Trends in Bioinformatics

ISSN 1994-7941

science alert

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

Trends in Bioinformatics 3 (1): 11-19, 2010
ISSN 1994-7941

In silico Secondary Structure Prediction Method (Kalasalingam University Structure Prediction Method) using Comparative Analysis

[1]A. Mugilan, [1]Ajitha, [1]M. Cathrin, [1]M. Kumar, [1]Devi and [2]Thinagar
[1]Department of Biotechnology, Kalasalingam University,
Krishnankoil- 626 190, India
[2]Sandford International School, Addis Ababa, Ethiopia

Abstract: Protein secondary structures mean regular patterns in natural 3D structures such as ALPHA-helix and BETA-strand and protein secondary structure prediction is to estimate them from amino acid sequences. The secondary structure prediction not only becomes the base to infer structural properties from structurally unknown proteins, but also is useful as the constraint to predict 3D structures. The trial to predict protein secondary structures has been started from 1970's and has gradually but steadily advanced until now. Today, average prediction accuracy rate exceeds 80% and over, so it can be said the prediction becomes a reliable and practical method. Here are summarized fundamental approaches to the secondary structure prediction, their recent development and the cautionary notes for their practical use. We had compared 72 proteins of known structure from their relationship between amino acid sequences and secondary structures. In this study, we are going to propose a server implementing a method to improve the accuracy in protein secondary structure prediction. This method is completely based on the prediction result, which is obtained by the online prediction tools to have a combined prediction of higher quality.

Key words: Secondary structure prediction, amino acid sequence, DSSP, prediction accuracy, comparative analysis

INTRODUCTION

Protein secondary structure prediction determines the regions of secondary structure in a protein at the level of α-helix, β-sheet and random coil, from information present in the primary protein sequence. The sequential information of proteins has been increasing many folds than their three-dimensional counterpart. Any vital information obtained by the analysis of the three-dimensional structure in terms of sequence will have definite impact on the structure prediction methods and will have added value in this field of research, due to sequence automation and genome project. Analysis of amino acid doublets, triplets and quadruplets using SWISS-PROT sequence database has implications on the significance of the deviated doublets, triplets and quadruplets in the structural aspect of proteins. Based on the information derived from the known three dimensional structures, methods were developed to predict the secondary structural elements of proteins, such as α-helix, β-strand and random structures (Chou and Fasman, 1978; Wu *et al*., 2009; Bhattacharjee and Biswas,

Corresponding Author: Arul Mugilan, Department of Biotechnology, Kalasalingam University, Krishnankoil-626 190, India

2009; Pal *et al*., 2003). These methods suffered from a lack of data. Prediction (Zheng and Kurgan, 2008; Narang *et al*., 2005) was performed based on amino acid singlet information derived from relatively few known three-dimensional structures. The accuracy of prediction is between 56 to 60% (Kabsch and sander, 1984; Berbalk *et al*., 2009). The problem in these methods has been the inclusion of structures used to derive parameters in the set of structures used to assess the accuracy of the method. Amino acid doublets and triplets information were also used in the early works for secondary structure prediction (Periti *et al*., 1967; Ptitsyn and Finkelstein, 1983; Kabat and Wu, 1974) and the number of proteins used in deriving the parameters were comparatively small due to the non-availability of enough three-dimensional structure. In the triplet parameter generation, (Nagano, 1977; Floudas, 2007) has grouped the 20 amino acids into 7 types leading to a total of 343 parameters. Kabsch and Sander (1983) have developed an algorithm to assign secondary structures for the structure solved proteins based on their X-ray crystal structure coordinates, which is commonly known as DSSP (Dictionary of Secondary Structure Prediction). Single sequence methods and multiple sequence alignment methods are the two eyes for the secondary structure prediction (Jaroszewski , 2009). Recently lot of structure prediction methods are available through internet. From all these methods we are using some famous methods like HNN, SOPMA, ISPBD, SSPRO and PHD.

