Protein Secondary Structure Prediction Using Support Vector Machine With a PSSM Profile and an Advanced Tertiary Classifier

Hae-Jin Hu[1], Phang C. Tai[2], Jieyue He[3], Robert Harrison[1,2] and Yi Pan[1*]
[1]*Department of Computer Science,* [2]*Department of Biology*
Georgia State University, Atlanta, GA 30303-4110, USA
[3]*Department of Computer Science, Southeast University, Nanjing 210096, China*
**Corresponding author, Phone: 404-651-0649, Email: pan@cs.gsu.edu*

Abstract

In this study, the support vector machine (SVM) is applied as a learning machine for the secondary structure prediction. As an encoding scheme for training the SVM, position-specific scoring matrix (PSSM) is adopted. To improve the prediction accuracy, three optimization processes such as encoding scheme, sliding window size and parameter optimization are performed. For the multi-class classification, the results of three one-versus-one binary classifiers (H/E, E/C and C/H) are combined using our new tertiary classifier called SVM_Represent. By applying this new tertiary classifier, the Q_3 prediction accuracy reaches 89.6% on the RS126 dataset and 90.1% on the CB513 dataset. Also the Segment Overlap Measure (SOV) is 85.0% on the RS126 dataset and 85.7% on the CB513 dataset. Compared with the existing best prediction methods, our new prediction algorithm improves the accuracy about 13% in terms of Q_3 and SOV, the two most commonly used accuracy measures.

1. Introduction

The protein secondary structure prediction is a crucial intermediate step for the protein tertiary structure prediction. The recent trend of secondary structure prediction studies is mostly based on the neural network or the support vector machine (SVM).

This research was supported in part by the U.S. National Institutes of Health (NIH) under grants R01 GM34766-17S1, and P20 GM065762-01A1, and the U.S. National Science Foundation (NSF) under grants ECS-0196569, and ECS-0334813. This work was also supported by the Georgia Cancer Coalition and used computer hardware supplied by the Georgia Research Alliance.

In this study, SVM is used as a machine learning tool for the prediction of secondary structure. As an encoding scheme for SVM, position-specific scoring matrix (PSSM) profile generated by PSI-BLAST search is adopted after comparing the performance with other encoding schemes, such as hydrophobicity matrix or combined matrix of orthogonal and BLOSUM62 matrix. To improve the prediction accuracy, sliding window size and SVM parameter values are optimized. For the multi-class classification, the results of three one-versus-one binary classifiers (H/E, E/C and C/H) are combined using our new tertiary classifier called SVM_Represent.

2. Method

2.1. Training and testing data sets

To compare the results with previous results [1, 2], RS126 and CB513 data set are used. The RS126 data set is a non-homologous set sharing less than 25% sequence identity which is proposed by Rost & Sander [3]. The CB513 data set is created by Cuff and Barton and it is also a non-homologous set [4]. With these data set, the seven-fold cross validation test was done. In the seven-fold cross validation test, one subset is chosen for testing and remaining 6 subsets are used for training and this process is repeated until all the subsets are chosen for the testing.

2.2. Optimization Processes

For an encoding scheme optimization, three different encoding schemes, such as hydrophobicity matrix, the combined orthogonal and Blosum62 matrix and PSSM matrix are tested. These three encoding schemes are applied to the SVM using the sliding window method. In this sliding window method, the information about the local interactions among neighboring residues can be embedded together. To

find the optimal window size, different window lengths ranging from 7 to 21 residues are tested. As a kernel function of SVM, radial basis function (RBF) kernel is selected based on the test results of the previous studies [1, 2]. To select the optimal parameter value γ in RBF kernel and the cost factor C (penalty of the misclassified data), different γ and C pairs are tested.

The performance of the prediction scheme is evaluated with two measures. The first one is the most commonly used three-state overall percentage of correctly predicted residues, Q_3. The second measure is a Segment Overlap Measure (SOV). It is developed by Rost et al.(1994) and modified by Zemla et al.(1999) to evaluate the quality of a prediction in a more realistic manner.

2.3. Binary classifier construction

Three one-versus-one classifiers (H/E, E/C and C/H) were constructed. Here, the name 'one' in one-versus-one classifier refers to positive class and negative class respectively. For example, the classifier E/C classifies the testing sample as sheet or coil.

2.4. Tertiary classifier design

In this research, to combine the output from the binary classifiers for secondary structure prediction., new tertiary classifier called SVM_Represent is developed. In this scheme, the classifier with the absolute maximum distance is chosen as the representative classifier for the final decision of the class. Based on the value sign of this representative classifier, the final class is chosen. For example, if the values of the decision function of the each one-versus-one classifiers (H/E, E/C, C/H) are -1.7, 0, and -2.5 respectively, the binary classifier with highest absolute value, here C/H classifier, can be chosen for deciding the final class. Once this representative classifier is selected, the final class is assigned based on the value of this classifier. In this example, since the value of C/H classifier shows negative, the final class is assigned as helix.

3. Result analysis

As can be observed from Table 1, by applying the new tertiary classifier SVM_Represent, the Q_3 prediction accuracy reaches 89.6% on the RS126 dataset and 90.1% on the CB513 dataset. Also the Segment Overlap Measure (SOV) is 85.0% on the RS126 dataset and 85.7% on the CB513 dataset. Compared with the existing best prediction methods, our new prediction algorithm improves the accuracy about 13% in terms of Q_3 and SOV, the two most commonly used accuracy measures.

Table 1. Accuracy comparison with other research results

Method	Q_3 (%)	Q_H (%)	Q_E (%)	Q_c (%)	SOV94 (%)	SOV99 (%)
PHD (RS126)	70.8	72.0	66.0	72.0	73.5	-
SVMfreq (RS126)	71.2	73.0	58.0	73.0	74.6	-
SVMpsi (RS126)	76.1	77.2	63.9	81.5	79.6	72.0
SVMpsi (CB513)	76.6	78.1	65.6	81.1	80.1	73.5
SVM_Represent (RS126)	89.6	86.9	81.0	95.3	-	85.0
SVM_Represent (CB513)	90.1	89.3	81.4	94.9	-	85.7

PHD result is obtained by Rost and Sander [3] and and SVMfreq result is obtained by Hua and Sun [2].
SVMpsi result is obtained by Kim and Park [1]. SVM_Represent is a new method proposed by this study.

4. References

[1] H. Kim and H. Park, "Protein Secondary Structure Prediction Based on an Improved Support Vector Machines Approach," *Protein Eng*, vol. 16, pp. 553-560, 2003.

[2] S. Hua, and Z. Sun, "A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach," *J. Mol. Biol*, vol. 308, pp. 397-407, 2001.

[3] B. Rost, and C. Sander, "Prediction of protein secondary structure at better than 70% accuracy.," *J. Mol. Biol.*, vol. 232, pp. 584-599, 1993.

[4] J. A. Cuff and G. J. Barton, "Evaluation and improvement of multiple sequence methods for protein secondary structure prediction.," *Proteins*, vol. 34, pp. 508-19, 1999.