Protein Secondary Structure Prediction Based on an Improved Support Vector Machines Approach

Hyunsoo Kim and Haesun Park *

Dept. of Computer Science and Engineering
Univ. of Minnesota, Minneapolis, MN 55455, U.S.A.

September, 2002
Revised January and June, 2003

Abstract

The prediction of protein secondary structure is an important step in the prediction of protein tertiary structure. A new protein secondary structure prediction method SVMpsi is developed to improve the current level of prediction by incorporating new tertiary classifiers and their jury decision system, and the PSI-BLAST PSSM profiles. Additionally, efficient methods to handle unbalanced data and a new optimization strategy for maximizing the Q_3 measure are developed. The SVMpsi produces the highest published Q_3 and SOV94 scores on both the RS126 and CB513 data sets to date. For a new KP480 set, the prediction accuracy of SVMpsi was Q_3 = 78.5% and SOV94 = 82.8%. Moreover, the blind test results for 136 non-redundant protein sequences which do not contain homologues of training data sets, were Q_3 = 77.2% and SOV94 = 81.8%. The SVMpsi results in CASP5 illustrate that it is another competitive method to predict protein secondary structure.

Keywords: protein structure prediction / secondary structure / support vector machines / PSSM / directed acyclic graph scheme

*To whom correspondence should be addressed. This work was supported in part by the National Science Foundation grant CCR-0204109. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF). A part of this work was carried out while the authors were visiting the Korea Institute for Advanced Study, Seoul, Korea, from January 2002 to August 2002. Email: hpark@cs.umn.edu

Introduction

The study of protein secondary structure plays an important role in protein tertiary structure prediction with the ab-initio method or protein fold recognition by providing additional constraints (Baldi *et al.*, 1999; Baker and Sali, 2001; Russell *et al.*, 1996). Many methods have been applied to predict secondary structure solely from the protein sequence, including DSC (King and Sternberg, 1996) based on linear discrimination, NNSSP (Salamov and Solovyev, 1995) based on the k-way nearest-neighbor method, PREDATOR (Frishman and Argos, 1997) based on the internal pair wise alignment method rather than a global multiple alignment, and PSIPRED (Jones, 1999) & Jnet (Cuff and Barton, 2000) based on neural networks.

A support vector machine (SVM) constructs an optimal separating hyperplane which maximizes the margin (i.e. the distance between the hyperplane and the nearest data point of each class) by mapping the input space into a high dimensional feature space. The mapping is determined by a kernel function. Training with SVMs has crucial advantages including fast convergence, typically about 1-2 order of magnitude faster than neural networks (NNs) (Ding and Dubchak, 2001), tending not to over fit, and the ability to find the problem formulation as a quadratic convex function minimization that is easier to solve (Burges and Schölkopf, 1997; Burges, 1998; Cristianini and Shawe-Taylor, 2000; Osuna *et al.*, 1997; Vapnik, 1995; Vapnik, 1998). The previous study for secondary structure prediction using support vector machines (Hua and Sun, 2001) achieved good results by using the frequency profiles with evolutionary information and removing the influence of noise and outliers by discarding a fraction of samples which are hard to predict because they are located near the optimal separating hyperplane. However, the prediction level is not sufficient to favorably compare with the recent results of the neural network approaches.

The recent approaches based on neural networks, for example PSIPRED and Jnet, have been successfully advanced by PSI-BLAST PSSM (position-specific scoring matrix) profiles (Jones, 1999) derived from sequences that have remote similarities, by an iterative strategy. Jones (Jones, 1999) expected that other secondary structure prediction methods will show measurable improvements in accuracy by using PSI-BLAST profiles instead of using the multiple sequence alignment approach, and Hua and Sun (Hua and Sun, 2001) already pointed out that it is possible to achieve significant improvement by incorporating PSI-BLAST generated profiles in the SVMs approach, as well.

In this paper, we show the improvement of prediction accuracy by new tertiary classifiers and their jury decision system, efficient methods to handle unbalanced data, and a new optimization strategy for support vector machines that maximizes the Q_3 measure. We also apply this to PSI-BLAST profiles, in order to improve the current prediction level and to show that the support vector machine approach is a valid method for secondary structure prediction. We also investigate a new way to reduce the

influence of noise and outliers by using the theoretical relationships in the soft margin support vector machine. The training sets with an unbalanced number of data items in each class can produce an ill-balanced binary classifier that may have low recall for the smaller class. If we use an ill-balanced binary classifier, it may not produce a good final prediction result in spite of high prediction accuracy in each binary classifier, which constitutes the cascaded tertiary classifier. We adopted the one-versus-one scheme and directed acyclic graph (DAG) scheme (Heiler, 2002) for handling three class problems since these demonstrate better performance results for multi-classifications (Hsu and Lin, 2002). We built the jury decision system for the all designed tertiary classifiers to get better prediction accuracy. Here, we will show that SVMpsi can achieve the most accurate published Q_3 and SOV94 scores on the RS126 (Rost and Sander, 1993) and CB513 (Cuff and Barton, 1999) data sets. In the fifth critical assessment of structure prediction (CASP5) experiment, we predicted the most accurate structure for 5 proteins compared to the other groups. The average Q_3 and SOV_3 scores for SVMpsi were 79.10% and 79.38%, respectively. The results demonstrate that SVMpsi is one of the most promising methods for protein secondary structure prediction.

Materials and Methods

Vector Space Representation of Proteins

The secondary structure is assigned from the experimentally determined tertiary structure by DSSP (Kabsch and Sander, 1983), STRIDE (Frishman and Argos, 1995), or DEFINE (Richards and Kundrot, 1988). We use DSSP since it has been the most widely used secondary structure definition. It has eight secondary structure classes: H(α-helix), G(3_{10}-helix), I(π-helix), E(β-strand), B(isolated β-bridge), T(turn), S(bend) and - (rest). Then reduction from eight classes to three states of helix (**H**), sheet (**E**), and coil (**C**) is done by using one of the following methods:

(1) H,G and I to **H** ; E to **E** ; all other states to **C**
(2) H,G to **H** ; E,B to **E** ; all other states to **C**
(3) H,G to **H** ; E to **E** ; all other states to **C**
(4) H to **H** ; E,B to **E** ; all other states to **C**
(5) H to **H** ; E to **E** ; all other states to **C**

The 8-to 3-state reduction method can alter the apparent prediction accuracy (Cuff and Barton, 1999). Though we can expect an accuracy increase by using method (5), we used different methods for different data sets to provide fair comparison of our results to other methods. The details are discussed in the subsection where we present our data sets.

