Artificial Neural Network aided Protein Structure Prediction

Arundhati Deka , Kandarpa Kr. Sarma
Dept. of Electronics and Communication Engineering
Gauhati University, Guwahati-781014
Assam, India

ABSTRACT
Protein structure prediction plays a vital role in drug design and biotechnology. Understanding protein structures is necessary to determine the function of a protein and its interaction with DNA, RNA and Enzymes. Experimental techniques such as NMR Spectroscopy and X-ray Crystallography have been the main source of information about protein structures. But these conventional methods are now replaced by Machine learning methods such as Artificial Neural Network (ANN) and Support Vector Machine (SVM)s. In this paper, ANNs are used as a two level classifier to estimate the tertiary structure of proteins. ANNs are trained to make them capable of recognizing the primary sequences and DSSP codes of protein structures and their association with the secondary structure is derived. Based on majority selection, the final secondary structure is evaluated. These secondary structures can be further used as inputs to classify between the basic tertiary folds and subclasses of tertiary folds.

General Terms
Artificial Intelligence

Keywords
DSSP codes,

1. INTRODUCTION
Protein 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 estimating its 3D structure, as well as its function. Tertiary structure of a protein can be predicted from its primary structures i.e. from the amino acid sequences or from the residues. Basically proteins have three structures---primary, secondary and tertiary. The sequences of amino acids are called primary structures [1]. Secondary structure is the spatial arrangement and regularities of amino acids with respect to each other. The secondary structure has 3 regular forms: helical , extended β sheets and loops or reverse turns or coils. 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 . A typical protein contains about 32% alpha helices, 21% beta sheets and 47% loops or non regular structures [1]. 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.

In this paper, a machine learning approach has been proposed in which ANNs are trained to make them capable of recognizing the primary sequences and the DSSP codes of the protein structures and their association with the secondary structure is derived. Based on majority selection, the final secondary structure is evaluated. The secondary structures are further used as inputs to classify between the basic tertiary folds and subclasses of tertiary folds. ANNs function as a two level classifier for the proposed work. Some of the previous works done are[1]-[5].

2 . THEORETICAL CONCEPTS
Every 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 tertiary structure is the 3D fold of the protein molecule comprising of secondary structure elements: alpha (α) helices, beta (β) sheets and loops .Based on the maximum element composition, the tertiary structure assumes three different topologies. The three basic topologies are alpha topology, beta topology and mixed topology[6]. These three basic topologies are further classified into the sub-topologies. In the protein tertiary structure prediction, the inputs are the DSSP codes while the output is the predicted topology. The Dictionary of Protein Secondary Structure (DSSP) is commonly used to describe the protein secondary structure with single letter codes.

3. NECESSITIES OF TSP
The function of a specific bio-molecule (protein) is mostly known from its molecular structure .Tertiary structure prediction (TSP) can determine the structure of the viral proteins which leads to the design of drugs for specific viruses.TSP provides Structure function relationship. It means that a particular protein structure is responsible for a particular function. 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 [7].

