International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:2, No:6, 2008

Protein Secondary Structure Prediction

Manpreet Singh, Parvinder Singh Sandhu, and Reet Kamal Kaur

Abstract**—**Protein structure determination and prediction has been a focal research subject in the field of bioinformatics due to the importance of protein structure in understanding the biological and chemical activities of organisms. The experimental methods used by biotechnologists to determine the structures of proteins demand sophisticated equipment and time. A host of computational methods are developed to predict the location of secondary structure elements in proteins for complementing or creating insights into experimental results. However, prediction accuracies of these methods rarely exceed 70%.

Keywords**—**Protein, Secondary Structure, Prediction, DNA, RNA.

I. The Genetic Material

DNA is the main constituent of genetic material with in a body. DNA is converted into RNA and then to Protein in the process of Gene Expression.

A. DNA (Deoxyribonucleic Acid)

DNA (deoxyribonucleic acid) is the genetic material. This is a profoundly powerful statement to molecular biologists. It is the information stored in DNA that allows the organization of inanimate molecules into functioning of living cells and organisms that are able to regulate their internal chemical composition, growth, and reproduction. As a direct result, it is also what allows us to inherit our mother's curly hairs, our father's blue eyes, and even our uncle's too large nose. The various units that govern those characteristics at the genetic level, be it chemical composition or nose size, are called genes [6].

B. RNA (Ribonucleic Acid)

RNA is a nucleic acid consisting of nucleotide monnomers that plays several important roles in the processes that translate genetic information from deoxyribonucleic acid (DNA) into protein products; RNA acts as a messenger between DNA and the protein synthesis complexes known as ribosomes, forms vital portions of ribosomes, and acts as an essential carrier molecule for amino_acids to be used in protein synthesis. RNA is very similar to DNA, but differs in a few important structural details: RNA is single stranded, while DNA is double stranded. Also, RNA nucleotides contain ribose sugars while DNA contains deoxyribose and RNA uses predominantly uracil instead of thymine present in DNA. RNA is transcribed from DNA by enzymes called RNA polymerases and further processed by other enzymes. RNA serves as the template for translation of genes into proteins, transferring amino acids to the ribosome to form proteins, and also translating the transcript into proteins.

Messenger RNA (mRNA): Messenger RNA is RNA that carries information from DNA to the ribosome sites of protein synthesis in the cell. In eukaryotic cells, once mRNA has been transcribed from DNA, it is "processed" before being exported from the nucleus into the cytoplasm, where it is bound to ribosomes and translated into its corresponding protein form with the help of tRNA. In prokaryotic cells, which have not partition into nucleus and cytoplasm compartments, mRNA can bind to ribosomes while it is being transcribed from DNA. After a certain amount of time the message degrades into its component nucleotides, usually with the assistance of ribonucleases.

Transfer RNA (tRNA): Transfer RNA is a small RNA chain of about 74-95 nucleotides that transfers a specific amino acid to a growing polypeptide chain at the ribosomal site of protein synthesis during translation. It has sites for amino-acid attachment and an anticodon region for codon recognition that binds to a specific sequence on the messenger RNA chain through hydrogen bonding. It is a type of non-coding RNA.

II. Proteins and their Structure

Proteins are complex organic compounds that consist of amino acids joined by peptide bonds. Proteins are essential to the structure and function of all living cells and viruses. Many proteins function as enzymes or form subunits of enzymes. Some proteins play structural or mechanical roles. Some proteins function in immune response and the storage and transport of various ligands. Proteins serve as nutrients as well; they provide the organism with the amino acids that are not synthesized by that organism.

Proteins are amino acid compounds and the composition of amino acids in a protein defines the three dimensional form that the protein folds to. These structures are unique in the sense that a given sequence of amino acids always folds into almost the same structure under the same environmental conditions (pressure, temperature, pH etc. There are exceptions but that is very rare.). Structures of proteins are investigated under four primary groups as shown in Fig. 1.

Primary Structure is the sequence of amino acids in the protein. Secondary Structure is the composition of common patterns in the protein. Some patterns are frequently observed

Manpreet Singh is with the Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana (e-mail: mpreet78@yahoo.com).

Dr. Parvinder Singh Sandhu is with Department of CSE, Rayat and Bahara Institute of Engineering and Technology, Ropar (e-mail: parvinder.sandhu@gmail.com).

