Prediction of protein secondary structure with two-stage multi-class SVMs

Minh N. Nguyen
BioInformatics Research Centre, School of Computer Engineering
Nanyang Technological University, Singapore

Jagath C. Rajapakse
BioInformatics Research Centre, School of Computer Engineering
Nanyang Technological University, Singapore
Biological Engineering Division, Massachusetts Institute of Technology, USA

Abstract: Bioinformatics techniques to protein secondary structure (PSS) prediction mostly depend on the information available in amino acid sequences. In this paper, we propose two-stage Multi-class Support Vector Machine (MSVM) approach where the second MSVM predictor is introduced at the output of the first stage MSVM to capture the contextual relationship among secondary structure elements in order to minimize the generalization error in the prediction. By using position specific scoring matrices generated by PSI-BLAST, the two-stage MSVM approach achieves Q_3 accuracies of 78.0% and 76.3% on the RS126 dataset of 126 nonhomologous globular proteins and the CB396 dataset of 396 nonhomologous proteins, respectively, which are better than the scores reported on both datasets to date. By using MSVM, the present prediction scheme significantly achieves 2-6% and 3-15% of improvement in Q_3 and Sov accuracies, respectively, on the two datasets. On larger blind-test datasets from PSIPRED, CASP 4, and EVA datasets, two-stage MSVM approach achieves Q_3 accuracies from 77% to 79.5%.

Keywords: protein structure; secondary structure prediction; support vector machines (SVMs); multi-class SVMs.

1 INTRODUCTION

Proteins are large molecules having a central role in coordinating living processes and constituting to the bulk of living organisms. As such, understanding the mechanisms of proteins' interaction and operations is vital to the study of many diseases. Key functional aspects of proteins depend on their 3-dimensional (3-D) structure (1). Therefore, the knowledge of a protein's structure and its components and their relation to its function are highly useful in the advances of life sciences. Unfortunately, the protein structure prediction problem is a combinatorial optimization problem, which so far has an eluded solution, because of the exponential number of potential solutions. One of the current approaches is to predict the protein secondary structure (PSS), which is linear representation of the full knowledge of the 3-D structure, and, thereafter, predict the 3-D structure (1; 2). The usual goal of secondary structure prediction is to classify a pattern of residues in amino acid sequences to a pattern of protein secondary structure elements: an α-helix (H), β-strand (E), or coil (C, the remaining type).

Many computational techniques have been proposed in the literature to solve the PSS prediction problem. The statistical methods are mostly based on the likelihood (3; 4; 5), k-nearest neighbor (6; 7), or Bayesian (8) techniques. Neural networks use residues in a local neighborhood, as inputs, to predict the secondary structure at a particular location of an amino acid sequence by finding an arbitrary non-linear mapping (9; 10; 11; 12). The consensus approaches combine different classifiers, in parallel, into a single superior predictor (13; 14; 15). The method of Cuff and Barton employed a majority voting scheme to combine predictions from different techniques (13). More complex approaches for combining different methods based on neural networks, linear discrimination (14), and multi-category SVM (15) have been studied. Combining evolutionary information with the existing methods has provided to be effective and efficient for predicting PSS (16). Recently, Meiler and Baker proposed using the information of 3-D structure and PSI-BLAST profiles as inputs to a neural network (17). The SVMs have been applied

to PSS prediction, by combining several binary classifiers (18; 19; 20).

Despite the existence of many approaches, the success rates of the existing approaches to PSS prediction are insufficient. Most existing secondary structure prediction techniques are single-stage approaches in the sense that they predict secondary structures directly from an amino acid sequence, which are not capable of effectivly taking into consideration the contextual relationships among the secondary structures of residues. Rost and Sander proposed the PHD approach using two Multi-Layer Perceptrons (MLP) in cascade, where the second stage MLP improved the accuracy of the prediction by capturing the contextual relations among the secondary structures from the output of the first stage (9). This approach, however, increases the size of the effective input window, but does not improve the generalization capability of the prediction of unseen patterns. The Bayesian approach provides a framework to account for non-local interactions among amino acid residues (8), where the inferences are based on the generalized probability distributions incorporating prior probabilities of segments of secondary structure elements. Nevertheless, this technique is unable to incorporate useful information from multiple sequence alignments or PSI-BLAST profiles in the prediction scheme. The PSI-BLAST profiles contained more useful information than single sequences and multiple sequence alignments: the probability of each residue residing at a specific position is computed; the amount of significant information of each sequence is weighted and more distant homologues are found (12).

This paper investigates the use of two-stage Multi-class Support Vector Machines (MSVMs) for PSS prediction by using a second MSVM to enhance the output of a single-stage MSVM approach. The input to the two-stage MSVM is based on the position specific scoring matrices generated by PSI-BLAST profiles of the input sequence. We show that the MSVM at the second stage minimizes the generalization error made by the first stage by incorporating the contextual relationships among the PSS elements. The experiments are presented to demonstrate the selection of the kernels and their parameters. Improvements of accuracy of PSS prediction up to 2.3% were achieved on the RS126 dataset (9) and the CB396 dataset (13), respectively, compared to the previously reported best results for the same datasets. The new prediction scheme achieved improvements of 2-6% and 3-15% of Q_3 and Sov accuracies, respectively, on the RS126 and CB396 datasets. On larger blind-test datasets from PSIPRED, CASP4, and EVA datasets, two-stage MSVMs obtained Q_3 accuracies from 77.0% to 79.5%.

2 TWO-STAGE MSVM APPROACH

The two-stage MSVM approach uses two MSVMs in cascade to predict secondary structures of residues in amino acid sequences.

Let us denote the given amino acid sequence by $\mathbf{r} = (r_1, r_2, \ldots, r_n)$ where $r_i \in \Omega_R$ and Ω_R is the set of 20 amino acid residues, and $\mathbf{t} = (t_1, t_2, \ldots t_n)$ denote the corresponding secondary structure sequence where $t_i \in \Omega_T$ and $\Omega_T = \{H, E, C\}$; n is the length of the sequence. The prediction of the PSS sequence, $\mathbf{t}$, from an amino acid sequence, $\mathbf{r}$, is the problem of finding the optimal mapping from the space of Ω_R to the space of Ω_T .

Let $\mathbf{v}_i$ be the 21-dimensional feature vector representing the residue r_i where 20 units are the values from raw matrices of PSI-BLAST profiles ranging from [0, 1] and the other is used for padding to indicate an overlapping end of the sequence (12). Let the input pattern to the MSVM approach at site i be $\mathbf{r}_i = (\mathbf{v}_{i-h_1^1}, \mathbf{v}_{i-h_1^1+1}, \ldots, \mathbf{v}_i, \ldots, \mathbf{v}_{i+h_2^1})$ where $\mathbf{v}_i$ denotes the center element, h_1^1 and h_2^1 denote the width of window on the two sides; $w_1 = h_1^1 + h_2^1 + 1$ is the neighborhood size around the element i.

2.1 First Stage

For the first stage, we use a MSVM scheme proposed by Crammer and Singer (21). For PSS prediction, this method constructs three discriminant functions by solving one single optimization problem, which can be formulated as follows.

Minimize

$$\frac{1}{2}\sum_{k\in\Omega_T}(\mathbf{w}_1^k)^T\mathbf{w}_1^k + \gamma^1\sum_{j=1}^{N}\xi_j^1$$

subject to the constraints

$$\mathbf{w}_1^{t_j}\phi^1(\mathbf{r}_j) - \mathbf{w}_1^k\phi^1(\mathbf{r}_j) \geq c_j^k - \xi_j^1 \qquad (1)$$

where t_j is the secondary structural type of residue r_j corresponding to the the training vector $\mathbf{r}_j$, $j = 1, 2, \ldots, N$, $\mathbf{w}_1^{t_j}$ and $\mathbf{w}_1^k$ are the weight vectors of class t_j and k, and $c_j^k = \begin{cases} 0 & \text{if } t_j = k \\ 1 & \text{if } t_j \neq k \end{cases}$.

