Protein Structure Prediction

METHODS IN MOLECULAR BIOLOGY™

John M. Walker, SERIES EDITOR

419. Post-Transcriptional Gene Regulation, edited by *Jeffrey Wilusz, 2008*
418. Avidin–Biotin Interactions: *Methods and Applications,* edited by *Robert J. McMahon, 2008*
417. Tissue Engineering, Second Edition, edited by *Hannsjörg Hauser and Martin Fussenegger, 2007*
416. Gene Essentiality: *Protocols and Bioinformatics,* edited by *Andrei L. Osterman, 2008*
415. Innate Immunity, edited by *Jonathan Ewbank and Eric Vivier, 2007*
414. Apoptosis in Cancer: *Methods and Protocols*, edited by *Gil Mor and Ayesha Alvero, 2008*
413. Protein Structure Prediction, Second Edition, edited by *Mohammed Zaki and Chris Bystroff, 2008*
412. Neutrophil Methods and Protocols, edited by *Mark T. Quinn, Frank R. DeLeo, and Gary M. Bokoch, 2007*
411. Reporter Genes for Mammalian Systems, edited by *Don Anson, 2007*
410. Environmental Genomics, edited by *Cristofre C. Martin, 2007*
409. Immunoinformatics: *Predicting Immunogenicity* In Silico, edited by *Darren R. Flower, 2007*
408. Gene Function Analysis, edited by *Michael Ochs, 2007*
407. Stem Cell Assays, edited by *Mohan C. Vemuri, 2007*
406. Plant Bioinformatics: *Methods and Protocols,* edited by *David Edwards, 2007*
405. Telomerase Inhibition: *Strategies and Protocols,* edited by *Lucy Andrews and Trygve O. Tollefsbol, 2007*
404. Topics in Biostatistics, edited by *Walter T. Ambrosius, 2007*
403. Patch-Clamp Methods and Protocols, edited by *Peter Molnar and James J. Hickman, 2007*
402. PCR Primer Design, edited by *Anton Yuryev, 2007*
401. Neuroinformatics, edited by *Chiquito J. Crasto, 2007*
400. Methods in Lipid Membranes, edited by *Alex Dopico, 2007*
399. Neuroprotection Methods and Protocols, edited by *Tiziana Borsello, 2007*
398. Lipid Rafts, edited by *Thomas J. McIntosh, 2007*
397. Hedgehog Signaling Protocols, edited by *Jamila I. Horabin, 2007*
396. Comparative Genomics, *Volume 2*, **edited by** *Nicholas H. Bergman, 2007*
395. Comparative Genomics, *Volume 1*, **edited by** *Nicholas H. Bergman, 2007*
394. Salmonella: *Methods and Protocols,* edited by *Heide Schatten and Abe Eisenstark, 2007*
393. Plant Secondary Metabolites, edited by *Harinder P. S. Makkar, P. Siddhuraju, and Klaus Becker, 2007*
392. Molecular Motors: *Methods and Protocols*, edited by *Ann O. Sperry, 2007*
391. MRSA Protocols, edited by *Yinduo Ji, 2007*
390. Protein Targeting Protocols, Second Edition, edited by *Mark van der Giezen, 2007*
389. *Pichia* Protocols, Second Edition, edited by *James M. Cregg, 2007*
388. Baculovirus and Insect Cell Expression Protocols, Second Edition, edited by *David W. Murhammer, 2007*
387. Serial Analysis of Gene Expression (SAGE): *Digital Gene Expression Profiling,* edited by *Kare Lehmann Nielsen, 2007*
386. Peptide Characterization and Application Protocols, edited by *Gregg B. Fields, 2007*
385. Microchip-Based Assay Systems: *Methods and Applications,* edited by *Pierre N. Floriano, 2007*
384. Capillary Electrophoresis: *Methods and Protocols,* edited by *Philippe Schmitt-Kopplin, 2007*
383. Cancer Genomics and Proteomics: *Methods and Protocols,* edited by *Paul B. Fisher, 2007*
382. Microarrays, Second Edition: *Volume 2, Applications and Data Analysis,* edited by *Jang B. Rampal, 2007*
381. Microarrays, Second Edition: *Volume 1, Synthesis Methods,* edited by *Jang B. Rampal, 2007*
380. Immunological Tolerance: *Methods and Protocols,* edited by *Paul J. Fairchild, 2007*
379. Glycovirology Protocols, edited by *Richard J. Sugrue, 2007*
378. Monoclonal Antibodies: *Methods and Protocols,* edited by *Maher Albitar, 2007*
377. Microarray Data Analysis: *Methods and Applications,* edited by *Michael J. Korenberg, 2007*
376. Linkage Disequilibrium and Association Mapping: *Analysis and Application,* edited by *Andrew R. Collins, 2007*
375. In Vitro Transcription and Translation Protocols: *Second Edition,* edited by *Guido Grandi, 2007*
374. Quantum Dots: edited by *Marcel Bruchez and Charles Z. Hotz, 2007*
373. Pyrosequencing® Protocols, edited by *Sharon Marsh, 2007*
372. Mitochondria: Practical Protocols, edited by *Dario Leister and Johannes Herrmann, 2007*
371. Biological Aging: *Methods and Protocols,* edited by *Trygve O. Tollefsbol, 2007*
370. Adhesion Protein Protocols*, Second Edition,* edited by *Amanda S. Coutts, 2007*
369. Electron Microscopy: *Methods and Protocols, Second Edition,* edited by *John Kuo, 2007*
368. Cryopreservation and Freeze-Drying Protocols, *Second Edition,* edited by *John G. Day and Glyn Stacey, 2007*
367. Mass Spectrometry Data Analysis in Proteomics, edited by *Rune Matthiesen, 2007*
366. Cardiac Gene Expression: *Methods and Protocols,* edited by *Jun Zhang and Gregg Rokosh, 2007*
365. Protein Phosphatase Protocols: edited by *Greg Moorhead, 2007*

