Journal of Computational Intelligence in Bioinformatics
ISSN 0973-385X Volume 6, Number 1 (2013) pp. 49-57

http://www.ripublication.com/jcib.htm

Comparision of Prediction of Structure of Protein of Soy Beans using Radial basis Function Neural Networks with other Methods for Rs126 and PDB Data Sets

Dr. K. Meena[**] and Dr. M.Manimekalai[*]

[**]*Vice Chancellor, Bharathidasan University, Tiruchirappalli, Tamil Nadu, India.*
[**]*Director, Department of MCA, Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India.*

Abstract

In this paper Prediction of structure of Protein of Soy Beans using Radial Basis Function Neural Networks for RS126 Data set and PDB Data set has been made and compared with other traditional methods namely Chou-Fasman, GOR, APSSP, PHD, Prospect and SSpro.The training and testing sets for both have been taken into consideration to train and test the networks respectively. The major parameter for finding the accuracy of the protein secondary structure prediction is the per-residue prediction accuracy, Q_3, which gives the percentage of all correctly predicted residues within the three-state (H, E, C) classes, and has also been employed for assessment of prediction approaches. The performance of the RBFNN protein secondary structure prediction models [1] is evaluated based on their prediction accuracy [2]. The accuracy of the developed approach is compared with other traditional methods to explore the performance of the proposed approach. It is found that the proposed techniques provide a prediction accuracy of about 81% which is very significant. The accuracy for different width of sliding windows. It clearly shows that, with the increase in the sliding window width the accuracy also increases.

Keywords: RBFNN, Prediction Accuracy, Training Set, Test Set, Sliding Window

Introduction

The secondary structure of a segment of polypeptide chain is defined as the local spatial arrangement of its main-chain atoms without regard to the conformation of its side chains or to its relationship with other segments. Alpha helices, beta sheets and turns are the most common secondary structures in proteins.

The other structures which cannot be classified as one of the standard three classes is grouped into a category called other or random coil. Regular secondary structure conformations in segments of a polypeptide chain occur when all the bond angles in that polypeptide segment are equal to each other, and all the bond angles are equal. The rotational angles for and bonds for common regular secondary structures are shown in the Table1

Table 1: Parameters of Regular Secondary Structures.

Structure			**n**	**P(Å)**	**A**	**H-bond(CO,HN)**
Right handed-helix	-57	47	3.6	5.4	13	i, i+2
310 -helix	-74	-4	3.0	6.0	10	i, i+3
pi–helix	-57	-70	4.4	5.0	16	i, i+4
Parallel beta strand	-119	113	2.0	604		
Antiparallel beta strand	-139	135	2.0	6.8		

In this table, n represents the number of residues per helical turn, p is the helical pitch, and A represents the atoms in H-bonded loop.

Training and Testing Sets for PDB Data Set

Table 2 shows a portion of the PDB Data set created for the Soybeans protein (1AVU). The Soybeans protein contains 181 amino acids, where each amino acid in the sequence is denoted by several lines in the file. The last three columns give the detailed orthogonal coordinates for X, Y and Z respectively. The coordinates are in the angstrom units. Each amino acid residue is composed of several atoms, the geometrical center coordinates of residues is used which was based on the atomic coordinates, to represent the residue position

Table 2: A Portion of PDB Data File for the soybeans protein.

Atom Serial Number	**Atom Name**	**Residue Name**	**Residue Number**	**Orthogonal Co-Ordinates**		
				X	**Y**	**Z**
1	N	ASP	1	3.484	35.111	24.104
2	CA	ASP	1	4.282	34.388	25.144
3	C	ASP	1	4.653	32.944	24.751
4	O	ASP	1	4.578	32.023	25.576
5	CB	ASP	1	3.553	34.407	26.497
6	CG	ASP	1	2.032	34.284	26.353
7	OD1	ASP	1	1.359	35.331	26.187
8	OD2	ASP	1	1.512	33.148	26.451
9	N	PHE	2	5.072	33.774	23.496
10	CA	PHE	2	5.604	31.514	22.971
11	C	PHE	2	6.892	31.024	23.592
12	O	PHE	2	7.761	31.825	23.964
13	CB	PHE	2	5.812	31.626	21.485
14	CG	PHE	2	4.557	31.486	20.715
15	CD1	PHE	2	3.47	30.838	21.276
16	CD2	PHE	2	4.455	31.982	19.43
17	CE1	PHE	2	2.299	30.682	20.562
18	CE2	PHE	2	3.29	31.835	18.704
19	CZ	PHE	2	2.28	31.183	19.267
20	N	VAL	3	6.96	29.702	23.776
21	CA	VAL	3	8.153	29.056	24.302
22	C	VAL	3	9.118	28.819	23.138

