Biosc.Biotech.Res.Comm. Special Issue Vol 14 No 06 (2021)

BBRC
Bioscience Biotechnology Research Communications

Analytical Model to Predict Protein Structure using Soft-Computing Approach

Hemashree Bordoloi[1,2], S.R. Nirmala[2] and Kandarpa Kumar Sarma[1]
[1]*Dept. of ECE, Gauhati University, Guhati India*
[2]*Dept. of ECE, Assam Don Bosco University, India*
[3]*School of ECE, KLE Tech. University, India*
[4]*Dept. of ECE, Gauhati University, India*
Corresponding author email: hemashree.bordoloi@dbuniversity.ac.in

ABSTRACT
Protein structure prediction is an important step towards elucidating its three-dimensional (3D) structures, as well as its function. There are four different protein structures: primary, secondary, tertiary and quaternary. This work describes an ANN model for secondary structure prediction in esophagus cancer linked proteins to gain insights on the proteins towards the understanding of the disease.

KEY WORDS: PROTEINS, AMINO ACIDS, SECONDARY STRUCTURE PREDICTION, ARTIFICIAL NEURAL NETWORK.

INTRODUCTION

Proteins play a crucial role in virtually almost all biological processes (Jiang Q et al. 2017). A protein is a natural polymer, made up of some monomers called amino acids joined together by peptide bonds. Basically, proteins have four different structures. These are primary, secondary, tertiary and quaternary. The primary structure has unique sequence of amino acids and biological information that determines the structure and function of a protein. In secondary structure, there are locally defined spatial arrangement and regularities of amino acids with respect to each other. The three secondary structures are– alpha helix, beta sheets and coil or loop which are influenced by the properties of each amino acid (Ma Y et al. 2018). Most protein structures are built up from combinations of secondary structure elements, alpha helices and beta strands which are connected by loop regions of various lengths and irregular shape (Wardah W et al. 2019).

A combination of the secondary structure elements forms the stable hydrophobic core of the molecule and the loop regions are at the surface of the molecule.Amino acids are the fundamental building units of proteins. Amino acids are the essential medium through which the human gene translates into proteins. Amino acids can be detected from their hydrophilicity and hydrophobicity value because for each amino acid these two values are unique (Bordoloi H. & Sarma K.K, 2012 and Yookesh, T.L 2020). Therefore, it is of interest to document the predicted secondary structure data in esophagus cancer linked proteins to gain insights on the proteins towards the understanding of the disease (S.Ranjeeth et al 2020).

METHODOLOGY

Proteins are made up of unique sequence of 20 different amino acids. Amino acids are the basis of proteins which in turn is the basis of all organisms (Haykin, S, 2003). In our proposedwork three ANN classifiers are used to detect the secondary structure of proteins. Unique BCD codes are used for coding each component or symbol in the chemical structure of these aminoacids. The 20 amino acids are coded with the help of the coded chemical structure based on their hydrophobic and hydrophilic

Biosc Biotech Res Comm P-ISSN: 0974-6455 E-ISSN: 2321-4007

Identifiers and Pagination
Year: 2021 Vol: 14 No (6) Special Issue
Pages: 324-327

Article Information
Received: 29th March 2021
Accepted after revision: 28th June 2021

DOI: *http://dx.doi.org/10.21786/bbrc/14.6.67*

indexes using the first classifier. Then the considered proteins are coded with the help of coded amino acids using classifier II. The system model shown in Fig 1comprises of 3 systems for detection. The first system (System I) provides the identification of the aminoacids. The second system (System II) uses the coded aminoacids as inputs. The primary structures of the proteins from the 2nd classifier is given to the third system (System III) as input which classifies the 3 secondary structures:alpha helix, beta sheets and coil or loop. The final structure is derived from the majority selection of the 3 secondary structures.

The methodology of our work is divided into three steps. These are–

- Detection of amino acids based on hydrophobicity and hydrophilicity value
- Detection of primary protein structure from the unique chain of amino acids
- Detection of secondary structure from the primary structure

In our work, we have considered a few proteins which are the biomarker of esophagus cancer. All the proteins are collected from NCBI database. Details of the proteins with their isomers are given in the Table 1 as shown below:

The system model of our proposedwork, uses three ANNs. The first ANN isconfigured to accept the chemical structures of 20 amino acids as the input parameter andprovide amino acids as the output. The second ANN accepts the amino acid sequences as theinput and provides the proteins as output. The third ANN is responsible for selecting the secondary protein structure. Several ANN configurations are used to ascertainthe best set up for the testing. ANNs are configured with one input layer, two hidden layer and one output layer. Coded amino acids are the result of the 1stclassifier which is coded on the basis of the average of the hydrophilic and hydrophobic indices of the amino acids. The second classifier will take amino acid sequences as its input. The amino acid sequence is unique for each and every protein and hence the primary structure of protein can be detected by the total sum of the sequence.500 samples are used for the training and 100 samples are used for the testing. Different training functions are used to measure the performance of the proposed model.