METHODS IN MOLECULAR BIOLOGY™

Protein Structure Prediction

Second Edition

Edited by

Mohammed J. Zaki
and
Christopher Bystroff

Rensselaer Polytechnic Institute, Troy, New York, USA

HUMANA PRESS TOTOWA, NEW JERSEY

999 Riverview Drive, Suite 208
Totowa, New Jersey 07512

www.humanapress.com

This publication is printed on acid-free paper. ♾
ANSI Z39.48-1984 (American Standards Institute) Permanence of Paper for Printed Library Materials

Production Editor: Rhukea Hussain
Cover design by Karen Schulz

For additional copies, pricing for bulk purchases, and/or information about other Humana titles, contact Humana at the above address or at any of the following numbers: Tel.: 973-256-1699; Fax: 973-256-8341; E-mail: humana@humanapr.com; or visit our Website: www.humanapress.com

Printed in the United States of America. 10 9 8 7 6 5 4 3 2 1
eISBN 978-1-59745-574-9

Library of Congress Control Number: 2007933144

Preface

For 40 years we have known the essential ingredients for protein folding – an amino acid sequence and water. But the problem of predicting the three-dimensional structure from its sequence has eluded computational biologists even in the age of supercomputers and high throughput structural genomics. Will we ever solve the "protein folding problem", or will we simply settle for a solution to the "protein prediction problem"? This book covers elements of both the data-driven comparative modeling approach to structure prediction and also recent attempts to simulate folding using explicit or simplified models. Despite the unsolved mystery of how a protein folds, advances are being made in predicting the interactions of proteins with other molecules, such as small ligands, nucleic acids, or other proteins. Also, rapidly advancing are the methods for solving the inverse folding problem, the problem of finding a sequence to fit a structure. This book focuses on the various computational methods for prediction, their successes, and their limitations, from the perspective of their most well-known practitioners. An overview of the chapters in this volume is given below.

