A Hybrid Method for Protein Secondary Structure Prediction

Shing-Hwang Doong   Chi-Yuan Yeh

Department of Information Management, ShuTe University
Yen-Chau, Kaohsiung 804, Taiwan
tungsh@mail.stu.edu.tw   jofu@ms17.hinet.net

Abstract-*Protein secondary structure can be used to help determine the tertiary structure via the fold recognition. Predicting the secondary structure from the protein sequence has attracted the attention of many researchers. Support Vector Machine (SVM) is a new learning algorithm based on statistical learning theory that has been successfully applied to the protein secondary structure prediction problem. However, the algorithm takes a long time to train the prediction model with a large data set. It becomes important to revise the method so that the time performance is improved while the accuracy performance is maintained. In this study, we implement a genetic algorithm to cluster the data set before the structure classification is predicted. Using position specific scoring matrix as part of the input, the hybrid method achieves good performances through 7-fold cross validation tests on a set of 513 non-redundant protein sequences (the CB513 data set). The result is comparable to that of the existing best prediction, yet the time spent is substantially reduced.*

Keyword: Secondary structure prediction, support vector machine, clustering.

1. Introduction

A protein is determined by the sequence (called the primary structure) of amino acids that make up the protein, and uses its tertiary structure (3-D structure) to carry out different biological tasks. Knowing the structure of a protein will take us less time and investment to develop new drugs. Besides experimental methods, one can use several knowledge-based methods to predict the tertiary structure of a protein from its primary structure. A protein frequently mutates faster in its sequence than its structure will evolve. Predicting a protein's tertiary structure from the sequence is nontrivial. The secondary structure of a protein is simpler than the tertiary structure and is found to have important influences on the tertiary structure. This is reflected in the fold recognition based paradigm for tertiary structure prediction.

The development of protein secondary structure prediction until now has almost 50 years of history; the prediction accuracy continues to rise [14].

Researchers proposed different solutions to improve the performance at every step.

The pre-processing of data is very important in secondary structure prediction and it influences the prediction accuracy critically [14, 16]. Several researches focused on this part to improve the accuracy performance. Rost & Sander [15] proposed the PHD algorithm using profiles from HSSP [18] as input and the prediction accuracy broke the 70% benchmark. Jones [9] proposed PSIPRED using Position Specific Scoring Matrix (PSSM) generated from PSI-BLAST [2] as input and the prediction accuracy broke the 75% benchmark.

On the computational side, researchers are devising new algorithms to improve the prediction results. Machine learning methods are frequently used to predict secondary structures because of the availability of more quality structure data and the fast advancement of the algorithms. Artificial neural networks (ANN) [9, 12, 15] and support vector machines (SVM) [8, 11] are two such methods.

Previous researches focused on the accuracy performance by using different inputs and prediction algorithms, however they did not look at the time performance problem. This problem is nontrivial when ANN or SVM is used to train a large data set. Clustering is a common technique used in data mining to preprocess a large data set. The objective of clustering is to partition data into clusters of similar data so that finer analysis can be performed in each cluster. This partitioning step may provide two advantages for the eventual analysis of data: (i) since each cluster is smaller in size than the original data set, the training time of a prediction analysis in each cluster may be substantially reduced; and (ii) since data in each cluster are more similar to one another, it might be easier to perform further analysis in the cluster by avoiding the noises problem.

In this study we propose a hybrid method to solve the time performance issue without affecting the accuracy performance. The paper is organized as follows. In section 2, we discuss the data and algorithms used in this study. Section 3 is devoted to the discussion of the experimental results. We conclude with a few remarks in section 4.

2. Materials and Methods

2.1. Data set

Many data sets are available for the experiment, e.g. RS126 [15], CB396 and CB513 [4] and CASP [12], etc. Different researchers have reported different prediction rates using different data sets. In order to compare with recent studies in secondary structure prediction, we selected the CB513 set in this study. This data set combines CB396 and RS126 by removing 9 sequences from the latter data set. Protein sequences in the CB513 set are non-redundant. There are 84119 residues in these protein sequences, and the percentages of helix, sheet and coil secondary structures are roughly 34.6%, 21.3% and 44.1%. A 7-fold cross validation is used in this study like previous researches [8, 11].

