COPYRIGHT © 2010 JCIT, ISSN 2078-5828 (PRINT), ISSN 2218-5224 (ONLINE), VOLUME 01, ISSUE 01, MANUSCRIPT CODE: 100713

Protein Secondary Structure Prediction using Feed-Forward Neural Network

M. A. Mottalib, Md. Safiur Rahman Mahdi, A.B.M. Zunaid Haque, S.M. Al Mamun, and Hawlader Abdullah Al-Mamun

Abstract— Neural network is one of the successful methods for protein secondary structure prediction. Day to day this technology is modified, improved, even other methods also combined with it to get better result. In this paper we trained feed-forward neural network with proteins for secondary structure prediction. Using Java Object Oriented Neural Engine (JOONE) our achieved accuracy for helix prediction is 71% and for sheet prediction is 65%. This paper is expected to benefit researchers in proteomics by presenting a summary of developments of neural network in this area.

Index Terms— α-helix, β-sheet, bioinformatics, feed-forward neural network.

1 Introduction

Protein structure prediction is the foundation of protein structural biology. Proteins are macromolecules made up from 20 different L-α-amino acids which fold into a particular three-dimensional structure that is distinctive to each protein. This three-dimensional structure is in charge for the function of a protein. The ultimate goal is to understand the function of the protein. So, it is essential to understand the protein structure. Biochemistry refers four distinct aspects of a protein's structure: Primary structure, Secondary structure, Tertiary structure and Quaternary structure. Protein Secondary Structure Prediction (PSSP) means to predict α-helix, β-strand and coils from the amino acid sequence of a protein.

Over the last 20 years, a huge number of works have been done for predicting secondary structure. A lot of strategies and methods have been used and most of them are probabilistic approach. The statistical methods were the very first method used on known protein structures to predict the protein secondary structure. Chou and Fasman, Garnier averaged the probabilities using small window in 1978 [1], [2]. Kabsch & Sander first defined the 3 categories of protein structure α-helix, β-strand and "other" in 1983 by the DSSP Program [3]. The first attempt of using neural networks in protein secondary structure prediction was done by Qian and Sejnowski at 1988. Later Kneller at 1990 [4] and Stolorz at 1992 [5] used the neural networks in various ways to predict the protein secondary structure. Zhang at 1992 [6] and Maclin and Shavlik at 1993 [7] used combination of neural networks and other methods in prediction. We found different levels of accuracy using varieties of methods. Using JOONE, for helix prediction we got 71% accuracy and 65% accuracy for sheet prediction.

The rest of the paper is organized as follows. Section 2 describes the basic concept of neural network. Section 3 describes on a high level, the methods reviewed and followed by their comparison. Section 4 and 5 describes methodology and result respectively. Finally, section 6 ends with a conclusion.

2 Basic Concept of Neural Network

The neural network technique is based on the study of biological nervous system. It is the study of building a computer model which is made of large number of simple, highly interconnected computational units (neurons) operates parallel. Each unit combines its input and according to some threshold value it generates output. Initially random connection strengths (weights) and thresholds (biases) are modified in repeated cycles by maximizing the accuracy of secondary structure assignment using the dataset of known protein structure. This is called the "training" phase. After this phase, the learned "knowledge" (which is actually derived weight and threshold value) is used in "test" phase to predict the unknown protein secondary structure. The network is composed of one input layer, one or more hidden layer and one output layer. Input layer encodes a moving window into amino acid sequence and central residue of the window is predicted. The computation is done in each input layer and output layer. The total input "Ei" to unit "i" is,

$$E_i = \sum_j W_{ij} S_j + b_i \quad (1)$$

Where "bi" is the bias of the unit and the output of each

• *Department of Computer Science and Information Technology, Islamic University of Technology (www.iuoic-dhaka.edu), Board Bazar, Gazipur - 1704, Bangladesh.*

E-mail: {mottalib, mahdi05, nashid, sharif05, hamamun}@iut-dhaka.edu

unit "i" is generated by:

$$S_i = F(E_i) = \frac{1}{1+e^{-E_i}} \quad (2)$$

After calculating each time the error is predicted using the function

$$E = \sum_c \sum_j (O_{J,C} - D_{J,C})^2 \quad (3)$$

until the error is reduced to some satisfactory value. The basic neural network model is given below:

