Multiscale Approaches to Protein Modeling
Andrzej Kolinski
Editor

Multiscale Approaches to Protein Modeling
Preface

Thanks to enormous progress in sequencing of genomic data, presently we know millions of protein sequences. At the same time the number of experimentally solved protein structures is much smaller, ca. 60,000. This is because of large cost of structure determination. Thus, the theoretical in silico prediction of protein structures and dynamics is essential for understanding the molecular basis of drug action, metabolic and signaling pathways in living cells, and designing new technologies in the life science and material sciences. Unfortunately, a “brute force” approach remains impractical. Folding of a typical protein (in vivo or in vitro) takes milliseconds to minutes, while the state-of-the-art all-atom molecular mechanics simulations of protein systems can cover only a time period of nanoseconds to microseconds. This is the reason for the enormous progress in the development of various multiscale modeling techniques applied to protein structure prediction, modeling of protein dynamics and folding pathways, in silico protein engineering, model-aided interpretation of experimental data, modeling of macromolecular assemblies, and theoretical studies of protein thermodynamics. Coarse-graining of the proteins’ conformational space is a common feature of all these approaches, although the details and the underlying physical models span a very broad spectrum.

This book contains comprehensive reviews of the most advanced multiscale modeling methods in protein structure prediction, computational studies of protein dynamics, folding mechanisms, and macromolecular interactions. The presented approaches span a wide range of the levels of coarse-grained representations, various sampling techniques, and a variety of applications to biomedical and biophysical problems. It was our intention to provide a collection of comprehensive reviews that could be used as a reference book for those who just are beginning their adventure with biomacromolecular modeling but also as a valuable source of more detailed information for those who are already experts in the field of biomacromolecular modeling and in related areas of computational biology or biophysics.

Proteins are linear copolymers composed of amino acids. Important ideas of polymer physics inspired the field of protein modeling. Chapter 1 explains some basic concepts of polymer conformational statistics and dynamics of chain molecules in context of simple lattice models. This chapter demonstrates how
these ideas could be employed in protein modeling. Chapter 2 describes application of a lattice-based protein model to the very challenging problem of protein docking. Chapter 3 provides a comprehensive overview of various coarse-grained protein-like and protein models. This chapter describes (among other approaches) probably the most rigorous system of physics-based reduced modeling of proteins. Coarse-grained, multiscale, protein modeling requires specific designs of interaction schemes. Chapters 4–6 provide in-depth overviews of various level force-fields for the reduced representations of protein conformational space, including knowledge-based statistical potentials. Chapters 7 and 8 (but also, in part, Chapters 3–5 and 12) describe a variety of applications of reduced models in the study of protein dynamics, folding pathways, molecular mechanisms of mechanical unfolding, and protein interactions. Chapter 9 gives an overview of the most effective sampling strategies in a reduced, although unrestricted conformational space. Chapters 10 and 11 present a very efficient philosophy of a conformational search, where the target structures are assembled from fragments excised from already known protein structures. These strategies proven to be very effective in the large-scale, automated in silico structure prediction. Chapter 12 describes a multiscale method, based on a high-resolution lattice model, for modeling protein folding pathways. Chapters 13 and 14 discuss the most important ideas and techniques of comparative modeling – the most effective and the most popular method for theoretical prediction of protein structures. These chapters provide also reviews of the model-quality assessment methods.

