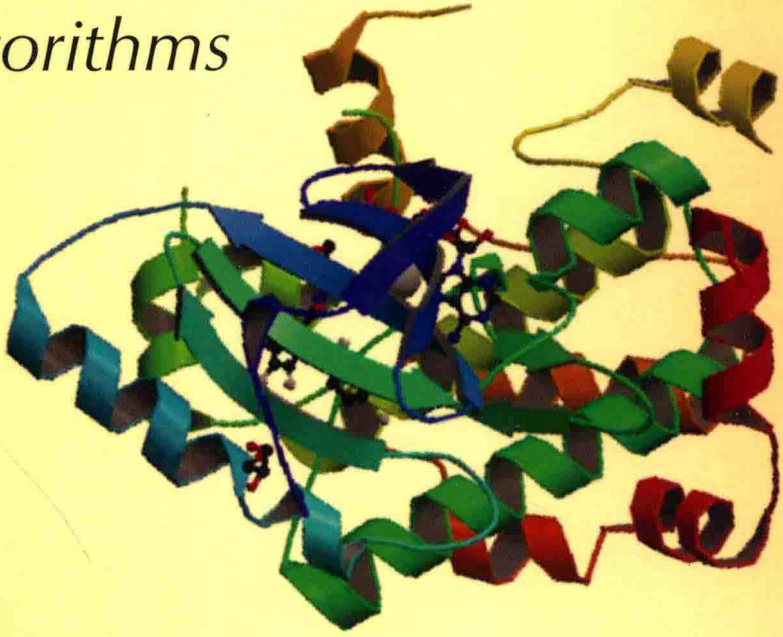
WILEY SERIES ON BIOINFORMATICS
Yi Pan and Albert Y. Zomaya, Series Editors

# *Introduction to* Protein Structure Prediction

## *Methods and Algorithms*



Edited by
Huzefa Rangwala
George Karypis

WILEY

# INTRODUCTION TO PROTEIN STRUCTURE PREDICTION

## Methods and Algorithms

Edited by

**HUZEFA RANGWALA**
**GEORGE KARYPIS**



WILEY

**A JOHN WILEY & SONS, INC., PUBLICATION**

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada



***Library of Congress Cataloging-in-Publication Data:***

Rangwala, Huzefa.
Introduction to protein structure prediction : methods and algorithms / Huzefa Rangwala, George Karypis.
p. cm.—(Wiley series in bioinformatics; 14)
Includes bibliographical references and index.
ISBN 978-0-470-47059-6 (hardback)
1. Proteins—Structure—Mathematical models. 2. Proteins—Structure—Computer simulation. I. Karypis, G. (George) II. Title.
QP551.R225 2010
572′.633—dc22
2010028352

Printed in Singapore

10 9 8 7 6 5 4 3 2 1

# INTRODUCTION TO PROTEIN STRUCTURE PREDICTION

## WILEY SERIES ON BIOINFORMATICS: COMPUTATIONAL TECHNIQUES AND ENGINEERING

**Series Editors, Yi Pan & Albert Zomaya**

**Knowledge Discovery in Bioinformatics: Techniques, Methods and Applications** / Xiaohua Hu & Yi Pan

**Grid Computing for Bioinformatics and Computational Biology** / Albert Zomaya & El-Ghazali Talbi

**Analysis of Biological Networks** / Björn H. Junker & Falk Schreiber

**Bioinformatics Algorithms: Techniques and Applications** / Ion Mandoiu & Alexander Zelikovsky

**Machine Learning in Bioinformatics** / Yanqing Zhang & Jagath C. Rajapakse

**Biomolecular Networks** / Luonan Chen, Rui-Sheng Wang, & Xiang-Sun Zhang

**Computational Systems Biology** / Huma Lodhi

**Computational Intelligence and Pattern Analysis in Biology Informatics** / Ujjwal Maulik, Sanghamitra, & Jason T. Wang

**Mathematics of Bioinformatics: Theory, Practice, and Applications** / Matthew He

**Introduction to Protein Structure Prediction: Methods and Algorithms** / Huzefa Rangwala & George Karypis

# PREFACE

## PROTEIN STRUCTURE PREDICTION

Proteins play a crucial role in governing several life processes. Stunningly complex networks of proteins perform innumerable functions in every living cell. Knowing the function and structure of proteins is crucial for the development of better drugs, higher yield crops, and even synthetic biofuels. As such, knowledge of protein structure and function leads to crucial advances in life sciences and biology. The motivation behind the structural determination of proteins is based on the belief that structural information provides insights as to their function, which will ultimately result in a better understanding of intricate biological processes.