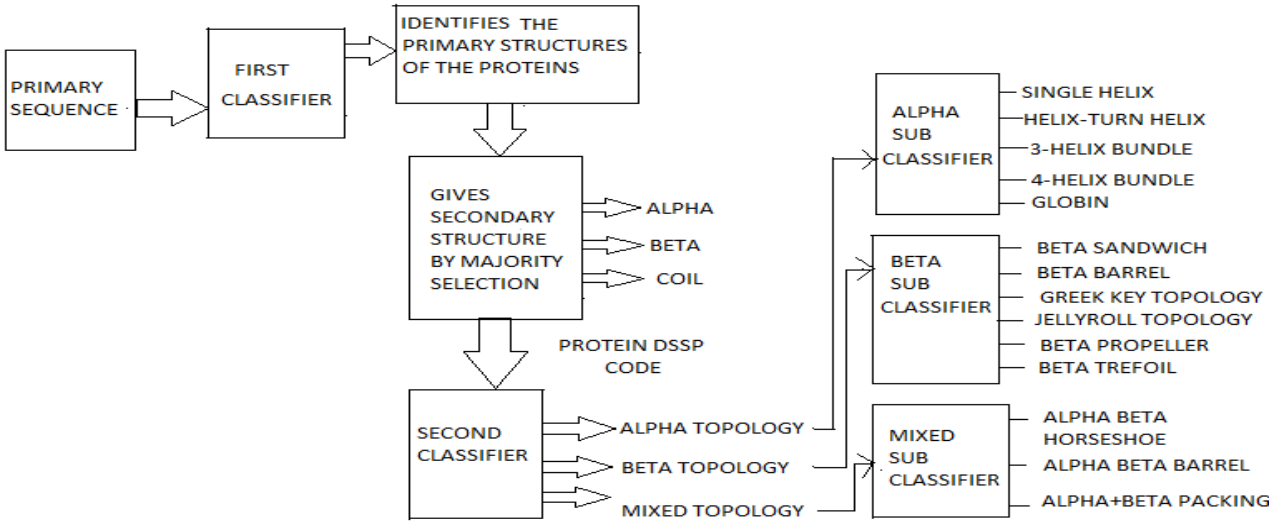

Fig 1: System model

4. BASICS OF ANN

Artificial neural network (ANN) is made up of interconnecting artificial neurons. The general function of a neural network is to produce an output pattern when given a particular input pattern, and is loosely related to the way the brain operates. Learning these mappings is done in conceptually the same way as the brain. Several types of neural networks exist but the most common one used has been the Multi-Layer Perceptron. Another ANN used in the work is Radial Basis Function (RBF) which is faster compared to MLP. The RBF uses a Bayesian decision making to process applied patterns. It has two hidden layers of which the first one provides a class distribution probability while the second one provides a decision depending upon the closeness the applied patterns shall have using a Gaussian spread function [8].

5. SYSTEM MODEL

The work done is summarized by the system model shown in Fig 1. The work is done in a two way approach to first confirm the secondary structure of proteins.

- **Proteins considered for the work:** In our work we have considered six proteins that are Myoglobin, Insulin, Hemoglobin, Porcine Pepsin, E.coli and Glyoxylase resistance Protein. These proteins are considered as they belong to different secondary categories.

- **Collection of dataset from database:** The primary amino acid codes of these proteins are collected from the Protein Data Bank(PDB). Moreover, the DSSP codes of the respective proteins collected.

- **Encoding of Proteins**: An Alphanumeric coding scheme is used for encoding each protein primary sequence. Each amino acid present in a particular

  protein is encoded by a unique alpha-numeric code. Then the considered proteins are coded with these coded amino acids. Each DSSP code present in a particular protein is also encoded by a unique alpha-numeric code.

- **Training and testing of the ANN**: The network is trained with the coded protein structures.

- **Extraction of 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 the final tertiary structure**: The DSSP codes of the proteins are fed as inputs for the second classifier. The second classifier classifies the proteins into the basic tertiary topologies.

Table I: ANN configuration set up

Parameters	**Specification**
Training function	'traingda'
No. of hidden layers	4
Learning function	'learngdm'
Maximum no. of epochs	1651
Performance goal	10^{-4}
Error function	'mse'
No. of training samples	1760

