International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 3 Issue 4, April-2014, pp: (18-23), Impact Factor: 1.147, Available online at: www.erpublications.com

Protein Secondary Structure Prediction Using Neural Network: A Comparative Study

Patel Mayuri Dinubhai[1], Dr. Hitesh B Shah[2]
[1]Information Technology Department, [2]Electronics & Communication Department
[12]G H Patel College of Engineering &Technology, Gujarat Technological University, Vallabh, Vidyanagar

Abstract: Bioinformatics or computational biology is field of science in which biology, computer science and information technology merges into a single discipline. In modern computation biology, research of protein secondary structure plays a major role in protein tertiary structure prediction. Protein structure prediction is depends on its amino acid sequence. Current studies prefer soft computing techniques for classification and regression task. Recently many researchers used various data mining and soft computing tool for protein structure prediction. Our objective is to enhance the prediction of 1D, 2D and 3D protein structure problem using Neural Network for solved linear and non-linear problems. The data base used for this problem is Protein data bank (PDB) select sets, RS126 and CB513.PDB is based on structural classification of protein (SCOP). All proteins in the PDB-40D that had more than 35% identity with proteins of the training set were excluded from the testing set.

Keywords: Bioinformatics, feature selection (FS), Scoop (Structural classification of protein), protein data bank (PDB), RS126 and CB513Neural Networks (NNs).

I. INTRODUCTION

Bioinformatics is an emerging and rapidly growing field of science. As a consequence, a large number of biological data are being collected due to genome-sequencing projects over the world. Therefore, computational tools are needed to analyze the collected data in the most efficient manner. For example, working on the Prediction of the biological functions of genes and proteins (or parts of them) based on structural data. Recently Neural Network have been a new and promising technique for machine learning. On some applications it has obtained higher accuracy than other existing method like chou-fasman and GOR method. In this paper [1],they have applied a new method for discovering regular patterns in data that is based on neural network models. The brain has highly developed pattern matching abilities and neural network models are designed to mimic them. This study was inspired by a previous application of network .learning to the problem of text-to-speech. In this paper, using Neural Network we exploit important issues of bioinformatics like: the prediction of 1-D,2-D and 3-D structure from amino acid sequence. The prediction of protein secondary structure and 3-D fold recognition is a challenging field strongly related with function determination which is of high interest for the biologists and the pharmaceutical industry.

Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry; it is highly important in medicine (for example, in drug design) and biotechnology(for example, in the design of novel enzymes).The primary structure refers to amino acid sequence is called primary structure. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The secondary structure consists of local folding regularities maintained by hydrogen bonds and is traditionally subdivided into three classes: alpha-helices (H), beta-sheets (E), and coil(C). Tertiary structure refers to three-dimensional structure of a single protein molecule. The alpha-helices and beta-sheets are folded into a compact globule. Quaternary structure is the arrangement of multiple folded protein or coiling protein molecules in a multi-subunit complex. Many pattern recognition and machine learning methods have been proposed to solve this issue. Surveys are, for example, some typical approaches are as follows: (i) statistical information (ii) physico-chemical properties [11] ;(iii) sequence patterns (iv) multi-layered neural networks [3,12]; (v) graph-theory (vi) multivariate statistics (vii) expert rules (viii) nearest-neighbour algorithms and (iv)support vector machine[1,2,10,13].

Among these machine learning methods, neural networks and Support Vector Machine may be the most popular and effective one for the secondary structure prediction. Up to now the highest accuracy is achieved by approaches using it. In this survey paper, we apply SVM for protein secondary structure prediction. We worked on similar data and encoding schemes as those in Protein Data Bank[1] and Rost &Sander [12](referred here as RS126) which has sharing less than 25% identity. The performance accuracy is verified by a ten-fold cross-validation. Ding and Dubchak [2] indicate that SVM easily returns comparable results as neural networks. Therefore, SVM and its various kernel function is a promising direction for classification and protein structure prediction.

