Trends in Bioinformatics

ISSN 1994-7941

science alert

ANSInet
an open access publisher
http://ansinet.com

Trends in Bioinformatics 7 (1): 1-6, 2014
ISSN 1994-7941 / DOI: 10.3923/tb.2014.1.6

Novel Method of Protein Structure Prediction (NPSPM) based on Short Range Interactions between Amino Acids

[1]Arul Mugilan, [1]Sherlyn Jemimah and [2]Preethi Jennifer
[1]Department of Bioinformatics, Karunya University, 641114, Coimbatore, Tamil Nadu, India
[2]Bharathidasan University, Tiruchirappalli, Tamil Nadu, India

Corresponding Author: Arul Mugilan, Department of Bioinformatics, Karunya University, 641114, Coimbatore, Tamil Nadu, India Tel: 91-960-0451716

ABSTRACT

Many methods have been developed to predict the secondary structure of protein sequences. Most methods predict the protein model as α helix-sheet-random coil structure, which is a simplistic view of protein structure. Very few predict other structural elements such as β-turns, 3/10 helix, bends etc. Currently, most approaches rely on neural networks to predict secondary structure. We propose a Novel Protein Structure Prediction Method (NPSPM), which uses DSSP (Dictionary of Protein Secondary Structure) and statistical techniques to generate Secondary Structure Prediction Parameters (SSPP) for single amino acids, all possible amino acid pairs and all possible amino acid triplets. These parameters can be used to predict the secondary structure elements present in the protein sequence. Our method shows better sensitivity and accuracy than other methods.

Key words: Protein structure prediction, secondary structure, deviation parameter, 3/10 helix prediction, β-turn prediction

INTRODUCTION

At present, there are more protein sequences than known protein 3D structures. Only 93788 structures are available in the PDB (Protein Data Bank) (database September 2013), compared to over 41 million sequence entries available in UniProtKB/trEMBL, July 2013). The experimental elucidation of the 3D structure of proteins is a time-consuming process (Schwede *et al*., 2003). Therefore, using *in-silico* protein structure prediction methods will bridge the gap between the amount of structural data and sequence data and accelerate drug discovery and protein function prediction.

Many *in-silico* secondary structure prediction methods have been developed based on neural networks. The earliest system is probably the classifier developed by Qian and Sejnowski (1988). Table 1 provides a list of secondary structure prediction methods for proteins. However, due to the non-availability of sufficient three-dimensional structures, the number of proteins used in estimating the parameters is comparatively small. Also, only a few methods are available for other structural elements such as β-turns, bends, 3-10 helices and random structures.

In our earlier method we had proposed Secondary Structure Prediction methods, SSPDP and ISPBD, for predicting the secondary structure (α-helix and β-sheet) of proteins from the amino acid sequence. β-turn prediction was carried out using data from 12 proteins (Mugilan and Veluraja, 2000; Mugilan, 2008). Combined structure prediction methods can also improve the prediction accuracy (Mugilan *et al*., 2010). Here the results of the prediction method to predict other

Table 1: A few protein secondary structure prediction methods

Method	Description	Citation
NNSSP	"nearest neighbour" approach combined with multiple sequence alignment	Yi and Lander (1993)
SOPMA	An improvement over the SOPM method	Geourjon and Deleage (1995)
PHD	profile-based neural networks) method with multiple sequence alignment input	Rost (1996)
PREDATOR	Predictions based on recognizing possibly hydrogen-bonded residues	Frishman and Argos (1996)
JPRED	Prediction using Jnet algorithm; consensus of DSC, NNSSP, PHD, PREDATOR, ZPRED and MULPRED methods	Cuff *et al.* (1998)
PSIPRED	Prediction based on position specific scoring matrices generated by PSI-BLAST	Jones (1999)
HNN	Hierarchical neural network prediction method based on Qian and Sejnowski's classifier and the NETtalk system	Guermeur *et al.* (1999)
Porter	Uses bidirectional recurrent neural networks and includes long-range information	Pollastri and Mclysaght (2005)
YASPIN	uses a neural network with 7 states (Hb, H, He, C, Eb, E and Ee) and a HMM to optimize the output	Lin *et al.* (2005)
ZPRED	Predicts distance between residue and center of the membrane (Z-coordinate) for membrane proteins using output of HMM and an ANN	Granseth *et al.* (2006)

structural elements (such as β-turns, bends and random coils), in addition to helices and sheets, are presented. The tool can be accessed at http://www.karunya.edu/bioinformatics/npspm.html.

