An Improved Statistical Model for Protein Secondary Structure Prognostication

Aswathy. Ravikumar, Saritha. R

Abstract— Genome sequencing projects continue to provide a flood of new protein sequences. Recently there have been advances in protein structure prognostication which in turn has improved fold recognition algorithms. Predicting the secondary structure of proteins is important in biochemistry because the 3D structure can be determined from the local folds that are found in secondary structures. Moreover, knowing the tertiary structure of proteins can assist in determining their functions. The problem of protein secondary structure prognostication with Hidden Markov Models is addressed here. Sequence family information is integrated via the combination of independent predictions of homologous sequences and a weighting scheme. Hidden Markov models were built for a representative set of just over 1,000 structures from the Protein Data Bank (PDB). The topology of the HMM was restricted to biologically meaningful building blocks.

Index Terms— HMM, protein structure, Markov Process

I. INTRODUCTION

A protein is a polymeric macromolecule made of amino acid building blocks arranged in a linear chain and joined together by peptide bonds. The linear polypeptide chain is called the primary structure of the protein. The primary structure is typically represented by a sequence of letters over a 20-letter alphabet. The function of a protein strongly depends of its 3D-structure. For instance, enzymes need to have tight spatial complementarities with their substrate. Thus knowledge of a protein structure gives relevant clues to its function. The prediction of the three-dimensional structure of a protein when only the amino-acid sequence is known has been a problem of major interest for many years .Approaches have ranged from purely *ab-initio* methods that are based entirely on physical chemical principles, to homology methods that are based primarily on the information available in sequence and structural databases. Secondary structure predictions may also be used to guide the design of site directed mutagenesis studies, and to locate potential functionally important residues[2]. The problem tackled is to provide a label for each residue in a protein sequence depending on its secondary structure. That is, whether the protein residue is part of an alpha-helix, a beta-sheet or some other structure. This is a first step towards predicting the structure and function of a protein from its sequence. Many machine learning methods have been applied to this neural networks [3, 4] and more recently support vector machines [5, 6].

The secondary structure is specified by a sequence classifying each amino acid into the corresponding secondary structure element (e.g., alpha, beta, or gamma). The secondary structure elements are further packed to form a tertiary structure depending on hydrophobic forces and side chain interactions, such as hydrogen bonding, between amino acids. The tertiary structure is described by the, and coordinates of all the atoms of a protein or, in a more coarse description, by the coordinates of the backbone atoms. Finally, several related protein chains can interact or assemble together to form protein complexes. These protein complexes correspond to the protein quaternary structure. The quaternary structure is described by the coordinates of all the atoms, or all the backbone atoms in a coarse version, associated with all the chains participating in the quaternary organization, given in the same frame of reference.

Protein structure prediction software is becoming an important proteomic tool for understanding phenomena in modern molecular and cell biology [7] and has important applications in biotechnology and medicine. Machine learning methods [8] that can automatically extract knowledge from the PDB are an important class of tools and have been widely used in all aspects of protein structure prediction. Here, the development and application of machine learning methods in protein structure prediction is discussed.

In spite of progress in robotics and other areas, experimental determination of a protein structure can still be expensive, labour intensive, time consuming, and not always possible. Some of the hardest challenges involve large quaternary complexes or particular classes of proteins, such as membrane proteins which are associated with a complex lipid bilayer environment. These proteins are particularly difficult to crystallize. Although membrane proteins are extremely important for biology and medicine, only a few dozen membrane protein structures are available in the PDB. Thus, in the remainder of this paper we focus almost exclusively on globular, non membrane proteins that are typically found in the cytoplasm or the nucleus of the cell, or that are secreted by the cell. methods in 1-D, 2-D, 3-D, and 4-D structure prediction. We focus primarily on supervised machine learning method hidden Markov models (HMMs).

