ISBN 978-981-11-3671-9
Proceedings of 2017 the 7th International Workshop on Computer Science and Engineering (WCSE 2017)
Beijing, 25-27 June, 2017, pp. 1353 -1357
doi: 10.18178/wcse.2017.06.234

A Prediction and Correction Model for Protein Secondary Structure Prediction

Yuming Ma, Jinyong Cheng, Yihui Liu+

School of Information Science, Qilu University of Technology, Jinan China

Abstract. Protein secondary structure prediction is one of the central topics in bioinformatics. Machine learning techniques have been widely applied to solve the problem, and many methods have gained substantial success in this research area. In this paper, we propose a prediction and correction model to improve the performance of secondary structure prediction. This model has a correction process on the basis of classification (SVM). Statistical analysis was carried out on the prediction results. These statistical results is used as a priori knowledge to find error patterns and design correction methods to correct it. The experimental results show that our proposed model can indeed improve the prediction accuracy.

Keywords: Protein secondary structure, support vector machine, Random Subspace.

1. Introduction

Protein is the basis of life, which is the most basic structure constitute substance and functional material in organisms. Proteins are the chief actors within the cell, they play a central role in most cellular functions such as gene regulation, metabolism and cell proliferation. Proteins serve as nutrients as well and they provide the organism with the amino acids that are not synthesized by that organism [1]. There are 20 different amino acids in nature that form proteins. Protein structures may be classified into four levels or classes: primary, secondary, tertiary, and quaternary structure [2]. The biological function of a protein is essentially associated with its structure, therefore protein structure prediction is a very important while challenging task in computational biology. However, despite decades of work, the gap between the number of known protein sequences and the number of known protein structures is widening [3].

Protein secondary structure prediction plays a major role in prediction of protein structure. Three basic local structures can be formed: α-helix, β-strand and random coil. There are also some other secondary structures, such as the 310-helix, π-helix, isolated β-bridge, turn and bend, but they are rare. So far, a variety of secondary structure prediction methods have been proposed in the literature. One of the first approaches for predicting protein secondary structure, due to Chou and Fasman, uses a combination of statistical and heuristic rules [4]. The GOR method formalizes the secondary structure prediction problem within an information-theoretic framework [5]. The second generation of methods exhibits better performance by exploiting protein databases, as well as statistic information about amino acid sub sequences and a variety of machine learning methods. For example, many neural network (NN) methods [6], [7], Hidden Markov model (HMM) [8], Support Vector Machines (SVM) [9] and K-nearest neighbours.

The prediction accuracy has been continuously improved over the years, especially by using hybrid or ensemble methods and incorporating evolutionary information in the form of profiles extracted from alignments of multiple homologous sequences [10], [11].

In this article, a prediction and correction model is proposed to improve the performance of secondary structure prediction. We add a correction process on the basis of classification (SVM) and statistical analysis

+ Corresponding author.
E-mail address: yxl@qlu.edu.cn.

of prediction results. In order to verify the effectiveness of the method, experiment was carried out on the data CB513 [12].

2. Data and Method

2.1. Dataset

The widely-used benchmark dataset CB513 is used to evaluate our method in this paper. The dataset CB513 is proposed by Cuff and Barton [12], with the aim of evaluating protein secondary structure prediction methods. It includes the CB396 dataset and almost all proteins of RS126 except nine homologous for which the S.D.score⩾5. It is one of the most used independent datasets in this field.

2.2. Method

In order to improve the accuracy of secondary structure prediction, a prediction and correction model (PCM) is presented. It has three steps: prediction, design correction rule, correction as shown in Fig. 1. The support vector machine (SVM) is utilized as predictor in this paper. The designing of correction rule is based statistics and analysis.

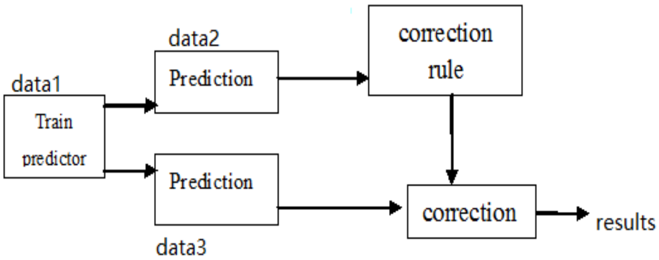


Fig. 1: The prediction and correction model (PCM).

The support vector machine (SVM) is originally a binary classification method developed by Vapnik and colleagues at Bell laboratories [13], with further algorithm improvements by others. SVM is the method that mapped feature vector into a high dimensional vector space, where a maximum margin hyper-plane is established in this space.