The minimization of the above formulation is simplified by solving the following quadratic programming problem (21):

$$\max_{\alpha_j^k} -\frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N}\mathcal{K}^1(\mathbf{r}_j,\mathbf{r}_i)\sum_{k\in\Omega_T}\alpha_j^k\alpha_i^k - \sum_{j=1}^{N}\sum_{k\in\Omega_T}\alpha_j^k c_j^k$$

$$\text{such that } \sum_{k\in\Omega_T}\alpha_j^k = 0 \text{ and } \alpha_j^k \leq \begin{cases} 0 & \text{if } t_j \neq k \\ \gamma^1 & \text{if } t_j = k \end{cases} \qquad (2)$$

where $\mathcal{K}^1(\mathbf{r}_i, \mathbf{r}_j) = \phi^1(\mathbf{r}_i)\phi^1(\mathbf{r}_j)$ denotes the kernel function and $\mathbf{w}_1^k = \sum_{j=1}^{N}\alpha_j^k\phi^1(\mathbf{r}_j)$. The input vectors, derived from a window of w_1 amino acid residues, are transformed to a higher dimensional space via the kernel function $\mathcal{K}^1$.

Once the parameters α_j^k are obtained from the optimization, the resultant discriminant function f_1^k of a test input vector $\mathbf{r}_i$ is given by

$$f_1^k(\mathbf{r}_i) = \sum_{j=1}^{N}\alpha_j^k\mathcal{K}^1(\mathbf{r}_i,\mathbf{r}_j) = \mathbf{w}_1^k\phi^1(\mathbf{r}_i). \qquad (3)$$

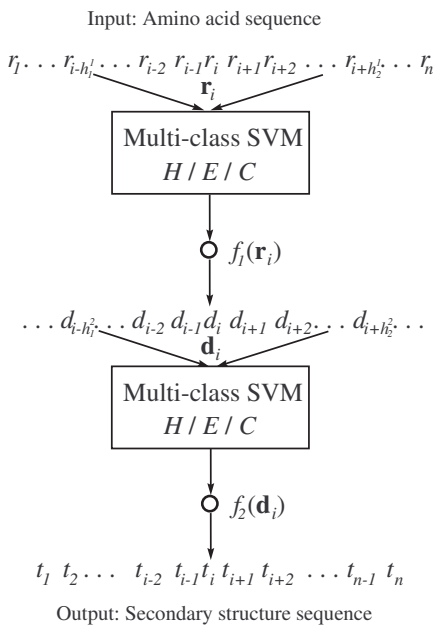

Figure 1: Two-stage MSVM approach for PSS prediction.

In a single-stage MSVM method, the secondary structural type t_i corresponding to the residue at site i, r_i, is determined by

$$t_i = \arg\max_{k \in \Omega_T} f_1^k(\mathbf{r}_i). \tag{4}$$

2.2 Second Stage

We extend the single-stage MSVM approach to PSS prediction by cascading another MSVM at the output of the single-stage approach. The second stage improves the accuracy of prediction because the secondary structure at a particular site of the sequence depends on the structures of the rest of the sequence, for example, it accounts for the fact that the strands span over at least three adjacent residues and helices consist of at least four consecutive residues (9; 22). This intrinsic relation cannot be captured by using only a single MSVM. Therefore, another predictor that minimizes the generalization error at the output of single-stage methods, by incorporating the contextual relationship among the protein structure elements, improves the prediction accuracy. Figure 1 represents the architecture of the two-stage MSVM approach for PSS prediction.

Consider a window of size w_2 at a site of the output sequence from the first stage; the vector at position i, $\mathbf{d}_i = (d^k_{i-h_1^2}, d^k_{i-h_1^2+1}, \ldots, d^k_i, \ldots, d^k_{i+h_2^2})$ where $w_2 = 3(h_1^2+h_2^2+1)$, and h_1^2 and h_2^2 are the length of the neighborhood on two sides. $d_i^k = 1/(1+e^{-f_1^k(\mathbf{r}_i)})$ and f_1^k denotes the discriminant function of the first stage. The logistic sigmoid function is selected to restrict the input units of the second stage to (0,1) interval.

The MSVM converts the input patterns, usually linearly inseparable, to a higher dimensional space by using the mapping ϕ^2 with a kernel function $\mathcal{K}^2(\mathbf{d}_i, \mathbf{d}_j) = \phi^2(\mathbf{d}_i)\phi^2(\mathbf{d}_j)$. As in the first stage, the outputs in the higher dimensional space are linearly combined by a weight vector, $\mathbf{w}_2$, to obtain the prediction output. The training set of exemplars for the second stage MSVM is $\Gamma^2_{\text{train}} = \{\mathbf{d}_j : j = 1, \ldots, N\}$ where N denotes the number of training exemplars. The vector $\mathbf{w}_2$ is obtained by solving the following convex quadratic programming problem, over all the patterns seen in the training phase (21).

The secondary structural type t_i corresponding to the residue r_i is determined by

$$t_i = \arg\max_{k \in \Omega_T} f_2^k(\mathbf{d}_i). \tag{5}$$

3 MINIMAL GENERALIZATION ERROR

In this section, we prove that the second stage MSVM minimizes the generalization error produced at the output of the first stage by incorporating the contextual relationships among secondary structure elements.

In the first stage , if the training and testing patterns, $(\mathbf{r}, t) \in \Gamma^1 \times \Omega_T$, are drawn independently and identically according to a probability distribution $\mathcal{P}_1$ based on the optimal value of window size w_1, where Γ^1 denotes the set of input patterns seen by the MSVM during both the training and testing phases and t denotes the desired ouput for input pattern $\mathbf{r}$, the generalization error, $\text{err}_{\mathcal{P}_1}(f_1)$, when $f_1(\mathbf{r}) = \arg\max_{k \in \Omega_T} f_1^k(\mathbf{r})$, is given by

$$\text{err}_{\mathcal{P}_1}(f_1) = \int L(f_1(\mathbf{r}), t) d\mathcal{P}_1(\mathbf{r}, t)$$

where the loss function

$$L(f_1(\mathbf{r}), t) = \begin{cases} 0 & \text{if } f_1(\mathbf{r}) = t \\ 1 & \text{if } f_1(\mathbf{r}) \neq t \end{cases}.$$

The aim of two-stage MSVM approach is to find an optimal function, f_2 such that $f_2(\mathbf{d}) = \arg\max_{k \in \Omega_T} f_2^k(\mathbf{d})$, to incorporate the contextual relationships among secondary structure elements in order to minimize the generalization error further in the prediction. Let Γ^2 denote the set of input patterns seen by the second stage MSVM in both training and testing phases. If, for both training and testing, the patterns $(\mathbf{d}, t) \in \Gamma^2 \times \Omega_T$ are drawn independently and identically according to a probability distribution $\mathcal{P}_2$ based on the optimal value of window size w_2, the generalization error, $\text{err}_{\mathcal{P}_2}(f_2)$, is given by

$$\text{err}_{\mathcal{P}_2}(f_2) = \int L(f_2(\mathbf{d}), t) d\mathcal{P}_2(\mathbf{d}, t).$$

If the input pattern $\mathbf{d}$ corresponds to a site i, then $\mathbf{d} = \mathbf{d}_i$ and $\mathbf{d}_i = ((1+e^{-f_1^k(\mathbf{r}_{i-h_1^2})})^{-1}, (1+e^{-f_1^k(\mathbf{r}_{i-h_1^2+1})})^{-1}, \ldots, (1+e^{-f_1^k(\mathbf{r}_i)})^{-1}, \ldots, (1+e^{-f_1^k(\mathbf{r}_{i+h_2^2})})^{-1})$. That is, the second stage takes into account the influences of the PSS elements of residues in the neighborhood into the prediction. Since there exists at least a discriminant function f_2^* such that $\text{err}_{\mathcal{P}_2}(f_2^*) = \text{err}_{\mathcal{P}_1}(f_1)$ when $h_1^2 = h_2^2 = 0$ (see Appendix) i.e the contextual information of secondary structures is not taken into account, the optimal function f_2 providing the smallest $\text{err}_{\mathcal{P}_2}(f_2)$ at optimal size of

neighborhood capturing the contextual information of secondary structures should ensure $\mathrm{err}_{\mathcal{P}_2}(f_2) \leq \mathrm{err}_{\mathcal{P}_2}(f_2^*) = \mathrm{err}_{\mathcal{P}_1}(f_1)$.