Primary structure

The primary structure refers to amino acid sequence is called primary structure. Each α-amino acid consists of a backbone part that is present in all the amino acid types, and a side chain that is unique to each type of residue. An exception from this rule is proline. The primary structure is held together by covalent or peptide bonds, which are made during the process of protein biosynthesis or translation. The primary structure of a protein is determined by the gene corresponding to the protein. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a process called translation. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. The sequence of a protein can be determined by methods such as Edman degradation or tandem mass spectrometry.

Protein secondary structure

The secondary structure consists of local folding regularities maintained by hydrogen bonds and is traditionally subdivided into three classes: alpha-helices(H), beta-sheets(E), and coil(C).Secondary structure contained localized and recurring fold of a polypeptide chain, where two main regular structures are the α-helix and β-sheet. Hydrogen bond is responsible for secondary structure-helix may be considered the default state for secondary structure. These secondary structures are defined by patterns of hydrogen bonds between the main-chain peptide groups. They have a regular geometry, being constrained to specific values of the dihedral angles ψ and φ on the Ramachandran plot.

Tertiary structure

The multi-class protein fold recognition or tertiary structure problem is central in molecular biology and it can be formulated as follows: given the primary structure of a protein, how the 3-D fold can be deduced from it. Tertiary structure is an important approach where same structure without relying on sequence similarity. Different types of methods have been developed for fold recognition [3]. These methods [8] are divided into two methodological approaches:

(a) The informatics based methods that involve the sequence based methods and the structure based methods, and
(b) The biophysics based methods.

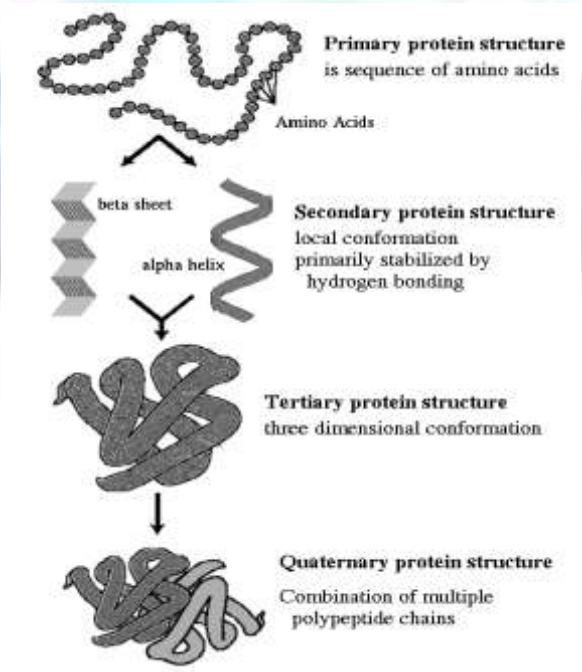

Fig 1.1 Four Levels of Protein Structure [13]

In fold recognition or tertiary sequence based methods is very common. Machine learning techniques, such as genetic algorithms, support vector machines [1, 2, 13], Using Fuzzy Rule-Based Classifier [9] and Multi Layer Perceptron12] Ensemble of Probabilistic Neural Networks [7], have been adopted to exploit protein sequence or secondary structure information. The amino acid composition (protein sequence), in specific, has been employed in many areas of bioinformatics, like protein structural class prediction [3], discrimination of DNA binding proteins and discrimination of outer membrane proteins. However, although significant improvement has been made in the field of fold recognition, the accuracy of the existing methods remains limited and there is a need to develop new methods.

Quaternary structure

Quaternary structure is the arrangement of multiple folded protein or coiling protein molecules in a multi-subunit complex. Many proteins are actually assemblies of more than one polypeptide chain, which in the context of the larger assemblage are known as protein subunits. In addition to the tertiary structure of the subunits, multiple-subunit proteins possess a quaternary structure, which is the arrangement into which the subunits assemble Enzymes composed of subunits with diverse functions are sometimes called holoenzymes, in which some parts may be known as regulatory subunits and the functional core is known as the catalytic subunit. Examples of proteins with quaternary structure include hemoglobin, DNA polymerase, and ion channels. Other assemblies referred to instead as multi-protein complexes also possess quaternary structure.