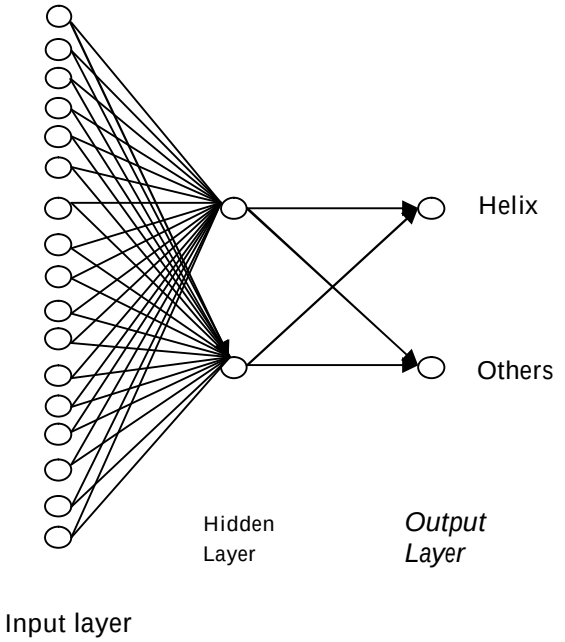

Fig. 1 Basic neural network

3 REVIEW OF PREVIOUS METHODS

Qian and Sejnowski were enormous in introducing a complete new era in protein secondary structure prediction. They used neural networks method which had more accuracy than other previous methods. They gained the accuracy of 64.3%. Qian et al. [8] and Holley et al. [9] worked based on the work of "Kabsch and Sander" [3]. Their work was also quite similar. But Holley et al. gained 63% accuracy which was 1.3% less than Qian et al. In [8] and [9] "supervised learning method" were used which was developed by Rosenblaft (1959) [10] and Widrow and Hoff (1960) [11]. Both [8] and [9] used "Feed forward neural network" and the "Back propagation learning algorithm". Qian et al. took the window size 13 which means the contiguous sequence of 13 amino acids where Holley et al. used the window size of 17. Both of them used 21 inputs. According to [8] the predicted best success rate regarding the number of hidden units is 40. But [9] got their peak success rate in prediction phase taking hidden layer size 2. Though taking more hidden units gave more accuracy to them in training phase but they were poor at test phase. Qian et al. used 3 output positions in the output layer for helix, sheet and coil determination where Holley et al. used only 2 units in output layer. Qian et al. used 106 proteins where subsets of them were used for testing and others were used for training. Other side Holley et al. used 62 proteins. For training first 48 proteins (83158 residues) were used and last 14 proteins (2441 residues) were used as test set. Qian et al. used "artificial structures" before using the real protein database. They measured it with first order artificial structures (no hidden layer) and also second order artificial structures (hidden layers present). They showed effect of noise in data and also effects of irrelevant weights [8]. Holley et al. did not use any artificial structure. They did one additional observation. They excluded the outputs that fall into a range centered near threshold 0.37. By this way they got significant improvement. By taking the strongest 31% of the database the prediction accuracy raise to 79%.They also did physicochemical encoding, means they characterized the amino acids for 48 proteins according to hydrophobicity, charge and backbone flexibility. Hence the accuracy gained for test set of 14 proteins is 61.1%.When they took 20 selected proteins whose structures with resolution better than 2.8 °A , crystallographic R factor less than or equal to 0.25 and sequence homology less than 50%. This case predictive accuracy was 63% and 34% of their strongest prediction was 76% accurate.

3.1 Over Fitting Problem

Though Qian et al. was more successful until they publish their work, but they had some problem like "over fitting". The problem occurred due to huge number of weight value needed to be deducted.

Rost and Sander, 1993 [12] tried to improve the system proposed by [8]. They used two methods to stop over fitting problem:

1. Early stopping which means training stops when the training error is below some threshold.
2. And also plotting different inputs in different networks and making average of the outcomes.

