 2, 34014 – Trieste, Italy

P. CARLONI(*)

SISSA, Scuola Internazionale Superiore di Studi Avanzati, IIT, Istituto Italiano di Tecnologia-SISSA Unit, and INFM-DEMOCRITOS – Via Beirut 2, 34014 – Trieste, Italy

Bioinformatics and MD approaches are highly complementary tools for structural predictions. We have used these two approaches together to investigate structural and functional properties of tetrameric voltage-gated ion channels for which structural information is completely (CNG channels) or partially (HCN channels) lacking. Besides their predicting power, MD simulations have allowed to provide insights on functional

(*) E-mail: carloni@sissa.it

properties of HCN channels. In addition, bioinformatics has allowed extending many of our findings obtained for a protein from one specific organism to the entire class of HCN and CNG channels, and MD to other proteins featuring the same fold but with totally different function.

1. – Introduction

The flux of K^+ , Na^+ and Ca^{2+} ions across the cell membrane plays a key role for several biological processes, including muscle contraction, nervous signal transmission and osmotic pressure. Specific membrane channels govern such flux allowing for fast ion permeation (up to $\sim 10^7$ ions s^{-1}) and they may be highly selective toward specific ions [1]. Ion channels “gate” (*i.e.* open) in response to different conditions, including lowering of pH, ligand binding and change of membrane voltage [2-8]. Those responding to the latter stimulus constitute the important superfamily of the voltage-gated (VG) ion channels [3, 4, 6]. VG ion channels can be either monomeric or tetrameric. The latter, far more structurally characterized than the former, are the focus of the present review. They include K^+ channels (also called K_v channels), the cyclic nucleotide-gated (CNG) and the hyperpolarization-activated cyclic nucleotide-gated (HCN) channels [9]. The transmembrane (TM) region of all monomeric VG channels features a homo- or hetero-tetrameric fold, in which four subunits encircle the ion-permeating pathway [3]. Each subunit features the same 6-TM-helix bundle topology (S1-S6, fig. 1). The pore region is constituted by the stretch of amino acids between S5 and S6 helices, which includes the pore or P-helix. The response to voltage changes is provided by the so-called voltage sensor, namely the S4 helix, which contains repeats of amino acids triplets a^+bc (a^+ = arginine or lysine; b, c = non polar residues). In particular, the motion of the voltage sensor S4 is believed to be coupled to that of S6 in the gating process [10], allowing for pore opening.

In K_v channels, a T1 cytoplasmic domain is attached to the N-term of the TM domain, whilst in the others a cytoplasmic cyclic nucleotide binding domain (CNBD) is linked to the C-term of the TM domain through the C-linker domain [11-14] (fig. 1).

The functional properties of VG tetrameric channels are markedly different. K^+ channels gate upon membrane depolarization, *i.e.* an increase of the membrane potential from its resting level (~ -5 mV). They are selective for K^+ against Na^+ ions. In contrast, CNG channels are permeable by both K^+ and Na^+ ions [9] and they are much less sensitive to changes of voltage (decrease of the membrane potential ~ -50 mV to ~ -60 mV). Probably for this reason, their opening is ruled by the binding of cyclic nucleotides ligands (cAMP or cGMP, see fig. 1D, E) at μ M concentration to the cytoplasmic side. Finally, HCN channels gate upon membrane hyperpolarization (decrease of the membrane potential ~ -50 mV to ~ -100 mV) and their gating is only modulated by the presence of cyclic nucleotides, which bind to a CNBD with the same topology of those of CNG channels [15]. These channels are permeable by K^+ and, to a less extent, by Na^+ ions.

The recent X-ray structure of a VG K^+ channel from rat [14] has provided the molecular basis for the function of these channels at the molecular level. Instead, X-ray

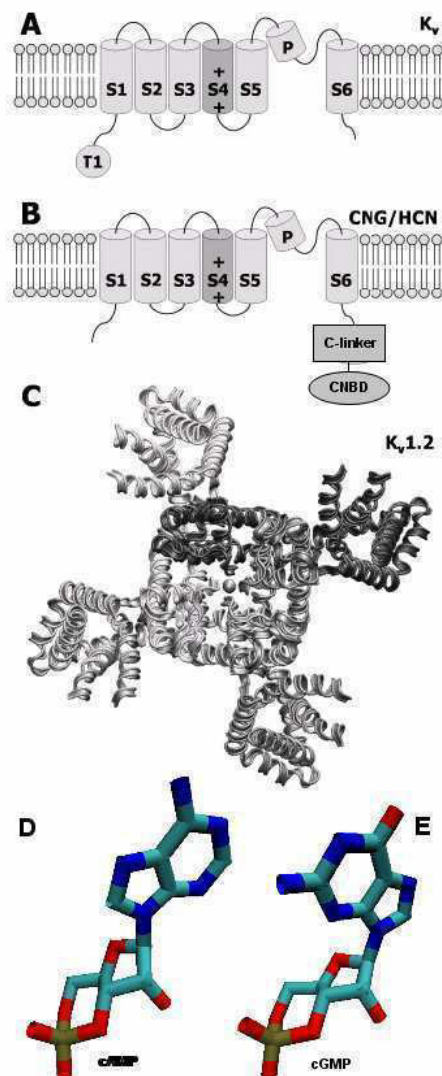


Fig. 1. – Topology of VG tetrameric ion channels subunits: Cartoon representation of K_v (A) and CNG/HCN (B) channels. (C) X-ray structure of the K^+ VG tetrameric channel [14]. The four subunits are differently colored and the tetramer is drawn along the ion permeating pathway. (D-E) Structures of cyclic adenosine monophosphate (cAMP) and cyclic guanosine monophosphate (cGMP).

structural information is presently not available for CNG channels. For HCN channels, the X-ray structure of the cytoplasmic domain is available [16]. In an attempt at providing insights on the intriguingly different functional properties among these channels, we have recently used computational tools (bioinformatics and molecular simulations) to investigate structure and dynamics of both the TM and cytoplasmic domains of these proteins. Some of our results are summarized in this review.

2. – Methods

Two computational approaches have been used here. The first uses structural bioinformatics tools, specifically: i) secondary structure predictions based on the primary sequence. They are based on statistical and physico-chemical methods, sequence pattern matching and evolutionary conservation [17]. ii) Homology modeling (HM), which constructs structural models (targets) that are homologous to other protein(s), whose 3D structure is known (templates). HM is based on the larger conservation of protein folds than primary sequences during evolution [18,19]. In fact, since the number of stable folds is limited by their thermodynamics stability [20], tertiary structures turn out to be more conserved in evolution than the amino acid sequences [21]. The similarity between amino acids, which provides the plausibility of a particular substitution, is defined mathematically according to specific criteria [22].

The second approach is based on classical molecular dynamics (MD) simulations, which explore the conformational space of biomolecules given an empirically derived interatomic potential energy function, by providing information on the motion of the atoms as a function of time. MD may be used to calculate thermodynamics parameters and to provide information on protein dynamics [23-25].

3. – Results

X-ray structures of *CNG channels* are so far lacking. We built structural models of CNG channels from bovine rod based on HM, with the aid of experimental constraints [26]. For the pore region, we used the KcsA K^+ channel from *Streptomyces lividans* [27], whose pore region has been suggested to share the same topology with CNG ion channels [28]. However, the sequence identity (not larger than 20%) between templates and target structure is very low. Thus, a limited reliability of the resulting HM structures is expected and a refinement of the structure was required. To this aim, as many as 50 structural constraints based on electrophysiological measurements of cysteine mutants were included [26]. Based on the resulting theoretical models of the open and closed states (fig. 2), we suggested that CNG channels open by a clockwise rotation of about 45° around the channel axis of N-terminal section at the cytoplasmic, cyclic nucleotide binding domain. The rotation is transmitted upwards on S6, which rotates by about 30° around the helix axis. S6 also bends around a hinge located nearly on Gly39. Because S6 interacts directly with the P-helix, this motion is transmitted to the pore helix, which rearranges its terminal residues. As a result, the lower part of the wall

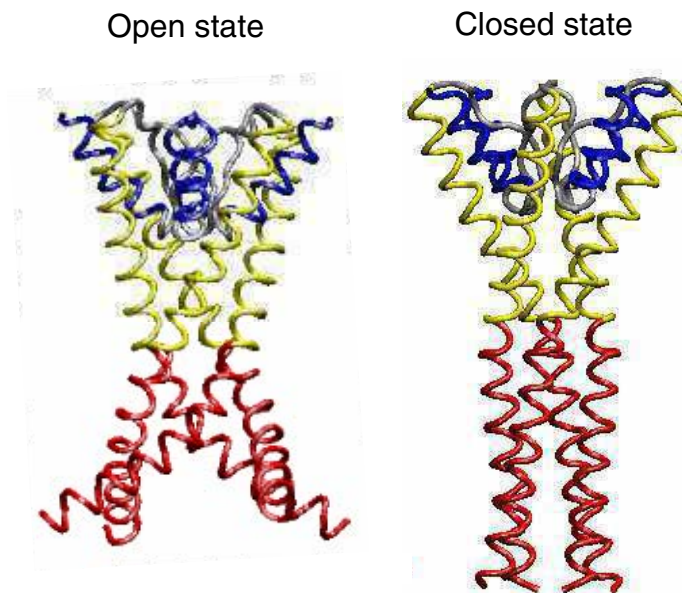


Fig. 2. – Cartoon representing the trace of the C α atoms of the tetrameric CNG channel in the open (left) and closed (right) conformations [26]. The P-helix, the filter region, the S6-helix and the N-terminal part of C-linker (A'-helix, at the bottom of the picture) are colored in blue, grey, yellow and red, respectively.

pore increases its diameter from $\sim 3 \text{ \AA}$ (as in K^+ channels) to $\sim 6 \text{ \AA}$. This finding is consistent with the absence of selectivity in these channels, as this pore is expected to allow permeation of both Na^+ and K^+ ions.

Bioinformatics provided insights also on the voltage sensor (S4 in fig. 1B). Indeed, multiple sequence alignment suggested that the S4 domain has a lower net positive charge with respect to K^+ channels. Second, predictions on the secondary structures suggest that the S4-S5 linker is shorter and probably forms a loop. Therefore, the motion of the S4 domain is not efficiently coupled to the S6 domain as in K^+ channels. Finally, the end portion of S6 is constrained by a large cytoplasmic domain (fig. 1). All these features may play a role for the fact that CNG channels are less sensitive to membrane voltage changes (fig. 3) [29].

Next, we have investigated structural features of the CNBD by HM and MD [30]. The model was based on the structure of the fairly homologous catabolite gene activator protein [31]. Again, owing to the low identity level obtained in the sequence alignment, data from site directed mutagenesis experiments on both, target and template was included to guide the construction of the structural model. The MD calculations provided a structural basis for the experimentally measured difference in activity between cAMP and cGMP. In addition, they provided support for the rearrangement of the terminal C-helix of the CNBD upon ligand binding and releasing proposed by Matulef *et al.* [7].

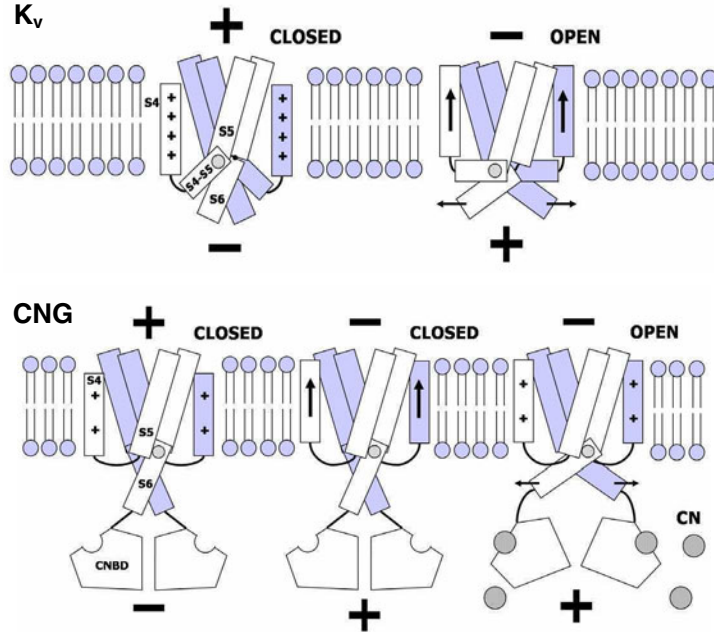


Fig. 3. – Response to membrane potential changes in VG K⁺ (top) and CNG channels (bottom). The cartoon shows that the charge of S4 is larger in K⁺ than in CNG channel and that the S4-S5 linker is α -helical in the first and a coil in the latter.

However, it should be mentioned that the recent X-ray structure of the same domain of HCN channels from [16] (see next paragraph), published after our work, call for further modeling based on this structure, which has a fold different from that of the template used by us. In addition, experiments in progress suggest that part of the model, in particular the C-linker and the CNBD, may be different from what emerges from our previous study (Torre *et al.*, personal communication).