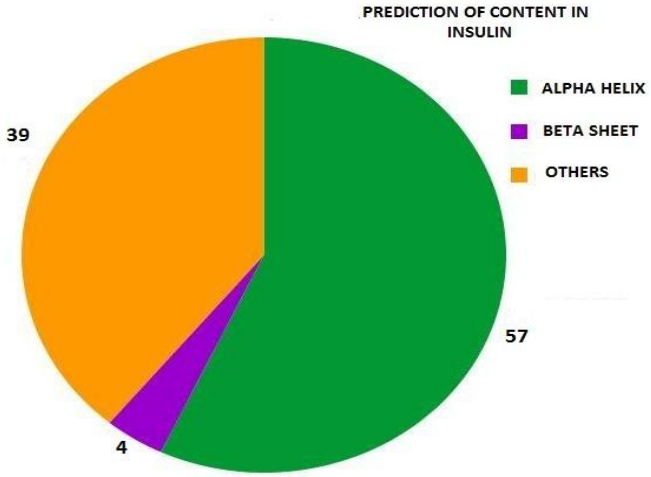

Fig 2: Percentage content in Insulin

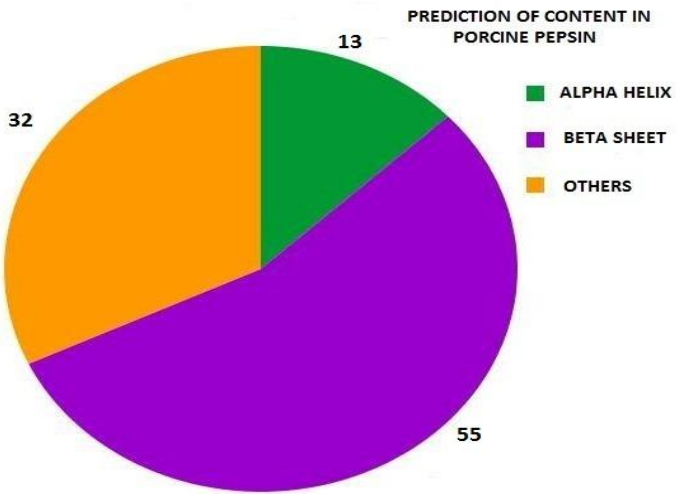

Fig 5: Percentage content in Porcine Pepsin

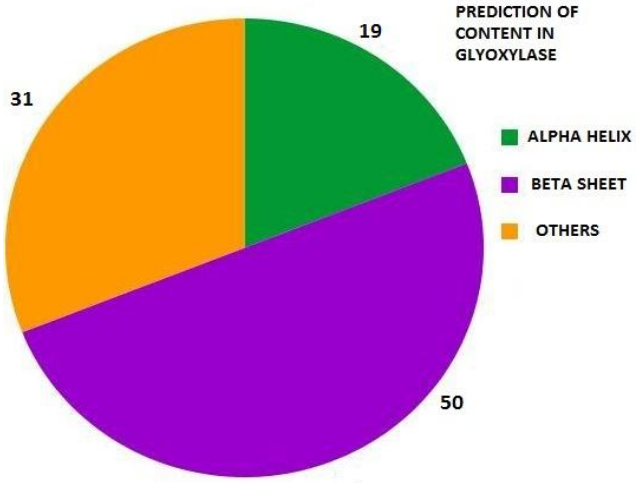

Fig 3: Percentage content in Glyoxylase

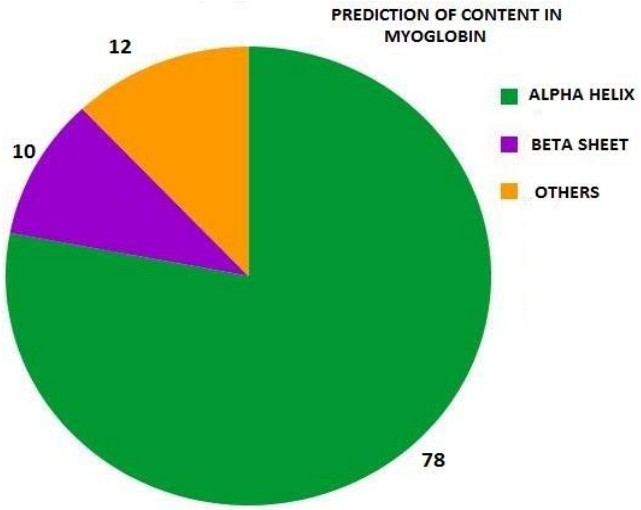

Fig 6: Percentage content in Myoglobin

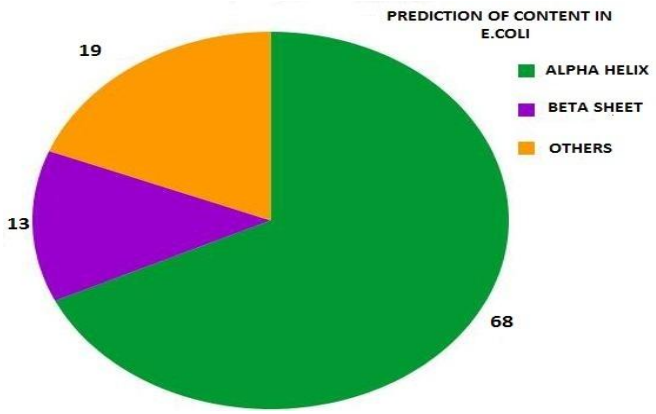

Fig 4: Percentage content in E. Coli

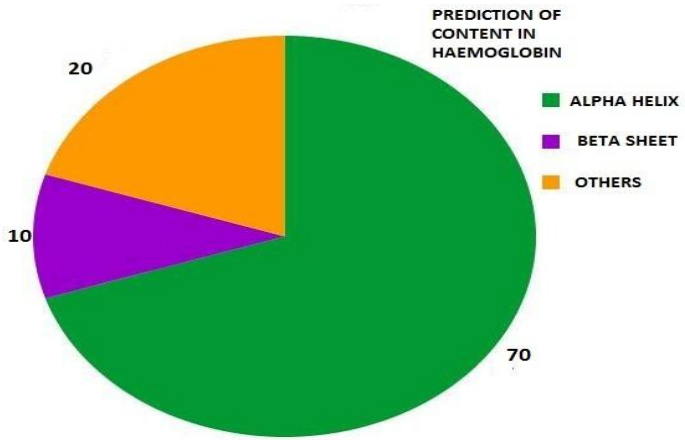

Fig 7: Percentage content in Hemoglobin

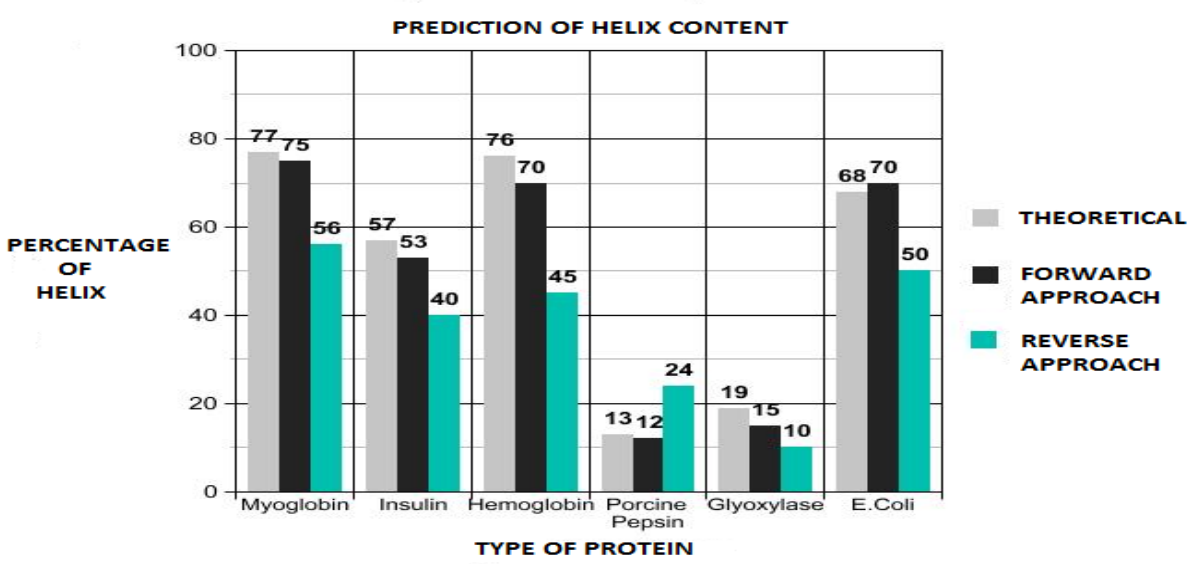

Figure 8: Performance comparison graph for helix prediction

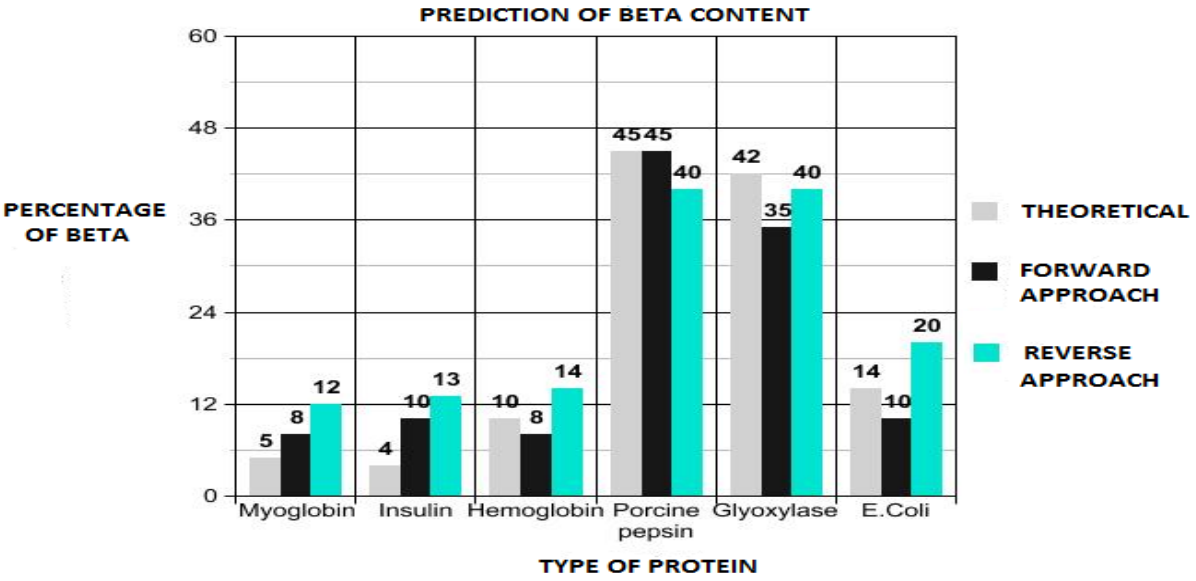

Figure 9: Performance comparison graph for beta sheet prediction

6. RESULTS

The ANN is trained with the protein structures. With these proteins during training the ANN shows (90% +) of accuracy . Using Gradient descent with Adaptive learning rate Algorithm, training is carried out. The ANN is given a performance goal of around 10^{-4} which is attained after certain number of sessions. The configuration of ANN is shown in table I.

Figures. 2-7 represent the pie diagrams showing the percentage of secondary structure content in the different proteins. The green color represents the percentage of helical content, the violet color represents the beta sheet and the orange color gives the percentage of the coil structures present in different proteins. The secondary structures have been confirmed by a two way approach. The first approach is to determine the secondary structure from the known primary amino acid sequences and the second approach is to determine the secondary structure from the 8 subtypes of DSSP codes. Figure.8 and Figure.9 shows the comparative bar graph for the prediction of helical content and prediction 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 beta category.

