Improved Protein Secondary Structure Prediction Using Support Vector Machine With a New Encoding Scheme and an Advanced Tertiary Classifier

Hae-Jin Hu, Yi Pan*, *Senior Member, IEEE*, Robert Harrison, and Phang C. Tai

***Abstract*—Prediction of protein secondary structures is an important problem in bioinformatics and has many applications. The recent trend of secondary structure prediction studies is mostly based on the neural network or the support vector machine (SVM). The SVM method is a comparatively new learning system which has mostly been used in pattern recognition problems. In this study, SVM is used as a machine learning tool for the prediction of secondary structure and several encoding schemes, including orthogonal matrix, hydrophobicity matrix, BLOSUM62 substitution matrix, and combined matrix of these, are applied and optimized to improve the prediction accuracy. Also, the optimal window length for six SVM binary classifiers is established by testing different window sizes and our new encoding scheme is tested based on this optimal window size via sevenfold cross validation tests. The results show 2% increase in the accuracy of the binary classifiers when compared with the instances in which the classical orthogonal matrix is used. Finally, to combine the results of the six SVM binary classifiers, a new tertiary classifier which combines the results of one-versus-one binary classifiers is introduced and the performance is compared with those of existing tertiary classifiers. According to the results, the Q_3 prediction accuracy of new tertiary classifier reaches 78.8% and this is better than the best result reported in the literature.**

***Index Terms*—Binary classifier, BLOSUM62, encoding scheme, orthorgonal matrix, Position Specific Scoring Matrix (PSSM), support vector machine (SVM), tertiary classifier.**

I. Introduction

FOR THE PAST few decades, there have been many studies focused on the prediction of protein structure. Since the direct prediction of protein tertiary structure was challenging, many approaches begin with the prediction of secondary structure and apply the results to predict the tertiary structure.

In the recent machine learning technologies for secondary structure prediction, the neural networks (NNs) or the support vector machines (SVMs) have been generally adopted for the learning tools. Among the studies of secondary structure prediction using NNs, the Porfile network from HeiDelberg (PHD) scheme [1] adopted three-layer feed-forward NNs with the inclusion of evolutionary information using multiple sequence alignments. And it showed outstanding performance of $Q_3 = 70.8\%$ on 126 nonhomologous data set (RS126). Besides the PHD scheme, there are many other approaches using different NN architectures. For example, Riis and Krogh [2] designed the highly structured NNs consisting of small neural networks for the prediction of three states of the secondary structure separately. With this scheme, they could avoid the overfitting problem effectively. Also, with the use of another NN to combine ensembles of the single-sequence networks and with the inclusion of multiple alignment information, they attained 71.3% cross validation accuracy on the RS126 set. Chandonia and Karplus [3] introduced a novel method for processing and decoding the protein sequence with NNs by using larger training data set, such as 681 nonhomologous proteins. And with the use of jury method, this scheme recorded 74.8% accuracy.

The SVM method is a comparatively new learning system which is developed by Vapnik and Cortes [4]. This machine uses hypothesis space of linear functions in a high-dimensional feature space, and it is trained with a learning algorithm based on optimization theory [5]. The superior features of this machine is that first it can avoid the overfitting effectively with the use of structural risk minimization. Second, the formulation can be simplified to a convex QP problem; the training can certainly converge to a global optimal [6]. Third, for the given data set, information can be condensed while training without loss of useful information [7]. Since this SVM has outperformed most other learning systems, including NNs in most pattern recognition problems [7], it has been gradually applied to pattern classification problems in biology.

One of the recent studies adopting this SVM learning machine for secondary structure prediction is the one which used frequency profiles with evolutionary information as an encoding scheme for SVM [7]. With this scheme, the authors claimed the prediction accuracy of $Q_3 = 73.5\%$ and Segment Overlap Measure (SOV) of SOV94 $= 76.2\%$ on the CB513 data set. Another approach is the one which adopted two layers of SVM with a weighted cost function for balanced training [8] and it presented prediction accuracy of $Q_3 = 71.5\%$ on C396 set. Also there was another scheme that incorporated PSI-BLAST Position Specific Scoring Matrix (PSSM) profiles as an input vector [9] and that

Manuscript received February 6, 2004; revised August 4, 2004. This work was supported in part by the U.S. National Institutes of Health under Grants R01 GM34766-17S1, and P20 GM065762-01A1, in part by the U.S. National Science Foundation under Grants ECS-0196569 and ECS-0334813, and in part by the Georgia Cancer Coalition using computer hardware supplied by the Georgia Research Alliance. *Asterisk indicates corresponding author.*

H.-J. Hu is with the Department of Computer Science, Georgia State University, Atlanta, GA 30303-4110 USA (e-mail: jpark1808@earthlink.net).

