 JQC, vol.1, no.1, pp.21-28, 2019

# Protein Secondary Structure Prediction with Dynamic Self-Adaptation Combination Strategy Based on Entropy

**Yuehan Du[1, 2], Ruoyu Zhang[1], Xu Zhang[1], Antai Ouyang[3], Xiaodong Zhang[4], Jinyong Cheng[1] and Wenpeng Lu[1, *]**

**Abstract:** The algorithm based on combination learning usually is superior to a single classification algorithm on the task of protein secondary structure prediction. However, the assignment of the weight of the base classifier usually lacks decision-making evidence. In this paper, we propose a protein secondary structure prediction method with dynamic self-adaptation combination strategy based on entropy, where the weights are assigned according to the entropy of posterior probabilities outputted by base classifiers. The higher entropy value means a lower weight for the base classifier. The final structure prediction is decided by the weighted combination of posterior probabilities. Extensive experiments on CB513 dataset demonstrates that the proposed method outperforms the existing methods, which can effectively improve the prediction performance.

**Keywords:** Multi-classifier combination, entropy, protein secondary structure prediction, dynamic self-adaptation.

## 1 Introduction

Protein secondary structure is the link between protein primary and tertiary structure. If the accuracy of protein secondary structure prediction reaches 0.8, the three-dimensional spatial structure of a protein molecule will be predicted accurately [Zhang, Tang, Zhang et al. (2003)]. Therefore, for a long time, protein secondary structure prediction has been an important method to study protein structure and function. Because it is a time-consuming work to determine protein structure by physical and chemical experiments, machine learning methods for determining protein structure become popular and are favored by researchers.

At present, the prediction of protein secondary structure mainly focuses on the following two aspects [Tang, Li, Zhang et al. (2013)]. One aspect is how to obtain the information of protein structure features effectively. There are a lot of physical and chemical

[1] School of Computer, Qilu University of Technology (Shandong Academy of Sciences), Jinan, 250353, China.

[2] Shandong Mental Health Center, Jinan, 250014, China.

[3] School of Computer Science and Engineering, Shandong University of Science and Technology, Qingdao, 266590, China.

[4] Jinan Intellectual Property Information Center, Jinan, 250099, China.

[*] Corresponding Author: Wenpeng Lu. Email: Wenpeng.Lu@qlu.edu.cn.

information, sequence information and other relevant information in proteins. Therefore, it is difficult to determine the correlation between a feature and structure. Moreover, if there are too many comprehensive feature information, redundant information will increase, leading to high dimension disaster, which will inevitably injure the prediction accuracy. The other aspect is how to select prediction algorithms and apply them to build pattern recognition classifier. These works usually adopts single classifier algorithm, lacking generalization ability.

Aiming at the problems in existing methods, a protein secondary structure prediction with dynamic self-adaptation combination strategy based on entropy is proposed, which comprehensively considers the performance differences of the classifiers and the uncertainty of samples. The method introduces the weight parameter of overall performance of a classifier based on entropy and the weight parameter of self-confidence of a classifier on a sample, which are utilized to improve weighted voting method with instance dynamic self-adaptation [Lu, Wu, Jian et al. (2018)]. Extensive experiments on CB513 dataset demonstrates the superiority of the proposed method over the existing combination methods, which can effectively improve the accuracy of protein secondary structure prediction.

The structure of this paper is as follows. We introduce the related work of some common methods for protein secondary structure prediction of multi-classifier combination in Section 2. Section 3 describes the implementation of protein secondary structure prediction method with dynamic self-adaptation combination strategy based on entropy. Empirical results are provided in Section 4. We conclude the work in Section 5.

## 2 Related work

In recent years, multi-classifier combination methods become popular in the field of machine learning [Bouziane, Messabih and Chouarfia (2015); Zheng and Li (2013); Ma, Liu, and Cheng (2018); Shi (2018); Yang, Tan and Zhang (2018); Yang, Chen, Chen et al. (2018); Xia, Yuan, Lv et al. (2018)], which has been applied on the prediction of protein secondary structure and has attracted more and more attention from researchers. The typical work on multi-classifier combination field can be roughly divided into two categories.

One category is homogeneous combination, which utilizes the same kind of base classifiers with different parameters to classify instances many times and combines the results, whose representative works include Bagging algorithm, Boosting or AdaBoosting algorithm. Zheng et al. proposed a Ma-Ada multi-classifier combination algorithm, which used SVM as a base classifier to conduct experiments on four datasets and achieved a better performance [Zheng and Li (2013)]. As the base classifier selected by this method is the same kind of classifier, it generally does not have strong generalization ability.