II. CLASSIFICATION METHOD

Neural Network (NNs)

In this Research, we use multi-layer perceptron in this research as a three-layer feed forward network with weight adjusted by conjugate gradient minimization factor. In NNs training there is always problem of generalization; the number of NNs parameters was adaptively adjusted to variable training set sizes by changing the number of hidden units. The perceptron classifies the input vector X into two categories. If the weights and threshold T are not known in advance, the perceptron must be trained. Ideally, the perceptron must be trained to return the correct answer on all training examples, and perform well on examples it has never seen. The training set must contain both type of data (i.e. with "1" and "0" output) which is shown in fig 2.1.

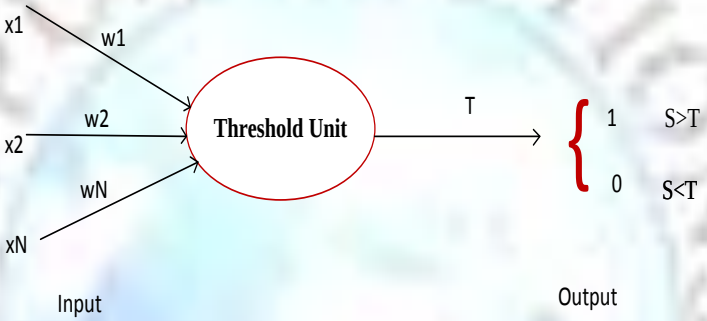

Fig 2.1: The perceptron

In this work, we use multi-layer perceptron as a three-layer feed forward network in figure 5 with weight adjusted by conjugate gradient minimization factor. Various NNs architecture were tested; but using three layer (1-hidden and 2-output layer) architecture achieves a good performance while having a minimum number of nodes.

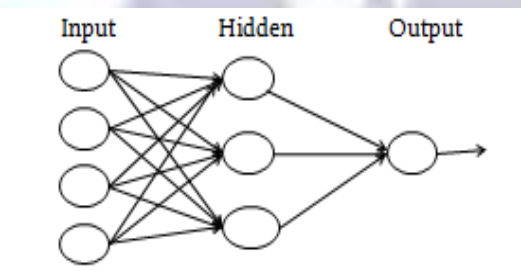

Fig 2.2: Multi-layer Perceptron

The perceptron computes the dot product S = xw. The output F is a function of S: it is often set discrete (i.e. 1 or 0), in which case the function is the step function.

For continuous output, often use a sigmoid:

$$F(x) = \frac{1}{1+e^{-x}} \qquad (1)$$

Neural networks are trained just like perceptron, by minimizing an error function:

$$E = \sum_{i=1}^{Ndata} (NN(x^i) - t(x^i))^2 \qquad (2)$$

III. RELATED MATERIALS

Dataset

We can use three dataset like PDB(protein data bank),RS126 and CB513 for comparisons in secondary structure prediction.

PDB: Protein data bank (PDB) select sets which is based on structural classification of protein (SCOP). All proteins in the PDB-40D that had more than 35% identity with proteins of the training set were excluded from the testing set. This dataset which we used was selected from the (Ding and Dubchak ,2001).In the database 128 folds, which have seven or more proteins and represent all major structural classes: α ,β, α +β and α/β. since the accuracy of any machine learning tool depends on the number of representative for training, we used 27 most populated fold in this research. This dataset is available on (http://ranger.uta.edu/~chqding/protein/).

RS126: The original set of 126 protein sequences proposed by Rost and Sander [12], currently containing a total of 26,846 amino acids (this number has varied slightly over the years due to changes and corrections in the PDB, 24,395 [1], with which to train and test secondary structure prediction algorithms. They defined non-redundancy to mean that no two proteins in the set share more than 25% sequence identity over a length of more than 80 residues.

CB513: A dataset of 513 sequences developed by Cuff and Barton [12] with the aim of evaluating and improving protein secondary structure prediction methods. It is, perhaps, one of the most used independent datasets in this field.