However, finding the global minimum of generalization error $\mathrm{err}_{\mathcal{P}_2}(f_2)$ is not a trivial problem because the form of the probability distribution $\mathcal{P}_2$ is unknown. We can instead consider the probably approximately correct (pac) bound, $\epsilon(N,\delta)$, of the generalization error satisfying

$$\mathcal{P}_2\{\Gamma^2_{\text{train}} : \exists\, f_2 \text{ such that } \mathrm{err}_{\mathcal{P}_2}(f_2) > \epsilon(N,\delta)\} < \delta.$$

This is equivalent to asserting that with a probability greater than $1-\delta$ over the training set Γ^2_{train} based on the optimal value of window size w_2, the generalization error of f_2 is bounded by

$$\mathrm{err}_{\mathcal{P}_2}(f_2) \leq \epsilon(N,\delta).$$

In the following theorems, we assume that both the training set $\Gamma^2_{\text{train}} \subset \Gamma^2$ and the testing set $\Gamma^2_{\text{test}} \subset \Gamma^2$ for the second stage contained N patterns, for simplicity. For the MSVM technique at the second stage, let $\mathbf{w}_2^{k/l}$ be the weight vector $\mathbf{w}_2^k - \mathbf{w}_2^l$ and, therefore, $\mathbf{w}_2^{k/l}\phi^2(\mathbf{d}) = \mathbf{w}_2^k\phi^2(\mathbf{d}) - \mathbf{w}_2^l\phi^2(\mathbf{d}) = f_2^k(\mathbf{d}) - f_2^l(\mathbf{d})$. The secondary structure of a residue r is not l if $\mathbf{w}_2^{k/l}\phi^2(\mathbf{d}) \geq 0$ or not k otherwise.

Definition 1 Let $\mathcal{F} = \{f_2^{k/l} : \mathbf{d} \to \mathbf{w}_2^{k/l}\phi^2(\mathbf{d}); \|\mathbf{w}_2^{k/l}\| \leq 1; \mathbf{d} \in \Gamma^2; k,l \in \Omega_T\}$ be a set of real valued functions defined on $S = \{\mathbf{d}_i \in \Gamma^2 : i = 1,\ldots,N\}$. We say that the set of points S is $\eta_2^{k/l}$-*shattered* by $\mathcal{F}$ if there exist real numbers $\zeta_i^{k/l}$, $i = 1,\ldots,N$, such that, for every binary classification, $q_i^{k/l} \in \{-1,+1\}$ on set S, there exists $f_2^{k/l} \in \mathcal{F}$ satisfying $f_2^{k/l}$ $\begin{cases} \geq \zeta_i^{k/l} + \eta_2^{k/l} & \text{if } q_i^{k/l} = +1 \\ \leq \zeta_i^{k/l} - \eta_2^{k/l} & \text{if } q_i^{k/l} = -1 \end{cases}$

Definition 2 The *fat-shattering dimension* $\mathrm{fat}_{\mathcal{F}}(\eta_2^{k/l})$ at scale $\eta_2^{k/l}$ is the size of the largest $\eta_2^{k/l}$-shattered subset of Γ^2.

Theorem 1 (23) Let $\mathcal{F} = \{f_2^{k/l} : \mathbf{d} \to \mathbf{w}_2^{k/l}\phi^2(\mathbf{d}); \|\mathbf{w}_2^{k/l}\| \leq 1; \mathbf{d} \in \Gamma^2; k,l \in \Omega_T\}$ be a set of discriminant functions defined on Γ^2 and restricted to points in a ball of m dimensions of radius R about the origin, that is, $f_2^{k/l}(\mathbf{d}) = f_2^k(\mathbf{d}) - f_2^l(\mathbf{d})$, $\phi^2(\mathbf{d}) \in \mathbb{R}^m$, and $\|\phi^2(\mathbf{d})\| \leq R$. Then, the fat-shattering dimension $\mathrm{fat}_{\mathcal{F}}(\eta_2^{k/l})$ at scale $\eta_2^{k/l}$ is bounded:

$$\mathrm{fat}_{\mathcal{F}}(\eta_2^{k/l}) \leq \left(\frac{R}{\eta_2^{k/l}}\right)^2.$$

Definition 3: The *decision directed acyclic graph* G on 3 classes H, E, and C, over $\mathcal{F}$ is a set of functions which can be implemented using a rooted binary directed acyclic graph with 3 leaves labeled by the classes H, E, and C, where each of 3 internal nodes is labeled with an element of $\mathcal{F}$ (see Figure 2).

Theorem 2 (24): Let G be a decision directed acyclic graph on 3 classes H, E, and C, with 3 decision nodes, H/E, E/C, and C/H, with margins $\eta_2^{k/l}$ and discriminant

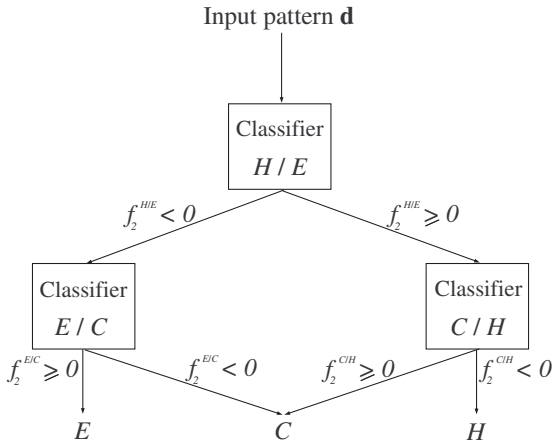

Figure 2: Illustration of a decision directed acyclic graph G with discriminant functions $f_2^{k/l} \in \mathcal{F}$ at decision nodes k/l where $k,l \in \Omega_T$ and the leaves labeled by three classes helix (H), strand (E), and coil (C).

functions $f_2^{k/l} \in \mathcal{F}$ at decision nodes $k/l \in G$, where $\eta_2^{k/l} = \min_{\mathbf{d}\in\Gamma^2_{\text{train}}} \frac{|\mathbf{w}_2^{k/l}\phi^2(\mathbf{d})|}{\|\mathbf{w}_2^{k/l}\|}$. Then, the following probability is bounded:

$$\mathcal{P}_2\{\Gamma^2_{\text{train}}, \Gamma^2_{\text{test}} : \exists\, G \text{ such that } \mathrm{err}_{\Gamma^2_{\text{train}}}(G) = 0 \text{ and } \mathrm{err}_{\Gamma^2_{\text{test}}}(G) > \epsilon(\delta)\} < \delta \quad (6)$$

where $\epsilon(\delta) = \frac{1}{N}\left(\sum_{k/l\in G} h^{k/l} \log\frac{4eN}{h^{k/l}} \log(4N) + \log\frac{2^3}{\delta}\right)$, $h^{k/l} = \mathrm{fat}_{\mathcal{F}}\left(\frac{\eta_2^{k/l}}{8}\right)$, $\mathrm{err}_{\Gamma^2_{\text{train}}}(G)$ and $\mathrm{err}_{\Gamma^2_{\text{test}}}(G)$ are the fractional error of G on the training set Γ^2_{train} and a random testing set Γ^2_{test}, respectively.

Theorem 3: Let G be a decision directed acyclic graph with discriminant functions $f_2^{k/l} \in \mathcal{F}$ at nodes k/l. Then, the generalization error of f_2 where $f_2(\mathbf{d}) = \arg\max_{k\in\Omega_T} f_2^k(\mathbf{d})$ in the probability distribution $\mathcal{P}_2$ is

$$\mathrm{err}_{\mathcal{P}_2}(f_2) = \mathrm{err}_{\mathcal{P}_2}(G).$$

Proof: This is equivalent to proving that for an arbitrary example $\mathbf{d} \in \Gamma^2$, $f_2(\mathbf{d})$ equals to the secondary structural type of $\mathbf{d}$, predicted by the decision directed acyclic graph G.

Firstly, consider $f_2(\mathbf{d}) = \arg\max_{k\in\Omega_T} f_2^k(\mathbf{d}) = H$. And there are two cases when $\arg\max_{k\in\Omega_T} f_2^k(\mathbf{d}) = H$:

- In the first case, $f_2^H(\mathbf{d}) > f_2^E(\mathbf{d}) > f_2^C(\mathbf{d})$. Starting at the root node H/E of G, the binary decision function of the classifier H/E is evaluated for the input pattern $\mathbf{d}$. The node is then exited via the right edge because $f_2^{H/E}(\mathbf{d}) = \mathbf{w}_2^{H/E}\phi^2(\mathbf{d}) = f_2^H(\mathbf{d}) - f_2^E(\mathbf{d}) > 0$. In the decision node C/H, the discriminant function $f_2^{C/H}(\mathbf{d})$ is evaluated. Since $f_2^{C/H}(\mathbf{d}) = f_2^C(\mathbf{d}) - f_2^H(\mathbf{d}) < 0$, the input $\mathbf{d}$ reaches to the leaf labeled H. Therefore, the secondary structural type of $\mathbf{d}$ predicted by G is H.

- In the second case, if $f_2^H(\mathbf{d}) > f_2^C(\mathbf{d}) > f_2^E(\mathbf{d})$ then we use an identical argument.

