Soft Computational Framework for Tertiary Protein Structure Prediction

Arundhati Deka
Deptt. of Electronics & Communication Engineering
GIMT, Guwahati, Assam, India

ABSTRACT

Protein structure prediction is turning out to be one of the major challenges in the field of bio-informatics. It is highly important in medicine, especially in drug design and biotechnology. Proteins, being the basic building unit of all organisms, require experimental techniques for prediction of related structures. Among available methods, soft-computational tools provide readily available solutions for making predictions with less complexity, higher reliability and less time. The Artificial Neural Network (ANN) is one such tool which is used for structure prediction of proteins. This method is a machine learning approach in which ANNs are trained to make them capable of recognizing the 8-level subclasses of secondary structure. After the subclasses are recognized in a given sequence, their association with 3-level secondary protein structures is derived. The final structure is obtained from a majority selection from the protein structure. The work is also done in the reverse way, by predicting the 3-level secondary structure from the primary structure .This is done to confirm the accuracy of the prediction. In this work, ANNs are used as classifier to predict the secondary structure.

Keywords

Protein structure prediction, proteinogenic amino acids, DSSP codes

1. INTRODUCTION

Protein structure prediction is the estimation of the 3-D structure of a protein from its amino acid sequence i.e. prediction of secondary, tertiary and quaternary structure from the primary structure.

The problem of protein structure prediction can be solved experimentally using methods such as NMR Spectroscopy and X-ray Crystallography[1].Experimental methods are the main source of information about protein structure and they can generate more accurate results.

However, they are also time consuming where the determination of the structure of a single protein can take months and they are expensive, laborious and need special instruments. Moreover and due to some limitations in the experimental methods, it is not always convenient to determine the protein structure experimentally.

This creates a big gap between the number of protein sequences and known protein tertiary structures. In order to bridge this gap, other methods are much needed to determine the protein structure. Scientists from many fields have worked to develop theoretical and computational methods which can help provide cost effective solutions for the protein structure prediction problem. Machine learning methods such as ANNs and SVMs are used over the traditional experimental methods.

In this paper, a machine learning approach is proposed in which ANNs are trained to make them capable of recognizing the 8-level subclasses of secondary structure and their association with 3-level secondary protein structures is derived. The final structure is obtained from a majority selection from the protein structure. In the reverse way, the 3-level secondary structure is also predicted from the primary structure.

2. BASIC THEORETICAL CONCEPTS

2.1 Protein structure

Proteins are the basis of all organisms. They take part in all biological processes inside the human body. All proteins are polymers of amino acids i.e. amino acids are the basic building blocks of protein .There are 20 different amino acids. Every protein is a unique chain of these 20 amino acids. They differ from one another in the number and sequence of amino acids. Depending on the number and sequence of amino acids, every protein has different shape and chemical properties [2]. The sequence of amino acids of a protein represents a biological information unit. This information creates the unique structure of proteins which determines its activity (chemical properties) in the organism. Basically proteins have four different structures [2]-

Primary: The primary structure refers to the unique amino acid sequence of the polypeptide chain. The primary structure is held together by covalent or peptide bonds.

Secondary: Secondary structure refers to highly regular local sub-structures. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. The three secondary structures are:-alpha helix, beta sheets and coil which are influenced by the properties of each amino acid.

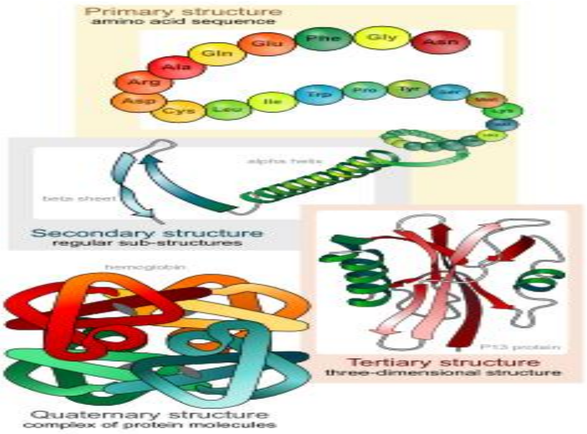

Figure 1: Structures of protein

Tertiary: Tertiary structure refers to three-dimensional structure of a single protein molecule. The alpha-helices and beta-sheets are folded into a compact globule.