We tested protein secondary structure prediction using PSI-BLAST profiles and designed classifiers for the three cluster problem based on the binary classifiers generated

by SVMs. The final position-specific scoring matrices from PSI-BLAST against the SWALL (Bairoch and Apweiler, 2000) non-redundant protein sequence database are used. We applied PFILT (Jones *et al.*, 1994; Jones and Swindells, 2002) to mask out regions of low complexity sequences, the coiled coil region, and transmembrane spans. For PSI-BLAST, the E-value threshold for inclusion of 0.001 and three iterations were applied to search the non-redundant sequence database.

The position specific scoring matrix has $20 \times N$ elements, where N is the length of the target sequence, and each element represents the log-likelihood of a particular residue substitution based on a weighted average of BLOSUM62 (Henikoff and Henikoff, 1992) matrix scores for a given alignment position in the template. The profile matrix elements in the range [-7,7] are scaled to the [0,1] range by using the following function:

$$f(x) = \begin{cases} 0.0 & \text{if } x \leq \text{-5} \\ 0.5 + 0.1x & \text{if -5} < x < 5 \\ 1.0 & \text{if } x \geq 5, \end{cases} \tag{1}$$

where x is the value from the raw profile matrix. We selected the above function after testing various scale functions to maximize the Q_3 score. As in the PHD coding scheme, we used a sliding window method (Qian and Sejnowski, 1988; Rost and Sander, 1993). In order to allow a window to extend over the N-terminus and the C-terminus, an additional 21st unit was appended for each residue. Therefore, each input vector has $21 \times w$ components, where w is the sliding window size. The window is shifted residue by residue through the protein chain. We constructed three one-versus-rest classifiers, each of which determines whether the secondary structure of the residue is a particular secondary state or not (H/∼H, E/∼E, C/∼C), and three one-versus-one classifiers (H/E, E/C, and C/H).

Prediction Accuracy Assessment

Several standard performance measures were used to assess prediction accuracy. Q_3 is a measure of the three-state overall percentage of correctly predicted residues:

$$Q_3 = \frac{\sum_{i \in \{H,E,C\}} \#\text{ of residues correctly predicted}_i}{\sum_{i \in \{H,E,C\}} \#\text{ of residues in class } i} \times 100. \tag{2}$$

The correlation coefficient (C_H, C_E, C_C) introduced by Matthews (Matthews, 1975) is

$$C_i = \frac{p_i r_i - u_i o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(r_i + u_i)(r_i + o_i)}}, \tag{3}$$

where p_i is the number of correctly predicted residues in conformation, r_i the number of those correctly rejected, u_i the number of the incorrectly rejected (false negative),

and o_i that of the incorrectly predicted to be in the class (false positive), for i = H,E,C. The per residue accuracy (Q_H, Q_E, Q_C; Q_H^{pre}, Q_E^{pre}, Q_C^{pre}) for each type of secondary structure (Hua and Sun, 2001) was also calculated as:

$$Q_i(\%) = \frac{\#\text{ of residues correctly predicted}_i}{\#\text{ of residues in class } i} \times 100, \tag{4}$$

and

$$Q_i^{pre}(\%) = \frac{\#\text{ of residues correctly predicted}_i}{\#\text{ of residues predicted}_i} \times 100, \tag{5}$$

where conformation state i can be H, E, or C.

The segment overlap measure (SOV) is a measure for evaluation of secondary structure prediction methods by secondary structure segment rather than individual residues (Rost and Sander, 1994; Zemla *et al.*, 1999). SOV is calculated as

$$SOV = \frac{1}{N} \sum_{i \in \{H,E,C\}} \sum_{S(i)} \left[\frac{minov(s_1, s_2) + \delta}{maxov(s_1, s_2)} \times len(s_1) \right] \times 100, \tag{6}$$

where $S(i)$ is the set of all overlapping pairs of segments (s_1, s_2) in conformation state i, $len(s_1)$ is the number of residues in segment s_1, $minov(s_1, s_2)$ is the length of the actual overlap and $maxov(s_1, s_2)$ is the total extent of the segment. The quality of match of each segment pair is taken as a ratio of the overlap of the two segments $minov(s_1, s_2)$ and the total extent of that pair $maxov(s_1, s_2)$. The definition of δ and the normalization factor N is different between SOV94 (Rost and Sander, 1994) and SOV99 (Zemla *et al.*, 1999). We calculated SOV94 for RS126 and CB513 to compare the results since PHD (Rost and Sander, 1994), PSIPRED (Jones, 1999) and SVMfreq (Hua and Sun, 2001) methods used SOV94.

Training and Testing Data Sets

For comparing our new results to some previously published results (Hua and Sun, 2001) that used a frequency based coding scheme, we selected non-homologous RS126 and CB513 data sets. The results show that the PSI-BLAST profiles are also helpful in improving accuracy in the SVM approach. The CB513 set includes the CB396 data set and almost all proteins of RS126 except 9 homologues for which the SD significance score is higher than 5 (Cuff and Barton, 1999). The SD score is a more stringent measure of sequence similarity than the percentage identity since it corrects for bias due to the length and composition of sequences.

We prepared a data set of 480 proteins by removing proteins from CB513 that have shorter than 30 residues and those that contained only a few sequences in the first iteration of PSI-BLAST. The 16 proteins that are shorter than 30 residues are removed since it has been shown that they do not have well defined secondary structure (Cuff

and Barton, 1999). The prepared KP480 data set may not be the same as the 480 data set of Jnet (Cuff and Barton, 2000), although they are generated from CB513 by removing proteins that are shorter than 30 residues. Each data set is divided into seven folds that have a similar number of proteins and similar composition of the secondary structure to perform cross validation tests.