Similar proofs hold for cases $f_2(\mathbf{d}) = \arg\max_{k\in\Omega_T} f_2^k(\mathbf{d}) = E$ or C. ■

Theorem 4 (25): Let $\mathrm{err}_{\mathcal{P}_2}(G)$ be the generalization error of G at the output of the first stage. Then

$$\mathcal{P}_2\{\Gamma^2_{\text{train}} :\ \exists\, G; \mathrm{err}_{\Gamma^2_{\text{train}}}(G) = 0 \text{ and } \mathrm{err}_{\mathcal{P}_2}(G) > 2\epsilon(N,\delta)\}$$

$$\leq 2\mathcal{P}_2\{\Gamma^2_{\text{train}}, \Gamma^2_{\text{test}} :\ \exists\, G \text{ such that } \mathrm{err}_{\Gamma^2_{\text{train}}}(G) = 0 \text{ and}$$

$$\mathrm{err}_{\Gamma^2_{\text{test}}}(G) > \epsilon(N,\delta)\}$$

Theorem 5: Suppose that we classify N patterns in the training set Γ^2_{train} using the MSVM method at the second stage with optimal values of weight vectors $\mathbf{w}_2^k$, $k \in \Omega_T$. Then, the pac bound of the generalization error, $\mathrm{err}_{\mathcal{P}_2}(f_2)$, is equal to: $\epsilon(\delta) =$

$$\frac{1}{N}\left(390R^2 \sum_{k\in\Omega_T} \|\mathbf{w}_2^k\|^2 \log(4eN)\log(4N) + 2\log\frac{2(2N)^3}{\delta}\right). \tag{7}$$

Proof: Since the margin $\eta_2^{k/l}$ is the minimum value of the distances from the instances labeled k or l to the hyperplane $\mathbf{w}_2^{k/l}\phi^2(\mathbf{d}) = 0$ at the second stage, we have, $\eta_2^{k/l} = \min_{\mathbf{d}\in\Gamma^2_{\text{train}}} \frac{|\mathbf{w}_2^{k/l}\phi^2(\mathbf{d})|}{\|\mathbf{w}_2^{k/l}\|} = \min_{\mathbf{d}\in\Gamma^2_{\text{train}}} \frac{|(\mathbf{w}_2^k - \mathbf{w}_2^l)\phi^2(\mathbf{d})|}{\|\mathbf{w}_2^k - \mathbf{w}_2^l\|} \geq \frac{1}{\|\mathbf{w}_2^k - \mathbf{w}_2^l\|}$. Therefore, the quantity $\sum_{k/l} \frac{1}{(\eta_2^{k/l})^2} \leq \sum_{k/l} \|\mathbf{w}_2^k - \mathbf{w}_2^l\|^2 = 3\sum_k \|\mathbf{w}_2^k\|^2 - \left(\sum_k \mathbf{w}_2^k\right)^2 \leq 3\sum_k \|\mathbf{w}_2^k\|^2$. Solving the optimization problems at second stage results in the minimization of the quantity $\sum_{k\in\Omega_T} \|\mathbf{w}_2^k\|^2$ which is directly related to the margin of the classifier. Plugging of the binary classifiers induced by $\mathbf{w}_2^{k/l}$ results a stepwise method for calculating the maximum among $\{f_2^k(\mathbf{d}) = \mathbf{w}_2^k\phi^2(\mathbf{d});\ k \in \Omega_T\}$ that is similar to the process of finding the secondary structure in the decision directed acyclic graph G. Let us apply the result of Theorem 2 for G with specified margin $\eta_2^{k/l}$ at each node to bound the generalization error $\mathrm{err}_{\mathcal{P}_2}(G)$. Since the number of decision nodes is 3 and the largest allowed value of $h^{k/l} = \mathrm{fat}_{\mathcal{F}}\left(\frac{\eta_2^{k/l}}{8}\right)$ is N, the number of all possible patterns of $h^{k/l}$'s over the decision nodes is bounded by N^3. We let $\delta_i = \delta/N^3$ so that $\sum_{i=1}^{N^3} \delta_i = \delta$. By choosing $\epsilon(\frac{\delta_i}{2})$

$$= \frac{1}{N}\left(195R^2 \sum_{k\in\Omega_T} \|\mathbf{w}_2^k\|^2 \log(4eN)\log(4N) + \log\frac{2(2N)^3}{\delta}\right)$$

$$\geq \frac{1}{N}\left(65R^2 \sum_{k/l\in G} \frac{1}{(\eta_2^{k/l})^2} \log(4eN)\log(4N) + \log\frac{2(2N)^3}{\delta}\right)$$

$$> \frac{1}{N}\left(\sum_{k/l\in G} \frac{R^2}{(\eta_2^{k/l}/8)^2} \log\frac{4eN}{h^{k/l}} \log(4N) + \log\frac{2^3}{\delta_i/2}\right)$$

from Theorem 1

$$> \frac{1}{N}\left(\sum_{k/l\in G} h^{k/l} \log\frac{4eN}{h^{k/l}} \log(4N) + \log\frac{2^3}{\delta_i/2}\right).$$

Theorem 2 ensures that the probability of any of the statements failing to hold is less than $\delta/2$. By using the result of the Theorem 4, the probability $\mathcal{P}_2\{\Gamma^2_{\text{train}} :\ \exists\ G \text{ s.t. } \mathrm{err}_{\Gamma^2_{\text{train}}}(G) = 0;\ \mathrm{err}_{\mathcal{P}_2}(G) > 2\epsilon(\delta_i/2)\}$ is bound to be less than δ. From Theorem 3, the pac bound of the generalization error $\mathrm{err}_{\mathcal{P}_2}(f_2)$ is therefore equal to $2\epsilon(\frac{\delta_i}{2}) = \frac{1}{N}(390R^2\sum_{k\in\Omega_T}\|\mathbf{w}_2^k\|^2 \log(4eN)\log(4N) + 2\log\frac{2(2N)^3}{\delta})$. ■

From Eq. (7), minimizing the quantity $\sum_{k\in\Omega_T}\|\mathbf{w}_2^k\|^2$ results in the minimization of the generalization error at the output of the first stage MSVM method. Since the MSVM at the second stage minimizes the error of the output of the first stage by solving the optimization problem, two stages are sufficient to find an optimal classifier for PSS prediction with a minimal generalization error, taking into account the contextual information of secondary structures.

3.1 Minimization of error by MLP

Two-stage Multi-Layer Perceptron (MLP) has been earlier used for PSS prediction (9). The aim of the PSS prediction using a MLP is to find an optimal function, f_1, taking into account the sequential relationships among amino acid residues to minimize the following so-called the *empirical risk functional* (26)

$$\mathrm{err}_{\Gamma^1_{\text{train}}}(f_1) = \frac{1}{N}\sum_{i=1}^{N} L(f_1(\mathbf{r}_i), t_i) \tag{8}$$

where $\mathbf{r}_i \in \Gamma^1_{\text{train}}$, t_i denotes the corresponding output for the training input pattern $\mathbf{r}_i$, and N is the number of patterns of Γ^1_{train}, $f_1(\mathbf{r}) = \arg\max_{k\in\Omega_T} f_1^k(\mathbf{r})$, and f_1^k is the transfer function of the network for secondary structure type k, which output is given by the output neuron representing the secondary structure $k \in \Omega_T$.

By using gradient-decent type training algorithm, the single-stage MLP method achieves a local minimum of the empirical risk functional $\mathrm{err}_{\Gamma^1_{\text{train}}}(f_1)$, which is not guaranteed to find the global minimum of the generalization error $\mathrm{err}_{\mathcal{P}_1}(f_1)$.

Since the MLP is not guaranteed to minimize the generalization error at the output of the first stage, the second neural network is unable to find the global minimum of the generalization error $\mathrm{err}_{\mathcal{P}_2}(f_2)$. Therefore, the two-stage MLPs are incapable of achieving an optimal predictor for PSS prediction as the test sets and training sets differ in PSS prediction. The addition of the second neural network at the output of the single-stage predictor effectively increases the size of the input window.

4 EXPERIMENTS AND RESULTS

The two-stage MSVM approach was implemented with the position specific scoring matrices generated by PSI-BLAST as inputs and tested on benchmark datasets by using seven-fold cross-validation. And the results were compared with the earlier methods for PSS prediction (3; 4; 5; 33; 9; 6; 34; 35; 10; 11; 13; 18; 19; 20). With the seven-fold cross-validation approximately one-seventh of the dataset was left out while training and, after training, the left one-seventh of the dataset was used for testing. We used BSVM library (29) to implement MSVM, which usually leads to faster convergence in large optimization problems.

