Two-Stage SVR Approach for Predicting Accessible Surface Areas of Amino Acids

Minh N. Nguyen **Jagath C. Rajapakse**

BioInformatics Research Centre
School of Computer Engineering
Nanyang Technological University, Singapore 639798

ABSTRACT

We address the problem of predicting solvent accessible surface area (ASA) of amino acid residues in protein sequences, without classifying them into buried and exposed types. A two-stage support vector regression (SVR) approach is proposed to predict real values of ASA from the position-specific scoring matrices (PSSMs) generated from PSI-BLAST profiles. By adding SVR as the second stage to capture the influences on the ASA value of a residue by those of its neighbors, the two-stage SVR approach achieves improvements of mean absolute errors up to 3.3%, and correlation coefficients of 0.66, 0.68, and 0.67 on the Manesh dataset of 215 proteins, the Barton dataset of 502 nonhomologous proteins, and the Carugo dataset of 338 proteins, respectively, which are better than the scores published earlier on these datasets. A web server for protein ASA prediction by using a two-stage SVR method has been developed and is available (http://birc.ntu.edu.sg/~ pas0186457/asa.html).

Contact: asjagath@ntu.edu.sg

Keywords: protein structure prediction; accessible surface area; solvent accessibility; support vector regression; PSI-BLAST

INTRODUCTION

Protein-protein interactions play a central role in numerous processes in biological cells and are one of the major areas of research in proteomics.[1] Understanding the mechanisms of protein-protein interactions is vital when addressing issues associated with the biological function and disease. In addition, protein three-dimensional (3D) structure prediction directly from amino acid sequences still remains as an open and important problem in life sciences;[2] The bioinformatics approaches first focus on predicting the secondary structure and/or the solvent accessibilities of a protein's structure which represents the one-dimensional projections of the complicated 3D structure.[2-4] The successful prediction of solvent accessibility is helpful in elucidating the relationship between protein structure and interactions.[5] The information of solvent accessibility in proteins leads to numerous insights into the organization of 3D structure.[6, 7] The studies of solvent accessibility have shown that the burial of core residues is a strong driving force in protein folding;[8] the prediction of exposed residues is valuable to the understanding of the function of a protein as the active sites of a protein are always located on its surface.[9] Ahmad et al.[10] demonstrated the importance of the role in solvent accessibility of amino acids in determining the probability of protein-DNA binding.

Many different techniques have been proposed for predicting relative solvent accessibility (RSA) of the residues in a given amino acid sequence. The RSA percentage (%) of an amino acid residue is defined as the ratio of the solvent accessible surface area (ASA) of the residue observed in the 3D structure to that observed in an extended tripeptide (Gly-X-Gly or Ala-X-Ala) conformation.[11] The approaches using Bayesian statistics,[12] th logistic functions,[13] and information theory[14] predict SA of a residue based only on single sequence in . Neural networks use residues in a local neighborhood, as inputs, to predict RSA of a residue at a particular site by extracting the information from a single sequence[15-17] multiple sequence alignments.[4, 18-20] The input information derived from multiple sequence alignments was used by support vector machines (SVM)[9] or as probability profiles.[21] Recently, the use of the position specific scoring matrices (PSSMs) generated from PSI-BLAST profiles has enhanced the prediction accuracies of methods using SVM,[22] regression methods using neural networks,[23] linear regression models,[24] and two-stage SVMs.[25]

All the above techniques, however, classify amino acid residues only into buried and exposed types based on different RSA thresholds. Thus, the applications and information provided by such RSA predictors are limited.[10] Furthermore, it is difficult to compare the importance of the accuracies of different methods in any subsequent applications. The results of prediction of ASA have significant impact in determining interacting residues in proteins and the prediction of protein-protein interactions.[5] Yuan and Huang have shown that it is more meaningful to know the real values of ASA than to know the residues as buried and exposed types since as ASA information directly reflects the degree to which the residues are in contact with the solvent molecules.[26] Moreover, some amino acid residues have significantly lower mean ASA values and, therefore, the classification of solvent accessibility at the same RSA threshold for all residues may not be justified.[10] Previously, we introduced a two-stage SVM for the prediction of RSA into two classes, buried or exposed, which gave substantial improvements of prediction accuracies.[25] However, our previous method is insufficient to predict the ASA, as a percentage of the surface area, of a residue in a given amino acid sequence since the classifiers used were crisp.

In this paper, we use support vector regressors (SVR) in a two-stage scheme combined with the evolutionary

information generated by PSI-BLAST profiles for ASA prediction. The SVR is an optimization technique, which creates regression functions of arbitrary type from a set of training data, based on SVMs that has strong foundations in statistical learning theory; as shown by Vapnik and Smola.[27-29] SVM and SVR are powerful and generally applicable tools in protein structure prediction[30] including solvent accessibility prediction.[9, 22, 25, 26] This is because many biological problems involve high-dimensional and noisy data; SVM and SVR, with their generalization capabilities, are known to behave well compared to other statistical or machine learning methods in handling such data. Recently, two approaches have been proposed to predict real values of ASA from amino acid sequences.[10, 26] Ahmad et al. proposed a neural network method to predict ASA values by finding an arbitrary complex mapping from a window of surrounding residues.[10] SVR has been applied to ASA prediction by using the information from a single sequence.[26] Nevertheless, they are single-stage approaches and do not account fully for the ASA values of the neighboring residues. Also, they use the conventional orthogonal encoding derived directly from the amino acid sequences as inputs for ASA prediction.

Our approach utilizes an SVR to predict the ASA values from the output predicted by the first stage SVR of ASA of residues. In this way, the influences on the ASA value of a residue by those of its neighbors are accounted for. The present approach improves the mean absolute errors by 3.1%, 2.8%, and 3.3% on the Manesh,[14] the Barton,[19] and the Carugo[31] datasets, respectively, compared to the previously reported best mean absolute errors using neural network[10] and single stage SVR methods.[26] The correlation coefficients between the predicted and observed ASAs are 0.66, 0.68, and 0.67 on the Manesh, Barton, and Carugo datasets, respectively, which are significantly better than those obtained by the methods of Ahmad et al.[10] and Yuan et al.[26]

MATERIALS AND METHODS

Dataset 1 (Manesh)

The set of 215 nonhomologous protein chains with no more than 25% pairwise-sequence identity and 50682 residues, used in the experiment of Manesh[14] and referred to as the Manesh dataset, was used to evaluate the accuracy of the prediction. The neural network method of Ahmad et al.[10] was developed and tested on this dataset. Our approach was implemented with the position specific scoring matrices (PSSMs) generated by PSI-BLAST as inputs. To objectively compare with the neural network method of Ahmad et al., the same test procedure was performed, using six-round tests (or three-fold cross-validation).[10] For three-fold cross-validation, the Manesh dataset was divided into three subsets of the same size. All six possible combinations of three subsets were then used for training, testing, and validation processes. The validation set was kept out of the training process to avoid the selection of extremely biased partitions of training and testing sets. The final results are then averaged to determine the accuracies of the method.