LIBSVM [14] is one of the most widely used SVM software. It implements the “one-against-one” approach for multiclass classification. In this article, we use LIBSVM to train the multi-classifier and to predict protein secondary structure.

Random subspace method (RSM) is one of ensemble construction techniques. It was proposed by Ho in 1998 [15]. RSM randomly samples a set of low dimensionality subspaces from the whole original high dimensional features space and then constructs a classifier on each smaller subspace and finally applies a combination rule for the final decision. RSM is a very simple and popular ensemble construction method.

Majority Voting Rule (MVR) is the simplest method to combine multiple classifiers. It does not consider any individual behaviour of each weak classifier. It only counts the largest number of classifiers that agree with each other.

3. Experiments and Results

The widely-used benchmark dataset CB513 is used to evaluate our method in this paper. It was divided into 3 subsets, S_1, S_2 and S_3, $S_i \cap S_j \neq \Phi$, i, j=1,2,3. They are train dataset, validation dataset, and test dataset respectively.

A sliding window method is used to consider a contiguous sequence of amino acids. Each residue is encoded by a vector of dimension 20*w, where w is the sliding window size and w is an odd number. The window is shifted residue by residue through the protein chain.

We use PSI-BLAST obtain the PSSM with three iterations and a cutoff E-value 0.01.The sliding window length w is set to 13. To use the first and the last six amino acids, we fill six zeros before and behind each protein sequence. The profile elements matrix are scaled to the [0,1] range by using the linear function defined as:

$$x' = (x - m_i)/(M_i - m_i) \tag{1}$$

where M_i and m_i represent the maximum and minimum values of the i attribute column vector, x is the raw profile value.

The PCM (Prediction and Correction Model)train the support vector machine(SVM) on the 260 dimension train dataset S_1, and tested it on validation dataset S_2 and test dataset S_3, then a variety of statistics on the prediction results of Statistical results were carried out. According to the above analysis, The correction method was designed according to the statistical results of validation dataset S_2.

3.1. Prediction

About the parameter selection of SVM, We use the RBF kernel, the form is $K(x, x_i) = \exp(-\gamma\|x - x_i\|^2)$, and the two parameters C, γ are decided by using the grid search method. The optimal value of the two parameters are 0.9956 and 0.065. We use the 260-dimensional PSSM data to perform the experiment with SVM classifier, the first row of Table 1 show the prediction accuracy of validation dataset S_2. The first row of Table 2 show the prediction accuracy of validation dataset S_3.

4. Majority Voting Experiment

RSM and MVR are used to combine multiple classifiers. This experiment is carried out on data set S_2 and S_3. Parameter r is the number of features selected from the 260-dimensional dataset randomly, so r<260. Parameter M is the number of the weak classifiers, and it is also the number of random selection. The experiments were carried out 4 times, we set M=10, and let r=170, r =200, r=220, r=240 respectively, the results show in Table 1 and Table 2. By comparison, we find that prediction accuracy of the combined classifier doesn't better than the prediction accuracy of 260-dimensional PSSM.

Table 1. The prediction accuracy on S_2

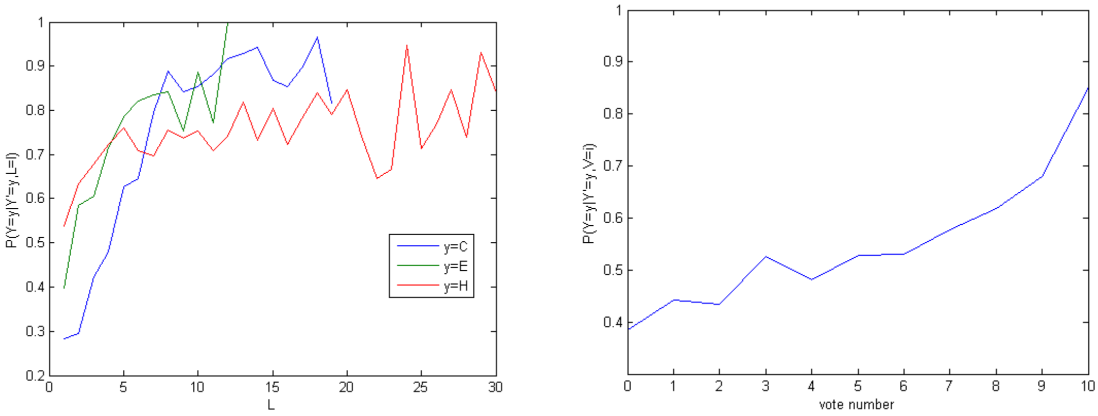
r	M	data	Q_3(%)	Q_H(%)	Q_E(%)	Q_C(%)
260	1	S_2	76.22	77.27	58.64	84.04
170	10	S_2	75.64	74.00	55.44	86.33
200	10	S_2	75.96	75.04	56.22	85.99
220	10	S_2	75.95	75.54	56.01	85.77
240	10	S_2	75.78	75.33	55.55	85.78

