The Third International Symposium on Optimization and Systems Biology (OSB'09)
Zhangjiajie, China, September 20–22, 2009

Voting for the Prediction of Protein Secondary Structure and Its Evaluation

Ying-Song Dong[3] Zhi-Song He[4] Zi-Liang Qian[5] Yu-Dong Cai[1,2,*]

[1] Institute of System Biology, Shanghai University, 99 ShangDa Road, Shanghai, 200244, China
[2] Department of Combinatorics and Geometry, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, 200031, China
[3] Department of Life Science and Technology, HuaZhong University of Science and Technology,1037 Luoyu Road, Wuhan 430074, China
[4] Department of Bioinformatics, College of Life Sciences, Zhejiang University, HangZhou, ZheJiang, 310058,China
[5] Bioinformatics Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China

Abstract Protein secondary structure prediction is one of the central topics in proteome analysis. Computational methods, developed for the prediction (classification) of protein secondary structures, have been improved substantially since 1990s, allowing us to investigate some of the computational classifiers and attempt to integrate them through voting. The study tries to evaluate whether and how much voting can improve the prediction accuracy.
In the research, 4 classifiers (i.e. predictors), SSpro, PSIpred, PHD and Prof, are selected since they produce some reasonably good prediction accuracies. Two voting methods are adopted to integrate these 4 classifiers – a simple majority voting by assigning data to a class that gains the majority votes, and a weighted majority voting which weights each vote by the prediction accuracy of the classifiers. The voting results show that including better-performed classifiers tends to improve the prediction while including poor-performed classifiers tend to deteriorate the prediction. More investigation could be carried out using more classifiers or more diverse classifiers in a future research.

Keywords protein secondary structure; prediction; classification; voting.

1 Introduction

Protein secondary structure, initially defined by Pauling in 1951 based on the hydrogen bond model [1], is one of the most important properties of a protein. A slight alter of the shapes could change their functions completely or cause them to be malfunctioning, resulting in various diseases. The local structure of a protein is referred to be the secondary structure, allowing people to assign each amino acid to the types of the secondary structures that it is part of. The most common secondary

* Corresponding author. These authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

structures of natural proteins are alpha helices and beta strands. Besides alpha helices and beta strands, six other types of structures are further proposed by DSSP (Define Secondary Structures of Proteins) [2], leading to totally 8 types. These 8 types of structures can be grouped into three larger classes – helix, strand and others. Because the other types of helices and strands, apart from alpha helices and beta strands, are rare, people also commonly classify the structures into alpha helices (H), beta strands (E), and random coil (C) instead of the broader definition. In this study, we treat the two categorizations the same, and refer to both as classifying the protein secondary structures into H, E and C.

The structures of a protein can be probed by X-ray crystallography or NMR spectroscopy. The accuracy of predicting the protein structures, based on the primary structures, is improved gradually [3], since more and more protein structures are accurately determined by experiments, which can be used to study and analyze a protein sequence with unknown structure. Because experimental determination of protein structures is tedious and expensive, and the expansion of the protein sequence database outgrows the expansion of the protein secondary structure database since proteins are sequenced rapidly in recent years, it would be an advantage if the protein secondary structures can be determined accurately by prediction. Many predictors (classifiers) are developed for the applications, among which the SSpro [4], PSIpred [5], PHD [6,7], and Prof [8,9] perform reasonably well with a prediction accuracy higher than 0.7. These predictors will be the focus of this paper.

And there is a famous integration software: Jpred3 (http://www.compbio.dundee.ac.uk/www-jpred/) [10,11] for the protein secondary structure prediction. However, we meet two problems when we use it: (1) The length of sequence is limited 800 residues. (2) Jpred3 is not allowed to download. Therefore, we need to develop a new integrating software to improve the accuracy.

Voting has long been recognized as a useful integration tool to improve the performance of a prediction system. Nearly all investigations find that if a decision gains the majority votes, that decision is more likely to be the right decision. These investigations are found in all kinds of research areas, including pattern recognitions [12-14], character and hand-writing recognitions [15], image analysis [16,17], credit card slip processing [18], speaker identification [19], and promoter prediction [20,21].

This research tries to investigate whether the integration of the above mentioned classifiers (SSpro, PSIpred, PHD, and Prof) through voting could improve the prediction accuracy. Two voting schemes, Simple Majority Voting (SMV) which counts the votes and allocates a data to the class that gains the majority votes, and Weighted Majority Voting (WMV) which weights each vote by the classifier's prediction accuracy, are applied for the prediction. The data obtained are divided into basic dataset, which is used to obtain the prediction accuracy of each individual classifier and for the voting evaluation, and an independent testing dataset, which is used purely for the voting evaluation, which will be described in greater detail in section 2.1. In the rest of the paper, section 2 describes the data, i.e. the protein sequences used in this study, and methods to process the data; section 3 presents the results with some discussion; finally a conclusion of the study is given in section 4.