MATERIALS AND METHODS

Input: Training dataset and test dataset: In an earlier publication 408 non-homologous proteins (25% or less homology) were chosen (Mugilan and Veluraja, 2000) from the Brookhaven Protein Data Bank using the PDB-SELECT sub-database (Hobohm *et al.*, 1992). Then, structural deviation parameters were generated for each of the possible combinations of singlets (single amino acids), doublets (pair of amino acids) and triplets (set of three consecutive amino acids in the sequence). Since 20 different amino acids comprise a protein sequence, there can be 20 possible singlets, 400 (i.e., 20^2) possible doublets and 8000 (i.e., 20^3) triplets (Mugilan and Veluraja, 2000).

For NPSPM, 12700 non-homologous protein 3D structures (with less than 25% similarity) from PDB were analyzed. The DSSP outputs of the PDB files were used to find structural information for the amino acids. Out of 12700 proteins, the 3D structures of 12650 proteins were analyzed to relate structural elements like helices, sheets, turns, 3/10 helices and bends to its amino acid composition, using the DSSP output available for the PDB entry. This was done by computing the occurrence of each possible combination of amino acids for each structural element from the training dataset.

Generating the deviation parameters: The secondary structural information for the selected proteins (12650 non-homologous proteins chains) was generated from the well-known software package DSSP of Kabsch and Sander (1983). The DSSP notation is provided in Table 2 below. For the DSSP output, each secondary structure (be it α-helix, β-strand, β-sheet, bend, or 3/10 helix) should have at least 3 consecutive amino acid residues. All other amino acids were considered to be a part of random structures. For random structures, the constraint imposed was that a minimum of three consecutive amino acids should be present in this structure. By imposing these conditions the secondary structures were identified from the DSSP output for all the selected proteins.

Table 2: DSSP notation for secondary structure elements

Secondary structure	DSSP notation
Extended strand (participates in β-ladder)	E
α-helix	H
Hydrogen-bonded turn	T
Bend	S
3/10 helix	G

The frequency of occurrence, denoted by P (X), for the amino acid singlets, doublets and triplets in the selected dataset was computed using the equation:

$$P(X)=\frac{\Sigma_{i=1}^{n} N_i(X)}{\Sigma_{i=1}^{n} Y_i}$$

Here X represents individual amino acid (A) for singlet, two consecutive amino acids (AB) for doublets and three consecutive amino acids (ABC) for triplets in the selected dataset. $N_i(X)$ is the number of counts for X in the ith protein, Y_i is equal to T_i for singlets, T_{i-1} for doublets, T_{i-2} for triplets where T_i is the total number of amino acids in the ith protein and n is the total number of proteins considered (12650 proteins). Based on the frequency of occurrence of amino acid singlets, doublets and triplets, one can work out the expected count for these entities (for each structural element) as:

$$C_{exp}(X)=P(X)\sum S_i$$

Here, S_i represents the occurrence of a particular entity (X) in a particular structural element in the protein i. The computed count (observed count) of singlets, doublets and triplets for the various secondary structural elements (α-helix, β-strand and random structures) are obtained from the output of DSSP by following the constraints mentioned earlier. The structural deviation parameters DP (X) for amino acid singlets, doublets and triplets in a particular secondary structural element are calculated as follows:

$$DP(X)=\frac{C_{comp}(X)-C_{exp}(X)}{C_{exp}(X)}\times 100$$

C_{comp} (X) and C_{exp} (X) are the computed and expected counts, respectively. Using the above formula the structural deviation parameters for amino acid singlets, doublets and triplets were computed for α-helices, β-strands, β-bends, β-turns and 3/10 helices. These parameters were normalized with respect to the α-helix parameters within the group. The normalized parameters can be used for structure prediction from the amino acid sequences. This is illustrated below.

Output: Model prediction: Consider the sequence 'TFRGQ' present in the protein 2B7U. To predict the structure for R we refer to the normalized parameters for the singlet R, doublets (FR and RG) and triplets (TFR, FRG and RGQ) (Table 3). The scores are added using an equation and the final sum for each structure is compared. The equation for calculating the sum is given below:

Table 3: Deviation Parameter Values for Model Calculation

	α-helix	β-sheet	β -turn	3/10 helix	β -bend	Random
R	1.000000	0.024235	0.129618	0.342096	0.164222	0.000000
FR	0.623158	0.831829	0.089402	0.414355	0.000000	0.215189
RG	0.000000	0.004492	0.95187	0.054299	0.324113	0.219431
TFR	0.113375	0.213542	1.000000	0.110096	0.000000	0.182284
FRG	0.30462	0.13577	0.89997	1.00000	0.000000	0.88525
RGQ	0.352101	1.000000	0.532258	0.15393	0.000000	0.304905
	2.393255	2.209872	3.603116	2.074776	0.488335	1.807059