II. PROTEIN STRUCTURE PROGNOSTICATION

A. Proteins

Proteins are made of simple building blocks called amino acids. An amino acid is a compound consisting of a carbon atom to which are attached a primary amino group, a carboxylic acid group, a side chain (R group), and an H atom. Also called an α amino acid. There are 20 different amino acids that can occur in proteins. Their names are abbreviated in a three letter code or a one letter code.

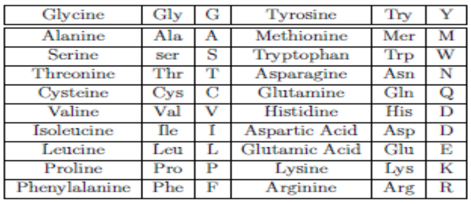
Manuscript received November 20, 2012.

Aswathy.Ravikumar, Department of Computer Science and Engineering, College of Engineering, Trivandrum, Kerala, India.

Saritha.R, Department of Computer Science and Engineering, College of Engineering, Trivandrum, Kerala, India.

Proteins are important for living organisms. For example, our body fabric is made of protein molecules. Some of the proteins are found in the food we eat. Proteins also help to build their own protein molecules. They serve as hormones, receptors, storage, enzymes and as transporters of particles in our bodies. The amino acids and their letter codes are given in Table 1

TABLE I AMINO ACIDS



Glycine	Gly	G	Tyrosine	Try	Y
Alanine	Ala	A	Methionine	Mer	M
Serine	ser	S	Tryptophan	Trp	W
Threonine	Thr	T	Asparagine	Asn	N
Cysteine	Cys	C	Glutamine	Gln	Q
Valine	Val	V	Histidine	His	D
Isoleucine	Ile	I	Aspartic Acid	Asp	D
Leucine	Leu	L	Glutamic Acid	Glu	E
Proline	Pro	P	Lysine	Lys	K
Phenylalanine	Phe	F	Arginine	Arg	R

B. Protein Structure

There are four different structure types of proteins, namely Primary, Secondary, Tertiary and Quaternary structures. Primary structure refers to the amino acid sequence of a protein. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. The two ends of the polypeptide chain are referred to as the carboxyl terminus (C-terminus) and the amino terminus (N-terminus) based on the nature of the free group on each extremity. Counting of residues always starts at the N-terminal end (NH_2-group), which is the end where the amino group is not involved in a peptide bond. The primary structure of a protein is determined by the gene corresponding to the protein. It provides the foundation of all the other types of structures. Secondary structure refers to the arrangement of connections within the amino acid groups to form local structures. α helix, β strand are some examples of structures that form the local structure. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups.Tertiary structure is the three dimensional folding of secondary structures of a polypeptide chain. The alpha-helices and beta-sheets are folded into a compact globule. The folding is driven by the non-specific hydrophobic interactions ,but the structure is stable only when the parts of a protein domain are locked into place by specific tertiary interactions, such as salt bridges, hydrogen bonds, and the tight packing of side chains and disulfide bonds Quaternary structure is formed from interactions of several independent polypeptide chains. The four structures of proteins are shown in Figure 1.

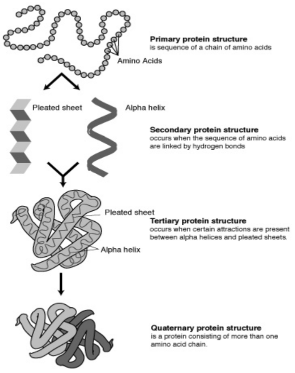

Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present between alpha helices and pleated sheets.
Alpha helix
Quaternary protein structure
is a protein consisting of more than one amino acid chain.

Fig. 1 Protein Structure

C. Secondary structure prediction/prognostication

Secondary structure prediction means predicting the secondary structure of a protein from its primary sequence. It is important because knowledge of secondary structure helps in the prediction of tertiary structure. This is very interesting for proteins whose sequences do not show any similarities with the sequences of proteins in the database. Secondary structure has two properties, hydrogen bond patterns, and backbone geometry. Hydrogen bonded features include turns, bridges, α helices, β ladders and β sheets while bends, hirality, SS bonds and Solvent exposure are features which are determined geometrically [9].