Another category is heterogeneous combination, which utilizes different kinds of base classifiers to build ensemble classifier, such as, probability-based methods, voting-based methods, result weighted voting and probability weighted voting. These methods utilize different classifier as base classifiers. How to precisely assign a suitable weight for each classifier is a key problem. Hafida et al. selected the BP neural network and support vector machine (SVM) as base classifier, and combined their results with nine methods, e.g., product, sum rules, which is experimented on RS126 and CB513 dataset [Bouziane,

Messabih and Chouarfia (2015)]. Homayouni et al. proposed a density-based learning framework to build a multi-classifier combination model to predict protein secondary structure [Homayouni and Mansoori (2017)]. Ma et al. proposed a protein secondary structure prediction method based on data segmentation and semi-random subspace, which trained base classifiers on the subspace data generated by the semi-random subspace method, and combined base classifiers by majority vote rule into ensemble classifiers on each subset. Multiple classifiers were trained on different subsets [Ma, Liu and Cheng (2018)]. These different classifiers were used to predict the secondary structures of different proteins according to the protein sequence length.

Lots of practical applications, experimental and theoretical achievements of some specific cases show that multi-classifier combination method is successful, which can achieve a better performance than a single classifier. For protein secondary structure prediction problem, multi-classifier combination methods also demonstrate its powerful ability.

## 3 Protein secondary structure prediction with dynamic self-adaptation combination strategy based on entropy

### *3.1 Multi-classifier combination framework*

In multi-classifier combination, each base classifier is regarded as an expert in the entire feature space, which outputs its judgment on each instance. With combination strategies or rules, the outputs of each base classifier are integrated. The general framework of multi-classifier combination is shown in Fig. 1. First, each base classifier classifies the instances and outputs posterior probability information, i.e., the Decision Profile (DP) [Kuncheva, Bezdek and Duin (2001)]. Then, with the combination strategy, the DP matrix is handled to combine the outputs of all base classifiers. Last, the label with maximum combined probability is returned as the final classification result.

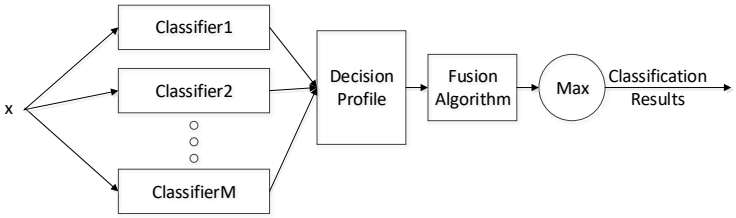


**Figure 1:** Framework of multi-classifier combination

### *3.2 Base classifier*

The base classifiers should have a high accuracy and be differentiated each other, so that they can generate complementary information. We investigate random forest classifier, RBF classifier and multi-classification SVM classifier. These methods are representative, whose principles are highly complementary. We respectively utilize these models to predict protein secondary structure. In the combination experiments, random forest classifier, RBF classifier and multi-classification SVM classifier are selected as the base classifiers of the combination algorithm.

### *3.3 Prediction with dynamic self-adaptation combination strategy based on entropy*

In this paper, a prediction with dynamic self-adaptive strategy based on entropy is proposed, which introduces two weight coefficients: (1) self-confidence of base classifier, (2) information entropy of instances uncertainty [Xia and Xu (2012)]. The model is described as:

$$\hat{s} = \underset{i=1,\ldots,c}{\operatorname{argmax}} \sum_{j=1}^{M} (\omega_j \times \beta_j \times d_{j,i}) \tag{1}$$

In Eq. (1), $\mathrm{d}_{\mathrm{j,i}}$ refers to the posterior probability of the i-th structure given by the j-th classifier, $\beta_{\mathrm{j}}$ is a weighted coefficient based on self-confidence of member classifier, which is decided with Eq. (2) and Eq. (3).

$$\beta_j = \begin{cases} 0.95 & if\ \max\ (d_{j,i}) \geq \theta_j, \\ 1-0.95 & otherwise, \end{cases} \tag{2}$$

$$\theta_j = \frac{1}{\mathrm{c}} \sum_{i}^{c} d_{j,i}, \tag{3}$$

where $\theta_{\mathrm{j}}$ is the average value of the posterior probabilities on all possible structures given by the j-th classifier. If the j-th base classifier outputs a posterior probability, which is greater than or equal to the average value $\theta_{\mathrm{i}}$, it means that the classifier is self confident to make a right judgment. Therefore, In Eq. (2), we assign a higher weight to the classifier, i.e., 0.95. Otherwise, a lower weight is assigned, i.e., 0.05.