The structure of *HCN channels* TM domain is so far lacking. Models of the pore region from mouse and sea urchin sperm were obtained by HM and MD calculations [32], making two major assumptions. i) In the closed state, the inner pore is assumed to be similar to that of all K⁺ channels for which an X-ray structure is available (VG tetrameric K⁺ channel KvAP from *Aeropyrum pernix* [13], the KcsA K⁺ channel from *Streptomyces lividans* [33] and the inward rectifier K⁺ channel KirBac1.1 from *Burkholderia pseudomallei* [34]). This assumption is fully justified by electrophysiological measurements. Thus, we can expect that the HM models, based on the three X-ray structures, are more reliable than those of CNG channels, for which this assumption could not be made. ii) In the open state, the S6 helix is bent further than it is in the closed state, as suggested (but not proven) by experimental data [32]. We used as template the X-ray structure of the MthK channel [27]. The structural models of the closed state turned out to be

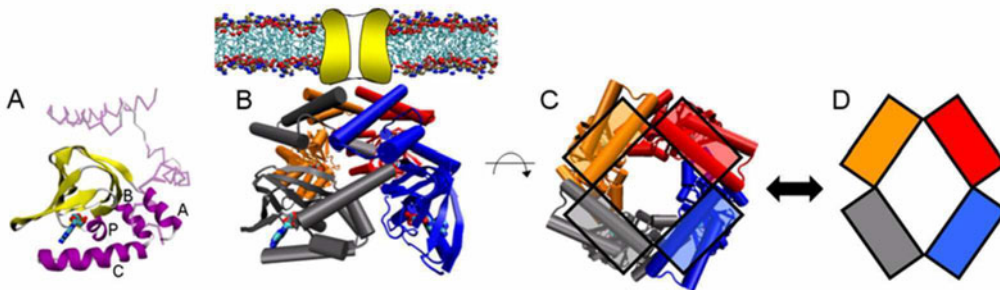


Fig. 4. – Structure of C-linker and CNBD of HCN2 from mouse [16]. (A) The crystallographic asymmetric units of the homotetramer are shown: α -helices are colored in violet and β -strands in yellow; a cAMP molecule is represented in sticks within its binding site and colored by atom type. The C-linker domain, drawn in thin trace, is at the top of the picture. The α -helices of the CNBD are labeled. (B) The four subunits of the homotetramer are differently colored and the position of the membrane is indicated. (C) The homotetramer is viewed from the membrane side and the colored rectangular shadows indicate the initial positions of the four subunits. (D) Diagram of the extreme orientation of the four subunits in the quaternary structure oscillation identified by MD simulations [36].

consistent with all the available electrophysiological data as well as the model of the open state in the filter region. Nevertheless the latter required the introduction of an appropriate, experimentally derived, constraint for the S6 helix [32, 35]. Our modeling suggested that the reduced ionic selectivity of HCN channels, relative to that of other channels, may be caused, at least in part, by the larger flexibility of the inner pore of HCN channels relative to that of K^+ channels.

As for S4, the presence of extensive repeats of leucines and other amino acids with similar packing properties suggested that gating occurs through a coordinated motion of S4, which acts on the S6 and, probably, the S5 domains. The whole process is expected to be slower than that of VG tetrameric K^+ channels, in which the S4-S6 helix are expected to move more freely because of the absence of the CNBD (fig. 1). This finding is fully consistent with the experimentally observed faster gating of K^+ channels relative to HCN channels [29].

Our investigation of the cytoplasmic domain has profited by the recently determined structure of the CNBD and the C-linker domain from mouse [16]. Four subunits are arranged around a four-fold symmetry axis, which is virtually perpendicular to the membrane plane. The C-linker is composed by six α -helices (A'-F'), establishing a large amount of intersubunit contacts and coupling the gating of the channel from the TM part to the CNBD. The CNBD is made up of four α -helices (A, P, B, C) with a β -roll between the A- and B-helices. The β -roll comprises eight β -strands in a jelly-roll-like topology, which together with the P and C-helices, define the cAMP binding site (fig. 4A). The structure was obtained in the ligand bound state. MD simulations [36] revealed that, in the absence of cAMP, the cytoplasmic domain of HCN2 remains in a flexible state

where the helical content of the CNBD is reduced, although no dramatic unfolding events are observed. On the other hand, simulations in the presence of cAMP highlighted large scale oscillations involving the whole holo-tetramer (fig. 4C, D) that putatively facilitate the opening of the channel, because they comport a weakening of the intersubunit interactions. Notably, these oscillations, which result from the rigidity of the CNBDs when bound to cAMP, were not detectable in the X-ray structure since the asymmetric unit included only one monomer of the tetramer and they were filtered out during the refinement process. Given the high sequence and structural conservation of CNBDs [37] it is expected that these modules share also a conserved allosteric mechanism. In fact, a very similar dynamical behavior has been retrieved in the homologous CNBD of the isoform II β of the regulatory subunit of protein kinase A (PKA) [38]. Although in the PKA context, the binding of cAMP drives the regulatory subunit to a more rigid state that allows for the releasing and activation of the catalytic subunit. Thus, molecular simulations allowed delineating a conserved allosteric mechanism able to generate completely different effects according to the quaternary environment in which CNBDs are inserted. This constitutes a nice example of how nature may reach a high degree of structural and functional diversity by combining the same protein modules in different arrangements.

4. – Concluding remarks

Structural bioinformatics and biomolecular simulations are very different computational approaches. The first uses biological concepts such as Darwinian evolution, whilst the other is based on the laws of physics; indeed the two techniques have been most often used separately by different communities to investigate protein structure and function. Yet, their combined approach turns out to be very powerful, as the information that can be obtained by each technique may be complementary to the other. In this review, we have shown that MD and bioinformatics provides complementary information of tetrameric VG channels for which structural information is lacking. We expect that this type of approach will become of increasing importance to elucidate the structure and function of ion channels, for which the structural information is still very limited.

REFERENCES

- [1] CORRY B. and CHUNG S. H., *Mechanisms of valence selectivity in biological ion channels*, *Cell. Mol. Life Sci.*, **63** (2006) 301.
- [2] CATTERALL W. A., *From ionic currents to molecular mechanisms: the structure and function of voltage-gated sodium channels*, *Neuron*, **26** (2000) 13.
- [3] SANDS Z., GROTTESI A. and SANSOM M. S., *Voltage-gated ion channels*, *Curr. Biol.*, **15** (2005) R44.
- [4] BEZANILLA F., *Voltage-gated ion channels*, *IEEE Trans. Nanobioscience*, **4** (2005) 34.
- [5] CHANG G., SPENCER R. H., LEE A. T., BARCLAY M. T. and REES D. C., *Structure of the MscL homolog from Mycobacterium tuberculosis: a gated mechanosensitive ion channel*, *Science*, **282** (1998) 2220.

- [6] BRADLEY J., REISERT J. and FRINGS S., *Regulation of cyclic nucleotide-gated channels*, *Curr. Opin. Neurobiol.*, **15** (2005) 343.
- [7] MATULEF K., FLYNN G. E. and ZAGOTTA W. N., *Molecular rearrangements in the ligand-binding domain of cyclic nucleotide-gated channels*, *Neuron*, **24** (1999) 443.
- [8] CUELLO L. G., ROMERO J. G., CORTES D. M. and PEROZO E., *pH-dependent gating in the *Streptomyces lividans* K⁺ channel*, *Biochemistry*, **37** (1998) 3229.
- [9] KAUPP U. B. and SEIFERT R., *Cyclic nucleotide-gated ion channels*, *Physiol. Rev.*, **82** (2002) 769.
- [10] SANSOM M. S., *Potassium channels: watching a voltage-sensor tilt and twist*, *Curr. Biol.*, **10** (2000) R206.
- [11] MANNUZZU L. M., MORONNE M. M. and ISACOFF E. Y., *Direct physical measure of conformational rearrangement underlying potassium channel gating*, *Science*, **271** (1996) 213.
- [12] MANNIKKO R., ELINDER F. and LARSSON H. P., *Voltage-sensing mechanism is conserved among ion channels gated by opposite voltages*, *Nature*, **419** (2002) 837.
- [13] JIANG Y. *et al.*, *X-ray structure of a voltage-dependent K⁺ channel*, *Nature*, **423** (2003) 33.
- [14] LONG S. B., CAMPBELL E. B. and MACKINNON R., *Crystal structure of a mammalian voltage-dependent Shaker family K⁺ channel*, *Science*, **309** (2005) 897.
- [15] CRAVEN K. B. and ZAGOTTA W. N., *CNG and HCN channels: two peas, one pod*, *Annu. Rev. Physiol.*, **68** (2006) 375.
- [16] ZAGOTTA W. N., OLIVIER N. B., BLACK K. D., YOUNG E. C., OLSON R. and GOUAUX E., *Structural basis for modulation and agonist specificity of HCN pacemaker channels*, *Nature*, **425** (2003) 200.
- [17] ROST B., *Review: protein secondary structure prediction continues to rise*, *J. Struct. Biol.*, **134** (2001) 204.
- [18] LECOMTE J. T., VULETICH D. A. and LESK A. M., *Structural divergence and distant relationships in proteins: evolution of the globins*, *Curr. Opin. Struct. Biol.*, **15** (2005) 290.
- [19] HILL E. E., MOREA V. and CHOTHIA C., *Sequence conservation in families whose members have little or no sequence similarity: the four-helical cytokines and cytochromes*, *J. Mol. Biol.*, **322** (2002) 205.
- [20] FINKELSTEIN A. V., *Cunning simplicity of a hierarchical folding*, *J. Biomol. Struct. Dyn.*, **20** (2002) 311.
- [21] RICHARDSON J. S., *beta-Sheet topology and the relatedness of proteins*, *Nature*, **268** (1977) 495.
- [22] BARKER W. C., KETCHAM L. K. and DAYHOFF M. O., *A comprehensive examination of protein sequences for evidence of internal gene duplication*, *J. Mol. Evol.*, **10** (1978) 265.
- [23] SIMONSON T., ARCHONTIS G. and KARPLUS M., *Free energy simulations come of age: protein-ligand recognition*, *Acc. Chem. Res.*, **35** (2002) 430.
- [24] KARPLUS M. and MCCAMMON J. A., *Molecular dynamics simulations of biomolecules*, *Nat. Struct. Biol.*, **9** (2002) 646.
- [25] DOLENC J., OOSTENBRINK C., KOLLER J. and VAN GUNSTEREN W. F., *Molecular dynamics simulations and free energy calculations of netropsin and distamycin binding to an AAAAAA DNA binding site*, *Nucleic Acids Res.*, **33** (2005) 725.
- [26] GIORGETTI A., NAIR A. V., CODEGA P., TORRE V. and CARLONI P., *Structural basis of gating of CNG channels*, *FEBS Lett.*, **579** (2005) 1968.
- [27] JIANG Y., LEE A., CHEN J., CADENE M., CHAIT B. T. and MACKINNON R., *Crystal structure and mechanism of a calcium-gated potassium channel*, *Nature*, **417** (2002) 515.

- [28] BECCHETTI A., GAMEL K. and TORRE V., *Cyclic nucleotide-gated channels. Pore topology studied through the accessibility of reporter cysteines*, *J. Gen. Physiol.*, **114** (1999) 377.
- [29] ANSELMi C., CARLONI P. and TORRE V., *Origin of functional diversity among tetrameric voltage-gated channels*, *Proteins*, **66** (2007) 136.
- [30] PUNTA M., CAVALLI A., TORRE V. and CARLONI P., *Molecular modeling studies on CNG channel from bovine retinal rod: a structural model of the cyclic nucleotide-binding domain*, *Proteins*, **52** (2003) 332.
- [31] WEBER I. T. and STEITZ T. A., *Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5 Å resolution*, *J. Mol. Biol.*, **198** (1987) 311.
- [32] GIORGETTI A., CARLONI P., MISTRIK P. and TORRE V., *A homology model of the pore region of HCN channels*, *Biophys. J.*, **89** (2005) 932.
- [33] DOYLE D. A. et al., *The structure of the potassium channel: molecular basis of K⁺ conduction and selectivity*, *Science*, **280** (1998) 69.
- [34] KUO A. et al., *Crystal structure of the potassium channel KirBac1.1 in the closed state*, *Science*, **300** (2003) 1922.
- [35] SHIN K. S., ROTHBERG B. S. and YELLEN G., *Blocker state dependence and trapping in hyperpolarization-activated cation channels: evidence for an intracellular activation gate*, *J. Gen. Physiol.*, **117** (2001) 91.
- [36] BERRERA M., PANTANO S. and CARLONI P., *cAMP Modulation of the cytoplasmic domain in the HCN2 channel investigated by molecular simulations*, *Biophys. J.*, **90** (2006) 3428.
- [37] BERMAN H. M., TEN EYCK L. F., GOODSSELL D. S., HASTE N. M., KORNEV A. and TAYLOR S. S., *The cAMP binding domain: an ancient signaling module*, *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 45.
- [38] PANTANO S., ZACCOLO M. and CARLONI P., *Molecular basis of the allosteric mechanism of cAMP in the regulatory PKA subunit*, *FEBS Lett.*, **579** (2005) 2679.