2 Materials and Methods

2.1 Data Preparation

The data in this research were drawn from http://wks16338.biology.ualberta.ca/seqsee.db.gz， which contains totally 16623 protein sequences. Some sequences are invalid as input for the four classifiers, which are excluded, leaving 15274 remaining sequences used by the study. These sequences can be found in supplemental material 1 and 2. We randomly divide them into two exclusive datasets (a basic dataset and a testing dataset) with a proportion of 4:1 to each other, resulting in 12221 sequences in the basic dataset and 3053 sequences in the testing dataset, respectively. The basic dataset is evaluated by Jackknife cross-validation to obtain the prediction accuracy for each individual classifier. The prediction accuracies are later fed back to weight the classifiers in voting in both the basic and testing dataset. Since the prediction accuracies are gained from the basic dataset and then later fed back to the basic dataset for weighting, the prediction might be biased. However, since the prediction accuracy is rather invariable to a large dataset, the bias is nearly neglectable, giving a good credibility for evaluating the voting using the basic dataset. For scrutiny, a testing dataset is independently used for testing by taking the accuracies obtained from the basic dataset for the voting.

2.2 Brief introduction to each individual predictor

4 predictors, SSpro, PSIpred, PHD, and Prof, are selected for the investigation. They are all reported to have a prediction accuracy higher than 0.7.

2.2.1 SSpro

Version: 4.01
Algorithm: one-dimensional recursive neural network
Datasets: NR database
Features: Using PSI-BLAST by filtering the sequence regions in the NR database, to make them of low complexity, finally, make the position-specific profile.

2.2.2 PSIpred

Version: 2.61
Algorithm: Two feed-forward neural networks
Datasets: NR database
Features: Make the final position-specific scoring matrix (log-odds values) from PSI-BLAST (after three iterations).

2.2.3 PHD

Version: 5.94
Algorithm: Neural networks
Datasets: NR database
Features: Make the multiple sequence alignment profile (by program MaxHom) and the conservation weight.

2.2.4 Prof

Version: 1.0
Algorithm: Neural networks and linear discrimination
Datasets: NR database
Features: Using PSI-BLAST with gap or without gap to search the sequence in the NR database, to make the position-specific profile.

From the above description, we can see that the algorithms and the features selection of the four software are similar. The only difference is the function usage and the parameters selection of PSI-BLAST.

2.3 Accuracy assessment

For the whole sequence, the prediction accuracy A_h is computed as:

$$A_h = \frac{n}{N} \tag{1}$$

where n is the number of residues that are correctly predicted, and N is the number of all the residues in the sequence.

The prediction accuracy A_c for all the sequences is computed as:

$$A_c = \frac{sum(A_h)}{M} \tag{2}$$

where $sum(A_h)$ is the sum of A_h of all sequences, and M is the number of all sequences, i.e. the average prediction accuracy for all sequences.

2.4 Algorithm integration scheme

Simple Majority Voting (SMV) and Weighted Majority Voting (WMV) will be introduced in detail in this section.

2.4.1 Simple Majority Voting (SMV)

Simple Majority Voting (SMV), like its name, is a simple integration scheme. Firstly, an algorithm set $S = \{ f_0, f_1, ..., f_h, ..., f_{N-1} \}$ is defined to contain the predictors used for the voting. Because each amino acid is assigned to have one of the three classes, H, E and C, each predictor will have its vote for each amino acid (the prediction class of an amino acid is the predictor's vote). The class of each amino acid, gaining the majority votes from all the predictors in S, is assigned to be the class of the amino acid. This can be formulated as follows. Let p_i denote the class of an amino acid n predicted by a predictor f_j, and let a counting function $X_s(a)$ be defined as:

$$X_s(a) = \begin{cases} 1 & a = s \\ 0 & a \neq s \end{cases} \quad \text{where } a \text{ and } s \text{ are the classes H, E and C.} \tag{3}$$

The count of total votes for class s can then be defined as

$$C_s = \sum_{i=0}^{N-1} X_s(p_i) \tag{4}$$

The predicted class for amino acid n using the algorithm set S is defined to be the class that gains the majority votes as

$$S(n) = \arg\max_{s \in \{H,E,C\}} C_s \tag{5}$$

If two or more classes gain the same vote, one of them is chosen arbitrarily. This research not only uses all the 4 predictors for the voting, but also uses all the subsets of the 4 predictors containing 2 or more than 2 predictors for the voting.