Reet Kamal Kaur is M.Tech Student in Department of CSE & IT, Guru Nanak Dev Engineering College, Ludhiana (e-mail: reetkamal1901@yahoo.co.in).

in the native states of proteins. This structure class includes regions in the protein of these patterns but it does not include the coordinates of residues. Tertiary Structure is the native state, or folded form, of a single protein chain. This form is also called the functional form. Tertiary structure of a protein includes the coordinates of its residues in three dimensional space. Quaternary Structure is the structure of a protein complex. Some proteins form a large assembly to function. This form includes the position of the protein subunits of the assembly with respect to each other.

There are a number of methods with varying resolution to determine the structure of proteins. For example, the primary structure can be determined by means of mass spectrometry, the secondary structure content (i.e. percentages of the common motifs) can be determined up to some certainty by means of circular diachronic spectroscopy and the tertiary structure can be determined by means of x-ray crystallography or NMR spectroscopy. These methods require more time and effort as the expected resolution from the method increases.

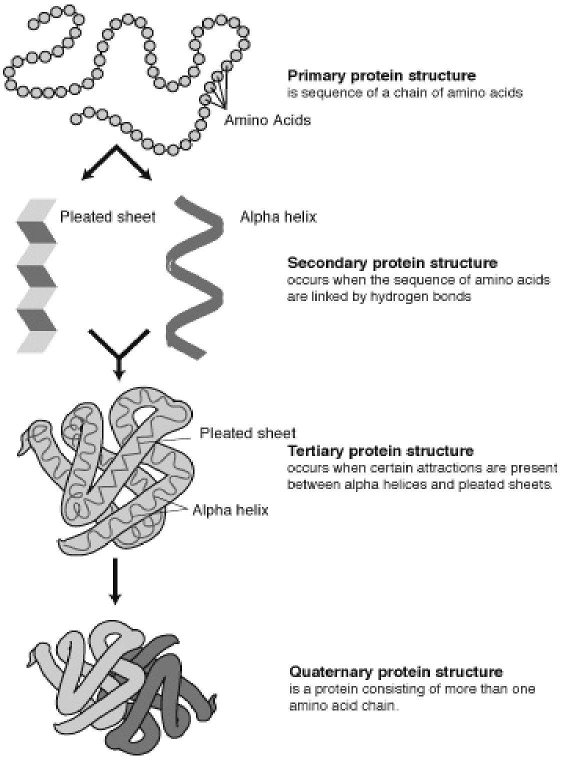
Primary protein structure
is sequence of a chain of amino acids
Amino Acids
Pleated sheet
Alpha helix
Secondary protein structure
occurs when the sequence of amino acids are linked by hydrogen bonds
Pleated sheet
Tertiary protein structure
occurs when certain attractions are present between alpha helices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

Fig. 1 Different representations of protein structure

There are also theory based methods in protein structure determination like homology modeling, threading or ab initio modeling. These methods are referred to as structure prediction methods. Homology modeling can be briefly described as fitting a known sequence to the experimentally determined three dimensional structure of a protein that is similar in sequence. Threading is fitting a sequence to a database of known structures using a heuristic scoring method and finding the most likely structure. Ab initio methods are methods that predict structure from scratch, i.e. they do not rely on known structure of the homologous proteins.

III. Secondary Structure Prediction

Given a protein sequence with amino acids a1a2 . . . an, the secondary structure prediction problem is to predict whether each amino acid ai is in an α−helix, a β−sheet, or neither. If you know (say through structural studies), the actual secondary structure for each amino acid, then the 3-state accuracy is the percent of residues for which your prediction matches reality. It is called “3-state” because each residue can be in one of 3 “states”: α, β, or other (O). Because there are only 3 states, random guessing would yield a 3-state accuracy of about 33% assuming that all structures are equally likely. There are different methods of prediction with various accuracies. Some of these methods are discussed below.

A. PHD

PHD [8] is the first method to break the 70% boundary on Q3 accuracies of secondary structure prediction methods. The two-level neural network structure in this work has been adopted by several other methods later (such as JNet [2] and PSIPRED [3]).

In this work, the authors prepared a set of non-homologous proteins and named it the RS126 set after the initials of their names (Rost and Sander) and the number of protein chains in it. If two proteins are at least 80 residues long and 25% of their sequences are identical, they are considered to be homologues and only one of them has been included in the set. For a reasonable measurement of the performance of the algorithm, it is necessary to have a non-redundant set of proteins.

