

RSC Biomolecular Sciences

Edited by Tamar Schlick

Innovations in Biomolecular Modeling and Simulations

Volume 1



RSC Publishing

Innovations in Biomolecular Modeling and Simulations Volume 1

Edited by

Tamar Schlick

*Department of Chemistry, Courant Institute of Mathematical Sciences,
New York University, New York, NY 10012, USA*



RSC Publishing

RSC Biomolecular Sciences No. 23

ISBN: 978-1-84973-461-5

ISSN: 1757-7152

A catalogue record for this book is available from the British Library

© Royal Society of Chemistry 2012

All rights reserved

Apart from fair dealing for the purposes of research for non-commercial purposes or for private study, criticism or review, as permitted under the Copyright, Designs and Patents Act 1988 and the Copyright and Related Rights Regulations 2003, this publication may not be reproduced, stored or transmitted, in any form or by any means, without the prior permission in writing of The Royal Society of Chemistry or the copyright owner, or in the case of reproduction in accordance with the terms of licences issued by the Copyright Licensing Agency in the UK, or in accordance with the terms of the licences issued by the appropriate Reproduction Rights Organization outside the UK. Enquiries concerning reproduction outside the terms stated here should be sent to The Royal Society of Chemistry at the address printed on this page.

The RSC is not responsible for individual opinions expressed in this work.

Published by The Royal Society of Chemistry,
Thomas Graham House, Science Park, Milton Road,
Cambridge CB4 0WF, UK

Registered Charity Number 207890

For further information see our web site at www.rsc.org

Printed in the United Kingdom by Henry Ling Limited, at the Dorset Press, Dorchester,
DT1 1HD

Innovations in Biomolecular Modeling and Simulations
Volume 1

RSC Biomolecular Sciences

Editorial Board:

Professor Stephen Neidle (Chairman), *The School of Pharmacy, University of London, UK*

Dr Marius Clore, *National Institutes of Health, USA*

Professor Roderick E Hubbard, *University of York and Vernalis, Cambridge, UK*

Professor David M J Lilley FRS, *University of Dundee, UK*

Titles in the Series:

- 1: Biophysical and Structural Aspects of Bioenergetics
- 2: Exploiting Chemical Diversity for Drug Discovery
- 3: Structure-based Drug Discovery: An Overview
- 4: Structural Biology of Membrane Proteins
- 5: Protein–Carbohydrate Interactions in Infectious Disease
- 6: Sequence-specific DNA Binding Agents
- 7: Quadruplex Nucleic Acids
- 8: Computational and Structural Approaches to Drug Discovery: Ligand–Protein Interactions
- 9: Metabolomics, Metabonomics and Metabolite Profiling
- 10: Ribozymes and RNA Catalysis
- 11: Protein–Nucleic Acid Interactions: Structural Biology
- 12: Therapeutic Oligonucleotides
- 13: Protein Folding, Misfolding and Aggregation: Classical Themes and Novel Approaches
- 14: Nucleic Acid–Metal Ion Interactions
- 15: Oxidative Folding of Peptides and Proteins
- 16: RNA Polymerases as Molecular Motors
- 17: Quantum Tunnelling in Enzyme-Catalysed Reactions
- 18: Natural Product Chemistry for Drug Discovery
- 19: RNA Helicases
- 20: Molecular Simulations and Biomembranes: From Biophysics to Function
- 21: Structural Virology
- 22: Biophysical Approaches Determining Ligand Binding to Biomolecular Targets: Detection, Measurement and Modelling
- 23: Innovations in Biomolecular Modeling and Simulations: Volume 1

How to obtain future titles on publication:

A standing order plan is available for this series. A standing order will bring delivery of each new volume immediately on publication.

For further information please contact:

Book Sales Department, Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge, CB4 0WF, UK

Telephone: +44 (0)1223 420066, Fax: +44 (0)1223 420247, Email: books@rsc.org

Visit our website at <http://www.rsc.org/Shop/Books/>

Preface

Modelers to the Rescue

“The purpose of models is not to fit the data but to sharpen the questions.”

