Auto-Coding of Amino Acid Sequences for Prediction of Protein Tertiary Structure

Arundhati Deka
Department of ECE, Gimt-Guwahati
E-mail: arundhati.deka@gmail.com

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*-RBF, DSSP codes*

INTRODUCTION

Proteins are the basic building blocks of every organism. Proteins are mainly composed of amino acids. In nature, there are 20 different types of amino acids. Amino acids share a common structure except for one chemical group (R, side chain) attached to the central carbon atom. Several amino acids combine together to form a protein molecule. Proteins with different functions have different amino acid sequences.

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.

In this paper, a soft computational framework has been proposed in which ANNs are trained to make them capable of recognizing the basic tertiary topologies. Furthermore, the subclasses of tertiary structure have also been classified. ANNs function as a two level classifier for the proposed work.

BIOLOGICAL CONCEPTS

Biological structure of protein

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. Basically proteins have four different structures:-

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.

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.

Significance of Tertiary Protein Structure Prediction

Knowledge of molecular (protein) structure helps us better understand the function of the molecule (protein).Identifying common structural elements helps us better understand and organize unrelated structures. Tertiary structure prediction 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 particular function. So by changing the structure of the proteins or by synthesizing new proteins,

Basic concept of tertiary 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. 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.

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: 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'.

Artificial Neural Network 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. Knowledge is acquired by the network through a learning (training) process. 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. A feed-forward ANN set-up called Multi Layer Perceptron (MLP). 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.

CODING OF PROTEINS

The work done is summarized as follows. It consists of several steps as described below:

Collection of data set:

In our work, we have considered 15 proteins, 5 proteins from each of the three topologies.

From the alpha topology proteins taken: Alamethicin RNA binding protein Rop, Ferritin, Cytochrome c and Annexin. Proteins Immunoglobulin, Beta2Microglobulin, Streptavadin, Plastocyanin and Satellite Tobacco Necrosis Virus are taken from the beta toplogy. From the mixed topology proteins Ribonuclease, Triose Phosphate Isomerase, Lysozome, Actidin and Histidine are taken.

Coding of amino acids

We have done this using bioinformatics toolbox in Matlab. For every type of amino acid, there is an in-built code/integer. Using these 20 specific integers for the 20 different amino acids we have replaced the amino acid sequences with the specifically defined alpha-numeric codes. The amino acids along with the in-built integers and alpha-numeric codes are shown in Table I.

Table 1: *Coding Of Amino Acids*

AMINO ACID	1-LETTER CODE	IN-BUILT INTEGER	ALPHA-NUMERIC CODE USED
Alanine	A	1	61
Arginine	R	2	72
Asparagine	N	3	75
Aspartic acid	D	4	64
Cysteine	C	5	63
Glutamine	Q	6	71
Glutamic acid	E	7	65
Glycine	G	8	67
Histidine	H	9	68
Isoleucine	I	10	69
Leucine	L	11	60
Lysine	K	12	78
Methionine	M	13	62
Phenylalanine	F	14	66
Proline	P	15	70
Serine	S	16	73
Threonine	T	17	74
Tryptophan	W	18	77
Tyrosine	Y	19	79
Valine	V	20	76
Unknown Character	?	0	12

Data Encoding:

The DSSP code of these proteins is collected from the Protein Data Bank 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.

RESULTS

Training and testing of the ANN:

The ANN block is trained with the coded proteins and then tested to obtain the results. Two classifiers Primary and

Secondary have been trained for the purpose. The primary classifier is trained with the DSSP codes of the 15 randomly selected proteins.

Determination of tertiary topology: The primary classifier classifies the proteins into three possible basic tertiary topologies. Three secondary classifiers are trained for the three classified topologies.

The proteins classified under alpha category are fed to the first secondary classifier. Similarly, for the other two categories the second and third secondary classifiers are used. The secondary classifiers classify the proteins into subclasses of tertiary folds.

In our work, two types of ANNs are used. Initially, the MLP is trained with the protein structures. The performance derived is found to be satisfactory using MLP. Still, we explored what additional advantages can be derived using the RBF which is faster and learns better than the MLP if properly configured.

The comparative performance of MLP and RBF configuration of ANN are shown in Table

Table 2: *Comparative Performance*

PARAMETERS	MLP	RBF
TRAINING SAMPLES	1920	1920
EPOCHS	1651	15
TIME	20 sec	3 sec
MSE	0.0001	6.2×10^{-32}

CONCLUSION

The proposed model functions with coded sequence of proteins. ANN is used as a two level predictor of tertiary structure. This work shows how multi level ANN classifier can be configured for protein structure prediction. The work further highlights the advantage of the RBF ANN over the MLP in terms of faster learning, speed of training and better accuracy of classification. The work may include more known and unknown protein structures which shall make it a reliable set up for research in bioinformatics.

REFERENCES

1. C. Branden and J. Tooze, "Introduction to protein structure", 2nd Ed., Garland Pub.,1999
2. 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.
3. 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.
4. 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.
5. Antony Joseph," Protein secondary structure prediction using Artificial Neural Networks", Department of Computer Science and Engineering, IIT Madras,2002

6. D. L. Nelson and Michael M.Cox,"Lehninger's principles of Biochemistry",4th Edition, 2009.
7. 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.