Folding inhibition of HIV Protease: An experimental glance

D. PROVASI and G. TIANA

Dipartimento di Fisica, Università di Milano - Milano, Italy
INFN, Sezione di Milano - via C. Celoria 16, 20133 Milano, Italy

R. A. BROGLIA

Dipartimento di Fisica, Università di Milano - Milano, Italy
INFN, Sezione di Milano - via C. Celoria 16, 20133 Milano, Italy
The Niels Bohr Institut, University of Copenhagen
Blegdamsvej 17, DK 2100 Copenhagen Ø, Denmark

1. – Introduction

Recently, a novel strategy to inhibit the action of pathogenic proteins has been identified on the basis of theoretical and computational arguments (see [1-5], as well as the contributions by R. Broglia *et al.* and by G. Tiana *et al.* in this volume). The strategy is grounded on the hierarchical picture of the folding process, a picture that has received strong and independent corroboration from different experimental techniques. In such hierarchy, local structured parts of the backbone are formed first, and then associate by means of a highly specific interaction, committing the protein to the basin of the native state. These local structures are ideal targets for the inhibition: a molecule that competes with the association of two such structures will sterically block the formation of the post-critical folding nucleus and eventually prevent the protein from reaching the native state. Moreover, a peptide featuring the same sequence of one local structure will likely be a good candidate for such a competitor, thereby providing a relatively general design strategy [6].

However, the computational analysis of such a complex problem can be followed on the appropriate time scale (10^{-3} seconds) only using simplified models, since modern realistic all-atom simulations can only cover some 10^{-9} seconds. Although the picture looks reasonable, its actual feasibility must thus be addressed experimentally. In particular, evidence should be presented that the proposed peptides indeed inhibit the catalytic action of the enzyme by binding to the corresponding local structures and disrupting sufficiently the fold. Once these primary results are achieved, other aspects of the method should be considered, in particular the specificity of the inhibition and its performance on drug-resistant strains. Moreover, before the strategy can be applied to cells, one should analyze several technical facets, such as the optimization of the solubility, the lack of toxicity and the delivery of the peptide to the interior of the membrane.

In this work we address only a part of these questions, presenting the results of assays of some of the peptides designed for the inhibition of HIV-1-Protease and an overall structural information from circular dichroism that is not inconsistent with the picture presented in [3]. We report here the results (published in [5]) on the peptide I₈₃₋₉₃ obtained from the 83-93 fragment of the enzyme, along with results for the shortened peptide I₈₃₋₉₂ and for the mutated N88D I₈₃₋₉₃ peptide.

Inhibition assays consist in a thorough characterization of the kinetic aspects of the reaction catalyzed by the enzyme and of its slow-down in presence of an inhibitor (the framework employed is described in subsects. 1'1 and 1'2). The kinetic footprint of the reaction gives only limited information about the actual molecular mechanism at work. To assess this question we performed structural analysis of the protein in the presence of the inhibitor, by means of circular dichroism spectroscopy (briefly illustrated in subsect. 1'3). The technical details of the experiments carried out are described in sect. 2, and the discussion of the results in sect. 3.

1'1. *The Michaelis-Menten framework.* – An enzymatic reaction is described by a set of nonlinear differential equations for the concentrations of the reagents, intermediate and final products. Within the Michaelis-Menten framework, the kinetics of the catalytic reaction is divided into two processes [7]. The enzyme and the substrate combine rapidly and reversibly to give a complex held together by noncovalent interactions; the chemical processes that occur afterwards are described as first-order processes, *i.e.* the rate of product formation is proportional to the amount of complex present in the sample.

Under these assumptions, the reaction dynamics is characterized by two parameters. The rate (turnover number) k_{cat} represents the number of substrate molecules converted to products per active site per unit time, or the number of times the enzyme “turns over” per unit time. The equilibrium concentration K_M is an apparent dissociation constant that may be treated as the overall dissociation constant of all enzyme-bound species.

The concentration of substrate in terms of the kinetic parameters can be obtained analytically (see, for instance [8]),

$$(1) \quad \frac{x_S(t)}{K_M} = W \left(\frac{x_S(0)}{K_M} \exp \left[-t \frac{v_M}{K_M} + \frac{x_S(0)}{K_M} \right] \right),$$

where W is the Lambert W -function (the inverse of $W \mapsto We^W$), and $v_M = k_{\text{cat}}x_E$ is referred to as the maximum velocity. Although the solution can be obtained in close form, it is customary and practical to use the initial speed of the reaction only. From expression (1) taking the derivative, we obtain

$$(2) \quad x'_S(0) = \frac{v_M x_S(0)}{x_S(0) + K_M},$$

so that the values of K_M and v_M can be obtained by fitting the values of $x'_S(0)$ measured for different values of the initial substrate concentration $x_S(0)$ ⁽¹⁾.

In presence of an inhibitor, different cases are possible, depending on whether the enzyme, the inhibitor and the substrate can form a ternary complex and the relative catalytic rate of the enzyme alone and complexed with the inhibitor. If the inhibitor I binds to the enzyme and prevents S from binding and vice versa, *i.e.* if I and S compete, we speak of *competitive inhibition*. Considering an additional equilibrium it is easy to see that in this case eq. (2) still holds, but now

$$(3) \quad K_M \rightarrow K_M \left(1 + \frac{x_I(0)}{K_I} \right),$$

where K_I measures the dissociation of the enzyme-inhibitor complex. Different inhibition kinetics occur if the inhibitor and the substrate bind simultaneously to the enzyme; in those cases, eq. (2) but both v_M and K_M are modified and depend on x_I .

Thus, by performing measures at different inhibition concentration x_I and at different substrate concentration x_S one can discriminate the various inhibition scenarios, and calculate K_I .

1.2. Spectrophotometric assay. – To test the *in vitro* activity of the protease, we employ a well established chromogenic substrate. This peptide is a 12-mer obtained by engineering a natural cleavage sequence of the protease; the central phenylalanine is modified into NO₂-phenylalanine, so that cleavage of the peptidic bond results in a shift of the absorbance peak from 280 to 256 nm, and the molar extinction coefficient at fixed wavelength changes. Letting $x_S(t)$ be the concentration of uncleaved substrate at time t , the absorbance of the solution is therefore given by

$$(4) \quad A(t, \hbar\omega) = \varepsilon_S(\hbar\omega)x_S(t) + \varepsilon_P(\hbar\omega)(x_S(0) - x_S(t)) + A_0(\hbar\omega),$$

⁽¹⁾ For the sake of clarity, it is also customary to turn (2) into a linear relationship by writing it as $(x'_S(0))^{-1} = v_M^{-1} + (x_S(0))^{-1}v_M^{-1}K_M$. In this way the saturation speed can be read as the intercept of the double reciprocal chart (usually called Lineweaver-Burk plot), and the dissociation concentration K_M as the point where the regression line crosses the abscissa. The distortion introduced by the non-linear transformation, however, prevents us from the usual assumptions of normality for the errors, thus making interval estimation and hypothesis testing unfeasible. We shall therefore present the results as Lineweaver-Burk plots, but calculate the estimates and confidence intervals using non-linear regression algorithms.

where $A_0(\hbar\omega)$ is the absorbance due to the molecules in the sample whose concentration remains fixed with time. Since

$$(5) \quad A'(t, \hbar\omega) = (\varepsilon_S(\hbar\omega) - \varepsilon_P(\hbar\omega))x'_S(t),$$

the speed of the reaction can be measured by measuring the variation in time of the absorbance, at a given value of $\hbar\omega$. The energy for which $\varepsilon_S - \varepsilon_P$ is larger is $\hbar\omega \simeq 4 \text{ eV}$, corresponding to a wavelength $\lambda \simeq 310 \text{ nm}$ where $\varepsilon_S - \varepsilon_P \simeq 500 \pm 90 (\text{M cm})^{-1}$.

1.3. Circular Dichroism. – Circular Dichroism (CD) relies on the difference in absorption of left and right circularly polarized light by chromophores which either possess intrinsic chirality or are placed in chiral environments. In the far UV region (5 eV–7 eV), corresponding to peptide bond absorption, the CD spectrum can be analyzed to give the overall fraction of regular secondary structural features such as α -helix and β -sheet.

If one transmits circularly polarised light through the sample and the left and right circularly polarised components are absorbed to the same extent, combination of the components would regenerate radiation polarised in the original plane. However, if one of the components is absorbed by the sample to a greater extent than the other, the resultant radiation (combined components) would now be elliptically polarised, *i.* Difference in absorbance of the two components ($\Delta A = A_L - A_R$) can be translated in ellipticity measured in degrees $\theta = \arctan(b/a)$, where b and a are the minor and major axes of the resultant ellipse.

The amount of protein required for CD measurements can be gauged from the need to keep the absorbance less than about one unit. Typical cell path-lengths for far-UV CD work are in the range 0.01 to 0.05 cm and protein concentrations are in the range 0.2 to 1 mg/ml. Depending on the design of the cell being used, the volume of sample required can range from about 1 ml to as little as 50 μl . Thus the amount of protein required to obtain an acceptable far-UV CD spectrum can be as little as 10 μg , but generally a 100 to 500 μg sample is required in order to explore the optimal conditions for recording spectra.

The main limitation of CD is that it only provides relatively low-resolution structural information. Thus, although far-UV CD can give reasonably reliable estimates of the secondary structure content of a protein (in terms of the proportions of α -helix, β -sheet and β -turns), it must be noted that these are overall figures and do not indicate which regions of the protein belong to a specific structural type.

CD has proved to be an ideal technique to monitor conformational changes in proteins which can occur as a result of changes in experimental parameters such as pH, temperature, binding of ligands etc. During the unfolding of the protein, the extent of the conformational changes can be analysed to provide quantitative estimates of the stability of the folded state of the protein; CD has been extensively used for this purpose.

2. – Materials and methods

Recombinant HIV-1 Protease, expressed in *E. Coli* was purchased from Bachem UK, Ltd. and contained five mutations to restrict autoproteolysis (Q7K, L33I, L36I) and to restrict cysteine thiol oxidation (C67A and C95A). The enzyme was stored at (-70°C) as a solution with concentration 0.1 mg/ml in dilute HCl, ($p\text{H} = 1.6$).

The chromogenic substrate described in subsect. 1.2 (HIV Protease Substrate III, Bachem UK Ltd.) with sequence H-His-Lys-Ala-Arg-Val-Leu-Phe(NO_2)-Phe-Glu-Ala-Nle-Ser- NH_2 was obtained as a 1 mg desiccate, diluted with 0.1 ml of DMSO, and stored at 20°C . The assay buffer was prepared, following ref. [9], by adding 0.8 mM NaCl, 1 mM EDTA and 1 mM dithiothreitol to a 20 mM phosphate buffer ($p\text{H} = 6$).

All peptides used in this study were synthesized by Fmoc solid phase peptide synthesis with acetyl and amide as terminal protection groups. Inhibitor peptides used were estimated to be $\geq 95\%$ pure by analytical HPLC after purification. It is to be noted that one of the control peptides is rather hydrophobic and only $\geq 70\%$ purity could be achieved.

Saquinavir (CAS number: 127779-20-8), a conventional potent protease inhibitor, has been obtained as a $100\text{ }\mu\text{M}$ solution and then diluted in buffer to achieve the needed concentrations.

For all the tests with peptides we ascribe to the following procedure: after that 1 mg of inhibitor peptide was dissolved in 100 ml of DMSO, 4 ml of this solution were then diluted with 16 ml of DMSO and 180 ml of the buffer used for assay. The obtained solution (150 mM of peptide I) was used for the experiments.

Ultraviolet CD spectra were recorded on a Jasco J-810 spectropolarimeter in nitrogen atmosphere at room temperature using 0.1 cm path-length quartz cell. The protein and the peptide were dissolved in a 20 mM phosphate buffer with 0.8 M NaCl at the same concentration used for the activity assays. The CD spectra were analyzed in terms of contribution of secondary structure elements.

3. – Discussion

As reported in [3], careful inspection of different pieces of information point to the 83-93 fragment of the protase (NIIGRNLLTQI) as a possible inhibitor. We thus start by describing the results obtained for the wild-type peptide I_{83-93} . The inhibitory activity of such peptide can be obtained by a spectrophotometric assay, assessing the enzymatic kinetics for different concentrations of the inhibitor $x_I = 0, 3, 10$ and $20\text{ }\mu\text{M}$. For each x_I , we ran the assay for several substrate concentrations ($100\text{ }\mu\text{M} \leq x_S \leq 600\text{ }\mu\text{M}$), as described in sect. 2. The results are presented in fig. 1 and in table I. From such data, an inhibition constant of $K_I = 2.58 \pm 0.78\text{ }\mu\text{M}$ published in [5] can be obtained, that confirms that the selected peptide indeed reduces the catalytic activity of the enzyme. Several other peptides of the same length (from the protease or not) have been checked not to give tangible effects, see [5] for details.

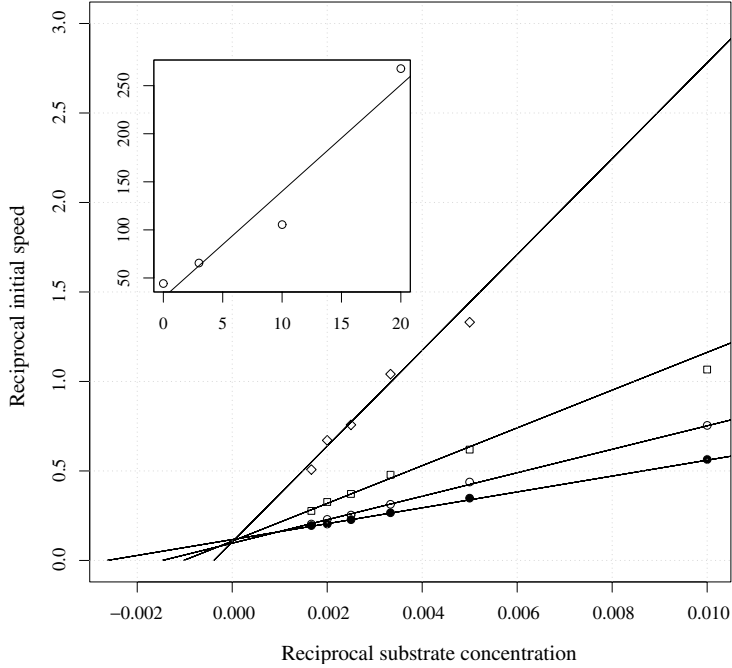


Fig. 1. – Double reciprocal plot of the initial speed of the HIV-1-Protease reaction for different values of substrate and inhibitor concentrations. In the inset, the values of the apparent $K_M v_M^{-1}$ for different inhibitor concentrations. Solid lines are obtained by a non-linear fit of the Michaelis-Menten equation.