—Samuel Karlin, 1983.

We are in an era of data deluge. Merchants know all about what we eat, drink, wear, and read; doctors subject us to numerous tests and procedures to extract numbers and images that are intended to capture our state of health; and large-scale genome sequencing companies are producing so much data that computer disks via FedEx, rather than Internet downloads, are the only viable mode of data transfer!¹ No wonder cartoonist Chris Wildt drew men and women of science pouring over paper files, computer screens, and endless diagrams before giving up: “Well, we’ve seen all the data and crunched the numbers... heads or tails?”

Sifting through biological and chemical data to extract knowledge, and eventually an understanding, is a task attributed today to bioinformaticians and other computational scientists through data mining approaches and computer modeling and simulation. As recently surveyed in a field perspective,² the trajectory of biomolecular modeling and simulation has evolved from initial euphoria and unrealistic expectations in the 1980s to a more mature, cautiously optimistic viewpoint bolstered by demonstrated successes (Figure P.1). Clearly, biomolecular modeling is not likely to supplant experimentation in the near future for predicting and understanding biomolecular structure and function; however, iterative program design between experiment and theory is becoming a reality for advancing the field. The problem of early unrealistic expectation and hype was not so much due to the inherent imperfections, approximations,

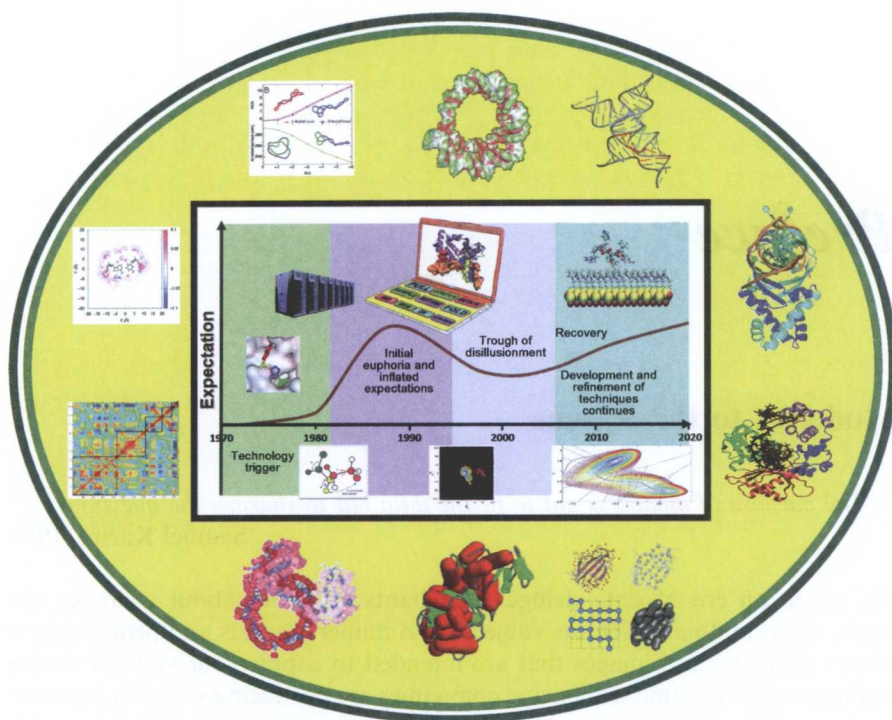


Figure P.1 Proposed expectation curve for the field of biomolecular modeling and simulation, with approximate timeline,² surrounded by representative images from work collected in these volumes. The field started when comprehensive molecular mechanics efforts began, and it took off with the increasing availability of fast workstations and later supercomputers. Following unrealistically high expectations and disappointments, the field is well on its fruitful trajectory of realistic progress and synergy between theory and experiment. The images taken from contributing chapters in this two-volume book are, *inside the box*: bottom right: contour plots connecting potential energy basins (A. E. Cardenas *et al.*); bottom middle: essential dynamics projection for HU protein/DNA junction complex (Wheatley *et al.*); top right: peptides on microarray surface (J. Feng *et al.*); and left middle: fragment grown for drug discovery applications (Z. Liu *et al.*). Outside the box, images clockwise from bottom right to top right are: proteins at various levels of resolution (N. Ceres *et al.*); coarse-grained model of the molecular motor myosin (E. R. Taylor); nucleosome chains (K. Rippe *et al.*); covariance matrix for protein/DNA/dNTP complex in DNA polymerases (M. Foley *et al.*); solvent charge density in an AT base pair plane (T. Luchko *et al.*); trefoil knot ensemble (G. Witz *et al.*); electrostatic potential of the nucleosome (R. C. Harris *et al.*); junction architecture in hairpin ribozyme (D. Lilley); structure of architectural protein HU bound to DNA (Wheatley *et al.*); and structure of DNA polymerase λ (M. Foley *et al.*).

and limitations of the models used, but rather an underestimation of the complexities of biological systems.

As we uncover the multi-layered levels of biological regulation, we are only beginning to understand what we actually do not understand. For this reason, modeling requires as much art as science in developing and appropriately applying, analyzing, and relating simulation results to the biological systems and associated experimental observations.