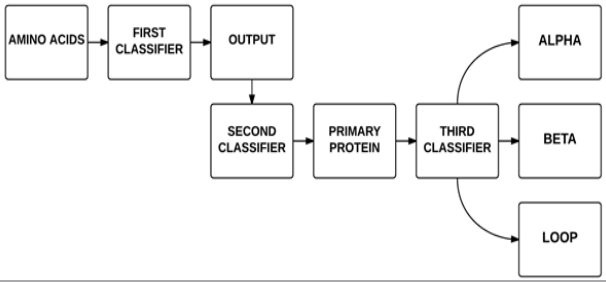
Third classifier is responsible for the detection of secondary structure. Secondary structure is detected on the basis of majority selection. The primary structures derived from the 2nd classifier are the input to the 3rd classifier. Each protein has the affinity to a particular secondary structure i.e. alpha, beta and coil. The final structure is derived from the majority selection of the three secondary structures as shown in Fig 2. In Fig 2, a is alpha secondary structure. B is beta secondary structure and c is coil secondary structure.

DISCUSSION

In our work we have used Artificial Neural Network(ANN) as the classifiers.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.For our work a fully connected MLP feedforward ANN is used. The weights of the network are updated by using Back Propagation algorithm. The ANN comprises of two hidden layer. Three ANN classifiers are designed for our work. The amino acids are classified by the first network, the second network classifies the protein primary structure and the secondary structure of proteins is classified by the third network.

The third ANN predicts the alpha, beta and coil segments. The training set for the 1stclassifier is amino acids, that for 2ndclassifier is the unique sequence of amino acids and the 3rdclassifier accepts primary protein structures as input. The three classifiers are connected to each other. Training is done to find a set of weights such that the squared error is minimized. Testing is the final step in prediction of secondary structure of proteins. Seven-fold cross validation is used to test prediction accuracy. In the seven-fold cross validation test, the data set is divided into seven segments. Six segments are used for training and one segment is used for testing. This process is repeated for seven times until all seven segments have been chosen as the testing set. In our work the trained network is tested with the coded data to obtain the results.

Figure 1: Proposed System Model



RESULTS

The secondary structure for DACT, E2F1, EGFR, HMGA1 and HMGA2 proteins are given in Fig 3 below:

The secondary structure detected for DACT, E2F1, EGFR, HMGA2 and HMGA1 protein is alpha secondary structure respectively. All these proteins are related to esophagus cancer. Although the primary structure of each protein is different but the secondary structure possess by all the five proteins with their isomers are same. Result of this mathematical model is compared with Chou&t Fasman Secondary Structure Prediction (CFSSP)Server which show the similarities with our system.

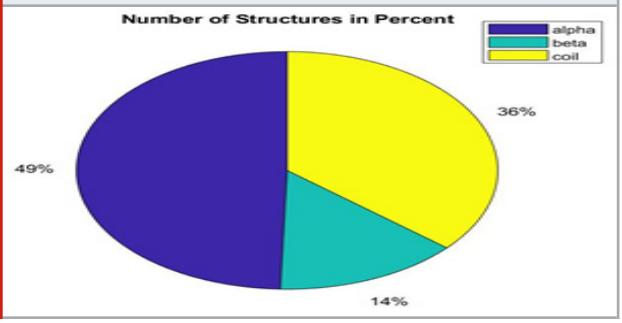
In case of primary structure, in DACT protein, Asparagine is same in all the isomers. The number of Methionine

is same in all the isomers of HMGA2 protein. In all the isomers of HMGA1 protein, the number of Arginine, Asparagine, Aspartate, Cysteine, Glutamate, Histidine, Isoleucine, Methionine, Phenylalanine, Tryptophan and Tyrosine is same.

