Tropical Agricultural Research Vol. 23 (1): 91 - 98 (2011)
Short communication

A Simple Comparison between Specific Protein Secondary Structure Prediction Tools

T.J. Koswatta, P. Samaraweera[1*] and V.A. Sumanasinghe[2]

Postgraduate Institute of Agriculture
University of Peradeniya
Sri Lanka

ABSTRACT. *A comparative evaluation of five widely used protein secondary structure prediction programs available in World Wide Web was carried out. Secondary structure data of ten proteins containing 190 secondary structure motifs were collected from Protein Data Bank (PDB). The amino acid sequences of the proteins were then evaluated using GOR, PSIPRED, HNN, PROF, and YASPIN secondary structure prediction tools and the results were compared with the structural information obtained from PDB. The study reveals considerable differences between results obtained from each program. Within the limit of this comparative study, PSIPRED showed the highest prediction accuracy with 77 % accuracy in α helix prediction and 70 % accuracy in β strand prediction. Furthermore, the level of accuracy varied with the length of the secondary structure motifs. Highest accuracies were obtained for α helices of 16-20 amino acids and β strands of 7-9 amino acids in length. The results suggest that, among the most frequently used software programs available in World Wide Web, PSIPRED is the tool that gives the best results for secondary structure prediction.*

Keywords: *Bioinformatics, proteins, secondary structure*

INTRODUCTION

Proteins are the most complex chemical substances in nature. They vary in shape, size and mobility. The structure of a protein is described at four different levels, namely, primary, secondary, tertiary, and quaternary structure. The primary structure is the sequence of amino acids in the protein chain. The secondary structure describes the local conformation of the segments of polypeptide backbone. The three-dimensional structures are produced by folding secondary structures into one or several domains. Since the functionality of the protein is determined by its three-dimensional structures, the knowledge about the structure of a protein may provide clues to its function. Moreover, it will also be helpful in understanding the role and responsibilities of the protein in the cell. (Zhu *et al.*, 2002)

Secondary structures are defined as the repetitive hydrogen-bonded shapes or substructures that make up sequentially proximal amino acids of proteins. Some of the most common protein structures are α helices and β strands. These structures are characterized by regular hydrogen bonding patterns that persist over three or more consecutive residues. In addition to

[1] Department of Molecular Biology and Biotechnology, Faculty of Science, University of Peradeniya, Sri Lanka

[2] Department of Agricultural Biology, Faculty of Agriculture, University of Peradeniya, Peradeniya, Sri Lanka

[*] Author for correspondence: psam@pdn.ac.lk

these two very abundant forms of secondary structures, there are several other less abundant structures, including β turns, Ω loops and 3/10 helices. The remaining unclassified or unclassifiable substructures are typically called random coil or more properly unstructured regions (Wishart, 2005). The α helix is the most abundant secondary structure in proteins. It has 3.6 amino acids per turn. Each amino acid in a α helix forms H-bond with the next third residue in the helix. The length of α helix varies from 5 to 40 amino acids, with average length of 10 residues. β Strands have fully extended conformation. They form β sheets by making H- bonds between an average of 5 -10 residues farther down in the chain (Mount, 2004).

Proteins are products of evolution. Their sequences are encoded histories of mutation and selection over millions of years. Therefore, understanding protein structures are very important for many studies. Two laboratory techniques, X-ray crystallography and Nuclear Magnetic Resonance (NMR) spectroscopy, can yield comprehensive structural information about protein at atomic resolution. These two techniques have advantages and disadvantages. Although they provide accurate results, due to complexity, expensiveness, and time-consuming nature of these experimental techniques the progress of protein structure determination has been slow. As a result, bioinformatics software programs have been developed to predict, evaluate, and visualize structures of proteins from their amino acid sequences. They are more flexible and avoid technical and time limitations of experimental methods.