In this two-volume book, *Innovations in Biomolecular Modeling and Simulations*, twenty-six teams of leading scientists in biomolecular structure and modeling present the state-of-the-art in the field, from both a methodological and application perspective.

Volume 1 describes recent developments in force fields and electrostatics, conformational sampling and rate calculations, and coarse grained modeling and multiscale models.

Volume 2 describes atomistic simulations of nucleic acids and nucleic acid complexes; coarse-grained modeling approaches to study large-scale DNA folding, knotting, sliding and hopping; and the latest approaches and achievements in structure-based drug design.

In more detail, Volume 1 begins with Scheraga's personal perspective of the field; having worked in the field of protein structure and molecular biophysics for 65 years, Scheraga's trajectory from mainly experimental to mostly theoretical work on conformational energy calculations of polypeptides and proteins is an inspiration. The second historical perspective is contributed by writer Pollack, who tells the story of the emergence of the freely available molecular dynamics program NAMD from the dream of one immigrant, Klaus Schulten, who arrived to America in 1987 with a supercomputer in his backpack. Through interviews and studies of other pioneers, she weaves in history, science, and personal stories to present a fascinating historical account of the early days of molecular dynamics taking us to current-day applications.

Turning into the force-field advances section of Volume 1, MacKerell and co-workers describe one of the most important areas of current development, namely incorporating the polarizable component of electrostatic interactions into modern force fields; they also describe procedures for parameterizing these forms and testing them on model systems as well as highly-charged nucleic acids, where improvements can be demonstrated. Case and co-workers continue discussions of the fundamental electrostatics interactions by focusing on approaches using integral equations to model the interactions of biomolecules with water and ions and improve upon other implicit-solvation techniques such as Poisson Boltzmann approximations. Smith and co-workers share their related methodological advances using both classical molecular mechanics and quantum-mechanical techniques to study complex systems of interest to the energy bioscience field using supercomputing resources; implicit solvation and efficient scaling on large-scale machines are required to study behavior of multi-million-atom systems.

The second group of papers in Volume 1 describes the latest algorithms to survey conformational changes, associated pathways, and reactions rates from

the molecular simulations. As reviewed recently,^{3,4} various clever stochastic as well as deterministic approaches have been developed and applied in recent years to increase the variety of configurations captured, encourage transitions, and extract essential quantitative information on those rare events. The methods include Milestoning and Forward Flux methods based on trajectory fragments, as described by Elber and co-workers, and conformational dynamics and Markov state models as detailed by Darve and co-workers. The alternative Monte Carlo approach for large systems is described by Mezei, who discusses special implementational considerations for macromolecules.

