www.gbtrp.com
International Journal of Biological Technology (2011) 2(2):83-87.
ISSN: 0976 – 4313

Original Article

Consensus Approach for increasing Accuracy of Protein Secondary Structure Prediction

Thushara Antony[1], Sharmila Baburam[2] and *Gnanendra Shanmugam[1]

[1]Department of Bioinformatics, Vivekanandha College of Arts and Sciences for Women, Namakkal, Tamil Nadu, India.
[2]Department of Microbiology, K.R. College of Arts & Science, Kovilpatti, Tamil Nadu, India.

Received: 19.07.2011; Revised: 28.07.2011; Accepted: 13.08.2011; Published: 15.08.2011.

Abstract

Protein structure prediction is one of high importance problem in biomedicine and biotechnology. In past years there has seen consolidation of protein secondary structure prediction have been suggested using different computational methods such as neural networks, machine learning and discriminate analysis. In the present paper, we have proposed a combination of secondary structure prediction method by combining four state of the art secondary structure prediction methods namely *PHD, PREDATOR, HNN and SOPMA* by a simple majority wins method. This simple consensus prediction gives an average *Q3 prediction accuracy* of 71.2%. This is a 0.9% improvement over PHD, which was the best single method reported to date. Further, the *Segment Overlap Accuracy (SOV)* is 72.4% for the consensus method. Presumably, the success of this simple consensus method is mainly due to the use of four best single methods and the noise-filtering properties of a consensus approach, which helps to ignore the training errors of single methods.

Keywords: Secondary Structure Prediction, Consensus prediction, Neural Networks, Machine learning.

Introduction

A long-term goal of the protein-folding problem is to be able to predict the folded three-dimensional structure of a protein from its amino acid sequence alone (Benner, 1989). Secondary structure prediction is often regarded as the initial starting point in predicting the three-dimensional structure of a protein (Boscott *et al.,*1993). Fundamentally, it attempts to classify amino acids in protein sequence according to their predicted local structure, which can be subdivided into three states: a-helix, b-sheets, or loops (Dalal *et al.,*1997) . However, the number of states may vary depending on the algorithm employed as Eight states namely H (α-helix), G (3_{10} - helix), I (π-helix), E (*β*-strand), B (isolated β-bridge), T (turn), S(bend), and - (the rest) (Kabsch and Sander, 1983).

The fundamental assumption on which all secondary structure prediction methods are based on that is there should be a correlation between amino acid sequence and secondary structure (Bystroff *et al*., 2000). Because the entire information for forming secondary structure is contained in the primary sequence any short stretch of amino acid sequence will preferentially adopt one kind of secondary structure over another. A protein secondary structure prediction algorithm assigns to each amino acid a structural state from a three-letter alphabet {H,E,C}(Schmidler et al., 2000). There are two types of algorithms in protein secondary structure prediction. A single-sequence algorithm does not use information about other similar (homologous) proteins. The algorithm should be applicable for a sequence with no sequence similarity to any other protein sequence. Algorithms of another type incorporate additional evolutionary information from multiple alignments or multiple alignment profiles, which are derived from homologous proteins (Rost and Sander, 1993; Frishman and Argos, 1997). Therefore, the prediction accuracy of such an algorithm should be higher than one of a single-sequence algorithm. The accuracy (sensitivity) of the current state-of-the-art single-sequence prediction methods approaches 70% (Aydin *et al.,* 2006). The accuracy of the state-of-the-art prediction methods that employ multiple alignments or alignment profiles is close to 80% (Baldi *et al.,* 1999). The secondary structure prediction performance can be further improved by consensus classifiers, in which different

prediction methods are combined to improve over a single method (Robles *et al.,* 2004 ; Guermeur *et al.,* 2003).

The commonly and widely used algorithms for of protein secondary structure prediction include i) Chou-Fasman and GOR methods (Chou and Fasman, 1974 ; Garnier et al., 1996), ii) Neural network models(Maclin and Shavlik, 1993; Riis and Krogh, 1996) iii) Nearest-neighbor methods (Yi and Lander, 1993). There is plethora of programs utilizing this algorithm and methods. To name a few includes DPM (Deleage and Roux,1987), DSC(King and Sternberg, 1996), GOR IV(Garnier et al., 1996), PHD(Rost,1996), SOPMA (Geourjon and Deleage,1995), PREDATOR (Frishman and Argos,1996) and HNN (Qian and Sejnowski,1988).

The purpose of this study is by using four state of the art secondary structure prediction methods namely PHD, PREDATOR, HNN and SOPMA by a simple majority wins method to correctly identify structure.