2.4.2 Weighted Majority Voting (WMV)

WMV (Weighted Majority Voting) is similar to SMV except that the predictor is weighted by the prediction accuracy A_c rather than weighted equally. The vote counting function in Eq (4) becomes

$$C_s = \sum_{i=0}^{N-1} X_s(p_i) \times A_c(f_i) \tag{6}$$

i.e. each vote is weighted by the A_c value. Other computation is performed in exactly the same way as the SMV.

3 Results and Discussion

3.1 Prediction accuracies of the 4 individual predictors

The basic set (12221 sequences) and the testing set (3053 sequences) were input into the four protein secondary structure predictors. The A_c values (defined in section 2.3) were computed as the prediction accuracies. For the reason of comparison, the A_c values of the testing set are also computed though they are not used to weight the predictors in WMV. The prediction results are shown in table 1.

Table 1: Prediction results of the four software

predictors	Prediction accuracies	
	Training set	Test set
SSpro	0.790464	0.79381
PSIpred	0.786289	0.789353
PHD	0.738047	0.740915
Prof	0.736249	0.736458

3.2 Results of SMV and WMV

For each of the two voting schemes, the prediction accuracies of all the possible combinations of the four predictors were calculated. Table 2 shows the prediction results for both SMV and WMV. If the voting improves the prediction accuracies in both the basic and testing set, the results are highlighted in blue color; in red color if

the prediction deteriorates, by comparing the prediction accuracies to those produced by the best-performed individual predictor involved in the voting.

Table 2: Prediction results of integration schemes

Software		Prediction accuracies	
		Basic set	Testing set
SMV	SSpro, PSIpred	0.789135	0.789201
	SSpro, PHD	0.764889	0.771089
	SSpro, Prof	0.752337	0.773998
	PSIpred, PHD	0.741337	0.74224
	PSIpred, Prof	0.767975	0.75431
	PHD, Prof	0.736548	0.740908
	SSpro, PSIpred, PHD	0.794713	0.797887
	SSpro, PSIpred, Prof	***0.796424***	***0.799435***
	SSpro, PHD, Prof	0.783552	0.785426
	PSIpred, PHD, Prof	0.779862	0.78197
	SSpro, PSIpred, PHD, Prof	0.79424	0.791228
WMV	SSpro, PSIpred	0.790464	0.79381
	SSpro, PHD	0.790464	0.79381
	SSpro, Prof	0.790464	0.79381
	PSIpred, PHD	0.786289	0.789353
	PSIpred, Prof	0.786289	0.789353
	PHD, Prof	0.738047	0.740915
	SSpro, PSIpred, PHD	0.794796	0.797914
	SSpro, PSIpred, Prof	0.796437	0.799487
	SSpro, PHD, Prof	0.785589	0.787649
	PSIpred, PHD, Prof	0.781795	0.78392
	SSpro, PSIpred, PHD, Prof	***0.797276***	***0.800541***

3.3 Discussion

The results show that not all integrations improve the prediction, while, in fact 5 integrations improve (highlighted in blue color) and 4 integrations deteriorate (highlighted in red color) the prediction in both the basic and the testing dataset. Predictors PHD and Prof are the poorer-performed predictors in the integration, and the prediction accuracy is more likely to deteriorate if both of them are involved in the voting. On the contrary, the prediction accuracy tends to improve if both SSpro and PSIpred, the best-performed predictors, are involved in the voting. Though some integrations improve the prediction accuracy, the improvement is very little, indicating that a simple voting does not help significantly in the prediction and, in order to gain a significantly better prediction, improvement of each individual predictor and/or adopting a more dedicate integration scheme, would be more viable. All the predictors adopted in this research apply artificial neural networks as their prediction algorithm. Using predictors with different prediction algorithms may further improve the voting, which could be considered as a future research.

In all the above applications, WMV performs better than or levels the SMV under the same conditions. When only two predictors are integrated through voting, the decision is actually made purely by the better-performed predictor in WMV, showing

better results than the SMV. The integration between two predictors must be performed in a lower level, e.g. taking the data distribution into consideration, in order to improve the effectiveness of the integration between two predictors. The results obtained by the basic dataset and the testing dataset are nearly completely consistent with each other except that when integrating all 4 predictors in SMV, the prediction accuracy is improved in the basic dataset while deteriorated in the testing dataset.