*Y. Pan is with the Department of Computer Science, Georgia State University, Atlanta, GA 30303-4110 USA (e-mail: pan@cs.gsu.edu).

R. Harrison is with the Department of Computer Science and the Department of Biology, Georgia State University, Atlanta, GA 30303-4110 USA (e-mail: rharrison@cs.gsu.edu).

P. C. Tai is with the Department of Biology, Georgia State University, Atlanta, GA 30303-4110 USA (e-mail: biopct@langate.gsu.edu).

Digital Object Identifier 10.1109/TNB.2004.837906

applied new tertiary classifiers. This scheme, which is called SVMpsi, showed the prediction accuracy of $Q_3 = 76.6\%$ and SOV94 $= 80.1\%$ on the CB513 data set.

In this paper, first, there is an attempt to increase the prediction accuracy of the secondary structure by applying the new encoding schemes, such as hydrophobicity matrix, BLOSSUM62 matrix, and the combined matrix with these or with the standard orthogonal matrix. For the comparison of the results with those of the previous studies [7], [9], the common data set of RS126 [1] is used. Second, to set up the optimal window size of the sliding window scheme, different window lengths from 5 to 19 are tested. Third, proper kernel function is chosen based on the results of the previous studies [7], [9] and the kernel function parameter and the regularization parameter C of the SVM are optimized. Finally, to combine the results of the SVM binary classifiers, a new tertiary classifier is designed and the performance is compared with those of several existing tertiary classifiers [7], [9].

II. Methods

A. Secondary Structure Assignment

The secondary structure is decided from the experimentally determined tertiary structure with the schemes, such as DSSP [10], DEFINE [11], or STRIDE [12]. In this study, the most generally used DSSP scheme is adopted. The DSSP classifies the secondary structure into eight different classes: H (α- helix), G (3_{10}-helix), I (π-helix), E (β-strand), B (isolated β-bridge), T (turn), S (bend), and - (rest). These eight classes are reduced into three regular classes based on the following method: H, G and I to H; E to E; all others to C.

B. Training and Testing Data Sets

To compare the results of this study with previous results [7], [9], RS126 data set is used. The RS126 data set is proposed by Rost and Sander [1] and based on their definition, it is a nonhomologous set. The authors define nonhomologous as "no two proteins in the set share more than 25% sequence identity over a length of more than 80 residues" [7]. With this data set, the sevenfold cross validation test is done [2], [7], [13]. In the sevenfold cross validation test, one subset is chosen for testing and the remaining six subsets are used for training and this process is repeated until all the subsets are chosen for the testing.

C. Encoding Schemes

To train the SVM, a sliding window scheme is applied. In this sliding scheme, a window becomes one training pattern for predicting the structure of the residue at the center of the window. And in this training pattern, the information about the local interactions among neighboring residues can be embedded as a feature value. The feature value of each amino acid residue in a window means the weight (costs) of each residue in a pattern. In this study, several different weight assignment schemes are tested. Among them, the first simplest way is to use the traditional orthogonal encoding which assigns a unique binary vector to each residue, such as $(1, 0, 0 \ldots)$, $(0, 1, 0 \ldots)$, $(0, 0, 1 \ldots)$, and so on. In this method, the weights of all residues in a window are assigned to one equally. This simple orthogonal encoding scheme is used as a reference for comparison with different encoding schemes.

TABLE I
Nonpolar → Polar Distributions of Amino Acid Side Chains, pH 7 (kcal/mol) [14]

Amino Acids	Feature Values	Amino Acids	Feature Values
Isoleucine, I	4.92	Tyrosine, Y	-0.14
Leucine, L	4.92	Threonine, T	-2.57
Valine, V	4.04	Serine, S	-3.40
Proline, P	4.04	Histidine, H	-4.66
Phenylalanine, F	2.98	Glutamine, Q	-5.54
Methionine, M	2.35	Lysine, K	-5.55
Tryptophan, W	2.33	Asparagine, N	-6.64
Alanine, A	1.81	Glutamic acid, E	-6.81
Cysteine, C	1.28	Aspartic acid, D	-8.72
Glycine, G	0.94	Arginine, R	-14.92

Since this simple orthogonal scheme does not give detailed information into the SVM classifiers except the existence of the specific residues around, the new encoding scheme is required to enhance the performance of prediction. For this purpose, a few encoding schemes are designed to incorporate the physicochemical properties of amino acids into the training pattern. However, since this physicochemical property encoding or the simple combination of this encoding with orthogonal vector does not show any improvement, those schemes are discarded at the preliminary screening process.