Besides the cross-validation tests for accessing the performance of the prediction method, we prepared a blind test set of 136 protein sequences that were not used in the training set. The test set was prepared using a structural similarity criterion so that it does not have any protein that is contained in the same fold family, i.e. the CATH (Orengo *et al.*, 1997) T-level, with the CB513 training set. Each protein sequence of the test set represents unique protein folds. Only highly resolved structures (resolution < 1.8Å) of which the length is more than 60 residues and less than 600 residues were included in the blind test set. The structural similarity criterion is more stringent than the SD score which is a measure of pairwise sequence similarity (Cuff and Barton, 1999; Jones, 1999). Thus, there is no pair of similar sequences between the training and blind test sets. We used 8-to 3-state reduction method (2) for the RS126 data set to provide a fair comparison of our results to the other methods such as PHD (Cuff and Barton, 1999), DSC, PREDATOR, NNSSP and their consensus method (Cuff and Barton, 1999), although PHD (Rost and Sander, 1993) and SVMfreq (Hua and Sun, 2001) methods based on frequency profiles used the reduction method (1). The 8-to-3 reduction method (4) was used for the KP480 set for the comparison with the Jnet result based on the 8-to-3 reduction method (4) (H to H, E,B to E, all other states to C). In Jnet, the 3_{10}-helix was removed since it represents a weak 1 kcal/mol hydrogen bond so that it does not represent core secondary structure. We adopted the 8-to-3 reduction method (2) for the 7-fold cross validation test of the CB513 data set and for the blind test of 136 non-redundant sequences, which is one of the most widely used, and to compare the prediction performance of our method to other methods.

Results

Parameter Optimization of the Prediction System

In an L_1 soft margin support vector machine (Vapnik, 1995; Vapnik, 1998), we need to select a kernel function and the regularization parameter C in each binary classifier, to construct a classifier for multiple classes. The primal formulation of the soft-margin SVMs maximize margin and minimize training error simultaneously by solving the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},\xi_i} \quad & \frac{1}{2}\mathbf{w}^t\mathbf{w} + C\sum_{i=1}^{n}\xi_i \\ s.t. \quad & y_i\left[\mathbf{w}^t\mathbf{x}_i + b\right] \geq 1-\xi_i,\ \xi_i \geq 0, \quad i = 1,\ldots,n, \end{aligned} \tag{7}$$

where $\mathbf{x}_i$ represents an input vector, $y_i = \pm 1$ according to whether $\mathbf{x}_i$ is in the positive or negative class, n is the number of the training data, and C is a parameter that controls the trade-off between margin and classification error represented by slack variables ξ_i's. The separating hyperplane in a mapped high dimensional feature space can be represented as $\mathbf{w}^t\varphi(\mathbf{x}) + b = 0$, where $\mathbf{w}$ is the solution of the primal formulation and $\varphi(\cdot)$ is a nonlinear function which maps the input space into a higher dimensional space.

The corresponding dual quadratic programming problem with the application of a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ can be written as

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \\ s.t. \quad & \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \le \alpha_i \le C, \quad i = 1, \ldots, n, \end{aligned} \tag{8}$$

where α_i are the solutions of the dual formulation. The dual formulation of the soft margin SVM with control parameter C shows that the influence of a single training example is limited by C. Our substantial tests show that the RBF (radial basis function) kernel, defined as,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2) \tag{9}$$

is appropriate for complex classification problems, when parameters γ and C are selected from the optimization process.

Since multi-class classification is based on binary classifiers in the support vector machine, the criteria for selecting the optimal parameters γ and C in each binary classifier play a critical role. A common practice is to choose the parameters that maximize the accuracy (i.e. maximize the number of correct predictions) in each binary classifier. Certainly the optimization criteria should depend on the performance measure of the final results, which we would like to optimize. In the case of the protein secondary structure prediction, two of the most commonly used performance measures are Q_3 and SOV_3. Using the example of Q_3 measure and the three class classifier that is built upon two binary classifiers, which determine the membership in H/~H (H vs. not H) in Step 1 and E/C (E vs. C) in Step 2, we now illustrate this point further. The total number t of training data items and Q_3 can be represented as

$$Q_3 = (p_H + p_E + p_C)/t, \quad \text{where} \quad t = p_H + r_H + u_H + o_H$$

using the notation introduced before. To reflect the fact that the value of Q_3 depends on the results from Steps 1 and 2, Q_3 may be rewritten in various ways including

$$Q_3 \quad = \quad \frac{p_H}{u_H + p_H} \times \frac{u_H + p_H}{t} + \frac{p_E + p_C}{t} \tag{10}$$

$$
\begin{aligned}
&= \frac{p_H}{t} + \frac{u_H + r_H}{t} \times \frac{p_E + p_C}{u_H + r_H} \\
&= \frac{p_H}{t} + \frac{\#(\sim H)}{t} \times \frac{p_E + p_C}{\#(\sim H)},
\end{aligned}
$$

where $\#(\sim H)$ denotes the number of data items not in H. The difficulty comes from the fact that the result in Step 2 depends on the result from Step 1 and there is no easy way to reflect this in the expression for Q_3. We have chosen our optimized parameters by fine tuning based on accuracy, recall, and the precision of each step. The recalls (R) and precision (P) for H and ∼H in Step 1 defined as

$$
\begin{aligned}
R_H &= \frac{p_H}{p_H + u_H},\ R_{\sim H} = \frac{r_H}{r_H + o_H}, \\
P_H &= \frac{p_H}{p_H + o_H},\ P_{\sim H} = \frac{r_H}{r_H + u_H}.
\end{aligned} \tag{11}
$$

Unlike in query processing, it is important to consider recalls and precision of both positive and negative classes in the classification. The optimized parameters chosen based on the results of each step and in both binary classifiers, are $\gamma = 0.05$ and $C = 1.0$ on RS126 set, $\gamma = 0.05$ and $C = 2.0$ on the KP480 set, and $\gamma = 0.05$ and $C = 2.5$ on the CB513 data set for the PSI-BLAST profiles.

