Prediction of Protein Structures, Functions, and Interactions

Edited by

JANUSZ M. BUJNICKI

Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland

and

Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland
Prediction of Protein Structures, Functions, and Interactions
Prediction of Protein Structures, Functions, and Interactions

Edited by

JANUSZ M. BUJNICKI
Laboratory of Bioinformatics and Protein Engineering,
International Institute of Molecular and Cell Biology,
Warsaw, Poland

and

Laboratory of Bioinformatics, Institute of Molecular Biology and
Biotechnology, Faculty of Biology, Adam Mickiewicz University,
Poznan, Poland
I dedicate this book with love to my parents.
Thank you for everything.
Contents

List of Contributors ix
Preface xiii

1 The Basics of Protein Sequence Analysis 1
  Katarzyna H. Kaminska, Kaja Milanowska and Janusz M. Bujnicki

2 First Steps of Protein Structure Prediction 39
  Karolina Majorek, Łukasz Kozłowski, Marcin Jakalski
  and Janusz M. Bujnicki

3 Automated Prediction of Protein Function from Sequence 63
  Meghana Chitale, Troy Hawkins and Daisuke Kihara

4 Template Based Prediction of Three-dimensional Protein Structures:
  Fold Recognition and Comparative Modeling 87
  Jan Kosiński, Karolina L. Tkaczuk, Joanna M. Kasprzak and
  Janusz M. Bujnicki

5 Template-free Predictions of Three-dimensional Protein Structures:
  From First Principles to Knowledge-based Potentials 117
  Dominik Gront, Dorota Latek, Mateusz Kurcinski and Andrzej Kolinski

6 Quality Assessment of Protein Models 143
  Björn Wallner and Arne Elofsson

7 Prediction of Molecular Interactions from 3D-structures: From Small
  Ligands to Large Protein Complexes 159
  Kengo Kinoshita, Hidetoshi Kono and Kei Yura

8 Structure-based Prediction of Enzymes and Their Active Sites 187
  James W. Torrance and Janet M. Thornton

9 The Prediction of Macromolecular Complexes by Docking 211
  Sjoerd J. de Vries, Marc van Dijk and Alexandre M.J.J. Bonvin
10 Protein Function Prediction via Analysis of Interactomes 231
   Elena Nabieva and Mona Singh

11 Integrating Prediction of Structure, Function, and Interactions 259
   Michael Tress, Janusz M. Bujnicki, Gonzalo Lopez and Alfonso Valencia

Index 281
List of Contributors

Alexandre M.J.J. Bonvin, Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University, Padualaan 8, 3584CH, Utrecht, The Netherlands

Janusz M. Bujnicki, Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland, and Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland

Meghana Chitale, Department of Computer Science, College of Science, Purdue University, West Lafayette, IN, 47907, USA

Sjoerd J. de Vries, Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University, Padualaan 8, 3584CH, Utrecht, The Netherlands

Arne Elofsson, Center for Biomembrane Research and Stockholm Bioinformatics Center, Stockholm University, SE-106 91 Stockholm, Sweden

Dominik Gront, Faculty of Chemistry, University of Warsaw, Pateura 1, 02-093 Warsaw, Poland

Troy Hawkins, Department of Biological Sciences, College of Science, Purdue University, West Lafayette, IN, 47907, USA

Marcin Jąkalski, Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland

Katarzyna H. Kaminska, Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland, and Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland

Joanna M. Kasprzak, Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland
x  List of Contributors

Daisuke Kihara, Department of Biological Sciences/Computer Science, Markey Center for Structural Biology, College of Science, Purdue University, West Lafayette, IN, 47907, USA

Kengo Kinoshita, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Andrzej Kolinski, Faculty of Chemistry, University of Warsaw, Pateura 1, 02-093 Warsaw, Poland

Hidetoshi Kono, Computational Biology, Quantum Beam Science Directorate, Japan Atomic Energy Agency, 8-1 Umemidai, Kizugawa, Kyoto 619-0215, Japan