Breakthroughs in large-scale sequencing have led to a surge in the available protein sequence information that has far outstripped our ability to characterize the structural and functional characteristic of these proteins. Several research groups have been working on determining the three-dimensional structure of the protein using a wide variety of computational methods. The problem of unraveling the relationship between the amino acid sequence of a protein and its three-dimensional structure has been one of the grand challenges in molecular biology. The importance and the far reaching implications of being able to predict the structure of a protein from its amino acid sequence is manifested by the ongoing biennial competition on "Critical Assessment of Protein Structure Prediction" (CASP) that started more than 16 years ago. CASP is designed to assess the performance of current structure prediction methods and over the years the number of groups that have been participating in it continues to increase.

This book presents a series of chapters by authors who are involved in the task of structure determination and using modeled structures for applications involving drug discovery and protein design. The book is divided into the following themes.

## BACKGROUND ON STRUCTURE PREDICTION

Chapter 1 provides an introduction to the protein structure prediction problem along with information about databases and resources that are widely used. Chapters 2 and 3 provide information regarding two very important initiatives in the field: (i) the structure prediction flagship competition (CASP), and (ii) the protein structure initiative (PSI), respectively. Since many of the approaches developed have been tested in the CASP competition, Chapter 2 lays the foundation for the need for such an evaluation, the problem definitions, significant innovations, competition format, as well as future outlook. Chapter 3 describes the protein structure initiative, which is designed to determine representative three-dimensional structures within the human genome.

## PREDICTION OF STRUCTURAL ELEMENTS

Within each structural entity called a protein there lies a set of recurring substructures, and within these substructures are smaller substructures. Beyond the goal of predicting the three-dimensional structure of a protein from sequence several other problems have been defined and methods have been developed for solving the same. Chapters 4–6 provide the definitions of these recurring substructures called local alphabets or secondary structures and the computational approaches used for solving these problems. Chapter 6 specifically focuses on a class of transmembrane proteins known to be harder to crystallize. Knowing the pairs of residues within a protein that are within contact or at a closer distance provides useful distance constraints that can be used while modeling the three-dimensional structure of the protein. Chapter 7 focuses on the problem of contact map prediction and also shows the use of sophisticated machine learning methods to solve the problem. A successful solution for each of these subproblems assists in solving the overarching protein structure prediction problem.

## TERTIARY STRUCTURE PREDICTION

Chapters 8–11 discuss the widely used structure prediction methods that rely on homology modeling, threading, and fragment assembly. Chapters 8–9 discuss the problems of fold recognition and remote homology detection that attempt to model the three-dimensional structure of a protein using known structures. Chapters 10 and 11 discuss a combination of threading-based approaches along with modeling the protein in parts or fragments and usually helps in modeling the structure of proteins known not to have a close homolog within the structure databases. Chapter 12 is a survey of the hybrid methods that use a combination of the computational and experimental methods to achieve high-resolution protein structures in a high-throughput manner.

Chapter 17 provides information about the challenges in modeling transmembrane proteins along with a discussion of some of the widely used methods for these sets of proteins.

Chapter 13 describes the loop prediction problem and how the technique can be used for refinement of the modeled structures. Chapters 14 and 15 assess the modeled structures and provide a notion of the quality of structures. This is extremely important from a biologist's perspective who would like to have a metric that describes the goodness of the structure before use. Chapter 19 provides insights into the different conformations that a protein may take and the approaches used to sample the different conformations.

