PROTEIN STRUCTURE PREDICTION
Bioinformatic Approach

edited by
IGOR F. TSIGELNY

Preface:

Prediction of protein structure is very important today. Whereas more than 17,000 protein structures are stored in PDB, more than 110,000 proteins are stored only in SWISSPROT. The ratio of solved crystal structures to a number of discovered proteins to about 0.15, and I do not see any improvement of this value in the future. At the same time development of genomics has brought an overwhelming amount of DNA sequencing information, which can be and already is used for constructing the hypothetical proteins. This situation shows the great importance of protein structure prediction. The field is growing very rapidly. A simple analysis of publications shows that the number of articles having the words 'protein structure prediction' has almost doubled since 1995.

So many really great ideas are used as a basis for the current prediction systems. Some of them will evolve into the next generation of the prediction software but some of them, even very promising, will be lost and rediscovered in the future. Here we tried to include the variety of methods representing the most interesting concepts of current protein structure prediction.

This compendium of ideas makes this book an invaluable source for scientists developing prediction methods. In many cases authors describe successful prediction methods and programs that make this book an invaluable source of information for numerous users of prediction software.

The first chapter describes the protein structure prediction program PROSPECT that produces a globally optimal threading alignment for a typical threading energy function, and allows users to easily incorporate experimental data as constraints into the threading process. PROSPECT also provides a confidence assessment of a threading result based on a neural network. The second chapter presents a protein fold-recognition method that selects the best fold model for a given protein sequence from a library of structural hidden Markov models (HMMs). The HMMs are built from protein structures following their modular decomposition into the secondary structure elements and representing those elements by a pre-designed set of submodels. The third chapter describes a method to fold proteins into simplified three-dimensional structures constructed from small fragments cut out of a representative set of known three-dimensional structures. The three-dimensional protein structures and fragments are represented in a simplified form as a sequence of angle pairs, one angle pair per residue. Chapter 4 describes the application of HMM constructed on the basis of structural alignments for protein structure prediction. An example system HMM-SPECTR is given with the description of different types of HMMs based on structural alignments. Chapter 5 reviews the different

methods of extraction of information from multiple sequence alignments and illustrates how to use them as a primary source of information. The chapter describes the application of rarely used features such as sequence conservation, variations between sub-families, correlation between the patterns of mutation of pairs of positions, and the distribution of apolar residues for structure prediction. Chapter 6 illustrates how knowledge of protein three-dimensional structure can be used to identify homologues of known structure, generate sequence-structure alignments and assist model building. It describes the programs: HOMSTRAD, a database of structure-based alignments for protein families of known structure, JOY, a program to annotate local environments in structure-based alignments and FUGUE, a program to perform sequence-structure homology recognition. Chapter 7 proposes a different concept of sequence homology. This concept is derived from a periodicity analysis of the physicochemical properties of the residues constituting proteins primary structures. The analysis is performed using a front-end processing technique in automatic speech recognition by means of which the cepstrum (measure of the periodic wiggliness of a frequency response) is computed that leads to a spectral envelope that depicts the subtle periodicity in physicochemical characteristics of the sequence.

Chapter 8 describes the building block protein folding model. Via a building block assigning algorithm, sequence comparisons and weighting scheme, building blocks are assigned to a target protein sequence. The problem of the ‘building block’ is very important for both protein folding modeling and protein structure prediction. Authors of several chapters in this book propose different ‘building blocks’ for discretization of the prediction process. In most cases they do not discuss the physical properties of these blocks, paying attention only to the information coding. The approach of the authors of the chapter 8 clearly defining the physical and informational properties of the building blocks looks very promising.

Chapter 9 describes a new fold recognition method called FROST (Fold Recognition Oriented Search Tool). It includes 1D and 3D comparison and a database of representative three-dimensional structures. The chapter uses information theory concepts for embedding of a number of sequence and structure parameters in one scoring function. This approach makes this chapter very elegant and useful for the developers of protein structure prediction systems. Chapter 10 continues the discussion of how to combine different levels of resolution and representation of a protein and the rationalizations of score functions for protein structure prediction. The statistical mechanical parameters are used together with purely empirical and even ad hoc parameters.

Chapter 11 describes one of the most effective HMM system for biological applications—SAM. The chapter shows an approach to fold recognition that relies on HMMs for both selecting the template and for aligning the target to the template. The technique has been used successfully in three of the Critical Assessment of Structure Prediction (CASP) experiments.

Chapter 12 discusses the important link between genomic information and protein structure. The chapter describes the clues that could be used to help infer the evolutionary

relationship via structural similarity and improve the ability to predict the biochemical function. The first such clue is a positional conservation along the genome, i.e., nearby genes tend to be structurally related more often than expected by chance alone. The second such clue is present in expression data: genes that are correlated in expression are more apt to share a common fold than two randomly chosen genes.

Chapter 13 proposes a comprehensive system for computer based drug design. The chapter describes the program HMM-ELONGATOR, which predicts putative protein targets based on a set of peptides shown to bind a drug molecule from combinatorial libraries.

Chapter 14 on the basis of three examples shows how the use of fold recognition helped biologists in planning and devising experiments and in generating verifiable hypotheses. This chapter describes the meta-predictor approach for protein structure prediction. Chapter 15 describes in details the Structure Prediction Meta Server that collects prediction models from many high quality services and translates them into standard formats enabling convenient analysis of the results. The Meta Server offers an infrastructure for the creation of automated jury algorithms, which analyze the set of results for the user and calculate the reliability score for a consensus prediction. Chapter 16 describes a new method for fold recognition, Pcons that utilizes the "consensus analysis." This chapter shows the advantages of Pcons based on the large scale benchmarking.

Chapter 17 starts the part of the book devoted to the concepts of structural alignment. It is obvious that proper structural alignment of proteins is the cornerstone of the majority of prediction methods. This chapter introduces several new views of protein fold space which will help to further understand protein evolution and interpret structural similarities. Differences between the manual (SCOP) and automated (CE) approaches to the structural classification problem are described. Chapter 18 discusses the design principles of a structure alignment system that can be used for structure prediction assessments. This system is based on a hierarchical representation of a protein shape. Such a representation makes the system suitable for effective alignments of structures with low similarity. Chapter 19 describes the Monte-Carlo approach to the construction of multiple structural alignments. Chapter 20 describes the specific example, where an alignment of eukaryotic protein kinases generated using the combinatorial extension algorithm (CE) is compared with a manually derived alignment. Implications for CE are discussed, as well as implications for automated structural alignment in general.

Overwhelmed by current errands, proposals, and papers, we mostly do not think in global terms of our place in building of knowledge, building of science. Nevertheless it is going on and in one way or another we build the structure of scientific knowledge. If the scientific articles are the 'bricks' in this building, books are the cornerstones.

I would like to thank all the authors for devoting their time to the writing of the chapters. I hope this book will be useful to professionals and students in the field.