Prediction of Protein Relative Solvent Accessibility with Two-Stage SVM approach

Minh N. Nguyen **Jagath C. Rajapakse**
BioInformatics Research Centre
School of Computer Engineering
Nanyang Technological University, Singapore 639798

ABSTRACT

Information on Relative Solvent Accessibility (RSA) of amino acid residues in proteins provides valuable clues to the prediction of protein structure and function. A two-stage approach with Support Vector Machines (SVMs) is proposed, where an SVM predictor is introduced to the output of the single-stage SVM approach to take into account the contextual relationships among solvent accessibilities for the prediction. By using the position specific scoring matrices, generated by PSI-BLAST, the two-stage SVM approach achieves accuracies upto 90.4% and 90.2% on the Manesh dataset of 215 protein structures and the RS126 dataset of 126 nonhomologous globular proteins, respectively, which are better than the highest published scores on both datasets to date. A web server for protein relative solvent accessibility prediction using two-stage SVM method has been developed and is available at: http://birc.ntu.edu.sg/~ pas0186457/rsa.html

Contact: asjagath@ntu.edu.sg

Keywords: protein structure prediction, solvent accessibility, support vector machines, PSI-BLAST.

INTRODUCTION

The knowledge of protein structures is valuable for understanding mechanisms of diseases of living organisms and for facilitating discovery of new drugs. Protein structure can be experimentally determined by NMR spectroscopy and X-Ray crystallography techniques or by molecular dynamic simulations. However, the experimental approaches are marred by long experimental time, prone to difficulties, and expensive and, therefore, limited to small proteins (Mount, 2001). Bioinformatics approaches are recently being sought to predict Relative Solvent Accessibility (RSA) to help elucidate the relationship between protein sequence and structure, and, thereby, predict the three-dimensional (3-D) structure of proteins (Chandonia & Karplus 1999, Rost & Sander 1994). The studies of solvent accessibility have shown that the hydrophobic free energies of proteins are directly related to the accessible surface area of both polar and non-polar groups of amino acid in proteins (Ooi et al. 1987). Chan and Dill (Chan & Dill 1990) have discovered the burial of core residues is a strong driving force in protein folding. Furthermore, the RSA prediction gives insight into the organization of 3-D structure: the position of protein hydration sites playing an important part in protein's function that can be predicted based on solvent accessibility (Ehrlich et al. 1998); the information of solvent accessibility has improved the prediction of protein subcellular location as the distribution of solvent accessibilities is correlated with its subcellular environments (Andrade et al. 1998). One of the objectives in RSA prediction is to classify a pattern of residues in amino acid sequences to a pattern of RSA types: buried (B) and exposed (E) residues.

Many different techniques have been proposed for RSA prediction, which broadly fall into the following categories: (1) Bayesian, (2) neural networks, and (3) information theoretical approaches. The Bayesian methods provide a framework to take into account local interactions among amino acid residues, by extracting the information from single sequences or multiple sequence alignments to obtain posterior probabilities for RSA prediction (Thompson & Goldstein 1996). Neural networks use residues in a local neighborhood, as inputs, to predict RSA of a residue at a particular location by finding an arbitrary non-linear mapping (Pascarella et al. 1999, Li & Pan 2001, Pollastri et al. 2002, Ahmad & Gromiha 2002). The information theoretical approaches use mutual information between the sequences of amino acids and of solvent accessibility values, derived from a single amino acid residues or pairs of residues in a neighborhood for RSA prediction (Naderi-Manesh et al. 2001). Recently, variants of these approaches with increased prediction accuracies have been proposed: Gianese *et al.* predicted RSA of a residue based on probability profiles computed on amino acid residues in the neighborhood (Gianese et al. 2003); Adamczak *et al.* proposed using neural networks based regression to find continuous approximations to RSA values (Adamczak et al. 2004).

Despite the existence of many approaches, the current success rates of existing techniques to RSA prediction are insufficient; further improvement of the accuracy is necessary. Most existing techniques for RSA prediction are single-stage approaches in the sense that the solvent accessibility type is directly predicted from amino acid sequences or profiles derived thereof. They suffer from the limited size of the local neighborhood used in the prediction; the sequential relationships among the solvent accessibilities of residues are not taken into account. In this paper, we propose a two-stage approach to RSA prediction by using a second predictor, an Support Vector Machine (SVM) classifier, introduced at the end of a single-stage RSA prediction scheme. The aim of the second stage is to take into account the influence on the RSA of a residue by those of its neighbors.

SVMs have been earlier shown to perform well in multiple areas of biological analysis (Cristianini & Shawe-Taylor 2000) including RSA prediction (Yuan et al. 2002, Kim &

Park 2004), which have strong foundations in statistical learning theory; as shown by Vapnik (Vapnik 1995, Vapnik 1998), SVMs implement a classifier that is capable of minimizing structural risk. Furthermore, SVMs offer several associated computational advantages such as the lack of local minima and a solution completely encompassed by the set of support vectors. In addition, SVMs scale well for large scale problems, which are particularly attractive for predicting structures of large protein sequences (Cristianini et al. 2000). Also, the generalization capability of SVMs is well suited for the prediction of RSA of novel amino acid sequences. All previous SVM approaches to RSA prediction are single-stage approaches.

By using two-stage SVM approach, based on the position specific scoring matrices generated by PSI-BLAST, substantial improvements of prediction accuracies upto 7.6% and 4% were achieved on the Manesh (Naderi-Manesh et al. 2001) and the RS126 (Rost et al. 1994) datasets, respectively, compared to previously reported accuracies (Gianese et al. 2003, Kim et al. 2004).

MATERIALS AND METHODS

Dataset 1 (RS126)

The set of 126 nonhomologous globular protein chains, used in the experiment of Rost and Sander (Rost et al. 1994) and referred to as the RS126 set, was used to evaluate the accuracy of the prediction. Many current generation RSA prediction methods have been developed and tested on this dataset which is available at *http://www.rtc.riken.go.jp/~shandar/netasa/rvp-net/rs-126/*. The two-stage SVM approach was 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. With seven-fold cross validation, approximately one-seventh of the dataset was left out while training and, after training, the one-seventh of the dataset was used for testing. In order to avoid the selection of extremely biased partitions, the RS126 set was divided into subsets of same size and composition of each type of RSA.

Dataset 2 (Manesh)

[Table 1 is to be included here.]