Dataset 2 (Barton)

The second dataset was generated by Cuff and Barton,[19] consisting of 502 nonhomologous protein chains with more than 83,000 residues, and is referred to as the Barton dataset. The dataset contained protein sequences with less than 25% homology. We adopted three-fold cross-validation with the same training and testing subsets used in the methods of Ahmad et al. and Yuan et al. in order to objectively compare the prediction accuracy of the two-stage SVR approach with the results of those earlier methods.[10, 26]

Dataset 3 (Carugo)

The third dataset was generated by Carugo,[31] consisting of 338 nonhomologous monomeric protein crystal structure extracted from Protein Data Bank, and is referred to as the Carugo dataset. This dataset contained protein sequences with no more than 25% pairwise-sequence identity. The three-fold cross-validation with the same training and testing subsets from the previous method[10] was used to provide an objective comparison of the prediction accuracy.

The Manesh, the Barton, and the Carugo datasets are available *http://gibk21.bse.kyutech.ac.jp/rvp-net/all-data.tar.gz.* The present method predicted the real values of ASA based on the PSSMs generated from PSI-BLAST profiles while the methods of Ahmad et al. and Yuan et al. only used the information of single sequences. The PSI-BLAST profiles contained more useful information than single sequences: 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.[32]

ASA and Prediction Accuracy Assessment

The absolute values of ASA in the Manesh dataset were obtained by using the Analytical Surface Calculation (ASC) program[33] with the van der Waals radii of the atoms given by Ooi et al.[34] The absolute ASA values for the Barton and Carugo datasets were computed with the Dictionary of Protein Secondary Structure (DSSP) program.[35] The programs used to compute the absolute ASA values of amino acid datasets are in consistent with those used by other authors whose methods are compared against the present approach. The normalized ASA values calculated by dividing the ASA value with the corresponding value for the extended Ala-X-Ala conformation of the different amino acid types is used as the measure for the ASA values. The absolute values of the ASA are transformed back by multiplying with the same normalization constants.

To measure the prediction accuracy of the proposed method, the mean absolute error and Pearson's correlation coefficient between the predicted and experimentally observed ASA values are calculated. The mean absolute error of the prediction is the absolute difference between the predicted and observed values of relative ASA values per residue in the sequence.[10] The Pearson's correlation coefficient is defined as the ratio of the covariance between the predicted and observed ASA values per residue to the product of the standard deviations.[10]

Two-Stage SVR Approach

[Figure 1 is to be included here.]

This section describes our approach which utilizes the two SVRs in cascade for predicting ASA values of amino acid residues in a protein sequence. 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 ASA $a_i \in \mathrm{R}$ takes a real value; n is the length of the sequence. The prediction of the sequence of ASA values, $\mathbf{a}$, 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 R^n. The architecture of the two-stage SVR prediction approach is illustrated in Figure 1.

First, the values of raw matrices of PSI-BLAST[36] to use as inputs to the first stage SVR, are obtained from non-redundant (NR) database at *ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/*, the version as of April 7, 2004, containing 2,745,128 sequences. The low-complexity regions, transmembrane regions, and coil-coil segments are then filtered from the NR database by using PFILT program.[32] Finally, the E-value threshold of 0.001, 3 iterations, BLOSUM62 matrix, a gap open penalty of 11, and a gap extended penalty of 1 were used for searching the non-redundant sequences to generate position specific scoring matrix (PSSM) profiles.

Let v_i be a vector representing a 21-dimensional codi… of the residue r_i where 20 elements take the valu… PSSM profiles for each type of the residue, rang… [0, 1],[22] and the last element is used as the p… spac… to indicate the end of the sequence; th… elemen… is set to 1 to indicate the end of the s…quenc… or 0, …therwise. The input pattern to the predi… at …te i, therefore, consists of a vector $\mathbf{r}_i$ of the profiles …m a neighborhood: $\mathbf{r}_i = (v_{i-h_1}, v_{i-h_1+1}, \ldots, v_i, \ldots, v_{i+h_1})$ where … represents the size of the neighborhood on either side of the window.

Let $\{(\mathbf{r}_j, q_j) : j = 1, 2, \ldots, N\}$ denote the set of all training exemplars where q_j denotes the desired real value of ASA of residue r_i and N is the number of training patterns. The first stage for ASA prediction consists of a SVR predictor that maps the input patterns to real values of ASA. The input vectors are transformed to a hidden-space via a kernel function, K^1 and then linearly combined to derive the outputs by using a weight vector $\mathbf{w}_1$ and a bias b_1.[28, 29] The SVR uses a more general type of loss function than that of SVM, the so-called Vapnik's ε-insensitive loss,[28, 29] to construct an analogue of the soft margin in the space of the target values $q \in \mathrm{R}$.

The vector $\mathbf{w}_1$ and b_1 are then determined to minimize the error in the prediction during the training phase, that are found by maximizing the following quadratic function to evaluate scalars α_j, α_j^*, $j = 1, 2, \ldots, N$:

$$\begin{aligned} \max_{\alpha,\alpha^*} & -\varepsilon\sum_{j=1}^{N}(\alpha_j + \alpha_j^*) + \sum_{j=1}^{N}(\alpha_j - \alpha_j^*)q_j \\ & -\frac{1}{2}\sum_{j=1}^{N}\sum_{i=1}^{N}(\alpha_j - \alpha_j^*)(\alpha_i - \alpha_i^*)q_j q_i K^1(\mathbf{r}_j, \mathbf{r}_i) \end{aligned} \quad (1)$$

$$\text{subject to } 0 \le \alpha_j, \alpha_j^* \le \gamma^1 \text{ and } \sum_{j=1}^{N}(\alpha_j - \alpha_j^*) = 0$$

where ε is the tolerance of the error; only the deviations of Vapnik's ε-insensitive loss function larger than ε are considered as errors. The kernel function $K^1(\mathbf{r}_j, \mathbf{r}_i) = \phi^1(\mathbf{r}_i)\phi^1(\mathbf{r}_j)$ denotes the kernel and ϕ^1 represents the mapping function to the higher dimension; γ^1 is a positive constant used to decide the trade-off between the training errors and model complexity.[28, 29]