While atomistic models are essential for describing detailed events at the fine level, coarse-grained models are required to describe events relevant to very large complex systems like the chromatin fiber or major rearrangements involving topological changes such as DNA knotting. For such problems, simplified representations are needed, and the challenge is not only to construct useful models that approximate many aspects of the systems, while detailing others, but to eventually telescope between the macroscopic and microscopic views. Examples of such approaches are described by Lavery and coworkers for proteins, by Taylor and co-workers for the molecular motor myosin, by Radhakrishnan and co-workers for protein-protein and protein-membrane networks, by Pettitt and co-workers for complexes at material surfaces, and by Fish and co-workers in the context of large-scale industrial applications.

Volume 2 begins with state-of-the-art examples of atomistic simulations of nucleic acids and nucleic acid complexes. Lankas describes the nuts and bolts of atomistic simulations of DNA in explicit solvent with a viewpoint into coarse-grained models and a focus on understanding sequence and anharmonic effects. Neidle and co-workers describe applications to quadruplex systems of nucleic acids, their complex hierarchical arrangements as studied by classical dynamics, enhanced sampling methods, and hybrid classical/quantum approaches, as well their relevance for supplementing current static structure resolution approaches. Rohs, Fenley and co-workers discuss electrostatic and structural aspects of DNA that have tailored DNA for interactions with proteins, as deciphered from atomistic studies of nucleic acids and their complexes. Schlick and co-workers describe insights into DNA polymerase repair mechanisms as deduced from atomistic dynamics and principal component analysis of several polymerase enzymes; the resulting correlated motions suggest mechanisms that are hybrid between traditional induced fit and conformational selections models. Mukerji, Beveridge and co-workers report their latest study for a structural prediction from molecular dynamics of a novel protein/DNA complex involving a four-way junction.

Continuing to RNA at atomic resolution, Sponer and co-workers describe the state-of-the-art in simulating RNAs by all-atom molecular dynamics, a field that has lagged behind RNA's famous cousin due to force-field limitations, among others. Lilley then describes the fascinating and hierarchical folding of RNA junctions, important since junctions organize the global architecture of RNA.

Events that require different kinds of methods are then investigated in the second part of Volume 2. Witz and Stasiak describe studies of DNA knots and catenanes using topological frameworks explored by Monte Carlo and Brownian dynamics simulations. Rippe and co-workers describe Monte Carlo simulations of nucleosome chains using simple coarse-grained models that capture essential features of the systems. Levy and Marcovitz then describe a study of DNA sliding on the genomic level.

The book concludes with overviews of drug design approaches, from both industrial (Singh and co-workers) and academic perspectives (Amaro and co-workers, McCammon). McCammon's personal perspective focuses on the identification of two anti-viral drug targets for AIDS using a structure and dynamics-based approach.

The range of subjects and methods as reflected by these exciting contributions indicate not only the ingenuity of the methods and wide range of tools available to the molecular modeler, but also the endless questions that arise from the modeling. These questions, in turn, serve to advance the field toward an ultimate understanding of the complex interaction between biomolecular structure and function.

The data deluge is unlikely to cease in the near future. Nonetheless, our modeling approaches will undoubtedly become better and more diverse, allowing us to tackle yet more challenging problems and unravel complex relationships, patterns, and long-time events in biomolecules.

Tamar Schlick
New York

References

1. A. Pollack, DNA Sequencing Caught in Deluge of Data, 30 Nov., *The New York Times*, 2011, page C1.
2. T. Schlick, R. Collepardo-Guevara, L. A. Halvorsen, S. Jung and X. Xiao, Biomolecular Modeling and Simulation: A Field Coming of Age. *Quart. Rev. Biophys.*, 2011, **44**, 191–228.
3. T. Schlick, Monte Carlo, Harmonic Approximation, and Coarse-Graining Approaches for Enhanced Sampling of Biomolecular Structure. *F1000 Biol. Rep.*, 2009, **1**, 48.
4. T. Schlick, Molecular-Dynamics Based Approaches for Enhanced Sampling of Long-Time, Large-Scale Conformational Changes in Biomolecules. *F1000 Biol. Rep.*, 2009, **1**, 51.