The main objective is comparative analysis of various databases and prediction of new database with better prediction accuracy based on comprehensive study of several tools available for protein secondary structure prediction.

MATERIALS AND METHODS

This study was conducted at Kalasalingam University, Bioinformatics Laboratory in 2007-2009.

HNN

Hierarchial neural network prediction method can be seen as an improvement on the famous classifier developed by QIAN and SEJNOWSKI derived from the system NET-TALK (Sivan *et al*., 2007). This is mainly made up of two types of network; they are sequence to structural network and structure to structure network. This improvement mainly deals with two points. They are many technical tricks have been used to increase the content on which the prediction is made and concominatantly decrease by two orders of magnitude of the number of parameters. Many physico-chemical data have been explicity incorporated in the predictors used by the structure to structure network.

SOPMA

Self optimized prediction method is based on the homologue method of Levin *et al*. (1993). This method correctly predicts 69.5% of amino acids for a three description of the secondary structure (α-helix, β-sheet and random coil) in a whole database containing 126 chains of non-homologous proteins on combination of SOPMA and PHD methods correctly predicts 82.2% of residues.

ISPBD

In my earlier methods, ISPBD (Innovative Structure Prediction using Bioinformatics Databases), was developed to predict the secondary structure of the proteins from amino acid sequences using the generated structure prediction parameters. In the above

method β-sheet is much better prediction than SSPDP (Mugilan and Veluraja, 2000) PHD (King *et al*., 1997; Rost *et al*., 1994) DSC (Rost and Sander, 1993), NNSSP (Salamov and Solovyev, 1995) and NNPREDICT (Kneller *et al*., 1990) methods.

PHD

The prediction (King *et al*., 1997; Rost *et al*., 1994) is well balanced between alpha-helix, beta-strand and the loop-65% of the observed strand residues are predicted correctly (Rost *et al*., 1994). The accuracy in predicting the content of three secondary structure types is comparable to that of circular dichroism spectroscopy.

SSPRO

The SSPRO belongs to the scratch protein predictor and are shown in many versions. It is a server of protein secondary structure prediction (Pollastri and Mclysaght, 2005). The SSPRO includes the direct incorporation of homologous proteins secondary structure and probabilistic methods to improve the accuracy. Its output is different from that of all the other methods.

METHOD OF CALCULATION

We have retrieved the sequential information from selected 377 non-homologous proteins PDB databases for this analysis. Collected primary sequences of the selected protein were submitted to the various structure prediction servers like HNN, SOPMA, ISPBD, SSPRO and PHD. Secondary structural information of the selected proteins were carried out from the output of the various secondary structure methods using PYTHON programming. In other way secondary structural information were collected from DSSP output for the selected proteins. Prediction Accuracy (PA) for various methods were obtained from the following formula.

$$\text{Prediction accuracy (PA)} = \frac{\text{No. of correctly predicted residues}}{\text{Total No. of residues}}$$

Using the prediction accuracy we find out a new method KLUSP. Prediction Accuracy (PA) calculated chart for the various methods are shown in Fig. 1.

