Two-Stage Multi-Class Support Vector Machines to Protein Secondary Structure Prediction

M.N. Nguyen and J.C. Rajapakse

Pacific Symposium on Biocomputing 10:346-357(2005)

TWO-STAGE MULTI-CLASS SUPPORT VECTOR MACHINES TO PROTEIN SECONDARY STRUCTURE PREDICTION

M. N. NGUYEN AND J. C. RAJAPAKSE
BioInformatics Research Centre,
School of Computer Engineering,
Nanyang Technological University, Singapore 639798
E-mail: asjagath@ntu.edu.sg

Bioinformatics techniques to protein secondary structure (PSS) prediction are mostly single-stage approaches in the sense that they predict secondary structures of proteins by taking into account only the contextual information in amino acid sequences. In this paper, we propose two-stage Multi-class Support Vector Machine (MSVM) approach where a MSVM predictor is introduced to the output of the first stage MSVM to capture the sequential relationship among secondary structure elements for 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 highest scores published on both datasets to date.

1 Introduction

One of the major goals of bioinformatics is to predict the three-dimensional (3-D) structure of a protein from its amino acid sequence. 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, which broadly fall into three categories: (1) statistical methods, (2) neural network approaches, and (3) nearest neighbor methods. The statistical methods are mostly based on likelihood techniques [3,4,5]. Neural networks use residues in a local neighborhood or a window, as inputs, to predict the secondary structure at a particular location of an amino acid sequence by finding an appropriate non-linear mapping [6,7,8,9]. The nearest neighbor approach often uses the k-nearest neighbor techniques [10,11]. The consensus approaches that combine different classifiers, parallely,

into a single superior predictor have been proposed for PSS prediction [12,13]. Support Vector Machines (SVMs) have been earlier applied to PSS prediction [14,15]; one of the drawbacks in these approaches is that the methods do not take into account the sequential relationship among the protein secondary structure elements. Additionally, SVM methods only construct a multi-class classifier by combining several binary classifiers.

Most existing secondary structure techniques are single-stage approaches, which are unable to find complex relations (correlations) among structural elements in the sequence. This could be improved by incorporating the interactions or contextual information among the elements of the sequences of secondary structures. We argue that it is feasible in enhancing the present single-stage MSVM approach farther by augmenting with another prediction scheme at their outputs and propose to use MSVM as the second-stage. By using the position specific scoring matrices generated by PSI-BLAST, the two-stage MSVM approach significantly achieves Q_3 accuracies of 78.0% and 76.3% on the RS126 and CB396 datasets, based on a seven-fold cross validation.

2 Two-Stage MSVM Approach

In the two-stage MSVM approach, we use 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 $r_i \in \Sigma_R$ and Σ_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 $t_i \in \Sigma_T$ and $\Sigma_T = \{H, E, C\}$; 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 Σ_R^n to the space of Σ_T^n .

Let $\mathbf{v}_i$ be the vector representing 21-dimensional coding of 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 the padding space to indicate the overlapping end of the sequence [9]. 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 $\mathbf{v}_i$ denote 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

A MSVM scheme has been proposed by Crammer and Singer [16]. For PSS prediction, this method constructs three discriminant functions but all are

obtained by solving one single optimization problem, which can be formulated as follows:
Minimize

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

subject to the constraints

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

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

We find the minimization of the above formulation by solving the following quadratic programming problem [16]:

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

$$\text{such that }\sum_{k\in\Sigma_T}\alpha_j^k=0 \text{ and } \alpha_j^k\leq\begin{cases}0 & \text{if } t_j\neq k\\ \gamma^1 & \text{if } t_j=k\end{cases} \tag{3}$$

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 $\mathbf{w}_1^k=\sum_{j=1}^{N}\alpha_j^k\phi^1(\mathbf{r}_j)$.