Methods

Seven different secondary structure prediction methods were analyzed and each is briefly described here.

Deleage G, and Roux B, DPM (Double Prediction Method) algorithm uses two approaches to produce the final result - first it predicts the protein structural class and then the secondary structure for the sequence.

King RD, and Sternberg MJ., DSC (Discrimination of protein Secondary structure Class) is based on dividing secondary structure prediction into the basic concepts and then use of simple and linear statistical methods to combine the concepts for prediction.

The GOR method, named for the three scientists who developed it - *G*arnier, *O*sguthorpe, and *R*obson - is an information theory-based method (Garnier *et al.,* 1996). The GOR method takes into account not only the probability of each amino acid having a particular secondary structure, but also the conditional probability of the amino acid assuming each structure given that its neighbors assume the same structure.

B Rost, PHD is a 3-level artificial neural network. The different levels consist of a sequence to secondary structure network, with a window of 13 amino acids, a structure to structure network, with a window of 17 amino acids, and finally an arithmetic average over a number of independently trained networks.

SOPMA (Self-Optimized Prediction Method with Alignment) is based on the homologue method of Levin et al. (1986). The improvement takes place in the fact that SOPMA takes into account information from an alignment of sequences belonging to the same family (Geourjon and Deleage, 1995).

PREDATOR (Frishman and Argos, 1996) is a secondary structure prediction method based on recognition of potentially hydrogen-bonded residues in a single amino acid sequence. This method predicts from single or multiple sequences.

The HNN (Hierarchical Neural Network) prediction method can be seen as an improvement on the famous classifier developed by Qian and Sejnowski, As its predecessor, it is made up of two networks: a sequence-to-structure network and a structure-to-structure network. The prediction is only based on local information.

Consensus Prediction Method

The observed Q3 accuracy of DPM, DSP and G2OR was lower than the other methods, so a consensus was calculated only from HNN, PHD, PREDATOR and SOPMA. According to the NPS@ web server's consensus prediction algorithm the standard consensus was calculated by examining the prediction for each method, at each position and taking the most popular state. (for example is a residue had the following predictions HNN,PHD, PREDATOR, for helices and SOPMA for strand, then the consensus prediction would be Helix. If there was no consensus for a particular residue, the result from the PHD method was used.

Accuracy Calculation

Two methods were applied to assess the accuracy of the predictions. Average Q3 and Segment Overlap. Q3 is a measure of the overall percentage of predicted residues, to observe:

$$Q_3 = \sum (i = H, E, C) \frac{predicted_i}{Observed_i} X100$$

Segment overlap calculation64 was performed for each data set. Segment overlap values attempt to capture segment prediction, and vary from an ignorance level of 37% (random protein pairs) to an average 90% level for homologous protein pairs. Segment overlap is calculated by:

$$Sov = \frac{1}{N} \sum_s \frac{minov\ (S_{obs}\ ;\ S_{pred}) + \delta}{maxov\ (S_{obs}\ ;\ S_{pred})} X\ len\ (S_1)$$

Where *N* is the total number of residues, *minov* is the actual overlap, with *maxov* is the extent of the segment. δ is the accepted variation which assures a ratio of 1.0 where there are only minor deviations at the ends of segments.

Results and Discussion

Recent improvements in the prediction accuracy have been accomplished not only by incorporating evolutionary information, but also by combining the results of single, independent secondary structure prediction methods into a consensus prediction. In this respect, the prediction accuracy has been checked and methods that taken into account for multiple alignments are 70% correct for a three-state description of secondary structure. Three cases need to be distinguished when forming the consensus sequence per amino acid according to the three possible secondary states a-helix (H), b-strand (E) and other/ loop (L).

In the literature there are different standards for reducing DSSP 8-state (H,C,B,E,T,S,G,I) assignments to 3 states (H,C,E). It was found that changing the reduction method can alter the apparent prediction accuracy by over 3% on average. Although we were unable to train the methods using different 8 to 3 state reductions, testing all methods with different reduction methods showed that consensus prediction method consistently gave higher accuracy.

We investigated a variety of techniques for combining the prediction methods, in an attempt to raise the average Q3. All possible combinations of methods were tried to calculate the consensus, but no combination of methods improved upon the average Q3 of the consensus of HNN, PHD, PREDATOR and SOPMA.