## FUNCTIONAL INSIGHTS

Certain parts of the protein structure may be conserved and interact with other biomolecules (e.g., proteins, DNA, RNA, and small molecules) and perform a particular function due to such interactions. Chapter 16 discusses the problem of ligand-binding site prediction and its role in determining the function of the proteins. The approach uses some of the homology modeling principles used for modeling the entire structure. Chapter 18 introduces a computational model that detects the differences between protein structure (modeled or experimentally-determined) and its modeled mutant. Chapter 20 describes the use of molecular dynamic-based approaches for modeling mutants.

## ACKNOWLEDGEMENTS

We wish to acknowledge the many people who have helped us with this project. We firstly thank all the coauthors who spent time and energy to edit their chapters and also served as reviewers by providing critical feedback for improving other chapters. Kevin Deronne, Christopher Kauffman, and Rezwan Ahmed also assisted in reviewing several of the chapters and helped the book take a form that is complete on the topic of protein structure prediction and exciting to read. Finally, we wish to thank our families and friends.

We hope that you as a reader benefit from this book and feel as excited about this field as we are.

HUZEFA RANGWALA

GEORGE KARYPIS

# CONTRIBUTORS


NIR BEN-TAL, Department of Biochemistry and Molecular Biology, Tel Aviv University, Tel Aviv, Israel

AURÉLIE BORNOT, Institut National de la Santé et de la Recherche Médicale, UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Paris Diderot, Paris, France

ALEXANDRE G. DE BREVERN, Institut National de la Santé et de la Recherche Médicale, Université Paris Diderot, Institut National de la Transfusion Sanguine, 75015, Paris, France

JIANLIN CHENG, Computer Science Department and Informatics Institute University of Missouri, Columbia, MO 65211

FENG DING, Department of Biochemistry and Biophysics University of North Carolina—Chapel Hill, NC 27599

NICHOLAS E. DIXON, School of Chemistry, University of Wollongong, NSW 2522, Australia

NIKOLAY V. DOKHOLYAN, Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599

ESHEL FARAGGI, Indiana University School of Informatics, Indiana University-Purdue University Indianapolis, and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202

KRZYSZTOF FIDELIS, Protein Structure Prediction Center, Genome Center, University of California, Davis, Davis, CA

ANDRAS FISER, Department of Systems and Computational Biology and Department of Biochemistry, Albert Einstein College of Medicine, Bronx, NY 10461

NARCIS FERNANDEZ-FUENTES, Leeds Institute of Molecular Medicine, University of Leeds, Leeds, UK

ADAM GODZIK, Program in Bioinformatics and Systems Biology, Sanford-Burnham Medical Research Institute, La Jolla, CA 92037

THOMAS HUBER, The University of Queensland, School of Chemistry and Molecular Biosciences, QLD, Australia

AGNEL PRAVEEN JOSEPH, Institut National de la Santé et de la Recherche Médicale, UMR-S 665, Dynamique des Structures et Interactions des Macromolécules Biologiques (DSIMB), Université Paris Diderot, Paris, France

KAZUHIKO KANOU, School of Pharmacy, Kitasato University, Tokyo 108-8641, Japan

GEORGE KARYPIS, Department of Computer Science, University of Minnesota Minneapolis, MN 55455

CHRIS KAUFFMAN, Department of Computer Science, University of Minnesota, Minneapolis, MN 55455

BOSTJAN KOBE, The University of Queensland, School of Chemistry and Molecular Biosciences, Brisbane, Australia

ANDRIY KRYSHTAFOVYCH, Protein Structure Prediction Center, Genome Center, University of California, Davis, Davis, CA

ALBERTO J.M. MARTIN, Complex and Adaptive Systems Lab, School of Computer Science and Informatics, UCD Dublin, Ireland

MAJID MASSA, Department of Bioinformatics and Computational Biology, George Mason University, Manassas, VA 20110

LIAM J. MCGUFFIN, School of Biological Sciences, The University of Reading, Reading, UK

CATHERINE MOONEY, Shields Lab, School of Medicine and Medical Science, University College Dublin, Ireland

JOHN MOULT, Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD 20850

DMITRI MOURADOV, The University of Queensland, School of Chemistry and Molecular Biosciences, QLD, Australia