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

$$f_1^k(\mathbf{r}_i)=\sum_{j=1}^{N}\alpha_j^k\mathcal{K}^1(\mathbf{r}_i,\mathbf{r}_j)=\mathbf{w}_1^k\phi^1(\mathbf{r}_i) \tag{4}$$

Let $f_1(\mathbf{r}_i)=\arg\max_{k\in\Sigma_T}f_1^k(\mathbf{r}_i)$. In the single-stage MSVM method, the secondary structural type t_i corresponding to the residue at site i, r_i, is determined by

$$t_i=f_1(\mathbf{r}_i) \tag{5}$$

The function, f_1, discriminates the type of PSS, based on the features or interactions among the residues in the input pattern. With optimal parameters, the MSVM attempts to minimize the generalization error in the prediction. If the training and testing patterns are drawn independently and

identically according to a probability distribution $\mathcal{P}$, then the generalization error, $\mathrm{err}_{\mathcal{P}}(f_1)$, is given by

$$\mathrm{err}_{\mathcal{P}}(f_1) = \mathcal{P}\{(\mathbf{r}, t) : \ f_1(\mathbf{r}) \neq t; (\mathbf{r}, t) \in \Gamma^1 \times \{H, E, C\}\}$$

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 $\mathbf{r}$.

2.2 Second Stage

We extend the single-stage MSVM technique by cascading another MSVM at the output of the present single-stage approach to improve the accuracy of prediction. This is because the secondary structure at a particular position of the sequence depends on the structures of the rest of the sequence, i.e., it accounts for the fact that the strands span over at least three adjacent residues and helices consist of at least four consecutive residues [6]. This intrinsic relation cannot be captured by using only single-stage approaches alone. Therefore, another layer of classifiers that minimize the generalization error of the output of single-stage methods by incorporating the sequential relationship among the protein structure elements improves the prediction accuracy.

Consider a window of w_2 size at a site of the output sequence of 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 $w_2 = 3(h_1^2 + h_2^2 + 1)$, $d^k_i = 1/(1 + e^{-f_1^k(\mathbf{r}_i)})$, and f_1^k denotes the discriminant function of the first stage. The application of the logistic sigmoid function to the outputs of the first stage has the advantage of constraining the input units of the second stage to the (0,1) interval that is similar to the range of the input units of the first stage. The purpose of this choice is to easier determine parameters for optimal performance. 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 hidden outputs in the higher dimensional space are linearly combined by a weight vector, $\mathbf{w}_2$, to obtain the prediction output. Let the training set of exemplars for the second stage MSVM be $\Gamma^2_{\text{train}} = \{\mathbf{d}_j : j = 1, \ldots, N\}$. The vector $\mathbf{w}_2$ is obtained by solving the following convex quadratic programming problem, over all the patterns seen in the training phase [16].

Let $f_2(\mathbf{d}_i) = \arg\max_{k \in \Sigma_T} f_2^k(\mathbf{d}_i)$. The secondary structural type t_i corresponding to the residue r_i is determined by

$$t_i = f_2(\mathbf{d}_i) \tag{6}$$

If the set of input patterns for the second stage MSVM in both training and testing phases is denoted by Γ^2, the generalization error of the two-stage MSVM approach, $\mathrm{err}_{\mathcal{P}}(f_2)$, is given by

$$\mathrm{err}_{\mathcal{P}}(f_2) = \mathcal{P}\{(\mathbf{d}, t) : \ f_2(\mathbf{d}) \neq t; (\mathbf{d}, t) \in \Gamma^2 \times \{H, E, C\}\}$$

If the input pattern $\mathbf{d}$ corresponds to a site i, then $\mathbf{d} = \mathbf{d}_i = ((1 + e^{-f_1^k(\mathbf{r}_{i-h_1^2})})^{-1}, (1 + e^{-f_1^k(\mathbf{r}_{i-h_1^2+1})})^{-1}, \ldots, (1 + e^{-f_1^k(\mathbf{r}_i)})^{-1}, \ldots, (1 + e^{-f_1^k(\mathbf{r}_{i+h_2^2})})^{-1})$. That is, the second stage takes into account the influences of the PSS values of residues in the neighborhood into the prediction. It could be easily shown that there exists a function f_2 such that $\mathrm{err}_{\mathcal{P}}(f_2) = \mathrm{err}_{\mathcal{P}}(f_1)$ when $h_1^2 = h_2^2 = 0$.