The success of [12] in use of alignment, they feed the multiple alignments to the network in profile manner. For every position of amino acid frequency vector is fed to network. The database in [12] used was 150 representative protein chain of known structure. Their database had not more than 30% similarity where, Zhang [6] used database of 49% homologous and Qian et al. used 46% homologous protein. According to Rost et al. the 130 chain is divided into 7 partitions (which is called 7-fold cross validation) and they are calculated separately and making average of their accuracy. Cross validation is important because accuracy is mostly dependent on which set is chosen as test set. Thus they achieved 69.7% of the three states prediction accuracy.

3.2 Multiple Sequence Alignment

Rost et al. mentioned that with appropriate cutoffs applied in a multiple sequence alignment, all structurally similar proteins can be grouped into a family and approximate structure of the family can be predicted. They used the known protein structure to make the family in training

phase. The family profile of amino acid frequencies at each alignment position was fed into network and they got the prediction accuracy 6% more.

3.3 Balance Training

The distribution of secondary structure types in globular protein is uneven. Approximately there was 32% α-helix, 21% β-strand and 47% loop in the database. So, prediction of loop was easy. So, Rost et al. tried to train the network in equal proportion and they got better result. Later, they used of 2 level network and got benefit from it. Though Qian et al. also used 2 level network but there was no improvement.

3.4 Jury of Networks

It is completely new idea by "Rost and Sander". Jury of networks predict by simple majority vote of a set of 12 different networks. Using this they got 2% improvement in overall accuracy. This was an effect of noise reduction which mitigated the bad effects of incomplete optimization. They had the overall accuracy 69.7% which is better than previous methods [8], [9], [6]. They used five prediction methods. The result is shown in Figure 2.

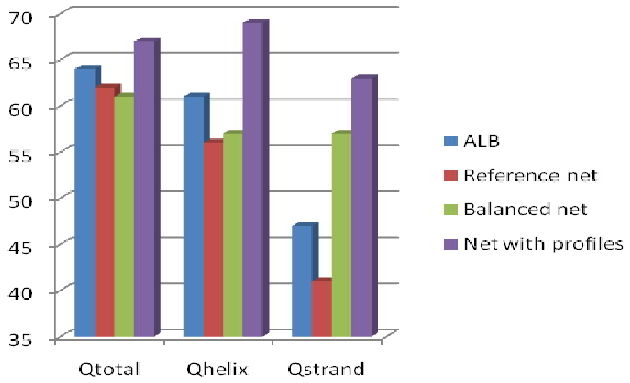

Fig. 2 Testing five secondary-structure prediction methods on the same set of proteins reveals the contribution of different devices to the improvement of accuracy [12].

For completely new protein, 72% of the observed helical and 68% of strand residues was predicted correctly [12] and the overall accuracy was 70.3%. Their method had some limitations too. The result is very poor with non-homologous protein and it is not applicable for membrane, non globular or non-water soluble proteins. They did their method of sequence profiles and neural networks at 1993. Soren et al. used a method, using structural neural network and multiple sequence alignments at 1996 [13]. First, they used "single structured network" and got the accuracy 66.3%. After that, they applied multiple sequence alignment and the accuracy become 71.3%. They used the database of 126 non-homologous globular proteins and 72% residue of the database gave 82% of prediction accuracy [13]. They got the accuracy of 66-67% in single sequence prediction which was 3-4% better than "fully connected network" method. After completing the whole system the result become 71.3% which was quite identical to Rost et al.

3.5 Genetic-neural System

Recently a Hybrid Genetic-Neural System has been introduced by Armano et al. in 2005 [14]. They made a system MASSP3 (Multi agent Secondary Structure Prediction with Post-processing) which does the overall processing. They used "Feed forward ANN" layer that performs a structure-to-structure prediction. Armano et al. [14] used the same database as [15] where the training set contains 1180 sequence obtained for PDB database. Their proteins were more than 25% homologous. In their test set there were 126 non-redundant protein and they used the moving window of size 15. Several Experiments were performed by them:

3.5.1 Optimization of Genetic Experts

They used 600 experts which were randomly generated by guards [14]. They used BLAST-based encoding to get the input. Their hidden layer contained 10-25 neurons and they used back-propagation algorithm. They also filtered inputs by guards. The accuracy they got is 69.1%.