The second dataset, generated by Manesh (Naderi-Manesh et al. 2001) consisted of 215 nonhomologous protein chains and is referred to as the Manesh dataset. The dataset contained proteins with less than 25% homology and is available at *http://www.rtc.riken.go.jp/~shandar/netasa/rvp-net/manesh-215/*. The NETASA prediction method (Ahmad et al. 2002) was developed and tested on this dataset. A set of 30 proteins containing 7545 residues was selected for training (see Table 1). The remaining 185 proteins with 43137 residues were used for testing. We adopted these training and testing sets in order to provide an objective comparison of the prediction accuracy of the two-stage SVM approach with the results of the NETASA method (Ahmad et al. 2002) and the probability profile approach of Gianese *et al.* (Gianese et al. 2003). The two-stage SVM predicted the RSA types based on the position specific scoring matrices generated by PSI-BLAST. The PSI-BLAST profiles contained probabilities of residues, taking into account the significance of each sequence and distant homologues (Jones 1999).

RSA and Prediction Accuracy Assessment

RSA percentage (%) of an amino acid residue is defined as the ratio of the solvent accessible surface area of the residue observed in the 3-D structure to that observed in an extended tripeptide (Gly-X-Gly or Ala-X-Ala) conformation. The value of RSA lies between 0% and 100% with 0% corresponding to a fully buried type and 100% to the fully exposed type. The type of the solvent accessibility of an amino acid residue is considered as buried (B) if the RSA value of the residue is smaller than a specified threshold c% or an exposed (E), otherwise. We demonstrate our approach with a range of thresholds of RSA: 0%, 5%, 9%, 10%, 16%, 20%, 25%, and 50%. The residue solvent accessible surface areas of the RS126 set were computed with the DSSP program (Kabsch & Sander 1983). The ASC program (Eisenhaber & Argos 1993) with the van der Waals radii of the atoms (Ooi et al. 1987) was used to compute the residue solvent accessible surface areas for the Manesh dataset. The Ala-X-Ala oligopeptide in an extended conformation instead of Gly-X-Gly is used to calculate RSA in the Manesh dataset. The definition of RSA and programs used to compute it are consistent with those used by other authors whose methods are compared against the proposed approach.

The prediction accuracy is measured by the percentage of correctly predicted types of solvent accessibility of residues (Rost et al. 1994); the sensitivity score indicates the proportion of exposed (E) residues that are correctly predicted as E; the specificity measures the proportion of buried (B) residues that are correctly predicted as B. By changing the thresholds of RSA definition of the prediction, we get a range of sensitivities and specificities, which leads to Receiver Operation Characteristics (ROC) that plots sensitivity versus one minus specificity. The ROC curves offer for comparisons among different prediction methods irrespective of the threshold for determination of solvent accessibility type.

Single-Stage SVM Approach

This section describes how a sequence of RSA types is predicted from an amino acid sequence by using an SVM classifier. Let us denote the 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 the corresponding solvent accessibility sequence by $\mathbf{a} = (a_1, a_2, \ldots, a_n)$ where $a_i \in \Omega_A$; n is the length of the sequence. The prediction of the sequence of RSA types, $\mathbf{a}$, from an amino acid sequence, $\mathbf{r}$, is the problem of finding the optimal mapping from the space of Ω_R^n to the space of Ω_A^n .

Firstly, the values of raw matrices of PSI-BLAST (Altschul et al. 1997), used as inputs to first stage SVM, are obtained from NR (non-redundant) database that is available at *ftp://ftp.ncbi.nih.gov/blast/db/FASTA/*. The low-complexity regions, transmembrane regions, and coil-coil segments are then filtered from the NR database by PFILT program (Jones 1999). Finally, the E-value threshold of 0.001, three iterations, BLOSUM62 matrix, a gap open penalty of 11, a gap extended penalty of 1 are used for searching the non-redundant sequence database to generate position specific scoring matrix (PSSM) profiles. These arguments are consistent with those used in other methods (Jones 1999, Kim et al. 2004). Let v_i be a vector representing a 21-dimensional coding of the residue r_i where 20 elements take the values from PSSM profiles ranging from [0, 1] (Kim et al. 2004) and the last element is used as the padding space to indicate the end of the sequence; the padding element is set to 1 to indicate the end of the sequence or 0, otherwise. The SVM, a binary classifier *B/E*, is constructed to predict whether the solvent accessibility of a residue at a site belongs to a particular type, *B* or *E*. The input pattern to the predictor at site i consists of a vector $\mathbf{r}_i$ of profiles from a neighborhood: $\mathbf{r}_i = (v_{i-h_1}, v_{i-h_1+1}, \ldots, v_i, \ldots, v_{i+h_1})$ where h_1 represents the size of the neighborhood on either side.

The SVM transforms the input vectors to a higher dimension via a kernel function, K^1, and linearly combines to derive the outputs with a weight vector, $\mathbf{w}_1$. The function K^1 and vector $\mathbf{w}_1$ are determined to minimize the error in the prediction during the training phase. Let $\Gamma^1_{\text{train}} = \{(\mathbf{r}_j, q_j) : j = 1, 2, \ldots, N\}$ denote the set of all training exemplars where q_j denotes the desired classification, *B* or *E*, for the input pattern $\mathbf{r}_j$ such that the output of SVM is -1 if the correct RSA type is *B* or +1 if the type is *E*. When N is the number of training patterns patterns, the vector, $\mathbf{w}_1$, is determined by scalars α_j, $j = 1, 2, \ldots, N$, that are found by maximizing the following quadratic function Q_1:

$$Q_1 = \sum_{j=1}^{N} \alpha_j - \frac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} \alpha_j \alpha_i q_j q_i K^1(\mathbf{r}_j, \mathbf{r}_i) \qquad (1)$$

subject to $0 \le \alpha_j \le \gamma^1$ and $\sum_{j=1}^{N} \alpha_j q_j = 0$. $K^1(\mathbf{r}_j, \mathbf{r}_i) = \phi^1(\mathbf{r}_i)\phi^1(\mathbf{r}_j)$ denotes the kernel function and ϕ^1 represents the mapping function to higher dimension; γ^1 is a positive constant used to decide the trade-off between the training error and the margin of the classifier (Vapnik 1995, Vapnik 1998).

The weight vector is then given by $\mathbf{w}_1 = \sum_{j=1}^{N} q_j \alpha_j \phi^1(\mathbf{r}_j)$. Once the parameters α_j are obtained from the above algorithm, the resulting discriminant function, say f_1, is given by

$$\begin{aligned} f_1(\mathbf{r}_i) &= \sum_{j=1}^{N} q_j \alpha_j K^1(\mathbf{r}_j, \mathbf{r}_i) + b_1 \\ &= \mathbf{w}_1 \phi^1(\mathbf{r}_i) + b_1 \end{aligned} \qquad (2)$$

where the bias b_1 is chosen so that $q_j f_1(\mathbf{r}_j) = 1$ for any j with $0 < \alpha_j < \gamma^1$.