3 Minimal Generalization Error

In this section, we find a function f_2 that minimizes the generalization error $\mathrm{err}_{\mathcal{P}}(f_2)$ when connecting another MSVM predictor at the output of the existing predictor. The optimal function f_2 providing the smallest $\mathrm{err}_{\mathcal{P}}(f_2)$ ensures $\mathrm{err}_{\mathcal{P}}(f_2) \leq \mathrm{err}_{\mathcal{P}}(f_1)$. However, finding the global minimum of generalization error $\mathrm{err}_{\mathcal{P}}(f_2)$ is not a trivial problem because the form of the probability distribution $\mathcal{P}$ is unknown. We can instead consider the probably approximately correct (pac) bound, $\epsilon(N, \delta)$, of the generalization error satisfying

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

This is equivalent to asserting that with probability greater than $1 - \delta$ over the training set $\Gamma^2_{\mathrm{train}}$, the generalization error of f_2 is bounded by

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

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

Theorem 3.1. [17] Let $\mathcal{F} = \{f_2^{k/l} : \mathbf{d} \rightarrow \mathbf{w}_2^{k/l}\phi^2(\mathbf{d}); \|\mathbf{w}_2^{k/l}\| \leq 1; \mathbf{d} \in \Gamma^2; k, l \in \Sigma_T\}$ be restricted to points in a ball of m dimensions of radius R about the origin, that is $\phi^2(\mathbf{d}) \in \mathbb{R}^m$ and $\|\phi^2(\mathbf{d})\| \leq R$. Then the fat-shattering dimension is bounded by

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

Theorem 3.2. [18] 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 functions $f_2^{k/l} \in \mathcal{F}$ at decision nodes k/l, where $\eta_2^{k/l} = \min_{\mathbf{d}\in\Gamma^2_{\text{train}}} \frac{|\mathbf{w}_2^{k/l}\phi^2(\mathbf{d})|}{\|\mathbf{w}_2^{k/l}\|}$, k and $l \in \Sigma_T$. Then, the following probability is bounded by

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

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

Theorem 3.3. Let G be a decision directed acyclic graph with discriminant functions $f_2^{k/l} \in \mathcal{F}$ at nodes k/l, k and $l \in \Sigma_T$. Then, the generalization error of f_2 where $f_2(\mathbf{d}) = \arg\max_{k\in\Sigma_T} f_2^k(\mathbf{d})$ in the probability distribution $\mathcal{P}$ is

$$\text{err}_{\mathcal{P}}(f_2) = \text{err}_{\mathcal{P}}(G)$$

Proof. This can be easily proved 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. ■

Theorem 3.4. [19] Let $\text{err}_{\mathcal{P}}(G)$ be the generalization error of G at the output of the first stage. Then

$$\mathcal{P}\left\{\Gamma^2_{\text{train}} : \exists G \text{ such that } \text{err}_{\Gamma^2_{\text{train}}}(G) = 0 \text{ and } \text{err}_{\mathcal{P}}(G) > 2\epsilon(N,\delta)\right\} \leq$$

$$2\mathcal{P}\left\{\Gamma^2_{\text{train}}, \Gamma^2_{\text{test}} : \exists G \text{ such that } \text{err}_{\Gamma^2_{\text{train}}}(G) = 0 \text{ and } \text{err}_{\Gamma^2_{\text{test}}}(G) > \epsilon(N,\delta)\right\}$$

Theorem 3.5. Suppose we classify a random N examples in the training set Γ^2_{train} using the MSVM method at second stage with optimal values of weight vectors $\mathbf{w}_2^k$, $k \in \Sigma_T$. Then, the generalization error $\text{err}_{\mathcal{P}}(f_2)$ with

probability greater than $1-\delta$ is bound to be less than

$$\epsilon(N,\delta) = \frac{1}{N}\left(390R^2\sum_{k\in\Sigma_T}\|\mathbf{w}_2^k\|^2\log(4eN)\log(4N) + 2\log\frac{2(2N)^3}{\delta}\right)$$