Table 4: Comparison of results for secondary structure prediction

Amino acid	DSSP	NPSPM	SOPMA	YASPIN	HNN	PHD
L	E	E	E	E	E	E
T	E	E	E	E	E	E
F	E	E	R	R	E	R
R	T	T	R	R	R	R
G	T	T	R	R	R	R
Q	E	E	R	R	R	R
V	E	E	R	E	E	E
T	E	E	R	E	E	E
T	E	E	E	E	E	E
V	E	E	E	E	E	E
R	E	E	E	E	E	E
I	E	E	E	E	E	E

$$Sum_h = (R)_{sh} + \frac{1}{2}\left[(FR)_{dh} + (RG)_{dh}\right] + \frac{1}{2}\left[(TFR)_{th} + (FRG)_{th} + (RGQ)_{th}\right]$$

Where:

- Sum_h is the sum of singlet, doublet and triplet parameters for the α-helix
- $(R)_{sh}$ = Singlet normalized parameter
- $(FR)_{dh}$ and $(RG)_{dh}$ = Doublet normalized parameters
- $(TFR)_{th}$ and $(FRG)_{th}$ and $(RGQ)_{th}$ = Triplet normalized parameters

From Table 3, it can be seen that the sum for **β**-turn shows the maximum value (3.603116). Thus R is predicted to be a **β**-turn. The same equation is adapted for other structural elements as well. This is compared with other prediction methods in Table 4.

RESULTS

To test our method, the structures of 50 proteins (not included in the training dataset) were predicted using the normalized parameters. The result for 2B7U is shown in the Fig. 1 in comparison with the results for SOPMA, YASPIN, HNN and PHD. The percentages of α-helical structures were found to be comparable for all methods. **β**-strands were predicted to have a higher proportion by our method. A detailed comparison done between the results of SOPMA and NPSPM for 1W9Z is shown in the Fig. 2 below. All elements were given a slightly higher but comparable predicted value by NSPSM. Detailed results for all 50 proteins are available at the website http://www.karunya.edu/bioinformatics/npspm.html.

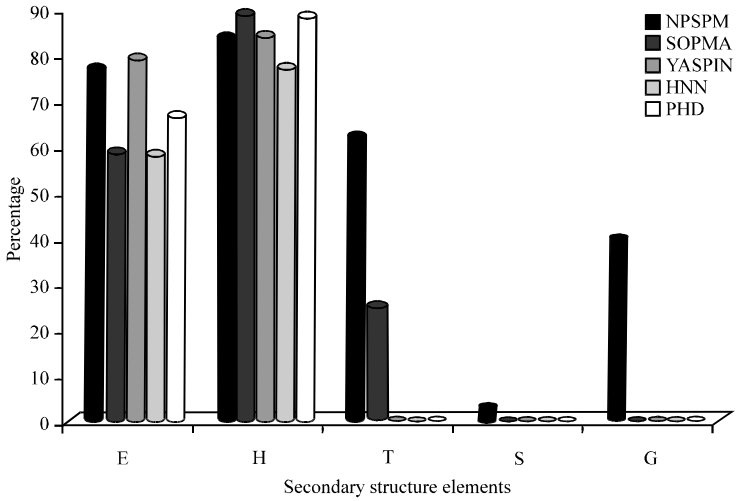

Fig. 1: Comparison of results for 2B7U

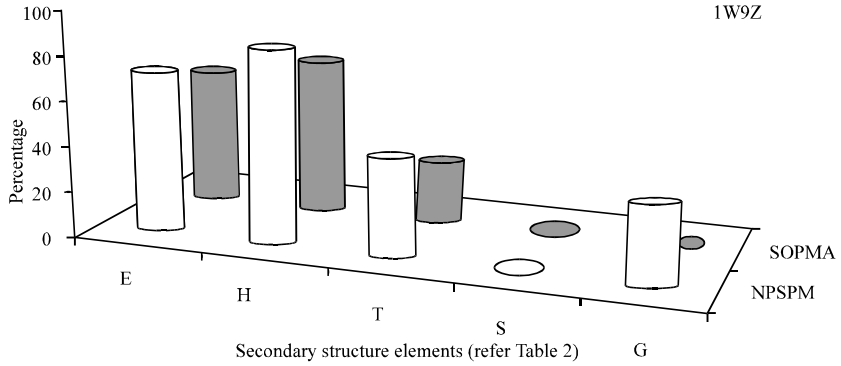

Fig. 2: Comparison of results for 1W9Z between SOPMA and NPSPM

DISCUSSION

The NPSPM method is comparatively better than the earlier methods SSPDP (Mugilan and Veluraja, 2000), KLUSP (Mugilan *et al*., 2010) and ISPBD (Mugilan, 2008) as it predicts **β**-strands and turns in addition to the **α**-helices and **β**-sheets. The previous method only predicts helices and sheets.