In the single-stage SVM method, the solvent accessibility type a_i corresponding to the residue at site i, r_i, is determined by

$$a_i = \begin{cases} E & \text{if } f_1(\mathbf{r}_i) \ge 0 \\ B & \text{otherwise} \end{cases} \qquad (3)$$

The function, f_1, discriminates the type of RSA, based on the features or interactions among the residues in the input pattern. With optimal parameters, the SVM attempts to minimise the generalization error in the prediction. If the training and testing patterns are drawn independently and identically according to a probability P_1, then the generalization error, err_{P_1}, is given by

$$\mathrm{err}_{P_1}(f_1) = P_1\{(\mathbf{r}, q) : \operatorname{sign}(f_1(\mathbf{r})) \ne q; (\mathbf{r}, q) \in \Gamma^1\}$$

where Γ^1 denotes the set of input patterns seen by the SVM during both the training and testing phases. In the following sections, we demonstrate that this error can be minimised by connecting another predictor at the output of the SVM predictor.

Two-Stage SVM Approach

[Figure 1 is to be included here.]

The single-stage approach takes only the interactions among amino acid residues in the neighborhood into the prediction scheme. The RSA type of a residue is also influenced by those in its neighborhood. A second SVM predictor is used in the two-stage approach to predict the RSA type of a residue by using the predictions from the first-stage, capturing the sequential relationships among the RSA values in the neighborhood. The architecture of the two-stage SVM prediction approach is illustrated in figure 1.

The second SVM classifier improves the accuracy of the single-stage RSA prediction schemes by taking into account the sequential relationships among the RSA values of residues into the prediction. The second-stage SVM processes the estimated RSA values at the first stage and minimizes the generalization error by incorporating the contextual information among RSA values. Rost and Sander (Rost et al. 1994) proposed a simple method to incorporate the sequential relationships of the estimated RSA types, in which an averaging filter is employed to take the average of neighboring outputs of the first neural network at each amino acid residue; then, the solvent accessibility is predicted as the type with the largest average. Two-stage SVM approaches were previously proposed for protein secondary structure prediction (Nguyen & Rajapakse 2003, Guo et al. 2004).

The second stage SVM processes the output of the discriminant functions of the first stage to enhance the prediction. At the site i, the input to the second SVM is given by a vector $\mathbf{d}_i = (d_{i-h_2}, d_{i-h_2+1}, \ldots, d_i, \ldots, d_{i+h_2})$ where h_2 is the length of the neighborhood on either side and $d_i = 1/(1+e^{-f_1(\mathbf{r}_i)})$. The SVM converts the input patterns, usually linearly inseparable, to a higher dimensional space by using the mapping ϕ^2 with a kernel function $K^2(\mathbf{d}_i, \mathbf{d}_j) = \phi^2(\mathbf{d}_i)\phi^2(\mathbf{d}_j)$.

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 SVM be $\Gamma^2_{\text{train}} = \{(\mathbf{d}_j, q_j) : j = 1, 2, \ldots, N\}$. The kernel function K^2 and vector $\mathbf{w}_2$ are obtained by solving the following convex quadratic programming problem, over all the patterns seen in the training phase:

$$\max_{\beta} \sum_{j=1}^{N} \beta_j - \frac{1}{2}\mathbf{w}_2^T \mathbf{w}_2 \tag{4}$$

such that $0 \le \beta_j \le \gamma^2$ and $\sum_{j=1}^{N} \beta_j q_j = 0$

where $\mathbf{w}_2 = \sum_{j=1}^{N} q_j \beta_j \phi^2(\mathbf{d}_j)$.

The discriminant function, f_2, at the second stage is given by

$$\begin{aligned} f_2(\mathbf{d}_i) &= \sum_{j=1}^{N} q_j \beta_j K^2(\mathbf{d}_j, \mathbf{d}_i) + b_2 \\ &= \mathbf{w}_2 \phi^2(\mathbf{d}_i) + b_2 \end{aligned} \tag{5}$$

where the bias b_2 is chosen so that $q_j f_2(\mathbf{d}_j) = 1$ for any j with $0 < \beta_j < \gamma^2$. The solvent accessibility type a_i corresponding to the residue r_i is given by

$$a_i = \begin{cases} E & \text{if } f_2(\mathbf{d}_i) \ge 0 \\ B & \text{otherwise} \end{cases} \tag{6}$$

If the set of input patterns for the second-stage SVM in both training and testing phases is denoted by Γ^2, the generalization error of the two-stage SVM approach, $\text{err}_{P_2}(f_2)$, is given by

$$\text{err}_{P_2}(f_2) = P_2\{\{(\mathbf{d}, q) : \ \text{sign}(f_2(\mathbf{d})) \ne q; (\mathbf{d}, q) \in \Gamma^2\}$$

If the input pattern $\mathbf{d}$ corresponds to a site i, then $\mathbf{d} = \mathbf{d}_i = ((1+e^{-f_1(\mathbf{r}_{i-h_2})})^{-1}, \ldots, (1+e^{-f_1(\mathbf{r}_i)})^{-1}, \ldots, (1+e^{-f_1(\mathbf{r}_{i+h_2})})^{-1})$ That is, the second stage takes into account the influences of the RSA values of residues in the neighborhood into the prediction. It could be easily conjectured that, if the RSA type of a residue depends on those of its neighborhood, $\text{err}_{P_2}(f_2) \le \text{err}_{P_1}(f_1)$ where the equality occurs when $h_2 = 0$.

RESULTS

[Table 2 is to be included here.]

[Table 3 is to be included here.]

For SVM classifiers, a window size of 13 amino acid residues $h_1 = 6$ gave optimal results in the [9, 21] range for the first stage and a window size of width 21, $h_2 = 10$, in the [11, 27] range gave the optimal accuracy for the second stage. The kernels selected were Gaussian functions, $K(\mathbf{x}, \mathbf{y}) = e^{-\sigma\|\mathbf{x}-\mathbf{y}\|^2}$ with the parameters: $\sigma = 0.1, \gamma^1 = 1.0$ at the first stage, and $\sigma = 0.15, \gamma^2 = 1.0$ at the second stage, which were determined empirically for optimal performances in [0.01, 0.5] and [0.1, 2] ranges, respectively. In the literature, the Gaussian kernel has been used in many classification problems (Cristianini et al. 2000). The main reason is that it can result in complex (but smooth) decision function, and therefore has the ability to better fit the data where simple discrimination by using a hyperplane or a low-dimensional polynomial surface is not possible. The use of Gaussian kernel showed the best performance when the dimension of feature space is infinite (Schollkopf et al. 1997) and gave better results over the linear and polynomial kernels for RSA prediction (Kim et al. 2004). The Gaussian kernels have shown faster convergence than linear kernels for large and complex training sets of RSA problem. The SVM method was implemented using sequential minimization algorithm (Platt 1999) which is simple to implement without needing storage for matrices or to invoke an iterative numerical routine for each sub-problem.