The three dimensional structures are responsible for the functional characteristics of proteins.

Quaternary: It is composed of two or more subunit of tertiary structures.

The different structures of protein are shown in Fig. 1.[3]

2.2 Proteinogenic amino acids

The word proteinogenic means "protein building". Proteinogenic amino acids can be assembled into a polypeptide through a process called translation [4]. Proteinogenic amino acids are those amino acids that can be found in proteins.

There are 22 standard amino acids, but only 21 are found in eukaryotes. Of the 22, 20 are directly encoded by the universal genetic code. Humans can synthesize 11 of these

20 from each other or from other molecules of intermediary metabolism. The other 9 must be consumed in the diet, and so are called essential amino acids. Amino acids are molecules containing an amine group, a carboxylic acid group and a side-chain that varies between different amino acids. The key elements of an amino acid are carbon, hydrogen, oxygen, and nitrogen.[5]

2.3 Folding of proteins

Proteins are an important class of biological macromolecules present in all organisms. Each protein has a unique linear sequence of amino acids, also called a polypeptide. This amino acid sequence contains information that guides the protein to fold up into a unique shape. To be able to perform their biological function, proteins fold into one or more specific spatial conformations. To understand the functions of proteins at a molecular level, it is often necessary to determine their 3-D structure. The folding of protein is shown in Fig. 2.[6]

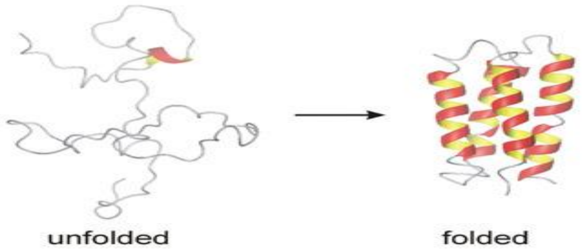

Figure 2: Folding of protein

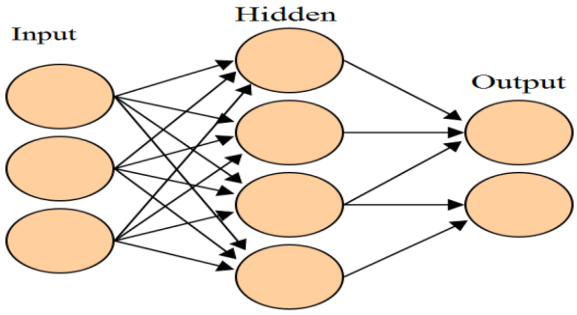

Figure 3: Multi-layer Perceptron

The unique structure of each protein is required for the specific task that it must perform in a living organism [7]. In genomic bioinformatics, the function of a particular protein can be predicted by homology technique. It follows the rule that if the structure of protein x, whose function is known, is homologous to the structure of gene y, whose function is unknown, one could infer that y may share x's function. the parts of a protein responsible for structure formation and interaction with others can be determined by homology modeling technique. In the prediction of protein structure this information is used once the structure of a homologous protein is known. For reliable prediction of protein structure this is the only way till date.

2.4 DSSP codes

The Dictionary of Protein Secondary Structure (DSSP) is commonly used to describe the protein secondary structure with single letter codes.

There are eight types of secondary structure that DSSP defines [8]: G = 3-turn helix , H = 4-turn helix ,I = 5-turn helix, T = hydrogen bonded turn , E = extended strand in parallel and/or anti-parallel beta-sheet conformation, B = residue in isolated beta-bridge ,S = bend . Amino acid residues which are not in any of the above conformations are assigned as the eighth type, ‘Coil'.