In a soft margin SVM, the support vectors satisfy the following relationships:

$$
\begin{aligned}
0 < \alpha_i < C &\longleftrightarrow \text{SV with } \xi_i = 0 \\
\alpha_i = C &\longleftrightarrow \text{SV with } \xi_i > 0 \\
\alpha_i = 0 &\quad\ \text{else.}
\end{aligned} \tag{12}
$$

where SV means a support vector and a training point $\mathbf{x}_i$ is the support vector only when $\alpha_i \neq 0$. To reduce the influence of noise and outliers, after finding the support vectors from the training stage, support vectors that are close to the optimal separating hyperplane or on the other side of the hyperplane can be partly or totally removed by ignoring those with $\alpha_i = C$ in the retraining stage. However, there was no significant effect of this strategy on the prediction results when our encoding scheme based on the PSI-BLAST was used, which shows that the PSI-BLAST profiles in secondary structure prediction is robust in the presence of noise and outliers.

Optimal Window Length for Binary Classifiers

The optimal window length of the sliding window coding scheme was obtained by testing the accuracy for the various window sizes. When the window size is too short, it may lose some important classification information and prediction accuracy, while a too long window size may suffer from inclusion of unnecessary noise. For convenience, we call our method using PSI-BLAST profiles SVMpsi and Hua and Sun's method (Hua

and Sun, 2001) based on the frequency profiles approach SVMfreq. Table 1 shows that the optimal window length of SVMpsi is much longer than that of SVMfreq on the RS126 data set. The prediction accuracy for the binary classifier does not dramatically change when the window length over 15 is used, which shows that the SVMpsi method effectively dealt with noise. We chose the window length of 15 for all results in this paper. It is slightly smaller than the sliding window length 17 of the first layer neural network for sequence to structure prediction in Jnet (Cuff and Barton, 2000).

Tertiary Classifier Design

There are many ways to design a tertiary classifier for secondary structure prediction based on binary classifiers. We used several methods proposed by Hua and Sun (Hua and Sun, 2001) to compare our results to theirs. Their methods are based on three one-versus-rest binary classifiers (H/∼H, E/∼E, C/∼C) and three one-versus-one binary classifiers (E/C, C/H, H/E). Three cascade tertiary classifiers, SVM_TREE1(H/∼H, E/C), SVM_TREE2(E/∼E, C/H), and SVM_TREE3(C/∼C, H/E), were made up of two binary classifiers. In the SVM_MAX_D tertiary classifier, the class for a testing sample was assigned as that corresponding to the largest positive distance to the optimal separating hyperplane among SVM_TREE1, SVM_TREE2, and SVM_TREE3 classifiers. The SVM_VOTE classifier combines all six binary classifiers using a simple voting principle: the testing sample was predicted to be in state i if the largest number of the six binary classifiers classified it as state i. SVM_JURY used the jury technique to combine all the results of the tertiary classifiers discussed above.

We designed two additional tertiary classifiers based on a one-versus-one scheme and a directed acyclic graph scheme (Heiler, 2002). The one-versus-one classifier for secondary structure prediction chooses the majority results based on three classifiers H/E, E/C, and C/H. Many test results show that one-versus-one classifiers are more accurate than one-versus-rest classifiers due to the fact that the one-versus-rest scheme often needs to deal with two data sets with very different sizes, i.e., unbalanced training data (Hsu and Lin, 2002; Heiler, 2002). However, a potential problem of the one-versus-one scheme is that the voting scheme might suffer from incompetent classifiers. For example, while the test point is helix (H), the result from the one-versus-one classifier E/C that is not related to helix inappropriately contributes to the decision. We can reduce this problem by using the directed acyclic graph (DAG) scheme that can classify a new data point after 2 binary classifications for 3 class problems without influence from incompetent classifiers. For example, if the testing point is predicted to be E (not C) from E/C classifier, then H/E the classifier is applied, while if the point is predicted to be not sheet (∼E) from E/C classifier, C/H the classifier is applied to determine if it is coil or helix. We developed the JURY2 classifier, which combines the results of SVM_MAX_D, SVM_VOTE, ONEvsONE, and DAG.

The results from the one-versus-one scheme and the DAG scheme were better than

those of SVM_TREE1, SVM_TREE2, or SVM_TREE3. Moreover, the results were comparable to those of SVM_MAX_D or SVM_JURY prediction though they used only one-versus-one classifiers for decisions instead of all 6 binary classifiers (See Table 3). This shows that the one-versus-one scheme or DAG scheme that utilizes only one-versus-one classifiers is a good approach in the three-class classification problem, such as protein secondary structure prediction, since we can reduce the computational complexity and the difficulty of big unbalanced classification by using one-versus-one binary classifiers rather than one-versus-rest binary classifiers.

Handling Unbalanced Data

For handling unbalanced data, we used different penalty parameters in the SVM formulation (Osuna *et al.*, 1997):

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ s.t. \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\ & 0 \le \alpha_i \le C_+, \text{ if } y_i = 1 \\ & 0 \le \alpha_i \le C_-, \text{ if } y_i = -1. \end{aligned} \tag{13}$$

Using different penalty parameters (C_+ and C_-), we can resolve the situation when the recall value of the smaller class is too small to produce good secondary structure prediction. We optimized the weight parameters for each binary classifier in order to produce optimal Q_3 or SOV_3.

Reliability Index and Support Vectors

When using machine learning approaches for the prediction of the secondary structure of a new sequence, it is important to know the prediction reliability. We used a reliability index (RI) to assign the prediction reliability. We can expect that the sample that has a large positive distance to the optimal separating hyperplane has large probability of belonging to the positive class. The reliability index is defined as:

$$RI = \begin{cases} 0 & \text{if d(i)} < 0.1 \\ \text{INT(d(i)/0.134)} & \text{if } 0.1 \le \text{d(i)} < 1.2 \\ 9 & \text{if d(i)} \ge 1.2, \end{cases} \tag{14}$$

where d(i) means the distance of the sample in state i to the optimal separating hyperplane of the binary classifier. The thresholds in the reliability index definition are chosen to make the percentage of residues about 22% for $RI = 9$ and about 12% for

$RI = 0$. Figure 2 shows the average accuracy (Q_3) and the percentage of residues covered against the cumulative reliability index from the SVMpsi method for the 136 blind test set proteins. 49.8% of all residues were predicted with $RI = 5$ or greater and 92% of them were correctly predicted. 22.8% of all residues were predicted with $RI = 9$ and 96% of them were correctly predicted. The ratio of the number of support vectors to all training samples for each 6 binary classifiers is below 50%, except in the C/∼C binary classifier. This shows that the PSI-BLAST profiles made classification easier than the multiple alignment frequency based profiles that had a ratio of about 50%. We developed the protein secondary structure predictor that is based on the SVMlight (Joachims, 1999). The single predicted α-helix is modified to the same secondary structure of the more reliable prediction (larger distance to optimal separating hyperplane) at the previous and next residue, since the occurrence of a single helix is not realistic.