Table 2 shows the performances of different solvent accessibility predictors and two-stage SVM approach on the RS126 set. Two-stage SVMs with PSI-BLAST profiles achieved accuracies of 90.2%, 83.5%, 81.3%, and 79.4% at thresholds of 0%, 5%, 9%, and 16%, respectively, which are the highest scores on the RS126 set to date. Compared to the newest method of Kim and Park, using single-stage SVM (Kim et al. 2004), the two-stage SVM method significantly obtained 4%, 3.7%, and 1.6% higher prediction accuracies at 0%, 5%, and 16% thresholds, respectively. On the RS126 dataset, the accuracies were improved by 4.5% and 4.3% at thresholds of 9% and 16% compared to the results of the probability profiles approach of Gianese *et al.* (Gianese et al. 2003). The prediction accuracy of two-stage SVMs outperformed the results by the multi-layer perceptron networks of PHDacc method proposed by Rost and Sander (Rost et al. 1994) at all thresholds.

Table 3 shows the performance of two-stage SVM approach on the Manesh dataset based on PSI-BLAST profiles and comparison with other solvent accessibility predictors. The best performance was shown by the cascade of two SVMs. On the Manesh dataset, the accuracies were significantly improved by 2.5%, 8.3%, 9.8%, 7.8% and 3.2% for 0%, 5%, 10%, 25%, and 50% thresholds, respectively, compared to the results of NETASA method (Ahmad et al. 2002). Comparing two-stage SVMs to the probability profiles method (Gianese et al. 2003), substantial gains of 0.9% to 7.6% of prediction accuracy were observed for different thresholds.

[Figure 2 is to be included here.]

[Figure 3 is to be included here.]

[Figure 4 is to be included here.]

[Figure 5 is to be included here.]

Figures 2 and 3 present the distributions of prediction scores obtained by two-stage SVMs for the benchmark Manesh and RS126 datasets with a 5% threshold based on PSI-BLAST profiles. The ROC curves on the Manesh and RS126 datasets for single-stage and two-stage SVM approaches at different thresholds are illustrated in figures 4 and 5. As shown, the prediction accuracy of two-stage SVMs outperformed the single-stage SVM methods for RSA prediction at all thresholds.

[Table 4 is to be included here.]

For RSA prediction, the accuracy of two-stage SVMs using PSI-BLAST profiles is significantly higher than results obtained by using multiple sequence alignments. For example, the accuracy of two-stage SVM method on RS126 dataset was only 78.6% at a threshold of 5% based on multiple sequence alignments. As mentioned (Jones 1999), PSI-BLAST profiles contain more information of homologous protein structures than multiple sequence alignments. Additionally, improvements of accuracies are observed when larger sequences or more homologous profiles are used in training. As shown in table 4, by using a set of 205 proteins instead of 30 proteins for training, the prediction accuracies of 10 sequences, obtained from the tails of the histogram in Figure 2 (1lts, 1nba, 1afw, 3cox, 2wsy, 7rsa, 1amm, 1mai, 1knb, 1kte), were improved at a threshold of 5%. These observations suggest that the performance of two-stage SVM method based on PSI-BLAST profiles for a novel amino acid sequence suffers if it lacks in the homologous structures in the training set. For a completely new protein whose homologous proteins are not used in training, two-stage SVM method predicts its solvent accessibilities with a low accuracy. For our knowledge, Rost and Adamczak (Rost et al. 1994, Adamczak et al. 2004) concluded that the overall performance of any method based on evolutionary profiles suffers when very remote or no homologues are included.

[Table 5 is to be included here.]

Table 5 lists the properties of 20 amino acids and their average occurrence and probabilities for exposure and error in RSA prediction on the Manesh dataset at a 25% threshold. Nelson and Cox (Nelson & Cox 2000) based on the polarity or tendency to interact with water of R group at biological pH to group 20 amino acids into five main classes. According to the statistical data, amino acids, Ala, Val, Leu, Ile, Phe, Cys were easy to predict while Gly, Pro, Trp, Thr, Arg, His were difficult to predict by two-stage SVMs. As shown, two-stage SVM method frequently predicted A, V, L, I, M, F, W, Y, C to be buried, and G, P, S, T, N, Q, K, R, H, D, E to be exposed. The statistical data confirms that the non-polar R groups (hydrophobic) tend to be buried, i.e., in the interior of a protein, and the polar R groups (hydrophilic) tend to be on the surface (exposed), except for G, P, and C (Chen et al. 2004). This is because two Cys are readily oxidized to form a disulfide bond and disulfide-linked residues are hydrophobic. Chen (Chen et al. 2004) also explained the reasons that Pro and Gly tend to be exposed from their structures. The results from table 5 suggest that the amino acid residues that tend to be buried (A, V, L, I, M, F, W, Y, C) are predicted with higher accuracies than exposed ones (G, P, S, T, N, Q, K, R, H, D, E)

As shown in tables 2 and 3, predictions were best for buried residues, e.g., 90.2% and 90.4% of the completely buried sites were correctly predicted at a threshold of 0% on RS126 and Manesh datasets, respectively. The two-stage SVM method achieved the highest prediction accuracy for the extreme case of fully buried type because the accessibility of completely buried residues is best conserved in 3-D homologous structures (Rost et al. 1994). Residues in α-helix and β-strand structure segments were predicted better than ones in coil segments, e.g., 80.7%, 82.2%, and 77.5% residues were correctly predicted in. α-helix, β-strand, and coil segments, respectively, on the Manesh dataset at a 25% threshold.

[Table 6 is to be included here.]

We also estimated the effect of the growing size of NR databases used to generate position scoring matrices by PSI-BLAST on the accuracy of two-stage SVM method. Two NR databases were used: one as of December 22, 2003 with 1,581,064 sequences and a newer version as of April 7, 2004 with 2,745,128 sequences. The different results of two-stage SVMs on two NR databases were not significant (see Table 6).