Contents

Volume 1

Beginnings

Chapter 1	Personal Perspective	3
	<i>Harold A. Scheraga</i>	
	References	6
Chapter 2	Fashioning NAMD, a History of Risk and Reward: Klaus Schulten Reminisces	8
	<i>Lisa Pollack</i>	
2.1	Introduction	8
2.2	Early Influences of Molecular Dynamics	9
2.3	Building a Parallel Computer	10
2.4	Transporting a Supercomputer	12
2.5	Rejection	13
2.6	Collaborations with Computer Scientists	13
2.7	Graduate Students Revolt	15
2.8	Factors that Shaped NAMD	15
2.9	Continual Evolution of NAMD	16
2.10	Conclusions	17
	Acknowledgements	18
	References	19

RSC Biomolecular Sciences No. 23

Innovations in Biomolecular Modeling and Simulations: Volume 1

Edited by Tamar Schlick

© Royal Society of Chemistry 2012

Published by the Royal Society of Chemistry, www.rsc.org

Force Fields and Electrostatics

Chapter 3	Towards Biomolecular Simulations with Explicit Inclusion of Polarizability: Development of a CHARMM Polarizable Force Field based on the Classical Drude Oscillator Model	23
	<i>C. M. Baker, E. Darian and A. D. MacKerell Jr</i>	
3.1	Introduction	23
3.2	Polarizable Force Fields	26
3.3	The CHARMM Drude Polarizable Force Field	27
3.3.1	The Model	27
3.3.2	Parametrization Strategy	33
3.3.3	Biomolecular Building Blocks	38
3.3.4	Case Study: The Nucleic Acids	41
3.4	Conclusions	45
	Acknowledgements	46
	References	46
Chapter 4	Integral Equation Theory of Biomolecules and Electrolytes	51
	<i>Tyler Luchko, In Suk Joung and David A. Case</i>	
4.1	Introduction	51
4.1.1	Some History	52
4.2	Integral Equation Theory for Simple Liquids	54
4.2.1	Ornstein–Zernike Equation	54
4.2.2	Closures	55
4.2.3	1D-RISM	56
4.2.4	3D-RISM	58
4.2.5	Long Range Asymptotics	59
4.2.6	Solvation Free Energy	61
4.2.7	Numerical Approaches	62
4.3	Pure Solvent Structure and Thermodynamics	63
4.4	Macromolecules and Other Solutes	69
4.4.1	Ion Binding to Crown Ethers	69
4.4.2	Ion Binding to Proteins	72
4.4.3	Ion Atmosphere Around Duplex DNA	73
4.4.4	Solvent Distributions	75
4.5	Conclusions	76
4.6	Methods	78
4.6.1	Parameters	78
4.6.2	Bulk Electrolytes	79
4.6.3	Crown-ether	79
4.6.4	DNA	79
	Acknowledgements	80
	References	80

Chapter 5	Molecular Simulation in the Energy Biosciences	87
	<i>Xiaolin Cheng, Jerry M. Parks, Loukas Petridis, Benjamin Lindner, Roland Schulz, Hao-Bo Guo, Goundla Srinivas and Jeremy C. Smith</i>	
5.1	Introduction	87
5.2	Methodological Development: An Order N Poisson-Boltzmann Solver	88
5.2.1	A Tree Code-Based Generalized Born (tGB) Method	91
5.2.2	Scaling of a Multimillion-atom Lignocellulose Simulation on a Petascale Supercomputer	95
5.2.3	Comparison of Simulations with Different Electrostatic Methods	96
5.2.4	Scaling	97
5.3	Applications	97
5.3.1	Biofuel Research	97
5.3.2	Bioremediation: Mechanisms of Bacterial Mercury Resistance	104
5.4	Conclusions	111
	Acknowledgements	111
	References	112