As another approach to include the physicochemical properties, hydrophobicity property is selected as the main feature among other properties, such as polarity, charge or size. As can be seen from Table I [14], the hydrophobicity can be expressed as the free energy (kilocalories per mole) of transfer of amino acid side chains from cyclohexane to water. In other words, the amino acids with the positive values of free energy in transferring cyclohexane to water are hydrophobic and the ones with negative values are hydrophilic. Based on these values, and with the use of the following function, the hydrophobicity matrix is formulated.

$$\text{Hydrophobicity\_matrix}[i][j] = \frac{\text{abs}(\text{Hydrophob\_ Index}[i] - \text{Hydrophob\_Index}[j])}{20}$$

where the denominator 20 is used to convert the data range into $[0, 1]$.

Next, the BLOSUM62 matrix [15] encoding scheme is applied into SVM. The BLOSUM62 matrix is originally made by Henikoff and Henikoff [15] and this is a measure of differences between two distantly related proteins. Namely, the values in the BLOSUM62 matrix mean "log-odds" scores for the possibility that a given amino acid pair will interchange with each other. In this research, this BLOSUM62 matrix is applied as an encoding scheme by converting its data range to $[0, 1]$.

In addition, the previous schemes are combined together with the following ways and this is to obtain the optimal encoding scheme which offers the most informative feature to predict the secondary structure:

- Orthogonal matrix + Hydrophobicity matrix;
- BLOSUM62 matrix + Hydrophobicity matrix;
- Orthogonal matrix + BLOSUM62 matrix;
- Orthogonal matrix + BLOSUM62 matrix added with the positional information inside a window.

Among the above four combinations, the fourth combination is the same as the third combination except the fact that different weights are applied based on the positions inside a window. In other words, in the third combination, even though each amino acid has 20 different "log-odds" scores, those values are always same regardless of the position inside a sliding window. Therefore, by assigning different weights based on their positions, the machine can be trained with more specific information.

D. Parameter Optimization of the Binary Classifier

The kernel function is selected based on the previous studies [7], [9]. For example, Hua and Sun [7] has proved that the radial basis function (RBF) kernel can provide superior performance in the generalization ability and convergence speed. Therefore, in this study, this RBF kernel, such as the following, is adopted:

$$K(x,y) = e^{-\gamma\|x-y\|^2}.$$

Once the kernel function is selected, the parameter of the kernel function γ and the regularization parameter C are optimized based on the process of the previous studies [7]–[9]. Namely, for the proper choice of C value and γ value of RBF kernel $e^{-\gamma\|x-y\|^2}$, the previous studies tested different γ and upper bound values of C over their own data sets and selected the pairs which show the best accuracy [7]–[9]. Similarly, in this study, different γ and C pairs are tested to find out the optimum parameter values.

E. Binary Classifier Construction

Six SVM binary classifiers, such as three one-versus-rest classifiers ($H/\sim H$, $E/\sim E$, and $C/\sim C$), and three one-versus-one classifiers (H/E, E/C, and C/H) are constructed based on the previous study [7]. Here, the name "one" in one-versus-rest classifier refers to positive class, and the name "rest" means negative class. Likewise, the name "one" in one-versus-one classifier refers to positive class and negative class respectively. For example, the classifier $H/\sim H$ classifies the testing sample as helix or not helix and the classifier E/C classifies the testing sample as sheet or coil.

F. Tertiary Classifier Design

To combine the outputs from the binary classifiers for secondary structure prediction in this research, a new tertiary classifier which combines the results of one-versus-one binary classifiers is designed. And the performance is compared with those of existing tertiary classifiers, including the tree-based classifiers [7], the *s*imple voting classifier which is called "SVM_VOTE" [7], the SVM_MAX_D [7] and the Directed Acyclic Graph (DAG)-based tertiary classifier [9].

In the tree-based tertiary classifier [7], three one-versus-rest binary classifiers ($H/\sim H$, $E/\sim E$, and $C/\sim C$) and three one-versus-one classifiers (H/E, E/C, and C/H) are combined together to form three cascade tertiary classifiers, such as TREE_HEC ($H/\sim H$, E/C), TREE_ECH ($E/\sim E$, C/H), and TREE_CHE ($C/\sim C$, H/E).

In the simple voting tertiary classifier (SVM_VOTE) [7], all six binary classifiers are combined by using a simple voting scheme in which the testing sample is predicted to be state i (i is among H, E, or C) if the largest number of the six binary classifiers classify it as state i. In case the testing samples have two classifications in each state, it is considered to be a coil.

In the SVM_MAX_D classifier [7], the three one-versus-rest classifiers ($H/\sim H$, $E/\sim E$, $C/\sim C$) are combined for handling the multiclass case. And the class of a testing sample (H, E or C) is assigned to the one which presents the largest positive distance from the optimal separating hyperplane (OSH). For example, if the distance values of the each one-versus-rest classifiers ($H/\sim H$, $E/\sim E$, $C/\sim C$) are −1.7, 1.2, and 2.5 respectively, as negative distance of $H/\sim H$ binary classifier does not give any information for decision, only two positive values (1.2, 2.5) are compared and finally, the class for the test sample is assigned to coil based on the largest positive distance.

