Protein Structure Prediction using Artificial Neural Network

Hemashree Bordoloi
Department of Electronics and Communication Technology
Gauhati University, Guwahati-781014

Kandarpa Kumar Sarma
Department of Electronics and Communication Technology
Gauhati University, Guwahati-781014

ABSTRACT
Protein secondary structure prediction is a problem related to structural bioinformatics which deals with the prediction and analysis of macromolecules i.e. DNA, RNA and protein. It is an important step towards elucidating its three dimensional structure, as well as its function. Secondary structure of a protein can be predicted from its primary structures i.e. from the amino acid sequences or from the residues though challenges exists. For these four methods are used. These are Statistical Approach, Nearest Neighbor method, Neural Network Approach and Hidden Markov Model Approach. The Artificial Neural Network (ANN) approach for prediction of protein secondary structure is the most successful one among all the methods used. In this method, ANNs are trained to make them capable of performing recognition of amino acid patterns in known secondary structure units and these patterns are used to distinguish between the different types of secondary structures. This work is related to the prediction of secondary structure of proteins employing artificial neural network though it is restricted initially to three structures only.

Keywords: Artificial Neural network, Amino acids, Protein structure prediction

1. INTRODUCTION
Protein secondary structure prediction is the most challenging and influencing area of research in the field of bioinformatics which deals with the prediction and analysis of macromolecules i.e. DNA, RNA and protein. Proteins are the fundamental molecules of all organisms. They are unique chains of amino acids that adopt unique three dimensional structures which allow them to carry out intricate biological functions and specifications of each protein is given by the sequence of amino acids [1].A unique one letter code is used to specify the amino acids. Basically proteins have three structures---primary, secondary and tertiary. The sequences of amino acids are called primary structures [2]. Secondary structure is the spatial arrangement and regularities of amino acids with respect to each other [3]. The secondary structure has 3 regular forms: helical (α helices), extended (β sheets) and loops or reverse turns or coils (0). From the secondary structure, three dimensional structures are derived and it is the tertiary structure of the protein. The three dimensional structure is responsible for the functional characteristics of proteins and it is termed as tertiary structure [2]. A typical protein contains about 32% alpha helices, 21% beta sheets and 47% loops or non regular structures [2]. Theoretically, it is not possible to predict 100% accurate protein structure because of the fact that there are 20 different amino acids and thus no. of ways to generate similar structure in proteins by different amino acids is much more[2]. The general approach to predict the secondary structure of a protein is done by comparing the amino acid sequence of a particular protein to sequences of the known databases. In protein secondary structure prediction, amino acid sequences are inputs and the resulting output is conformation or the predicted structure which is the combination of alpha helices, beta sheets and loops.

2. BASIC BIOLOGICAL CONCEPTS
The basic idea behind bioinformatics is the notion of homology. It is used to predict the function of a gene in genomic bioinformatics. The homology technique follows the rule that if the sequence of gene A, whose function is known, is homologous to the sequence of gene B, whose function is unknown, one could infer that B may share A's function. In the structural bioinformatics, homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins. In homology modeling technique, this information is used to predict the structure of a protein once the structure of a homologous protein is known. Presently, it remains the only way to predict protein structures reliably. For example, by determining the structures of viral proteins it would enable researchers design drugs for specific viruses [4].
The secondary structure has 3 regular forms: helical (α helices), extended (β sheets) and loops or reverse turns or coils (0). In the protein secondary structure prediction, the inputs are the amino acid sequences while the output is the predicted structure also called conformation, which is the combination of alpha helices, beta sheets and loops [4]. A typical protein sequence and its conformation class are shown below

Protein Sequence:
ABABABABCCQQFFFAAAQQAQQA
Conformation Class:
HHHH EEEE HHHHHHHH

where H means Helical, E means Extended, and blanks are the remaining coiled conformations. A typical protein contains about 32 % α helices, 21% β sheets and 47% loops or non-regular structure [4]. With a given protein sequence of amino acids a1,a2,....an, the problem of secondary structure prediction is to predict whether each amino acid ai is in a α-helix, a β-sheet or neither.