2.5 ANN as a soft computational tool

An ANN is a massively parallel distributed processor that has a natural propensity for storing experimental knowledge and making it available for use. It means that:

- *Knowledge is acquired by the network through a learning (training) process.*
- *The strength of the interconnections between neurons is implemented by means of the synaptic weights used to store the knowledge.*

The learning process is a procedure of adapting the weights with a learning algorithm in order to capture the knowledge. The aim of the learning process is to map a given relation between inputs and outputs of the network [9]. The figure of feed-forward ANN set-up called multi layer Perceptron is shown in Fig. 3.[10]

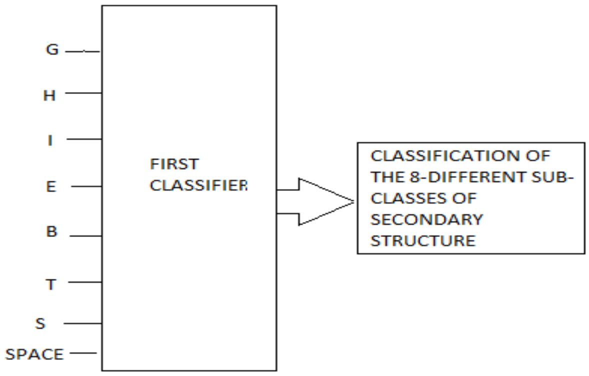
G
H
I
E
B
T
S
SPACE
FIRST CLASSIFIER
CLASSIFICATION OF THE 8-DIFFERENT SUB-CLASSES OF SECONDARY STRUCTURE

Figure 4: System Model for classifier 1 (1st approach)

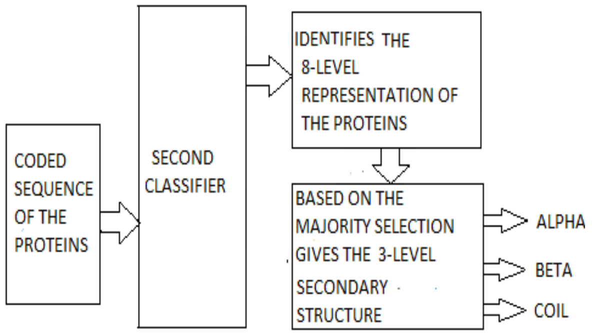
CODED SEQUENCE OF THE PROTEINS
SECOND CLASSIFIER
IDENTIFIES THE 8-LEVEL REPRESENTATION OF THE PROTEINS
BASED ON THE MAJORITY SELECTION GIVES THE 3-LEVEL SECONDARY STRUCTURE
ALPHA
BETA
COIL

Figure 5: System Model for classifier 2 (1st approach)

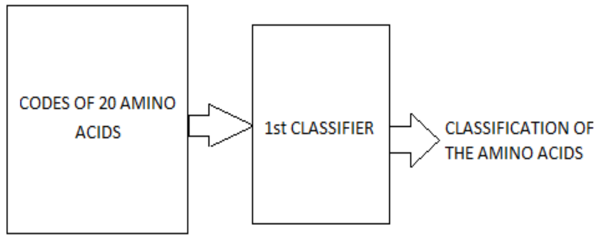

Figure 6: System Model for classifier 1 (2nd approach)

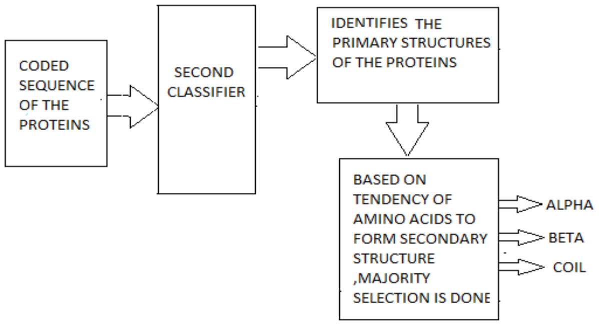
CODED SEQUENCE OF THE PROTEINS
SECOND CLASSIFIER
IDENTIFIES THE PRIMARY STRUCTURES OF THE PROTEINS
BASED ON TENDENCY OF AMINO ACIDS TO FORM SECONDARY STRUCTURE ,MAJORITY SELECTION IS DONE
ALPHA
BETA
COIL

Figure 7: System Model for classifier 2 (2nd approach)

3. PROPOSED MODEL FOR PSP

The work done is summarized by the system model shown in Fig. 4, Fig. 5, Fig. 6 and Fig. 7.