Table 2. The prediction accuracy on S_3

r	M	data	Q_3(%)	Q_H(%)	Q_E(%)	Q_C(%)
260	1	S_3	76.74	76.75	59.51	83.13
170	10	S_3	75.83	72.76	55.70	85.29
200	10	S_3	76.51	74.06	56.98	85.34
220	10	S_3	76.56	74.40	57.37	85.08
240	10	S_3	76.40	74.57	56.64	84.91

4.1. Statistics and analysis

A variety of statistics on the prediction results of validation dataset S_2 were carried out, such as the actual number and the predicted number of secondary structure etc. we found that the predicted number of E(β-strand) is less than the actual one number of E, but the predicted number of C(coil) is more than the actual number of C(coil).



(a) (b)

Fig. 2: (a) The accuracy increases with the increase of length(L) (b) The relationship between the of vote number and the accuracy of the prediction results

The accuracy of a single amino acid residues is related to the length (L) of predicted secondary structure segment, the accuracy increases with the increase of length (L). Fig. 2(a) show the result. The accuracy of a single amino acid residues is related to the position (P) in the predicted segment too. When the continuous fragment is long enough, the first position and the last position of the fragment have the higher error probability than other positions. Our purpose is to utilize these statistical results as a priori knowledge to find error patterns and design correction methods to correct it.

The vote number of RSM can help us to get more information about the prediction results of full 260 PSSM, Fig. 2(b) shows the relationship between the number of votes and the accuracy of the prediction results of full 260 PSSM. We can find that the higher the number of votes, the greater prediction accuracy of a single amino acid. we find the error predicted amino acid and then correct it on the basis of the relationship.

4.2. Error patterns

There are several error patterns which can be recognized possibly according to the length of the predicted fragment, the position in the predicted segment and the number of votes, Fig. 3 show 4 error patterns: (a) show when the length of the predicted fragment L≤2, the prediction accuracy are much lower than Q_3; (b)show several special position of the predicted fragment have low accuracy; (c) show several uncontinuous predicted secondary structure segment (d) show a whole predicted secondary structure segment are wrong, this is very difficult to identify it, because the length of the predicted fragment L>2, have no special position, and vote number is high.

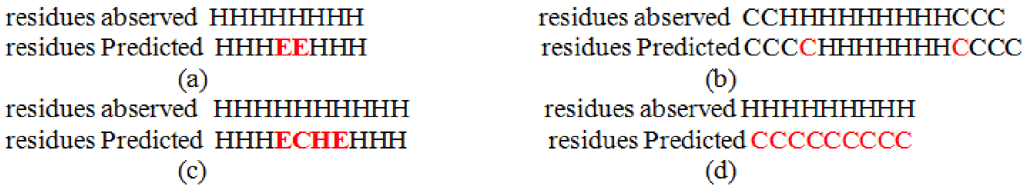

residues abserved HHHHHHHH
residues Predicted HHHEEHHH
(a)
residues abserved CCHHHHHHHHHCCC
residues Predicted CCCCHHHHHHHCCCC
(b)
residues abserved HHHHHHHHHH
residues Predicted HHHECHEHHHH
(c)
residues abserved HHHHHHHHH
residues Predicted CCCCCCCCC
(d)

Fig. 3: Four error patterns.

5. Correction

According to the above analysis, we designed the correction method:

Let Y'(i) is predicted secondary structure, L(i) is the length of fragment that Y(i) belong to, P(i) is the position of Y(i) in fragment, when P(i)=-1, the position of Y(i) is the first, when P(i)=1, the position of Y(i) is the last. V(i) is the vote number of Y(i) obtained from Majority Voting Rule and M=max(V(i)). O(i) is out results and i=1,2,3......l.

Rule 1 if V(i)=M then O(i)= Y'(i)

Rule 2(a) if L(i)≤m_1 and V(i)<k_1, then O(i)= Y'(i-1) or O(i)= Y'(i+1)

Rule 2(b) if L(i)≤m_1 and V(i)<k_1,and Y'(i) ≠'E', then O(i)= Y'(i-1) or O(i)= Y'(i+1)

Rule 3 if Y'(i)='H', Y'(i-1)='E' or Y'(i+1)='E',and L(i)≥m_2 and V(i)<k_2, and P(i)=1 or -1 then O(i)= Y'(i-1) or O(i)= Y'(i+1)

Here k_1,k_2,m_1,m_2 is an integer that can be adjusted of before correction.