In order to improve the solubility of the peptide, we tried two different routes. The most natural is to shorten the peptide, eliminating one or more hydrophobic residues from its terminals. We choose peptide I_{83-92} , obtained cutting the last isoleucine from peptide I_{83-93} . Another possible approach is to scan the many strains of HIV and introduce some of the mutations observed. One reasonable choice is to mutate the asparagine at site 88 with an aspartic acid, a frequently reported mutation. Both peptides show better

TABLE I. – Kinetic data for the HIV-1-PR hydrolyzation of the substrate without and with different concentrations x_I of inhibitor I_{83-93} . The values of K_M for the inhibited reactions have to be regarded as apparent dissociation constants. The values for v_M are reported as measured (mAbs/min) as well as converted using the differential extinction coefficient.

	K_M (μM)	v_M (mAbs/min)	v_M ($\mu\text{mol/s}$)
$x_I = 0 \mu\text{M}$	380 ± 80	8.57 ± 0.88	0.94 ± 0.19
$x_I = 3 \mu\text{M}$	680 ± 92	10.37 ± 0.8	1.14 ± 0.09
$x_I = 10 \mu\text{M}$	980 ± 290	9.29 ± 1.9	1.02 ± 0.27
$x_I = 20 \mu\text{M}$	2600 ± 2000	9.71 ± 11	1.07 ± 1.2

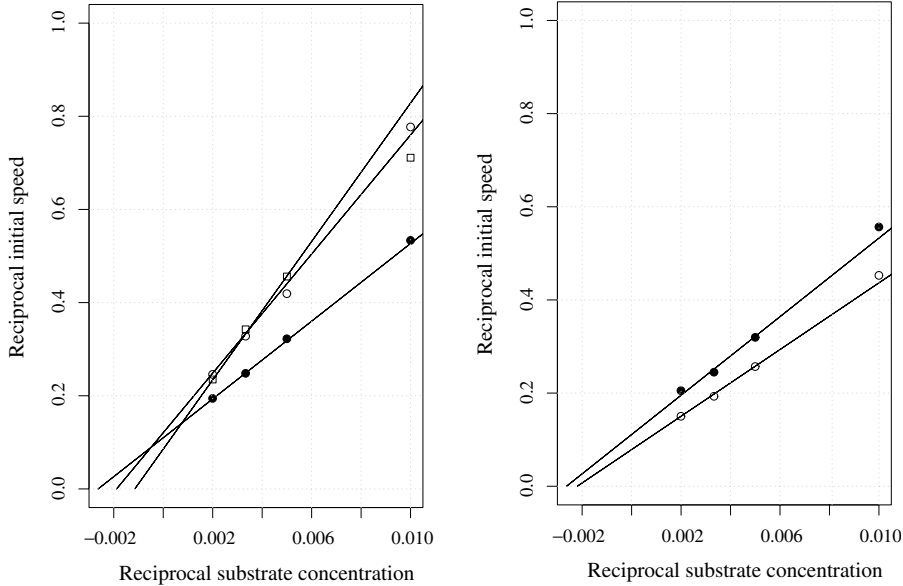


Fig. 2. – Double reciprocal plot of the initial speed of the HIV-1-Protease reaction for different values of substrate and inhibitor concentrations. In the left panel we report data for the shortened peptide I₈₃₋₉₂ and in the right one for the N88D mutated peptide.

solubility than the full-length wild-type (their respective solubility threshold is 40% and 90% higher than I₈₃₋₉₃). Inhibition data are presented in fig. 2, and in table II. The data is still consistent with a competitive inhibition, this time with higher inhibition constants: respectively $K_I \simeq 7.8 \pm 1.2 \mu\text{M}$ and $K_I \simeq 15.8 \pm 1.9 \mu\text{M}$.

Such findings, however, do not provide much information on the inhibition mechanism. Matter-of-factly, nothing in our results prove that the protein is not working because it

TABLE II. – Kinetic data for the HIV-1-PR hydrolyzation of the substrate without and with different concentrations x_I of the mutated inhibitor. The values of K_M for the inhibited reactions have to be regarded as apparent dissociation constants. The values for v_M are reported as measured (mAbs/min) as well as converted using the differential extinction coefficient.

	K_M (μM)	v_M (mAbs/min)	v_M ($\mu\text{mol/s}$)
I₈₃₋₉₂			
$x_I = 0 \mu\text{M}$	380 ± 80	9.11 ± 0.88	0.94 ± 0.19
$x_I = 3 \mu\text{M}$	530 ± 90	8.29 ± 0.8	1.14 ± 0.09
$x_I = 10 \mu\text{M}$	870 ± 300	11.71 ± 1.9	1.02 ± 0.27
N88D I₈₃₋₉₃			
$x_I = 0 \mu\text{M}$	380 ± 83	9.01 ± 0.97	0.94 ± 0.19
$x_I = 3 \mu\text{M}$	455 ± 90	12.71 ± 0.94	1.14 ± 0.09

TABLE III. – Values of K_M (in μM) for the HIV-1-PR hydrolyzation of the substrate with different concentrations of the peptidic inhibitor x_I and of the conventional protease inhibitor Saquinavir x_T .

	$x_T = 0 \text{ nM}$	$x_T = 5 \text{ nM}$	$x_T = 10 \text{ nM}$
$x_I = 0 \mu M$	487 ± 150	478 ± 190	612 ± 250
$x_I = 10 \mu M$	1260 ± 530	990 ± 400	1713 ± 620
$x_I = 20 \mu M$	1650 ± 700	1947 ± 970	1921 ± 800

is unfolded, and that it is unfolded because of the specific docking of the peptide onto its predicted binding site.

To provide partial insight into this problem, we followed two different complementary strategies. First, we ran competitive inhibition of the enzyme with our peptide with the well-known transition state analogue Saquinavir. The results are reported in table III. Although the number of points is too slow to draw statistically significant conclusions, the fact that for a fixed value of x_I the inhibition constant of Saquinavir is constant, indicates that the active side pocket is not formed.

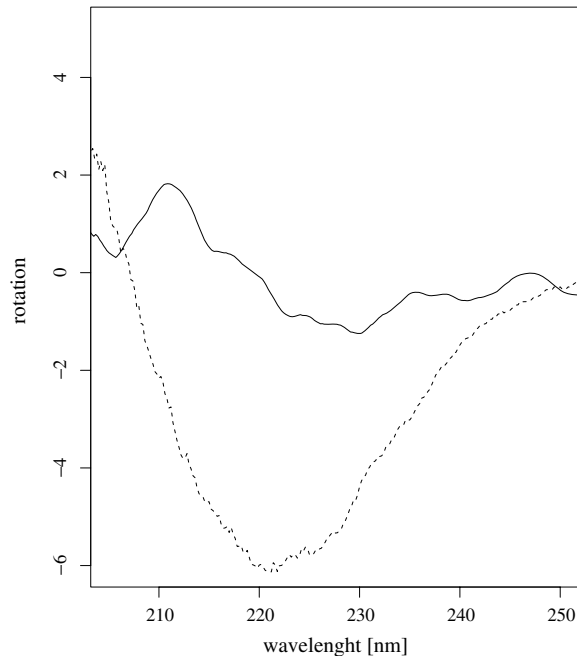


Fig. 3. – The circular dichroism spectrum of the protease (dashed curve) and of the solution composed of the protease and peptide I (continuous curve) in the ratio 1:3, from which the spectrum of the peptide has been subtracted.

Second, we measured and analyzed the change in circular dichroism spectrum upon incubation with the peptide. The CD spectrum of the protease under the same conditions used for the activity assay indicate a β -sheet content of 30%, consistent with that of the native conformation. Figure 3 also displays the CD spectrum of the solution of the protease plus the peptide inhibitor (to which the spectrum of the peptide itself has been subtracted) at the same concentrations and under the same conditions as those of the activity assay. Analysis indicates that the protein has an average content of β -sheet of 14%, and thus is, to a large extent, in a non-folded conformation.

4. – Conclusions

The experimental results presented in this work provide strong evidence that a peptide featuring the same sequence of a fragment of the HIV-1-Protease inhibits the enzyme itself, with an inhibition constant in the micromolar range. Slight modification of the same peptide also have similar effectiveness.

Although the experiments cannot probe at the atomic level the inhibition mechanism at work, circular dichroism assays are consistent with the theoretical and computational prediction of a folding inhibition.

Despite these strongly encouraging results, many questions remain unanswered. First, a complete structural description of the inhibitor-protease complex is still lacking. Such information can be obtained, for instance, from nuclear-magnetic-resonance spectroscopy. Secondly, it would be interesting to assess how specific the inhibition is, in particular with respect to other proteins that display the same fold, and thus are likely to share the same folding driving forces.

In the present case, moreover, medical applications of the method are of interest. In another round of experiments, the peptide has been tested on human peripheral blood mononuclear cells, infected with *ex vivo* HIV isolates, and proved effective (see the contribution of S. Rusconi *et al.* in this volume). Such results open the possibility for a number of interesting experiments to assess the performance of the non-conventional inhibitor on resistant strains where conventional ones fail to stop viral replication.

REFERENCES

- [1] BROGLIA R. A., TIANA G. and BERERA R., Resistance proof, folding-inhibitor drugs, *J. Chem. Phys.*, **118** (2003) 4754.
- [2] BROGLIA R. A., TIANA G. and PROVASI D., Simple model of protein folding and of non-conventional drug design, *J. Phys. Condens. Matter*, **R16** (2004) 111.
- [3] BROGLIA R. A., TIANA G., SUTTO L., PROVASI D. and SIMONA F., Design of HIV-1-PR inhibitors which do not create resistance: blocking the folding of single monomers, *Protein Sci.*, **14** (2005) 2668.
- [4] TIANA G., BROGLIA R. A., SUTTO L. and PROVASI D., Design of a folding inhibitor of the HIV-1 Protease, *Mol. Simul.*, **31** (2005) 765.
- [5] BROGLIA R. A., PROVASI D., VASILE F., OTTOLINA G., LONGHI R. and TIANA G., A folding inhibitor of the HIV-1 Protease, *Proteins*, **62** (2006) 928.

- [6] BROGLIA R. A., TIANA G., SUTTO L., PROVASI D., PERELLI V., Low-Throughput Model Design of Protein Folding Inhibitors, *Proteins*, **67** (2007) 469.
- [7] FERSHT A., *Structure and mechanism in protein science: a guide to enzyme catalysis and protein folding* (W. H. Freeman and Company, New York) 1999.
- [8] SCHNELL S. and MENDOZA C., Closed Form Solution for Time dependent Enzyme Kinetics, *J. Theor. Biol.*, **187** (1997) 207.
- [9] TOMASZEK T. A. jr., MAGAARD V. W., BRYAN H. G., MOORE M. L. and MEEK T. D., Chromophoric peptide substrates for the spectrophotometric assay of HIV1 protease, *Biochem. Biophys. Res. Commun.*, **168** (1990) 274-280.

Susceptibility to a non-conventional (folding) protease inhibitor of human immunodeficiency virus Type 1 isolates *in vitro*

S. RUSCONI(*), M. LO CICERO, S. FERRAMOSCA, F. SIRIANNI, M. GALLI and M. MORONI

Dipartimento di Scienze Cliniche "Luigi Sacco", Sezione di Malattie Infettive e Immunopatologia, Università degli Studi - Milano, Italy

A. E. LAFACE, E. CESANA and A. CLIVIO

Dipartimento di Scienze Precliniche LITA Vialba, Università degli Studi - Milano, Italy

G. TIANA and D. PROVASI

*Dipartimento di Fisica, Università degli Studi - Milano, Italy
and INFN, Sezione di Milano - Milano, Italy*

R. A. BROGLIA

*Dipartimento di Fisica, Università degli Studi - Milano, Italy
and INFN, Sezione di Milano - Milano, Italy*

The Niels Bohr Institute, University of Copenhagen - Copenhagen, Denmark

1. – Introduction

The generation of HIV-1 drug resistant strains depends on either the inability of HIV-1 reverse transcriptase (RT) to proofread nucleotide sequences during replication or the genetic recombination when different viruses infect the same cell [1-3]. Drug resistance is the main problem and the eradication of HIV-1 infection from the host is not long-term effective. Many efforts have been done to develop new compounds able to block the

(*) E-mail: stefano.rusconi@unimi.it

HIV-1 replication. Currently, most of the treatments used in clinical practice are aimed to inhibit the HIV-1 enzymes.