The weight vector is then given by $\mathbf{w}_1 = \sum_{j=1}^{N}(\alpha_j - \alpha_j^*)\phi^1(\mathbf{r}_j)$. In the first stage, once the parameters α_j and α_j^* are obtained from the above algorithm, the r…lting ASA value, say $a_i' \in \mathrm{R}$, is given by

$$\begin{aligned} a_i' &= \sum_{j=1}^{N}(\alpha_j - \alpha_j^*)K^1(\mathbf{r}_j, \mathbf{r}_i) + b_1 \\ &= \mathbf{w}_1\phi^1(\mathbf{r}_i) + b_1 \end{aligned} \quad (2)$$

…e sin…le-stage SVR approach takes only the features or the …interactions among amino acid residues in the n…ghborhood into the prediction scheme, which is unable t… sufficiently take into account the contextual information about solvent accessibilities. The ASA value of a residue is also influenced by those values of the residues in its neighborhood. A second SVR predictor is used in the two-stage approach to enhance the ASA values prediction by using the predictions from the first-stage as inputs to take into account the contextual information among ASA values in the neighborhood. Recently, the two-stage methods have yielded substantial improvements of the accuracies compared to the single-stage methods for secondary structure[37, 38] and RSA[25] predictions of the proteins. Rost and Sander[4] first proposed a simple method to incorporate the sequential relationships of the estimated solvent accessibilities, in which an averaging filter was employed to take the average of neighboring outputs of the first neural network at each amino acid residue and then, the solvent accessibility is predicted as the type with the largest average. Cuff and Barton[19] proposed an 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 solvent accessibilities from the output of the first stage. Zhou and Shan[39] used two neural networks in cascade for prediction of protein-protein interaction sites from sequence profiles of neighboring residues and solvent exposures.

The second stage SVR processes the output of the first stage SVR to enhance the prediction of ASA values. At the site i, the input to the second SVR is given by a vector

$\mathbf{a}_i^{'} = (a_{i-h_2}^{'}, a_{i-h_2+1}^{'} \ldots, a_i^{'}, \ldots, a_{i+h_2}^{'})$ where h_2 is the length of the neighborhood on either side. The SVR converts the input patterns, usually linearly inseparable, to a higher dimensional space by using the mapping ϕ^2 with a kernel function $K^{\,2}(\mathbf{a}_i^{'}, \mathbf{a}_j^{'}) = \phi^2(\mathbf{a}_i^{'})\phi^2(\mathbf{a}_j^{'})$. As in the first stage, the hidden outputs in the higher dimensional space are linearly combined with a weight vector $\mathbf{w}_2$ and a bias b_2 to obtain the final prediction.

Let the training set of exemplars of the second stage SVR be $\{(\mathbf{a}_j^{'}, q_j) : j = 1, 2, \ldots, N\}$. The weight vector $\mathbf{w}_2$ and the value b_2 are obtained by solving the following convex quadratic programming problem over all the patterns seen in the training phase:

$$\max_{\beta, \beta^*} -\varepsilon \sum_{j=1}^{N} (\beta_j + \beta_j^*) + \sum_{j=1}^{N} (\beta_j - \beta_j^*) q_j - \frac{1}{2} \sum_{j=1}^{N} \sum_{i=1}^{N} (\beta_j - \beta_j^*)(\beta_i - \beta_i^*) q_j q_i K^{\,2}(\mathbf{a}_j^{'}, \mathbf{a}_i^{'}) \quad (3)$$

such that $0 \le \beta_j, \beta_j^* \le \gamma^2$ and $\sum_{j=1}^{N} (\beta_j - \beta_j^*) = 0$.

After obtaining $\beta_j, \beta_j^*, j = 1, 2, \ldots, N$, the weights are given by $\mathbf{w}_2 = \sum_{j=1}^{N} (\beta_j - \beta_j^*) \phi^2(\mathbf{a}_j^{'})$.

At the output of the second stage, the resulting ASA value a_i corresponding to the residue r_i is given by

$$a_i = \sum_{j=1}^{N} (\beta_j - \beta_j^*) K^{\,2}(\mathbf{a}_j^{'}, \mathbf{a}_i^{'}) + b_2 \quad (4)$$
$$= \mathbf{w}_2 \phi^2(\mathbf{a}_i^{'}) + b_2$$

RESULTS

The SVR method was implemented using LIBSVM library,[40] which usually leads to faster convergence in large optimization problems. For two-stage SVR method, a window size of 13 amino acid residues, $h_1 = 6$, was selected for the first stage, and a window size of width 21, $h_2 = 10$, was used for the second stage. These parameters were heuristically derived and are consistent with the optimal values used in the two-stage SVM method.[25] The parameter ε of Vapnik's ε-insensitive loss function was set as 0.001.

The Gaussian kernel $K(\mathbf{x}, \mathbf{y}) = e^{-\sigma \|\mathbf{x}-\mathbf{y}\|^2}$ showed superior performance over the linear and polynomial kernels for solvent accessibility prediction,[9, 22, 25] the reasons being that (1) the Gaussian kernel can result in complex (but smooth) decision functions and therefore has the ability to better fit the data where a simple discrimination by using a hyperplane or a low-dimensional polynomial surface is not possible and (2) the prediction is more dependent on the residues in a local neighborhood rather than those at distant locations. The parameters of the Gaussian kernel and SVR: $\sigma = 0.01, \gamma^1 = 2.0$ at the first stage, and $\sigma = 0.1, \gamma^2 = 1.0$ at the second stage were determined empirically for optimal performances in [0.01, 0.1] and [0.5, 2] ranges, respectively.

[Table 1 is to be included here.]

Table 1 shows a comparison of performances of the present approach and neural network method of Ahmad et al.[10] and single-stage SVR of Yuan et al.[26] for ASA value prediction on the Manesh, Barton, and Carugo datasets. On the Manesh dataset, two-stage SVR with PSI-BLAST profiles achieved mean absolute error of 14.9%. Compared to the neural network method of Ahmad et al., using single sequence input, the two-stage SVR method significantly improved on the mean absolute error by 3.1%. On the Manesh dataset, the Pearson's correlation coefficient of 0.68 was achieved by two-stage SVR, which is substantially higher than the result of the neural network approach of Ahmad et al.[10] On the Barton dataset of 502 proteins, the mean absolute errors were improved by 3.1% and 2.8% by the present method compared to the results of the neural network of Ahmad et al.[10] and the single-stage SVR method of Yuan et al.[26] using single sequence inputs, respectively. On the Barton dataset, the Pearson's correlation coefficient of 0.66 was observed by the two-stage SVR, which is better than those achieved by the methods of Ahmad et al. and Yuan et al. [10, 26] On the Carugo set of 338 proteins, the two-stage SVR approach significantly improved on the mean absolute error by 3.3% and the Pearson's correlation coefficient by 0.19 compared to the method of Ahmad et al.[10]

[Table 2 is to be included here.]