In the DAG-based tertiary classifier [9], three one-versus-one classifiers (H/E, E/C, and C/H) are combined based on the previous test results [16], [17]. According to the authors [9], one-versus-one classifiers are more accurate than one-versus-rest classifiers, since they handle two data sets with similar sizes. In this scheme [9], if the class is predicted to be E from E/C classifier, H/E classifier is combined. On the other hand, if the class is predicted to be not sheet ($\sim E$) from E/C classifier, C/H classifier is combined to determine the final class.

A new tertiary classifier of this study is similar to the SVM_MAX_D classifier in that the maximum distance is used for the decision. But unlike the SVM_MAX_D classifier, this scheme combines the three one-versus-one binary classifiers (H/E, E/C, and C/H) which give more information than one-versus-rest binary classifiers. In other words, in one-versus-one classifier, both positive and negative values are meaningful to assign a final class but in one-versus-rest classifier, negative value does not provide any specific information for the decision. In this scheme, no matter what the distance values are positive or negative, the classifier with the absolute maximum distance is chosen as the representative classifier for the final decision of the class. And the final class is assigned based on the value of this classifier. 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.

G. Prediction Accuracy Evaluation Methods

There are several standard evaluation methods of secondary structure prediction performance. In this study, Q_3 and SOV [18] are adopted for the performance evaluation, since these are the most widely used assessing methods.

Q_3 is one of the most commonly used performance measure in the protein secondary structure prediction and it refers to the

TABLE II
TESTING ACCURACY BASED ON DIFFERENT WINDOW LENGTH

Window Size	5	7	9	11	13	17	19	s*
Binary Classifier								
H/~H	73.4	75.3	75.9	76.5	76.8	76.7	76.6	13
E/~E	79.4	79.9	80.4	80.7	80.7	80.7	80.7	11
C/~C	69.8	70.3	70.6	70.5	70.6	70.4	70.4	13
H/E	71.2	73.1	74.0	74.8	75.5	75.7	75.7	17
E/C	75.5	75.8	75.8	75.7	76.0	76.3	76.3	17
C/H	71.9	73.9	74.3	74.7	75.1	75.2	74.9	17

The results are on the RS126 with the orthogonal encoding. The s^* value is the optimal window length for each binary classifier. Combined results of sevenfold cross validation are shown.

three-state overall percentage of correctly predicted residues. This measure is defined as

$$Q_3 = \frac{\sum\limits_{i \in \{H,E,C\}} \#\text{ of residues correctly predicted}_i}{\sum\limits_{i \in \{H,E,C\}} \#\text{ of residues in class } i} \times 100.$$

Based on the above equation, the per-residue accuracy for each type of secondary structure (Q_H, Q_E, Q_C) can be obtained as

$$Q_i = \frac{\#\text{ of residues correctly predicted in state } i}{\#\text{ of residues in state } i} \times 100$$
$$i \in \{H, E, C\}.$$

SOV was developed by Rost *et al.* [19] and modified by Zemla *et al.* [18] to evaluate the quality of a prediction in a more realistic manner by assessing the prediction by segment. SOV is calculated as follows [18]:

$$\text{SOV} = \frac{1}{N} \sum_{i \in \{H,E,C\}} \sum_{s(i)} \left[\frac{\min \text{ov}(s_1, s_2) + \delta(s_1, s_2)}{\max \text{ov}(s_1, s_2)} \times \text{len}(s_1) \right] \times 100$$

where

N normalization value;
$S(i)$ set of all overlapping pairs of segments (s_1, s_2) in state i;
len(s_1) number of residues in segment s_1,
$\min \text{ov}(s_1, s_2)$ length of the actual overlap;
$\max \text{ov}(s_1, s_2)$ total extent of the segment, and $\delta(s_1, s_2)$ is given as

$$\delta(s_1, s_2) = \min \left\{ (\max \text{ ov}(s_1, s_2) - \min \text{ ov}(s_1, s_2)), \min \text{ ov}(s_1, s_2), \text{int}\left(\frac{\text{len}(s_1)}{2}\right), \text{int}\left(\frac{\text{len}(s_2)}{2}\right) \right\}.$$

SOV94 [19] and SOV99 [18] are different in the definition of δ and the normalization factor N. In this study, SOV99 is adopted for RS126 to compare the results with SVMpsi [9], which claims the best performance in the protein secondary structure prediction.