Some of the computationally-based methods that can be used to achieve the secondary predictions include Neural Networks, Support Vector Machines and Nearest Neighbour Methods. Here we discuss a optimised method HMM- Hidden Markov Model.

III. HIDDEN MAKOV MODEL

The Hidden Markov Model (HMM) is a powerful statistical tool for modelling generative sequences that can be characterised by an underlying process generating an observable sequence. HMMs have found application in many areas interested in signal processing, and in particular Speech processing, but have also been applied with success to low level NLP tasks such as part-of-speech tagging, phrase chunking, and extracting target information from documents. Profile hidden Markov models (HMMs) are similar to simple sequence profiles, but in addition to the amino acid frequencies they contain the position specific probabilities for inserts and deletions along the multiple sequence alignment. The logarithms of these probabilities are in fact equivalent to position specific gap penalties [10]. Not surprisingly, profile HMMs perform better than sequence profiles in the detection of homologous proteins and in the quality of alignments [11,12], but despite the success of profile-profile lignment methods, the generalization to HMM-HMM comparison has not been done until recently

A. Markov model

Markov model is a finite set of states, each of which is associated with a (generally multidimensional) probability distribution. Transitions among the states are governed by a set of probabilities called transition probabilities. In a particular state an outcome or observation can be generated, according to the associated probability distribution. It is only the outcome, not the state visible to an external observer and therefore states are ``hidden" to the outside; hence the name Hidden Markov Model. The formal definition of a HMM is as follows: $\lambda = \{\pi, A, B\}$

- π= Initial Probabilities
- A = Transition Matrix
- B= Emission Matrix
- k= number of states
- n= number of distinct observation symbols
- Π is $k \times 1$
- A is $k \times k$
- B is $k \times n$

B. The HMM Library

Our model library included about 1,000 (now 1,312) structures from PDB, the core of which was a representative set of PDB structures. For each of these structures we constructed an HMM using the associated HSSP alignment of the structure and its homologs as the initial basis. This alignment and the corresponding HMM parameters were re-estimated using standard HMM methods in combination with priors over amino acids and transition probabilities in various structural environments. The transition priors allowed us to incorporate general structural information, such as the low.

C. HMM: Application to Secondary Structure

In a Markovian sequence, the character appearing at position t only depends on the k preceding characters, k being the order of the Markov chain. Hence, a Markov chain is fully defined by the set of probabilities of each character given the past of the sequence in a k-long window: the transition matrix. In the hidden Markov model, the transition matrix can change along the sequence. The choice of the transition matrix is governed by another Markovian process, usually called the hidden process. Hidden Markov models are thus particularly useful to represent sequence heterogeneity. These models can be used in predictive approaches: some algorithms like the Viterbi algorithm and the forward-backward procedure allow to recover which transition matrix was used along the observed sequence.

The hidden process to be recovered is the secondary structure of the protein. The observed process is the amino-acid sequence. The hidden chain process is a first order Markov chain. Each hidden state is characterized by a distribution of amino-acids. Due to the large alphabet size, the order of the observed chain is 0, which means that amino-acids are independent conditionally on the the hidden process. The prediction is achieved with the forward/backward algorithm. The simplest model for three classes prediction is a HMM with three hidden states, each state accounting for a secondary structure class.

- *Model of α-helices:* A well-characterized sequence motif in α helices is the amphiphilic motif, i.e., a succession of two polar residues and two a polar residues.

- *Model of β-strand:* There is no strong motif characterizing β-strands similar to the amphipatic motif for α-helices. Characteristic motifs are found using a statistical approach based on exceptional words.

D . Result

As shown in fig 2 the amino acid sequence was classified as α helix, β sheet, coil.

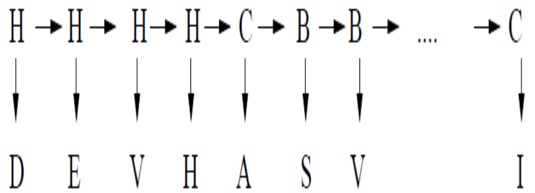


Fig. 2 Secondary structure prediction via a hidden Markov model. The upper line represents the secondary structure along a protein sequence: H for a residue in α-helix, B for β-strand, C for coil. The lower line represents the amino acid sequence of the protein. This is the observed sequence.