It consists of several steps as described below:

The work done is a two way approach to confirm the secondary structure. In the first approach, secondary structure is predicted from the 8-level subclass representation of secondary structure .In the second approach, a reverse way is followed, i.e secondary structureis predicted from the primary structure.

The work is done in several steps:-

3.1 First approach

- Collecting data set: In our work we have considered six proteins that are Myoglobin, Insulin, Hemoglobin, Porcine pepsin, E.coli and Glyoxylase resistance Protein. The DSSP code of these proteins are collected from the protein data bank(PDB).
- Coding of Proteins: To code the proteins an alphanumeric coding scheme is used. Each subclass of secondary structure is coded with a unique alphanumeric code. Then the considered proteins are coded with these coded subclasses.
- Training and testing of the protein structure: The network is trained with the six coded proteins and then tested to obtain the results .Two classifiers have been trained for the purpose. The first classifier is trained with the eight different subclasses G, H, I, E, B, S, T, SPACE. The classifier identifies these eight subclasses. The second classifier is trained with the coded sequence of the six proteins. This classifier identifies the six proteins. Based on the majority selection, the secondary structure of the proteins is also extracted.

3.2 Second approach/Reverse approach

- Collecting amino acid sequences of the proteins from the PDB: The amino acid sequences of the six proteins are collected from the protein data bank.
- Coding the proteins: To code the proteins an alphanumeric coding scheme is used. Each amino acid is coded with a unique alphanumeric code. Then the considered proteins are coded with these coded amino acids.
- Training and testing of the protein structure: The first classifier is trained with the 20 coded amino acids. This classifier identifies the amino acids. The second classifier is trained with the amino acid sequences of the six proteins .It identifies the primary structures of the proteins.
- Extracting the secondary structure: The secondary structure of these proteins is evaluated based on the tendencies of the amino acids to form different secondary structures.
- Derivation of final structure: Depending on majority selection from the secondary structure , the final structure is obtained.

4. RESULTS

The ANN is trained with six protein structures. With these proteins during training the ANN shows 100 percent accuracy while the learning phase with the same set. Using Gradient descent with Adaptive learning rate Algorithm, training is carried out. The ANN is given a performance goal of around 10^{-9}which is attained after certain number of sessions. The configuration of ANN and performance of ANN are shown in table I and tableII.

The performance training graphs for the 1st approach are shown in Fig.8, Fig. 9 and Fig. 10.

The performance training graphs for the 2nd approach are shown in Fig.11, Fig. 12 and Fig. 13.

Fig.14 and Fig.15 shows the comparative bar graph for the percentage of helical content and percentage of beta content present in different proteins obtained from the two way approach. The grey bar represents the theoretical value, the black bar represents the percentage value obtained by forward approach and the cyan blue bar gives the value from reverse approach.

From the bar charts it can be estimated that the proteins Myoglobin, Insulin, Hemoglobin and E. Coli have higher helical content than the proteins Porcine Pepsin and Glyoxylase. On the other hand proteins Porcine Pepsin and Glyoxylase have higher beta content compared to the other four proteins. Hence, based on majority content, the secondary structure of the proteins can be confirmed. Proteins Myoglobin, Insulin, Hemoglobin and E. Coli are alpha proteins while pepsin porcine and Glyoxylase fall under the category.

Table I: Configuration of ANN

ANN	MLP
No. of Training samples	651
Training type	TRAINGDA
Max. no of epochs	2433
No. of hidden layers	4

Table II: Performance of ANN

EPOCH	**TIME**	**MSE**
497	5 sec	10^{-4}
819	8 sec	10^{-5}
1042	10 sec	10^{-7}
2433	23 sec	10^{-9}

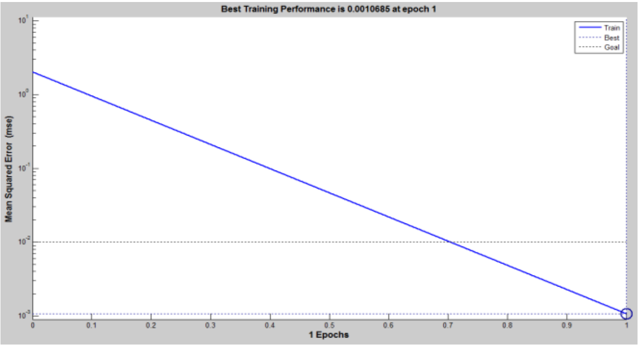

Figure 8: Performance graph for 8-subclasses (1st approach)