A web server for protein relative solvent accessibility prediction using two-stage SVM method has been developed and is available at: *http://birc.ntu.edu.sg/~pas0186457/rsa.html*. A set of 30 proteins containing 7545 residues (see Table 1) was selected for training two-stage SVM method presented on the web server.

DISCUSSION AND CONCLUSION

The existing bioinformatics techniques for RSA prediction are mostly single-stage approaches which predict the RSA types of residues, based on only the information available in amino acid sequences. We demonstrated a two-stage approach, by using SVMs, that utilises the output predicted by single-stage prediction schemes and improve the accuracy of RSA prediction. In this way, the influences on the RSA value of a residue by those of its neighbors are accounted for. This is because the solvent accessibility 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 buried or exposed type consists of at least two consecutive residues. Therefore, another layer of SVM classifier incorporating the contextual relationship among the solvent accessibility characteristics makes the prediction more realistic in terms of predicted mean lengths of solvent accessibility elements. The analysis of prediction results from single-stage and two-stage SVM methods showed that the second stage SVM ultimately cleans the output prediction of the first stage SVM, mostly by removing isolated buried or exposed residue.

SVMs are more suitable for prediction of RSA values because they minimize the generalization error in the prediction. We showed that the generalization error made in the first stage is further minimized by the second stage of the two-stage approach. SVM is an optimal classifier for the second stage in terms of the margin of separation; it attempts to minimize not only the empirical risk of known sequences but also the actual risk for unknown sequences. Two stages of SVMs are sufficient to find an optimal classifier for RSA prediction as the second stage SVM attempts to minimize the generalization error of the first stage by solving the optimization problem at the second stage.

Recently, Kim and Park (Kim et al. 2004) suggested to use the information of the position specific scoring matrices generated by PSI-BLAST as inputs to SVMs for RSA prediction. By combining PSI-BLAST profiles, the present approach achieved better results than the methods using information from single sequences and multiple sequence alignments. Compared to the method of Kim and Park, our method showed a considerable improvement in the accuracy of prediction. By incorporating the state-of-the-art methods based on PSI-BLAST profiles and SVM in a two-stage approach, we are able to report the best accuracies to date for RSA prediction on the tested datasets. The RSA elements of residues predicted by our approach could facilitate the prediction of the structure and function of amino acid sequences.

References

R. Adamczak, A. Porollo, and J. Meller, Accurate prediction of solvent accessibility using neural networks based regression, *Proteins: Structure, Function, and Bioinformatics*, vol 56, pp 753-767, 2004.

S. Ahmad and M. M. Gromiha, NETASA: neural network based prediction of solvent accessibility, *Bioinformatics*, vol 18:6, pp 819-824, 2002.

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

M. A. Andrade, S. I. O'Donoghue, B. Rost, Adaptation of protein surfaces to subcellular location, *Journal of Molecular Biology*, vol 276, pp 517-525, 1998.

H. S. Chan and K. A. Dill, Origins of structure in globular proteins, *Proc. Natl Acad. Sci. USA*, vol 87, pp 6388-6392, 1990.

J. Chandonia and M. Karplus, New methods for accurate prediction of protein secondary structure, *Protein Engineering*, vol 35, pp 293-306, 1999.

H. Chen, H. X. Zho, X. Hu, and I. Yoo, Classification comparison of prediction of solvent accessibility from protein sequences, *the 2nd Asia-Pacific Bioinformatics Conference*, 2004.

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

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.

F. Eisenhaber and P. Argos, Improved strategy in analytical surface calculation for molecular systems-handling of singularities and computational efficiency, *J. Comp. Chem*, vol 14, pp 1272-1280, 1993.

L. Ehrlich, M. Reczko, H. Bohr, and R. C. Wade, Prediction of water-binding sites on proteins using neural networks, *Protein Engineering*, vol 11, pp 11-19, 1998.

G. Gianese, F. Bossa, and S. Pascarella, Improvement in prediction of solvent accessibility by probability profiles, *Protein Engineering*, vol 16:12, pp 987-992, 2003.

M.H.M. Giorgi, S. Hazout, and P. Tuffery, PredAcc: prediction of solvent accessibility, *Bioinformatics*, vol 15, pp 176-177, 1999.

J. Guo, H. 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.

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

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.

H. Kim and H. Park, Prediction of protein relative solvent accessibility with support vector machines, *Proteins: Structure, Function, and Bioinformatics*, vol 54:3, pp 557-562, 2004.

X. Li and X. M. Pan, New method for accurate prediction of solvent accessibility from protein sequence, *Proteins: Structure, Function, and Genetics*, vol 42, pp 1-5, 2001.

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

H. Naderi-Manesh, M. Sadeghi, S. Araf, and A.A.M. Movahedi, Predicting of protein surface accessibility with information theory, *Proteins: Structure, Function, and Genetics*, vol 42, pp 452-459, 2001.

D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, Worth Publishsers, New York, 2000.

M. N. Nguyen and J. C. Rajapakse, Two-stage support vector machines for protein secondary structure prediction, *Neural, Parallel and Scientific Computations*, vol 11, pp 1-18, 2003.

G. Pollastri, P. Baldi, P. Fariselli, and R. Casadio, Prediction of coordination number and relative solvent accessibility in proteins, *Proteins: Structure, Function, and Genetics*, vol 47, pp 142-153, 2002.

T. Ooi, M. Oobatake, G. Nemethy, and H. A. Scheraga, Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides, *Proc. Natl Acad. Sci. USA*, vol 84, pp 3086-3090, 1987.

J. C. Platt, Using sparseness and analytic QP to speed training of support vector machines, *Proc. Advances in Neural Information Processing Systems 11*. Cambridge, MA:MIT Press, 1999.

S. Pascarella, R. .D. Persio, F. Bossa, and P. Argos, Easy method to predict solvent accessibility from multiple protein sequence alignments, *Proteins: Structure, Function, and Genetics*, vol 32, pp 190-199, 1999.

B. Rost and C. Sander, Conservation and prediction of solvent accessibility in protein families, *Proteins: Structure, Function, and Genetics*, vol 20, pp 216-226, 1994.

B. Scholköpf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik, Comparing support vector machines with gaussian kernels to radial basis function classifiers, *IEEE Trans. Sign. Processing*, vol 45, pp 2758-2765, 1997.

M. J. Thompson and R. A. Goldstein, Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes, *Proteins: Structure, Function, and Genetics*, vol 47, pp 142-153, 1996.