Although sequence similarity would suggest structure similarity between homologous proteins, in some cases, analyzing proteins and their relationship through identifying sequence similarities by sequences alignment could give false results. However, structure similarities are better indicative of similarities of proteins functions because in functionally related protein structures are far more conserved than the sequence. Therefore structure comparisons allow identification of evolutionary relationship that would otherwise be unidentifiable via sequence comparison alone (Baxevanis & Outellette, 2005).

Efficient automatic methods for protein structure prediction are becoming increasingly important because of the influx of nucleotide sequence data arising from sequencing projects. Therefore, a number of bioinformatics programs for secondary structure prediction of proteins have been developed. All secondary structure prediction methods assume that there exists a correlation between the amino acid sequence and the secondary structure. The usual assumption is that a given short stretch of sequence may be more likely to form one kind of secondary structure than another. Many bioinformatics based secondary structure prediction methods examine a sequence window of 3-17 residues and assume that the central amino acid in the window will adopt a confirmation that is determined by side groups of all the amino acids in the window. This window size is within the range of 5-40 residues in α helices and 5-10 residues in β strands which are often found in proteins (Mount, 2004).

A brief overview of the tools

GOR IV (http://npsa-pbil.ibcp.fr/cgibin/npsa_automat.pl?page=npsa_gor4.html):
The GOR (**G**arnier, **O**sguthorpe, and **R**obson) method is an information-based method for the prediction of secondary structures of proteins. There is no defined constant decision. GOR IV uses all possible pair frequencies within the window of 17 amino acid residues (Garnier *et al.*, 1996). It was developed in the late 1970's shortly after the simpler Chou-Fasman method (Chou & Fasman, 1974). Like Chou-Fasman method, GOR method is also

based on probability parameters derived from empirical studies of known protein tertiary structures solved by X-ray crystallography. The program gives two outputs: one is eye-friendly and gives the sequence and the predicted secondary structure in parallel rows, with symbols H= α helix, E=extended or β strand and C=coil; the second gives the probability values for each secondary structure at each amino acid position. The predicted secondary structure is the one with the highest probability-compatible structure with a predicted helix segment of at least four residues and a predicted extended segment of at least two residues (Garnier *et al.*, 1996).

HNN (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_nn.html):
The HNN (Hierarchical Neural Network) prediction method employs two networks to predict structures: a sequence-to-structure network and a structure-to-structure network. Thus, the prediction is only based on local information. Neural network methods are trained to recognize amino acid patterns by providing a data set containing known structures. The algorithm then identifies the structures present in unknowns.

YASPIN (http://www.ibi.vu.nl/programs/yaspinwww/):
YASPIN is a Hidden Neural Network (HNN) secondary structure prediction method. It uses a feed-forward perception network with one hidden layer to predict the secondary structure present in the query sequence. Then, a Hidden Markov Model (HMM) filters these predictions (Lin *et al.*, 2005). The prediction results are converted into 3-state secondary structure predictions (‘H’- α helix, ‘E’- β strand and ‘-’-other). The YASPIN neural network uses the soft max transition function with a window of 15 residues. For each residue in this window, 20 units are used for the scores in the position-specific scoring matrix (PSSM) and one unit is used to mark where the window spans the terminals of protein chains (Lin *et al.*, 2005).

PROF (http://www.aber.ac.uk/~phiwww/prof/):
Ouali and King (2000) developed the PROF system. It is used in Discrimination of Secondary structure Class (DSC) program to predict secondary structures. DSC is based on decomposing secondary structure prediction into the basic concepts and then using simple and linear statistical methods to combine the concepts for prediction (Ross and Sternberg, 1996). PROF program provides result only via mail. It does not give URL link for graphical view. PROF Result page summarizes the whole sequence in a horizontal way and gives the probability of each amino acid and predicted structure.

PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/):
The PSIPERD program incorporates PSIPRED, GenTHREADER, and MEMSAT2 methods for protein structure prediction. This prediction method employs two feed-forward neural networks, which perform an analysis on the output obtained from PSI-BLAST. The PSIPRED server allows users to submit a protein sequence, perform a prediction of their choice, and receive the results of the prediction both textually via e-mail and graphically via the web. It also allows the user to download graphical representation in PDF file format (McGuffin *et al.*, 2000).

The accuracy of the protein secondary structure prediction programs differs from each other. For molecular biologist, correct structure prediction is a more important factor for understanding protein function, reconstructing protein structures, studying protein–protein interactions and rationally designing drugs (Duan *et al.*, 2008). Therefore, it is of immense importance to identify the best software, which has higher prediction accuracy. This study gives a comparison of five different freely accessible secondary structure prediction programs.

MATERIALS AND METHODS

All analyses were performed on a computer with Intel dual core 2.0 GHz processors and 1 GB RAM. The operating system was Windows XP. The accuracy of the secondary structures prediction programs were evaluated by comparing the results obtained from the programs with the data retrieved from the Protein Data Bank.

Secondary structure prediction programs

The secondary structure prediction tools evaluated were YASPIN, PSIPRED, GOR IV, HNN and PROF. The selection of the five programs was done randomly based on accessibility among the free software programs.

Test data

In this study, Protein Data Bank (PDB) (http://www.pdb.org/pdb/home/home.do) was the source for identification of exact secondary structures. As the test dataset, ten unrelated proteins with complete structure information were selected from PDB (PDB accession numbers: 1GEJ, 2WSK, 3MQ4, 3KH9, 3I6L, 3OE0, 1SB7, 3M7S, 3ODD, and 2Z8R) and their secondary structure information and amino acid sequences were retrieved.

Testing Secondary Structure Prediction Tools

The amino acid sequences of the test proteins were submitted to the five secondary structure prediction servers, namely, YASPIN, PSIPRED, GOR IV, HNN, and PROF. The predicted secondary structure elements were divided into several groups depending on their sequences length. The α helixes were categorized into four groups with sequence lengths 6-10, 11-15, 16-20, and 21-26 amino acids. The β strands were divided into three groups with sequence lengths 4-6, 7-9, and >10 amino acids. Then, experimentally determined secondary structures of the ten proteins in PDB were compared with results obtained from the secondary structure prediction tools. The predictions were divided into different classes based on classification of McGuffin and Jones (2003) as described by Lin *et al.* (2005). However, the method was modified to obtain five classes. They were non-prediction, wrong prediction, over prediction (prediction of more than 125 % amino acid residues at a particular secondary structure position), under prediction (prediction of less than 75 % amino acids residues at a particular secondary structure position) and fair prediction. The five prediction types are illustrated in Fig. 1.

```
SEQ   HPVMINYLKQLGITALELLPVAQFASEPRLQRMGLGYNPVAMFALHPAYACSPETA
PDB   SCHHHHHHHHHTCCEEEESCEEEEECCHHHHTTTCCCCHHHHHHHCGGGCSSEEEE
TOOL  CCHHHHHHHHHCHHHHHHCCEEEECHHHHHHCCCCCCCCCEEEEECCCCCCCCCC
PREDICT  FFFFFFFFF   WWWW  UUUU OOOOOO       WWWWWWW       NNNN
```

Fig. 1. Schematic representation of five protein secondary structure prediction types. **N**, non-prediction; **W**, wrong prediction; **U**, under prediction; **O**, over prediction; **F**, fair prediction

The accuracy of secondary structure prediction tools was expressed using the formula:

$$\textit{Accuracy} = \frac{\textit{Number of fair predictions}}{\textit{Total number of predictions}}$$

Number of fair predictions (correct predictions) is the number of secondary structures that were both present in PDB structures and predicted by the prediction tool.

RESULTS AND DISCUSSION

A total of 78 α helix and 112 β strands motifs in the ten sequences were considered in analysis. The prediction accuracies of GOR, PSIPRED, HNN, PROF, and YASPIN on these 190 motifs are presented in Fig. 2 for α helices and Fig. 3 for β strands.