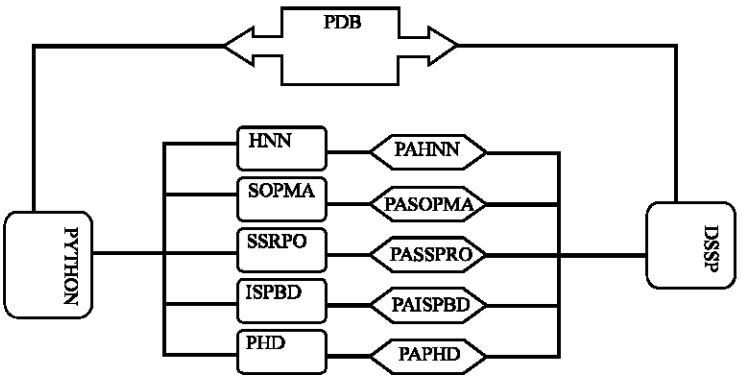

Fig. 1: Flow chart for Prediction accuracy calculation

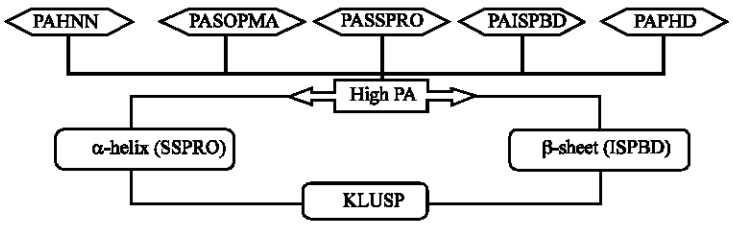

Fig. 2: Flow chart for Prediction accuracy calculation

Flow chart for the *in silico* secondary structure prediction method indicates that, KLUSP was rendered based on comparative analysis (Fig. 2). Chart explains that on comparison of all prediction accuracy the higher prediction accuracy (High PA) for α-helix and β-sheet were found out and combine together to carried out KLUSP.

RESULTS AND DISCUSSION

Sequential information for 377 non-homologous proteins were collected from the PDB databases. The collected sequences were submitted to the various secondary structure methods server page and found out the structural information for the above sequences. Structural information for selected proteins (377) were also collected from the DSSP out put (Fig. 3).

Output of DSSP and various secondary structure prediction methods of model sequence were plotted. Prediction accuracy for various methods were calculated based on DSSP results using the above formula. α -helix and β-sheet Prediction accuracy for selected proteins were calculated using python programming. Prediction accuracy results of 10 proteins were calculated (Table 1).

Average prediction accuracy for α-helix and β-sheet for selected proteins (377) were calculated. α-helix average prediction accuracy for HNN, SOPMA, ISPBD, SSPRO and PHD are 64.12, 78.79, 79.62, 77.59 and 80.39, respectively and β-sheet average prediction accuracy for HNN, SOPMA, ISPBD, SSPRO and PHD are 50.65, 57.95, 78.21, 67.26 and 64.56, respectively.

Among the methods used for our comparison, SSPRO was known to be of highest prediction accuracy for helix and ISPBD has highest Prediction accuracy for sheet.

We found out the *in silco* method (KLUSP) from the output of secondary structure prediction methods SSPRO and ISPBD. In the first step, we have to assign sequence for structure prediction. It was submitted to SSPRO and ISPBD methods. α-helical structural information from SSPRO output and β sheet structural information from ISPBD output were alone taken from among all the α-helix and β-sheet structural elements. Thus, combining these two methods we can generate a new method known as KLUSP. Now by comparing KLUSP with various secondary structure prediction methods and then prediction accuracy was calculated for the above methods. Structural information of the various methods and DSSP output for 1G3P protein were analysed. In the Table 2 sequence number 6-10 favour the α-helical structure in SSPRO and KLUSP methods, but the sequence number 18-20 favour the β-sheet in ISPBD and KLUSP methods. For the above information overall prediction accuracy for KLUSP is much better than ISPBD and SSPRO methods. The result clearly indicate that KLUSP has higher accuracy than the SSPRO and PHD method.

Fig. 3: Comparison of secondary structures

Table 1: Prediction accuracy results of randomly selected 10 proteins

	HNN		SOPMA		ISPBD		PHD		SSPRO	
Protein name	%α	%β	%α	%β	%α	%β	%α	%β	%α	%β
1A2PA	68.1	44	81.8	80	71	60	72.7	48	77.2	60
1AMM	40	46.8	80	62	78	69.6	80	67.1	80	60.1
1ARCA	82.8	46.3	91.4	55.5	85	65	85.7	61.1	97.1	61.1
1AWD	75	41.3	100	58.6	90	91	91.6	89.6	100	68.9
1BKO	68.5	67.2	88.7	59.3	80	83	80.9	50	88.7	71.8
1BPI	100	42.8	100	57.1	96	94	100	92.8	87.5	71.4
1BS9	64.2	50	74.2	60	81	65	82.8	53.3	43.3	64.3
1BXO	36.3	51.4	37.7	59.5	52	83	36.3	80.1	45.4	71.2
1BYQA	91.5	68.7	96.2	72.9	93.2	82.5	94.9	79.2	94.9	72.9
1CEX	72.4	93.1	75.3	75.8	78	89	78.3	86.2	76.8	58.6