In order to test the performance of the proposed approach using RBFNN totally about 39 Soybeans protein sequences (about 19287 amino acid residues) and their spatial distances are used as training sets. These were obtained from Protein Data Bank (PDB), whose archive contains macromolecular structure data about proteins, nucleic acid, protein-nucleic acid complexes and viruses. The keyword Soybeans is used to identify all the sequence, and their corresponding secondary structure parameters (α-helix, -strand, and coil) are determined [3]. The Soybeans protein sequence consisting of 71 amino acid residues was used to evaluate the RBFNN optimized by GA. The proposed approach is used to predict the first 41 amino acid and the protein secondary structure. The training and testing procedures of a GA based RBFNN were based on the neural network tool box of MATLAB.

Prediction Accuracy

The major parameter for finding the accuracy of the protein secondary structure prediction is the per-residue prediction accuracy, Q_3, which gives the percentage of all correctly predicted residues within the three-state (H, E, C) classes, and is almost always employed for assessment of prediction Approaches[4]. They are

$$Q3=\frac{\text{number of residues predicted}}{\text{total number of residues in the data set}}$$

$$QH=\frac{\text{number of alpha}-\text{helix residues predicted}}{\text{total number of alpha}-\text{helix residues in the data set}}$$

$$QE=\frac{\text{number of extended beta}-\text{helix residues predicted}}{\text{total number of extended beta}-\text{helix residues in the data}}$$

$$QC=\frac{\text{number of extended beta}-\text{helix residues predicted}}{\text{total number of extended beta}-\text{helix residues in the data set}}$$

Segment Overlap Score

The **Segment OVerlap score (SOV)** depends on the average overlap between the observed and the predicted segments instead of the average per-residue accuracy. The SOV measures provide more elaborate scoring, in which the predictions that have high per-residue accuracy but deviate from experimental segment length distributions that are assigned lower scores. For instance, the definition of the SOV measure for the protein structures like
-helices is as follows:

$$SOV_\alpha=1/N_\alpha \sum \text{minOV}(s_1, s_2) + \delta(s_1, s_2)/\underline{\text{maxOV}}(s_1, s_2)$$

Here, s_1 and s_2 are the observed and predicted secondary structure segments in the -helix. s is the number of all segment pairs (s_1, s_2), where s_1 and s_2 have at least one residue in a -helix state in common, minOV(s_1, s_2), is the length of the actual overlap of s_1 and s_2 and maxOV(s_1, s_2), is the length of the total extent for which either of the segments s_1 or s_2 has a residue in the protein structure state. N is the total number of amino acid residues observed in the-helix conformation. The definition of (s_1, s_2), is as follows:

$$\delta(s_1, s_2)=\min\begin{Bmatrix} \max OV(s_1 - s_2) - \min OV(s_1, s_2) \\ \min OV(s_1, s_2) \\ \text{int}(0.5 * len(s_1)) \\ \text{int}(0.5 * len(s_2)) \end{Bmatrix}$$

Here, len(s_1) is the number of amino acid residues in the segment s_1. The segment overlap measure for all three states, SOV (%), is similar to the Q_3 (%) sensitivity measure. Hence SOV is given as.

$$SOV(\%) = \frac{1}{N} \sum_{i \in H,E,C} \frac{\min OV(s1, s2) + \delta(s1, s2) \times len(s)}{\max OV(s1, s2)} \times 100$$

Here, s_1 and s_2 are the observed and predicted secondary structure segments in state i. N is the total length of protein sequences under consideration.