After a non-redundant set of sequences is compiled, the multiple sequence alignments and secondary structure assignments of each sequence have been retrieved from a database called HSSP [5]. The secondary structure definition algorithm used in this work is DSSP [4] (DSSP is the method of secondary structure definition in HSSP). The 8-to-3 state reduction scheme used is given in Table I.

PHD uses a two layered feed-forward neural network [2] for sequence-to-structure prediction. The input to this network is a frame of 13 consecutive residues. Each residue is represented by the frequencies of residues in the column of multiple sequence alignment which corresponds to that residue. That is to say, the residues in the homologous proteins that correspond to the residue in the query protein are selected and frequencies of each type of residue are calculated and input to the network. This means each residue introduces 20 inputs to the neural network. Also, one more input is used for each residue in the frame for the cases that the frame extends over the N or C terminus of the protein. One final input is added for each residue called the conservation weight [8]. This weight represents the quality of a multiple sequence alignment (i.e. the number of aligned sequences and the similarity of the residues at that position in the alignment). So every residue is represented by 20+1+1=22 inputs, thus the sequence-to-structure network has 13x22 input nodes. The output of this network is 3 weights, one for each of the helix, strand and loop states.

TABLE I
8-TO-3 STATE REDUCTION SCHEME USED IN PHD METHOD

Reduction	DSSP Code	Description
H	H	Alpha-helix
H	G	3-helix (3/10 helix)
H	I	5-helix (pi helix)
C	B	residue in isolated beta-bridge
E	E	Extended strand, participates in beta ladder
C	T	Hydrogen bonded turn
C	S	Bend
C	(no code)	Loop/coil

The structure-to-structure prediction part of the algorithm is also implemented as a two layered feed-forward network. This time the input to the network is a frame of 17 consecutive residues. Each residue is represented by the 3 weights from the output of sequence-to-structure part plus one other weight for the cases that the frame extends over the N or C terminus of the protein. The conservation weights are added here too. This means each residue is represented by 3+1+1=5 nodes and this makes a total of 17x5 input nodes to the structure-to-structure network. Output of this step is again 3 weights for each of the possible states.

B. JNet

JNet [2] algorithm uses the same network structure used in PHD method. The difference of this algorithm is that it utilizes an expanded set of protein chains, another 8-to-3 state reduction scheme and a number of new methods for generating multiple sequence alignments.

TABLE II
8-TO-3 STATE REDUCTION SCHEME USED IN JNET METHOD

Reduction	DSSP Code	Description
H	H	alpha-helix
C	G	3-helix (3/10 helix)
C	I	5-helix (pi helix)
E	B	Residue in isolated beta-bridge
E	E	extended strand, participates in beta ladder
C	T	hydrogen bonded turn
C	S	Bend
C	(no code)	Loop/coil

The multiple sequence alignments in this method have been obtained by running PSI-BLAST searches on different databases and by aligning the sequences using different techniques. The secondary structure definition algorithm used in this work is also DSSP [4]. The 8-to-3 state reduction scheme used is given in Table II.

The sequence-to-structure part of this algorithm is, like PHD [8], a neural network. In this case the input frame is 17 residues long. At this step the various networks were trained which utilizes different representations of the columns of multiple sequence alignments.

C. PSIPRED

PSIPRED [3] is a neural network based method, which has three components. The difference of this method is that it conducts homology searches on a different database and uses a different set of proteins for training and testing. It also represents the multiple sequence alignments only as PSI-BLAST position specific scoring profiles.

The network structure is simplified with respect to PHD and JNet methods. The sequence-to-structure part of the method is a back-propagation neural network. The input to this part is a frame of 15 residues. The residues are represented by the PSI-BLAST scoring matrices. This neural network has 75 hidden nodes and 3 output nodes.

The output of the sequence-to-structure network is fed to the structure-to-structure network in frames of 15 residues. This network has 60 hidden nodes and 3 output nodes for the final prediction.

The performance of this method is not directly comparable to PHD or JNet since the same data set with those methods was not utilized during its development. Its Q3 accuracy is 76.5%. This method has, however, proven to be more successful than the others in the third Critical Assessment of Techniques for Protein Structure Prediction (CASP) [15,16] experiment [3].

D. GORV

GORV [5] is a secondary structure prediction method based on information theory and Bayesian statistics. Unlike other methods mentioned previously, this method does not use real valued encodings of multiple sequence alignments.