HNN: Hierarchial neural network, SOPMA: Seif optimized prediction method, ISPBD: Innovative structure prediction using bioinformatics databases, PHD: Prediction method, SSPRO: Scratch protein predictor

The following proteins are randomly selected and compared with new method KLUSP. Prediction accuracy of proteins 1LIT, 1XNB, 1A7S, 1G3P, 1RHS and 3SEB is shown in Table 3. Prediction accuracy on analyzing these proteins with KLUSP, it is on the whole good. Average prediction accuracy of α-helix for various methods HNN, SOPMA, ISPBD, PHD, SSPRO and KLUSP is 67.18, 85.15, 78.5, 71.42, 88.77 and 88.7. In the result KLUSP is comparable to SSPRO.

Similarly, Table 4 indicates the average prediction accuracy of β-sheet for various methods HNN, SOPMA, ISPBD, PHD, SSPRO and KLUSP is 50.72, 56.23, 73.23, 70.57, 63.25 and 72.9. From the result KLUSP is comparable to ISPBD.

Overall (α-helix + β-Sheet) prediction accuracy of various methods for the above selected proteins are shown in Fig. 4. From overall prediction accuracy of KLUSP for the protein 1LIT and 1XNB has been comparable to SSPRO. But our method has higher prediction accuracy than other methods like HNN, SOPMA, ISPBD, PHD and SSPRO.

Average prediction accuracy(overall) for various methods HNN, SOPMA, ISPBD, PHD, SSPRO and KLUSP is 58.95, 70.69, 75.87, 70.99, 76.01 and 80.80, respectively. The above result clearly indicates that our method (KLUSP) is better than other methods. So predictions made by KLUSP better than a prediction using for instance of all methods.

Table 2: Structural information of the various methods and DSSP output for 1 G3P protein

SEQ	DSSP	HNN	SOPMA	ISPBD	PHD	SSPRO	KLUSP
A			H				
E			H				
T			H				
V	H	H	H			H	H
E	H	H	H			H	H
S	H	H	H			H	H
C	H	H	H	H	E	H	H
L	H	H	H	H	E	H	H
A		H					
K							
S							
H							
T	E						
E	E						
N	E						
S	E		H	E			E
F	E		H	E			E
T	E		H	E			E
N			H				
V			H				
W	E		H		E		
K	E						
G							
D							
E							
T							
Q	E			E	E	E	E
C	E			E	E	E	E
Y	E		E	E	E	E	E
G	E		E	E	E		E
T	E		E	E	E		E
W	E	E	E	E	E	E	E
V	E	E		E	E	E	E
P	E	E		E	E		E
I	E	E	E	E	E	E	E
G	E	E	E	E			E
L	E	E	E				
A		E	E		E	E	E
I					E		E
P					E		E
E							
N							

H: alpha helix, E: beta sheet, SEQ: sequence, Amino acid: A,E,T,V,E,S, HNN: Hierarchial neural network, SOPMA: Seif optimized prediction method, ISPBD: Innovative structure prediction using bioinformatics databases, PHD: Prediction method, SSPRO: Scratch protein predictor

Table 3: Average prediction accuracy for alpha

	Percentage of prediction			Alpha-helix		
Proteins	HNN	SOPMA	ISPBD	PHD	SSPRO	KLUSP
1LIT	67.5	87.5	89	90	91.6	91.6
1XNB	35.7	64.3	60	57	71.4	71.2
1A7S	77.8	88.8	87	87	88.8	88.8
1G3P	100	100	84	46	100	100
1RHS	62.8	82.8	85	86	87.1	87
3SEB	59.3	87.5	66	63	93.7	93.6

