Book reviews

Protein Structure Prediction: Bioinformatics Approach

Edited by Igor F. Tsigelny
International University Line, La Jolla, CA; ISBN 0 963 68177 X; 480 pp.; US$129.95; 2002

'Protein Structure Prediction: Bioinformatics Approach', edited by Igor F. Tsigelny, presents some of the most important ideas and state-of-the-art methods in the task of detecting and predicting protein's 3D structure, given its amino acid sequence. With the ever-increasing number of researchers and practitioners involved in protein studies, Tsigelny recognised the need to fortify the foundations of a field that still abounds with challenging problems, both theoretical and practical. The book contains very useful theoretical models, conceptual explanations, in-depth analyses and insights, but mainly it emphasises the algorithms and models that showed success in practical situations. It addresses a wide range of topics, on several levels; it provides the big picture of the field, subtleties of individual approaches, and predictions about its future.

There are 20 chapters in the book, written by 50 authors. Although the traditional divisions of fold-prediction methods distinguish among homology modelling, fold recognition and *ab initio* techniques, many approaches successfully coalesce principles of all three categories.

Chapter 1 describes PROSPECT, a fold recognition method that optimally aligns a target sequence (a protein sequence whose structure is needed) to a protein, or peptide, with a known fold based on a knowledge-based energy function. It then estimates the significance of the alignment. Chapters 2–4 implicitly exploit a broadly accepted model of hierarchical folding according to which proteins initially fold into local elements of secondary structure and then adopt a minimum free energy 3D conformation that stabilises both its secondary and tertiary structure. Chapter 2 presents a method where hidden Markov models (HMMs) are constructed over aligned secondary structure blocks. Using a probabilistic framework it recognises a fold model and then attempts to find the correct structural alignment within the subset of all structures. Approaches described in Chapters 3 and 4 further combine libraries of sequence profiles or HMMs constructed from aligned structural fragments to find the most likely local similarities. Then, for example, simulated annealing is used to move blocks in 3D space until the conformation hits the minimum energy state.

Chapter 5 discusses ways of using homology information for various tasks in structure prediction. Multiple sequence alignments are used to detect informative positions such as conserved positions or positions with correlated mutations. Informative positions are found to be spatially clustered and therefore can be used to guide threading. Chapter 6 presents another group of methods for fold recognition through sequence homology coupled with sequence-to-structure alignment. The approach from Chapter 7 exploits a commonly used similarity between speech recognition and homology (or fold) detection. Proteins are represented as time series of residue hydrophobicity indices and similarities are sought in the spectral domain. Chapters 8 and 9 further elaborate on the approaches based on the building block model. In Chapter 8, the hierarchical folding concept of proteins is mimicked and folding is assumed to be governed by electrostatic interactions, including hydrophobic effect. In Chapter 9, the authors combine information theory and algorithmic approaches. The building blocks are 'cores', ie parts of the fold conserved within a family, which are subsequently arranged using a combinatorial optimisation algorithm. A classical threading approach is presented in Chapter 10. A target sequence is aligned to each element in a library of structural templates with physically based scoring system. A widely used HMM-based software, SAM, is described in Chapter 11. This chapter covers the most important concepts of SAM used in homology detection, fold recognition and sequence alignment. An effective scheme for estimating statistical significance is outlined.

Chapter 12 starts a series of generally shorter chapters. It presents an interesting concept of incorporating genomic and expression data so as to improve detecting structural homology. The methods are based on an observation that structural homology tends to be conserved in nearby genes and genes with correlated expression patterns. Chapter 13 presents an approach for detecting and structurally characterising drug targets from a series of short, weakly significant peptide matches, experimentally verified for drug binding activity. Chapter 14 describes a few practical situations where fold recognition methods find their applications. Together with Chapters 15 and 16 it also describes techniques of combining models for improved prediction results. Although model combination is a well-known machine learning concept, 'averaging' server outputs is a non-trivial task. Details of the methods achieving expert-human quality are presented. Finally, Chapters 17–20 provide analyses and new algorithms of one typical 'block' of a structure prediction procedure – structure alignment block. New insights, algorithms, scoring functions and measures of quality are presented in these chapters.

