METHODS IN MOLECULAR BIOLOGY

For further volumes:
http://www.springer.com/series/7651

For over 35 years, biological scientists have come to rely on the research protocols and methodologies in the critically acclaimed *Methods in Molecular Biology* series. The series was the first to introduce the step-by-step protocols approach that has become the standard in all biomedical protocol publishing. Each protocol is provided in readily-reproducible step-by-step fashion, opening with an introductory overview, a list of the materials and reagents needed to complete the experiment, and followed by a detailed procedure that is supported with a helpful notes section offering tips and tricks of the trade as well as troubleshooting advice. These hallmark features were introduced by series editor Dr. John Walker and constitute the key ingredient in each and every volume of the *Methods in Molecular Biology* series. Tested and trusted, comprehensive and reliable, all protocols from the series are indexed in PubMed.

Prediction of Protein Secondary Structure

Second Edition

Edited by

Andrzej Kloczkowski

The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA

Lukasz Kurgan

Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA

Eshel Faraggi

Department of Physics, Indiana University, Indianapolis, IN, USA

Humana Press

Editors
Andrzej Kloczkowski
The Steve and Cindy Rasmussen
Institute for Genomic Medicine
Nationwide Children's Hospital
Columbus, OH, USA

Lukasz Kurgan
Department of Computer Science
Virginia Commonwealth University
Richmond, VA, USA

Eshel Faraggi
Department of Physics
Indiana University
Indianapolis, IN, USA

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-4195-8 ISBN 978-1-0716-4196-5 (eBook)
https://doi.org/10.1007/978-1-0716-4196-5

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Preface

The first edition of our book *Prediction of Protein Secondary Structure* was published in October 2016. Subsequently, Springer-Nature asked us to prepare the second edition of this book. From the original Editorial team (Yaoqi Zhou, Andrzej Kloczkowski (A.K.), Eshel Faraggi (E.F.), and Yeudong Yang), two (A.K. and E.F.) have accepted this invitation and invited Lukasz Kurgan as a new Co-Editor. Although this is a second edition of the book, it differs substantially from second editions of other published titles that typically cover the same scope as the original edition; this book focuses on adding new developments in the field. While the title of this volume remains *Prediction of Protein Secondary Structure*, it adds completely new materials written chiefly by different authors when compared to the coverage of its first edition.

Since 2016, there has been enormous amount of progress in protein structure prediction due to the development, release, and applications of deep learning-based methods, such as AlphaFold and Protein Language Models that use Natural Language Processing (NLP) techniques.

In the post-AlphaFold age, one of the most critical challenges in molecular biology is to understand how boundary conditions and the one-dimensional (1D) sequence of amino acid determines three-dimensional (3D) structure and function of a given protein and how these aspects change as the sequence changes. After more than 50 years of efforts, reliable computational protein models that reach quality of experimental structure are now available for most proteins. But even when static computational 3D models of proteins are available, and more so when they do not (e.g., because these proteins are intrinsically disordered), connecting sequence and structure to dynamic and function remains challenging. Secondary structure (SS) and other one-dimensional (1D) structural properties, such as the accessible surface area and intrinsic disorder, help predict 3D models. Many methods that predict 1D structure from sequence have been developed as an intermediate step or a substitute for the 3D structure prediction. These 1D characteristics can be structural and functional properties that are expressed by a one-dimensional vector along the protein sequence. They provide ways to represent information that can be more readily processed, for example, by machine learners.

In the secondary structure, the protein backbone structure is annotated by a few states, such as helices, sheets, or coils. Torsion angles can also characterize the protein's local structure. In addition to the structural properties of the backbone, protein structures can be described by global structural properties. These properties depend on interactions between multiple residues far apart in the sequence. Examples include the solvent-accessible surface area and the number of residues in a given spatial region, which are relevant to proteins' tertiary packing and function. More recently, predicting intrinsic disorder (lack of well-defined structure under physiological conditions) and one-dimensional functional properties (e.g., residues that bind ligands and posttranslational modification sites) and the effects of sequence variation on function have received increasing attention. Correspondingly, this book expands the first edition to cover a broad spectrum of related topics ranging from those related directly to protein secondary structure, to methods for the prediction of 3D structure, the deep learning, and large-scale protein language models for the prediction of other 1D properties of proteins that include protein-nucleic acid and protein-protein interactions, intrinsic disorder, and noncoding RNA fragments and their secondary structure and applications as medical biomarkers and therapeutic targets.