Since the regression method using neural networks of Adamczak et al. has been introduced mainly for RSA prediction,[23] we transform the RSA values predicted by the method of Adamczak et al. to ASA values by multiplying with the corresponding value for the extended Gly-X-Gly conformation to compare with the results of our method. Table 2 shows the comparison of our approach with the method of Adamczak et al. on 199 proteins of Manesh dataset, containing no more than 25% pairwise-sequence identity, which are different from 860 proteins used for training of the method of Adamczak et al. We did not attempt to remove structures from 199 sequences that might be homologous to 860 proteins in the training set. On this testing set, the two-stage SVR approach performed better than the regression method using neural networks of Adamczak et al. although the number of training proteins used in three-fold cross-validation of the present work is much smaller than 860 proteins used by Adamczak et al.

[Table 3 is to be included here.]

Further, the real ASA values predicted by the two-stage SVR approach were converted into solvent states (buried and exposed) to compare previous methods proposed for RSA prediction. The results of the experiments performed on the Manesh dataset with the training set of 30 proteins and the testing set of 185 proteins[25] are shown in Table 3. We adopted these training and testing sets in order to provide a fair comparison of the prediction accuracy of the two-stage SVR approach with the results of NETASA method,[17] the probability profile (PP) approach of Gianese

et al.,[21] and two-stage SVM method[25] for RSA prediction. The information theoretical method of Manesh et al.[14] used a full jack-knife validation to estimate the prediction accuracy on the Manesh dataset. The prediction accuracies were improved up to 2.9%, 7.3%, and 5.7% at different thresholds by the present approach compared to the results of the methods of Manesh et al., Ahmad et al., and Gianese et al., respectively. Also, the present method achieved prediction accuracies that are comparable with the results of two-stage SVM method for RSA prediction.[25] The mean absolute error of 15.1% was observed by two-stage SVR on this dataset, indicating the ability of the method to generalize well with a small training set.

[Table 4 is to be included here.]

Table 4 shows the mean absolute errors of residues in buried and exposed parts on the Barton dataset of 502 proteins. The mean absolute errors of exposed residues are 17.4%, 17.8%, 18.8%, and 19.4% at thresholds of 5%, 10%, 20% and 25%, respectively. As shown in Table 3, two-stage SVR achieved better prediction accuracy than previous methods for RSA prediction at a threshold of 50%. These results show that the two-stage SVR approach also performs well on exposed residues compared to previous methods (see Tables 1, 3 and 4).

For real value of ASA prediction, the accuracy of two-stage SVR method using PSI-BLAST profiles is significantly higher than the results obtained by using the information from single sequences.[26] As mentioned,[32] PSI-BLAST profiles contain more information of homologous protein structures than single sequences. Rost and Adamczak[4, 23] have suggested that the overall performance of any method based on evolutionary profiles suffers when very remote or no homologues sequences are included. Therefore, the performance of two-stage SVR method based on PSI-BLAST profiles for a novel amino acid sequence suffers if it lacks in the homologous structures in the training set.

[Figure 2 is to be included here.]

[Figure 3 is to be included here.]

Figure 2 presents the distribution of mean absolute errors resulted in the two-stage SVR for the benchmark Barton dataset of 502 nonhomologous proteins, based on PSI-BLAST profiles. The distribution of mean absolute errors for individual proteins is related to their lengths. As illustrated in Figure 3, 93.7% of long protein sequences (>150 amino acids where 150 is the median) and 72.0% of short protein sequences ($\leq$ 150 amino acids) were predicted with mean absolute errors less than 18%. These observations concur with the findings of Yuan et al. that the ASA values of the small proteins are more difficult to predict.[26]

[Table 5 is to be included here.]

The 5% of with lowest mean absolute errors and the 5% with largest mean absolute errors were selected from the tails of the histogram in Figure 2 for further analysis to investigate why ASA values of residues in such sequences are difficult to predict by two-stage SVR method. As seen in Table 5, the reasons for the largest mean absolute errors of the predictions are that they are (1) short sequences with the mean length of 50.4, compared to the others and (2) had lower hydrophilic residues with the mean hydrophobicity value of –0.5. These results suggest that if a novel protein has a short length and a large negative mean hydrophobicity value, i.e., with the most of its amino acids being hydrophilic, the ASA values of the residues are difficult to predict. The Pearson's correlation coefficient between mean hydrophobicity scale and mean absolute error on the Barton dataset of 502 proteins was computed to be –0.22 (see Table 5). The negative value of the correlation coefficient indicates that mean absolute error decreased with the increase in mean hydrophobicity scale.

[Figure 4 is to be included here.]

[Figure 5 is to be included here.]

Figure 4 shows the observed ASA values and absolute errors of predicted values of the protein 1TND:B which had the lowest mean absolute error of 10.1% by using two-stage SVR. Further, a single-stage SVR resulted in the mean absolute error of 11.0% for the prediction of ASA of this protein. As seen, most residues in the highly or completely buried regions are well predicted. Figure 5 presents the observed ASA and absolute errors of predicted values of the protein 2MEV:4 which had the largest mean absolute error, 40.7%, of the prediction with the two-stage SVR method. The single stage SVR produced a mean absolute error of 41.5% for the protein 2MEV:4. The poorly predicted regions are from position 6 (G) to 12 (P) and from position 38 (Q) to 58 (A), had most residues in highly exposed regions.

[Table 6 is to be included here.]

Table 6 lists the properties of 20 amino acids, their average occurrence, hydrophobicity scales, and mean absolute error in ASA prediction on the Barton dataset. Nelson and Cox,[41] based on the polarity or tendency to interact with water grouped 20 amino acids into five main class; hydrophobicity scales[42] combining hydrophobicity and hydrophilicity of R groups are used to measure the tendency of an amino acid to seek an aqueous environment (negative value) or a hydrophobic environment (positive value). According to the statistical data, the ASA values of residues in amino acids, Val, Leu, Ile, Phe, and Cys are easy to predict while Gly, Pro, Ser, Asn, Asp, and Glu are difficult to predict by the two-stage SVR method. The results from Table 6 suggest that the nonpolar (hydrophobic) residues tending to be in the interior of a protein (buried), Ala, Val, Leu, Ile, Met, and Pro are predicted with lower errors than the polar and uncharged (hydrophilic) residues tending to be on the surface (exposed): Ser, Thr, Asn, and Gln except for Gly, Pro, and Cys. This is because two Cys residues are readily oxidized to form a disulfide bond and disulfide-linked residues are strongly hydrophobic.[43] Gly tends to be exposed as it contributes little in general to the stability of folded proteins.[41] Pro commonly appears at exposed sites in proteins, such as loops, turns, and N-terminal first turn of helix.[41] This result is in agreement with the earlier analyses of protein stability, that indicate the structural information to be very important for the prediction of the stability of exposed mutations, while residue information is sufficient for buried mutations.[44]