Matthew's Correlation Coefficient

This Correlation Coefficient can also be used to measure the prediction accuracy. This measurement is given by

$$\rho_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FN_i)(TP_i + FP_i)(TN_i + FP_i)(TN_i + FN_i)}}$$

Where *TP*, *FP*, *FN*, *TN* are the number of true positives, false positives, false negatives, and true negatives for class i {H, E, C}, respectively. The result is a value between –1 and 1, such that 1 shows complete agreement, -1 show complete disagreement, and 0 show that the prediction is uncorrelated with the results [5].

Accuracies based on Different Sliding Window Widths

Table 2 shows the accuracy for different width of sliding windows. It is seen that the accuracy by a single residue (width= 1) information is low, (i.e.) only 57.28. However prediction accuracies show dramatic improvement for the other widths, which shows the surrounding residues information play a vital role in the secondary structure process. From Table 3, it is seen that width 9 shows the best performance with an accuracy of about 77.44%.

Table 3: Accuracy vs. Different Sliding Windows Width in PDB.

width	Training set Q_3 (%)	Testing Set Q_3 (%)	Q_H(%)	Q_E(%)	Q_C(%)
1	67.43	57.28	61.26	38.07	58.24
3	91.56	69.11	69.08	51.26	74.30
5	97.32	74.44	71.87	59.72	79.81
7	98.21	76.28	71.04	63.40	84.04
9	99.83	77.44	73.84	60.42	81.38
11	100	76.98	69.43	64.82	81.83

The bar chart given in Figure 1 represents the accuracies for the different sliding window width. It clearly shows that, with the increase in the sliding window width the accuracy also increases.

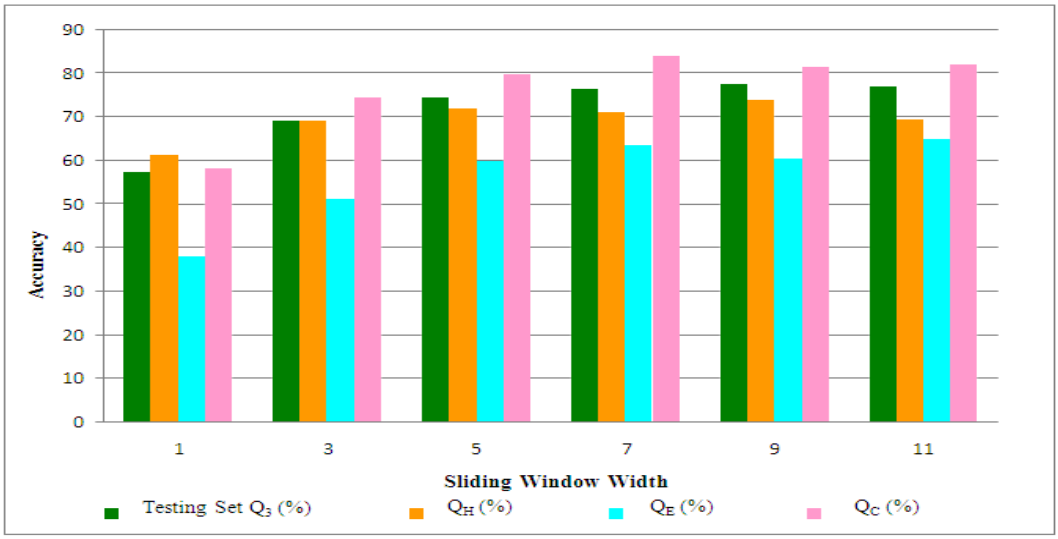


Figure 1: Accuracies depending on different sliding window width.

Comparing Accuracies of various prediction methods

Several approaches, such as Chou and Fasman method[6], GOR method[7], PhD method and etc., have been used in the experiment for the predicting the secondary structure of Soybeans.

The results are obtained and tabulated. Table 4 shows the detailed accuracies of different methods and indicate that the developed approach using RBFNN, conformational classification method, has a better prediction accuracy compared to other traditional approaches.