CHRISTINE ORENGO, Department of Structural and Molecular Biology, University College London, London UK

SHASHI BHUSHAN PANDIT, Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology, Atlanta, GA 30318

GIANLUCA POLLASTRI, Complex and Adaptive Systems Lab, School of Computer Science and Informatics, UCD Dublin, Ireland

HUZEFA RANGWALA, Department of Computer Science, George Mason University, Fairfax, VA 22030

BURKHARD ROST, Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY 10032

AMBRISH ROY, Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109

MAYA SCHUSHAN, Department of Biochemistry and Molecular Biology, Tel Aviv University, Tel Aviv, Israel

AMARDA SHEHU, Department of Computer Science, George Mason University, Fairfax, VA 22030

MAYUKO TAKEDA-SHITAKA, School of Pharmacy, Kitasato University, Tokyo 108-8641, Japan

ISTVÁN SIMON, lntsitute of Enzymology, BRC, Hungarian Academy of Sciences, Budapest, Hungary

JEFFREY SKOLNICK, Center for the Study of Systems Biology, School of Biology, Georgia Institute of Technology Atlanta, GA 30318

ALLISON N. TEGGE, Computer Science Department and Informatics Institute, University of Missouri, Columbia, MO 65211

GENKI TERASHI, School of Pharmacy, Kitasato University, Tokyo 108-8641, Japan

GÁBOR E. TUSNADY, Intsitute of Enzymology, BRC, Hungarian Academy of Sciences, Budapest, Hungary

HIDEAKI UMEYAMA, School of Pharmacy, Kitasato University, Tokyo 108-8641, Japan

IOSIF I. VAISMAN, Department of Bioinformatics and Computational Biology, George Mason University, Manassas, VA 20110

IAN WALSH, Complex and Adaptive Systems Lab, School of Computer Science and Informatics, UCD Dublin, Ireland

ZHENG WANG, Computer Science Department, University of Missouri, Columbia, MO 65211

SITAO WU, Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109

SHUANGYE YIN, Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, NC 27599

YANG ZHANG, Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109

HONGYI ZHOU, Center for the Study of Systems Biology, School of Biology Georgia Institute of Technology, Atlanta, GA 30318

YAOQI ZHOU, Indiana University School of Informatics, Indiana University-Purdue University Indianapolis, and Center for Computational Biology and Bioinformatics, Indiana University School of Medicine, Indianapolis, IN 46202

# CONTENTS

# CHAPTER 1

# INTRODUCTION TO PROTEIN STRUCTURE PREDICTION

HUZEFA RANGWALA
Department of Computer Science
George Mason University
Fairfax, VA

GEORGE KARYPIS
Department of Computer Science
University of Minnesota
Minneapolis, MN

Proteins have a vast influence on the molecular machinery of life. Stunningly complex networks of proteins perform innumerable functions in every living cell. Knowing the function and structure of proteins is crucial for the development of improved drugs, better crops, and even synthetic biofuels. As such, knowledge of protein structure and function leads to crucial advances in life sciences and biology.

With recent advances in large-scale sequencing technologies, we have seen an exponential growth in protein sequence information. Protein structures are primarily determined using X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy, but these methods are time consuming, expensive, and not feasible for all proteins. The experimental approaches to determine protein function (e.g., gene knockout, targeted mutation, and inhibitions of gene expression studies) are low-throughput in nature [1,2]. As such, our ability to produce sequence information far outpaces the rate at which we can produce structural and functional information.

Consequently, researchers are increasingly reliant on computational approaches to extract useful information from experimentally determined three-dimensional (3D) structures and functions of proteins. Unraveling the

*Introduction to Protein Structure Prediction: Methods and Algorithms*,
Edited by Huzefa Rangwala and George Karypis

relationship between pure sequence information and 3D structure and/or function remains one of the fundamental challenges in molecular biology.

Function prediction is generally approached by using inheritance through homology [2], that is, proteins with similar sequences (common evolutionary ancestry) frequently carry out similar functions. However, several studies [2–4] have shown that a stronger correlation exists between structure conservation and function, that is, structure implies function, and a higher correlation exists between sequence conservation and structure, that is, sequence implies structure (sequence → structure → function).