The contributing authors are world-wide recognized experts. Some of them (Bujnicki and Zhang) are leaders in the field of protein structure prediction, as assessed by the recent (CASP6–CASP8) community-wide experiments in a blind structure prediction. Others also developed very successful methods for the protein structure prediction (Scheraga, Liwo, Feig, and Kihara). Several of the authors of this book developed very efficient coarse-grained interaction schemes for protein models based on either an evolutionary knowledge approach (Jernigan and Scheraga have built theoretical foundations of this class of approaches, but others also contributed significantly: Feig and Micheletti) or a physics-based approach (Scheraga, Liwo, Feig, and Irback). Among the authors are also the world top leaders of comparative modeling (Bujnicki, Zhang, Tramontano, and Kihara) and automated structure prediction (Zhang and Bujnicki) – the structure prediction server created by Zhang is the best till date. The book presents also the state-of-the-art methods of evaluation of quality of the theoretical protein models (Tramontano and Kihara). Recently, a significant progress has been achieved in multiscale modeling of protein dynamics and folding mechanisms. The authors of the chapters dealing with this class of problems are also world-class leaders (Scheraga, Liwo, Irback, Feig, Cieplak, Jernigan, and Micheletti). The conformational search strategies are crucial in protein modeling. Developers of the most efficient computational techniques and strategies are also among the authors (Hansmann, Scheraga, and others).

Warsaw, Poland

Andrzej Kolinski
Contents

1 Lattice Polymers and Protein Models ........................................ 1
   Andrzej Kolinski

2 Multiscale Protein and Peptide Docking .................................. 21
   Mateusz Kurcinski, Michał Jamroz, and Andrzej Kolinski

3 Coarse-Grained Models of Proteins: Theory and Applications .... 35
   Cezary Czaplewski, Adam Liwo, Mariusz Makowski, Stanisław Oldziej, and Harold A. Scheraga

4 Conformational Sampling in Structure Prediction and Refinement with Atomistic and Coarse-Grained Models ............. 85
   Michael Feig, Srinivasa M. Gopal, Kanagasabai Vadivel, and Andrew Stumpff-Kane

5 Effective All-Atom Potentials for Proteins ............................... 111
   Anders Irbäck and Sandipan Mohanty

6 Statistical Contact Potentials in Protein Coarse-Grained Modeling: From Pair to Multi-body Potentials ...................... 127
   Sumudu P. Leelananda, Yaping Feng, Pawel Gniewek, Andrzej Kloczkowski, and Robert L. Jernigan

7 Bridging the Atomic and Coarse-Grained Descriptions of Collective Motions in Proteins ........................................... 159
   Vincenzo Carnevale, Cristian Micheletti, Francesco Pontiggia, and Raffaello Potestio

8 Structure-Based Models of Biomolecules: Stretching of Proteins, Dynamics of Knots, Hydrodynamic Effects, and Indentation of Virus Capsids ..................................................... 179
   Marek Cieplak and Joanna I. Sułkowska

9 Sampling Protein Energy Landscapes – The Quest for Efficient Algorithms ......................................................... 209
   Ulrich H. E. Hansmann
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Protein Structure Prediction: From Recognition of Matches with Known Structures to Recombination of Fragments</td>
<td>231</td>
</tr>
<tr>
<td></td>
<td>Michal J. Gajda, Marcin Pawlowski, and Janusz M. Bujnicki</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Genome-Wide Protein Structure Prediction</td>
<td>255</td>
</tr>
<tr>
<td></td>
<td>Srayanta Mukherjee, Andras Szilagyi, Ambrish Roy, and Yang Zhang</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>Multiscale Approach to Protein Folding Dynamics</td>
<td>281</td>
</tr>
<tr>
<td></td>
<td>Sebastian Kmiecik, Michal Jamroz, and Andrzej Kolinski</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Error Estimation of Template-Based Protein Structure Models</td>
<td>295</td>
</tr>
<tr>
<td></td>
<td>Daisuke Kihara, Yifeng David Yang, and Hao Chen</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Evaluation of Protein Structure Prediction Methods: Issues and Strategies</td>
<td>315</td>
</tr>
<tr>
<td></td>
<td>Anna Tramontano and Domenico Cozzetto</td>
<td></td>
</tr>
<tr>
<td>Index</td>
<td></td>
<td>341</td>
</tr>
</tbody>
</table>
Contributors

Janusz M. Bujnicki Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland; Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland, iamb@genesilico.pl

Vincenzo Carnevale Institute for Computational Molecular Science, Temple University, Philadelphia, PA, USA, vincenzo.carnevale@temple.edu