HIV-1 protease is an aspartic enzyme consisting of two identical 99-amino acid monomers and cleaves the *gag* and *pol* viral poly-proteins in the maturation step during the replication cycle of the virus. Inhibition of this step prevents the maturation and production of new viral particles [4-6].

Gene *pol*, which codify for HIV-1 protease, presents highly conserved genetic regions containing important functional domains of the HIV-1 PR, such as the N and C termini, that are essential for dimerization, the active site, the flap and the LES (Local Elementary Structures) regions [7-10]. LES play a key role because lead the proteic folding and are important in the stability of the HIV-1 PR structural conformation (see also Broglia *et al.* and Tiana *et al.* contributions to this volume).

Based on theoretical approaches and optical enzymatic assays [11-13], see also Provasi *et al.* contribution to this volume), we designed a new synthetic peptide characterized by the same sequence as the LES associated with segment 83-92 (NIIGRNLLTQ) able to interact with the PR and potentially inhibit the folding process.

To clarify the *in vitro* efficacy of the p-LES BRU 83-92⁽¹⁾ we performed tests of susceptibility on PBMC (peripheral blood mononuclear cells) from pooled HIV-1 seronegative donors. To better understand the action mechanism of the p-LES BRU 83-92, we also performed Western-blot analyses evaluating the capacity of the peptide to interfere with the Gag-Pol poly-protein processing and its ability to inhibit the maturation of HIV-1 structural proteins like p24.

2. – Experimental strategy

Peripheral blood mononuclear cells (PBMC) taken from the patients were cultivated with PBMC from healthy donors according to the method described by Johnson *et al.* [14]. Cell-free supernatants were assayed twice a week by a p24 antigen-production ELISA (Perkin Elmer, Wellesley, MA, USA).

We used a viral isolate endowed of high replication rate (IT). Viral titration was carried out in PBMC and the viral titer, measured as the 50% tissue-culture infectious dose (TCID₅₀/mL), was calculated using the method of Reed & Muench [15].

In each drug study, 3- or 4-day phytohaemoagglutinin (PHA)-stimulated BMC were exposed to the HIV-1 inoculum (1000 TCID₅₀/mL per 10⁶ cells) without a subsequent wash. Drugs were added simultaneously. Cells were suspended in a 1.0 mL final volume of R-20 medium supplemented with 10% interleukin-2 in 24-well tissue culture plates, and incubated in a humidified atmosphere with 5% CO₂ at 37°C. In all experiments culture medium was changed twice weekly so that 0.5 mL of cell suspension was resuspended in 1.0 mL of fresh medium that contained the original drug concentration(s).

⁽¹⁾ BRU: Broglia Research Unit (Broglia, Tiana, Provasi).

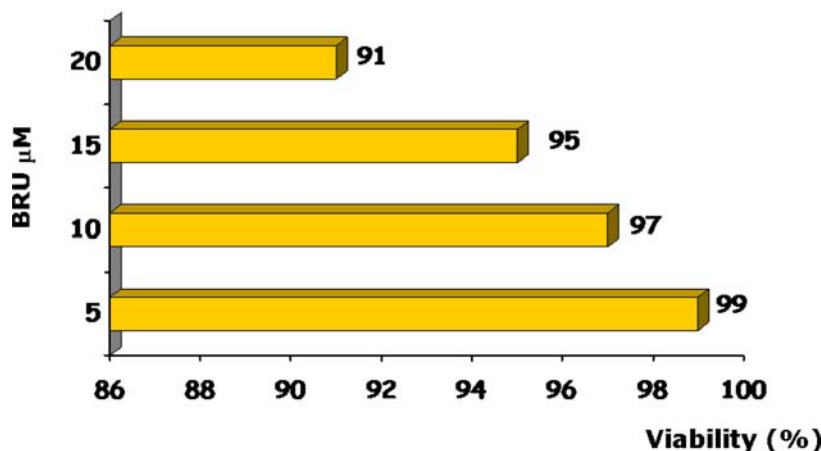


Fig. 1. – Cellular toxicity of BRU at different concentrations on PBMCs after 7 days of culture.

BRU 83-92 was tested at four different concentrations (5, 10, 15 and 20 μM), while ATV was used at 10, 1, 0.1 and 0.01 μM .

The drug concentrations that inhibited the viruses were evaluated in PBMC according to the method described previously [16, 17]. Each single drug and each combination was tested in duplicate, and each experiment was repeated at least twice. In addition, uninfected drug-treated toxicity controls were maintained at the highest concentration for each agent studied. We also maintained viruses without cells or drugs for the entire duration of the experiments in order to take into account the viral carryover.

On day 7, the cultures were ended and cell-free supernatants were harvested for p24 antigen-production ELISA, for measurement of virus replication. Cell proliferation and viability were assessed by the Trypan Blue exclusion method.

To determine the action mechanism of BRU, we performed western-blot assays. The virus-enriched pellets (after supernatant microcentrifugation) were resuspended in SDS sample buffer in the presence of β -mercaptoethanol and electrophoresed in a 12.5% Acrylamide-polyacrylamide gel. Proteins were transferred to a PVDF membrane and detected by using HIV-specific antibodies against p24 (Aalto, Ireland) and secondary antibodies labelled with alkaline phosphatase. Bands detection was performed through the Total Lab program.

3. – Results

In all our experiments none of the two drugs tested showed toxicity and reached 50% cellular toxicity (CC_{50}). In particular, BRU 83-92 displayed a reduced toxicity even at the highest concentration (20 μM) as reported in fig. 1.

As far as the susceptibility is concerned, we tested our peptide against an HIV-1 isolate endowed with a rapid replication capacity (IT strain).

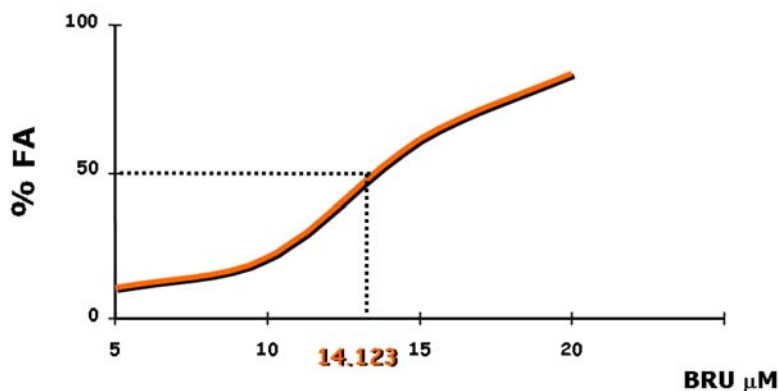


Fig. 2. – Replication rate under drug pressure and IC_{50} value.

On this isolate, BRU 83-92 exhibited a concentration-dependent activity and the IC_{50} value and viral replication rate are reported in fig. 2.

Subsequently, we performed other susceptibility assays to compare the BRU 83-92 and ATV activity at concentration of drug equal to their IC_{50} values (12-14 μM and 4 nM, respectively). We also maintained virus without cells or drugs for the entire duration of the experiment in order to take into account the viral carryover.

Both drugs were able to strongly inhibit the viral replication. Notably, BRU 83-92 was able to inhibit more than 92% of the viral replication, while ATV showed an 80% inhibitory effect (fig. 3).

At last, we carried out Western-blot assays such as to value the activity of BRU 83-92 on the inhibition of the p24 production. Comparing the electropherograms derived from

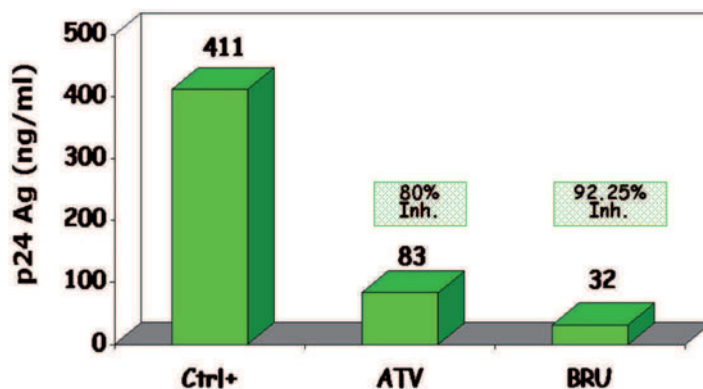


Fig. 3. – Production of p24 antigen in the absence and presence of drugs and their inhibitory effect.

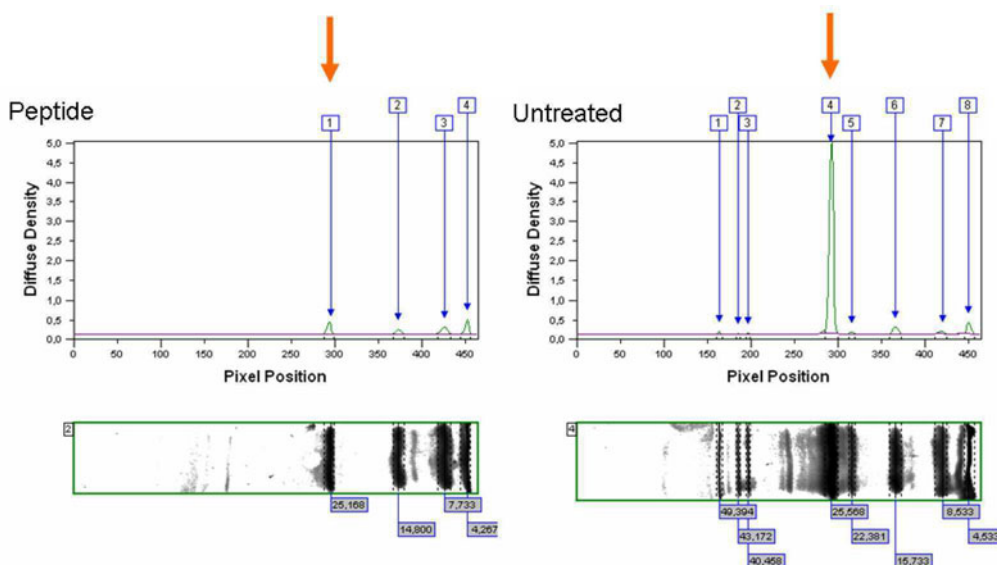


Fig. 4. – Electropherogram relating to the production of p24 antigen under drug pressure and in the absence of BRU 83-92.

the Western-blot analysis we were able to show the reduction of p24 yield under drug pressure at IC_{50} value by our peptide.

As is possible to see in fig. 4, the experimental condition in the presence of BRU 83-92 showed a conspicuous decrement of the p24 signal intensity either on the electropherogram or the gel picture compared with those of the untreated condition. This decrement had been evidenced in regard to the area, the volume, and the height of the peak belonging to the detected signal. These findings had been consistent in multiple experiments.

4. – Conclusions

Protease inhibitors (PI) have a pivotal role in combination therapy against HIV-1, thus it is mandatory to exploit strategies using better compounds directed against this enzyme, taking into consideration the problem of cross-resistance.

The drugs that had been lately presented showed very interesting characteristics from the chemical side and resistance pattern. Serious attempts to explore the feasibility of drug therapeutics with these compounds are ongoing. The clinical trials of the newest among these drugs are always preceded by clarifying pharmacokinetic issues in healthy volunteers. Many among these PI are, or will be in the near future, involved in phase-III head-to-head randomized clinical trial with the PI which are currently considered the ideal comparison in optimized antiretroviral regimens.

New compounds within the class of PI are strongly needed to improve treatment efficacy, reduce toxicity, and provide wider therapeutic options for patients who failed

multiple regimens with current antiviral agents. Over the next 5–10 years, the new agents will need to be proven active against a range of viral variants that are resistant to current PI, less likely to cause the side effects associated with current PI (*e.g.* lipid abnormalities), and show a better bioavailability, such as be administered as prodrugs. If all, or a consistent part of, this will be accomplished, a better care of HIV-infected individuals is granted thus causing a higher quality of life and a longer survival.

The compound described here (BRU 83-92), based on the concept of inhibiting the proteic folding of HIV-1 protease, could be a very useful therapeutic tool in the future clinical arena. Its characteristics could also overcome the occurrence of drug resistance, which often impairs the success of antiretroviral regimens. In fact, preliminar results underlined a proper inhibition effectiveness of BRU 83-92 in a multiresistant viral isolate as compared to a conspicuous loss of inhibitor activity of the control drug ATV, which in any case was not effective on the patient infected with this variant of the virus.

Furthermore, it was found that BRU 83-92 displays an excellent therapeutic/toxic ratio, with antiviral levels well below toxic concentrations. A correspondence between the p24 Ag ELISA and the Western blot detection has been demonstrated. The p24 protein quantification among the different experimental conditions has been evident, particularly between the infected control and the treatment with BRU 83-92. These characteristics underline a possible new approach for the inhibition of HIV-1 infection.