Table 1. Proteins with Accession Number

1	NP_999627.2	DACT
2	NP_001273279.1	DACT
3	XP_011533809.1	DACT
4	NP_005216.1	E2F1
5	NP_958439.1	EGFR
6	NP_958440.1	EGFR
7	NP_958441.1	EGFR
8	NP_001333870.1	EGFR
9	NP_003474.1	HMGA2
10	NP_003475.1	HMGA2
11	NP_001287847.1	HMGA2
12	NP_001287848.1	HMGA2
13	NP_001317119.1	HMGA2
14	NP_002122.1	HMGA1
15	NP_665906.1	HMGA1
16	NP_665908.1	HMGA1
17	NP_665909.1	HMGA1
18	NP_665910.1	HMGA1
19	NP_665912.1	HMGA1
20	NP_001306006.1	HMGA1
21	NP_001306007.1	HMGA1
22	NP_001306008.1	HMGA1
23	NP_001306009.1	HMGA1
24	NP_001306010.1	HMGA1
25	NP_001306011.1	HMGA1

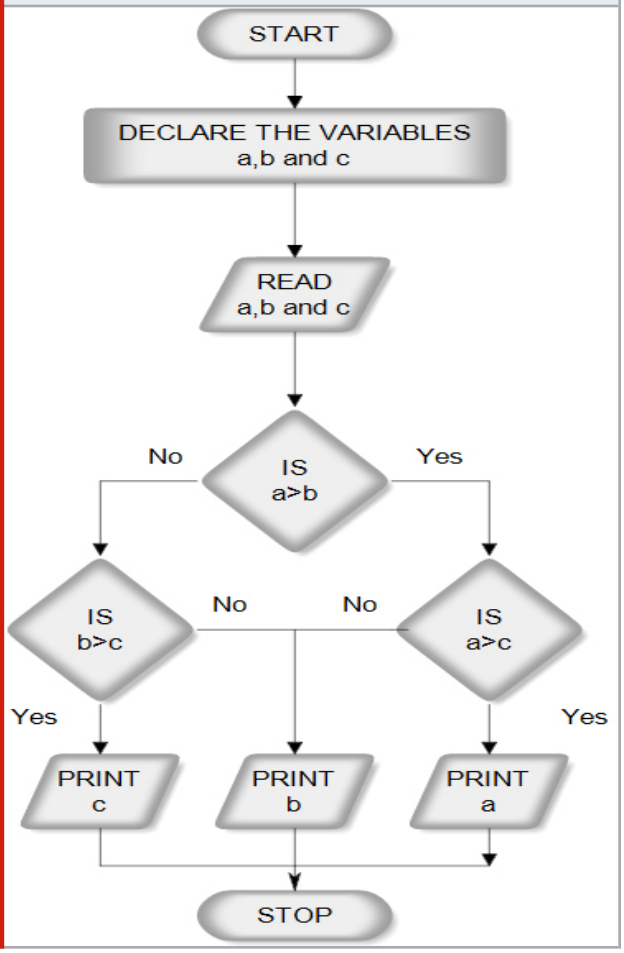
Figure 3: Percentage of secondary structures in DACT, E2F1, EGFR, HMGA1 and HMGA2 protein showing alpha as majority one



CONCLUSION

Proteins play an important role in almost all biological processes. Protein structure prediction has been a focus subject in bioinformatics just due to the importance of protein structure in full understanding the biological and chemical activities of organisms. In our work we have predicted the secondary structure of various proteins which are the biomarker of esophagus cancer. We describe an ANN model for secondary structure prediction in esophagus cancer linked proteins to gain insights on the proteins towards the understanding of the disease.

Figure 2: Flowchart for detection of Secondary Structure



REFERENCES

Bordoloi, H. and Sarma, K.K., 2012. Protein structure prediction using multiple artificial neural network classifier. In Soft Computing Techniques in Vision Science (pp. 137-146). Springer, Berlin, Heidelberg.

Jiang, Q., Jin, X., Lee, S.J. and Yao, S., 2017. Protein secondary structure prediction: A survey of the state of the art. Journal of Molecular Graphics and Modelling, 76, pp.379-402.

http://www.ncbi.nlm.nih.gov

Kubat, M., 1999. Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. The Knowledge Engineering Review, 13(4), pp.409-412.

Ma, Y., Liu, Y. and Cheng, J., 2018. Protein secondary structure prediction based on data partition and semi-random subspace method. Scientific reports, 8(1), pp.1-10.

Ranjeeth, S. and Latchoumi, T.P., 2020. Predicting Kids Malnutrition Using Multilayer Perceptron with

Stochastic Gradient Descent. Rev. d'Intelligence Artif., 34(5), pp.631-636.
Wardah, W., Khan, M.G., Sharma, A. and Rashid, M.A., 2019. Protein secondary structure prediction using neural networks and deep learning: A review. Computational biology and chemistry, 81, pp.1-8.
Yookesh, T.L., Boobalan, E.D. and Latchoumi, T.P., 2020, March. Variational Iteration Method to Deal with Time Delay Differential Equations under Uncertainty Conditions. In 2020 International Conference on Emerging Smart Computing and Informatics (ESCI) (pp. 252-256). IEEE.