Hao Chen Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN, USA, chen177@purdue.edu

Marek Cieplak Institute of Physics, Polish Academy of Sciences, Warsaw, Poland, mc@ifpan.edu.pl

Domenico Cozzetto Department of Biochemical Sciences, “Sapienza” University of Rome, Rome, Italy, d.cozzetto@cs.ucl.ac.uk

Cezary Czaplewski Faculty of Chemistry, University of Gdańsk, Gdańsk, Poland; Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, USA, czarek@chem.univ.gda.pl

Michael Feig Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA; Department of Chemistry, Michigan State University, East Lansing, MI, USA, feig@msu.edu

Yaping Feng Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA, USA; L.H.Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA, ypfeng@iastate.edu

Michal J. Gajda European Molecular Biology Laboratories, Hamburg Outstation, Hamburg, Germany, mgajda@embl-hamburg.de

Pawel Gniewek L.H.Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA; Laboratory of Theory of Biopolymers, Faculty of Chemistry, University of Warsaw, Warsaw, Poland, gniewko.pablo@gmail.com
Contributors

**Srinivasa M. Gopal** Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA, srini@msu.edu

**Ulrich H. E. Hansmann** Department of Physics, Michigan Technological University, Houghton, MI, USA, hansmann@mtu.edu

**Anders Irbäck** Computational Biology & Biological Physics, Department of Theoretical Physics, Lund University, Lund, Sweden, anders@thep.lu.se

**Michał Jamroz** Faculty of Chemistry, University of Warsaw, Warsaw, Poland, jamroz@chem.uw.edu.pl

**Robert L. Jernigan** Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA, USA; L.H.Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA, jernigan@iastate.edu

**Daisuke Kihara** Department of Biological Sciences, College of Science; Department of Computer Science, College of Science; Markey Center for Structural Biology, Purdue University, West Lafayette, IN, USA, dkihara@purdue.edu

**Andrzej Kloczkowski** Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA, USA; L.H.Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA, kloczkow@iastate.edu

**Sebastian Kmiecik** Faculty of Chemistry, University of Warsaw, Warsaw, Poland, sebastian.kmiecik@gmail.com

**Andrzej Kolinski** Faculty of Chemistry, University of Warsaw, Warsaw, Poland, kolinski@chem.uw.edu.pl

**Mateusz Kurcinski** Faculty of Chemistry, University of Warsaw, Warsaw, Poland, mkurcinski@gmail.com

**Sumudu P. Leelananda** Department of Biochemistry, Biophysics, and Molecular Biology, Iowa State University, Ames, IA, USA; L.H.Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, IA, USA, sumudu@iastate.edu

**Adam Liwo** Faculty of Chemistry, University of Gdańsk, Gdańsk, Poland; Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, USA, adam@chem.univ.gda.pl

**Mariusz Makowski** Faculty of Chemistry, University of Gdańsk, Gdańsk, Poland; Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, USA, momo@chem.univ.gda.pl
Contributors

**Cristian Micheletti** Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy; Democritos CNR-IOM and Italian Institute of Technology (SISSA Unit), Trieste, Italy, michelet@sissa.it

**Sandipan Mohanty** Jülich Supercomputing Centre, Forschungszentrum Jülich GmbH, Jülich, Germany, s.mohanty@fz-juelich.de

**Srayanta Mukherjee** Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA; Center for Bioinformatics, University of Kansas, Lawrence, KS, USA, srayanta@umich.edu

**Stanisław Ołdziej** Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, USA; Laboratory of Biopolymer Structure, Intercollegiate Faculty of Biotechnology, University of Gdańsk and Medical University of Gdańsk, Gdańsk, Poland, stan@biotech.univ.gda.pl

**Marcin Pawlowski** Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland, marcinp@genesilico.pl

**Francesco Pontiggia** Department of Biochemistry, Brandeis University, Waltham, MA, USA, pontiggi@brandeis.edu

**Raffaello Potestio** Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy, potestio@sissa.it

**Ambrish Roy** Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA; Center for Bioinformatics, University of Kansas, Lawrence, KS, USA, ambryroy@umich.edu