Discussion

Comparison with Results of Other Methods

Table 2 shows that the accuracy of the binary classifier is significantly improved with SVMpsi. It is interesting that the accuracy of the H/E binary classifier is improved by more than 9%, while that of the C/∼C binary classifier is improved by 4.72% on the RS126 data set. Whenever a binary classifier involves the class of coil (C) (C/∼C, C/H, C/E), the prediction accuracy was lower. This seems to be due to the fact that the class C involves states that are not as well defined, and therefore items that belong to class C do not seem to have high within class consistency. As shown in Table 4, the Q_3 and SOV94 of the SVMpsi method based on the PSI-BLAST profiles are higher than those of the SVMfreq method based on the frequency profiles with multiple sequence alignments (Hua and Sun, 2001) as well as PHD, DSC, PREDATOR, and NNSSP for the RS126 data set. The Q_3 of SVMpsi outperforms the Bi-directional Recurrent Neural Network (BRNN) proposed by Baldi *et al.* (Baldi *et al.*, 1999) of 72.0% on the RS126 set. We can expect higher accuracy if the SVMpsi method is used as a component of the consensus method in conjunction with other good predictors, such as PSIPRED and Jnet. Q_3 scores for RS126 and CB513 are improved by 4.9% and 3.1%, respectively, and SOV94 scores are improved by 5.0% and 3.9%, respectively, compared to Hua and Sun's results. The improvement is much more than that of Jnet (3.1%) and comparison with PHD (Cuff and Barton, 2000). We can say that the improvement of SVMs with PSI-BLAST is higher than that of NNs with the PSI-BLAST profiles. The SVMpsi method achieves the highest published Q_3 and SOV94 values on both the RS126 and CB513 data sets to date. On the blind test of 136 protein sequences, the weighted average accuracy by sequence length, SOV94, and SOV99 scores were 77.2%, 81.8%, and 73.9%, respectively. Jones' PSIPRED method based on neural networks (Jones, 1999), which

used the PSI-BLAST profiles, achieved an overall pre-residue accuracy of $Q_3 = 76.5\%$ and SOV94 = 73.5% on his test set which includes 187 sequences after training with over 1,000 protein structures. Our results cannot be compared directly with those of PSIPRED since they used a different training set and test proteins that contain some sequences of the CB513 data set. Cuff and Barton (Cuff and Barton, 2000) showed the $Q_3 = 75.2\%$ from cross-validated predictions of their 480 non-redundant testing set proteins when the PSI-BLAST profiles was used. We obtained $Q_3 = 78.5\%$, SOV99 = 75.6%, and SOV94 = 82.8% from 7-fold cross validated predictions on our KP480 non-redundant test set. A direct comparison of performance between the two methods was not possible because the prepared KP480 data set may not be exactly the same as the 480 data set of Jnet. However, it shows that the SVM approach is another good method to perform secondary structure prediction. The Q_3 on the KP480 data set is higher than that of the CB513 data set. This is expected because it helps to increase prediction accuracy to remove the sequences that are shorter than 30 residues and to use the 8- to 3-state reduction method (4) (H to helix, E,B to sheet, all other states to coil) instead of reduction method (2).

CASP5 Experiment

We could improve the support vector machine approach for the protein secondary structure prediction (Hua and Sun, 2001) by new tertiary classifiers and their jury decision, an efficient method to handle unbalanced data, and PSSM profiles. This was promoted to improve the current level of prediction using SVMs, since neural network approaches have been studied for various structure and profiles by many researchers. It is not fair compare only the absolute Q_3 values when they are trained by different datasets. To evaluate our method, we participated in the recent fifth CASP (Critical Assessment of Structure Prediction) experiment in 2002. Figure 1 shows the results for the predictions that were submitted to the CASP5. The average Q_3 and SOV_3 scores for SVMpsi were 79.10% and 79.38%, respectively. We predicted the most accurate structure for 5 proteins compared to the other groups. This was ranked 4th among all participating groups. 21 groups could predict the most accurate structure at least for 1 protein. The 1st rank group predicted the most accurate structure for 7 proteins. It is not possible to say the exact rank with respect to average Q_3 or SOV_3 since the target size is small and the leading groups submitted different numbers of proteins. However, it shows that the SVMpsi method can at least match the current levels of prediction.

Further Improvements and Other Applications

We have focused on the contribution of the local interaction to the protein secondary structure using a sliding window scheme in this paper. The tertiary interactions between residues far apart in sequence but close in three dimension can be considered

(Baldi *et al.*, 1999) to improve prediction accuracy. However, the prediction accuracy of the secondary structure higher than 79% was obtained in CASP5 even when only the local contribution was considered. This shows that the local sequence environment of a residue substantially determines its secondary structure. It is possible that our method can be improved by considering long-range interaction. The SVMpsi method can also be improved by using larger training sets that contain new protein structure information, since the CB513 dataset used for the current SVMpsi was developed in 1999. It may require more memory to store data points while obtaining the optimal separating hyperplane. The prediction takes quite a long time if the ratio between the number of support vectors and the data points is large. The optimization of kernel parameters may become quite difficult due to computing time. A remaining problem is to handle huge training datasets using the SVMs approach. The Neural Networks (NNs) approach suffers from the local minima, determination of appropriate structure of neural networks, and too many parameters. Though SVMs also suffer from the kernel choice, we have shown that it is a comparable method for protein secondary structure prediction. This approach can be used for the biologically important, relevant problems, such as prediction of solvent accessibility and disulfide bonding state and connectivity. Though the local information is already effectively implemented by the sliding window, it is important to consider the long range interactions that are a major driving force underlying remote contact. Using the acquired information, it is quite possible to improve the accuracy of the prediction of protein tertiary structure. This will be useful to study protein folding processes as well.