The following chapters are included in this volume of *Methods in Molecular Biology*:

The first chapter is written by George D. Rose of the Johns Hopkins University, USA, one of the pioneers in protein structure prediction. This chapter gives an excellent overview of proteins' most important secondary structure element, the α-helix. It starts from the historically first model of α-helix, developed by Pauling, and goes to the most recent developments in understanding the α-helix by deep learning and AlfaFold2 predictions.

Adam Liwo from the University of Gdansk, Poland, and his collaborators discuss the role of secondary structure as a scaffold for building the tertiary structure of protein. Both template-based (assisted) modeling and free modeling for proteins with novel folds are overviewed. Particular emphasis is given to the physics-based UNRES force field developed by Liwo and Scheraga and the energy terms of UNRES that govern secondary structure formation.

In Chap. 3, Jianlin Cheng from the University of Missouri-Columbia, USA, with collaborators, describes how protein secondary structure prediction can be improved by applying deep language models and transformer networks. They developed a novel TransPross method based on the transformer network and attention mechanism commonly used in natural language processing that significantly improves protein secondary structure prediction. The TransPross algorithm is freely available on GitHub.

Aleksandra E. Badaczewska from the Iowa State University, USA, and Andrzej Kolinski from the University of Warsaw, Poland, discuss the significant role of secondary structure in large-scale protein modeling with coarse-grained methods. The CABS model developed in the past by the Kolinski group is too slow for modeling large-scale transitions in large protein systems. Therefore, they created a lower-resolution, faster, coarse-grained modeling tool called SURPASS. The application and performance of the SURPASS methodology are discussed in the chapter.

Chapter 6, authored by Michal Brylinski and his colleagues from the Louisiana State University, USA, surveys contributions of machine learning to the development of computational methods for the prediction of protein structure and function. Motivated by recent developments in this area, they focus on the deep neural networks and natural language processing inspired algorithms and cover a broad range of predictive targets, such as secondary and 3D structures, residue-residue contacts, subcellular localization, and protein function annotations.

Daisuke Kihara from the Purdue University, USA, with collaborators, describe how the identification of protein secondary structure and the detection of DNA/RNA location in cryo-electron microscopy (cryo-EM) and cryo-electron tomography (cryo-ET) maps at intermediate or low-resolution enables significantly better interpretations of such maps. Kihara's lab developed Emap2sec software to identify the secondary protein structures and Emapsec+ to identify additional locations of DNA or RNA molecules in the cryo-EM maps. Both methods are publicly available as web servers.

In Chap. 7, Liam J. McGuffin and Ahmet Gurkan Genc from the University of Reading, UK, discuss impact of AlphaFold2 and the subsequent and related tools on the protein structure prediction field. This chapter provides easy to read history of the protein structure prediction and structure refinement areas and covers key methods and advancements, which include generative models, protein language models, co-evolutionary approaches, and graph neural network. The authors examine impact of the Critical Assessment of Structure Prediction (CASP) experiments.

Eshel Faraggi with colleagues from across the USA analyze targets from the latest CASP15 experiment for the blind prediction of 3D protein structures from amino acid sequences. They focus on cases where the AlphaFold2 method, the top predictor for CASP15, failed. They identify six such targets and compare these results with predictions of SPINE-X and Seder; two methods developed by the authors of this chapter and participated in CASP15. They find that the poor quality of the AlphaFold2 predictions results from incorrectly assigned secondary structures. They also observe that SPINE-X and Seder sometimes give better predictions by improving detection of the secondary structure elements.