Overview of Protein Structure Prediction

In the first chapter, entitled "A Historical Perspective of Template-Based Protein Structure Prediction," Jun-tao Guo, Kyle Ellrott, and Ying Xu give a comprehensive, as well as historical, account of protein structure prediction. They touch upon methods spanning threading, fold recognition, homology modeling, ab initio methods, and their hybrids. They also discuss recent progress in the worldwide blind structure prediction evaluation experiments like CASP and its cousin for automated servers, CAFASP.

In the second chapter "The Assessment of Methods for Protein Structure Prediction," Anna Tramontano, Domenico Cozzetto, Alejandro Giorgetti, and Domenico Raimondo take a critical look at extant methods for protein structure prediction and assess how well they perform. They focus on automatic assessment methods as well as the CASP challenges and discuss their limitations and trade-offs.

Template-Based Methods

In the third chapter "Aligning Sequences to Structures," Liam J. McGuffin discusses the current approaches to template-based fold prediction. The goal here is to align new protein sequences to library of known/template folds. Liam also shows a step-by-step guide to template alignment.

In the fourth chapter "Protein Structure Prediction Using Threading," Jinbo Xu, Feng Jiao, and Libo Yu discuss approaches for protein threading. After setting up the general requirements for protein structure prediction by threading, they specifically focus on their successful new method called RAPTOR, which combines linear programming with machine learning approaches.

Structure Alignment and Indexing

In the fifth chapter "Algorithms for Multiple Protein Structure Alignment and Structure-Derived Multiple Sequence Alignment," Maxim Shatsky, Ruth Nussinov, and Haim J. Wolfson, present methods to recognize the structural core common to a set of proteins through multiple structure alignment. They also discuss how to align multiple sequences with knowledge derived from structural alignment.

In the sixth chapter "Indexing Protein Structures using Suffix Trees," Feng Gao and Mohammed J. Zaki describe a new approach to 3D database searching for protein sub-structures. Given a large set of proteins, they extract local structural features, which are converted into a set of symbols, which can be indexed using a traditional suffix tree. They show how one can rapidly retrieve approximately similar protein substructure matching a query protein.

Protein Features Prediction

In the seventh chapter "Hidden Markov Models for Prediction of Protein Features," Christopher Bystroff and Anders Krogh present a comprehensive overview of Hidden Markov Models (HMMs), which are used extensively in protein structure/sequence algorithms. They specifically focus on the applications of HMMs to predict signal peptides, secondary and local structure, and transmembrane helices.

In the eighth chapter "The Pros and Cons of Predicting Protein Contact Maps," Lisa Bartoli, Emidio Capriotti, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio discuss methods to predict protein contact maps. Contact maps are "simplified" 2D representations of the 3D proteins structure yet, they retain most of the important features for protein folding. They discuss the strengths and weaknesses of the contact map representation and highlight ways to improve contact map predictions.

In the ninth chapter "Road Map Methods for Protein Folding," Mark Moll, David Schwarz, and Lydia E. Kavraki give a comprehensive survey of "roadmap" approaches to protein folding. Roadmap methods, inspired by motion planning techniques in robotics research, provide a model for understanding and predicting the folding mechanism or pathway.

Methods for De Novo Structure Prediction

In the tenth Chapter "Scoring Functions for De Novo Protein Structure Prediction Revisited," Shing-Chung Ngan, Ling-Hong Hung, Tianyun Liu, and Ram Samudrala, provide a thorough review of both physics-based and knowledge-based scoring functions for conformational samples in de novo protein structure prediction.

In the eleventh chapter "Protein–Protein Docking: Overview and Performance Analysis," Kevin Wiehe, Matthew W. Peterson, Brian Pierce, Julian Mintseris, and Zhiping Weng focus on Fast Fourier Transform-based methods for protein docking. They specifically focus on the ZDOCK algorithm and study its performance on benchmark datasets and study its strengths and weaknesses through regression analysis.