4 Conclusion

We introduce SMV (Simple Majority Voting) and WMV (Weighted Majority Voting) for the integration of predictors for the prediction of protein secondary structures. We show that the integration of better-performed predictors tends to improve the prediction, and including the poorer-performed predictors tends to reduce the prediction accuracies. WMV performs better than or levels the SMV in all the integrations of the predictors. Integration of two predictors or poorer-performed predictors will most likely reduce the prediction accuracies, the problem of which could only be solved by applying different integration schemes e.g. taking the data distribution into consideration instead of simply counting the votes. Only 4 predictors are adopted in this research, and all of them use artificial neural networks as their prediction algorithm. A future research could consider more predictors and/or more diverse predictors for the voting. Different integration schemes may also be devised for a further improvement of the prediction/classification performance.

Acknowledgements

We would like to thank the authors of the 4 predictors SSpro, PSIpred, PHD, and Prof.

References

[1] C Branden and J Tooze (1999). Introduction to Protein Structure 2nd ed. Garland Publishing: New York, NY.

[2] Kabsch W, Sander C Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983 Dec;22(12):2577-637

[3] Burkhard Rost. Journal of Structural Biology 134**,** 204–218 (2001). Review: Protein Secondary Structure Prediction Continues to Rise.

[4] Cheng J, Randall AZ, Sweredoski MJ, Baldi P. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W72-6. SCRATCH: a protein structure and structural feature prediction server.

[5] McGuffin LJ, Bryson K, Jones, D.T. (2000). Bioinformatics. 16, 404-405. The PSIPRED protein structure prediction server.

[6] Rost B, Yachdav G, Liu J. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W321-6. The PredictProtein server.

[7] Rost B. Methods Enzymol. 1996;266:525-39. PHD: predicting one-dimensional protein structure by profile-based neural networks.

[8] Ouali M, King RD. Protein Sci. 2000 Jun;9(6):1162-76. Cascaded multiple classifiers for secondary structure prediction.

[9] Andreas Karwath and Ross D King. BMC Bioinformatics 2002, 3:11. Homology Induction: the use of machine learning to improve sequence similarity searches.
[10] Cuff J. A., and Barton G. J. (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. JPred combines various prediction methods.
[11] Cole, C., J.D. Barber, and G.J. Barton, The Jpred 3 secondary structure prediction server. NUCLEIC ACIDS RESEARCH, 2008. 36: p. 197-201.
[12] Lam L. and Suen C. Y. Application of majority voting to pattern recognition: An analysis of its behavior and performance. IEEE Trans. Pattern Analysis and Machine Intelligence, 27(5):553 – 568,1997.
[13] Lam L. and Suen C. Y. A theoretical-analysis of the application of majority voting to pattern-recognition. In Proc. 12th IAPR Int. Conf. on Pattern Recognition, Conf. B: Pattern Recognition and Neural Networks, 2: 418 – 420, Jerusalem,Israel,1994.
[14] Rahman A. F. R. and Fairhurst M. C. Exploiting second order information to design a novel multiple expert decision combination platform for pattern classification. Electronics Letters,33(6):476 – 477,1997.
[15] Burkhard Rost1* and Volker A. Eyrich1. PROTEINS: Structure, Function, and Genetics Suppl 5:192–199 (2001). EVA: Large-Scale Analysis of Secondary Structure Prediction.
[16] Ho T. K., Hull J. J. and Srihari S. N. Combination of Decisions by Multiple Classifiers in Structured Document Image Analysis, pages 188 – 202. S-V,1992. H. S. Baird,H. Bunke and K. Yamamoto(Eds.).
[17] Rohlfing T., Russakoff D. B., and Maurer C. R. Performance-Based Classifier Combination in Atlas-Based Image Segmentation Using Expectation-Maximization Parameter Estimation. IEEE Transactions on Medical Imaging, 23(8): 983-994, 2004
[18] Paik J., Jung S. and Lee Y. Multiple combined recognition system for automatic processing of credit card slip applications. In Proceedings of the Second International Conference on Document Analysis and Recognition (Cat. No.93TH0578-5), pages 520 – 523,1993.
[19] Altincay H. and Demirekler M., “An information theoretic framework for weight estimation in the combination of probabilistic classifiers for speaker identification,” Speech Commun., 30(4): 255–272, 2000.
[20] Liu R and States D. J. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. Genome Res. 2002 Mar;12(3):462-469.
[21] Hong-Hee Won, Min-Ji Kim (2008) EnsemPro: An ensemble approach to predicting transcription start sites in human genomic DNA sequences, Genomics 91:259 – 266.