As seen in Table 6, Cys and Ile are predicted with the least mean absolute errors, 9.9% and 9.7%, respectively, which may be due to the fact that Cys and Ile residues are usually present in the interior of a protein. The results from Table 6 also confirm that the most difficult predictions are Gly and Asn with mean absolute errors, 19.6% and 20.2%, respectively, which may be due to their conformation flexibility and variability.[10] Furthermore, all of the aromatic residues, Phe, Trp, and Tyr, are predicted with low mean absolute errors because they are usually located in buried or partially buried regions and form stable conformations.[10] All of the charged residues, Lys, Arg, His, Asp, and Glu, are predicted with high mean absolute errors because most these residues present on the surface of a protein. Further, as shown in Table 6, the prediction accuracy of the two-stage SVR outperformed the single-stage SVR for ASA prediction. Comparing the two-stage SVR to the single-stage SVR method, the improvements of mean absolute errors were observed for all 20 amino acid residues.

DISCUSSION AND CONCLUSION

Earlier, we proposed a two-stage SVM approach to predict the residues present in an amino acid sequence as buried or exposed type. However, such an approach gives a limited information about the ASA values available from the structural data. Furthermore, the existing bioinformatics techniques for ASA prediction are single-stage approaches that they predict the ASA values of residues, based on only the information derived from single sequences. In this study, we demonstrated a two-stage approach, by using SVRs, that utilizes the output predicted by a single-stage prediction scheme to further improve the accuracy of ASA prediction. The aim of the second stage SVR is to take into account the influence on the ASA value of a residue by those of its neighbors. 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.[25] Therefore, another layer of SVR classifier incorporating the contextual relationship among the solvent accessibility characteristics enhances the prediction of ASA values, predicted by the first stage.

SVR is more suitable for the prediction of ASA values because it minimizes the generalization error in the prediction.[28, 29] In addition, the SVR method offers several associated computational advantages such as the lack of local minima and a solution completely encompassed by a set of support vectors. Two stages of SVRs are sufficient to find an optimal classifier for the prediction of ASA values as the second stage SVR attempts to minimize the generalization error produced by the first stage.[45]

By combining the evolutionary information generated from PSI-BLAST profiles as inputs, the present approach achieved better results than the methods using information from single sequences. The present method reported the best accuracies to date for the ASA prediction on the tested datasets. Our experiments on different datasets confirmed that the improvements by using the two-stage SVR approach are consistent and do not depend on the test data chosen. The ASA values of residues predicted by our approach could facilitate the prediction of the secondary structure,[46] the protein-protein interactions,[5] and the function of amino acid sequences, which applications are worthwhile for further investigating. Furthermore, due to the significance of the ASA prediction, even a slight enhancement is vital and may lead to better techniques for further improvement of solvent accessibility prediction.

A web server for protein ASA prediction using two-stage SVR method has been developed and is available at: *http://birc.ntu.edu.sg/~pas0186457/asa.html*. A set of 30 proteins containing 7545 residues from the Manesh dataset[25] was selected for training two-stage SVR method presented on the web server.

REFERENCES

1. Archakov AI, Govorun VM, Dubanov AV, Ivanov YD, Veselovsky AV, Lewi P, Janssen P. Protein-protein interactions as a target for drugs in proteomics. Proteomics. 2003;3:380 – 391.

2 Mount DW. Bioinformatics: sequence and genome analysis. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2001.

3 Chandonia J, Karplus M. New methods for accurate prediction of protein secondary structure. Protein Eng 1999;12:293 – 306.

4 Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216 – 226.

5. Raih MF, Ahmad S, Zheng R, Mohamed R. Solvent accessibility in native and isolated domain environments: general features and implications to interface predictability. Biophys Chem. 2005;114:63 – 69.

6. Ehrlich L, Reczko M, Bohr H, Wade RC. Prediction of waterbinding sites on proteins using neural networks. Protein Eng 1998;11:11 – 19.

7. Andrade MA, O' Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. J Mol Biol 1998;276:517 – 525.

8. Chan HS, Dill KA. Origins of structure in globular proteins. Proc Natl Acad Sci USA 1990;87:6388 – 6392.

9. Yuan Z, Burrage K, Mattick J. Prediction of protein solvent accessibility using support vector machines. Proteins 2002;48:566 – 570.

10. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629 – 635.

11. Samanta U, Bahadur RP, Chakrabarti P. Quantifying the accessible surface area of protein residues in their local environment. Protein Eng 2002;15:659 – 667.

12. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics

and optimized residue substitution classes. Proteins 1996;47:142 – 153.

13. Giorgi MHM, Hazout S, Tuffery P. PredAcc: prediction of solvent accessibility. Bioinformatics 1999;15:176 – 177.

14. Naderi-Manesh H, Sadeghi M, Araf S, Movahedi AAM. Predicting of protein surface accessibility with information theory. Proteins 2001;42:452 – 459.

15. Holbrook SR, Muskal SM, Kim SH. Predicting surface exposure of amino-acids from protein-sequence. Protein Eng 1990;3:659 – 665.

16. Li X, Pan XM. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1 – 5.

17. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819 – 824.

18. Pascarella S, Persio RD, Bossa F, Argos P. Easy method to predict solvent accessibility from multiple protein sequence alignments. Proteins 1999;32:190 – 199.

19. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502 – 511.

20. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142 – 153.

21. Gianese G, Bossa F, Pascarella S. Improvement in prediction of solvent accessibility by probability profiles. Protein Eng 2003;16: 987 – 992.

22. Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins 2004;54:557 – 562.

23. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks based regression. Proteins 2004;56:753 – 767.

24. Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. Journal of Computational Biology 2005;12:355-369.

25. Nguyen MN, Rajapakse JC. Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins 2005;59:30-37.

26. Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. Proteins 2004; 57:558 – 564.

27. Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag; 1995.

28. Vapnik V. Statistical learning theory. New York: Wiley; 1998.

29. Smola A, Schölkopf B. A tutorial on support vector regression. NeuroCOLT Technical Report, NC-TR-1998-030, http://www.neurcolt.com; 1998.

30. Cristianini N, Shawe-Taylor J. An introduction to support vector machines. Cambridge, UK: Cambridge University Press; 2000.

31. Carugo O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. Protein Eng 2000;13:607 - 609.

32. Jones DT. Protein secondary structure prediction based on position specific scoring matrices. J Mol Biol 1999;292:195 – 202.