Acknowledgements

The authors would like to thank University of Minnesota Supercomputing Institute (MSI) for intensive numerical computing. This work is partly achieved as a visiting scholar in the Korea Institute for Advance Study from Jan. 2002 to Aug. 2002. We would also thank Prof. Thorsten Joachims for making SVMlight software widely available, James A. Cuff & Prof. Geoffrey J. Barton for providing the data set, and Prof. David T. Jones for the PFILT software and his kind help.

References

Bairoch,A. and Apweiler,R. (2000) *Nucleic Acids Res.,* **28**, 45–48.

Baker,D. and Sali,A. (2001) *Science,* **294**, 93–96.

Baldi,P., Brunak,S., Frasconi,P., Pollastri,G. and Soda,G. (1999) *Bioinformatics,* **15**, 937–946.

Burges,C.J.C. (1998) *Data Mining and Knowledge Discovery,* **2**, 121–167.

Burges,C.J.C. and Schölkopf,B. (1997) In Mozer,M., Jordan,M. and Petsche,T. (eds), *Advances in Neural Information Processing Systems.* MIT Press, Cambridge, vol. 9, pp. 375–381.

Cristianini,N. and Shawe-Taylor,J. (2000) *Support Vecor Machines and other kernel-based learning methods.* University Press, Cambridge.

Cuff,J.A. and Barton,G.J. (1999) *Proteins: Struct. Funct. Genet.,* **34**, 508–519.

Cuff,J.A. and Barton,G.J. (2000) *Proteins: Struct. Funct. Genet.,* **40**, 502–511.

Ding,C.H.Q. and Dubchak,I. (2001) *Bioinformatics,* **17**, 349–358.

Frishman,D. and Argos,P. (1995) *Proteins: Struct. Funct. Genet.,* **23**, 566–579.

Frishman,D. and Argos,P. (1997) *Proteins: Struct. Funct. Genet.,* **27**, 329–335.

Heiler,M. (2002) *Optimization Criteria and Learning Algorithms for Large Margin Classifiers.* Diploma Thesis, University of Mannheim.

Henikoff,S. and Henikoff,J.G. (1992) *Proc. Natl Acad. Sci. USA,* **89**, 10915–10919.

Hsu,C.W. and Lin,C.J. (2002) *IEEE Transactions on Neural Networks,* **13**, 415–425.

Hua,S.J. and Sun,Z.R. (2001) *J. Mol. Biol.,* **308**, 397–407.

Joachims,T. (1999) In Schölkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods - Support Vector Learning.* MIT Press, Cambridge, pp 41–56.

Jones,D.T. (1999) *J. Mol. Biol.,* **292**, 195–202.

Jones,D.T. and Swindells,M.B. (2002) *Trends Biochem. Sci.,* **27**, 161–164.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1994) *Biochemistry,* **33**, 3038–3049.

Kabsch,W. and Sander,C. (1983) *Biopolymers,* **22**, 2577–2637.

King,R.D. and Sternberg,M.J.E. (1996) *Protein Sci.,* **5**, 2298–2310.

Matthews,B.W. (1975) *Biochim. Biophys. Acta,* **405**, 442–451.

Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) *Structure,* **5**, 1093–1108.

Osuna,E., Freund,R. and Girosi,F. (1997). Technical Report AI Memo 1602 MIT A.I. Lab.

Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.,* **202**, 865–884.

Richards,F.M. and Kundrot,C.E. (1988) *Proteins: Struct. Funct. Genet.,* **3**, 71–84.

Rost,B. and Sander,C. (1993) *J. Mol. Biol.,* **232**, 584–599.

Rost,B. and Sander,C. (1994) *J. Mol. Biol.,* **235**, 13–26.

Russell,R.B., Copley,R.R. and Barton,G.J. (1996) *J. Mol. Biol.,* **259**, 349–365.

Salamov,A.A. and Solovyev,V.V. (1995) *J. Mol. Biol.,* **247**, 11–15.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory.* Springer-Verlag, New York.

Vapnik,V. (1998) *Statistical Learning Theory.* John Wiley & Sons, New York.

Zemla,A., Venclovas,C., Fidelis,K. and Rost,B. (1999) *Proteins: Struct. Funct. Genet.,* **34**, 220–223.

Figure Legends

Figure 1. Bar graph showing the distribution of Q_3 and SOV scores for a set of 54 test proteins in CASP5. The y axis represents the number of proteins.

Figure 2. (a) Average secondary structure prediction accuracy (Q_3). (b) The percentage of amino acids covered with reliability index of greater or equal to the values shown on the x axis. For example, for residues with reliability index of greater or equal to 9, the average accuracy is 96.0%, and the percentage of residues with this index is 22.8%. Predictions are for a blind test set of 136 proteins.

classifier	$l = 7$	$l = 9$	$l = 11$	$l = 13$	$l = 15$	$l = 17$	$l = 19$	l^*
H/∼H	86.08	86.70	87.18	87.36	87.46	87.40	87.37	15
E/∼E	85.44	85.90	86.02	86.16	86.07	86.27	86.07	17
C/∼C	77.34	77.52	77.47	77.74	77.92	77.70	77.71	15
E/C	81.43	81.42	81.84	81.74	81.85	81.78	81.51	15
C/H	83.87	84.64	84.79	84.84	84.88	84.98	84.74	17
H/E	88.03	89.26	89.91	90.17	90.24	90.06	89.88	15

Table 1: Dependency of testing accuracy on window length for each binary classifier. The results are on the RS126 with PSI-BLAST profiles and SVM with RBF kernel where $\gamma = 0.05$ and $C = 1.0$. The l^* value is the optimal window length for each binary classifier. Combined results of 7-fold cross validation are shown.