The proposed techniques provide a prediction accuracy of about 81% which is very significant.

Table 4: Accuracy of Different Methods in PDB Data Set.

Prediction Accuracies of Various Techniques		
Methods	**Accuracy %**	**Errsig (Accuracy)**
Chou-Fasman	59	-
GOR	65	-
APSSP	70	±2.2
PHD	73	±1.9
Prospect	73.5	±2.6
SSpro	77	±2.2
Proposed RBFNN Method	81	±3.2

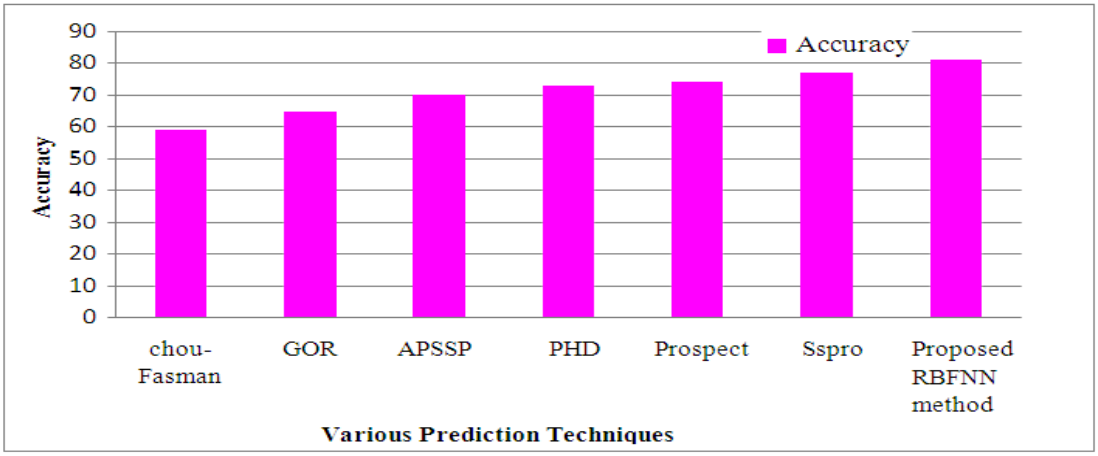


Figure 2: Accuracy of Various Prediction Techniques in PDP Data set.

The bar chart given in Figure 2 represents the performance accuracy of various prediction techniques in predicting the secondary structure of Soybeans. It is clearly observed form the chart that, the proposed prediction method using RBFNN provides higher accuracies compared to the other earlier prediction techniques.

Training and Testing Sets for Rs 126 Data Set

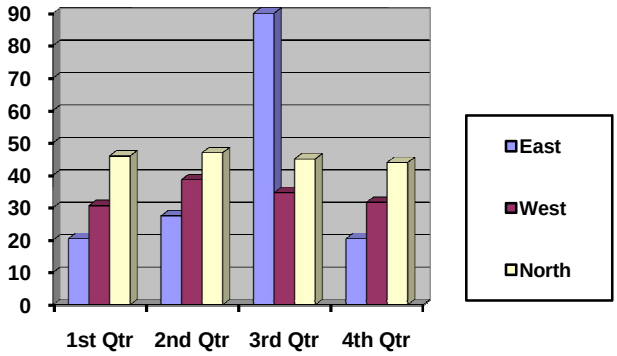
The data set used to evaluate the accuracy of the classifiers is the set 126 non-homologous globular protein chains used in the experiment of Rost and Sander, which is commonly referred to as the RS126 set. The dataset contained 23349 residues with 32% α-helix, 23% β-strand, and 45% coil.Various prediction methods like Chou and Fasman method, GOR method, PhD method and etc. have been used in the experiments using the RS126 data set for the prediction of secondary structure.

Performance Comparison of various prediction methods

Table 5 shows the detailed accuracies of different methods and indicate that the developed approach using RBFNN, conformational classification method, has a better accuracy than other conventional methods.

Table 5: Prediction accuracies of various techniques in rs 126 data set.