CONCLUSION

The average percentage of prediction for 50 protein sequences is compared graphically in Fig. 3. The results are similar for all the methods in case of helices; NPSP and YASPIN show similar values for sheets. Although, SOPMA can predict turns, no other prediction method can account for turns, 3/10 helices or bends. Our method can be used to predict turns, 3/10 helices and bends as well as **α**-helices.

Overall, NPSPM shows better sensitivity comparable to other methods for sheets and helices. NPSPM also shows good results for turns, 3/10 helices and bends, which are accounted for by other methods. We conclude that our method may be used for good structure prediction from sequence data. We hope this method will prove very useful for researchers working on questions in drug design, protein structure prediction, biochemistry and pharmacology.

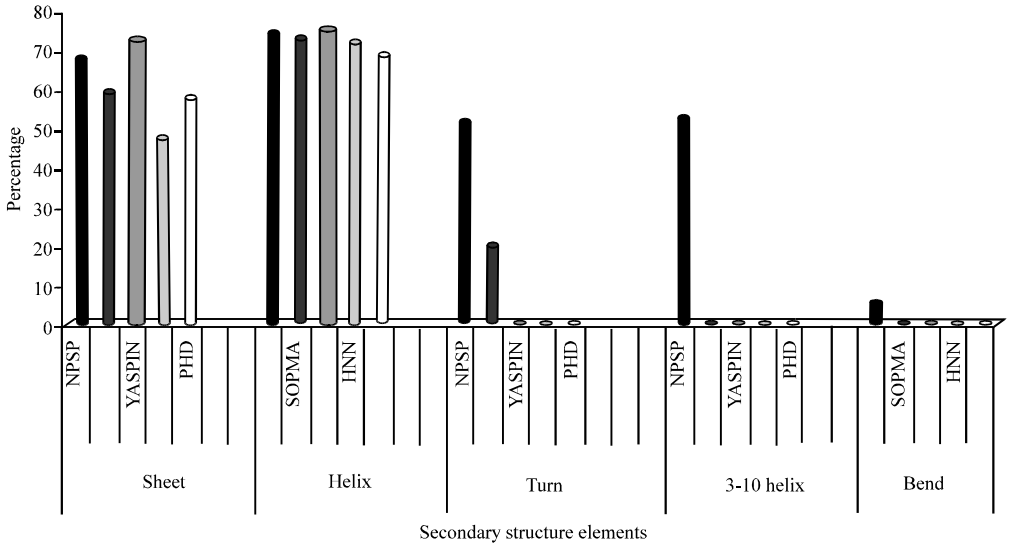

Fig. 3: Average percentage of prediction for 50 protein sequences not included in training set

ACKNOWLEDGMENTS

We thank DST, New Delhi for generously funding this work and Karunya University, Coimbatore for providing the facilities required for the project.

REFERENCES

Hobohm, U., M. Scharf, R. Schneider and C. Sander, 1992. Selection of representative protein data sets. Protein Sci., 1: 409-417.

Kabsch, W. and C. Sander, 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers, 22: 2577-2637.

Mugilan, A., Ajitha, M. Cathrin, M. Kumar, Devi and Thinagar, 2010. *In silico* secondary structure prediction method (Kalasalingam University structure prediction method) using comparative analysis. Trends Bioinform., 3: 11-19.

Mugilan, S.A., 2008. Applications of bioinformatics databases to predict the secondary structure. Proceedings of the International Conference on Advanced Computer Theory and Engineering, December 20-22, 2008, Phuket, pp: 847-851.

Mugilan, S.A. and K. Veluraja, 2000. Generation of deviation parameters for amino acid singlets, doublets and triplets from three-dimentional structures of proteins and its implications for secondary structure prediction from amino acid sequences. J. Biosci., 25: 81-91.

Qian, N. and T.J. Sejnowski, 1988. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol., 202: 865-884.

Schwede, T., J. Kopp, N. Guex and M.C. Peitsch, 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res., 31: 3381-3385.