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

34. Ooi T, Oobatake M, Nemethy G, Scheraga HA. Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci USA 1987;84:3086 – 3090.

35. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577 – 263.

36. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389 – 3402.

37. Nguyen MN, Rajapakse JC. Two-stage support vector machines for protein secondary structure prediction. Neural Parallel Sci Comput 2003;11:1 – 18.

38. Guo J, Chen H, Sun Z, Lin Y. A novel method for protein secondary structure prediction using dual-layer SVM and profiles. Proteins 2004;54:738 – 743.

39. Zhou HX, Shan Y. Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 2001; 44:336 – 343.

40. Hsu CW, Lin CJ. A comparison on methods for multi-class support vector machines. IEEE Transactions on Neural Networks 2002; 13:415 – 425.

41. Nelson DL, Cox MM. Lehninger principles of biochemistry. New York: Worth; 2000.

42. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982; 157:105 – 132.

43.Chen H, Zho HX, Hu X, Yoo I. Classification comparison of prediction of solvent accessibility from protein sequences. Paper presented at the 2nd Asia-Pacific Bioinformatics Conference, Dunedin, New Zealand: 2004.

44. Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A. Role of structural and sequence information for predicting protein stability changes: comparison

between buried and partially buried mutations. Protein Eng 1999;12:549 – 555.

45. Nguyen MN, Rajapakse JC. Two-stage multi-class SVMs for protein secondary structure prediction. Pacific Symposium on Biocomputing, Hawaii, USA: 2005.

46. Adamczak R, Porollo A, Meller J. Combining prediction of secondary structure and solvent accessibility in proteins. Proteins 2005;59:467 – 475 .

Method / Datasets	Manesh		Barton		Carugo	
	Mean Absolute Error (%)	Correlation Coefficient	Mean Absolute Error (%)	Correlation Coefficient	Mean Absolute Error (%)	Correlation Coefficient
Ahmad et al.[10] (NN)	18.0	0.50	18.8	0.48	19.0	0.48
Yuan et al.[26] (SVR)	-[a]	-	18.5	0.52	-	-
Two-stage SVR	14.9	0.68	15.7	0.66	15.7	0.67

Table 1: Comparison of performances of two-stage SVR approach in predicting real values of ASA, based on PSSMs generated from PSI-BLAST profiles, with other approaches on the Manesh, Barton, and Carugo datasets. [a]Dashes indicate that the corresponding results were not available from the literature.

Method / Accuracies	Mean Absolute Error (%)	Correlation Coefficient
[a]Adamczak et al.[23] (NN)	15.0	0.68
Two-stage SVR	14.6	0.69

Table 2: Comparison of performances of two-stage SVR approach in predicting real values of ASA with the regression method using neural networks of Adamczak *et al*. on 199 proteins from the Manesh dataset. [a]The method of Adamczak et al. was trained on a large dataset of 860 proteins.

Method / Threshold	5%	9%	10%	16%	20%	25%	36%	50%	60%	70%	80%	90%
[a]Manesh et al.[14]	-[b]	75.9	-	75.5	-	74.4	74.1	-	-	-	-	-
Ahmad et al.[17] (NETASA)	74.6	-	71.2	-	-	70.3	-	75.9	-	-	-	-
Gianese et al.[21] (PP)	75.7	-	73.4	-	-	71.6	-	76.2	-	-	-	-
Nguyen and Rajapakse[25] (Two-stage SVM)	82.9	-	81.0	-	78.6	78.1	-	79.1	83.4	-	-	-
Two-stage SVR	81.1	78.7	78.5	77.9	77.6	77.3	76.9	79.5	84.3	89.9	95.0	97.5
Rost and Sander[4] (PHDacc)	-	74.6	-	75.0	-	-	-	-	-	-	-	-
Cuff and Barton[19] (Jnet)	79.0	-	-	-	-	75.0	-	-	-	-	-	-

Table 3: Comparison of performances of two-stage SVR approach in RSA prediction at different thresholds based on PSSMs generated by PSI-BLAST, with other methods on the Manesh dataset with the training set of 30 proteins and the testing set of 185 proteins. [a]The information theoretical method of Manesh et al. used a jack-knife validation to estimate the prediction accuracy. [b]Dashes indicate that the corresponding results were not available from the literature.

Threshold / Accuracies	Mean Absolute Error (%)	
	Buried	Exposed
5%	11.5	17.4
10%	12.0	17.8
20%	12.4	18.8
25%	12.4	19.4

Table 4: The mean absolute errors of residues in buried and exposed parts at different thresholds on the Barton dataset of 502 nonhomologous proteins.

Properties of the Proteins	In the Whole Dataset	Within the Lowest 5% MAE	Within the Largest 5% MAE
Mean length	166.9	155.1	50.4
Mean hydrophobicity value	-0.3	0.1	-0.5
Correlation coefficient between mean hydrophobicity and MAE	-0.22	-	-

Table 5: The mean length and mean hydrophobicity value of proteins within the lowest 5% mean absolute errors (MAE) and within the largest 5% mean absolute errors in ASA prediction compared to those in the whole dataset of the Barton dataset; the Pearson's correlation coefficient of mean hydrophobicity scale and MAE is given for the whole dataset as the numbers of proteins in the lowest 5% and the highest 5% of MAE are small to compute the correlation coefficients with a significant accuracy.

Amino acid		Occurrence (%)	Hydrophobicity scales	Mean Absolute Error (%)	
				Single-stage SVR	Two-stage SVR
Non-polar group (hydrophobic)					
Gly	G	7.9	-0.4	20.1	19.6
Ala	A	8.8	1.8	14.8	14.4
Val	V	7.0	4.2	11.1	10.7
Leu	L	8.5	3.8	11.1	10.8
Ile	I	5.6	4.5	10.1	9.7
Met	M	2.1	1.9	12.5	12.1
Pro	P	4.7	-1.6	17.9	17.7
Aromatic group (hydrophobic)					
Phe	F	3.9	2.8	11.6	11.2
Trp	W	1.5	-0.9	12.7	12.4
Tyr	Y	3.6	-1.3	13.2	12.9
Polar, uncharged group (hydrophilic)					
Ser	S	6.2	-0.8	19.3	18.8
Thr	T	6.0	-0.7	17.1	16.7
Cys	C	0.9	2.5	10.5	9.9
Asn	N	4.8	-3.5	20.7	20.2
Gln	Q	3.7	-3.5	17.9	17.6
Positively charged (hydrophilic)					
Lys	K	6.0	-3.9	16.5	16.4
Arg	R	5.5	-4.5	17.1	17.0
His	H	2.2	-3.2	15.5	15.4
Negatively charged (hydrophilic)					
Asp	D	6.0	-3.5	19.9	19.5
Glu	E	6.1	-3.5	18.9	18.3

Table 6: The properties of 20 amino acids: their average occurrences, hydrophobicity scales, and the mean absolute error in ASA prediction on the Barton dataset.