In the final chapter "Molecular Dynamics Simulations of Protein Folding," Angel E. Garcia describes the Replica Exchange Molecular Dynamics (REMD) method for molecular dynamics simulation. He illustrates the effectiveness of the REMD method on the folding of a small protein.

Mohammed J. Zaki
Chris Bystroff

Contents

Contributors

Lisa Bartoli • *CIRB/Department of Biology, University of Bologna, Bologna, Italy*

Christopher Bystroff • *Department of Biology, Rensselaer Polytechnic Institute, Troy NY, USA*

Emidio Capriotti • *CIRB/Department of Biology, University of Bologna, Bologna, Italy*

Rita Casadio • *CIRB/Department of Biology, University of Bologna, Bologna, Italy*

Domenico Cozzetto • *Department of Biochemical Sciences, University of Rome "La Sapienza", Rome, Italy*

Kyle Ellrott • *Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA*

Piero Fariselli • *CIRB/Department of Biology, University of Bologna, Bologna, Italy*

Feng Gao • *Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA*

Angel E. Garcia • *Department of Physics, Applied Physics and Astronomy, Rensselaer Polytechnic Institute, Troy, NY 12180*

Alejandro Giorgetti • *Department of Biochemical Sciences, University of Rome "La Sapienza", Rome, Italy*

Jun-tao Guo • *Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA*

Ling-Hong Hung • *Department of Microbiology, University of Washington School of Medicine, Seattle, WA, USA*

Feng Jiao • *School of Computer Science, University of Waterloo, Waterloo, Canada*

Lydia E. Kavraki • *Computer Science Department, Rice University, Houston, TX, USA*

Anders Krogh • *The Bioinformatics Centre, Inst. Mol. Biol. and Physiology, University of Copenhagen, Denmark*

Tianyun Liu • *Department of Microbiology, University of Washington School of Medicine, Seattle, WA, USA*

Pier Luigi Martelli • *CIRB/Department of Biology, University of Bologna, Bologna, Italy*
Liam Jones McGuffin • *The University of Reading, Reading, UK*
Julian Mintseris • *Bioinformatics Program, Boston University, Boston, MA, USA*
Mark Moll • *Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA*
Shing-Chung Ngan, • *Department of Microbiology, University of Washington School of Medicine, Seattle, WA, USA*
Ruth Nussinov • *Sackler Inst. of Molecular Medicine, Tel Aviv University, Tel Aviv, Israel and Basic Research Program, SAIC-Frederick, Inc., Frederick, MD, USA*
Matthew W. Peterson • *Department of Biomedical Engineering, Boston University, Boston, MA, USA*
Brian Pierce • *Bioinformatics Program, Boston University, Boston, MA, USA*
Domenico Raimondo • *Department of Biochemical Sciences, University of Rome "La Sapienza", Rome, Italy*
Ram Samudrala • *Department of Microbiology, University of Washington School of Medicine, Seattle, WA, USA*
David Schwarz • *Computer Science Department, Rice University, Houston, TX, USA*
Maxim Shatsky • *School of Computer Science, Tel Aviv University, Tel Aviv, Israel*
Anna Tramontano • *Department of Biochemical Sciences, University of Rome "La Sapienza", Rome, Italy*
Zhiping Weng • *Department of Biomedical Engineering, Boston University, Boston, MA, USA*
Kevin Wiehe • *Bioinformatics Program, Boston University, Boston, MA, USA*
Haim J. Wolfson • *School of Computer Science, Tel Aviv University, Tel Aviv, Israel*
Jinbo Xu • *Toyota Technological Institute at Chicago, Chicago, IL, USA*
Ying Xu • *Department of Biochemistry and Molecular Biology, University of Georgia, Athens, GA, USA*
Libo Yu • *Bioinformatics Solutions Inc., Waterloo, Canada*
Mohammed J. Zaki • *Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY, USA*