Sampling and Rates

Chapter 6	Enhancing the Capacity of Molecular Dynamics Simulations with Trajectory Fragments	117
	<i>Alfredo E. Cardenas and Ron Elber</i>	
6.1	Introduction	117
6.2	Trajectory Fragments: General Overview	120
6.3	Challenges in Using Trajectory Fragments	123
6.3.1	Milestoning	123
6.3.2	Forward Flux Sampling (FFS)	126
6.4	Applications of Trajectory Fragments: Equilibrium and Kinetics	129
6.5	Examples	132
6.6	Conclusions	135
	References	136
Chapter 7	Computing Reaction Rates in Bio-molecular Systems Using Discrete Macro-states	138
	<i>Eric Darve and Ernest Ryu</i>	
7.1	Introduction	138
7.2	Transition Path Sampling	144

7.2.1	Reactive Flux and Transition State Theory	144
7.2.2	Transition Path Sampling	146
7.2.3	Transition Interface Sampling	148
7.2.4	Forward Flux Sampling	149
7.2.5	Milestoning	151
7.2.6	Milestoning Using Optimal Milestones	152
7.3	Conformation Dynamics and Markov State Models	156
7.3.1	Conformation Dynamics	156
7.3.2	Perron Cluster Cluster Analysis	157
7.3.3	Markov State Models	160
7.4	Non-equilibrium Umbrella Sampling and Reactive Trajectory Sampling	163
7.4.1	Non-equilibrium Umbrella Sampling	163
7.4.2	Reactive Trajectory Sampling	165
7.4.3	Optimal Cells	170
7.4.4	Metastability, Convergence, and the Multi-colored Algorithm	172
7.5	Analysis of Statistical Errors in Markov State Models	174
7.5.1	Eigenvectors and Eigenvalues of the Transition Matrix	175
7.5.2	Sensitivity of Eigenvalues and Systematic Errors	178
7.5.3	Statistical Errors	182
7.6	Some Numerical Benchmarks for Markov State Models and Reactive Trajectory Sampling	186
7.6.1	Set up of Numerical Simulations	186
7.6.2	Numerical Benchmarks	188
7.7	Conclusion	193
7.8	Technical Proofs	195
	Acknowledgements	197
	References	197

Chapter 8 Challenges in Applying Monte Carlo Sampling to Biomolecular Systems **207**

M. Mezei

8.1	Introduction	207
8.2	Basic Ideas of Monte Carlo Sampling	208
8.3	Why ‘Bother’ with Monte Carlo?	209
8.4	Correlated Moves	210
8.5	Cooperative Potentials	212
8.6	Long-range Energy Contributions	212
8.7	Parallelization	213
8.8	Conclusion	215
	References	215

Coarse Graining and Multiscale Models

Chapter 9	Coarse-grain Protein Models	219
	<i>N. Ceres and R. Lavery</i>	
9.1	Introduction	219
9.2	Coarse-grain Representations	220
9.2.1	All-atom and United-atom Models	220
9.2.2	Bead Models	220
9.2.3	Lattice Models	222
9.2.4	Polymer and Material Models	222
9.3	Formulating the Energy	222
9.3.1	Gō Potentials	222
9.3.2	Elastic Network Models	223
9.3.3	Statistical Potentials	225
9.3.4	Physics-based Potentials	228
9.3.5	Hybrid Potentials	229
9.3.6	Solvent	229
9.4	Applications: From Folding to Function	230
9.4.1	Protein Folding	230
9.4.2	Structure Prediction	232
9.4.3	Mechanical and Dynamical Properties	234
9.4.4	Protein-protein Interactions	235
9.4.5	Large Protein Assemblies	237
9.5	Conclusions	238
	Acknowledgements	239
	References	239
Chapter 10	Generalised Multi-level Coarse-grained Molecular Simulation and its Application to Myosin-V Movement	249
	<i>William R. Taylor and Zoe Katsimitsoulia</i>	
10.1	Introduction	249
10.2	Implementation	251
10.2.1	Overview	251
10.2.2	Steric Exclusion	252
10.2.3	Polymers and Cross-linking	253
10.2.4	Geometric Regularisation	255
10.2.5	Shape Specification	256
10.2.6	Implementation	256
10.3	Example Application	257
10.3.1	Model Construction	258
10.3.2	Driver Construction	264
10.3.3	Simulation	268