	RS126		CB513	
classifier	SVMfreq*	SVMpsi*	SVMfreq†	SVMpsi†
H/∼H	80.36	87.46	83.02	86.75
E/∼E	81.25	86.27	83.39	86.69
C/∼C	73.20	77.92	75.52	78.40
E/C	76.69	81.85	78.32	81.84
C/H	77.63	84.98	79.97	84.83
H/E	80.87	90.24	83.08	90.52

Table 2: Accuracy of each binary classifiers in the corresponding optimal window length. SVMfreq*, SVMpsi*: results obtained on the RS126 set using 8- to 3-state reduction method (1). SVMfreq†, SVMpsi†: results obtained on the CB513 set using 8- to 3-state reduction method (2). SVMpsi results are obtained by PSI-BLAST profiles and SVM_JURY2 tertiary classifier. The results of SVMfreq* and SVMfreq† are from Hua and Sun (2001). Combined results of 7-fold cross validation are shown. SVMpsi is the new method proposed in this paper.

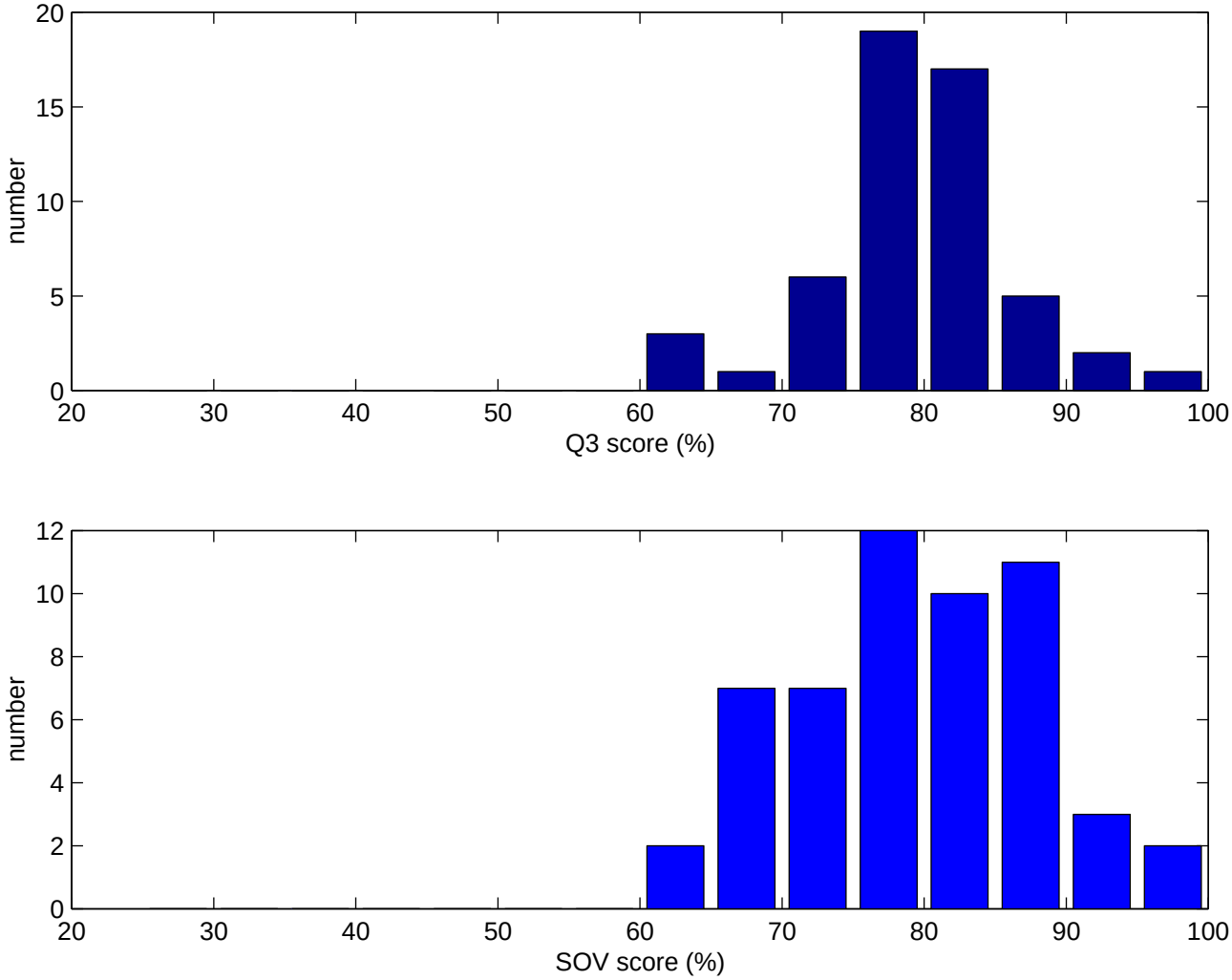
classifier	Q_3	Q_H	Q_E	Q_C	Q_H^{pre}	Q_E^{pre}	Q_C^{pre}	C_H	C_E	C_C	SOV94
SVM_TREE1	76.0	77.2	65.4	80.8	84.1	74.4	71.5	0.67	0.59	0.55	80.6
SVM_TREE2	76.1	78.2	64.2	80.7	83.9	74.8	71.6	0.68	0.58	0.54	80.0
SVM_TREE3	75.6	78.1	66.2	80.0	82.5	71.4	72.7	0.67	0.58	0.54	81.8
SVM_MAX_D	76.4	78.3	65.6	80.6	83.7	74.4	72.3	0.68	0.59	0.56	82.2
SVM_VOTE	76.5	78.1	65.6	81.1	84.4	74.8	72.1	0.68	0.60	0.56	80.2
SVM_JURY	76.5	78.2	65.6	80.9	84.6	74.8	72.1	0.68	0.60	0.56	80.1
ONEvsONE	76.5	77.9	65.6	81.0	84.5	74.5	72.0	0.68	0.60	0.55	80.0
DAG	76.4	78.1	66.0	80.6	84.3	74.1	72.2	0.68	0.59	0.55	81.6
SVM_JURY2	76.6	78.1	65.6	81.1	84.4	74.8	72.1	0.68	0.60	0.56	80.1

Table 3: Accuracy of tertiary classifiers on the CB513 set. Combined results of 7-fold cross validation are shown.