REFERENCES

- [1] PRESTON B., Reverse transcriptase fidelity and HIV-1 variation, *Science*, **275** (1997) 228.
- [2] GU Z., GAO Q., FAUNT E. A. and WAINBERG A., Possible involvement of cell fusion and viral recombination in generation of human immunodeficiency virus variants that display dual resistance to AZT and 3TC, *J. Gen. Virol.*, **76** (1995) 2601.
- [3] SRINIVASAN A., YORK D., JANNOUN-NASR R., KALYANARAMAN S., SWAN D., BENSON J., BOHAN C., LUCIW P. A., SCHNOLL S., ROBINSON R. A., DESAI S. M. and DEVARE S. C., Generation of hybrid human immunodeficiency virus by homologous recombination, *Proc. Natl. Acad. Sci. U.S.A.*, **86** (1989) 6388.
- [4] TOTH G. and BORICS A., Flap opening mechanism of HIV-1 protease, *J. Mol. Graph. Model.*, **24** (2006) 465.
- [5] DAVIS D. A., NEWCOMB F. M., STARKE D. W., OTT D. E., MIEYAL J. M. and YARCHOAN R., Thioltransferase (glutaredoxin) is detected within HIV-1 and can regulate the activity of glutathionylated HIV-1 protease in vitro, *J. Biol. Chem.*, **272** (1997) 25935.
- [6] DAVIS D. A., NEWCOMB F. M., MOSKOVITZ J., WINGFIELD P. T., STAHL S. J., KAUFMAN J., FALES H. M., LEVINE R. L. and YARCHOAN R., HIV-2 protease is inactivated after oxidation at the dimmer interface and activity can be partly restored with methionine sulfoxide reductase, *Biochem. J.*, **346** (2000) 305.
- [7] WU T. D., SCHIFFER C. A., GONZALES M. J., TAYLOR J., KANTOR R. and CHOU S. *et al.*, Mutation pattern and structural correlates in immunodeficiency virus type 1 protease following different protease inhibitor treatments, *J. Virol.*, **77** (2003) 4836.
- [8] CECCHERINI-SILBERSTEIN F., ERBA F., GAGO F., BERTOLI A., FORBICI F., BELLOCCHI M. C., GORI C., D'ARRIGO R., MARCON L., BALOTTA C., ANTINORI A., D'ARMINIO MONFORTE A. and PERNO C. F., Identification of the minimal conserved structure of HIV-1 protease in the presence and absence of drug pressure, *AIDS*, **18** (2004) 11.

- [9] BROGLIA R. A. and TIANA G., Reading the three-dimensional structure of lattice model designed proteins from their amino acids sequence, *Proteins*, **45** (2001) 421.
- [10] TIANA G. and BROGLIA R. A., Folding and design of dimeric proteins, *Proteins*, **49** (2002) 82.
- [11] BROGLIA R. A., TIANA G., SUTTO L., PROVASI D. and SIMONA F., Design of HIV-1-PR inhibitors that do not create resistance: blocking the folding of single monomers, *Protein Sci.*, **14** (2005) 2668.
- [12] BROGLIA R. A., PROVASI D., VASILE F., OTTOLINA G., LONGHI R. and TIANA G., A folding inhibitor of the HIV-1 protease, *Proteins*, **62** (2006) 928.
- [13] BROGLIA R. A., TIANA G., L. SUTTO L., PROVASI D. and PERELLI V., Low-throughput model design of protein folding inhibitors, *Proteins*, **67** (2007) 469.
- [14] JOHNSON V. A., BARLOW M. A. and MERRILL D. P. *et al.*, Three-drug synergistic inhibition of HIV-1 replication *in vitro* by zidovudine, recombinant soluble CD4 and recombinant interferon-alpha A, *J Infect. Dis.*, **161** (1990) 1059.
- [15] DULBECCO R., Endpoint methods – measurements of the infectious titer of a viral sample. In *Virology*, edited by DULBECCO R. and GINSBERG H. S. (Lippincott, Philadelphia, PA) 1998, pp. 22–5.
- [16] RUSCONI S., DE PASQUALE M. P. and MILAZZO L. *et al.*, *In vitro* effects of continuous pressure with zidovudine (ZDV) and lamivudine on a ZDV-resistant HIV-1 isolate, *AIDS*, **11** (1997) 1406.
- [17] RUSCONI S., DE PASQUALE M. P. and MILAZZO L. *et al.*, Loss of lamivudine resistance in a zidovudine and lamivudine dual-resistant human immunodeficiency virus type 1 (HIV-1) isolate after discontinuation of *in vitro* lamivudine drug pressure, *Antiviral Therapy*, **3** (1998) 203.

This page intentionally left blank

International School of Physics “Enrico Fermi”

Villa Monastero, Varenna

Course CLXV

4–14 July 2006

“Protein Folding and Drug Design”

Directors

Ricardo A. BROGLIA
Dipartimento di Fisica
Università di Milano
Via Celoria 16
20133 MILANO
Italy
Tel.: ++39 02 50317231
Fax: ++39 02 50317487
broglia@mi.infn.it

Luis SERRANO
European Molecular Biology Laboratory
(EMBL)
Meyerohofstrasse 1
D-69117 HEIDELBERG
Germany
Tel.: ++49 6221 3878320
Fax: ++49 6221 3878519
serrano@EMBL-Heidelberg.de

Scientific Secretary

Guido TIANA
Dipartimento di Fisica
Università di Milano
Via Celoria 16
20133 MILANO
Italy
Tel.: ++39 02 50317221
Fax: ++39 02 50317487
guido.tiana@mi.infn.it

Lecturers

Paolo CARLONI
SISSA
Via Beirut 4
34014 TRIESTE
Italy
Tel.: ++39 040 3787407
Fax: ++39 040 3787528
carloni@sissa.it

Amos MARITAN
Dipartimento di Fisica “G. Galilei”,
Università di Padova
Via Marzolo 8
35131 PADOVA
Italy
Tel.: ++39 049 8277175
Fax: ++39 049 8277102
maritan@pd.infn.it

Kenneth MERZ
Department of Chemistry
Penn State University
152 Davey Laboratory
UNIVERSITY PARK, PA 16802
USA
Tel.: ++1 352 3926973
Fax: ++1 352 3928758
merz@psu.edu

Leonid MIRNY
MIT
Health Sciences and Technology Division
bldg 16-343
77 Massachussets Ave.
CAMBRIDGE, MA 02139
USA
Tel.: ++1 617 4524862
Fax: ++1 617 2532514
leonid@mit.edu

Harold A. SCHERAGA
Baker Laboratory of Chemistry
and Chemical Biology
Cornell University
ITHACA, NY 14853-1301
USA
Tel.: ++1 607 2554034
Fax: ++1 607 2544700
has5@cornell.edu

Eugene SHAKHNOVICH
Department of Chemistry
and Chemical Biology
Harvard University
12 Oxford Street
CAMBRIDGE, MA 02138
USA
Tel.: ++1 617 4954130
Fax: ++1 617 4965948
eugene@belok.harvard.edu

Wilfred van GUNSTEREN
ETH
Laboratorium für Physikalische Chemie
ETH Hönggerberg HCI
CH-8093 ZÜRICH
Switzerland
Tel.: ++41 44 6325501
Fax: ++41 1 6321039
wfvgn@igc.phys.chem.ethz.ch

Gennady VERKHIVKER
Pfizer, Inc.
La Jolla laboratories
10614 Science Center Drive
SAN DIEGO, CA 92121-1111
USA
Tel.: ++1 858 3420392
gverkhiv@ucsd.edu,
gverkhivker@san.rr.com

Peter James WINN
EMBL Research GmbH
Villa Bosch
Schloss-Wolfsbrunnenweg 31c
69118 HEIDELBERG
Germany
Tel.: ++49 6221 533269
Fax: ++49 6221 533298
peter.winn@embl-r.villa-bosch.de

Peter WOLYNES
Department of Chemistry
and Biochemistry
University of California San Diego
9500 Gilman Drive
LA JOLLA, CA 92093-0303
USA
Tel.: ++1 858 8224825
Fax: ++1 858 5347687
pwolynes@chem.ucsd.edu

Seminar Speakers

Giorgio COLOMBO
CNR-ICRM
Via Mario Bianco 9
20131 MILANO
Italy
Tel.: ++39 02 28500031
Fax: ++39 02 28901239
g.colombo@icrm.cnr.it

Paolo DE LOS RIOS
Laboratoire de Biophysique Statistique
École Polytechnique Fédérale de Lausanne
1013 LAUSANNE
Switzerland
Tel.: ++41 21 6930510-09
Fax: ++41 21 6939523
paolo.delosrios@epfl.ch

Michele PARRINELLO
Department of Chemistry
and Applied Biosciences
ETH Zürich
USI - Campus, Via Giuseppe Buffi 13
CH - 6900 LUGANO
Switzerland
Tel.: ++41 58 6664801
Fax: ++41 58 6664817
parrinello@phys.chem.ethz.ch

Students

Alessandro BARDUCCI
Dipartimento di Chimica
Università di Firenze
Via della Lastruccia 3
50019 SESTO FIORENTINO (FI)
Italy
Tel.: ++39 055 4573082
Fax: ++39 055 4573077
alessandro.barducci@unifi.it

Tetyana BOGDAN
Department of Chemistry
University of Cambridge
Lensfield Road
CAMBRIDGE CB2 1EW
UK
Tel.: ++44 01223 336530
tvb20@cam.ac.uk

Massimiliano BONOMI
Department of Chemistry and Biosciences
ETH
Via Buffi 13
6900 LUGANO
Switzerland
Tel.: ++41 58 6664803
Fax: ++41 58 6664817
mbonomi@phys.chem.ethz.ch

Davide BRANDUARDI
Department of Chemistry and Biosciences
ETH
Via Buffi 13
6900 Lugano
Switzerland
Tel.: ++41 58 6664805
Fax: ++41 58 6664817
davide.branduardi@phys.chem.ethz.ch

Carlo CAMILLONI
Dipartimento di Fisica
Università di Milano
Via Celoria 16
20133 MILANO
Italy
Tel.: ++39 02 50317654
Fax: ++39 02 50317487
carlo.camilloni@mi.infn.it

Paolo CERRI
Dipartimento di Fisica
Università di Milano
Via Celoria 16
20133 MILANO
Italy
Tel.: ++39 02 50317654
Fax: ++39 02 50317487
paolo.cerri@mi.infn.it

Maria COTALLO ABAN
Instituto de Biocomputacion
y Fisica de Sistemas Complejos
Corona de Aragon 42
50009 ZARAGOZA
Spain
Tel.: ++34 976 562212
Fax: ++34 976 562215
mcotallo@unizar.es

Lucia DANDREA
Dipartimento di Fisica
Università di Trento
Via Sommarive 14
38050 Povo (TN)
Italy
Tel.: ++39 0461 882043
Fax: ++39 0461 882014
ldandrea@science.unitn.it

Elsa Fernanda DE SOUSA HENRIQUES
Frankfurt Institute for Advanced Studies
Max-von-Laue Strasse 1
60438 FRANKFURT A.M.
Germany
Tel.: ++49 69 79847502
Fax: ++49 69 79847611
henriques@fias.uni-frankfurt.de

Giovanni GRAZIOSO
Istituto di Chimica Farmaceutica
e Tossicologica
Università di Milano
Viale Abruzzi 42
20131 MILANO
Italy
Tel.: ++39 02 50317570
Fax: ++36 02 50317574
giovanni.grazioso@unimi.it

Dominik GRONT
Faculty of Chemistry
Warsaw University
Pasteura 1
02-093 WARSAW
Poland
Tel.: ++48 22 8220211 ext. 310
dgront@chem.uw.edu.pl

Carlo GUARDIANI
Centro Interdipartimentale per lo Studio
di Dinamiche Complesse
Università di Firenze
Via Sansone 1
50019 SESTO FIORENTINO (FI)
Italy
Tel.: ++39 055 4572278
Fax: ++39 055 4572340
carlo.guardiani@unifi.it,
guardiani@fi.infn.it

Marius HATLO
School of Chemical Engineering
University of Manchester
PO Box 88
Sackville St.
MANCHESTER M 60 1QD
UK
Tel.: ++44 787 0121616
Fax: ++44 161 3064399
marius.hatlo@postgrad.manchester.ac.uk

Fernando Enrique HERRERA
SISSA
Via Beirut 2-4
34014 TRIESTE
Italy
Tel.: ++39 040 3787321
Fax: ++39 040 3787528
herrera@sissa.it

Mateusz KURCINSKI
Faculty of Chemistry, Warsaw University
Pasteura 1
02093 WARSAW
Poland
Tel.: ++48 22 8220211 ext. 310
Fax: ++48 22 8225996
mkurc@chem.uw.edu.pl

Ylva IVARSSON
Department of Biochemistry and Organic
Chemistry BMC
University of Uppsala
Box 576
75123 UPPSALA
Sweden
Tel.: ++46 18 4714312
Fax: ++46 18 558431
ylva.ivarsson@biokemi.uu.se

Luca LANZANÒ
Dipartimento di Metodologie Fisiche
e Chimiche per l'Ingegneria
Università di Catania
Viale Andrea Doria 6
95125 CATANIA
and
INFN Laboratori Nazionali del Sud
Via Santa Sofia 62
95123 CATANIA
Italy
Tel.: ++39 095 542305
lanzano-l@lns.infn.it

Svetlana KHATUNTSEVA
Centre "Bioengineering"
Russian Academy of Science
Prospekt 60 - letiya Oktyabrya 7/1
117312 MOSCOW
Russia
Tel.: ++7 495 1356219
Fax: ++7 495 1350571
sdepp@mail.ru

Dorota LATEK
Faculty of Chemistry, Warsaw University
Pasteura 1
02-093 WARSAW
Poland
Tel.: ++48 22 8220211
Fax: ++48 22 8225996
pledor@chem.uw.edu.pl