2.2. Clustering of protein sequences

Clustering is used to partition a large data set into clusters so that each cluster contains more homogeneous data. Traditional techniques include two categories: hierarchical and partitional. Hierarchical techniques split or merge data in order to form a dendrogram. Since hierarchical methods generally consume more computing resources and they do not provide cluster representative points, we selected partitional methods to cluster training sets of the CB513. A partitional method uses cluster representative points (called medoids) to attract cluster members. A performance measure is used to judge the goodness of the partition. We will use a k-means like performance measure in our study.

First of all, the similarity between two proteins is obtained by aligning the sequences using dynamic programming (DP). The PAM250 matrix is used to measure the substitution rates of amino acids in DP. Let $sim(s_i, s_j)$ denote the alignment score between sequences s_i and s_j; the larger this number is, the more similar these two sequences are according to the PAM250 matrix. Suppose we are partitioning the full CB513 data set of protein sequences into c clusters, then we measure the clustering performance given by a set V of c medoids $R_1, …, R_c$ according to the following function:

$$f(V) = \sum_{i=1}^{513} sim(s_i, R_{j(i)}) \qquad (1)$$

Here $R_{j(i)}$ is the protein from V that is most similar to the sequence s_i according to the similarity measure defined above. Each medoid R_j is a protein sequence from the CB513 set, and our objective is to maximize this performance measure by finding a proper set V. This performance measure is similar to that of the famous k-means method. Though it is a combinatorial optimization problem – choosing the best combination of c sequences from the 513 sequences to maximize the performance measure, an exhaustive search method is not appropriate when c is moderate. We employ a robust search algorithm, the genetic algorithm, to find the near optimal solution for the clustering problem. In this study, we set c to 3 since it produced the best prediction accuracy in a sample study. Another reason for setting this number of c is explained in section 3.2 regarding the jury method.

2.3. Data encoding

Several methods are available to assign the secondary structure of a protein sequence with known tertiary structure, e.g. DSSP [10] and DEFINE [13], etc. We selected DSSP as it is the most widely used secondary structure definition program. DSSP assigns residues to eight different classes, which are H (α-helix), G (3_{10}-helix), I (π-helix), E (β-strand), B (isolated β-bridge), T (turn), S (bend), and - (the rest). Reducing from eight classes to three classes of helix (H), sheet (E), and coil (C) is an important step in the encoding of structure data [16]. Two popular reduction methods are: (i) H, G, I to H; E to E; all other states to C; and (ii) H, G to H; E, B to E; all other states to C [8]. In CB513, method (ii) yields many discrete states such as CEC, CEH, and HEC while method (i) does not. So, reduction method (i) is adopted in this study.

2.3.1. Position Specific Scoring Matrix (PSSM). As Rost and Sander have pointed out in [16] that using evolutionary information such as the profile as the input for structure prediction can improve 5-10% over the second generation prediction method that uses the sliding window alone. Since sequence data generally mutate faster than the structure data, if we can detect more homologous sequences with low sequence identity, then the inferred evolutionary information will be very useful in the structure prediction. Altschul et al. [2] proposed the Position Specific Iterated BLAST (PSI-BLAST) algorithm to detect more homologous sequences than BLAST [1], which was used to construct the profile in Rost and Sander's study [15]. Thus the PSSM generated from PSI-BLAST may contain more evolutionary information than the profile or the sequence alone, and will provide more useful information for the prediction algorithm [9, 11]. The PSI-BLAST program should use a large database of protein sequences to obtain useful evolutionary information in PSSM. We used the stand-alone edition of PSI-BLAST and the NCBI non-redundant (nr) database to get the PSSM of each sequence in CB513. During this preprocessing step, a setting of 3-iteration and the default E-value were adopted in PSI-BLAST.