3.5.2 Input Encoding

The population was evolved using covering, single point crossover and mutation operations. The GA performed 60 epochs and the final population contained 550 experts. They filtered the population by removing those who did not match more than 0.1% of the overall inputs used for training. The result was obtained 71.8% accurate.

3.5.3 Expert's Specialization Technique

In this technique, the training was done over whole database for first 5 epochs. In the next epochs only inputs selected by guard were feed. This way the accuracy is raised to 73.2%.

3.5.4 Post-Processing Technique

The most successful one and Armano et al, emphasized most on this. It had a single MLP with moving window of 21 amino acids. They used Low pass Gaussian filter ($\sigma = 0.5$) to encode the output of Multiple experts. The improved accuracy is raised to 76.1%. The improvement is shown in table I.

TABLE I

RESULTS ON THE RS126 TEST SET IN ACCORDANCE WITH THE TRAINING STRATEGIES AND ENCODING TECHNIQUES THAT HAVE BEEN EXPERIMENTED [14]

Experiment	Accuracy
Random population	69.1
Generally-selected population	71.8
Improved experts' specialization technique (global + local)	73.2
Using PSI-BLAST profiles	74.7
Using post-processing	76.1

4 METHODOLOGY

In this section, we describe the test set and the results we obtained for the prediction of the transmembrane helix using feed-forward network. For experimenting we have used a component based neural network framework built in Java named Java Object Oriented Neural Engine (JOONE).

4.1 Data Set

For the protein data we used proteins from the PDB (Protein Data Bank) data sets. We have classified the protein according to their structure, their size and their hydrophobicity. If the particular residue is helix then we have given 1 as the output value, otherwise 0. Table II shows the classification.

TABLE II
AMINO ACID CLASSIFICATION

Criteria	Amino Acids	Value
Electrically charged Side chain	Arg, His, Lys, Asp, Glu	0;0
polar but uncharged side chains	Ser, Thr, Asn, Gln, Tyr	0;1
Special cases	Cys, Gly, Pro	1;0
with hydrophobic side chains	Ala, Ile, Leu, Met, Phe, Trp, Val	1;1

4.2 Training with JOONE

We have trained around 20 proteins with JOONE. We have taken 2 nodes in input layers, 3 nodes in hidden layers, 1 node in output layer, learning rate 0.9, momentum 0.1, and epochs 10000. Table III shows the parameters used in the simulation and Figure 3 shows the network.

Table III
PARAMETERS USED IN EXPERIMENTS

Parameter	Value
Training pattern	4980
Epochs	10000
Learning rate	0.9
Momentum	0.1

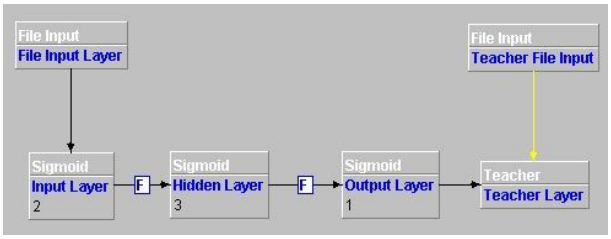

Fig. 3 Feed-forward network used in JOONE.

5 RESULT AND DISCUSSION

5.1 Helix Prediction

We have tested the network with 1R2N, 1MGY, 1JV6, 1S8J and other similar tansmembrane proteins. In total, there are 1456 residues of helix and the network predicts 1048 of them including 186 false positive. So the accuracy is 71%. Accuracy varies with the number of nodes in input layer and sequence similarity of training proteins. Figure 4 shows the learning curve.

5.2 Sheet Prediction

We have tested the network with 1SIP, 2SAM, 1AZ5, 1YTJ and other similar proteins. In total, there are 2040 residues of sheet and the network predicts 1320 of them. So the accuracy is 65%. Accuracy varies with the number of nodes in input layer and sequence similarity of training proteins. Figure 4 shows the learning curve.

4 CONCLUSION

In this paper, we worked with only helix and sheet prediction. Here, we used the feed-forward network architecture. Future experiment can be done which includes coil prediction. Still many challenges remain, requiring the development of alternate strategies to complement/improve existing techniques.