Sebastian KMIECIK
Faculty of Chemistry, Warsaw University
Laboratory of Theory of Biopolymers
Pasteura 1
02-093 WARSAW
Poland
Tel.: ++48 22 8220211
Fax: ++48 22 8225996
sekmi@chem.uw.edu.pl

Jessica SILTBERG LIBERLES
Department of Molecular Biology
University of Wyoming
LARAMIE, WY 82071
USA
Tel.: ++1 307 7665534
Fax: ++1 307 7665098
jessica.liberles@biomed.uib.no

Stefano LUCCIOLI
Istituto dei Sistemi Complessi del CNR
Via Madonna del Piano
50019 SESTO FIORENTINO (FI)
Italy
Tel.: ++39 055 5226626
Fax: ++39 055 5226683
`stefano.luccioli@fi.isc.cnr.it`

Francesco MARINI
Dipartimento di Fisica
Università di Milano
Via Celoria 16
20133 MILANO
Italy
Tel.: ++39 02 50317221
Fax: ++39 02 50317487
`franz.marini@mi.infn.it`

Inesa MESROPYAN
Institute of Physics
Georgian Academy of Sciences
6 Tamaraskvili str.
0177 TBILISI
Georgia
Tel.: ++99 59 3344597
Fax: ++99 53 2391494
`imesropian@yahoo.com`

Elham MIRMOMTAZ
ICTP-Elettra
Strada Costiera 11
34014 TRIESTE
Italy
Tel.: ++39 040 3758755
`emirmomt@ictp.it`,
`elmir@ch.iut.ac.ir`

Simon MITTERNACHT
Department of Theoretical Physics
Lund University
Solvegatan 14A
SE-22362 LUND
Sweden
Tel.: ++46 46 2223494
Fax: ++46 46 2229686
`simon@thep.lu.se`

Irina MOREIRA
Computational Chemistry Group
Faculdade de Ciencias
Universidade de Porto
Rua do Campo Alegre 687
4169-007 PORTO
Portugal
Tel.: ++351 22 6082955
Fax: ++351 22 6082959
`irina.moreira@fc.up.pt`

Franco Maria NERI
Dipartimento di Fisica
Università di Parma
Viale delle Scienze 7/A
43100 PARMA
Italy
Tel.: ++39 0521 905493
Fax: ++39 0521 905223
`neri@fis.unipr.it`

Chi Celestine NGANG
Department of Medical Biochemistry
and Microbiology
University of Uppsala
Biomedical Center
Box 582
SE-75123 UPPSALA
Sweden
Tel.: ++46 18 4714359
Fax: ++46 18 4714673
`celestine.chi@imbim.uu.se`

Marija RAKONJAC
Department of Medical Biochemistry
and Biophysics
Karolinska Institute
Division of Physiological Chemistry II
Scheeles vg 2
17177 STOCKHOLM
Sweden
Tel.: ++46 8 52487642
Fax: ++46 8 736043
marija.rakonjac@ki.se

Aleksandr SAHAKYAN
Molecule Structure Research Center
Azatutian 26
YEREVAN 375014
Armenia
Tel.: ++37 410 287423
Fax: ++37 410 282267
aleksahak@yahoo.com

Andrea SORANNO
Dipartimento di Chimica, Biochimica
e Biotecnologie Mediche
Università di Milano
Via F.lli Cervi 93
20090 SEGRATE (MI)
Italy
Tel.: ++39 02 50330329
Fax: ++39 02 50330365
andrea.soranno@gmail.com

Ludovico SUTTO
Dipartimento di Fisica
Università di Milano
Via Celoria 16
20133 MILANO
Italy
Tel.: ++39 02 50317221
Fax: ++39 02 50317487
ludovico.sutto@mi.infn.it

Rui TRAVASSO
University of Lisbon
CFTC, Complexo Interdisciplinar
Av. Prof Gama Pinto 2
1649-003 LISBOA
Portugal
rui@cii.fc.ul.pt

Antonio TROVATO
Dipartimento di Fisica
Università di Padova
Via Marzolo 8
35131 PADOVA
Italy
Tel.: ++39 049 8277159
Fax: ++39 049 8277102
trovato@pd.infn.it

Semen TRYGUBENKO
Department of Chemistry
University of Cambridge
Lensfield Road
CAMBRIDGE CB2 1EW
UK
Tel.: ++44 1223 336530
Fax: ++44 1223 336362
sat39@cam.ac.uk

Valeria VETRI
Dipartimento di Scienze Fisiche
ed Astronomiche
Università di Palermo
Via Archirafi 36
90123 PALERMO
Italy
Tel.: ++39 091 6234219
Fax: ++39 091 6162461
valeria.vetri@fisica.unipa.it

Semen YESYLEVSKYY
Department of Physics of Biological
Systems
Institute of Physics
prosp. Nauky 46
03039 KIEV
Ukraine
Tel.: ++38 044 5259851
Fax:
yesint3@yahoo.com

Marco ZAMPARO
Dipartimento di Fisica
Politecnico di Torino
Corso Duca degli Abruzzi 24
10129 TORINO
Italy
Tel.: ++39 011 5647384
Fax: ++39 011 5647399
marco.zamparo@polito.it

Observers

Marco BUSCAGLIA
Dipartimento di Chimica, Biochimica
e Biotecnologie Mediche
Università di Milano
Via F.lli Cervi 93
20090 SEGRATE (MI)
Italy
Tel.: ++39 02 50330352
Fax: ++39 02 50330365
marco.buscaglia@unimi.it

Luca NARDO
Dipartimento di Fisica e Matematica
Università dell'Insubria
Via Valleggio 11
22100 COMO
Italy
Tel.: ++39 031 2386272
Fax: ++39 031 2386119
luca.nardo@uninsubria.it

Giovanni NICO
Istituto Applicazioni del Calcolo del CNR
Via Amendola 122/D
70126 BARI
Italy
Tel.: ++39 080 5929753
Fax: ++39 080 5929770
g.nico@ba.iac.cnr.it

Oleg BOLENSKI
Frankfurt Institute for Advanced Studies
Max-von-Laue Strasse 1
60438 FRANKFURT
Germany
Tel.: ++49 69 79847623
Fax: ++49 69 79847611
o.obolensky@fias.uni-frankfurt.de

Filippo PULLARA
Dipartimento di Scienze Fisiche
ed Astronomiche
Università di Palermo
Via Archirafi 36
90123 PALERMO
Italy
Tel.: ++39 091 6234302
Fax: ++39 091 6173889
pullara@fisica.unipa.it

PROCEEDINGS OF THE INTERNATIONAL SCHOOL OF PHYSICS “ENRICO FERMI”

Course I (1953)

Questioni relative alla rivelazione delle particelle elementari, con particolare riguardo alla radiazione cosmica
edited by G. PUPPI

Course II (1954)

Questioni relative alla rivelazione delle particelle elementari, e alle loro interazioni con particolare riguardo alle particelle artificialmente prodotte ed accelerate
edited by G. PUPPI

Course III (1955)

Questioni di struttura nucleare e dei processi nucleari alle basse energie
edited by C. SALVETTI

Course IV (1956)

Proprietà magnetiche della materia
edited by L. GIULOTTO

Course V (1957)

Fisica dello stato solido
edited by F. FUMI

Course VI (1958)

Fisica del plasma e relative applicazioni astrofisiche
edited by G. RIGHINI

Course VII (1958)

Teoria della informazione
edited by E. R. CAIANIELLO

Course VIII (1958)

Problemi matematici della teoria quantistica delle particelle e dei campi
edited by A. BORSELLINO

Course IX (1958)

Fisica dei pioni
edited by B. TOUSCHEK

Course X (1959)

Thermodynamics of Irreversible Processes
edited by S. R. DE GROOT

Course XI (1959)

Weak Interactions
edited by L. A. RADICATI

Course XII (1959)

Solar Radioastronomy
edited by G. RIGHINI

Course XIII (1959)

Physics of Plasma: Experiments and Techniques
edited by H. ALFVÉN

Course XIV (1960)

Ergodic Theories
edited by P. CALDIROLA

Course XV (1960)

Nuclear Spectroscopy
edited by G. RACAH

Course XVI (1960)

Physicomathematical Aspects of Biology
edited by N. RASHEVSKY

Course XVII (1960)

Topics of Radiofrequency Spectroscopy
edited by A. GOZZINI

Course XVIII (1960)

Physics of Solids (Radiation Damage in Solids)
edited by D. S. BILLINGTON

Course XIX (1961)

Cosmic Rays, Solar Particles and Space Research
edited by B. PETERS

Course XX (1961)

Evidence for Gravitational Theories
edited by C. MØLLER

Course XXI (1961)

Liquid Helium
edited by G. CARERI

Course XXII (1961)

Semiconductors
edited by R. A. SMITH

Course XXIII (1961)

Nuclear Physics
edited by V. F. WEISSKOPF

Course XXIV (1962)

Space Exploration and the Solar System
edited by B. ROSSI

Course XXV (1962)

Advanced Plasma Theory
edited by M. N. ROSENBLUTH

Course XXVI (1962)

Selected Topics on Elementary Particle Physics

edited by M. CONVERSI

Course XXVII (1962)

Dispersion and Absorption of Sound by Molecular Processes

edited by D. SETTE

Course XXVIII (1962)

Star Evolution

edited by L. GRATTON

Course XXIX (1963)

Dispersion Relations and their Connection with Casuality

edited by E. P. WIGNER

Course XXX (1963)

Radiation Dosimetry

edited by F. W. SPIERS and G. W. REED

Course XXXI (1963)

Quantum Electronics and Coherent Light

edited by C. H. TOWNES and P. A. MILES

Course XXXII (1964)

Weak Interactions and High-Energy Neutrino Physics

edited by T. D. LEE

Course XXXIII (1964)

Strong Interactions

edited by L. W. ALVAREZ

Course XXXIV (1965)

The Optical Properties of Solids

edited by J. TAUC

Course XXXV (1965)

High-Energy Astrophysics

edited by L. GRATTON

Course XXXVI (1965)

Many-body Description of Nuclear Structure and Reactions

edited by C. L. BLOCH

Course XXXVII (1966)

Theory of Magnetism in Transition Metals

edited by W. MARSHALL

Course XXXVIII (1966)

Interaction of High-Energy Particles with Nuclei

edited by T. E. O. ERICSON

Course XXXIX (1966)

Plasma Astrophysics

edited by P. A. STURROCK

Course XL (1967)

Nuclear Structure and Nuclear Reactions

edited by M. JEAN and R. A. RICCI

Course XLI (1967)

Selected Topics in Particle Physics

edited by J. STEINBERGER

Course XLII (1967)

Quantum Optics

edited by R. J. GLAUBER

Course XLIII (1968)

Processing of Optical Data by Organisms and by Machines

edited by W. REICHARDT

Course XLIV (1968)

Molecular Beams and Reaction Kinetics

edited by CH. SCHLIER

Course XLV (1968)

Local Quantum Theory

edited by R. JOST

Course XLVI (1969)

Physics with Intersecting Storage Rings

edited by B. TOUSCHEK

Course XLVII (1969)

General Relativity and Cosmology

edited by R. K. SACHS

Course XLVIII (1969)

Physics of High Energy Density

edited by P. CALDIROLA and H. KNOEPFEL

Course IL (1970)

Foundations of Quantum Mechanics

edited by B. D'ESPAGNAT

Course L (1970)

Mantle and Core in Planetary Physics

edited by J. COULOMB and M. CAPUTO

Course LI (1970)

Critical Phenomena

edited by M. S. GREEN

Course LII (1971)

Atomic Structure and Properties of Solids

edited by E. BURSTEIN

Course LIII (1971)

Developments and Borderlines of Nuclear Physics

edited by H. MORINAGA

Course LIV (1971)

Developments in High-Energy Physics

edited by R. R. GATTO

Course LV (1972)

Lattice Dynamics and Intermolecular Forces

edited by S. CALIFANO

Course LVI (1972)