10.4	Conclusions	269
	Acknowledgements	270
	References	270
Chapter 11	Top-down Mesoscale Models and Free Energy Calculations of Multivalent Protein-Protein and Protein-Membrane Interactions in Nanocarrier Adhesion and Receptor Trafficking	272
	<i>Jin Liu, Neeraj J. Agrawal, David M. Eckmann, Portonovo S. Ayyaswamy and Ravi Radhakrishnan</i>	
11.1	Introduction	273
11.2	Mesoscale Model for Functionalized Nanocarrier Targeting to Cells in Drug Delivery	274
11.2.1	Parameter Estimation in the Top-down Modeling Strategy	276
11.2.2	Calculation of the Absolute Binding Free Energy	276
11.2.3	Calculation of Potential of Mean Force (PMF) of NC Binding to EC	278
11.2.4	Comparison with Experiments	279
11.3	Mesoscale Model for Protein-Induced Membrane Curvature	280
11.3.1	Membrane Curvature Model and Conformational Sampling	281
11.3.2	Calculation of Relative Free Energies of Curved Membranes	283
11.3.3	Application to Clathrin-Mediated Endocytosis	285
11.4	Future Outlook	286
	Acknowledgements	287
	References	287
Chapter 12	Studying Proteins and Peptides at Material Surfaces	293
	<i>Jun Feng, Gillian C. Lynch and B. Montgomery Pettitt</i>	
12.1	Introduction	293
12.2	Experimental Surface Chemistry and Immobilization Strategies	295
12.3	Surface Design	297
12.4	Theoretical and Simulation Studies of Proteins at Surfaces	303
12.5	Concluding Remarks	308
	Acknowledgements	308
	References	309

Chapter 13	Multiscale Design: From Theory to Practice	321
	<i>J. Fish, V. Filonova and Z. Yuan</i>	
13.1	Introduction	321
13.2	RC ² Formulation	323
13.2.1	Computational Continua	323
13.2.2	Residual-free Computational Unit Cell Problem	324
13.2.3	The Coarse-scale Weak Form	330
13.2.4	Coarse-scale Discretization	332
13.3	Model Verification	333
13.4	Multiscale Characterization	336
13.4.1	Formulation of an Inverse Problem	336
13.4.2	Characterization of Model Parameters in Reduced Order Computational Continua	339
13.4.3	Multistep Model Characterization	340
13.5	Multiscale Design System Architecture	341
13.6	Conclusions	342
	Acknowledgements	343
	References	343
Subject Index		345

Volume 2

Atomistic Simulations of Nucleic Acids and Nucleic Acid Complexes

Chapter 1	Modelling Nucleic Acid Structure and Flexibility: From Atomic to Mesoscopic Scale	3
	<i>Filip Lankaš</i>	
1.1	Introduction	3
1.2	Models Based on Pseudoatoms	4
1.3	Rigid Base and Basepair Models	8
1.3.1	Fitting the Base Frame	11
1.3.2	Internal Coordinates	13
1.3.3	Nonlocal Models	15
1.3.4	Example: DNA A-tract	18
1.3.5	Anharmonic Effects	23
1.4	Coarser Models and Other Approaches	25
1.5	Concluding Remarks	27
	Acknowledgements	28
	References	28