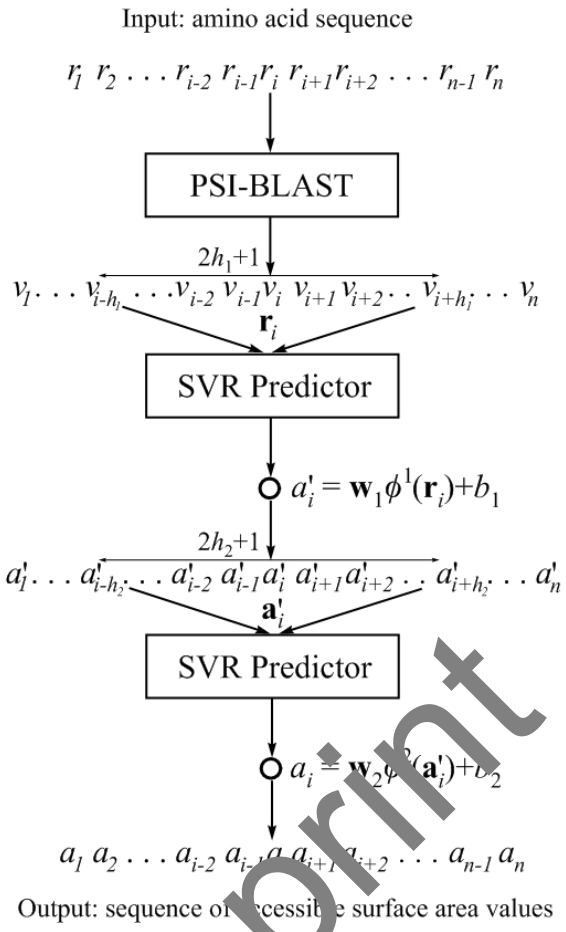


Figure 1: Two-stage SVR approach for the prediction of the real value accessible surface area (ASA).

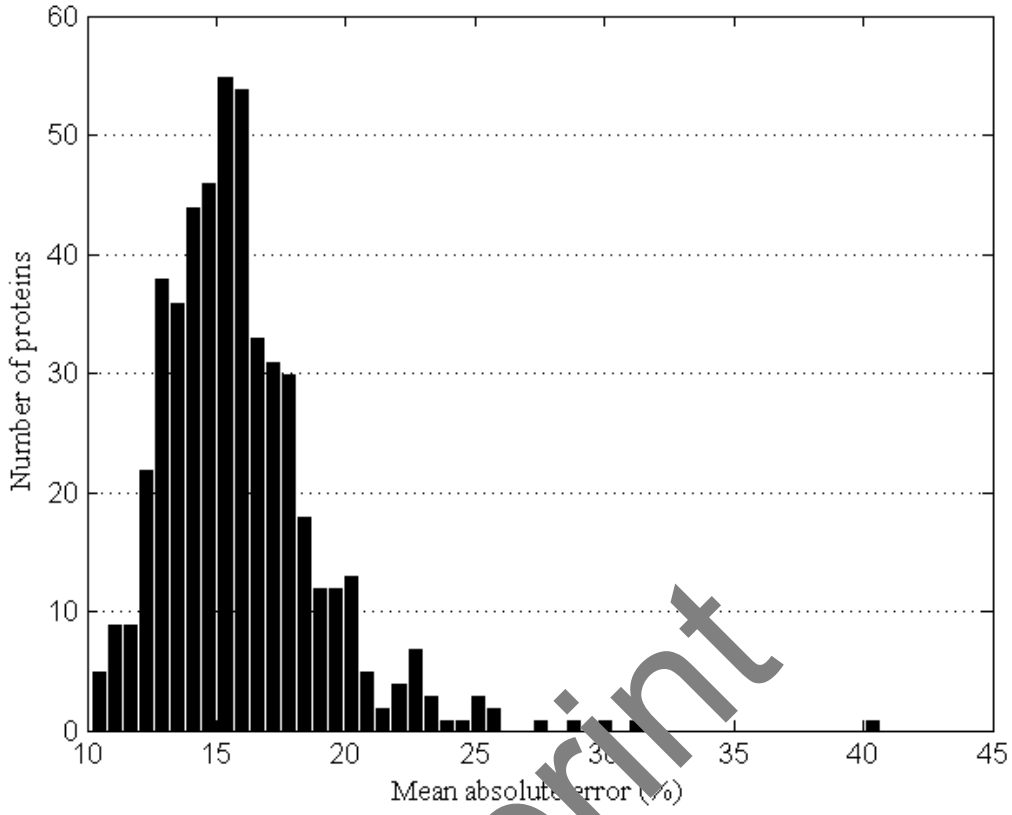


Figure 2: The distribution of protein mean absolute errors obtained by two-stage SVR method in predicting ASA values for the benchmark 502 proteins of the Barton data set.