HNN: Hierarchial neural network, SOPMA: Seif optimized prediction method, ISPBD: Innovative structure prediction using bioinformatics databases, PHD: Prediction method, SSPRO: Scratch protein predictor

Table 4: Average prediction accuracy for beta

	Percentage of prediction			Beta Sheet		
Protein name	HNN	SOPMA	ISPBD	PHD	SSPRO	KLUSP
1LIT	55	55	69	67.5	57.3	67.2
1XNB	51.3	59.4	76	74.7	58.5	76
1A7S	36.3	66.2	80	75.2	79.2	80
1G3P	40.2	16.8	53.2	51.9	35.1	53.2
1RHS	72.1	79.1	90.2	82.8	88.5	90
3SEB	49.4	60.9	71	71.3	60.9	71

HNN: Hierarchial neural network, SOPMA: Seif optimized prediction method, ISPBD: Innovative structure prediction using bioinformatics databases, PHD: Prediction method, SSPRO: Scratch protein predictor

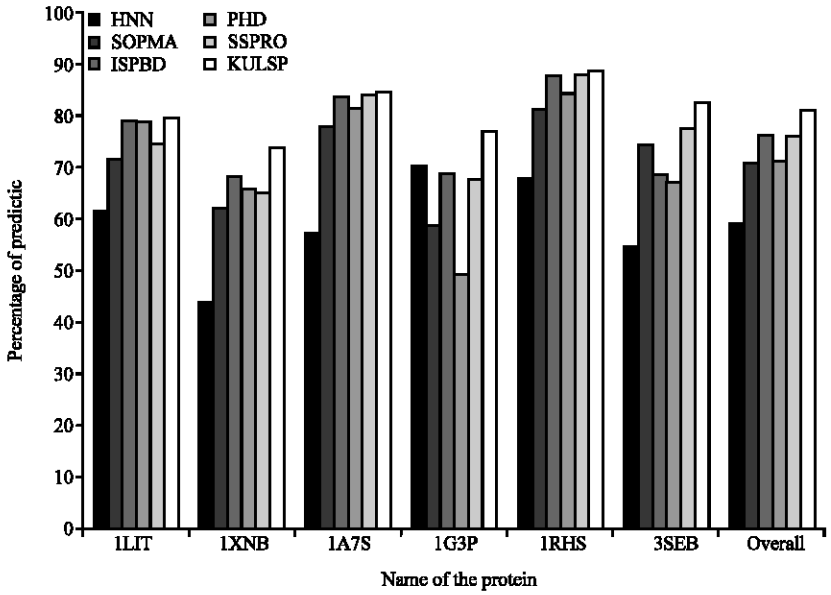

Fig. 4: Overall (α-helix + β-sheet) prediction accuracy of various methods

DISCUSSION

Protein Secondary structure prediction is an important step towards understanding how proteins fold in three dimensions. Our recent analysis of many proteins with the databases HNN, SOPMA, GORIV, PHD, SSPRO is done for keeping the target for improvement in protein secondary structure prediction. These Bioinformatics Databases are libraries of life sciences information, collected from scientific experiments, published literature, high throughput experiment technology and computational analysis. They contain information from research areas including Genomics, Proteomics, Metabolomics, Microarray gene expression, Phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures. Biological database design, development and long-term management is a core area of the discipline of Bioinformatics. So predictions made by KLUSP, better than a prediction using for instance of all methods. Some of the significance of our KLUSP method is overall accuracy of 81% and Beta sheets predict with an accuracy of 84%.

CONCLUSION

Our new method completely revolutionized protein secondary structure predictions, KLUSP, taking it into an area where, it actually is very useful. For instance you will achieve higher accuracy in secondary structure prediction from the modern methods than other methods so far we compared. And the accuracy is so high that it is often the first method used when trying to predict the structure of a protein. In future, by using this higher predicting method, we can improve the design of drug and molecular modeling.