2.3.2. Relative frequencies of secondary structures in residues. Even though the three secondary structures H, E and C appear in typical data sets with

the ratio approximately equal to 3:2:5 [16], they do appear differently in each residue. Table 1 from [3] lists the relative frequencies of the three secondary structures in each residue. This statistics will be used as part of the input data to the prediction problem.

Table 1. Relative frequencies of secondary structures in each residue [3]

Residue	Helix (H)	Sheet (E)	Coil (C)
A	1.41	0.72	0.82
R	1.21	0.84	0.90
N	0.76	0.48	1.34
D	0.99	0.39	1.24
C	0.66	1.40	0.54
Q	1.27	0.98	0.84
E	1.59	0.52	1.01
G	0.43	0.58	1.77
H	1.05	0.80	0.81
I	1.09	1.67	0.47
L	1.34	1.22	0.57
K	1.23	0.69	1.07
M	1.30	1.14	0.52
F	1.16	1.33	0.59
P	0.34	0.31	1.32
S	0.57	0.96	1.22
T	0.76	1.17	0.90
W	1.02	1.35	0.65
Y	0.74	1.45	0.76
V	0.90	1.87	0.41

2.3.3. Encoding of the input data. In this study, the PSSM of a sliding window around and the relative structure frequencies of the center residue will be used as the input to the prediction algorithm. The length of the window affects the prediction accuracy [8]. If the window is too short, then useful information from neighboring residues is lost. On the other hand, if the length is too large, then noises may affect the prediction accuracy. In most cases, the length of the sliding window is between 7 and 17, and the best length is usually obtained by the trial-and-error procedure. In this study, the length of the sliding window was between 11 and 17, and we found that 15 yielded the best prediction result. A sliding of 15 will be used in the following study.

2.4. Support vector machines

Vapnik and his coworker, based on statistical learning theory, proposed a novel method called Support Vectors Machines (SVMs) to perform data classification and regression [19]. Because of the high performance, SVM is receiving the attentions of more researchers in bioinformatics.

Many learning algorithms including Artificial Neural Networks (ANNs) implement the empirical risk minimization (ERM) principle to learn the prediction model. The empirical risk $R_{emp}[f]$ in eqn. (2) is given by the fitting error of the model f with the training data.

$$R_{emp}[f] = \frac{1}{l} \sum_{i=1}^{l} | f(x_i) - y_i | \qquad (2)$$

Vapnik proved a type of error estimate in the following equation,

$$R[f] \le R_{emp}[f] + capacity \qquad (3)$$

where the expected risk $R[f] = \int | f(x) - y | \, dP(x, y)$ is the average actual error according to the model f over the test samples drawn from the distribution $P(x, y)$. The training samples are assumed to have the same distribution function. SVM tries to control both the empirical error and the generalization error (controlled by the *capacity* term in eqn. (3)) at the same time. Using the structure risk minimization (SRM) principle, SVM finds a balance between the fitting power of a learning function on the training data and the complexity of the learning function [19]. Therefore SVM can avoid the overfitting problem frequently encountered in ANN. Because of this special property of SVM, we employ SVM as the classification algorithm in the secondary structure prediction problem.

SVM was originally designed for binary classification. Since proteins have three different types of secondary structures according to our reduction method above, some modification of the SVM usage is necessary. Researchers have proposed different methods to solve the multi-class problem [6, 7]. One of them is to combine several binary classifiers to construct the tertiary classifier. This type of solution will be called the combination method. The other type of solution is to solve the multi-class problem directly by extending the original SVM theory. We will call the latter type the decomposition method.

We used the open source software BSVM [6, 7] to train a model from the training data, and made prediction on the test data. A radial basis function (RBF) kernel was adopted for the BSVM, and we used a soft margin to handle the noises. This left us with two hyper-parameters C (the regularization parameter that controls the weight of the fitting error) and γ (the width of the Gaussian function) to determine. BSVM provides a tool to determine these values optimally. It was found that the best result was given by the choice where $C = 1.5$ and $\gamma = 0.15$.