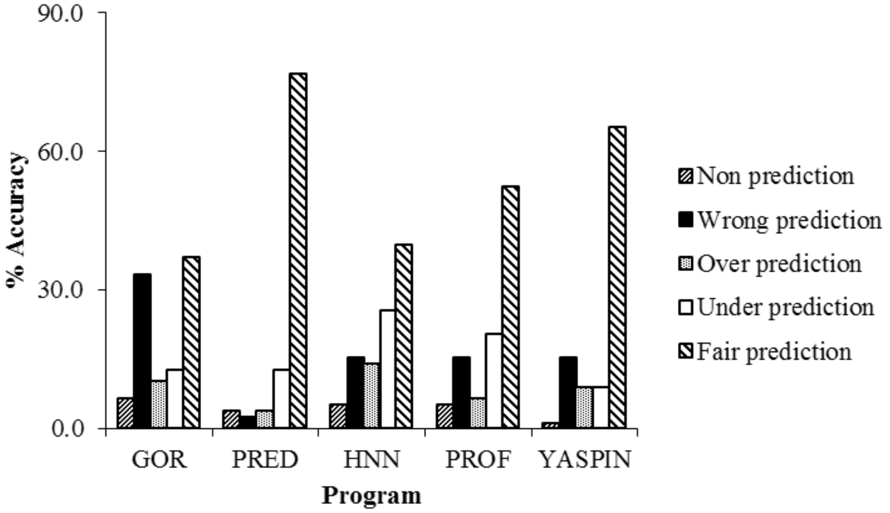

Fig. 2. Comparison of α helix prediction accuracy different secondary structure prediction programs

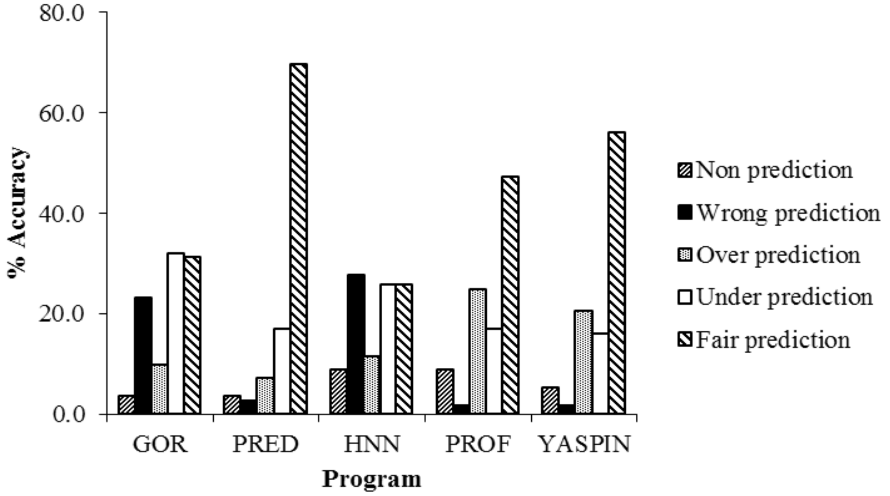

Fig. 3. Comparison of β strand prediction accuracy different secondary structure prediction programs

All the secondary structure prediction programs have higher prediction accuracy for α helix region than the β strands regions. Among the five secondary structure predication programs PSIPRED and YASPIN software programs have higher prediction accuracy of secondary structure than the other three. PSIPRED has the highest prediction accuracy compared to other four programs with 77 % accuracy of prediction in α helix regions and 70 % in β strands. YASPIN has the second highest accuracy of 65 % in α helix region prediction and 56 % in β strands prediction. PROF program accuracy was almost 53 % in α helix region and 47 % in β strand region. GOR and HNN had less than 40 % accuracy (Fig. 2 & Fig. 3).