Table- 1: Comparison of predicted secondary structure results of Seven Different Methods from NPS@ server

Sl.No	PDBID	Length	DPM		DSC		GOR		HNN		PHD		PREDATOR		SOPMA	
			H	S	H	S	H	S	H	S	H	S	H	S	H	S
1	154L	185	22	50	53	21	60	26	65	26	66	29	82	0	74	26
2	1AAZ	87	24	8	25	11	16	24	20	26	29	19	30	16	29	19
3	1ADD	349	145	34	134	35	124	76	134	54	151	46	160	49	178	41
4	1ADE	431	135	51	95	92	148	87	187	61	123	113	143	78	167	80
5	1AHB	246	36	92	80	46	98	44	111	49	92	56	87	55	89	54
6	1ALK	449	156	26	89	90	153	64	127	56	110	79	119	69	127	78
7	1AMP	291	42	72	94	29	78	50	94	38	97	53	103	46	113	38
8	1AOR	605	210	40	166	78	210	76	208	93	184	83	222	78	229	85
9	1AOZ	552	38	167	0	185	69	168	90	152	18	155	33	196	53	157
10	1ASW	161	66	10	62	10	48	41	50	26	50	26	60	27	55	22
11	1ATP	20	6	3	0	8	0	8	3	2	3	1	0	0	6	6
12	1AVH	320	128	28	0	0	192	26	165	49	232	12	224	0	231	9
13	1AYA	101	29	14	14	26	29	24	25	23	26	38	21	22	24	30
14	1BAM	213	92	41	95	15	80	28	103	18	96	28	65	47	100	36
15	1BCX	185	2	65	6	92	0	89	17	62	9	96	7	37	11	69
16	1BDO	80	23	35	14	14	18	22	19	19	0	0	15	19	17	24
17	1BET	107	9	51	0	78	7	56	27	22	3	61	0	62	16	31
18	1BFG	146	46	13	0	54	19	42	31	27	9	48	9	29	22	33
19	1BNC	449	196	41	175	54	184	71	161	102	186	77	152	96	192	72
20	1BOV	69	14	28	0	52	14	26	13	22	9	35	13	31	14	29
21	1BPH	30	7	5	9	0	0	14	7	8	12	3	18	3	11	7
22	1BRS	80	44	18	38	4	47	6	56	3	0	0	44	11	50	10
23	1BSD	80	10	17	26	3	18	12	17	11	18	13	17	12	30	12
24	1CBG	490	140	51	130	47	131	90	201	52	171	68	176	68	172	76
25	1CEI	93	32	9	40	3	39	11	39	10	44	4	36	5	39	6

Table- 2: Comparison of predicted secondary structure results of Four Methods from NPS@ server used for consensus prediction Method

Sl.No	PDBID	Length	HNN		PHD		PREDATOR		SOPMA		consensus	
			H	S	H	S	H	S	H	S	H	S
1	154L	185	65	26	66	29	82	0	74	26	62	16
2	1AAZ	87	20	26	29	19	30	16	29	19	27	13
3	1ADD	349	134	54	151	46	160	49	178	41	146	41
4	1ADE	431	187	61	123	113	143	78	167	80	136	68
5	1AHB	246	111	49	92	56	87	55	89	54	81	52
6	1ALK	449	127	56	110	79	119	69	127	78	111	62
7	1AMP	291	94	38	97	53	103	46	113	38	94	35
8	1AOR	605	208	93	184	83	222	78	229	85	197	68
9	1AOZ	552	90	152	18	155	33	196	53	157	18	156
10	1ASW	161	50	26	50	26	60	27	55	22	55	16
11	1ATP	20	3	2	3	1	0	0	6	6	2	4
12	1AVH	320	165	49	232	12	224	0	231	9	218	9
13	1AYA	101	25	23	26	38	21	22	24	30	23	26
14	1BAM	213	103	18	96	28	65	47	100	36	90	20
15	1BCX	185	17	62	9	96	7	37	11	69	6	70
16	1BDO	80	19	19	0	0	15	19	17	24	16	16
17	1BET	107	27	22	3	61	0	62	16	31	1	57
18	1BFG	146	31	27	9	48	9	29	22	33	12	30
19	1BNC	449	161	102	186	77	152	96	192	72	171	70
20	1BOV	69	13	22	9	35	13	31	14	29	11	31
21	.1BPH	30	7	8	12	3	18	3	11	7	9	5
22	1BRS	80	56	3	0	0	44	11	50	10	45	5
23	1BSD	80	17	11	18	13	17	12	30	12	17	6
24	1CBG	490	201	52	171	68	176	68	172	76	163	55
25	1CEI	93	39	10	44	4	36	5	39	6	39	4

The comparison of Secondary structure predicted from DSC, DPM, GOR, HNN, PHD, PREDATOR and SPOMA, prediction were shown in Table 1. The four prediction methods used for the consensus secondary prediction method was chosen based on calculated Q3 accuracy. The four methods used for consensus prediction were compared with the consensus prediction and found to be that the accuracy of the prediction has increased and the same is tabulated in Table 2. The method with the highest average accuracy of 25 proteins was PHD with 70.3. While the new combination of HNN, PHD, PREDATOR and SPOMA presented here shows an improvement by 0.9% from 70.3 to 71.2% (Table -3).