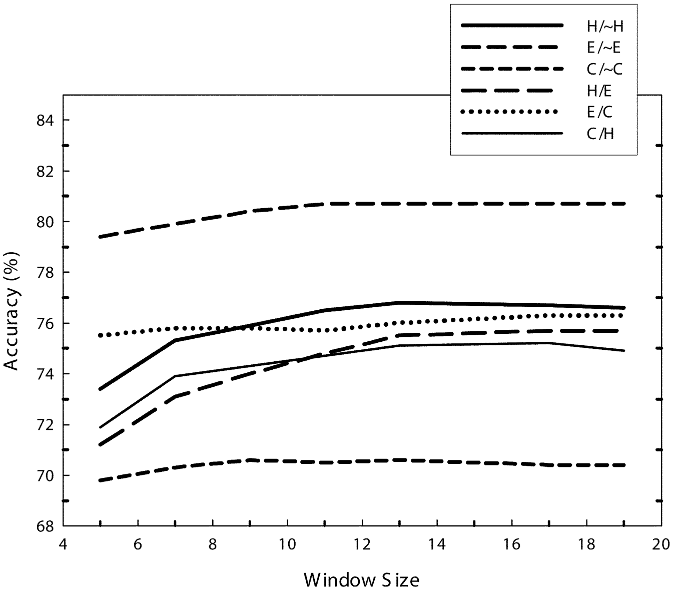


Fig. 1. Testing accuracy based on different window lengths.

III. Experimental Results

A. Window Size Optimization

The optimal window size of the sliding window scheme is obtained by testing different window lengths from 5 to 19. For this optimization, RS126 data set is used with the orthogonal encoding scheme. And via sevenfold cross validation test, the window length 13 is adopted as an optimal window size for all six binary classifiers.

In Table II, testing accuracy based on different window lengths is shown. The optimal window length (s^*) for each binary classifier is determined on the RS126 set using the orthogonal encoding scheme. As can be seen from Table II and Fig. 1, for all six binary classifiers, once the window length is over 13, accuracy values converge. While all three one-versus-one binary classifiers show the highest accuracy in window size 17, the improvement is less than 0.3% when compared with the case of size 13. Therefore, in this study, the

TABLE III
ACCURACY COMPARISON OF BINARY CLASSIFIERS WITH OTHER METHODS

Binary Classifier	RS126		
	SVMfreq	SVMpsi	SVMob
H/~H	80.4	87.5	78.8
E/~E	81.3	86.3	80.9
C/~C	73.2	77.9	70.7
H/E	76.7	81.9	76.6
E/C	77.6	85.0	76.5
C/H	80.9	90.2	75.8

The results of SVMfreq are from Hua and Sun [7] and the SVMpsi results are obtained by PSI-BLAST profiles [9]. SVMob is the new profile of this study with the combined matrix of orthogonal and BLOSUM62.

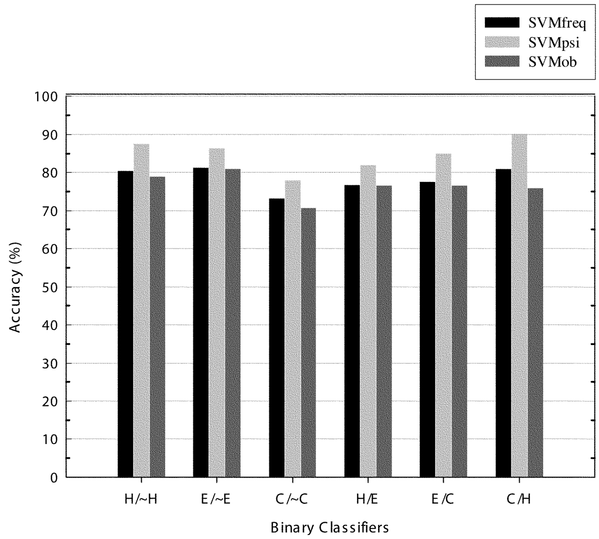


Fig. 2. Accuracy comparison of binary classifiers with other methods.

window length 13 is adopted as an optimal window size for all six binary classifiers.

B. Encoding Schemes

To determine the most informative encoding scheme, several encoding schemes, including orthogonal matrix, hydrophobicity matrix, BLOSUM62 matrix, and the combination of these matrices, are tested. According to the test results, the combined matrix of orthogonal and BLOSUM62 matrix shows 78.8% accuracy and this value is the highest one among all the encoding schemes. Also, this accuracy is 2% higher than the result of standard orthogonal encoding scheme. Therefore, this combined matrix of orthogonal and BLOSUM62 matrix is adopted to train all six binary classifiers.