**Harold A. Scheraga** Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY, USA, has5@cornell.edu

**Andrew Stumpff-Kane** Department of Biochemistry and Molecular Biology, Michigan State University, Michigan, USA, stumpff.kane@gmail.com

**Joanna I. Sulikowska** Institute of Physics, Polish Academy of Sciences, Warsaw, Poland; CTBP, University of California, Gilman Drive 9500, La Jolla, San Diego, CA, USA, isulkow@physics.ucsd.edu

**Andras Szilagyi** Center for Bioinformatics, University of Kansas, Lawrence, KS, USA; Institute of Enzymology, BRC, Hungarian Academy of Sciences, Budapest, Hungary, szia@enzim.hu

**Anna Tramontano** Department of Biochemical Sciences, “Sapienza” University of Rome, Rome, Italy; Istituto Pasteur – Fondazione Cenci Bolognetti, “Sapienza” University of Rome, Rome, Italy, anna.tramontano@uniroma1.it

**Kanagasabai Vadivel** Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA, vksabai@msu.edu
Yifeng David Yang Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN, USA, yang41@purdue.edu

Yang Zhang Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA; Center for Bioinformatics, University of Kansas, Lawrence, KS, USA, zhng@umich.edu
Chapter 1
Lattice Polymers and Protein Models

Andrzej Kolinski

Abstract The size of conformational space of chain polymers is enormous. Much has been learned about polymer structure, thermodynamics, and dynamics by theoretical considerations and numerical study of simple lattice models. Self-avoiding random walks on a lattice provide a good approximation for the excluded volume effect and nature of the coil–globule transition. Semiflexible polymers on a lattice exhibit two-state collapse transition that captures some essential features of the all-or-none folding transition of small globular proteins. More complex, decorated with some structural details, lattice polymers provide a very powerful means for study of protein dynamics and thermodynamics and protein structure prediction.

1.1 Reduced Models of Chain Molecules

The torsional rotations, only around the main-chain backbone bonds, make the conformational space of chain molecules enormous in size (Flory 1969). For a chain containing $N$ single bonds, the number of conformations is in the range of $q^N$, where $q$ is approximately equal to the number of distinct low-energy regions of the rotational potential. For a polyethylene chain, $q$ would be 3. Obviously, when $N$ is hundreds or many thousands, a detailed conformational analysis becomes impractical. Impractical are also detailed all-atom computer simulations, unless only very local conformational changes require examination. Thus, in order to make the problem tractable, simplified models have often been designed and studied (Milik et al. 1990; Kolinski and Skolnick 1996), either from statistical analyses or/and by computer simulations. As it will become apparent later, the statistical analysis itself is of rather limited utility and in typical cases requires quite drastic simplifications. Usually, it is difficult to estimate a priori the effect of such simplifications on the final results.

A. Kolinski (✉)
Faculty of Chemistry, University of Warsaw, Warsaw, Poland
e-mail: kolinski@chem.uw.edu.pl
Let us consider two extremely simple models of polymers, one for idealized conformational statistics and the second for the first level of approximation for chain dynamics. These models can be solved rigorously by simple analytical considerations (Flory 1969). The first is the freely jointed chain (sometimes it is also called “the random flight model”). The freely jointed chain (see Fig. 1.1) consists of \( n \) segments of equal length \( l \). Mutual orientations of the segments are completely uncorrelated. It is well known that for a sufficiently large number of segments, the mean-square end-to-end distance \( \langle R^2 \rangle \) of such a chain scales with the number of segments as \( \langle R^2 \rangle = l^2 n \). This result closely resembles the central formula obtained for a Brownian particle theory, where the mean-square displacement is proportional to time. It is also easy to show that the mean-square radius of gyration (a quantity that is easier to measure experimentally than the \( \langle R^2 \rangle \)) is related to \( \langle R^2 \rangle \) as \( \langle S^2 \rangle = \langle R^2 \rangle / 6 \). The distribution of the end-to-end distance and distribution of the segment density is Gaussian. Such an ideal polymer random coil is frequently called the Gaussian chain, although the freely jointed chain is not uniquely Gaussian since other types of chains can also follow Gaussian statistics.