In addition, both α helix and β strand structures which contained a large number of amino acid residues were more accurately predicted than those with fewer residues (Table 1 & Table 2). The α helix regions that contained less than five amino acid residues and β strands that contained less than four amino acid residues were predicted with less accuracy. The helix regions with more than 20 amino acid residues showed more than 60 % prediction accuracy by all five tools. β Strands regions that contained more than seven residues gave both under predictions and fair predictions.

Table 1. Number of predicted positions and percentage predictions of α helical regions of different lengths.

	GOR		PRED		HNN		PROF		YASPIN	
	No.	**(%)**	**No.**	**(%)**	**No.**	**(%)**	**No.**	**(%)**	**No.**	**(%)**
6-10 amino acids:										
Non prediction	5	12.82	3	7.69	4	10.26	4	10.26	1	2.56
Wrong prediction	16	41.02	2	5.13	8	20.51	9	23.08	8	20.51
Over prediction	6	15.38	2	5.13	6	15.38	5	12.82	6	15.38
Under prediction	2	5.13	6	15.38	11	28.2	6	15.38	2	5.13
Fair prediction	10	25.64	26	66.67	10	25.64	15	38.46	22	56.41
11-15 amino acids:										
Non prediction	0	0	0	0	0	0	0	0	0	0
Wrong prediction	4	18.18	0	0	2	9.09	1	4.54	1	4.54
Over prediction	2	9.09	1	11.11	3	13.63	0	0	1	4.54
Under prediction	5	22.72	3	13.63	3	13.63	6	27.27	2	9.09
Fair prediction	11	50	18	81.81	14	63.63	15	68.18	18	81.81
16-20 amino acids:										
Non prediction	0	0	0	0	0	0	0	0	0	0
Wrong prediction	1	14.28	0	0	0	0	1	14.28	0	0
Over prediction	0	0	0	0	2	28.57	0	0	0	0
Under prediction	2	28.57	1	14.28	2	28.57	2	28.57	2	28.57
Fair prediction	4	57.14	6	85.71	3	42.86	4	57.14	5	71.43
>20 amino acids:										
Non prediction	0	0	0	0	0	0	0	0	0	0
Wrong prediction	5	50	0	0	2	20	1	10	3	30
Over prediction	0	0	0	0	0	0	0	0	0	0
Under prediction	1	10	0	0	4	40	2	20	1	10

Fair prediction	4	40	10	100	4	40	7	70	6	60

Previous studies show that YASPIN has the highest prediction accuracy than the PSIPRED (Lin *et al.,* 2005). The differences in the prediction type employed here could contribute to the change of accuracy observed between present study and that of Lin *et al.* (2005). In addition, the method of calculation of accuracy is also different between the two studies. The study by Lin *et al.* (2005) was carried out according to the Q3, SOV, and Matthew's correlations accuracy measures but in the present study only a simple mathematical equation was used for convenience and simplicity.

Table 2. Number of predicted positions and percentage predictions of β strands of different lengths

	GOR		PRED		HNN		PROF		YASPIN	
	No.	**(%)**	**No.**	**(%)**	**No.**	**(%)**	**No.**	**(%)**	**No.**	**(%)**
4-6 amino acids:										
Non prediction	4	5.48	2	2.74	8	10.96	8	10.96	5	6.85
Wrong prediction	23	31.51	2	2.74	19	26.03	2	2.74	1	1.37
Over prediction	11	15.07	6	8.22	13	17.81	25	34.25	18	24.66
Under prediction	17	23.29	9	12.33	15	20.55	10	13.7	8	10.96
Fair prediction	18	24.66	54	73.97	18	24.66	28	38.35	41	56.16
7-9 amino acids:										
Non prediction	0	0	2	6.67	2	6.67	2	6.67	1	3.33
Wrong prediction	3	10	1	3.33	9	30	0		0	0
Over prediction	0	0	2	6.67	0	0	3	10	5	16.67
Under prediction	13	43.33	6	20	9	30	5	16.67	7	23.33
Fair prediction	14	46.67	19	63.33	10	33.33	20	66.67	17	56.67
>10 amino acids:										
Non prediction	0	0	0	0	0	0	0	0	0	0
Wrong prediction	0	0	0	0	3	33.33	0	0	1	0
Over prediction	0	0	0	0	0	0	0	0	0	0
Under prediction	6	66.67	4	44.44	5	55.56	4	44.44	3	33.33
Fair prediction	3	33.33	5	55.56	1	11.11	5	55.56	5	55.56