7. CONCLUSION

This work reflects the uniqueness of the proposed model functioning with coded sequence of proteins. Especially, we formulated a framework which is unique in the sense that it uses coded sequence of proteins and applied to ANN for prediction of 3-level secondary structures from the 8-level secondary structures. Further, the work is extended to include the two level ANN predictor. Multi level ANN classifier is configured for tertiary protein structure prediction. The work highlights the advantage of RBF ANN over the MLP in terms of faster learning, speed of training and better accuracy of classification. It shows how multi level ANN classifier can be configured for protein structure prediction.. The work can be extended for prediction of tertiary structure using Recurrent Neural Networks (RNN), Linear Vector Quantization (LVQ) as classifiers for faster training and better accuracy of classification. More unknown protein structures can be included to make the proposed system robust and reliable for research in bioinformatics.

7. REFERENCES

[1] H. Bordoloi and K. K. Sarma, ``Protein Structure Prediction Using Multiple Artificial Neural Network Classifier'', as a Chapter of a volume titled *Soft Computing Techniques in Vision Science*, Studies in Computational Intelligence, 2012, Volume 395/2012, pp. 137-146, DOI: 10.1007/978-3-642-25507-6_12 , 2012.

[2] H. Bordoloi and K. K. Sarma, ``Protein Structure Prediction using Artificial Neural Network''**,** *IJCA Special Issue on Electronics, Information and Communication Engineering* ICEICE (3), pp. 24-26, December 2011. Published by Foundation of Computer Science, New York, USA.

[3] A.Deka, H.Bordoloi and K. K. Sarma, "ANN-aided Tertiary Protein Structure Prediction using Certain Coding Techniques and Known Secondary Structures", in *Proceedings of International Conference on Electronics and Communication Engineering(ECE)*, 2012.

[4] A. Deka and K. K. Sarma,"Soft Computational Framework for Tertiary Protein Structure Prediction", International Journal of Electronics Signals and Systems (IJESS), ISSN:2231-5969, Vol.1, Issue 3

[5] A. Deka and K. K. Sarma," Tertiary Protein Structure Prediction using Artificial Neural Network as a Two-level Classifier" , to appear in the proceedings of 3rd International Conference on Computer and Communication Technology,ICCCT-2012.

[6] H. Mathkour and M. Ahmad, "An integrated approach for protein structure prediction using artificial neural network", in *Proceedings of Second International Conference on Computer Engineering and Applications*, 2010

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

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

[9] C.Kehyayan, N.Mansour, H.Khachfe, "Evolutionary Algorithm for Protein Structure Prediction",in *Proceedings of International Conference on Advanced Computer Theory and Engineering*, 2008

[10] 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.

[11] S.Kushwaha and M.Shakya, "Multi-Layer Perceptron Architecture for Tertiary Structure Prediction of helical content of proteins from peptide sequences",*in Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing*, 2009