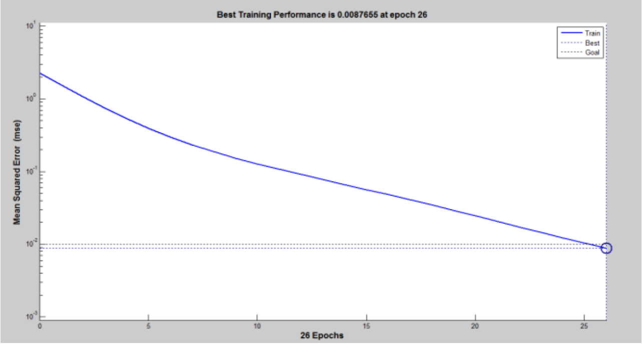

Figure 9: Performance graph for Myoglobin, Insulin, Hemoglobin (1st approach)

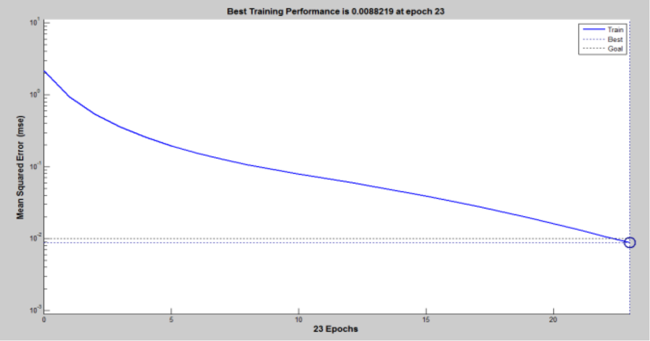

Figure 10 : Performance graph for porcine, glyoxylase, E.coli (1st approach)

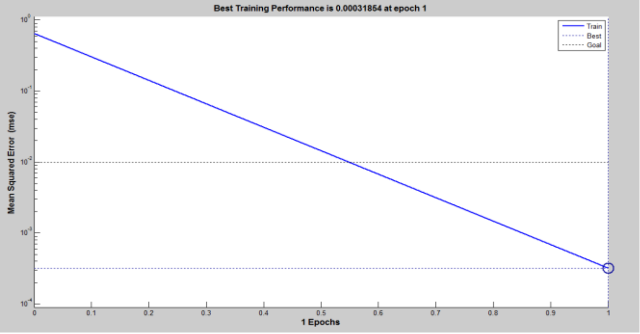

Figure 11: Performance graph for 20 amino acids (2nd approach)

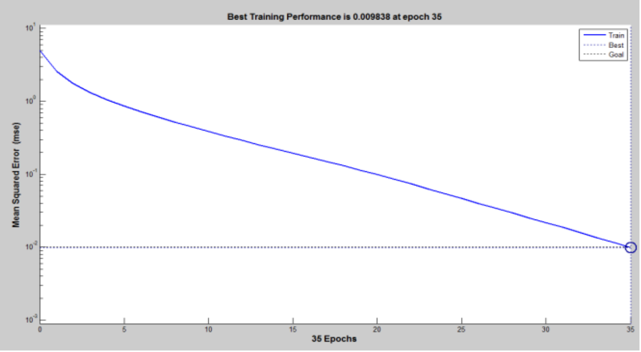

Figure 12: Performance graph for Myoglobin, Insulin, Hemoglobin (2nd approach)

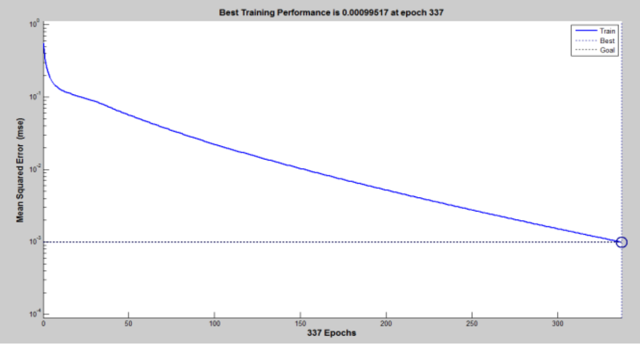

Figure 13: Performance graph for Porcine pepsin, Glyoxalase, E.coli (2nd approach)