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	-
PHD†	73.5	-	-	-	73.5	-
DSC	71.1	-	-	-	71.6	-
PREDATOR	70.3	-	-	-	69.9	-
NNSSP	72.7	-	-	-	70.6	-
CONCENSUS	74.8	-	-	-	74.5	-
SVMpsi* (RS126)	76.1	77.2	63.9	81.5	79.6	72.0
SVMfreq†	73.5	75.0	60.0	79.0	76.2	-
SVMpsi† (CB513)	76.6	78.1	65.6	81.1	80.1	73.5
SVMpsi‡ (KP480)	78.5	80.2	65.8	83.5	82.8	75.6

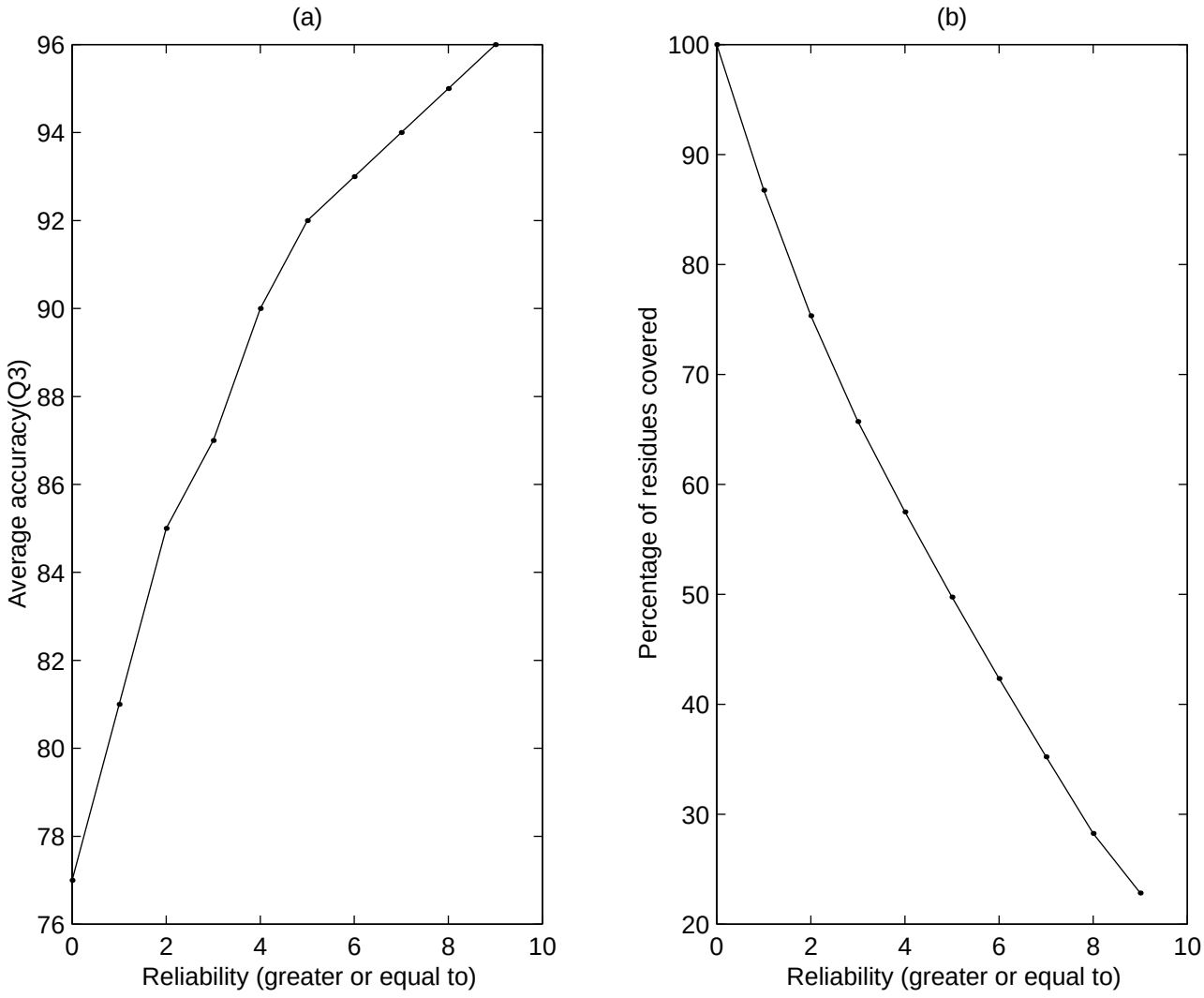
Table 4: All results used 8- to 3-state reduction method (2) except PHD* and SVMfreq* of which results were obtained on the RS126 set using 8- to 3-state reduction method (1), and SVMpsi‡ of which results obtained on the KP480 set using 8- to 3-state reduction method (4). PHD†, DSC, PREDATOR, NNSSP, CONCENSUS: results obtained on the RS126 from Cuff & Barton (1999). PHD*: results obtained from Rost and Sander (1993) and Rost *et al.* (1994). SVMfreq†, SVMpsi†: results obtained on the CB513 set. SVMfreq*, SVMfreq†: results obtained from Hua and Sun (2001). SVMpsi*, SVMpsi†, SVMpsi‡: results obtained on the RS126, CB513, and KP480 data set, respectively. Combined results of 7-fold cross validation are shown. SVMpsi is the new method proposed in this paper.

Figure 1


number
Q3 score (%)
0
5
10
15
20
30
40
50
60
70
80
90
100

number
SOV score (%)
0
2
4
6
8
10
12
20
30
40
50
60
70
80
90
100

Figure 2


(a)
Average accuracy(Q3)
Reliability (greater or equal to)
96
94
92
90
88
86
84
82
80
78
76
0
2
4
6
8
10
(b)
Percentage of residues covered
Reliability (greater or equal to)
100
90
80
70
60
50
40
30
20
0
2
4
6
8
10