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

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

Z. Yuan, K. Burrage, and J. Mattick, Prediction of protein solvent accessibility using support vector machines, *Proteins: Structure, Function, and Genetics*, vol 48, pp 566-570, 2002.

1aba	1abr	1bdo	1beo	1bib	1bmf	1bnc	1btm	1btn	1cem
1ceo	1cew	1cfy	1chd	1chk	1cyx	1dea	1del	1dkz	1dos
1fua	1gai	1gpl	1gsa	1gtm	1hav	2i1b	2sns	3grs	3mdd

Table 1: The list of 30 proteins used for training the single-stage and two-stage SVM approaches.

Method / Threshold	0%	5%	9%	16%
Rost and Sander, 1994 (PHDacc)	86.0	-	74.6	75.0
Gianese *et al.*, 2003 (PP)	-	-	76.8	75.1
Kim and Park, 2004 (Single-stage SVM)	86.2	79.8	-	77.8
Two-stage SVMs	90.2	83.5	81.3	79.4

Table 2: Comparison of performances of two-stage SVM approach with other methods in RSA prediction on the RS126 dataset with position specific scoring matrices generated by PSI-BLAST. The notation **-** indicates that the corresponding result was not available from the literature.

Method / Threshold	0%	5%	10%	20%	25%	50%
Ahmad and Gromiha, 2002 (NETASA)	87.9	74.6	71.2	-	70.3	75.9
Gianese *et al.*, 2003 (PP)	89.5	75.7	73.4	-	71.6	76.2
Two-stage SVMs	90.4	82.9	81.0	78.6	78.1	79.1
Giorgi *et al.*, 1999 (PredAcc)	85.0	-	-	-	70.7	-
Cuff and Barton, 2000 (Jnet)	86.6	79.0	-	-	75.0	-
Li and Pan, 2001	-	-	-	71.5	-	-
Pollastri *et al.*, 2002 (BRNN)	86.5	81.2	-	-	77.2	-
Adamczak *et al.*, 2004 (SABLE)	-	76.8	77.5	77.9	77.6	-

Table 3: Comparison of performances of two-stage SVM approach in RSA prediction based on position specific scoring matrices generated by PSI-BLAST, with other methods on the Manesh dataset.

Training set	1lts	1nba	1afw	3cox	2wsy	7rsa	1amm	1mai	1knb	1kte
30 proteins	70.9	71.4	71.8	73.6	75.2	91.1	92.0	93.3	93.3	93.3
205 proteins	72.6	71.8	71.8	73.9	76.9	91.0	93.1	94.1	93.3	93.3

Table 4: Comparison of performances of two-stage SVM approach on 10 proteins (1lts, 1nba, 1afw, 3cox, 2wsy, 7rsa, 1amm, 1mai, 1knb, 1kte) based on PSI-BLAST profiles with two different training sets of 30 and 205 proteins at a 5% threshold.

Amino acid		Occurrence (%)	Exposure (%)	Error in RSA prediction (%)
Non-polar R group (hydrophobic)				
Gly	G	7.5	55.8	27.1
Ala	A	7.7	39.7	19.0
Val	V	6.8	16.7	16.2
Leu	L	8.8	14.1	15.7
Ile	I	5.7	12.1	14.8
Met	M	2.2	20.8	21.7
Pro	P	4.5	64.8	27.5
Aromatic R group (hydrophobic)				
Phe	F	4.1	10.5	16.5
Trp	W	1.4	12.3	25.1
Tyr	Y	3.8	18.8	24.8
Polar, uncharged R group (hydrophilic)				
Ser	S	5.9	63.7	24.7
Thr	T	5.6	53.2	25.6
Cys	C	1.6	1[illegible].5	15.1
Asn	N	4.6	74.4	25.0
Gln	Q	3.9	79.4	24.6
Positively R charged (hydrophilic)				
Lys	K	6.1	84.5	19.8
Arg	R	4.8	72.5	28.3
His	H	2.2	51.2	30.5
Negatively R charged (hydrophilic)				
Asp	D	6.3	80.9	23.2
Glu	E	6.4	84.7	21.4

Table 5: The properties of 20 amino acids: their average occurrences, probabilities of exposures, and the error in RSA prediction on the Manesh dataset at a 25% threshold.

Database / Threshold	0%	5%	10%	20%	25%	50%
1,581,064 NR	90.2	82.8	80.9	78.6	78.1	79.0
2,745,128 NR	90.4	82.9	81.0	78.6	78.1	79.1

Table 6: Comparison of performances of two-stage SVM approach on the Manesh dataset based on position specific scoring matrices generated by PSI-BLAST with two different NR databases.

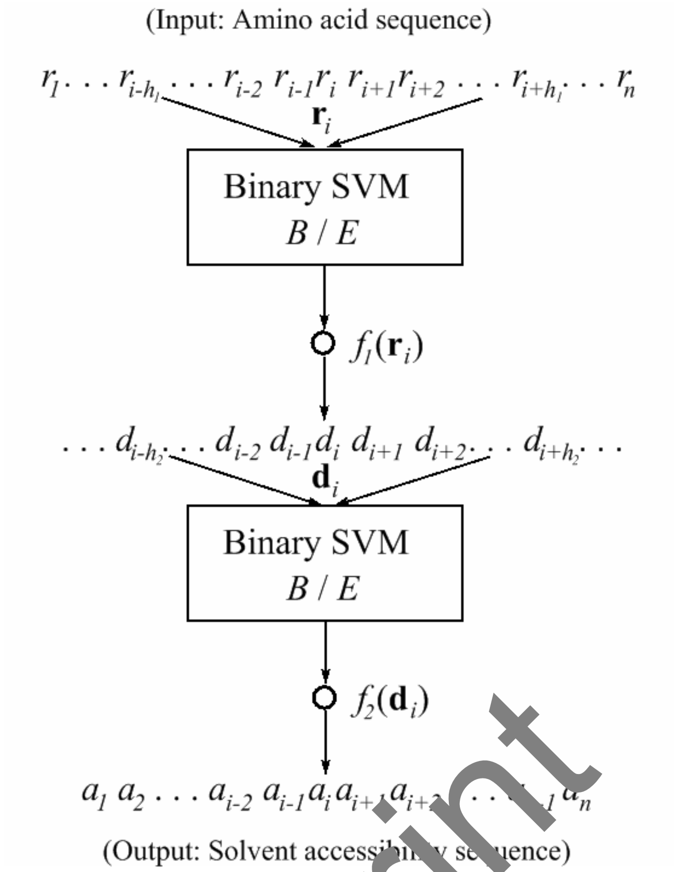
(Input: Amino acid sequence)
$r_1 \ldots r_{i-h_1} \ldots r_{i-2}\ r_{i-1} r_i\ r_{i+1} r_{i+2} \ldots r_{i+h_1} \ldots r_n$
$\mathbf{r}_i$
Binary SVM
B / E
$f_1(\mathbf{r}_i)$
$\ldots d_{i-h_2} \ldots d_{i-2}\ d_{i-1} d_i\ d_{i+1}\ d_{i+2} \ldots d_{i+h_2} \ldots$
$\mathbf{d}_i$
Binary SVM
B / E
$f_2(\mathbf{d}_i)$
$a_1\ a_2 \ldots a_{i-2}\ a_{i-1} a_i a_{i+1} a_{i+2} \ldots a_{n-1} a_n$
(Output: Solvent accessibility sequence)

Figure 1: Two-stage SVM approach for RSA prediction.