Jan Kosiński, Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland

Łukasz Kozłowski, Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland

Mateusz Kurcinski, Faculty of Chemistry, University of Warsaw, Pateura 1, 02-093 Warsaw, Poland

Dorota Latek, Faculty of Chemistry, University of Warsaw, Pateura 1, 02-093 Warsaw, Poland

Gonzalo Lopez, Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C. Melchor Fernandez Almagro, 3, 28029 Madrid, Spain

Karolina Majorek, Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland

Kaja Milanowska, Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, ul. Umultowska 89, 61-614 Poznan, Poland

Elena Nabieva, Department of Computer Science & Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Princeton, NJ 08544, USA

Mona Singh, Department of Computer Science & Lewis-Sigler Institute for Integrative Genomics, Carl Icahn Laboratory, Princeton University, Princeton, NJ 08544, USA

Janet M. Thornton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK
Karolina L. Tkaczuk, Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, ul. Ks. Trojdena 4, 02-109 Warsaw, Poland

James W. Torrance, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Michael Tress, Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C. Melchor Fernandez Almagro, 3, 28029 Madrid, Spain

Alfonso Valencia, Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C. Melchor Fernandez Almagro, 3, 28029 Madrid, Spain

Marc van Dijk, Bijvoet Center for Biomolecular Research, Science Faculty, Utrecht University, Padualaan 8, 3584CH, Utrecht, The Netherlands

Björn Wallner, Center for Biomembrane Research and Stockholm Bioinformatics Center, Stockholm University, SE-106 91 Stockholm, Sweden

Kei Yura, Center for Informational Biology, Ochanomizu University, 2-1-1, Otsuka, Bunkyo, Tokyo 112-8610, Japan
Preface

Nucleotide sequences of nucleic acids (DNA or RNA) and amino acid sequences of proteins are the most widely used type of biological information. Nucleic acid sequences are relatively easy to determine experimentally, and can be used to predict the sequences of proteins they encode. Thanks to advances in high-throughput DNA sequencing, the number of sequences in public databases grows exponentially. Thereby, sequence databases became one of the largest data resources in biology, providing researchers with a wealth of information, and challenging computer scientists to develop tools that would allow for rapid discrimination between relevant and irrelevant information for various types of analyses.

While DNA sequence is mostly a carrier of genetic information, genes typically exert their function through the RNA and protein molecules they encode. A growing body of evidence points toward the importance of various RNA molecules that do not encode proteins, but fulfill various regulatory and/or catalytic roles. However, proteins that are produced by translation of coding RNAs remain the most important type of macromolecules required for the vital functions of the cell. Therefore, this book focuses on computational analyses of proteins, although many of these analyses are relevant also to nucleic acids (e.g. sequence alignment and prediction of protein-nucleic acid complexes).

Proteins perform most essential structural, enzymatic, transport, and regulatory functions in the cell. Protein functions are strictly determined by their structures, which can be organized into four levels of hierarchies with increasing complexity. These levels are: primary, secondary, tertiary, and quaternary structure. Above the structure of individual proteins and complexes there is another level of complexity, namely networks of interactions. The description of this hierarchy extends beyond the area of traditional structural bioinformatics and enters the realm of systems biology. The chapters in this book are meant to provide summaries of the concepts underlying protein sequence-structure-function relationships at these different levels of organization, and to serve as a comprehensive resource of state-of-the-art methods (as of 2008) for the corresponding bioinformatics analyses. The book is aimed at advanced graduate students, postdocs and faculty members. Although individual chapters have been written by several different authors and each can stand on its own, the book has been designed as a cohesive whole rather than a collection of independent reviews.

The primary structure corresponds to a linear sequence of amino acid residues linked together by peptide bonds into a polypeptide chain. The C-α atoms and peptide bond atoms form the main chain backbone, and other atoms protrude away as side chains. Functionally and evolutionarily relevant elements of primary structure are sequence motifs, domains and modules. In this volume, the bioinformatic tools for protein sequence analysis, including detection of motifs and domains, are reviewed in Chapter 1 by Kaminska et al., while...