**Fig. 1.1** An example of the freely jointed chain

The simplifications of the physical properties of real polymers assumed in the freely jointed chain model are essentially of two types. First, the correlations between the chain segments, especially between those that are close to one another along the chain contour, are an important property of polymers and strongly depend on their chemical structure. As long as these correlations extend only to a distance small in comparison with the chain length, it is relatively straightforward to generalize the model by introducing various approximation of the local chain stiffness related to sometimes complex profiles of the rotational potential energy. All the short-range (short distance along the chain contour) correlations do not change the general picture. For all such ideal models \( \langle R^2 \rangle = Cl^2 n \), and the value of the prefactor \( C \) depends on the shape of the rotational potential and the temperature.

Approximations of the second type are much more significant and much more difficult to deal with. Namely, all ideal chains neglect the effective interactions between the chain segments that are far away from one another along the chain
but close to each other in space. On the most trivial level, the fact that two segments
cannot occupy the same element of space must be taken into account. A rigorous
analytical treatment of such “real” chains is not possible, although approximate the-
ories exist (de Gennes 1979). Probably, the most famous is Flory’s mean-field theory
(Flory 1953). The theory assumes that a balance between intramolecular interactions
and those with solvent defines the average coil size. A quasi-chemical approxima-
tion is employed and an average Gaussian density of segments is assumed. The
resulting formula describes the chain dimension as a function of temperature:

\[ \alpha^5 - \alpha^3 = \text{const.} \left(1 - \frac{\Theta}{T}\right) n^{1/2} \]  

where \( \alpha \) is the so-called expansion factor and is defined as

\[ \alpha^2 = \frac{\langle R^2 \rangle}{\langle R_0^2 \rangle} \]

with \( \langle R_0^2 \rangle \) denoting the ideal chain dimensions.

Note that for \( T = \Theta \) the chain dimensions become identical with the dimen-
sions of the ideal chain. Thus, idea behind Flory’s “theta” (\( \Theta \)) temperature closely
resembles the Boyle temperature for real gases. At temperatures below \( \Theta \), the chain
undergoes a transition to a dense globular state, and this transition is somewhat sim-
ilar to the gas–liquid transition of small molecule systems. However, the transition
for flexible polymers is continuous and has most of the features of a second-
order phase transition (Kolinski et al. 1987b). At high temperatures (see Eq. (1.1))
\( \langle R^2 \rangle \sim n^{6/5} \), and the average chain dimensions are much larger than for an equiva-
 lent ideal chain. Interestingly, despite a rather poor estimation of chain entropy and
internal energy, Flory’s theory gives quite an accurate estimation of the free energy
and conformational properties of chain molecules. Such a cancellation of errors is
quite typical of mean-field-type theories.

Ideal chain statistics provides a zero-order picture of the protein denatured state,
while Flory’s theory is a zero-order approximation for the folding (or collapse) tran-
sition. The approximation is quite crude for several reasons. First, protein chains are
relatively stiff polymers and the limit of infinitely long chains is hardly satisfied
even for large proteins (Creighton 1993). Second, proteins are heteropolymers with
highly specific patterns of intramolecular interactions (Branden and Tooze 1991).
Even in the random coil state, there is a significant extent of residual structure.
Thus, the mean-field theory is hardly applicable. We will address these issues later
in more detail.

Somewhat analogous to ideal chain statistics, models for ideal chain dynamics
were designed. Probably the best known of these is the Rouse model (Rouse 1953),
shown in a schematic fashion in Fig. 1.2. It assumes that a flexible polymer chain
can be represented as a chain of points joined by harmonic springs of equal strength.
This model is analytically solvable. The results are quite interesting. For short times,
when the average displacements of chain segments must be small in comparison
with the coil size a single segment moves according to