Methods	**Accuracy %**	**Errsig (Accuracy)**
Chou-Fasman	60	-
GOR	66	-
APSSP	72	±2.3
PHD	73.5	±1.8
Prospect	74.5	±2.7
SSpro	78.5	±2.3
Proposed RBFNN Method	82.5	±3.3



In RS 126 set, the prediction accuracy of the proposed approach is 82.5% which is very higher than the conventional prediction approaches.
The bar chart in Figure 3 represents the accuracies of various prediction methods in RS 126 data set. From the bar chart it is clearly seen that the proposed RBFNN approach provides a better results than all the other traditional prediction techniques.

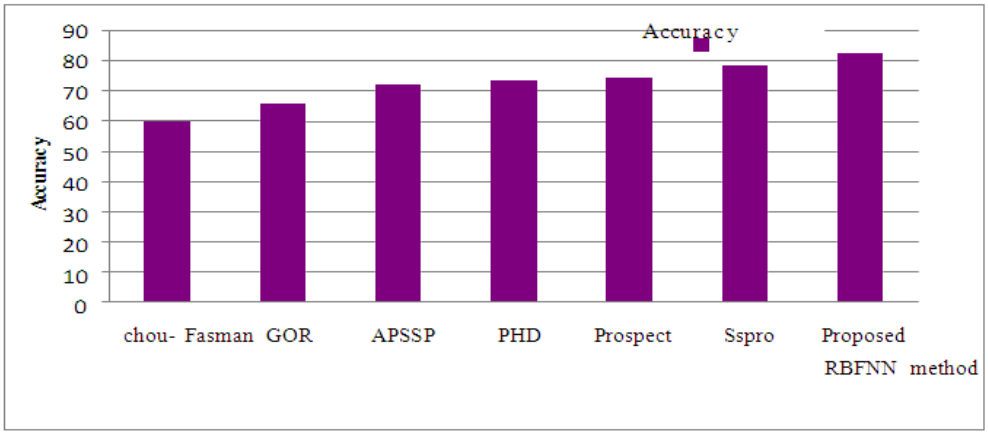


Figure 3: Accuracy of Various Prediction Techniques in RS 126 Data set.

Conclusion

The interest in secondary structure prediction is influenced by the extraordinary pace of discovery of new protein sequences in genome sequencing experiments[8]. The proposed method can be a useful technique in producing good variety of crops and therefore, it may assist the agricultural researchers in future. This can also be used as an effective technique for drug discovery[9]. This will lead to the discovery of new variety of crops. This approach will lead to research in the field of microbiology and bio technology.

References

[1] Boscott, P.E., Barton, G.J. and Richards, W.G., "Secondary Structure Prediction for Modelling by Homology", PEDS, Vol. 6, Issue 3, pp.261–266, January 1993.

[2] Marti-Renom et al., "Comparative protein structure modeling of genes and genomes," Rev. Biophys. Biomol, Struct, 29:291-325, 2000.

[3] Solovyev, V.V. and Salamov, A.A., "Predicting alpha-helix and beta-strand segments of globular proteins". Computer Applications in the Biosciences, 10, 661-669, 1994.

[4] Pollastri, G., Przybylski, D., Rost, B. and Baldi, P., "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*", 47(2), 228–235, 2002.

[5] Karlin, S., Bucher, P., "Correlation analysis of amino acid usage in protein classes", Proc Natl Acad Sci, USA 1992.

[6] Chou, P.Y., Fasman, G.D., "Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins". *Biochemistry* 13(2):211-22, 1974.

[7] Garnier, J.J., Gibrat, J.F. and Robson, "GOR method for predicting protein secondary structure from amino acid sequence, Methods Enzymol", 266, 540–550, 1996.

[8] Marti-Renom et al., "Comparative protein structure modeling of genes and genomes," Rev. Biophys. Biomol, Struct, 29:291-325, 2000

[9] Rost, B., "Rising accuracy of protein secondary structure prediction", D.Chasman, Ed., "Protein structure determination, analysis, and modeling for drug discovery", New York: Dekker, pp. 207–249, 2003.