4.1 Dataset 1 (RS126)

The set of 126 nonhomologous globular protein chains, used in the experiment of Rost and Sander (9) and referred to as the RS126 set, was used to evaluate the accuracy of the predictors. Many current generation secondary structure prediction methods have been developed and tested on this dataset. The RS126 set is available at *http://www.compbio.dundee.ac.uk/∼www-jpred/data/pred_res/126_set.html*.

4.2 Dataset 2 (CB396)

The second dataset generated by Cuff and Barton (13) at the European Bioinformatics Institute (EBI) consisted of 396 nonhomologous protein chains and was referred to as the CB396 set. Cuff and Barton used a rigorous method consisting of the computation of the similarity score to derive their non-redundant dataset. The CB396 set is available at *http://www.compbio.dundee.ac.uk/∼www-jpred/data/*.

4.3 Dataset 3 (PSIPRED)

We performed an extensive experiment on the PSIPRED dataset with 2245 protein chains for PSS prediction to compare two-stage MSVMs with PSIPRED method (12). The PSIPRED set is available at *ftp://bioinf.cs.ucl.ac.uk/pub/psipred/old*. We use the same PSI-BLAST profiles from the PSIPRED dataset to provide an objective comparison of two methods.

4.4 CASP4 and EVA Datasets

[Table 1 is to be included here.]

Two-stage MSVM approach has been tested more on CASP4 dataset (*http://predictioncenter.org/casp4/Casp4.html*) and a set of 64 new proteins provided by EVA (*http://cubic.bioc.columbia.edu/eva/*) (as shown in Table 1) based on the position specific scoring matrices generated by PSI-BLAST. These testing sets contains sequences with no homology to the proteins of the training set extracted from PSIPRED dataset in 1999 (12). The above training and testing sets were used to provide fair evaluation of the present approach.

4.5 Protein secondary structure definition

The type of the secondary structure for each residue in the training and testing sets was assigned from DSSP (30) which is the most widely used secondary structure definition. The eight types, H(α-helix), G(3_{10}-helix), I(π-helix), E(β-strand), B(isolated β-bridge), T(turn), S(bend), and -(rest), were reduced to three classes, α-helix (H), β-strand (E) and coil (C), by using the following: H and G to H; E and B to E; all others states to C. Recent methods for PSS prediction have used this reduction for evaluation (5; 6; 7; 8; 9; 10; 11; 12; 13; 14; 15; 17; 18; 19; 20) We adopted this reduction method to provide an objective comparison of the prediction accuracy of two-stage MSVMs to the results of other methods.

4.6 Prediction accuracy assessment

We have used several measures to evaluate the prediction accuracy. The Q_3 accuracy indicates the percentage of correctly predicted residues of three types of secondary structure (13):

$$Q_3(\%) = \frac{\sum_{t\in\Omega_T} a_t}{\sum_{t\in\Omega_T} b_t} \times 100, \tag{9}$$

where a_t is the number of correctly predicted residues in structure t, and b_t is the number of residues observed in type t. The Q_H, Q_E, and Q_C accuracies represent the percentages of correctly predicted residues of each type of secondary structure (13).

A complementary measure of prediction accuracy is obtained from the Matthews' correlation coefficients (31) for each of the three secondary structures: ρ_H, ρ_E, and ρ_C.

Segment overlap measure (Sov) gives the accuracy by counting the predicted and observed segments and measuring their overlap (32).

$$\text{Sov} = \frac{1}{n}\sum_{s} \frac{\text{minov}(s_{obs}; s_{pred}) + \delta(s_{obs}; s_{pred})}{\text{maxov}(s_{obs}; s_{pred})} \times \text{len}(s_{obs}), \tag{10}$$

where n is the total number of residues, s_{obs} and s_{pred} are the observed and predicted secondary structure segments respectively, and $\text{len}(s_{obs})$ is the number of residues in the segments s_{obs}.

We stick to the "per-protein" evaluation to provide an objective comparison with the previous approaches.

4.7 Results

[Table 2 is to be included here.]

[Table 3 is to be included here.]

[Table 4 is to be included here.]

Tables 2 and 3 show the performance of the single-stage MSVM and two-stage MSVM approaches, using different neighborhood input windows on the RS126 and CB396 datasets. As seen, the neighborhood windows, $w_1 = 15$ ($h_1^1 = h_2^1 = 7$), and $w_2 = 21$ ($h_1^2 = 2$ and $h_2^2 = 4$), gave the optimal accuracies at the first stage and the second stage, respectively. Table 4 shows that the prediction accuracies of two-stage MSVM with window lengths w_1 when the second stage window fixed at $w_2 = 21$ on the RS126 and CB396 datasets. The result confirmed that the window size at the input stage w_1 was optimal at 15.

[Table 5 is to be included here.]

[Table 6 is to be included here.]

Table 5 shows PSS prediction accuracies of the MSVM method with the Gaussian kernels $\mathcal{K}(\mathbf{x}, \mathbf{y}) = e^{-\sigma\|\mathbf{x}-\mathbf{y}\|^2}$ at $\sigma^1 = 0.05$ and $\sigma^2 = 0.01$, linear kernel $\mathcal{K}(\mathbf{x}, \mathbf{y}) = \mathbf{xy}$, and polynomial kernels $\mathcal{K}(\mathbf{x}, \mathbf{y}) = (\mathbf{xy} + c)^d$ at $d = 2, 3, 4$ and $c = 1$, on the RS126 and CB396 datasets with $\gamma^1 = \gamma^2 = 0.5$. The Gaussian kernel outperforms the other kernels. The kernel parameters were empirically determined for Gaussian kernel for the single-stage MSVM and two-stage MSVM approaches on the CB396 dataset (Table 6). As seen in Table 6, the optimal parameters were $\sigma^1 = 0.05$ for the first stage and $\sigma^2 = 0.01$ for the second stage, with $\gamma^1 = \gamma^2 = 0.5$.

[Figure 3 is to be included here.]

Figure 3 shows a comparison of performances of different secondary structure prediction methods on the RS126 dataset based on single sequences, multiple sequence alignments, and PSI-BLAST profiles. The methods are GOR I (1^{st} generation) (3), GOR III (2^{nd}) (4), GOR IV (2^{nd}) (5), Zpred (2^{nd}) (33), PHD (3^{rd}) (9), NNSSP (3^{rd}) (6), PREDATOR (3^{rd}) (34), DSC (3^{rd}) (35), Riis and Krogh (3^{rd}) (10), BRNN (3^{rd}) (11), Jpred (3^{rd}) (13), SVMfreq (3^{rd}) (18), and SVMpsi (3^{rd} ext) (19). The results of Zpred, NNSSP, PREDATOR, DSC, and Jpred methods on the RS126 dataset were obtained from Cuff and Barton (13) and the results of the refined neural network proposed by Riis and Krogh, SVMfreq, SVMpsi, BRNN, and PHD methods were obtained from their original publications (10; 18; 19; 11; 9). We implemented and tested GOR I, GOR III, and GOR IV techniques on the RS126 dataset with single sequences. The best Q_3 accuracy of 78.0% was achieved by the two-stage MSVM using the PSI-BLAST profiles. Comparing two-stage MSVM to two cascaded MLP networks of PHD method (9), a substantial gain of 7.2% of Q_3 accuracy was observed. The two-stage MSVM method obtained 6.8% and 1.9% higher Q_3 scores compared to the SVM methods of Hua and Sun (18) and Kim and Park (19), respectively.

[Figure 4 is to be included here.]

Figure 4 shows the performance of two-stage MSVM approach on the CB396 dataset based on multiple sequence alignments and PSI-BLAST profiles, compared to other approaches. Two-stage MSVM approach achieved 76.3% of Q_3 accuracy, which is the highest scores reported on the CB396 set to date. Compared to the recent method of Guo *et al.* using dual-layer binary SVMs (20), the two-stage MSVM method showed 2.3% higher Q_3 score. A direct comparison of Sov accuracy between the two methods was impossible because we used the Sov measure of Zemla *et al.* (32), which makes the evaluation of PSS prediction structurally more meaningful than Sov94 measure (36) used in the Guo *et al.*'s method.

[Table 7 is to be included here.]

As shown in Table 7, we performed an extensive experiment on the PSIPRED dataset for PSS prediction using two-stage MSVMs. For our knowledge, the PSIPRED dataset with 2252 protein chains is one of the largest dataset used for training and testing of PSS prediction. On this dataset, two-stage MSVM approach achieved 79.4% of Q_3 accuracy while the accuracy of PSIPRED method was previously reported to be 78.3% (12).