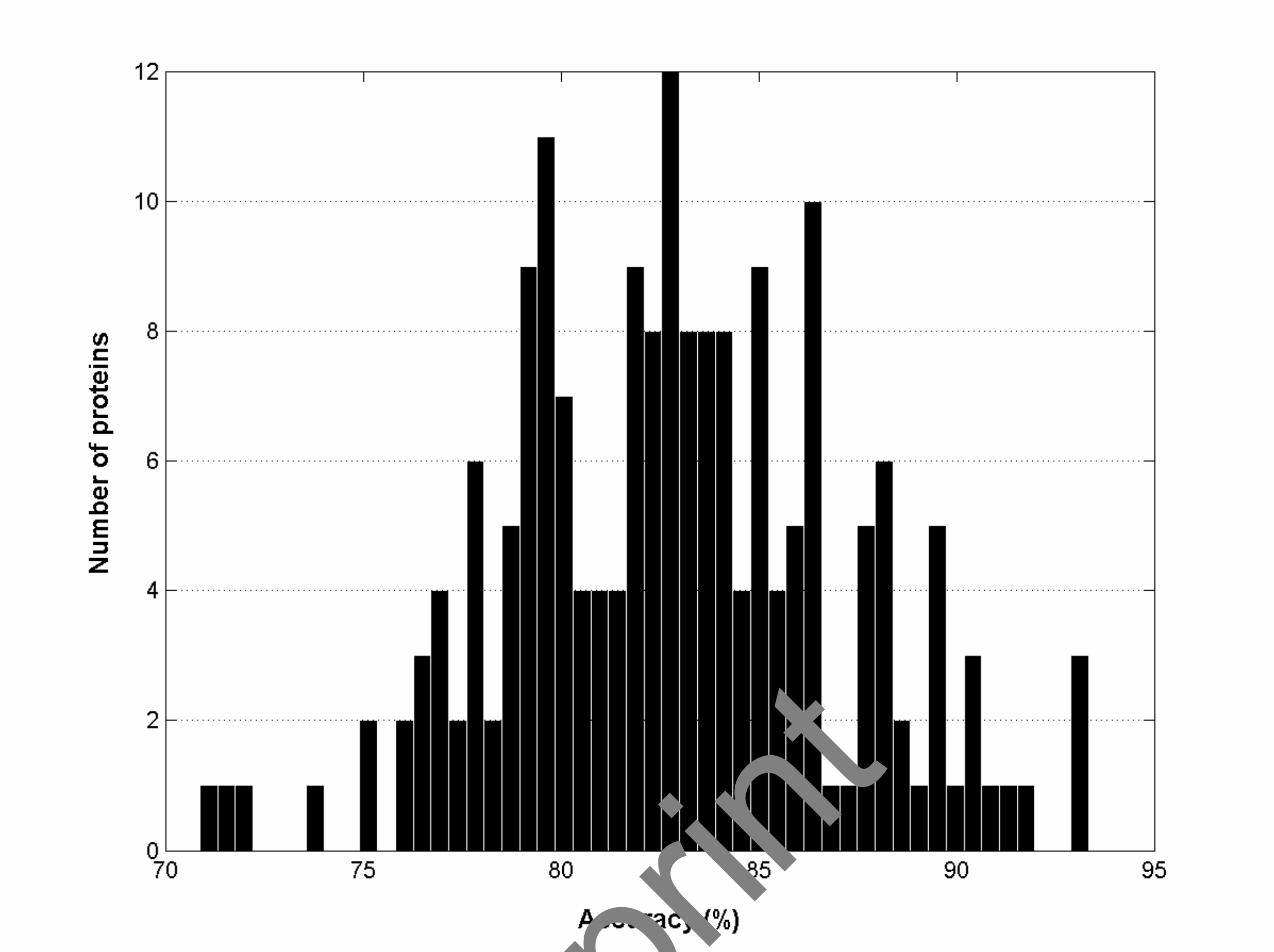

Figure 2: The distribution of prediction scores obtained by two-stage SVMs for the benchmark 185 proteins of the Manesh dataset at a 5% threshold based on PSI-BLAST profiles.