## 1.1. INTRODUCTION TO PROTEIN STRUCTURES

In this section we introduce the basic definitions and facts about protein structure, the four different levels of protein structure, as well as provide details about protein structure databases.

### 1.1.1. Protein Structure Levels

Within each structural entity called a protein lies a set of recurring substructures, and within these substructures are smaller substructures still. As an example, consider hemoglobin, the oxygen-carrying molecule in human blood. Hemoglobin has four domains that come together to form its quaternary structure. Each domain assembles (i.e., folds) itself independently to form a tertiary structure. These tertiary structures are comprised of multiple secondary structure elements—in hemoglobin's case α-helices. α-Helices (and their counterpart β-sheets) have elegant repeating patterns dependent upon sequences of amino acids.

***1.1.1.1. Primary Structure.*** Amino acids form the basic building blocks of proteins. Amino acids consists of a central carbon atom ($C_\alpha$) attached by an amino ($NH_2$), a carboxyl (COOH) group, and a side chain (R) group. The side chain group differentiates the various amino acids. In case of proteins, there are primarily 20 different amino acids that form the building blocks. A protein is a chain of amino acids linked with peptide bonds. Pairs of amino acid form a peptide bond between the amino group of one and the carboxyl group of the other. This polypeptide chain of amino acids is known as the primary structure or the protein sequence.

***1.1.1.2. Secondary Structure.*** A sequence of characters representing the secondary structure of a protein describes the general 3D form of local regions. These regions organize themselves independently from the rest of the protein into patterns of repeatedly occurring structural fragments. The most dominant local conformations of polypeptide chains are α-helices and β-sheets. These local structures have a certain regularity in their form, attributed to the hydrogen bond interactions between various residues. An α-helix has a coil-like

structure, whereas a β-sheet consists of parallel strands of residues. In addition to regular secondary structure elements, irregular shapes form an important part of the structure and function of proteins. These elements are typically termed coil regions.

Secondary structure can be divided into several types, although usually at least three classes (α-helix, coils, and β-sheet) are used. No unique method of assigning residues to a particular secondary structure state from atomic coordinates exists, although the most widely accepted protocol is based on the Dictionary of Protein Secondary Structure (DSSP) algorithm [5]. DSSP uses the following structural classes: H (α-helix), G ($3_{10}$-helix), I (π-helix), E (β-strand), B (isolated β-bridge), T (turn), S (bend), and – (other). Several other secondary structure assignment algorithms use a reduction scheme that converts this eight-state assignment down to three states by assigning H and G to the helix state (H), E and B to a the strand state (E), and the rest (I, T, S, and –) to a coil state (C). This is the format generally used in structure databases.

***1.1.1.3. Tertiary Structure.*** The tertiary structure of the protein is defined as the global 3D structure, represented by 3D coordinates for each atoms. These tertiary structures are comprised of multiple secondary structure elements, and the 3D structure is a function of the interacting side chains between the different amino acids. Hence, the linear ordering of amino acids forms secondary structure; arranging secondary structures yields tertiary structure.

***1.1.1.4. Quaternary Structure.*** Quaternary structures represent the interaction between multiple polypeptide chains. The interaction between the various chains is due to the non-covalent interactions between the atoms of the different chains. Examples of these interactions include hydrogen bonding, van Der Walls interactions, ionic bonding, and disulfide bonding.

Research in computational structure prediction concerns itself mainly with predicting secondary and tertiary structures from known experimentally determined primary structure or sequence. This is due to the relative ease of determining primary structure and the complexity involved in quaternary structure.

### 1.1.2. Protein Sequence and Structure Databases

The large amount of protein sequence information, experimentally determined structure information, and structural classification information is stored in publicly available databases. In this section we review some of the databases that are used in this field, and provide their availability information in Table 1.1.

***1.1.2.1. Sequence Databases.*** The Universal Protein Resource (UniProt) [6] is the most comprehensive warehouse containing information about protein