2.5. A hybrid method for secondary structure prediction

A hybrid method for the secondary structure prediction problem is proposed as follows:

1. Partition the training set according to the clustering section by finding the proper medoids.
2. Train a SVM prediction model using the

data encoding method stated above for each cluster of the training set.

3. Assign a test sequence to a proper cluster and use the prediction model from that cluster to predict the secondary structure of the sequence.

Further information of this hybrid approach will be detailed in the next section.

3. Results and Discussions

3.1. Prediction accuracy assessment

A couple of indicators may be used to evaluate the performance of secondary structure prediction, e.g. the three-state single residue accuracy (Q_3); for i = H, E or C, the percentage of residues observed in state i (Q_i^{obs}), the percentage of residues predicted in state i (Q_i^{pre}), Mathews correlation coefficient (c_i), and the segment overlap measurement (Sov). Rost & Sander [17] proposed the Sov in 1994 as a way to measure the prediction accuracy. In 1999 Zemla et al. [20] modified the original Sov definition by redefining $\delta(s_1, s_2)$ and the normalization factor $N(i)$. Zemla's definition is more rigorous than Rost & Sander's original definition. In order to distinguish these two definitions, we refer to Rost & Sander's definition as $Sov94$ and Zemla's definition as $Sov99$.

3.2. Comparison with other studies

Table 2 summarizes results of 7-fold cross validation tests from various studies using SVM on the CB513 data set. Except the first two rows, all other results are obtained in this study. We can inspect the results from three perspectives: (i) the input variables for the SVM classifier (the PSSM + structure frequencies vs. PSSM vs. profiles); (ii) the method used to construct the tertiary classifier (the combination method vs. decomposition method); and (iii) the segmentation of data set (clustered vs. non-clustered data set).

In the input variables perspective, Hua & Sun (row 1) used profiles of the sliding window as input; Kim & Park (row 2) used PSSM of the sliding window as input; we used both the PSSM of the sliding window and the relative structure frequencies of the center residue as input. From rows 1, 2, 3 and 5, we can see that using PSSM is more advantageous than the profiles approach (3% increase in accuracy), and adding the relative frequencies will improve the result slightly (1% increase).

Regarding the construction of tertiary classifiers, both Hua & Sun [8] and Kim & Park [11] used the combination method to construct tertiary classifiers. We used the decomposition (row 3) and combination (row 5) methods to solve the multi-class problem in SVM. The combination method is a little bit more accurate than the decomposition method.

In the segmentation of data set, all previous studies including Hua & Sun and Kim & Park used the non-clustered training set in the cross validation test. Training a SVM prediction model is converted to a quadratic programming problem [19]. When the training set is large, it takes the algorithm substantial time to learn the model. For example, using the decomposition method took us more than a week to finish a 7-fold cross validation test on the CB513 without the clustering preprocessing. On the other hand, if we adopt the hybrid method, the total time for a 7-fold cross validation test is less than a day. In rows 7-10 of Table 2, global alignments were used in DP to compute the similarity measure between two protein sequences, while in rows 11-14, we used the local alignments in DP.

We now describe the hybrid approach in further details. First of all, we separated the CB513 data set into the training set (6/7 of CB513) and the test set (the remaining CB513). The roles of training and test sets will be rotated according to a 7-fold cross validation procedure. The partitional clustering method of section 2.2 was applied to the training set (approximately 440 protein sequences) of CB513. We assumed three clusters are to be located in the partitioning, and used the genetic algorithm to find the medoids and their associated cluster members. A prediction model was built for each cluster via the decomposition approach by using the BSVM software. In rows 7 and 11 of Table 2, each sequence from the test set was categorized to the cluster corresponding to the nearest medoid, and the prediction model from that cluster was used to predict the structure of this test sequence.