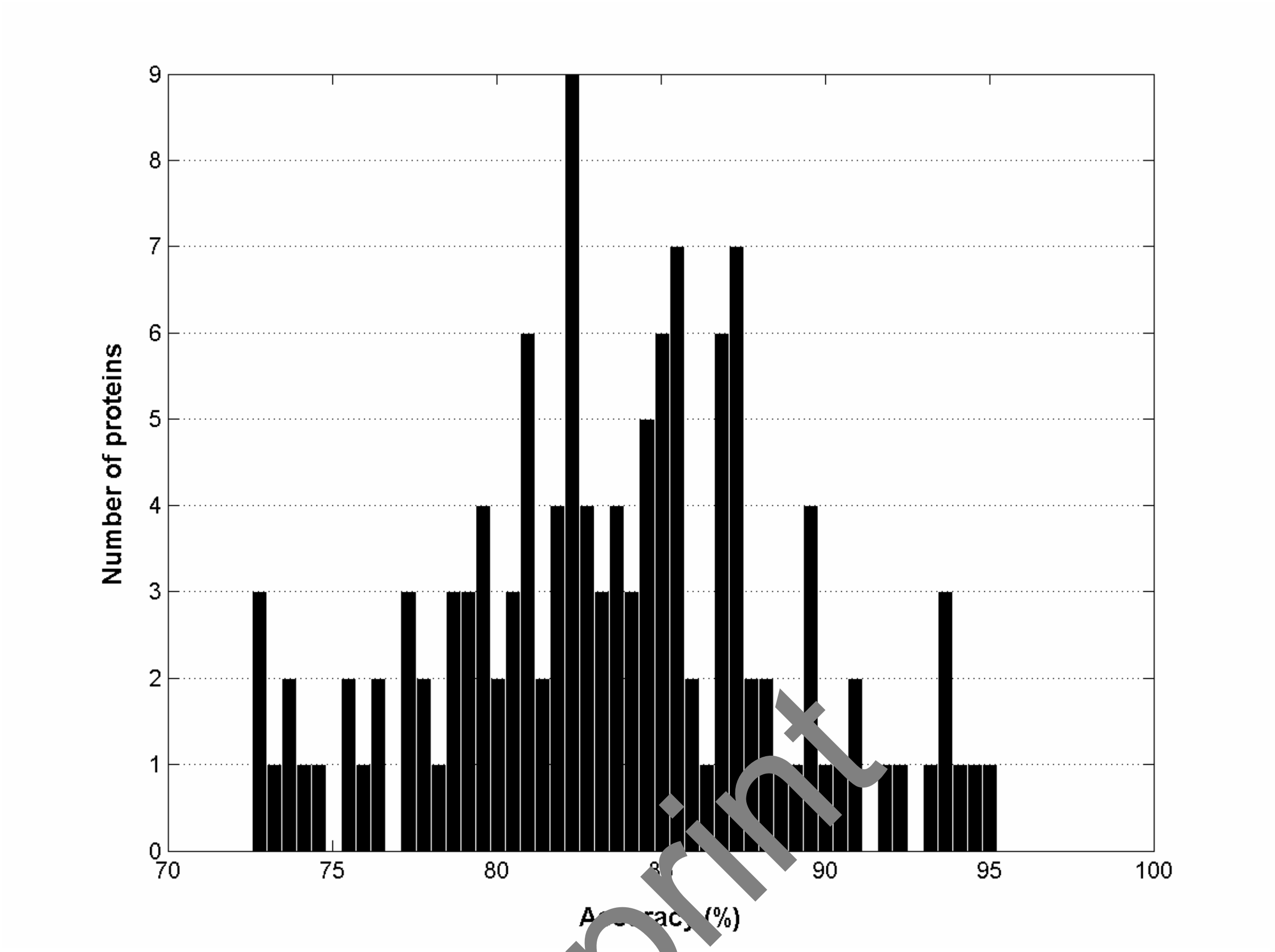

Figure 3: The distribution of prediction scores obtained by two-stage SVMs for the benchmark RS126 dataset at a 5% threshold based on PSI-BLAST profiles.

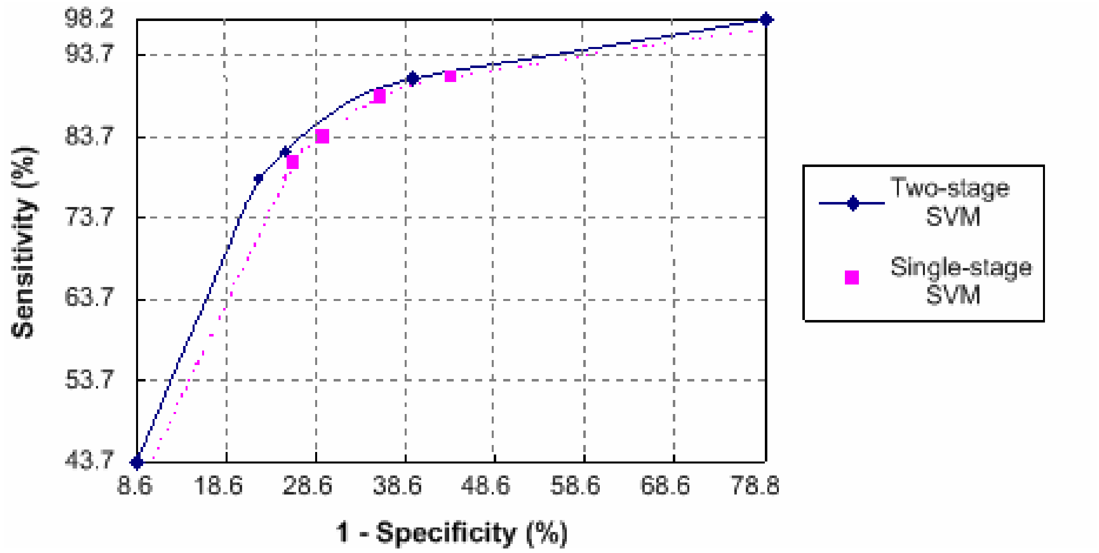

Figure 4: The ROC curves on the Manesh dataset for single-stage and two-stage SVM approaches for RSA prediction.

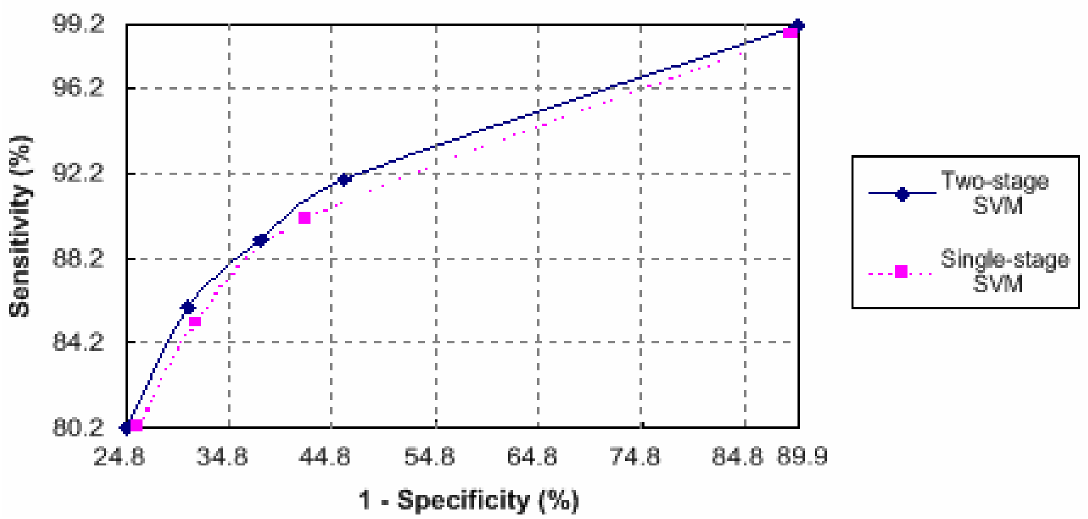

Figure 5: The ROC curves on the RS126 dataset for single-stage and two-stage SVM approaches for RSA prediction.