REFERENCES

[1] P. Y. Chou and G. D. Fasman, "Prediction of the secondary structure of proteins from their amino acid sequence," *Advanced Enzymol*, 1978, pp. 47, 45-148.

[2] J. Garnier, D. J. Osguthorpe, and B. Robson, "Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins". *Journal of Molecular Biology*, 1978, 120:97-120.

[3] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features." *Biopolymers*, 1983a, pp. 2577 - 2637.

[4] D. Kneller, F. Cohen, and R. Langridge, "Improvements in protein secondary structure prediction by an enhanced neural network," *Journal of Molecular Biology*, 1990.

[5] P. Stolorz, A. lapedes, and Y. Xia, "Predicting protein secondary structure using neural net and statistical methods," *Journal of Molecular Biology*, 1992.

[6] X. Zhang, J. Mesirov, and D. Waltz, "Hybrid system for protein secondary structure prediction," *Journal of Molecular Biology*, 1992.

[7] R. Maclin and J. Shavlik, "Using knowledge based neural networks to improve algorithms: Refining the chou-fasman algorithm for protein folding," *Journal of Molecular Biology*, 1993.

[8] N. Qian and T. J. Sejnowski, "Predicting the secondary structure of globular proteins using neural network models," *Journal of Molecular Biology*, 1988.

[9] H. L. Holley and M. Karplus, "Protein secondary structure prediction with a neural network," Proc. Natl. Acad. Sci.USA, 1989.

[10] F. Rosenblatt, "Mechanization of thought processes," *Journal of Molecular Biology*, vol. 1, 1959.

[11] R. M. Widrow and M. E. Hoff, "In institute of radio engineers, western electronic show and convention," Convention Record, part 4, pp. 96–104, 1960.

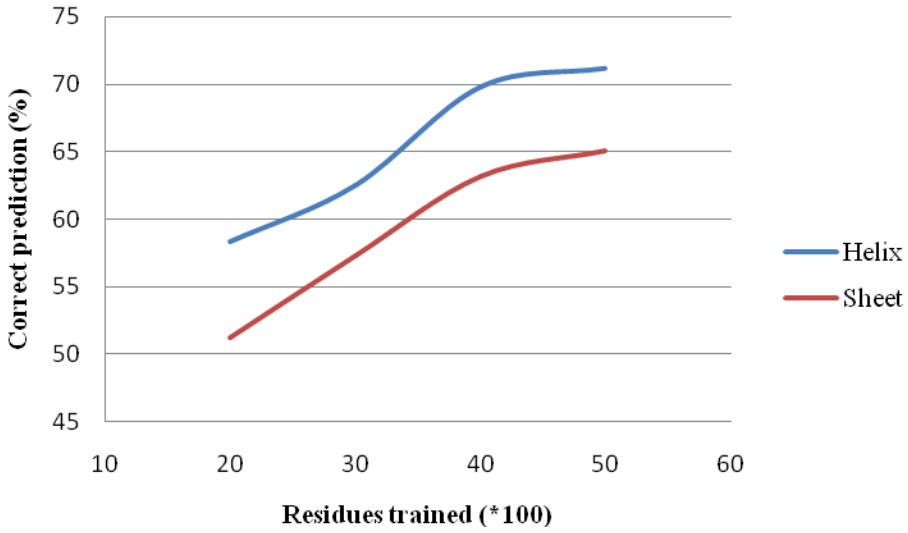

Fig. 4 Learning curve for feed-forward network. The percentage of correct predicted helix and sheet is plotted as a function of the number of amino acids presented during training

[12] B. Rost and C. Sander, "Improved prediction of protein secondary structure by use of sequence profiles and neural networks," in National Academy of Sciences of the United States of America, 1993a.

[13] S. kamaric Riis and A. Krogh, "Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments," *Journal of Computational Biology*, 1996.

[14] G. Armano, G. Mancosu, L. Milanesi, A. Orro, M. Saba, and E. Vargiu, "A hybrid genetic-neural system for predicting protein secondary structure," BMC Bioinformatics, 2005.

[15] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using neural networks and profiles," *Proteins*, 2002.