GORV uses the CB513 [1] data set. Secondary structure assignments were taken from DSSP. The 8-to-3 state reduction scheme used is given in Table III. This scheme does not take into account the 3/10 helices, which are not so rare1 (3%). Thus the published results are not comparable with the other methods using CB513 set. We have checked to see that this reduction scheme may add at least 2.44% to the performance of a prediction using one of the other reduction schemes and exactly the same methods other than that (same training algorithm, same multiple sequence alignments etc.).

The sequence-to-structure part of this method has 66.9% single-sequence Q3 accuracy 2. When multiple sequence alignments are incorporated to the algorithm, the accuracy rises to 73.4%. The individual contribution of the filtering part is not stated.

TABLE III
8-TO-3 REDUCTION SCHEME USED IN GORV METHOD

Reduction	DSSP Code	Description
H	H	alpha-helix
C	G	3-helix (3/10 helix)
C	I	5-helix (pi helix)
C	B	residue in isolated beta-bridge
E	E	Extended strand, participates in beta ladder
C	T	hydrogen bonded turn
C	S	Bend
C	(no code)	Loop/coil

E. Chou-Fasman Method

If you were asked to determine whether an amino acid in a protein of interest is part of a α-helix or β-sheet, you might think to look in a protein database and see which secondary structures amino acids in similar contexts belonged to. The Chou-Fasman method (1978) is a combination of such statistics-based methods and rule-based methods. Here are the steps of the Chou-Fasman algorithm:

1. Calculate propensities from a set of solved structures. For all 20 amino acids i, calculate these propensities by:

$$\frac{\Pr[i|\beta\text{-sheet}]}{\Pr[i]} \qquad \frac{\Pr[i|\alpha\text{-helix}]}{\Pr[i]} \qquad \frac{\Pr[i|\text{other}]}{\Pr[i]}$$

That is, we determine the probability that amino acid i is in each structure, normalized by the background probability that i occurs at all.

2. Once the propensities are calculated, each amino acid is categorized using the propensities as one of: helix-former, helix-breaker, or helix-indifferent. (That is, helix-formers have high helical propensities, helix-breakers have low helical propensities, and helix-indifferent have intermediate propensities.) Each amino acid is also categorized as one of: sheet-former, sheet-breaker, or sheet- indifferent. For example, it was found (as expected) that glycine and pralines are helix-breakers.

3. When a sequence is input, find nucleation sites. These are short subsequences with a high-concentration of helix-formers (or sheet-formers). These sites are found with some heuristic rule (e.g. “a sequence of 6 amino acids with at least 4 helix-formers, and no helix-breakers”).

4. Extend the nucleation sites, adding residues at the ends, maintaining an average propensity greater than some threshold.

5. Step 4 may create overlaps; finaly, we deal with these overlaps using some heuristic rules.

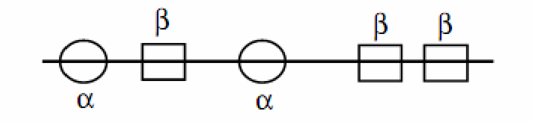

Fig. 2 In the Chou-Fasman method, nucleation sites are found along the protein using a heuristic rule, and then extended

REFERENCES

[1] Cuff, J. A. and Barton, G.J. “Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins, 34, 1999, pp. 508-519.

[2] Cuff, J.A. and Barton G.J. “Application of multiple sequence alignment profiles to improve protein secondary structure prediction” Proteins, 40, 2000, pp. 502-511.

[3] Jones, D.T. “Protein secondary structure prediction based on position-specific scoring matrices” Journal of Molecular Biology, 292, 1999, pp. 195 -202.

[4] Kabsch, W. and Sander, C. “Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features” Biopolymers, 22, 1983, pp. 2577-2637.

[5] Kloczkowski, A., Ting, K.L., Jernigan, R.L. and Garnier, J. “Combining the GORV algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence” Proteins, 49, 2002, pp. 154-166.

[6] Krane, D. and Raymer, M. (2003), “Fundamental Concepts of Bioinformatics”, Pearson Education, New Delhi, pp.1-314.

[7] Moult, J., Hubbard, T., Bryant, S.H., Fidelis, K. and Pedersen, J.T. “Critical assesment of methods of protein structure prediction (CASP): round II”. Proteins, S1, 1992, pp. 2-6.

[8] Rost, B. and Sander, C. (1993) “Prediction of protein secondary structure at better than 70% accuracy” Journal of Molecular Biology, 232, pp. 584-599.

[9] Sander, C. and Schneider, R. “Database of homology-derived structures and the structural meaning of sequence alignment" Proteins, 9, 1991, 56-68.