The documentation of results by each program also has differences in appearance. PSIPRED has graphical and colored output of results, which make it very clear and easy for further analysis. It also provides the facility to download result in PDF format. GOR and HNN both give the same graphical appearance and percentage of occurrence of each secondary structure in query sequence. However, the disadvantage of GOR and HNN web tools is that their accuracy of predictions is less than that of others. YASPIN does not provide any graphical view as PSIPRED; it gives the predicted secondary structure aligned with the query sequence used.

CONCLUSIONS

The present study shows that there is a considerable variation in the performance of currently available secondary structure prediction tools. Among the most frequently used software in the World Wide Web, PSIPRED is the best program for secondary structure prediction.

In general, not all software programs provide 100 % accuracy in prediction of structure, but they recognize a sizeable portion of secondary structures present in query sequences. As the amount of amino acid sequencing data are gradually increased with time, the efficiency of prediction will become more important for evaluating function of the unknown proteins because the structure of a protein is a key factor for identification of protein function. This information might provide guidance for selecting a better protein secondary structures prediction program.

REFERENCES

Baxevanis, A.D. and Outellette, B.F.F. (2004). Bioinformatics: A parctical guide to the Analysis of Genes and Proteins. Wiley-Interscience, New York, USA. pp. 223-252.

Chou, P.Y. and Fasman, G.D. (1974) Prediction of protein conformation. Biochemistry (Mosc.) *13*, 222–245.

Duan, M., Huang, M., Ma, C., Li, L. and Zhou, Y. (2008). Position-specific residue preference features around the ends of helices and strands and a novel strategy for the prediction of secondary structures. Protein Science *17*, 1505-1512.

Garnier, J., Gibrat, J.F. and Robson B. (1996). GOR secondary structure prediction method version IV. pp. 540-553. In: R.F. Doolittle (Ed.) Computer Methods for Macromolecular Sequence Analysis, Methods in Enzymology, vol. 266. Academic Press, USA.

Lin, K., Simossis, V.A., Taylor, W.R. and Heringa, J. (2005). A simple and fast secondary structure prediction method using hidden neural networks. Bioinformatics *21*, 152-159.

McGuffin, L.J., Bryson, K., and Jones, D.T. (2000). The PSIPRED protein sturcture prediction server. Bioinformatic Applications Note *16(4)*, 404-405.

McGuffin, L.J. and Jones, D.T. (2003). Benchmarking secondary structure prediction for fold recognition. Proteins, *52*, 166–175.

Mount, D.M., (2004). Bioinformatics: Sequnence and Genome Analysis, 2nd Edition), Cold Spring Harbor Laboratary Press, New York, USA., pp. 409-469.

Ouali, M. and King, R.D. (2000). Cascaded multiple classifiers for secondary structure prediction. Protein Science *9*, 1162-1176.

Ross, D.K., and Sternberg, M.J.E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Science. 5, 2298-2310.

Wishart, D.S. (2005). Metabolomics: The Principles and Potential Applications to Transplantation. American J. of Transplantation, *5*, 2814–2820.

Zhu. H, Yoshihara, I. and Yamamori, K. (2002). Prediction of Protein Secondary Structure by Multi-Modal Neural Networks. Proceedings of the International Joint Conference on Neural Networks 2002, 280-285.