Feature vector extraction

Feature vector extraction method performed using machine learning tool and its variation. It is a pre-Processing step for protein secondary prediction and protein fold(tertiary structure) and approach focuses on modifying data set to improve the accuracy of the classification..It is extracted from original (primary)sequence based on three descriptors: 'Composition', composition of three constituents(e.g. polar, neutral and hydrophobic residues in Hydrophobicity);'Transition', the transition of frequencies(polar to neutral and neutral to hydrophobic,etc.); and 'Distribution', the distribution pattern of constituents. We are extracting three classes α, β and coil using Soft Computing technique from original amino acid sequence. This is given in fig 3.1.

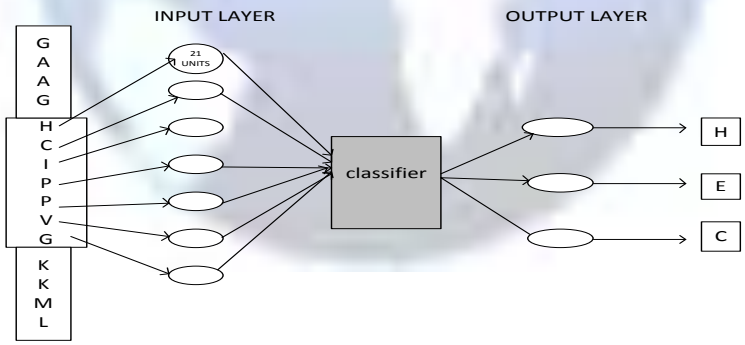

Fig 3.1: Input and Output Coding for Protein fold problem [13]

Evaluation of Prediction Accuracy

The most common measure for the secondary structure prediction is the overall three-state accuracy

(Q3). It is defined as the ratio of correctly predicted residues to the total number of residues in the database under consideration [2].Q3 is calculated by:

$$Q_3 = \frac{\sum_{i\in(H,E,C)} \#of\ \ residues\ correctly\ predicted_i}{\sum_{i\in(H,E,C)} \#of\ \ residues\ in\ class\ i} \times 100 \qquad (3)$$

Where, Q_3 =Total accuracy

International Journal of Enhanced Research in Management & Computer Applications, ISSN: 2319-7471
Vol. 3 Issue 4, April-2014, pp: (18-23), Impact Factor: 1.147, Available online at: www.erpublications.com

IV. PROTEIN STRUCTURE PREDICTION

Primary structure prediction

There are four main methods to perform this reduction process:

(1) DSSP: H to H; E to E; all other states to C;
(2) DSSP: H, G to H; E, B to E; all other states to C;
(3) DSSP: H, G, I to H; E to E; all other states to C;
(4) DSSP: H, G to H; E to E; all other states to C;

In this article, we adopt the strictest method (2).Using this method we are getting the primary structure.

Original Sequence of protein

KSEEELANAFRIFDKNADGYIDIEELGEILRATG

CCEEECCCCCCCCCCCCCHCCCCEECHECCCCCH

Primary structure

Secondary structure prediction

In secondary structure our motivation is translate
primary structure into three main classes: helix (H), strand (E) and coil(C).using Feed forward Neural Network with various database we are classifies it which is display in table 4.1

Table 4.1 Testing set for Secondary structure prediction using Neural Network

Dataset		**RS126**	**PDB**	**CB513**
No.of proteins		21147	48415	75707
Testing set	Helix(H) %	11.40	11.40	9.96
	Strand(E) %	5.46	0.0546	6.57
	Coil(C) %	83.14	83.14	83.48
Simulation time(Sec)		51.436	310.76	231.27

Table 4.2 Training set for Secondary structure prediction using Neural Network

Dataset		**RS126**	**PDB**	**CB513**
No.of proteins		21147	48415	75707
Training set	Helix(H) %	10.31	10.31	10.41
	Strand(E) %	5.82	5.82	6.06
	Coil(C) %	83.87	83.87	83.53
Simulation Time (sec)		51.436	310.76	231.27