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 $M = \sum_{k,l}\frac{1}{(\eta_2^{k/l})^2} \leq \sum_{k,l}\|\mathbf{w}_2^k-\mathbf{w}_2^l\|^2 \leq 3\sum_k\|\mathbf{w}_2^k\|^2$. Solving the optimization problems at second stage results in the minimization of the quantity M which is directly related to the margin of the classifier. Plugging the binary classifiers induced by $\mathbf{w}_2^{k/l}$ results a stepwise method for calculating the maximum among $\{f_2^k(\mathbf{d}) = \mathbf{w}_2^k\phi^2(\mathbf{d})\}$ 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 (3.2) for G with specified margin $\eta_2^{k/l}$ at each node to bound the generalization error $\text{err}_{\mathcal{P}}(G)$. Since the number of decision nodes is 3 and the largest allowed value of $a^{k/l}$ is N, the number of all possible patterns of $a^{k/l}$'s over the decision nodes is bounded by N^3. We let $\delta_i = \delta/N^3$ so that the sum $\sum_{i=1}^{N^3}\delta_i = \delta$. By choosing $\epsilon(N, \frac{\delta_i}{2})$

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

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

from Theorem (3.1)

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

Theorem (3.2) ensures that the probability of any of the statements failing to hold is less than $\delta/2$. By using the result of the Theorem (3.4), the probability $\mathcal{P}\{\Gamma^2_{\text{train}} : \exists G;\ \text{err}_{\Gamma^2_{\text{train}}}(G) = 0;\ \text{err}_{\mathcal{P}}(G) > 2\epsilon(N,\delta_i/2)\}$ is bound to be less than δ. From Theorem (3.3), the generalization error $\text{err}_{\mathcal{P}}(f_2)$ with probability greater than $1-\delta$ is bound to be less than $2\epsilon(N,\delta_i/2)$. ■

Minimizing the quantity $\sum_{k\in\Sigma_T} \|\mathbf{w}_2^k\|^2$, that is, maximizing the value of margin $\eta_2^{k/l}$ results in the minimization of the generalization error of the single stage MSVM method. Minimization of $\sum_{k\in\Sigma_T} \|\mathbf{w}_2^k\|^2$ is done by solving the convex quadratic programming problem of MSVM. As shown in the result of Theorem (3.5), two-stage MSVMs are sufficient for PSS prediction because they minimize both the generalization error $\mathrm{err}_{\mathcal{P}}(f_1)$ based on interactions among amino acids and the generalization error $\mathrm{err}_{\mathcal{P}}(f_2)$ of the output of the single-stage MSVM by capturing the contextual information of secondary structure.

4 Experiments and Results

4.1 Dataset 1 (RS126)

The set 126 nonhomologous globular protein chains, used in the experiment of Rost and Sander [6] and referred to as the RS126 set, was used to evaluate the accuracy of the classifiers. The RS126 set is available at *http://www.compbio.dundee.ac.uk/∼www-jpred/data/pred_res/126_set.html*. The single-stage and two-stage MSVM approaches were implemented, with the position specific scoring matrices generated by PSI-BLAST, and tested on the dataset, using a seven-fold cross validation to estimate the prediction accuracy.

4.2 Dataset 2 (CB396)

The second dataset generated by Cuff and Barton [12] 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 on the computation of the similarity score to derive their nonredundant dataset. The CB396 set is available at *http://www.compbio.dundee.ac.uk/∼www-jpred/data/*. The single-stage and two-stage MSVM approaches have been used to predict PSS based on the position specific scoring matrices generated by PSI-BLAST.

4.3 Protein secondary structure definition

The secondary structure states for each structure in the training and testing sets were assigned from DSSP [20] that is the most widely used secondary structure definition. The eight states, 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 method: H and G to H; E and B to E; all others states to C.

4.4 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 states of secondary structure [12]. The Q_H, Q_E, Q_C accuracies represent the percentage of correctly predicted residues of each type of secondary structure [12]. Segment overlap measure (Sov) gives accuracy by counting predicted and observed segments, and measuring their overlap [21].