Experimental Gravitation

edited by B. BERTOTTI

- Course LVII (1972)
History of 20th Century Physics
edited by C. WEINER
- Course LVIII (1973)
Dynamics Aspects of Surface Physics
edited by F. O. GOODMAN
- Course LIX (1973)
Local Properties at Phase Transitions
edited by K. A. MÜLLER and A. RIGAMONTI
- Course LX (1973)
C*-Algebras and their Applications to Statistical Mechanics and Quantum Field Theory
edited by D. KASTLER
- Course LXI (1974)
Atomic Structure and Mechanical Properties of Metals
edited by G. CAGLIOTI
- Course LXII (1974)
Nuclear Spectroscopy and Nuclear Reactions with Heavy Ions
edited by H. FARAGGI and R. A. RICCI
- Course LXIII (1974)
New Directions in Physical Acoustics
edited by D. SETTE
- Course LXIV (1975)
Nonlinear Spectroscopy
edited by N. BLOEMBERGEN
- Course LXV (1975)
Physics and Astrophysics of Neutron Stars and Black Hole
edited by R. GIACCONI and R. RUFFINI
- Course LXVI (1975)
Health and Medical Physics
edited by J. BAARLI
- Course LXVII (1976)
Isolated Gravitating Systems in General Relativity
edited by J. EHLERS
- Course LXVIII (1976)
Metrology and Fundamental Constants
edited by A. FERRO MILONE, P. GIACOMO and S. LESCHIUTTA
- Course LXIX (1976)
Elementary Modes of Excitation in Nuclei
edited by A. BOHR and R. A. BROGLIA
- Course LXX (1977)
Physics of Magnetic Garnets
edited by A. PAOLETTI
- Course LXXI (1977)
Weak Interactions
edited by M. BALDO CEOLIN
- Course LXXII (1977)
Problems in the Foundations of Physics
edited by G. TORALDO DI FRANCA
- Course LXXIII (1978)
Early Solar System Processes and the Present Solar System
edited by D. LAL
- Course LXXIV (1978)
Development of High-Power Lasers and their Applications
edited by C. PELLEGRINI
- Course LXXV (1978)
Intermolecular Spectroscopy and Dynamical Properties of Dense Systems
edited by J. VAN KRANENDONK
- Course LXXVI (1979)
Medical Physics
edited by J. R. GREENING
- Course LXXVII (1979)
Nuclear Structure and Heavy-Ion Collisions
edited by R. A. BROGLIA, R. A. RICCI and C. H. DASSO
- Course LXXVIII (1979)
Physics of the Earth's Interior
edited by A. M. DZIEWONSKI and E. BOSCHI
- Course LXXIX (1980)
From Nuclei to Particles
edited by A. MOLINARI
- Course LXXX (1980)
Topics in Ocean Physics
edited by A. R. OSBORNE and P. MALANOTTE RIZZOLI
- Course LXXXI (1980)
Theory of Fundamental Interactions
edited by G. COSTA and R. R. GATTO
- Course LXXXII (1981)
Mechanical and Thermal Behaviour of Metallic Materials
edited by G. CAGLIOTI and A. FERRO MILONE
- Course LXXXIII (1981)
Positrons in Solids
edited by W. BRANDT and A. DUPASQUIER
- Course LXXXIV (1981)
Data Acquisition in High-Energy Physics
edited by G. BOLOGNA and M. VINCELLI
- Course LXXXV (1982)
Earthquakes: Observation, Theory and Interpretation
edited by H. KANAMORI and E. BOSCHI
- Course LXXXVI (1982)
Gamow Cosmology
edited by F. MELCHIORRI and R. RUFFINI

Course LXXXVII (1982)

Nuclear Structure and Heavy-Ion Dynamics

edited by L. MORETTO and R. A. RICCI

Course LXXXVIII (1983)

Turbulence and Predictability in Geophysical Fluid Dynamics and Climate Dynamics

edited by M. GHIL, R. BENZI and G. PARISI

Course LXXXIX (1983)

Highlights of Condensed Matter Theory

edited by F. BASSANI, F. FUMI and M. P. TOSI

Course XC (1983)

Physics of Amphiphiles: Micelles, Vesicles and Microemulsions

edited by V. DEGIORGIO and M. CORTI

Course XCI (1984)

From Nuclei to Stars

edited by A. MOLINARI and R. A. RICCI

Course XCII (1984)

Elementary Particles

edited by N. CABIBBO

Course XCIII (1984)

Frontiers in Physical Acoustics

edited by D. SETTE

Course XCIV (1984)

Theory of Reliability

edited by A. SERRA and R. E. BARLOW

Course XCV (1985)

Solar-Terrestrial Relationship and the Earth Environment in the Last Millennia

edited by G. CINI CASTAGNOLI

Course XCVI (1985)

Excited-State Spectroscopy in Solids

edited by U. M. GRASSANO and N. TERZI

Course XCVII (1985)

Molecular-Dynamics Simulations of Statistical-Mechanical Systems

edited by G. CICCOTTI and W. G. HOOVER

Course XCVIII (1985)

The Evolution of Small Bodies in the Solar System

edited by M. FULCHIGNONI and Ľ. KRESÁK

Course XCIX (1986)

Synergetics and Dynamic Instabilities

edited by G. CAGLIOTI and H. HAKEN

Course C (1986)

The Physics of NMR Spectroscopy in Biology and Medicine

edited by B. MARAVIGLIA

Course CI (1986)

Evolution of Interstellar Dust and Related Topics

edited by A. BONETTI and J. M. GREENBERG

Course CII (1986)

Accelerated Life Testing and Experts' Opinions in Reliability

edited by C. A. CLAROTTI and D. V. LINDLEY

Course CIII (1987)

Trends in Nuclear Physics

edited by P. KIENLE, R. A. RICCI and A. RUBINO

Course CIV (1987)

Frontiers and Borderlines in Many-Particle Physics

edited by R. A. BROGLIA and J. R. SCHRIEFER

Course CV (1987)

Confrontation between Theories and Observations in Cosmology: Present Status and Future Programmes

edited by J. AUDOUZE and F. MELCHIORRI

Course CVI (1988)

Current Trends in the Physics of Materials

edited by G. F. CHIAROTTI, F. FUMI and M. TOSI

Course CVII (1988)

The Chemical Physics of Atomic and Molecular Clusters

edited by G. SCOLES

Course CVIII (1988)

Photoemission and Absorption Spectroscopy of Solids and Interfaces with Synchrotron Radiation

edited by M. CAMPAGNA and R. ROSEI

Course CIX (1988)

Nonlinear Topics in Ocean Physics

edited by A. R. OSBORNE

Course CX (1989)

Metrology at the Frontiers of Physics and Technology

edited by L. CROVINI and T. J. QUINN

Course CXI (1989)

Solid-State Astrophysics

edited by E. BUSSOLETTI and G. STRAZZULLA

Course CXII (1989)

Nuclear Collisions from the Mean-Field into the Fragmentation Regime

edited by C. DETRAZ and P. KIENLE

Course CXIII (1989)

High-Pressure Equation of State: Theory and Applications

edited by S. ELIEZER and R. A. RICCI

Course CXIV (1990)

Industrial and Technological Applications of Neutrons

edited by M. FONTANA and F. RUSTICHELLI

- Course CXV (1990)
The Use of EOS for Studies of Atmospheric Physics
edited by J. C. GILLE and G. VISCONTI
- Course CXVI (1990)
Status and Perspectives of Nuclear Energy: Fission and Fusion
edited by R. A. RICCI, C. SALVETTI and E. SINDONI
- Course CXVII (1991)
Semiconductor Superlattices and Interfaces
edited by A. STELLA
- Course CXVIII (1991)
Laser Manipulation of Atoms and Ions
edited by E. ARIMONDO, W. D. PHILLIPS and F. STRUMIA
- Course CXIX (1991)
Quantum Chaos
edited by G. CASATI, I. GUARNERI and U. SMILANSKY
- Course CXX (1992)
Frontiers in Laser Spectroscopy
edited by T. W. HÄNSCH and M. INGUSCIO
- Course CXXI (1992)
Perspectives in Many-Particle Physics
edited by R. A. BROGLIA, J. R. SCHRIEFFER and P. F. BORTIGNON
- Course CXXII (1992)
Galaxy Formation
edited by J. SILK and N. VITTORIO
- Course CXXIII (1992)
Nuclear Magnetic Double Resonance
edited by B. MARAVIGLIA
- Course CXXIV (1993)
Diagnostic Tools in Atmospheric Physics
edited by G. FIOCCO and G. VISCONTI
- Course CXXV (1993)
Positron Spectroscopy of Solids
edited by A. DUPASQUIER and A. P. MILLS jr.
- Course CXXVI (1993)
Nonlinear Optical Materials: Principles and Applications
edited by V. DEGIORGIO and C. FLYTZANIS
- Course CXXVII (1994)
Quantum Groups and their Applications in Physics
edited by L. CASTELLANI and J. WESS
- Course CXXVIII (1994)
Biomedical Applications of Synchrotron Radiation
edited by E. BURATTINI and A. BALERNA
- Course CXXIX¹ (1994)
Observation, Prediction and Simulation of Phase Transitions in Complex Fluids
edited by M. BAUS, L. F. RULL and J. P. RYCKAERT
- Course CXXX (1995)
Selected Topics in Nonperturbative QCD
edited by A. DI GIACOMO and D. DIAKONOV
- Course CXXXI (1995)
Coherent and Collective Interactions of Particles and Radiation Beams
edited by A. ASPECT, W. BARLETTA and R. BONIFACIO
- Course CXXXII (1995)
Dark Matter in the Universe
edited by S. BONOMETTO and J. PRIMACK
- Course CXXXIII (1996)
Past and Present Variability of the Solar-Terrestrial System: Measurement, Data Analysis and Theoretical Models
edited by G. CINI CASTAGNOLI and A. PROVENZALE
- Course CXXXIV (1996)
The Physics of Complex Systems
edited by F. MALLAMACE and H. E. STANLEY
- Course CXXXV (1996)
The Physics of Diamond
edited by A. PAOLETTI and A. TUCCIARONE
- Course CXXXVI (1997)
Models and Phenomenology for Conventional and High-Temperature Superconductivity
edited by G. IADONISI, J. R. SCHRIEFFER and M. L. CHIOFALO
- Course CXXXVII (1997)
Heavy Flavour Physics: a Probe of Nature's Grand Design
edited by I. BIGI and L. MORONI
- Course CXXXVIII (1997)
Unfolding the Matter of Nuclei
edited by A. MOLINARI and R. A. RICCI
- Course CXXXIX (1998)
Magnetic Resonance and Brain Function: Approaches from Physics
edited by B. MARAVIGLIA
- Course CXL (1998)
Bose-Einstein Condensation in Atomic Gases
edited by M. INGUSCIO, S. STRINGARI and C. E. WIEMAN
- Course CXLI (1998)
Silicon-Based Microphotonics: from Basics to Applications
edited by O. BISI, S. U. CAMPISANO, L. PAVESI and F. PRIOLO

¹This course belongs to the NATO ASI Series C, Vol. 460 (Kluwer Academic Publishers).

- Course CXLII (1999)
Plasmas in the Universe
 edited by B. COPPI, A. FERRARI and E. SINDONI
- Course CXLIII (1999)
New Directions in Quantum Chaos
 edited by G. CASATI, I. GUARNERI and U. SMILANSKY
- Course CXLIV (2000)
Nanometer Scale Science and Technology
 edited by M. ALLEGRI, N. GARCÍA and O. MARTI
- Course CXLV (2000)
Protein Folding, Evolution and Design
 edited by R. A. BROGLIA, E. I. SHAKHNOVICH and G. TIANA
- Course CXLVI (2000)
Recent Advances in Metrology and Fundamental Constants
 edited by T. J. QUINN, S. LESCHIUTTA and P. TAVELLA
- Course CXLVII (2001)
High Pressure Phenomena
 edited by R. J. HEMLEY, G. L. CHIAROTTI, M. BERNASCONI and L. ULIVI
- Course CXLVIII (2001)
Experimental Quantum Computation and Information
 edited by F. DE MARTINI and C. MONROE
- Course CXLIX (2001)
Organic Nanostructures: Science and Applications
 edited by V. M. AGRANOVICH and G. C. LA ROCCA
- Course CL (2002)
Electron and Photon Confinement in Semiconductor Nanostructures
 edited by B. DEVEAUD-PLÉDRAN, A. QUATTROPANI and P. SCHWENDIMANN
- Course CLI (2002)
Quantum Phenomena in Mesoscopic Systems
 edited by B. ALTSHULER, A. TAGLIACCOZZO and V. TOGNETTI
- Course CLII (2002)
Neutrino Physics
 edited by E. BELLOTTI, Y. DECLAIS and P. STROLIN
- Course CLIII (2002)
From Nuclei and their Constituents to Stars
 edited by A. MOLINARI, L. RICCATI, W. M. ALBERICO and M. MORANDO
- Course CLIV (2003)
Physics Methods in Archaeometry
 edited by M. MARTINI, M. MILAZZO and M. PIACENTINI
- Course CLV (2003)
The Physics of Complex Systems (New Advances and Perspectives)
 edited by F. MALLAMACE and H. E. STANLEY
- Course CLVI (2003)
Research on Physics Education
 edited by E.F. REDISH and M. VICENTINI
- Course CLVII (2003)
The Electron Liquid Model in Condensed Matter Physics
 edited by G. F. GIULIANI and G. VIGNALE
- Course CLVIII (2004)
Hadron Physics
 edited by T. BRESSANI, U. WIEDNER and A. FILIPPI
- Course CLIX (2004)
Background Microwave Radiation and Intrachuster Cosmology
 edited by F. MELCHIORRI and Y. REPHAELI
- Course CLX (2004)
From Nanostructures to Nanosensing Applications
 edited by A. D'AMICO, G. BALESTRINO and A. PAOLETTI
- Course CLXI (2005)
Polarons in Bulk Materials and Systems with Reduced Dimensionality
 edited by G. IADONISI, J. RANNINGER and G. DE FILIPPIS
- Course CLXII (2005)
Quantum Computers, Algorithms and Chaos
 edited by G. CASATI, D. L. SHEPELYANSKY, P. ZOLLER and G. BENENTI
- Course CLXIII (2005)
CP Violation: From Quarks to Leptons
 edited by M. GIORGI, I. MANNELLI, A. I. SANDA, F. COSTANTINI and M. S. SOZZI
- Course CLXIV (2006)
Ultra-Cold Fermi Gases
 edited by M. INGUSCIO, W. KETTERLE and C. SALOMON
- Course CLXV (2006)
Protein Folding and Drug Design
 edited by R. A. BROGLIA, L. SERRANO and G. TIANA
- Course CLXVI (2006)
Metrology and Fundamental Constants
 edited by T. W. HÄNSCH, S. LESCHIUTTA, A. J. WALLARD and M. L. RASTELLO