In order to avoid only considering the individual differences of the instances in the combination strategy, we further introduce the uncertainty measurement of the j-th classifier, i.e., information entropy, as shown in Eq. (4).

$$H_j(x) = -\sum_{j=1}^{M} d_{j,i} \log_2 d_{j,i}, \tag{4}$$

In other words, $H_j(x)$ is the uncertainty of the base classifier $f_j$. If the value of $H_j(x)$ is larger, it means that the classifier is more uncertain on the classification of the instance, indicating that the classifier has a worse classification ability on it, and the combination weight $\omega_j$ of the classifier for the instance $x$ is smaller, as shown in Eq. (5).

$$\omega_j = \frac{\exp(-H_j(x))}{\sum_{k=1}^{M} \exp(-H_k(x))}, \tag{5}$$

The proposed prediction with dynamic self-adaptation combination strategy based on

entropy, which considers both self-confidence of base classifier and information entropy of uncertainty. The comprehensive consideration provides the potential to achieve a better combination performance on protein secondary structure prediction.

## 4 Experiments

### *4.1 Dataset*

In order to verify the performance of our proposed combination strategy, the popular CB513 is chosen as the benchmark dataset, which is a widely used low-homology dataset [Cuff and Barton (1999)]. CB513 dataset contains 513 non-homologous protein sequences, whose sequence similarity is less than 0.25.

### *4.2 Classification criteria of secondary structures*

Protein secondary structures are usually divided into eight categories: G (310-helix), H( α -helix), I ( π -helix), B (isolated β -bridge), E ( β -stand), S (bend), T (hydrogen bonded turn) and the rest (apparently random conformations). The mainstream ideology of protein secondary structure prediction usually map the eight labels into three ones, i.e., H, E and C. In this paper, DSSP method is adopted, and eight structures are clearly classified into three ones, with the principle: H and G are Helices, denoted as H; E and B belong to Sheets, denoted as E; G, S, T, C and I belong to Coils, denoted as C.

### *4.3 Preprocessing of base classifier output*

The posterior probability outputted by the base classifier need to be preprocessed before the combination. The original value with large difference should be normalized. We use mapminmax function in MATLAB to normalize the posterior probability outputted by the base classifier.

### *4.4 Evaluation measures*

There are many evaluation measures for the prediction of protein secondary structure. Currently, the following measures are used.

#### *4.4.1 Overall prediction accuracy Q3*

At present, the most widely used accuracy rate refers to the total percentage of three secondary structures (residues) which be correctly predicted, which can be calculated from Eq. (6):

$$Q_3 = \frac{P_H + P_E + P_C}{N_H + N_E + N_C} \times 100\,\%, \tag{6}$$

where NH, NE and NC respectively represent the total number of residues whose secondary structure is H, E and C in the sequence, and PH, PE and PC respectively represent the number of residues which is correctly predicted as H, E and C structures.

*4.4.2 Three-state prediction accuracy Qi*

We use $Q_i$ to represent the prediction accuracy rate of each secondary structure which is correctly predicted as H, E or C structure. It can be calculated with Eq. (7):

$$Q_i = \frac{P_i}{N_i} \times 100\,\%, i \in \{H, E, C\} \quad , \tag{7}$$

where $P_i$ is the residue of structure *i* that is correctly predicted in the sequence, $N_i$ is the residue of structure *i* in the sequence. Structure *i* may be structure H, structure E or structure C.

*4.4.3 Matthews correlation coefficient MCC*

We use it to measure the quality of classifier classification, as shown in Eq. (8).

$$M_{cc_i} = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i)(TP_i + FN_i)(TN_i + FP_i)(TN_i + FN_i)}}, i \in \{H, E, C\}, \tag{8}$$

where $TP_i$ represents the residue base of the structure *i* that is correctly predicted. $TN_i$ represents the residue base of the $\bar{i}$ structure (not *i*) that is correctly predicted. $FP_i$ represents the residue base of the structure *i* that is actually $\bar{i}$, but is predicted as *i*. $FN_i$ represents the residue base of the structure *i* that is actually *i*, but is predicted as $\bar{i}$. *i* may be H structure, E structure or C structure.

***4.5 Result and analysis***

In the experiment, PSSM matrix of CB513 dataset is calculated by PSI-BLAST program. In order to ensure the reliability of the experimental results, the dataset uses seven-fold cross validation in the base classifier experiment stage. The performances of single classifier and our models are compared in Tab. 1.

**Table 1:** Experimental Results on CB513

| Methods | $Q_H$(%) | $Q_E$(%) | $Q_C$(%) | $Q_3$(%) | $Mcc_H$ | $Mcc_E$ | $Mcc_C$ |
|---|---|---|---|---|---|---|---|
| M-SVM$_{CS}$ | 79.01 | 62.74 | 79.20 | 75.50 | 0.675 | 0.576 | 0.562 |
| M-SVM$_{LLW}$ | 77.95 | 52.46 | 83.67 | 74.62 | 0.657 | 0.548 | 0.553 |
| M-SVM$_{WW}$ | 77.69 | 62.73 | 80.62 | 75.54 | 0.675 | 0.576 | 0.558 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| M-SVM$_{MSVM2}$ | 78.74 | 60.60 | 80.50 | 75.39 | 0.668 | 0.574 | 0.559 |
| RBFNN | 78.38 | 58.25 | 80.74 | 74.83 | 0.656 | 0.560 | 0.553 |
| RF | 71.35 | 49.93 | 81.62 | 71.32 | 0.594 | 0.541 | 0.516 |
| Our model | 78.79 | 59.83 | 81.79 | 75.76 | 0.675 | 0.578 | 0.576 |