4.5 Results

For MSVM classifier at the first stage, a window size of 15 amino acid residues ($h_1^1 = h_2^1 = 7$) was used as input for optimal result in the [7, 21] range. At the second stage, the window size of width 21 ($h_1^2 = 2$ and $h_2^2 = 4$) in the [9, 24] range gave the optimal accuracy. The kernel selected here was the radial basis function $\mathcal{K}(\mathbf{x}, \mathbf{y}) = e^{-\sigma\|\mathbf{x}-\mathbf{y}\|^2}$ with the parameters: $\sigma = 0.05$, $\gamma^1 = 0.5$ for MSVM at the first stage, and $\sigma = 0.01$, $\gamma^2 = 0.5$ for two-stage MSVMs, determined empirically for optimal performance in the [0.01, 0.5] and [0.1, 2] ranges, respectively. We used BSVM library [22], which leads to faster convergence for large optimization problem, to implement the multi-class technique.

In tables 1 and 2, the results of Zpred, NNSSP, PREDATOR, DSC and Jpred methods on the RS126 and CB396 datasets were obtained from Cuff and Barton [12]. The results of the refined neural network proposed by Riis and Krogh, SVM method of Hua and Sun, dual-layer SVM of Guo, BRNN, and PHD methods were obtained from their papers [6,7,8,14,15].

Table 1 shows the performance of the different secondary structure predictors and two-stage MSVM approach on the RS126 set. The best algorithm was found to be the cascade of two MSVMs with the PSI-BLAST profiles, which achieved 78.0% of Q_3 accuracy. Comparing two-stage MSVMs to two multi-layer perceptron networks of PHD method proposed by Rost and Sander [6], a substantial gain of 7.2% of Q_3 accuracy was observed. Compared to SVM method of Hua and Sun [14], the two-stage MSVM method obtained 6.8% higher Q_3 score.

Table 2 shows the performance of two-stage MSVMs with the CB396 dataset based on multiple sequence alignments and PSI-BLAST profiles. Two-stage MSVMs with PSI-BLAST profiles achieved 76.3% of Q_3 accuracy that is

Table 1. Comparison of performances of single-stage and two-stage MSVM approaches in PSS prediction on the RS126 dataset. The notation - indicates that the result cannot be obtained from the papers.

Method	Q_3	Q_H	Q_E	Q_C	Sov
Zvelebil *et al.* (Zpred) [23]	66.7	-	-	-	-
Rost and Sander (PHD) [6]	70.8	72.0	66.0	72.0	-
Salamov *et al.* (NNSSP) [10]	72.7	-	-	-	-
Frishman (PREDATOR) [24]	70.3	-	-	-	-
King and Sternberg (DSC) [25]	71.1	-	-	-	-
Riis and Krogh[7]	71.3	68.9	57.0	79.2	-
Baldi *et al.* (BRNN) [8]	72.0	-	-	-	-
Cuff and Barton (Jpred) [12]	74.8	-	-	-	-
Hua and Sun (SVM) [14]	71.2	73.0	58.0	75.0	-
Single-Stage MSVM	76.2	69.6	63.5	83.1	68.8
Two-Stage MSVMs	78.0	73.1	65.7	83.8	72.6

Table 2. Comparison of performances of single-stage and two-stage MSVM approaches in PSS prediction on the CB396 dataset with PSI-BLAST profiles.

Method	Q_3	Q_H	Q_E	Q_C	Sov
Zvelebil *et al.*(Zpred)[23]	64.8	-	-	-	-
Salamov *et al.* (NNSSP)[10]	71.4	-	-	-	-
Frishman (PREDATOR)[24]	68.6	-	-	-	-
King and Sternberg (DSC)[25]	68.4	-	-	-	-
Guo *et al.*(Dual-Layer SVM)[15]	74.0	79.3	69.3	72.0	-
Single-Stage MSVM	74.5	68.5	62.0	82.4	69.5
Two-Stage MSVMs	76.3	70.6	63.4	83.4	73.2

the highest scores on the CB396 set to date. Compared to the newest method of Guo *et al.* using dual-layer SVM [15], the two-stage MSVM method significantly obtained 2.3% higher Q_3 score. As shown, the prediction accuracy of two-stage MSVMs outperformed the result of single-stage MSVM method for PSS prediction.