Robert L. Jernigan and collaborators at the Iowa State University, USA, focus their chapter on an analysis of a minimal organism capable of self-replication, JCVI-Syn3. This synthetic minimal organism contains only 473 genes that code 434 proteins, while most of these genes still have unknown functions. They used their novel and highly successful PROST method that relies on the large protein language model, ESM1b, for functional annotation of the protein-coding genes. They use PROST to annotate 93% of the JCVI-Syn3 genes with high accuracy and found new functions for other genes.

In Chap. 10, Lukasz Kurgan and his colleagues from the Virginia Commonwealth University, USA, describe the DescribePROT database, a large online resource that provides easy access to 13 diverse 1D protein structure and function descriptors for 1.4 million proteins from 83 complete proteomes. They summarize the scope of this comprehensive database and different ways to access and query its contents, and comment on future plans to expand it.

M. Michael Gromiha and coworkers from the Indian Institute of Technology Madras and University of the Witwatersrand, South Africa, focus on the HIV protease, a 99-residue protein protease that is recognized as one of the potent targets for treating HIV infections. They overview various HIV protease subgroups, the structure of HIV protease, the catalytic mechanism of HIV-1 protease and its specificity, and the movement of the flaps, which is crucial for HIV-1 PR enzymatic activity. They discuss drug resistance in the HIV protease and the application of 3D-QSAR methods to study HIV protease–inhibitor interactions.

The next two chapters focus on the prediction of intrinsic disorder and its arguably most extreme form, which are disordered linkers. In Chap. 12, Lukasz Kurgan and his collaborators from the Nankai University, China, describe the flDPnn method that targets prediction of the intrinsic disorder. This is one of the most accurate disorder predictors based on the results from the last two Critical Assessment of Intrinsic Disorder (CAID) experiments, which is additionally capable of predicting several major functions of disorder. In Chap. 13, Zhenling Peng from the Shandong University, China, and her colleagues summarize the disordered linker predictor, APOD. The disordered linkers connect protein domains and structural elements within domains and they perform these functions without ever folding into a well-defined structure. This is in contrast to other disordered regions that undergo folding under certain circumstances, such as binding. APOD provides accurate predictions and is available as a web server. In both chapters, the authors explain the underlying predictive models and use case studies to showcase how to read and understand their predictions.

Jörg Gsponer and Nawar Malhis from the University of British Columbia, Canada, provide insights into the area of the analysis and prediction of linear interacting peptides (LIPs), a generic class of disordered regions that are involved in inter-molecular interactions. They perform empirical comparison of several relevant predictors using benchmark data from the CAID1 and CAID2 experiments. They find significant differences between predictors and recommend using multiple methods and performing manual analysis of their results to minimize false positive rate.

In Chap. 15, Jian Zhang and his collaborators from the Xinyang Normal University, China, and the Virginia Commonwealth University, USA, describe the SCRIBER tool that accurately predicts protein binding residues in protein sequences. The key feature of this method is the fact that it minimizes cross-predictions, which are defined as predictions of residues that interact with other ligand types (ions, nucleic acids, lipids, etc.) as protein binding. The authors include a case study that illustrates how to read and interpret predictions generated by the web server of SCRIBER.

Dukka B. KC and his research group at the Rochester Institute of Technology, USA, introduce their LMPTMSite (Language Model-based Post-Translational Modification Site) prediction platform. This platform includes accurate predictors of two types of the post-translational modification sites, S-nitrosylation and succinylation. The authors explain the underlying deep learning architectures, empirically compare their predictors with related tools, and present a case study to demonstrate how to use their tools.

Chapter 17, authored by Mubashir Hassan and collaborators, overviews protein aggregation phenomena and discusses various computational methods, databases, algorithms, and tools that predict the aggregation of proteins and peptides. Protein aggregation is directly related to misfolding and secondary structure. Proteins aggregate, forming complex amyloid structures that contain alpha helices and beta sheets. The authors discuss secondary structure propensity in the protein aggregation and diseases related to the protein aggregation.

M. Michael Gromiha and his coworkers discuss the recognition mechanism in protein-nucleic acid interactions, particularly focusing on the binding affinities of protein-DNA and protein-RNA complexes. They describe various databases of protein-nucleic acid complexes and interactions between proteins and nucleic acids. They also overview multiple algorithms and tools for extracting features from protein-nucleic acid complexes and nucleic acid structures, analyzing protein-DNA and protein-RNA complexes, and predicting protein-nucleic acid complex structures. The authors also discuss databases on protein-nucleic acid binding affinities and methods that analyze and predict these binding affinities.