Table-3: Difference between Q3 and SOV accuracies for each method.

Sl.No.	Method	Accuracy	
		Q3	SOV
1.	PHD	70.3	70.2
2.	HNN	69.5	66.3
3.	PREDATOR	68.6	69.8
4.	SOPMA	68.4	67.3
5.	CONSENSUS	71.2	72.4

The reported simple consensus approach based on the majority voting of solely four prediction methods can be superior to each of the seven single methods as well as to complex combinations of more than three single prediction methods as employed in Jpred. This method is yet is to be proven to work with distinct combinations of different prediction methods on large benchmark sets.

Conclusion

In this study we have proposed a combination of secondary structure prediction method, by combining four secondary structure prediction methods PHD, PREDATOR, HNN and SOPMA by a simple majority wins method. The predicted results of four methods were taken for the consensus secondary structure prediction. Presumably, the success of the method is mainly due to the use of four of the currently best single methods and the noise-filtering properties of a consensus approach, which helps to ignore the training errors of single methods.

References

Benner, S. A. 1989. Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure. *Adv.Enzyme Regu*., 28: 219-236.

Boscott, P. E., Barton, G. J. and Richards, W. G. 1993. Secondary structure prediction for homology modeling. *Prot. Engin*., 6:261-266.

Dalal, S., Balasubramanian, S. and Regan, L. 1997. Protein alchemy: changing α-sheet into β-helix. *Nat. Struct. Biol*., 4: 548-552.

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

Bystroff ,C., Thorsson, V. and Baker, D. 2000. HMMSTR: A hidden markov model for local sequence structure correlations in proteins. *J. Mol. Biol*., 301:173-190.

Schmidler, S. C., Liu, J. S. and Brutlag, D. L. 2000. Bayesian segmentation of protein secondary structure. *J. Comp. Biol*., 7: 233-248.

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

Frishman, D. and Argos, P.1997. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins*, 27: 329-335.

Aydin, Z., Altunbasak, Y. and Borodovsky, M. 2006. Protein secondary structure prediction for a single sequence using hidden semi-Markov models. *BMC Bioinform*., 7: 178.

Baldi, P., Brunak, S., Frasconi, P., Soda, G. and Pollastri, G. 1999. Exploiting the past and the future in protein secondary structure prediction. *Bioinform*., 15: 937–946.

Robles,V., Larrañaga, P., Pena, J., Menasalvas, E., Perez, M. and Herves, V. 2004. Bayesian networks as consensed voting system in the construction of a multi-classifier for protein secondary structure prediction. *Artif. Intell. Med.* (Special Issue in Data Mining in Genomics and Proteomics), 31:117-136.

Guermeur,Y., Pollastri, G., Elisseeff, A., Zelus, D., Paugam-Moisy, H. and Baldi, P. 2003. Combining protein secondary structure prediction models with ensemble methods of optimal complexity. *Neurocomput*., 56: 305-327.

Chou, P. Y. and Fasman, U. D. 1974. Prediction of protein conformation. *Biochem*., 13: 211-215.

Garnier, J., Gibrat, J.F. and Robson, B. 1996. GOR method for predicting protein secondary structure from amino acid sequence, *Meth. Enzymol*., 266: 540-553.

Maclin, R. and Shavlik, J. W. 1993. Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding. *Machine Learning*, 11: 195-215.

Riis, S. K. and Krogh, A. 1996. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. *J. Comp. Biol*., 3: 163-183.

Yi, T.M. and Lander, E. S. 1993. Protein Secondary Structure Prediction Using Nearest-neigbor Methods. *J. Mol. Biol*., 232: 1117-1129.

Deleage, G. and Roux, B. 1987. DPM: An algorithm for protein secondary structure prediction based on class prediction, Protein Eng., 1, 289-294.

King, R.D. and Sternberg, M.J. 1996. DSC: Identification and application of the concepts important for accurate and reliable protein secondary structure prediction, *Protein Sci*., 11: 2298-2310.

Rost, B. 1996. PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods in Enzymology*, 266: 525-539.

Geourjon, C. and Deleage, G. 1995. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci*.,11: 681-684.

Frishman, D. and Argos, P. 1996. PREDATOR: Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence, Protein Eng., 9, 133-142.

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