The prediction accuracy of Q_3 is reduced by about 1.5% using this basic hybrid approach. In order to minimize the possible errors caused by incorrectly categorizing a test sequence using the medoids, we also experimented a jury modification of the hybrid method. Since the prediction models of all clusters have been trained, we can quickly predict the structure of the test sequence using these three models. If two of the models predict the residue with the same structure say C (coil), then we will assign C to this residue. On the other hand, if these models assign three different structures, H, E and C, to the residue, then we will assign H to the residue. Using a 3-cluster segmentation allowed us to implement this jury assignment easily. Rows 9 and 13 of Table 2 show that the jury modification can increase the accuracy by about 1%.

The time spent in the genetic clustering of the training set is negligible compared to the time spent in training the SVM models. Clustering, as a preprocessing of the data, has reduced the training time substantially without sacrificing the accuracy performance too much when a jury modification of the hybrid method is implemented.

1183

The results in Table 2 also indicate that the prediction accuracies of H and C are higher than E; this finding is consistent with previous studies. Our Mathews correlation coefficients of the three classes are better than other's. The *Sov99* results of our experiment are lower than Kim & Park's result. When we examined our predicted results more carefully, we found that there were a few discrete states like CHE, EHE, HEC or HEH in the predicted results. Because a α-helix contains at least three consecutive Hs, these kinds of discrete states are unreasonable in the real world. Using this type of knowledge from molecular biology, we post-processed the prediction and the results are shown in rows with the KB suffix in Table 2. The *Sov99* scores of our experiments rise to beat Kim & Park's. It is also interesting to note that the Q_3 score and other indicators are improved as well.

4. Conclusions

From this study of the protein secondary structure prediction problem, we conclude that a few actions may be used to improve the accuracy and time performances of the prediction problem: (i) using proper variables as input can increase the accuracy, e.g. using the PSSM of the sliding window and the relative structure frequencies improves the prediction rate by about 1-3%; (ii) pre-processing the training set by a clustering procedure can substantially reduce the training time of SVM models; and (iii) post-processing the prediction by using knowledge from molecular biology can remove certain unreasonable cases from the prediction and hence improve the prediction accuracy.

Clustering, as a preprocessing step to reduce the data size and increase the homogeneity of data in a cluster, has been used frequently in data mining field before a classification analysis is performed. In this study, we proposed a hybrid method to predict the secondary structure of a protein sequence from its primary structure. We clustered the training set by using a partitional technique, trained SVM prediction models for the clusters, assigned test sequences to appropriate clusters and predicted the structure using the corresponding models. The basic hybrid method reduced the prediction rate by about 1.5%, while the experimental time was reduced substantially. Using the jury modification of the hybrid method improved the accuracy by about 1% with a small add up to the processing time.

One may argue that the more data available to train the SVM prediction model, the more accurate the model will predict. For example, using the full training set to build the SVM model seemed to predict better than the hybrid method. We do agree that the more sequence data available for finding homologous sequences in the PSI-BLAST program, the more evolutionary information the PSSM will contain for structure prediction. This is the reason why we used the nr database in NCBI to compute the PSSM. However, just like a classification problem in data mining, we may ask: will irrelevant sequences from other clusters cause the noises problem in addition to the lengthy training time issue in the secondary structure prediction problem?

Adding the clustering procedure as the preprocessing step of data classification may introduce two issues in the classification problem: (i) the cluster assignment problem. A test data may be categorized into the wrong cluster; and (ii) the compatibility issue between the clustering procedure and the classification algorithm. In the secondary structure prediction problem, we found that a jury method can be used to alleviate the false categorization problem. On the other hand, how to ensure that the clustering result is compatible with the later classification analysis remains an issue that merits further investigation in a hybrid method.

Acknowledgments

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Meyers, and D.T. Lipman, "Basic Local Alignment Search Tool," *J. Mol. Biol.*, vol 215, pp. 403-410, 1990.

[2] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Res.*, vol 25, pp. 3389-3402, 1997.

[3] T.E. Creighton, *Proteins: structures and molecular properties*, second edition, W. H. Freeman, New York, 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. 508-519, 1999.