C. Binary Classifiers

Based on the orthogonal and BLOSUM62 matrix combined encoding, six different binary classifiers are trained with sevenfold cross validation method. And the results are compared with other methods which use the same data set but different encoding schemes. In Table III and Fig. 2, the SVMfreq is the scheme which adopts the frequency matrix with multiple sequence alignments as the encoding profile [7]. The SVMpsi is the scheme which applies the PSSM obtained by PSI-BLAST searches as the encoding profile [9]. SVMob is the new encoding scheme of this study which adopts the combined matrix of orthogonal and BLOSUM62 matrix.

TABLE IV
ACCURACY OF TERTIARY CLASSIFIERS ON THE RS126 DATA SET

Tertiary Classifier	Q_3 (%)	Q_H (%)	Q_E (%)	Q_c (%)	SOV99 (%)
TREE_HEC	63.2	51.0	45.2	79.9	63.7
TREE_ECH	62.3	62.4	26.2	79.0	56.7
TREE_CHE	61.2	64.8	47.3	65.2	64.4
SVM_VOTE	68.3	62.9	40.7	84.8	56.1
SVM_MAX_D	63.2	61.0	40.1	75.5	53.1
DAG	63.2	59.2	41.6	76.0	53.7
SVM_REPRESNT.	78.8	77.5	59.1	88.9	71.1

Combined results of sevenfold cross validation are shown.

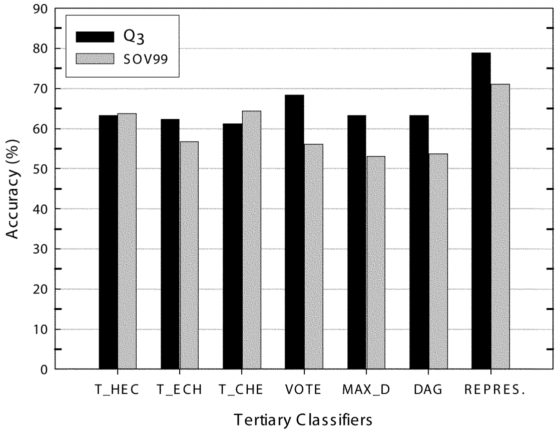


Fig. 3. Accuracy of tertiary classifiers on the RS126 data set.

As can be noticed from Table III and Fig. 2, the SVMpsi shows the best performance for all six binary classifiers and the SVMfreq represents the similar performance with SVMob, but slightly better. The accuracy of SVMob is about 5%–14% lower than SVMpsi and it is about 0.4%–5% lower than SVMfreq on RS126 data set. This fact infers that the encoding scheme of SVMpsi or SVMfreq might be a better choice for the performance of binary classifiers.

D. Tertiary Classifiers

In Table IV and Fig. 3, the performance of tertiary classifiers is compared by using the two most typical accuracy measures of Q_3 and SOV. According to the result of Table IV, the tertiary classifier of this study, SVM_REPRESENT, records the best performance among all. It provides Q_3 accuracy of 78.8% and SOV of 71.1%. Compared with the result of this study, the tree-based schemes, such as TREE_HEC, TREE_ECH and TREE_CHE, or the DAG-based scheme shows about 15% lower Q_3 accuracy. This is probably due to the fact that in these schemes, since the binary classifiers are combined with

TABLE V
ACCURACY COMPARISON WITH OTHER RESEARCH RESULTS ON THE RS126 DATA SET

Method	Q_3 (%)	Q_H (%)	Q_E (%)	Q_c (%)	SOV94 (%)	SOV99 (%)
PHD	70.8	72.0	66.0	72.0	73.5	-
SVMfreq	71.2	73.0	58.0	73.0	74.6	-
SVMpsi	76.1	77.2	63.9	81.5	79.6	72.0
SVMob	78.8	77.5	59.1	88.9	-	71.1

Combined results of sevenfold cross validation are shown. PHD result is obtained by Rost and Sander [1] and Rost *et al.* [19] and SVMfreq result is obtained by Hua and Sun [7]. and SVMpsi result is obtained by Kim and Park [9]. SVMob is the new method proposed by this study.

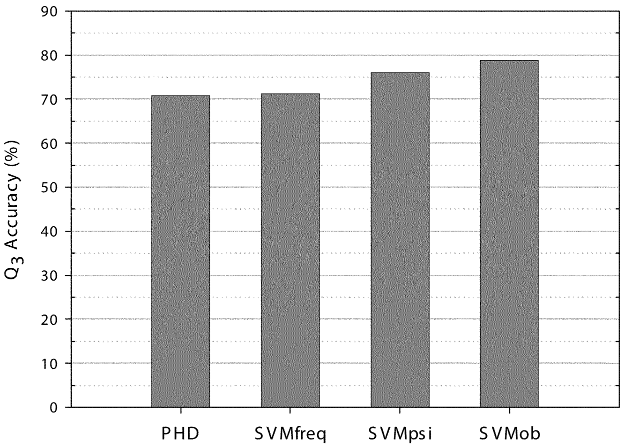


Fig. 4. Accuracy comparison with other research results on the RS126 data set.

multilayers, the error of upper layer binary classifier (false positive and false negative) can be propagated to the next layer.