The overall prediction accuracy (Q3) range obtained by each base classifier on the CB513 dataset is 71.32%-75.50%. Among them, Random Forest classifier (RF) is the worst, 71.32%, while M-SVMCS classifier is the best, 75.50%. Compared on the prediction accuracy of H structure, E structure and C structure, all the base classifiers output better prediction results for C structure, with the prediction accuracy range of 79.20%-81.62%. While the prediction accuracy of E structure was relatively worse, with the range of 49.93%-62.74%.

As shown in Tab. 1, our proposed protein secondary structure prediction with dynamic self-adaptation combination strategy based on entropy achieves the best performance, i.e., 75.76%. Besides, the results obtained by calculating the Matthews correlation coefficient $Mcc_H$, $Mcc_E$ and $Mcc_C$ are better than those obtained by the base classifier. This demonstrates that our models have a better classification quality.

## 5 Conclusion

In this paper, we propose a protein secondary structure prediction with dynamic self-adaptation combination strategy based on entropy, which assigns combination weights according to the entropy of posterior probabilities outputted by base classifiers. Extensive experiments on CB513 dataset demonstrates that the proposed method can effectively improve the prediction performance. Our future work is to verify the performance on more dataset and try to apply the combination strategy on other applications.

**Acknowledgement:** The research work is supported by the National Nature Science Foundation of China under Grant No. 61375013 and No. 61502259, and Taishan Scholar Program of Shandong Province in China (Directed by Prof. Yinglong Wang).

## References

**Bouziane, H.; Messabih, B.; Chouarfia, A.** (2015): Effect of simple ensemble methods on protein secondary structure prediction. *Soft Computing*, vol. 19, no. 6, pp. 1663-1678.

**Cuff, J. A.; Barton, G. J.** (1999): Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, vol. 34, no. 4, pp. 508-519.

**Homayouni, H.; Mansoori, E. G.** (2017): A novel density-based ensemble learning

algorithm with application to protein structural classification. *Intelligent Data Analysis*, vol. 21, no. 1, pp. 167-179.

**Kuncheva, L. I.; Bezdek, J. C.; Duin, R. P.** (2001): Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, vol. 34, no. 2, pp. 299-314.

**Lu, W.; Wu, H.; Jian, P.; Huang, Y.; Huang, H.** (2018): An empirical study of classifier combination based word sense disambiguation. *IEICE Transactions on Information and Systems*, vol. 101, no. 1, pp. 225-233.

**Ma, Y.; Liu, Y.; Cheng, J.** (2018): Protein secondary structure prediction based on data partition and semi-random subspace method. *Scientific Reports*, vol. 8, no. 1, pp. 9856.

**Shi, C.** (2018): A novel ensemble learning algorithm based on ds evidence theory for IOT security. *Computers, Materials & Continua*, vol. 57, no. 3, pp. 635-652.

**Tang, Y.; Li, C.; Zhang, Y.; Shang, J.; Zou, L.; Li, L.** (2013): Advanced studies on protein secondary structure prediction. *Progress in Modern Biomedicine*, vol. 13, no. 26, pp. 5180-5182.

**Xia, M.; Xu, Z.** (2012): Entropy/cross entropy-based group decision making under intuitionistic fuzzy environment. *Information Fusion*, vol. 13, no. 1, pp. 31-47.

**Xia, Z.; Yuan, C.; Lv, R.; Sun, X.; Xiong, N. N. et al.** (2018): A novel weber local binary descriptor for fingerprint liveness detection. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

**Yang, K.; Tan, T.; Zhang, W.** (2018): An evidence combination method based on dbscan clustering. *Computers, Materials & Continua*, vol. 57, no. 2, pp. 269-281.

**Yang, Y.; Chen, Y.; Chen, Y.; Bi, W.** (2018): A novel universal steganalysis algorithm based on the IQM and the SRM. *Computers, Materials & Continua*, vol. 56, no. 2, pp. 261-272.

**Zhang, H.; Tang, H.; Zhang, L.; Jin, L.; Tanng, Y.** (2003): Evaluation on predietion methods of protein secondary strueture. *Computers and Applied Chemistry*, vol. 20, no. 6, pp. 19-24.

**Zheng, B.; Li, L.** (2013): Protein structural class prediction based on multi-feature and Ma-Ada multi-classifier fusion. *Chinese Journal of Biomedical Engineering*, vol. 32, no. 5, pp. 580-587.