Table 3 show the comparision of the results before correction and after correction with Rule 1,2(a),3, Table 4 show the comparision of the results before correction and after correction with Rule 1, 2(b),3.

Table 3. Results of correction with Rule 1,2(a),3

data		Q_3(%)	Q_H(%)	Q_E(%)	Q_C(%)
S_2	Before correction	76.22	77.27	58.64	84.04
	After correction	76.66	77.70	59.01	84.51
S_3	Before correction	76.74	76.75	59.51	83.13
	After correction	77.01	77.29	59.42	83.30

Table 4. Results of correction with Rule 1,2(b),3

data		Q_3(%)	Q_H(%)	Q_E(%)	Q_C(%)
S_2	Before correction	76.22	77.27	58.64	84.04
	After correction	76.61	77.66	61.51	83.94
S_3	Before correction	76.74	76.75	59.51	83.13
	After correction	76.92	76.90	61.39	82.67

6. Conclusions

The protein secondary structure prediction is one of the most important tasks in bioinformatics. Although some new methods have improved the accuracy of prediction to some extent, there is still a long way to find more powerful predictors in this area. In this article, we propose a Prediction and Correction Model based on support vector machine and random subspace method. We add a correction process on the basis of SVM classification. The result of our experiments show that random subspace method (RSM) cannot improve the accuracy of prediction, and Prediction and Correction Model (PCM) can improve the accuracy of prediction. In this method, to find the error pattern and correct it is a key step. This method can improve the accuracy, but it is not significant. So, our future task is to study more effective methods to find and correct the error patterns, and improve the protein secondary structure prediction accuracy.

7. Acknowledgements

The work is supported by the National Natural Science Foundation of China (61375013), Shandong Provincial Natural Science Foundation, China (ZR2013FM020), and the National Natural Science Foundation of China (81402538).

8. References

[1] M.Zvelebil and J.O.Baum, Understanding Bioinformatics, Garland Science—Taylor & Francis Group, 2007.

[2] A. M. Lesk. Introduction to protein architecture .Oxford University Press, 2001.

[3] Yang, B., Qu, W., Xie, Y., and Zhai, Y. "Predicting protein second structure using a novel hybrid method." Expert Systems with Applications An International Journal, vol. 38, 2011, pp. 11657-11664.

[4] P. Y. Chou, and G. D. Fasman, "Prediction of protein conformation, "Biochemistry, vol.13, 1974, pp. 222-245.

[5] J. Garnier, J.F. Gibrat, and B. Robson, “GOR method for predicting protein secondary structure from amino acid sequence,” Methods Enzymol., vol. 266, 1995, pp. 540–553.

[6] Yao, Xin Qiu, H. Zhu, and Z. S. She. "A dynamic Bayesian network approach to protein secondary structure prediction." BMC Bioinformatics9.1 (2008), pp. 1-13.

[7] Qu W., Sui H., Yang B., Qian W. “Improving protein secondary structure prediction using a multi-modal BP method” Computers in Biology and Medicine, 41 (10)2011, pp. 946-959

[8] K. J.Won, T. Hamelryck, A. Prügelbennett and A. Krogh "An evolutionary method for learning HMM structure: prediction of protein secondary structure." BMC Bioinformatics, vol. 8, 2007, pp. 1-13.

[9] Mayoraz, Eddy, and E. Alpaydin. "Support Vector Machines for Multi-class Classification." Proceeding of the International Workshop on Artificial Neural Networks Idiap 1607, 1998, pp. 833-842.

[10] Guo, J., Chen, H., Sun, Z., and Lin, Y. "A novel method for protein secondary structure prediction using dual-layer SVM and profiles. " Proteins-structure Function & Bioinformatics 54.4, 2004, pp. 738-743.

[11] J. Zhou, and O. Troyanskaya, “Deep supervised and convolutional generative stochastic network for protein secondary structure prediction,” Presented at the 31st Int. Conf. Mach. Learn., Beijing, China, 2014

[12] Cuff, J. A., and G. J. Barton. "Application of multiple sequence alignment profiles to improve protein secondary structure prediction. " Proteins Structure Function & Bioinformatics vol. 40, 2000, pp. 502-511.

[13] Vapnik V.N., The Nature of Statistical Learning Theory, 2nd ed., Springer, 2000.

[14] Chang, Chih Chung, and C. J. Lin. "LIBSVM: A library for support vector machines." Acm Transactions on Intelligent Systems & Technology, vol. 2, 2007, pp. 389-396.

[15] Ho, Tin Kam. "The Random Subspace Method for Constructing Decision Forests." IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 20, 1998, pp. 832-844.