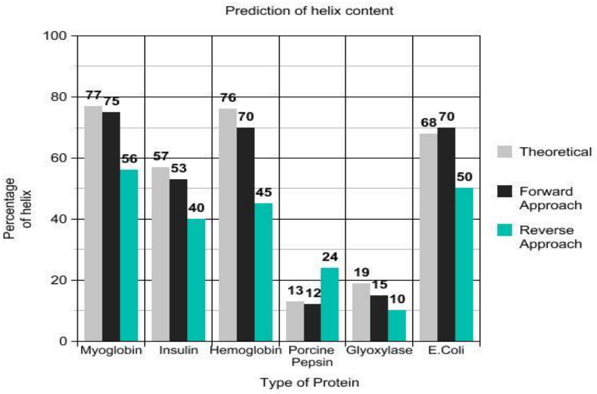

Figure 14: Performance comparison graph for helix prediction

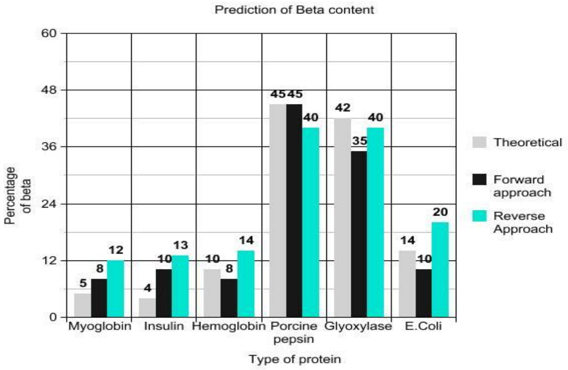

Figure 15: Performance comparison graph for beta prediction

5. CONCLUSION

This work reflects the uniqueness of the proposed model functioning with coded sequence of proteins and applied to ANN for prediction of 3-level secondary structures from the 8-level secondary structures. This work shows how multi level ANN classifier can be configured for protein structure prediction. Moreover, it also predicts the secondary structure from the reverse way, i.e from the primary structure. The work may be extended to go into deep insight into prediction theory and explore the three dimensional tertiary structure predictions from the secondary structures. It may include more known and unknown protein structures which shall make it a reliable set up for research in bioinformatics.

6. REFERENCES

[1] S.Kushwaha and M.Shakya, "A machine learning technique for Tertiary Structure Prediction of proteins from peptide sequences", in *Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing*, 2009.

[2] C.Branden and J.Tooze, "Introduction to protein structure", 2nd Ed.,Garland Pub.,1999

[3] Protein Structure. Wikipedia: http://en.wikipedia.org/wiki/Protein_structure

[4] D. L.Nelson and Michael M.Cox,"Lehninger's principles of Biochemistry",4th Edition, 2009.

[5] G. Pok, C. H. Jin and K. H. Ryu, "Correlation of Amino Acid Physicochemical Properties with Protein Secondary Structure Conformation", in *Proceedings of International Conference on BioMedical Engineering and Informatics*, 2008.

[6] Cornell and Scripps researchers cite evidence supporting theory of how proteins fold into their critical shapes. By Krishna Ramanujan http://www.news.cornell.edu/stories/Aug06/ProteinFoldingScheraga.kr.html

[7] S. Malkov, M. V. Zivkovic, M.s V. Beljanski, S. D. Zaric: ``Correlation of amion acids with secondary structure types,connection with amino acid structure", paper preview, May 24, 2005.

[8] 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", Department of Information and Computer Science Institute for Genomics and Bioinformatics University of California, Irvine

[9] S. Haykins,"Neural Networks, A Comprehensive Foundation", 2nd Ed., Pearson Education, New Delhi,2003.

[10] Image: Artificial Neural Network: http://en.wikipedia.org/wiki/File:Artificial_neural_network.svg