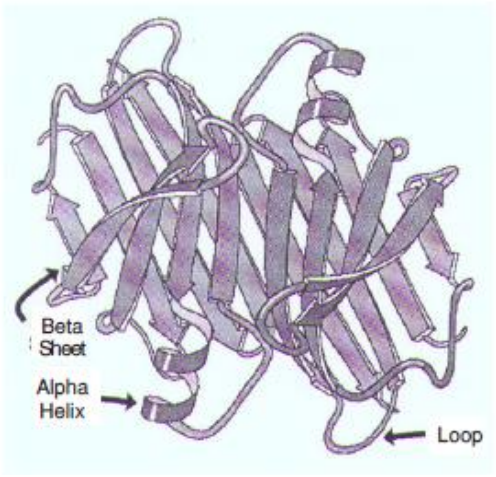

Figure 1: The protein secondary structure, which consists of alpha helices, beta sheets and loops [5]

If through the studies of structural biology one knows the actual secondary structure for each amino acid, then the 3-state accuracy is the percent of residues for which the prediction matches reality. It is called 3-state because each residue can be in one of 3 states :α, β or other (0)[6]

3. BASICS OF ARTIFICIAL NEURAL NETWORK

The human brain is basically an information processing system. That means it is a highly complex, nonlinear, parallel computer [8]. Based on the characteristics of human brain a massively parallel distributed processor called Artificial Neural Network (ANN) is proposed. It is made up of simple processing units called artificial neurons. Artificial neural network has a natural propensity to store experimental knowledge and making it available for use. It resembles the brain in two respects--

- Knowledge is acquired by the network from its environment through a learning process.
- Interneuron connection strengths known as synaptic weights, are used to store the acquired knowledge.[7]

4. NECESSITIES OF PSSP

PSSP is receiving importance in the recent area of research due to the following:

- Being the problem of structural bioinformatics, protein secondary structure prediction can provide prediction and analysis of macromolecules which are the basis of an organism.
- Protein secondary structure prediction (PSSP) provides structure function relationship. That is which particular protein structure is responsible for which particular function would be known by PSSP. So by changing the structure of the proteins or by synthesizing new proteins, functions could be added or removed or desired functions could be obtained [9].
- Structure of the viral proteins can be determined by PSSP and this determination of the structures of the viral proteins provides the way to design drugs for specific viruses.
- PSSP reduces the sequence structure gap [4].The sequence structure gap can be best described by giving the example of large scale sequencing projects such as Human Genome Project. In such type of projects, protein sequences are produced at a very fast speed which results in a large gap between the number of known protein sequences (>150,000) and the no. of known protein structures (>4,000).This gap is called sequence structure gap and PSSP can successfully reduce this gap.
- Experimental techniques are not capable of structure determination of some proteins such as membrane proteins. So the prediction of protein structure using computational tool is of great interest [10].

5. METHODOLOGY

Data Set: In our work we have considered three proteins that are hemoglobin, sickle cell anemia and Myoglobin. We have considered these three proteins because though their function is different, these three proteins are closely related to each other.

Coding of Proteins: Coding scheme is generated based on the chemical structure of the proteins. BCD codes are used for coding. There is a unique BCD code for each component or symbol in the chemical structure. Considering 20 amino acids, each amino acid is coded using the generated coding scheme. Then the three proteins i.e. hemoglobin, sickle cell anemia and Myoglobin are coded with the help of coded amino acids.

Architecture of Neural Network: A fully connected Multi Layer Perceptron (MLP) feedforward neural network is used for our proposed work. Backpropagation algorithm is used to update the weight of the network. Network comprises of only one hidden layer.

Training and Testing of ANN: The network is trained with three coded protein structures. Then testing is done with the coded data to obtain the results.

6. SYSTEM MODEL

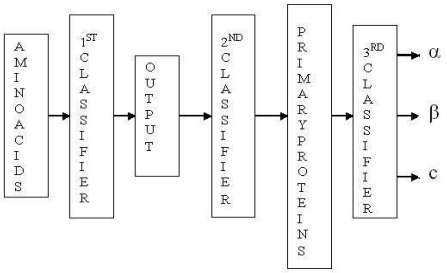

Figure2: System Model for proposed work

The system model as shown in figure 2 comprises of three classifiers. Here Artificial Neural Networks (ANN) is used as classifiers. In the 1st classifier amino acids are given as inputs and the 2nd classifier is same as 1st classifier but here amino acids sequences are given as input. The 3rd classifier is used to obtain secondary structure of proteins.

7. RESULTS

With the three protein structures the ANN during training shows 100% accuracy while validating the learning phase with the same set. The ANN is configured to handle an input of length 985 holding coded values of the protein structure. Three such samples

representing the classes where initially taken during training. The ANN successfully recognizes the required parameters. The training is carried out with (error) back propagation algorithm with Levenberg-Marquardt optimization. The ANN is given a performance goal of around 10^{-6} which is attained after certain number of session though the time taken is around 80 seconds. With variations of upto ±20% in the training sequence the ANN handles the recognition without any variation in performance. It shows its robustness to variations after it is trained properly. The results also validate the coding scheme used for the work.