[Table 8 is to be included here.]

Since CASP4 did not publish the secondary structure of the target protein sequence, we only test our method on target protein sequences that can be found secondary structures from Protein Data Bank (PDB). Two-stage MSVM approach achieved 77.0% of accuracy on 31 target proteins of CASP4 dataset (as shown in Table 8). This result shown that our method has the ability to perform well on new protein sequences.

[Table 9 is to be included here.]

Table 9 shows the performance of two-stage MSVMs on a set of 64 new proteins provided by EVA with PSI-BLAST profiles. The best prediction was achieved to be two-stage MSVMs: 79.5% of Q_3. On the testing set, the Q_3 accuracy of two-stage MSVMs was substantial higher than results of PHD (72.3%), Prof_King (71.5%), PHDpsi (72.7%), and PSIPRED (77.4%) methods.

[Table 10 is to be included here.]

Table 10 lists the properties of 20 amino acids and their average occurrences and errors of the PSS prediction and probabilities of the presence of α-helices, β-strands, and coils on CB396 dataset. According to the statistics, amino acids, Val, Ile, and Met were easy to predict while Trp, Cys, and His were difficult to predict by our method. The statistical data shows that the amino acid residues in the non-polar group (hydrophobic), Gly, ALa, Val, Leu, Ile, Met, and Pro, were predicted with higher accuracies than

the other ones. The results from Table 10 suggest that five amino acids, Ala, Met, Gln, Arg, and Glu strongly tend to be α-helix, and Val and Ile tend to be β-strand, while Gly, Pro, Ser, Asn, and Asp are strong coil formers.

4.8 Two-stage Approaches to PSS Prediction

[Table 11 is to be included here.]

To demonstrate the MSVM's capacity to minimize the generalization error at the output of the single stage predictors, we experimented with the two-stage predictors by connecting MSVM at the output of the GOR I, GOR III, and GOR IV methods. The new prediction schemes achieved ∼2%-6% and ∼3%-15% improvements in Q_3 and Sov accuracies, respectively, on the CB396 dataset (see Table 11). The poorer performance of PHD method, which uses the two cascade MLPs, indicates that the MSVMs are better at both stages than the feedforward neural networks. We can infer that by connecting MSVM as the second stage of single stage approaches, the accuary of the prediction improves. From all the two-stage approaches tested, though all results are not shown, the two-stage MSVM approach showed the best performance for PSS prediction.

5 DISCUSSION AND CONCLUSION

We introduced a two-stage MSVM approach to PSS prediction, which improved the prediction accuracy because the secondary structure at a particular position of a sequence depends not only on the amino acid residue at the particular location but also on the structural formations of the rest of the sequence. This intrinsic relation cannot be captured by using only single-stage predictors alone. Therefore, another layer of predictors, which predicts the output from the results of the single-stage methods improves the accuracy. This has also been shown in PHD approach with MLPs earlier but with less prediction accuracy. MLPs are not optimal for PSS prediction because they cannot generalize the prediction for unseen amino acid sequences. As shown, the MSVM method was an optimal classifier for the second stage because it minimizes not only the empirical risk of known sequences but also the actual risk of unknown sequences. Additionally, two stages were proven to be sufficient to find an optimal classifier for PSS prediction as the MSVM minimized the generalization error at the output of the first stage by solving the optimization problems at the second stage.

Since the proof of minimal generalization error of two-stage MSVM in the theoretical part applies for any values of window sizes, without loss of generality, we assumed that the training and testing patterns at the first stage and second stage are drawn independently and identically according to probability distributions based on the optimal values of window sizes w_1 and w_2, respectively. Choosing the optimal value of window sizes w_1 and w_2 in the theoretical part makes the samples closer to independent and identical distribution. In practice, the optimal values of window sizes at the first stage and the second stage are found empirically.

Furthermore, we have compared two-stage SVM techniques for PSS problem: one method based on binary classifications of Guo (20) and the other approach for multiclass problem by solving one single optimization problem. We found that the two-stage MSVM approach is more suitable for PSS prediction because of its capacity to lead faster convergence for large and complex training sets of sequences and solve the optimization problem in one step. By incorporating the state-of-the-art single-stage MSVM classifiers, the two-stage MSVM approach, presented here, demonstrated the best accuracy over the earlier approaches to PSS prediction on the tested data.

The kernels and their parameters were empirically determined as there do not exist simple methods to find them otherwise. The Gaussian kernels outperformed the other kernels in the experiments with benchmark datasets; Based on that since the Gaussian kernels are local, the local interactions of amino acid residues and structural elements were more involved the prediction of the secondary structures by our method than the long distance interactions. Investigation into the determination of optimal parameters of MSVMs at the two stages could further enhance the accuracies of prediction. The two-stage MSVM approach outperformed earlier techniques of PSS prediction on the tested datasets and had better generalization capabilities, which could be used to aid the prediction the 3D structures of proteins. Web server for the prediction of PSS from amino acid sequences by using the two-stage MSVM approach has been developed and is available at *http://birc.ntu.edu.sg/∼pas0186457*.

6 APPENDIX

We consider the case where $h_1^2 = h_2^2 = 0$ and $f_2^{k*}(\mathbf{d}) = \ln\frac{d^k}{1-d^k}$. It follows that $\mathbf{d} = \mathbf{d}_i$ and $f_2^{k*}(\mathbf{d}) = \ln\left(e^{-f_1^k(\mathbf{r}_i)}\right)^{-1} = f_1^k(\mathbf{r})$. Thus $f_2^*(\mathbf{d}) = f_1(\mathbf{r})$ where $f_2^*(\mathbf{d}) = \max_{k\in\Omega_T} f_2^{k*}(\mathbf{d})$. The generalization error, $\mathrm{err}_{\mathcal{P}_2}(f_2^*)$, can now be written as

$$\begin{aligned}
\mathrm{err}_{\mathcal{P}_2}(f_2^*) &= \int_{L=0} L(f_2^*(\mathbf{d}),t)d\mathcal{P}_2(\mathbf{d},t) \\
&+ \int_{L=1} L(f_2^*(\mathbf{d}),t)d\mathcal{P}_2(\mathbf{d},t), \\
&= \int_{L=1} L(f_2^*(\mathbf{d}),t)d\mathcal{P}_2(\mathbf{d},t), \\
&= \int_{L=1} L(f_1(\mathbf{r}),t)d\mathcal{P}_1(\mathbf{r},t), \\
&= \int L(f_1(\mathbf{r}),t)d\mathcal{P}_1(\mathbf{r},t), \\
&= \mathrm{err}_{\mathcal{P}_1}(f_1).
\end{aligned}$$

As a result, there exists at least a discriminant function f_2^* such that $\mathrm{err}_{\mathcal{P}_2}(f_2^*) = \mathrm{err}_{\mathcal{P}_1}(f_1)$.

REFERENCES

[1] P. Clote and R. Backofen, *Computational Molecular Biology*, Wiley and Sons, Ltd., Chichester, 2000.

[2] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, 2001.

[3] J. Garnier, D. J. Osguthorpe, and B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *Journal of Molecular Biology*, vol 120, pp 97–120, 1978.

[4] J. F. Gibrat, J. Garnier, and B. Robson, Further developments of protein secondary structure prediction using information theory, *Journal of Molecular Biology*, vol 198, pp 425–443, 1987.

[5] J. Garnier, J. F. Gibrat, and B. Robson, GOR method for predicting protein secondary structure from amino acid sequence, *Methods Enzymol*, vol 266, pp 541–553, 1996.

[6] A. A. Salamov and V. V. Solovyev, Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments, *Journal of Molecular Biology*, vol 247, pp 11–15, 1995.

[7] A. A. Salamov and V. V. Solovyev, Protein secondary structure prediction using local alignments, *Journal of Molecular Biology*, vol 268, pp 31–36, 1997.

[8] S. C. Schmidler, J. S. Liu, and D. L. Brutlag, Bayesian segmentation of protein secondary structure, *Journal of Computational Biology*, vol 7, pp 233–248, 2000.

[9] B. Rost and C. Sander, Prediction of protein secondary structure at better than 70% accuracy, *Journal of Molecular Biology*, vol 232, pp 584–599, 1993.

[10] S. K. Riis and A. Krogh, Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignment, *Journal of Computational Biology*, vol 3, pp 163–183, 1996.

[11] P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, Exploiting the past and the future in protein secondary structure prediction, *Bioinformatics*, vol 15, pp 937–946, 1999.