Table 4.3 Validation set for Secondary structure prediction using Neural Network

Dataset		**RS126**	**PDB**	**CB513**
No.of proteins		21147	48415	75707
validation set	Helix(H) %	10.10	10.10	10.26
	Strand(E) %	5.87	5.87	6.19
	Coil(C) %	84.04	84.04	83.55
Simulation time(sec)		51.436	310.76	231.27

CONCLUSION

Different algorithms are available for secondary structure prediction such as GOR (GarnierOsguthorpe-Robson) ,chou-Fashman and Neural Network. Here we can use three various dataset with Feed Forward Neural Network for secondary structure prediction. For implementation we can implement Neural Network in MATLAB-2010.Our future work is implement Support Vector machine (SVM) for same dataset.

REFERENCES

[1]. Chries H.Q.Ding and Inna Dubchak, "Multi-class protein fold recognition using Support Vector Machines and Neural Network,"Bioinformatics, vol. 17, no.4, pp. 349-358, 2001.
[2]. Jung-Yang Wang, "Application of Support Vector Machine in Bioinformatics," 2002.
[3]. Alessio Ceronia, Paolo Frasconia and Gianluca Pollastrib, "Learning protein secondary structure from sequential and relational data," Neural Networks 18, pp. 1029–1039 2005.
[4]. Jieyue He, Hae-Jin Hu, Robert Harrison , Phang C. Tai and Yi Pan b, "Transmembrane segments prediction and understanding using support vector machine and decision tree," Expert Systems with Applications 30, pp. 64-72, 2006.
[5]. Hany Alashwal, Safaai Deris and Razib M. Othman, "Comparison of Domain and Hydrophobicity Features for the Prediction of Protein-Protein Interactions using Support Vector Machines, "World Academy of Science, Engineering and Technology 7, pp.431-437, 2007.
[6]. Themis P. Exarchos, Costas Papaloukas Christos Lampros and Dimitrios I. Fotiadis, "Mining sequential patterns for protein fold recognition, "Journal of Biomedical Informatics 41 , pp. 165-179, 2008 .
[7]. Yuehui Chen, Xueqin Zhang, Mary Qu Yang and Jack Y. Yang," Ensemble of Probabilistic Neural Networks for Protein Fold Recognition,"IEEE Tran., pp.66-70, 2007 .
[8]. Robertas Damasevicius, "Analysis of Binary Feature Mapping Rules for Promoter Recognition in Imbalanced DNA Sequence Datasets using Support Vector Machine,"4th International IEEE Conference intelligent Systems , pp. 2008.
[9]. Eghbal G. Mansoori, Mansoor J. Zolghadri and Seraj D. Katebi, "Protein Superfamily Classification Using Fuzzy Rule-Based Classifier,"IEEE Trans. Nanobioscience, vol. 8,no. 1,pp- 92-99, MARCH 2009.
[10]. Ioannis K. Valavanis, George M. ,Spyrou and Konstantina S. Nikita, "A comparative study of multi-classification methods for protein fold recognition," Int. j. Comput. Intelligence in Bioinformatics and Systems Biology, vol. 1, no. 3, 2010.
[11]. Abdollah Dehzangi and Bahador Ganjeh Khosravi, "Introducing Novel Physicochemical Based Features to Enhance Protein Fold Prediction Accuracy,"International Conference On Computer Design And Appliations (ICCDA 2010), no.1, pp. 592-596, 2010.
[12]. Wu Qu, Haifeng Sui,Bingru Yang and Wenbin Qian, " Improving protein secondary structure prediction using a multi-modal BP method,"Computers in Biology and Medicine 41,pp. 946-959, 2011.
[13]. Anil Kumar Mandle, Pranita Jain and Shailendra Kumar Shrivastava, "Protein Structure Prediction Using Support Vector Machine," International Journal on Soft Computing (IJSC), vol.3,no. 1,pp. 67-78, February 2012.
[14]. Wolfgang Kabsch and Christian Sander, "Dictionary of Protein Secondary Structure:Pattern Recognition of Hydrogen-Bonded and Geometrical Features," Biopolymers, Vol. 22,pp=2577-2637, 1983.