'Protein Structure Prediction: Bioinformatics Approach' is a sound book from the perspectives of topic selection and of level of details contained in individual chapters. It is predominantly oriented toward presenting concepts and successful heuristics, but does not lack basic theory. I find it quite appropriate for such a new and dynamic field. It also gives many excellent insights by the authors and speculations about the future of the field. However, the book requires adequate fundamentals in bioinformatics or basic, occasionally advanced, familiarity with molecular biology, physics, machine learning, statistics and signal processing. It would be invaluable to molecular biologists and bioinformaticians regardless of whether they perform experimental research on individual proteins or large-scale computational studies. Understanding strengths and weaknesses of each particular method may be critical for interpreting results. Reading this book would also be a worthwhile experience for researchers entering bioinformatics and would help them to incorporate their own ideas into the field without duplicating previous efforts. It provides a high-quality text for special topic graduate seminars and could be used as supplemental material for a more comprehensive, graduate-level bioinformatics course.

This book, though excellent, should have included the role of intrinsically disordered (or unstructured) regions[1] in protein fold prediction. Except when bound to partners, disordered regions, especially the long ones, exist as ensembles of conformations under physiological conditions. Quite naturally, most of the authors demonstrate the success of their approaches. However, their analyses would have been even more informative if they had provided the reader with estimates of how often their methods produced poor structures or how often they were unable to produce statistically significant outputs. It would have been interesting if the structural form of such proteins (if known) had been provided. If, as estimated, ~10 per cent of residues in nature are in fact disordered,[2] then a sizeable fraction of proteins may create difficulties for the current techniques. Another topic this book would have benefited from is a review of the techniques based on molecular dynamics simulations. Although lagging behind the mainstream bioinformatics methods in terms of prediction accuracy and time, inclusion of molecular dynamics would have provided completeness of the topics covered as well as important comparisons.

To conclude, this book fully meets high scientific standards. Prediction of protein folding is a cornerstone of bioinformatics and therefore this book represents an important contribution to the field. For me, it was a rewarding read.

Predrag Radivojac, PhD
Center for Computational Biology and Bioinformatics
Indiana University School of Medicine

References

1. Wright, P. E. and H. J. Dyson (1999), 'Intrinsically unstructured proteins: Re-assessing the protein structure–function paradigm', *J. Mol. Biol.*, Vol. 293, pp. 321–331.
2. Romero, P., Obradovic, Z. and Dunker, A. K. (2000), 'Intelligent data analysis for protein disorder prediction', *Artif. Intel. Rev.*, Vol. 14, pp. 447–484.

Bioinformatics for Geneticists

Michael R. Barnes and Ian C. Gray
John Wiley and Sons, Chichester; ISBN 0 470 84394 2; 408 pp.; £45.00; 2003

'Bioinformatics for Geneticists' is a recently published book edited by Michael Barnes and Ian Gray, both at GlaxoSmithKline Pharmaceuticals, UK. This introductory textbook attempts to give geneticists a bioinformatics toolkit that they may not have acquired in traditional genetics training. For the most part, this book succeeds where it tries, and comes out as a useful addition to the library of a seasoned scientist, or even as a text for a graduate level course in genetic bioinformatics. There is extensive coverage here, from the basics, including internet databases and resources to the more complex, such as functional analysis of splice variation. The focus is quite applied: the book will not tell you how write your own version of your favourite program or algorithm, but it will probably tell you how and where to use it. Generally, the authors take what they call a web-based approach to bioinformatics, highlighting free databases and software available on the internet.

Opening the book is an introductory chapter by the editors introducing their version of 'genetic bioinformatics'. The chapter adequately introduces the material and puts the rest of the book into perspective. This is followed by a chapter focusing solely on internet resources. While having sections on what the best search engine is and how to use it may be a bit too broad, there is adequate coverage of many resources, some of which are less well known but still useful. Following this is a chapter devoted solely to human genetic variation databases and related resources. Coverage here is thorough, ranging from NCBI's dbSNP to mutation databases such as HGVBase and the HGMD. There are also links to marker databases such as dbSTS, UniSTS and the GDB. Once again the focus here is on tools and resources, not necessarily on the details of the underlying theory, although each chapter is furnished with many references. Finally this broad, introductory section is finished with a chapter focusing on finding and analysing genes. Included here is an introduction to genome data and the golden path assembly as well as other resources for genome sequence analysis.

The second section of the book focuses on complete genomes and their use in genetics. Here there are extensive lists of available genome sequences, visualisation tools and software. Included in this chapter are introductions to the powerful