[12] D. T. Jones Protein secondary structure prediction based on position-specific scoring matrices, *Journal of Molecular Biology*, vol 292, pp 195–202, 1999.

[13] J. A. Cuff and G. J. Barton, Evaluation and improvement of multiple sequence methods for protein secondary structure prediction, *Proteins*, vol 4, pp 508–519, 1999.

[14] M. Ouali and R. D. King, Cascaded multiple classifiers for secondary structure prediction, *Protein Science*, vol 9, pp 1162–1176, 1999.

[15] Y. Guermeur and D. Zelus, Combining protein secondary structure prediction methods with a new multi-category SVM, *International conference on Intelligent Systems for Molecular Biology*, San Diego, 2000.

[16] G. B. Fogel and D. W. Corne, *Evolutionary Computation in Bioinformatics*, Morgan Kaufmann, 2002.

[17] J. Meiler and D. Baker, Coupled prediction of protein secondary and tertiary structure, *Protein Science*, vol 100:21, pp 12105–12110, 2003.

[18] S. Hua and Z. Sun, A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach, *Journal of Molecular Biology*, vol 308, pp 397–407, 2001.

[19] 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.

[20] J. Guo, Y. Chen, Z. Sun, and Y. Lin, A novel method for protein secondary structure prediction using dual-layer SVM and profiles, *Proteins: Structure, Function, and Bioinformatics*, vol 54, pp 738–743, 2004.

[21] K. Crammer and Y. Singer, On the Learnability and Design of Output Codes for Multiclass Problems, *Machine Learning*, vol 47, pp 201–233, 2002.

[22] M. N. Nguyen and J. C. Rajapakse, Two-stage multi-class SVMs for protein secondary structure prediction, *Pacific Symposium on Biocomputing (PSB)*, Hawaii, USA, January 4-8, 2005.

[23] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, 2000.

[24] J. C. Platt, N. Cristianini and J. Shawe-Taylor, Large margin DAG's for multiclass classification, *in* Proc. Advances in Neural Information Processing Systems 12. Cambridge, MA:MIT Press, pp 547–553, 2000.

[25] V. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York, 1982.

[26] V. Vapnik, *Statistical Learning Theory*, Wiley and Sons, Inc., New York, 1998.

[27] B. Rost, Rising accuracy of protein secondary structure prediction, *Protein structure determination, analysis, and modeling for drug discovery*, (ed. D Chasman), New York: Dekker, pp 207–249, 2003.

[28] J. A. Cuff and G. J. Barton, Application of multiple sequence alignment profiles to improve protein secondary structure prediction, *Proteins: Structure, Function, and Genetics*, vol 40, pp 502–511, 2000.

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

[30] 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.

[31] B. Matthews, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys, Acta*, vol 405, pp 442–451, 1975.

[32] 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: Structure, Function, and Genetics*, vol 34, pp 220–223, 1999.

[33] M. J.J.M. Zvelebil, G. J. Barton, W. R. Taylor, and M. J.E. Sternberg, Prediction of protein secondary structure and active sites using the alignment of homologous sequences, *Journal of Molecular Biology*, vol 195, pp 957–961, 1987.

[34] D. Frishman and P. Argos, Knowledge-based secondary structure assignment, *Proteins: Structure, Function, and Genetics*, vol 23, pp 566–579, 1995.

[35] R. D. King and M. J.E. Sternberg, Identification and application of the concepts important for accurate and reliable protein secondary structure prediction, *Prot[illegible] Science*, vol 5, pp 2298–2310, 1996.

[36] B. Rost and C. Sander and R. Schneider, [illegible]edefining the goals of protein secondary structure prediction, *Journal of Molecular Biology*, vol 2[illegible] pp1[illegible]–26, 1994.

1hk9A	1i85A	1izmA	1j0wA	1lmmA	1mwqA	1nngA	1nnvA
1nslA	1nxhA	1oj5A	1oyiA	1p57A	1p5hA	1p94A	1pc0A
1pd3A	1pg6A	1pjvA	1pv6A	1pw4A	1px5A	1pzqA	1pzrA
1q3jA	1q3kA	1q68A	1q7sA	1q90L	1q90M	1q90N	1qw2A
1r2mA	1r4gA	1rh5B	1rhzA	1rifA	1rjiA	1rklA	1rocA
1rpuA	1rqtA	1rwsA	1s0yB	1s4zC	1s5lI	1s5lJ	1s5lL
1s5lM	1s5lX	1s5lZ	1s68A	1s6cB	1s7bA	1uf3A	1ufiA
1uhwA	1ujxA	1usmA	1v74A	1v74B	1vjqA	1ocsA	1rf8B

Table 1: The list of 64 proteins of the EVA dataset used for testing the two-stage MSVM approach.

Dataset	**Accuracy**	**Window w_1** 9	11	13	15	17	19	21	23	25	27
RS126	Q_3	75.8	76.0	76.1	**76.2**	**76.2**	76.0	75.7	75.5	75.2	75.0
	Sov	68.2	68.5	68.7	68.8	**68.9**	68.4	67.0	66.7	66.2	66.0
CB396	Q_3	74.2	74.3	**74.5**	**74.5**	74.3	74.2	73.9	73.7	73.5	73.2
	Sov	68.6	69.0	69.4	**69.5**	69.3	69.0	68.7	68.4	68.2	67.8

Table 2: Accuracies of PSS prediction by the single-stage MSVM approach on the RS126 and CB396 datasets. The window length indicates the size of neighborhood taken as the input for the single-stage approach.

Dataset	**Accuracy**	**Window w_2** 3	6	9	12	15	18	21	24	27	30
RS126	Q_3	76.4	76.9	77.2	77.5	77.7	77.9	**78.0**	77.9	77.8	77.8
	Sov	68.9	70.4	71.1	71.6	72.1	72.4	**72.6**	72.4	72.1	72.0
CB396	Q_3	74.6	75.2	75.5	75.8	76.0	76.1	**76.3**	**76.3**	76.2	76.0
	Sov	69.7	71.8	72.5	72.7	72.9	73.1	**73.2**	73.1	**73.2**	73.0

Table 3: Accuracies of PSS prediction by the two-stage MSVM approach on the RS126 and CB396 datasets at different neighborhood window sizes of the second stage. The size of the first stage neighborhood was taken as 15.

Dataset	Method	Window w_1 9	11	13	15	17	19	21	23	25	27
RS126	Single-stage	75.8	76.0	76.1	**76.2**	**76.2**	76.0	75.7	75.5	75.2	75.0
	Two-stage	77.6	77.8	77.8	**78.0**	**78.0**	77.9	77.7	77.6	77.4	77.2
CB396	Single-stage	74.2	74.3	**74.5**	**74.5**	74.3	74.2	73.9	73.7	73.5	73.2
	Two-stage	76.1	76.1	**76.3**	**76.3**	76.2	76.1	75.9	75.8	75.6	75.4

Table 4: Accuracies of PSS prediction by the single-stage and two-stage MSVM approaches with different window lengths w_1. The second stage window size w_2 was maintained at 21 on the RS126 and CB396 datasets.

Dataset	Method	Kernel Function Gaussian $\sigma_1 = 0.05$ $\sigma_2 = 0.01$	Linear	Polynomial $d = 2$	Polynomial $d = 3$	Polynomial $d = 4$
RS126	Single-stage	**76.2**	73.5	74.1	74.8	75.4
	Two-stage	**78.0**	76.4	76.8	77.1	77.3
CB396	Single-stage	**74.5**	72.6	72.9	73.1	73.4
	Two-stage	**76.3**	74.8	75.2	75.4	75.7

Table 5: Comparison of Q_3 accuracies of the single-stage and two-stage MSVM approaches with different type of kernel functions on RS126 and CB396 datasets with parameters $\gamma_1 = \gamma_2 = 0.5$. The neighborhood windows of size 15 and 21 were used at the first stage and second stage MSVMs, respectively.

Parameter γ	0.5				1.0	1.5	2.0
Gaussian σ	0.01	0.05	0.1	0.15	0.01	0.01	0.01
Single-stage	74.2	**74.5**	74.1	73.9	73.8	73.6	73.5
Two-stage	**76.3**	76.2	76.0	75.8	**76.3**	76.1	75.9

Table 6: Comparison of Q_3 accuracies of the single-stage and two-stage MSVM approaches with different parameters for Gaussian kernels on the CB396 dataset.