IV. CONCLUSIONS

For many years, protein-structure prediction was viewed as an important but distant goal. This perspective has changed. Dramatically with the recent explosion of sequence and structural information and because of computational advances in many different areas. These include pure sequence analysis, structure-based sequence analysis, the conformational analysis of proteins and the understanding of the energetic determinants of protein stability. Perhaps the most significant conclusion that can be reached from the presented results, is that a very simple method for secondary prediction based on a straightforward on method HMM. Since HMMs do not use pair wise contacts, they are more computationally efficient than threading models. Their minimal dependency on structure information also allows them to be used to search for remote homologs of protein families that contain no sequence with known structure. It may turn out that other techniques, which make real use of structural information, may be able to do better at finding very distant homologs, but we feel that the HMM methods can still be improved enough to remain competitive with more expensive methods.

REFERENCES

1. Moult J, Hubbard T, Fidelis K, Pedersen J: Critical assessment of methods of protein structure prediction (CASP): Round III.*Proteins* 1999, 37(suppl 3):2-6.
2. Zvelebil MJJM, Barton GJ, Taylor WR, Sternberg MJE. Prediction of protein secondary and active sites using the alignment of homologous sequences. J Mol Biol 1987;195:957–961.
3. G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, “Improving the Prediction of Protein Secondary Structure in Three and Eight Classes Using Recurrent Neural Networks and Profiles,” *PROTEINS: Structure, Fuction,and Genetics*, vol. 47, pp. 228–235, 2002.
4. K. Lin, V. A. Simossis, W. R. Taylor, and J. Heringa,“A simple and fast secondary structure prediction method using hidden neural networks,” *Bioinformatics*,vol. 21, no. 2, pp. 152–159, 2005.
5. J. J. Ward, L. J. McGuffin, B. F. Buxton, and D. T. Jones, “Secondary structure prediction with support vector machines,” *Bioinformatics*, vol. 19, no. 13, pp.1650–1655, 2003.
6. J. Guo, H. Chen, Z. Sun, and Y. Lin, “A NovelMethod for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles,” PROTEINS: Structure, function, and Bioinformatics, vol. 54, pp.738–743, 2004.
7. Petrey and B. Honig, “Protein structure prediction: Inroads to biology,”Mol. Cell., vol. 20, pp. 811–819, 2005.
8. P. Baldi and S. Brunak, Bioinformatics: The Machine Learning Approach,2nd ed. Cambridge, MA: MIT Press, 2001M. Young, The Techincal Writers Handbook. Mill Valley, CA: University Science, 1989.
9. Chandonia, J.M. & Karplus, M. (1995). Neural Networks for Secondary Structure and Structural Class Prediction. Protein Sci. 4: pp. 275-285.
10. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (1998) Biological sequence analysis: Probabilistic models of proteins and nucleic acids, Cambridge University Press, Cambridge.
11. Krogh, A., Brown, M., Mian, I. S., Sjölander, K., and Haussler, D. (1994) Hidden markov models in computational biology. Applications to protein modeling. J.Mol.Biol.,235,1501–1531.
12. Eddy, S. R. (1998) Profile hidden markov models. Bioinformatics,14,755–763

AUTHORS PROFILE

Aswathy.Ravikumar, 2nd year M.Tech student, Computer Science and Engineering, College of Engineering ,Trivandrum. Research Interests: Machine Learning, Bioinformatics, Software Engineering .

Saritha.R , M.Tech in computer Science & Engineering with specialization in Digital Image Computing ,has been the field of Technical education for almost 9 years, currently working as Assistant Professor in Computer Science and Engineering at College of Engineering, Trivandrum. Research Interests :Machine Learning, Pattern Recognition, Image Processing