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

[6] C. Hsu, and C. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Transactions on Neural Networks*, vol 13, pp. 415-425, 2002.

[7] C. Hsu, and C. Lin, "A simple decomposition method for support vector machines," *Machine Learning*, vol 46, pp. 291-314, 2002.

[8] 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.

[9] D.T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," *J. Mol. Biol.*, vol 292, pp. 195-202, 1999.

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

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

[12] G. Pollastri, D. Przybylski, B. Rost, and P. Baldi, "Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles," *Proteins*, vol 47, pp. 228-235, 2002.

[13] F.M. Richards, and C.E. Kundrot, "Identification of Structural Motifs from Protein Coordinate Data: Secondary Structure and First-Level Supersecondary Structure," *Proteins*, vol 3, pp. 71-84, 1988.

[14] B. Rost, "Review: protein secondary structure prediction continues to rise," *J. Struct. Biol.*, vol 134, pp. 204-218, 2001.

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

[16] B. Rost, and C. Sander, "Third generation prediction of secondary structures," *Methods Mol. Biol.*, vol 143, pp. 71-95, 2000.

[17] B. Rost, C. Sander, and R. Schneider, "Redefining the goals of protein secondary structure prediction," *J. Mol. Biol.*, vol 235, pp. 13-26, 1994.

[18] R.M. Schwartz, and M.O. Dayhoff, "Matrices for detecting distant relationships," *Atlas of Protein Sequence and Structure*, Nat. Biomed. Res. Found., Washington D.C., vol 5, pp. 353-358, 1978.

[19] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[20] A. Zemla, C. Venclovas, K. Fidelis, and B. Rost, "A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment," *Proteins*, vol 34, pp. 220-223, 1999

Table 2. Summary of prediction results from various experiments

Method	Q_3	Q_H^{obs}	Q_E^{obs}	Q_C^{obs}	Q_H^{pre}	Q_E^{pre}	Q_C^{pre}	C_H	C_E	C_C	Sov94	Sov99
1. Hua & Sun [8]	73.5	75.0	60.0	79.0	79.0	67.0	70.0	0.65	0.53	0.54	76.2	-
2. Kim & Park [11]	76.6	78.1	65.6	81.1	84.4	74.8	72.1	0.68	0.60	0.56	80.1	73.5
3. Decomposition	77.6	76.6	66.3	83.8	85.7	76.3	73.3	0.72	0.64	0.59	91.0	70.2
4. Decomposition + KB	77.8	76.6	65.9	84.4	86.0	77.6	73.2	0.72	0.64	0.60	84.9	75.0
5. Combination	77.7	77.4	64.1	84.6	84.8	78.1	73.4	0.72	0.64	0.60	89.7	71.5
6. Combination + KB	77.9	77.4	63.7	85.0	85.3	78.9	73.2	0.72	0.64	0.60	84.2	75.0
7. Global	76.1	74.3	65.2	82.8	84.7	73.9	72.1	0.70	0.62	0.57	89.4	69.0
8. Global + KB	76.2	74.3	64.9	83.3	85.1	74.8	71.8	0.70	0.62	0.57	83.9	73.2
9. Global + Jury	76.7	73.9	64.2	84.9	86.4	76.0	71.6	0.71	0.62	0.58	88.0	70.5
10. Global + Jury + KB	76.8	73.9	63.8	85.4	86.8	76.8	71.4	0.71	0.63	0.58	83.4	73.9
11. Local	76.0	74.6	61.5	84.1	84.2	76.3	71.3	0.70	0.61	0.57	88.1	68.9
12. Local + KB	76.1	74.4	60.9	84.7	84.8	77.3	71.0	0.70	0.61	0.57	82.3	72.4
13. Local + Jury	76.9	74.8	62.0	85.7	86.0	78.4	71.6	0.71	0.63	0.58	88.0	70.5
14. Local + Jury + KB	77.0	74.8	61.5	86.1	86.4	79.1	71.3	0.71	0.63	0.58	82.4	73.8