ACKNOWLEDGMENT

The author thanks Bioinformatics, KLU, who provided the computational facility for this study. The author would also like to thank the funding agency DST, Newdelhi.

REFERENCES

Berbalk, C., C.S. Schwaiger and P. Lackner, 2009. Accuracy analysis of multiple structure alignments. Protein Sci., 18: 2027-2035.

Bhattacharjee, N. and P. Biswas, 2009. Structural patterns in alpha helices and beta sheets in globular proteins. Protein Pept. Lett., 16: 953-960.

Chou, P.Y. and G.D. Fasman, 1978. Prediction of the secondary structure of protein from their amino acid sequences. Adv. Enzymol., 62: 45-148.

Floudas, C.A., 2007. Computational methods in protein structure prediction. Biotechnol. Bioeng, 97: 207-213.

Jaroszewski, L., 2009. Protein structure prediction based on sequence similarity. J. Mol. Biol., 569: 129-156.

Kabat, E.A. and T.T. Wu, 1974. Further comparison of predicted and experimentally determined structure of adenylate kinase. J. Mol. Bio., 71: 4217-4220.

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

Kabsch, W. and C. Sander, 1984. On the use of structure: identical pentapeptides can have Completely different conformations. J. Mol. Bio., 81: 1075-1078.

King, R.D., M. Saqi and M.J. Sternberg, 1997. DSC: Public domain protein secondary structure prediction. Comut. Appl. Biosci., 13: 473-474.

Kneller, D.G., F.E. Cohen and R. Langridge, 1990. Improvements in protein secondary Structure Prediction by an enhanced neural network. J. Mol. Biol., 214: 171-182.

Levin, J.M., S. Pascarella, P. Argosand and J. Garnier, 1993. Quantification of secondary structure prediction Improvement Using multiple alignment. Protein Eng., 6: 849-854.

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

Nagano, K., 1977. Triplet information in helix prediction applied to the analysis of super secondary structures. J. Molbio., 109: 251-274.

Narang, P., K. Bhushan, S. Bose and B. Jayaram, 2005. A computational pathway for bracketing native-like structures of small alpha helical globular proteins. Chem. Phys., 7: 2364-2375.

Pal, L., P. Chakrabarti and G. Basu, 2003. Sequence and structure patterns in proteins from an analysis of the shortest helices implications for helix nucleation. J. Mol. Biol., 326: 273-291.

Periti, P.F., G. Quagliarotti and A.M. Liquori, 1967. Recognition of alpha helical segments in proteins of known primary structure. J. Molbio., 24: 313-322.

Pollastri, G. and A. Mclysaght, 2005. A new accurate server for protein secondary structure prediction. Bioinformatics, 21: 1719-1720.

Ptitsyn, O.B. and A.V. Finkelstein, 1983. Theory of protein secondary structure and algorithm of its prediction. J. Bio., 22: 15-25.

Rost, B. and C. Sander, 1993. Prediction of protein secondary structure at better than 70% accuracy. J. Mol. Biol., 232: 584-599.

Rost, B., C. Sander and R. Schneider, 1994. PHD-an automatic server for protein secondary structure prediction. CABIOS, 10: 53-60.

Salamov, A.A. and V.V. Solovyev, 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiply sequence alignment. J. Mol. Biol., 247: 11-15.

Sivan, S., O. Filo and H. Sielmann, 2007. Application of expert networks for predicting protein Secondary structure. Bio. Mol. Eng., 24: 237-243.

Wu, T.Y., C.C. Hsieh, J.J. Hong, C.Y. Chen and Y.S. Tsai, 2009. IRSS a web-based tool for automatic layout and analysis of IRES secondary structure prediction and searching system in silico. BMC Bioinformatics, 10: 160-160.

Zheng, C. and L. Kurgan, 2008. Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinformatics, 9: 430-430.