In this respect, the fact that the accuracy of SVM_VOTE scheme is higher than these multilayer schemes can be explained. For the reason of the low accuracy of SVM_MAX_D scheme, since it combines three one-versus-rest binary classifiers ($H/\sim H$, $E/\sim E$, $C/\sim C$), it might not catch the useful information for decision. Namely, since these binary classifiers deal with two data sets with very different sizes, these are less accurate than one-versus-one classifiers. Moreover, the final classification depends only on the positive value of each binary classifier because the negative value cannot give any information to make a decision.

One more interesting thing to notice in Table IV is that there is no trend in SOV values among tertiary classifiers, such as the one found in Q_3 values. In other words, the order of SOV measure does not match with the order of Q_3 accuracy. According to the result of the previous study [8], high Q_3 accuracy does not always guarantee high SOV value. Therefore, we could conclude that SOV measure should be interpreted as an independent measure without finding some relationship with Q_3 accuracy and without trying to find some patterns among different classifiers.

IV. DISCUSSION

A. Result Comparison With Other Research

Table V and Fig. 4 show the result of accuracy comparison with other research. Here, PHD results are obtained by Rost and Sander [1] and Rost *et al.* [19], SVMfreq results are obtained by Hua and Sun [7], and SVMpsi results are obtained by Kim and Park [9]. SVMob is the new method proposed by this study in which orthogonal and BLOSUM62 combined matrix is applied as an encoding scheme.

When we recall the previous result of accuracy comparison with binary classifiers (Table III), the result of accuracy comparison with tertiary classifiers is quite different. Even though the performance of combined matrix of this study over the binary classifiers is not satisfactory, with the use of the new tertiary classifier, the Q_3 accuracy increases noticeably. In other words, the Q_3 accuracy of SVMob is the best among all on the RS126 data set and even 2.7% higher than that of the SVMpsi which claims the best performance so far. Also, when the SOV99 is applied, the performance of SVMob shows comparable performance with SVMpsi.

B. Potential Improvements

As can be seen from the previous performance comparison based on six binary classifiers, the encoding scheme of this study, the combined matrix of orthogonal and BLOSUM62 matrix, is not satisfactory. Also, we can see that the best result is obtained by the SVMpsi scheme, which applied PSSM as the encoding.

Therefore, if this PSSM encoding scheme is applied for the binary classifiers and if the results of these binary classifiers are combined with the tertiary classifier of this study, more improvement in performance could be expected.

As another problem, the training time of binary classifier should be mentioned. On a modern Linux server, a sevenfold cross validation for the RS126 data set took around two days. And for the CB513 set which was planned to be tested, due to time and processing power constraints, the training was not successful. Therefore, the parallelization of the numerical computing for SVM should be considered to resolve the problem related to the training speed. Otherwise, as the number of data set increases, this problem would be worse.

V. CONCLUSION

In this study, SVM learning machine is applied for the improvement of the prediction accuracy of the protein secondary structure. For this purpose, two new approaches are adopted. The first one is to optimize the encoding scheme for binary classifiers and the second one is to design a new tertiary classifier to combine the results of binary classifiers.

For the first approach, several different encoding schemes are applied and optimized. And the optimal window size for six SVM binary classifiers is set up by testing with different window lengths. Also, proper kernel function is selected based on the results of the previous studies and its parameter and the regularization parameter are optimized.

For the second approach, a new tertiary classifier which combines the results of one-versus-one binary classifiers is designed and its efficiency is compared with the existing tertiary classifiers.

Based on the result of performance comparison with previous studies, the optimized encoding scheme of this study, the combined matrix of orthogonal and BLOSUM62 matrix, showed lower performance than that of SVMfreq or SVMpsi. However, by applying the new tertiary classifier of this study, the performance is enhanced noticeably. Namely, the final Q_3 accuracy of this study is 2.7% higher than the result of SVMpsi which claims the highest accuracy so far. Even in the SOV value, the result of this study shows the comparable performance with SVMpsi.

The tertiary classifier designed for our research has immediate application in other areas where the tertiary classification can be decomposed into a set of binary classifications. This scheme could improve performance in many other areas such as pattern recognition, data mining, and machine learning.

Acknowledgment

The authors would like to thank Prof. T. Joachims for making SVMlight software available and Prof. B. Rost and C. Sander for providing the RS126 data set.

References

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

[2] S. K. Riis and A. Krogh, "Improving prediction of protein secondary structure suing structured neural networks and multiple sequence alignments," *J. Comput. Biol.*, vol. 3, pp. 163–183, 1996.