Some problems, however, were observed due to the larger data sets which the ANN classifier was given to handle. It generates certain computational constrains. Moreover, the ANN classifier in its present form finds it hard to make differentiation between the amino acids in the sequences. Also, the classifier faced some difficulties while detecting the starting and ending point between two amino acids. These shall be removed in subsequent stages of the work and extend it for the prediction of some unknown protein structure with subclasses.

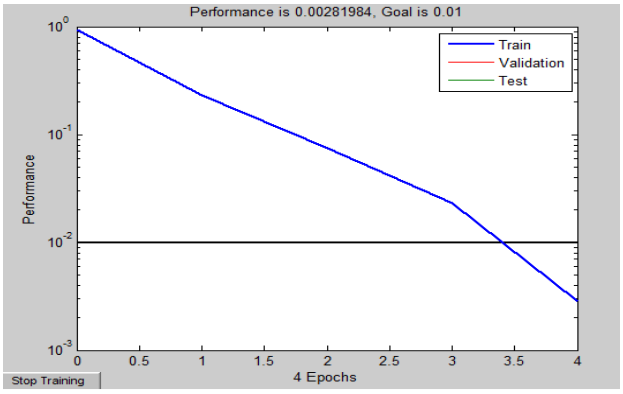

Figure 3: Performance graph for goal 0.01

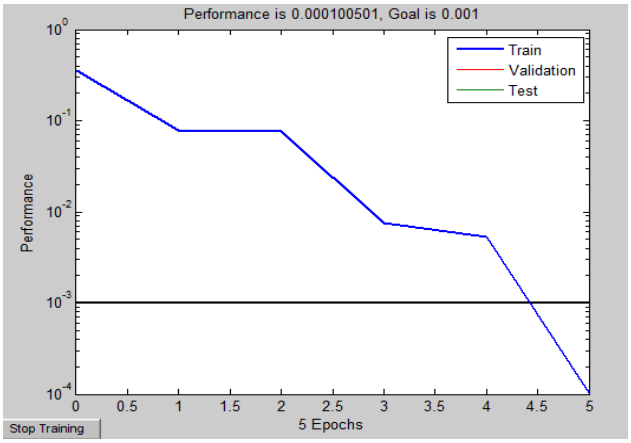

Figure 4: Performance graph for goal 0.001

8. CONCLUSIONS

This work shows the uniqueness of the proposed model functioning with coded amino acids and applied to ANN for prediction of protein secondary structures from three important primary forms

9. REFERENCES

[1] Z. Xiu-fen, P. Zi-shu, K. Lishan, Z. Chu-yu; The Evolutionary computation Techniques for Protein Structure Prediction: A Survey; "WU411 Wuhan University Journal of Natural Sciences", Article ID: 1007-1202(2003)01Pr0297-06; Vol. 8 No. 1B 2003.

[2] F. C. Bernstein, T. F. Koetzle, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T and Tasumi M; "The Protein Data Bank: a computer based archival file for macromolecular structures''; J Mol Biol. Vol. 112, pp. 535-542, 1977.

[3] S. A. Malekpour, S. Naghizadeh, H. Pezeshk, M. Sadeghi, C. Eslahchi; Protein secondary structure prediction using three neural networks and a segmental semi Markov model; "Elsevier Inc. All rights reserved.2008".

[4] S. Akkaladevi, A. K. Katangur, S. Belkasim and Y. Pan; Protein Secondary Structure Prediction using Neural Network and Simulated Annealing Algorithm; "Proceedings of the 26th Annual International Conference of the IEEE EMBS San Francisco, CA, USA • September 1-5, 2004".

[5] Protein Structure Image Gallery at "Biochemistry Department Website of Duke University. USA". (http://kinemage. biochem.duke.edu/).

[6] Mona Singh; Introduction to computational biology;11/16/2000

[7] Simon Haykin; Neural Networks: A Comprehensive Foundation, 2nd ed., Pearson Education, New Delhi, 2003

[8] Kandarpa Kumar Sarma; Speech Corpus of Assamese Numerals for Recognition using Artificial Neural Network; "September 5,Springer"

[9] S. N. V. Arjunan, S. Deris and R. MD Illias; Literature Survey of Protein Secondary Structure Prediction;" Jurnal Teknologi 63–72", Universiti Teknologi Malaysia,34(C) 2001

[10] Aramak Afzal; Applications of neural networks in protein structure prediction.