The last chapter, written by Andrzej Kloczkowski and collaborators, differs from the other chapters since it covers noncoding RNAs (ncRNAs) rather than proteins. The ncRNAs are nowadays one of the hottest topics, being linked to various diseases including cancers, neurodegenerative disorders, and psychiatric disorders. This chapter overviews multiple types of ncRNAs, secondary structures of RNAs linked to intramolecular base matching, databases, tools used to analyze ncRNAs, and ncRNA target prediction. The authors discuss applications of ncRNAs in treating different diseases, including their use as diagnostic and prognosis biomarkers and therapeutic targets. Finally, they also summarize clinical implementations of ncRNAs in cancer, neurodegenerative, cardiovascular, and infectious diseases.

This volume covers a comprehensive collection of methods, resources, and studies that target analysis and prediction of various structural and functional aspects of proteins and ncRNAs. We hope it will help readers to gain a better understanding of this dynamic and growing field of research and guide them in the quest to identify and use the best computational methods and resources.

Columbus, OH, USA	*Andrzej Kloczkowski*
Richmond, VA, USA	*Lukasz Kurgan*
Indianapolis, IN, USA	*Eshel Faraggi*

Contents

Contributors

ALEKSANDRA E. BADACZEWSKA-DAWID • *Office of Biotechnology, Iowa State University, Ames, IA, USA*
JAVAD BAGHIROV • *Department of Computer Science, Purdue University, West Lafayette, IN, USA*
SOUFIA BAHMANI • *Computer Science Department, Rochester Institute of Technology, Rochester, NY, USA; Michigan Technological University, Comptuer Science Department, Houghton, MI, USA*
SUSHMITA BASU • *Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA*
MICHAL BRYLINSKI • *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA; Center for Computation and Technology, Louisiana State University, Baton Rouge, LA, USA*
JIANLIN CHENG • *Electrical Engineering and Computer Science Department, University of Missouri, Columbia, MO, USA*
WEIHANG CHENG • *Department of Chemistry, Hubei University, Wuhan, Hubei, China*
CEZARY CZAPLEWSKI • *Faculty of Chemistry, University of Gdańsk, Fahrenheit Union of Universities, Gdańsk, Poland*
ŁUKASZ DZIADEK • *Faculty of Chemistry, University of Gdańsk, Fahrenheit Union of Universities, Gdańsk, Poland*
ESHEL FARAGGI • *Research and Information Systems, LLC, Indianapolis, IN, USA; Department of Physics, Indiana University, Indianapolis, IN, USA*
AHMET GURKAN GENC • *School of Biological Sciences, University of Reading, Reading, UK*
M. MICHAEL GROMIHA • *Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India; International Research Frontiers Initiative, School of Computing, Tokyo Institute of Technology, Yokohama, Japan*
JÖRG GSPONER • *Michael Smith Laboratories, the University of British Columbia, Vancouver, BC, Canada*
K. HARINI • *Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India*
MUBASHIR HASSAN • *The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children Hospital, Columbus, OH, USA*
GANG HU • *NITFID, School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China*
HAMID D. ISMAIL • *Computer Science Department, Rochester Institute of Technology, Rochester, NY, USA; North Carolina A&T State University, Computational Data Science and Engineering, Greensboro, NC, USA*
ROBERT L. JERNIGAN • *Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, IA, USA; Bioinformatics and Computational Biology Program Iowa State University, Ames, IA, USA*
KEJUE JIA • *Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA*