[3] J. M. Chandonia and M. Karplus, "New methods for accurate prediction of protein secondary structure," *Proteins*, vol. 35, pp. 293–306, 1999.

[4] V. Vapnik and C. Cortes, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–293, 1995.

[5] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*. Cambridge, U.K.: Cambridge Univ. Press, 2000.

[6] K. K. Chin, "Support vector machines applied to speech pattern classification," M.Phil. thesis, Cambridge Univ., Cambridge, U.K., 1999.

[7] 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.

[8] J. Casbon, "Protein secondary structure prediction with support vector machines," M.Sc. thesis, Univ. Sussex, Brighton, U.K., 2002.

[9] H. Kim and H. Park, "Protein secondary structure prediction based on an improved support vector machines approach," Dept. Comput. Sci. Eng., Univ. Minnesota, Minneapolis, 2003.

[10] C. 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] F. M. Richards and C. E. Kundrot, "Identification of structural motifs from protein coordinate data: Secondary structure and first-level super secondary structure," in *Proteins*, 1988, vol. 3, pp. 71–84.

[12] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," in *Proteins*, 1995, vol. 23, pp. 566–579.

[13] B. Rost and C. Sander, "Combining evolutionary information and neural networks to predict protein secondary structure," *Proteins*, vol. 19, pp. 55–72, 1994.

[14] A. Radzicka and R. Wolfenden, "Comparing the polarities of the amino acids: Side-chain distribution coefficients between the vapor phase, cyclohexane, 1-Octanol, and neutral aqueous solution," *Biochemistry*, vol. 27, pp. 1664–1670, 1988.

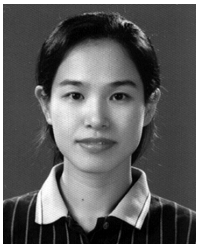
[15] S. Henikoff and J. G. Henikoff, "Amino acid substitution matrices from protein blocks," *Proc. Nat. Acad. Sci.*, vol. 89, pp. 10 915–10 919, 1992.

[16] C. C. Chang and C. J. Lin. (2003) LIBSVM. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[17] M. Heiler, "Optimization criteria and learning algorithms of large margin classifiers," Diploma thesis, Univ. Mannheim, Mannheim, Germany, 2002.

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

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



Hae-Jin Hu was born in Seoul, Korea. She received the M.S. degree in chemical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejon, in 1993 and the M.S. degree in computer science from Georgia State University, Atlanta, in 2003. She is currently working toward the Ph.D. degree in the Department of Computer Science, Georgia State University.

Her main research interests include protein engineering and molecular biology.

Yi Pan (S'90–SM'91) received the B.Eng. and M.Eng. degrees in computer engineering from Tsinghua University, Beijing, China, in 1982 and 1984, respectively, and the Ph.D. degree in computer science from the University of Pittsburgh, Pittsburgh, PA, in 1991.

He is currently a Professor in the Department of Computer Science, Georgia State University, Atlanta. He has published more than 80 journal papers. In addition, he has published over 90 papers in refereed conferences. He has also coedited 13 books (including proceedings) and contributed several book chapters. He has served as an editor-in-chief or editorial board member for eight journals and a guest editor for seven special issues. His research interests include parallel and distributed computing, optical networks, wireless networks, and bioinformatics. His recent research has been supported by the National Science Foundation, the National Institutes of Health, the Air Force Office of Scientific Research, the Air Force Research Laboratory, the Japan Society for Promotion of Science, the International Information Science Foundation, and the states of Georgia and Ohio.

Dr. Pan is listed in *Men of Achievement*, *Who's Who in the Midwest*, *Who's Who in America*, *Who's Who in American Education*, *Who's Who in Computational Science and Engineering*, and *Who's Who of Asian Americans*. He is an IEEE Distinguished Speaker (2000–2002), a Yamacraw Distinguished Speaker (2002), and a Shell Oil Colloquium Speaker (2002). He has delivered over 40 invited talks, including keynote speeches and colloquium talks, at conferences and universities worldwide. He has published 25 papers in various IEEE journals. He has served as an editor-in-chief or editorial board member for three IEEE Transactions.

Robert Harrison received the B.S. degree from Pennsylvania State University, University Park, in 1979 and the Ph.D. degree from Yale, New Haven, CT, in 1985.

He is an active researcher in computational chemistry and bioinformatics and currently an Associate Professor in computer science at Georgia State University, Atlanta.

Phang C. Tai received the Ph.D. degree in microbiology from the University of California, Davis.

He has done postdoctoral work at Harvard Medical School, Cambridge, MA, and is currently a Regents' Professor and Chair of the Department of Biology, Georgia State University, Atlanta. His research interest is in molecular biology and microbial physiology. His current research focuses on the mechanism of protein secretion across membranes.