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
Jones (PSIPRED)	78.3	-	-	-	-	-	-	-
Two-stage MSVM	79.4	76.2	79.0	62.2	82.1	0.67	0.60	0.61

Table 7: Comparison of performances of two-stage MSVM approach with PSIPRED method on the PSIPRED dataset.

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
Two-stage MSVM	77.0	72.1	73.2	47.8	79.2	0.61	0.67	0.57

Table 8: Performances of two-stage MSVM approach on 31 target proteins of CASP4 dataset.

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
Rost and Sander (PHD)	72.3	68.9	65.4	37.8	67.7	0.65	0.67	0.51
Przybylski and Rost (PHDpsi)	72.7	69.1	65.8	37.8	67.8	0.65	0.67	0.51
Ouali and King (Prof_King)	71.5	69.1	60.1	39.6	73.2	0.62	0.68	0.50
Jones (PSIPRED)	77.4	75.8	73.4	36.0	72.3	0.70	0.69	0.55
Two-stage MSVM	79.5	73.8	83.8	74.5	80.2	0.72	0.71	0.61

Table 9: Comparison of performances of two-stage MSVM approach with other methods on 64 new proteins of EVA dataset.

Amino acid		**Occurrence** (%)	**Error in PSS prediction** (%)	**Helix** (%)	**Strand** (%)	**Coil** (%)
Non-polar R group (hydrophobic)						
Gly	G	7.8	21.3	14.0	8.9	77.1
Ala	A	8.8	21.0	47.9	14.0	38.1
Val	V	6.9	20.2	30.9	42.7	26.4
Leu	L	8.6	21.4	47.1	24.9	28.0
Ile	I	5.5	20.6	36.4	37.8	25.8
Met	M	2.1	20.5	45.3	20.9	33.8
Pro	P	4.6	21.9	14.0	6.8	79.2
Aromatic R group (hydrophobic)						
Phe	F	3.9	23.6	33.6	30.9	35.5
Trp	W	1.5	26.2	33.0	27.0	40.0
Tyr	Y	3.7	25.1	34.6	28.2	37.2
Polar, uncharged R group (hydrophilic)						
Ser	S	6.1	24.3	27.2	13.8	59.0
Thr	T	5.9	25.3	25.4	24.3	50.3
Cys	C	1.5	28.0	26.0	25.8	48.2
Asn	N	4.7	22.4	24.1	9.9	66.0
Gln	Q	3.7	22.4	35.1	13.7	41.2
Positively R charged (hydrophilic)						
Lys	K	5.8	23.3	38.7	14.7	46.7
Arg	R	4.7	23.3	32.0	18.2	39.8
His	H	2.2	27.0	28.3	19.1	52.6
Negatively R charged (hydrophilic)						
Asp	D	6.0	22.7	28.1	8.5	63.4
Glu	E	6.1	22.6	49.3	12.5	38.3

Table 10: The properties of 20 amino acids: their average occurrences, the errors in prediction, and the occurrence probabilities of α-helices, β-strands, and coils, in the CB396 dataset.

Method	Q_3 (%)	Sov (%)	Q_H (%)	Q_E (%)	Q_C (%)	ρ_H	ρ_E	ρ_C
Single-Stage								
GOR I	64.3	58.5	73.0	41.6	66.1	0.44	0.43	0.43
GOR III	66.0	52.3	75.9	43.2	66.4	0.46	0.45	0.44
GOR IV	65.2	62.9	71.3	55.0	62.2	0.53	0.42	0.44
MSVM	74.5	69.5	68.5	62.0	82.4	0.61	0.59	0.55
Two-Stage								
GOR I - MSVM	67.2	63.3	63.6	51.7	75.5	0.50	0.47	0.47
GOR III - MSVM	71.3	67.1	68.6	55.1	77.8	0.57	0.52	0.41
GOR IV - MSVM	70.9	66.5	68.1	54.0	78.9	0.57	0.51	0.51
MSVM - MSVM	76.3	73.2	70.6	63.4	83.4	0.63	0.62	0.57
PHD	71.9	-	-	-	-	-	-	-

Table 11: Comparison of performances of single-stage and two-stage approaches with a MSVM at the second stage in the PSS prediction on the CB396 dataset.

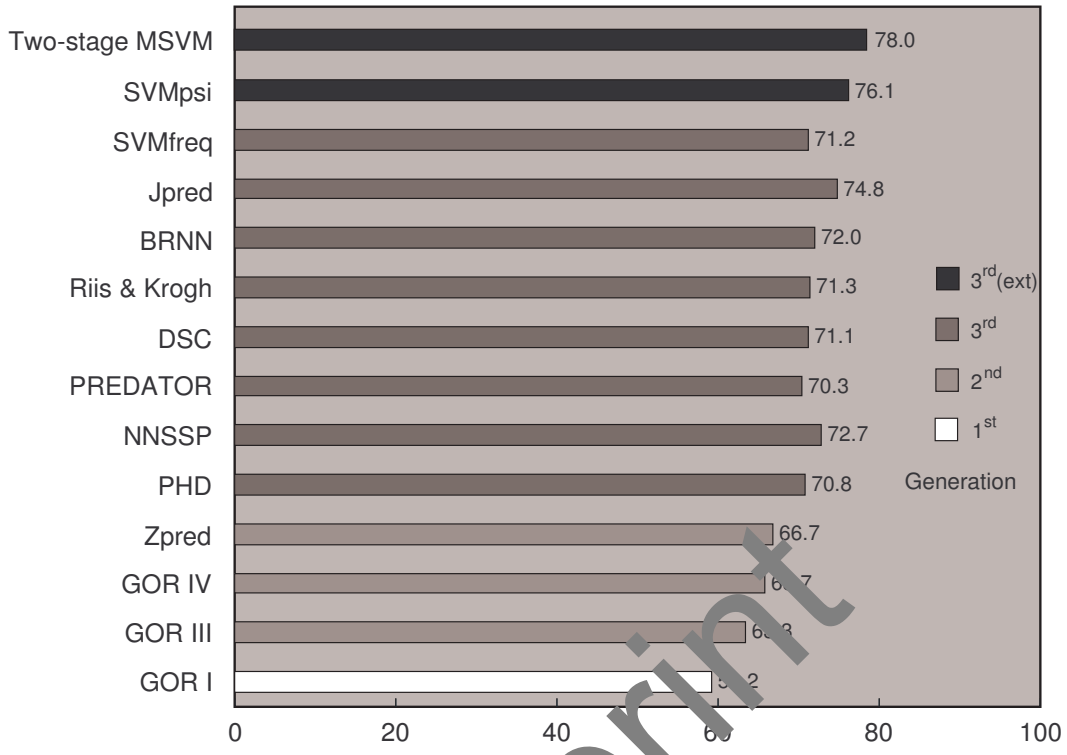

Figure 3: Comparison of Q_3 accuracies of different predictors in PSS prediction on RS126 dataset of 126 nonhomologous globular proteins. The classification of the approaches of PSS prediction as the first generation, second generation, and third generation is based on the paper of Rost (27). The notation (ext) indicates that the corresponding method uses position specific scoring matrices generated by PSI-BLAST instead of multiple sequence alignments.

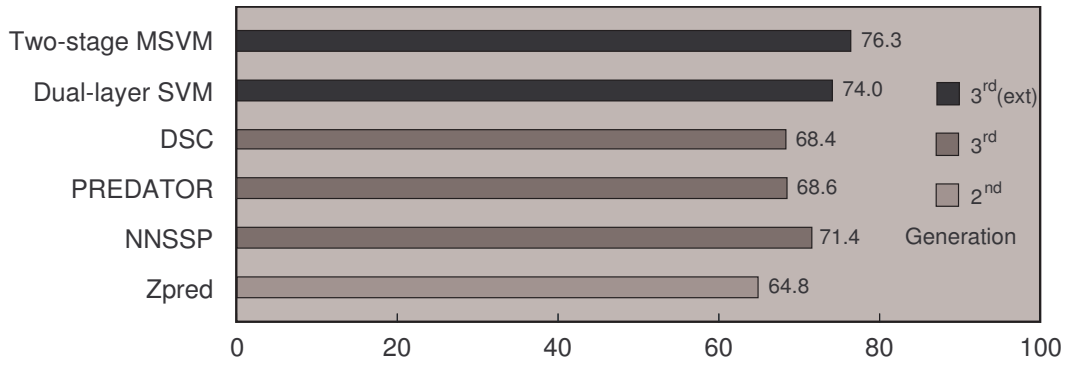

Figure 4: Comparison of Q_3 accuracies of different predictors of protein secondary structure on CB396 dataset of 396 nonhomologous proteins.