In order to avoid to gross overestimates of accuracy, we performed another test on CB396 dataset: we selected best parameters within each cross-validation step by dividing the training data into one for SVM learning and another for selection of window size, σ and γ parameters. The accuracies of the new evaluation approach were not significantly different from those shown

on Table 2 (76.5% of Q_3 and 72.9% of Sov). These results confirmed that the selected window size, sigma and gamma parameters in both learning stages were not biased by the test data chosen.

5 Discussion and Conclusion

We have introduced a two-stage MSVM approach to PSS prediction. With two-stage approaches, the accuracy of prediction is improved because secondary structure at a particular position of a sequence depends not only on the amino acid residue at a particular location but also on the structural formations of the rest of the sequence. Two-stage approach was first introduced in PHD approach which uses two MLPs in cascade for PSS prediction. MLPs are not optimal for this because the cannot generalize the prediction for unseen patterns. The outputs of single stages have been combined in parallel into a single superior predictor to improve upon the individual predictions [12,13]. However, these methods are depended on performances of individual single models and also do not overcome the limitation of single-stage methods. 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 of the output of single-stage by solving the optimization problems at second stage.

Furthermore, we have compared two-stage SVM techniques for PSS problem: one method based on binary classifications of Guo [15] and the other approach for multi-class problem by solving one single optimization problem. We found that the two-stage MSVMs is more suitable for protein secondary structure prediction because its capacity to lead faster convergence for large and complex training sets of PSS problem and solve the optimization problem in one step.

As proved analytically, two-stage MSVMs have the best generalization ability for PSS prediction, by minimizing the generalization error made in the first stage MSVM. However, since this scenario could not be compared with the other techniques as they stick to seven-fold cross-validation for evaluation, which does not test true generalization capabilities. Further, our comparisons with the other techniques were not complete due to the inaccessibility of previously used data and programs. Also, the kernels and SVM parameters were empirically determined as there do not exist any simple methods to find them otherwise. Investigation on two-stage MSVM parameters could further enhance accuracies.

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 *et al*, Journal of Molecular Biology **120**, 97 (1978).
4. J.F. Gibrat *et al*, Journal of Molecular Biology **198**, 425 (1987).
5. J. Garnier *et al*, Methods Enzymol **266**, 541 (1996).
6. B. Rost and C. Sander, Journal of Molecular Biology **232**, 584 (1993).
7. S.K. Riis and A. Krogh, Journal of Computational Biology **3**, 163 (1996).
8. P. Baldi *et al*, Bioinformatics **15** 937 (1999).
9. D.T. Jones, Journal of Molecular Biology **292**, 195 (1999).
10. A.A. Salamov and V.V. Solovyev, Journal of Molecular Biology **247**, 11 (1995).
11. A.A. Salamov and V.V. Solovyev, Journal of Molecular Biology **268**, 31 (1997).
12. J. A. Cuff and G.J. Barton, Proteins **4**, 508 (1999).
13. M. Ouali and R. King, Protein Science **9**, 1162 (1999).
14. S. Hua and Z. Sun, Journal of Molecular Biology **308**, 397 (2001).
15. J. Guo *et al*, Proteins **54**, 738 (2004).
16. K. Crammer and Y. Singer, Computational Learing Theory , 35 (2000).
17. N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines*, (Cambridge University Press, 2000).
18. J.C. Platt *et al*, Proc. Advances in Neural Information Processing Systems 12. Cambridge, MA:MIT Press **12**, 547 (2000).
19. V. Vapnik, *Estimation of Dependences Based on Empirical Data*, (Springer-Verlag, New York, 1982).
20. W. Kabsch and C. Sander, Biopolymers **22**, 2577 (1983).
21. A. Zemla *et al*, Proteins: Structure, Function, and Genetics **34** ,220 (1999).
22. C.W. Hsu and C.J. Lin, IEEE Transactions on Neural Networks **13**, 415 (2002).
23. M.J.J.M Zvelebil *et al*, Journal of Molecular Biology **195**, 957 (1987).
24. D. Frishman and P. Argos, Proteins: Structure, Function, and Genetics **23**, 566 (1995).
25. R.D. King and M.J.E Sternberg*et al*, Protein Science **5** 2298 (1996).