DUKKA B. KC • *Computer Science Department, Rochester Institute of Technology, Rochester, NY, USA*
DAISUKE KIHARA • *Department of Computer Science, Purdue University, West Lafayette, IN, USA; Department of Biological Sciences, Purdue University, West Lafayette, IN, USA*
MESIH KILINC • *Bioinformatics and Computational Biology Program Iowa State University, Ames, IA, USA*
ANDRZEJ KLOCZKOWSKI • *The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, Columbus, OH, USA; Department of Pediatrics, The Ohio State University, Columbus, OH, USA; Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA*
ANDRZEJ KOLINSKI • *Faculty of Chemistry, University of Warsaw, Warsaw, Poland*
SOWMYA RAMASWAMY KRISHNAN • *Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India*
LUKASZ KURGAN • *Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA*
MAI SUAN LI • *Institute of Physics, Polish Academy of Sciences, Warsaw, Poland*
XINGCHEN LIANG • *School of Computer and Information Technology, Xinyang Normal University, Xinyang, China*
AGNIESZKA G. LIPSKA • *Centre of Informatics Tri-city Academic Supercomputer and Network (CI TASK), Gdańsk University of Technology, Fahrenheit Union of Universities in Gdańsk, Gdańsk, Poland*
MENGMENG LIU • *Division of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA, USA*
ADAM LIWO • *Faculty of Chemistry, University of Gdańsk, Fahrenheit Union of Universities, Gdańsk, Poland*
EMILIA A. LUBECKA • *Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Fahrenheit Union of Universities in Gdańsk, Gdańsk, Poland*
YUXIAN LUO • *Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China*
NAWAR MALHIS • *Michael Smith Laboratories, the University of British Columbia, Vancouver, BC, Canada*
AMAL MALIK • *Institute of Molecular Biology and Biotechnology, The University of Lahore, Lahore, Pakistan*
LIAM J. MCGUFFIN • *School of Biological Sciences, University of Reading, Reading, UK*
XIALONG NI • *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA*
ZHENLING PENG • *Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China; Frontier Science Center for Nonlinear Expectations, Ministry of Education, Shandong University, Qingdao, China*
SURESH POKHAREL • *Computer Science Department, Rochester Institute of Technology, Rochester, NY, USA*
PAWEL PRATYUSH • *Computer Science Department, Rochester Institute of Technology, Rochester, NY, USA*
LIMENG PU • *Center for Computation and Technology, Louisiana State University, Baton Rouge, LA, USA*
GEORGE D. ROSE • *Johns Hopkins University, Baltimore, MD, USA*
YASIEN SAYED • *Protein Structure-Function Research Unit, School of Molecular and Cell Biology, University of the Witwatersrand, Johannesburg, South Africa*

MASAKAZU SEKIJIMA • *Department of Computer Science, Tokyo Institute of Technology, Yokohama, Japan*
SABA SHAHZADI • *The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children Hospital, Columbus, OH, USA*
ADAM K. SIERADZAN • *Faculty of Chemistry, University of Gdańsk, Fahrenheit Union of Universities, Gdańsk, Poland*
GOPAL SRIVASTAVA • *Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA*
SANKARAN VENKATACHALAM • *Department of Biotechnology, Bhupat and Jyoti Mehta School of Biosciences, Indian Institute of Technology Madras, Chennai, Tamil Nadu, India*
KUI WANG • *NITFID, School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China*
XIAO WANG • *Department of Computer Science, Purdue University, West Lafayette, IN, USA*
HAIYAN WU • *Research Center for Mathematics and Interdisciplinary Sciences, Shandong University, Qingdao, China*
TIANQI WU • *Electrical Engineering and Computer Science Department, University of Missouri, Columbia, MO, USA*
ZHONGHUA WU • *School of Mathematical Sciences and LPMC, Nankai University, Tianjin, China*
ZAINAB YASEEN • *Faculty of Science and Technology, University of Central Punjab, Lahore, Pakistan*
MUHAMMAD YASIR • *Department of Pharmacology, College of Medicine, Kangwon National University, Chuncheon, Republic of Korea*
JIAN ZHANG • *School of Computer and Information Technology, Xinyang Normal University, Xinyang, China*
BI ZHAO • *Genomics program, College of Public Health, University of South Florida, Tampa, FL, USA*
FENG ZHOU • *School of Computer and Information Technology, Xinyang Normal University, Xinyang, China*
HAN ZHU • *Department of Computer Science, Purdue University, West Lafayette, IN, USA*