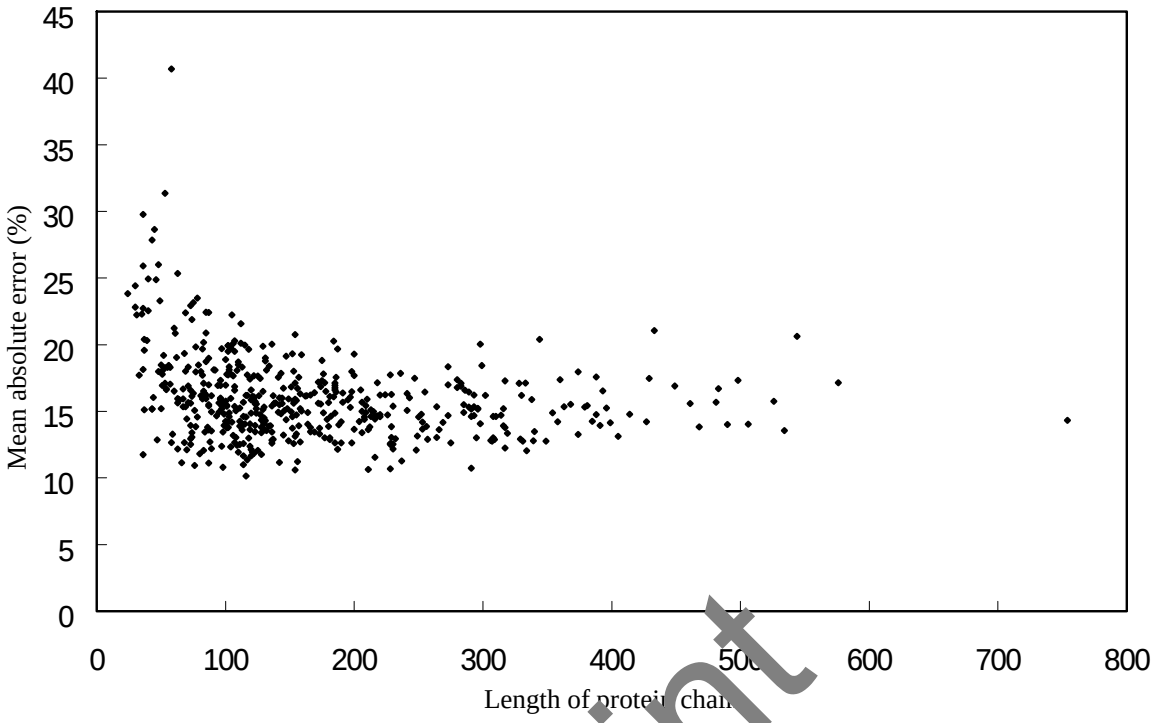
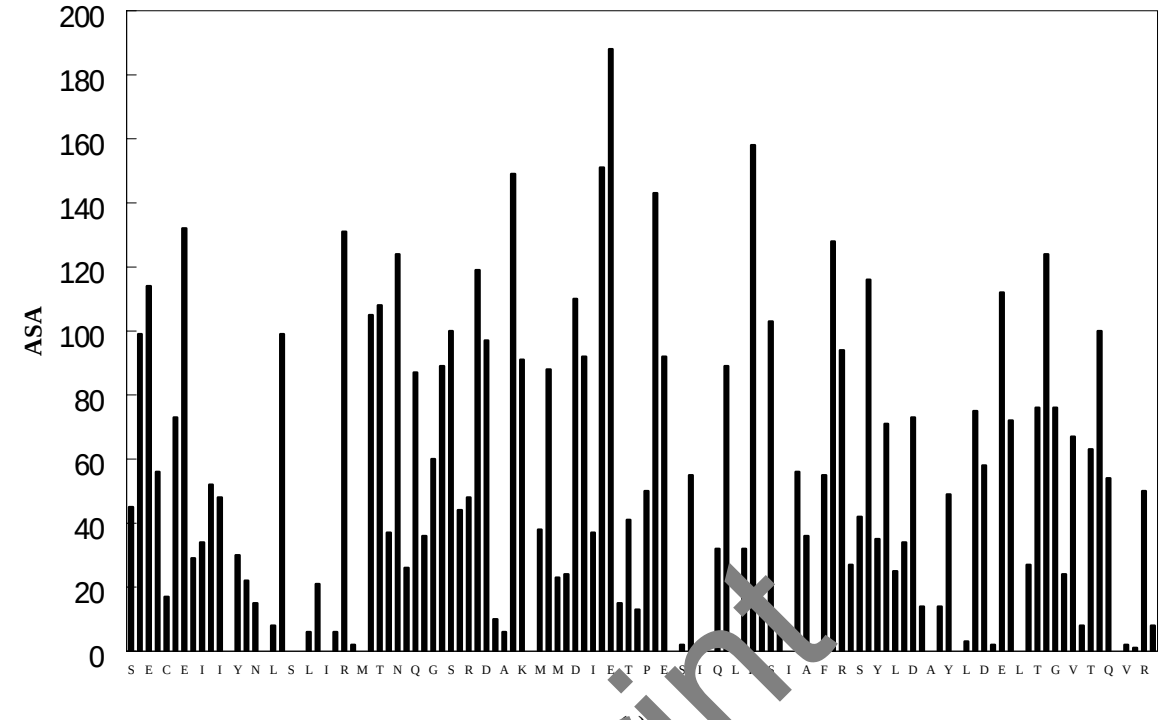
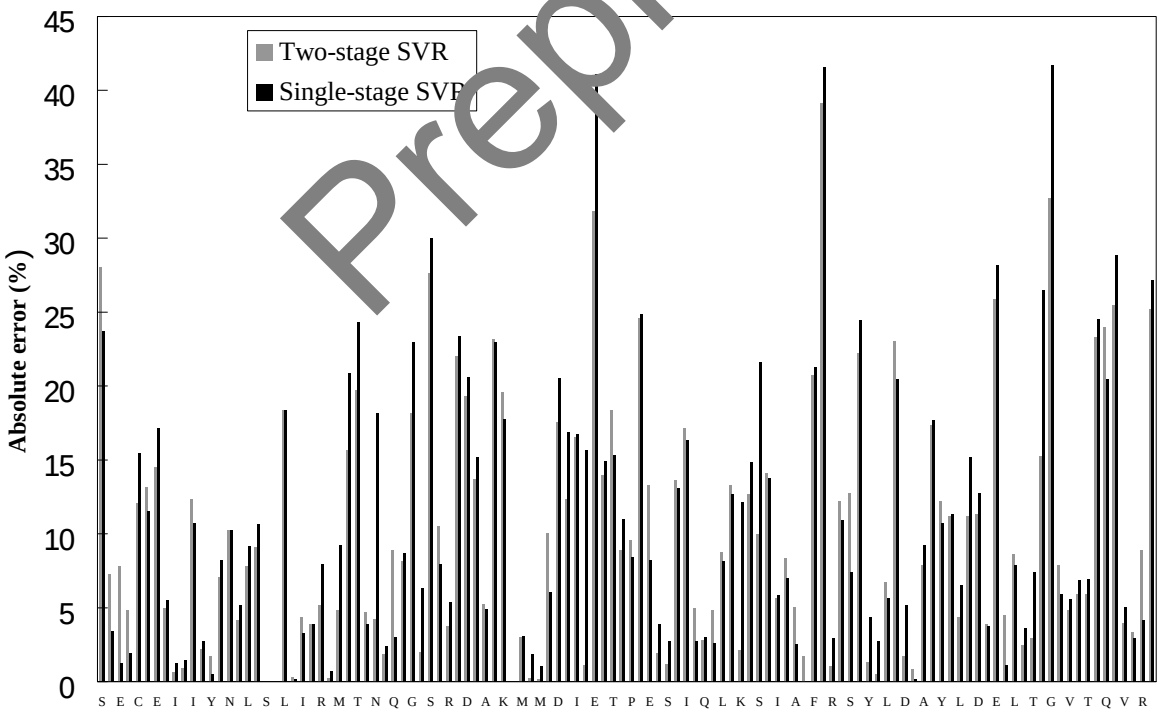


Figure 3: The distribution of protein mean absolute errors according to the lengths of protein chains in predicting the ASA values by two-stage SVR method for the benchmark Barton datas

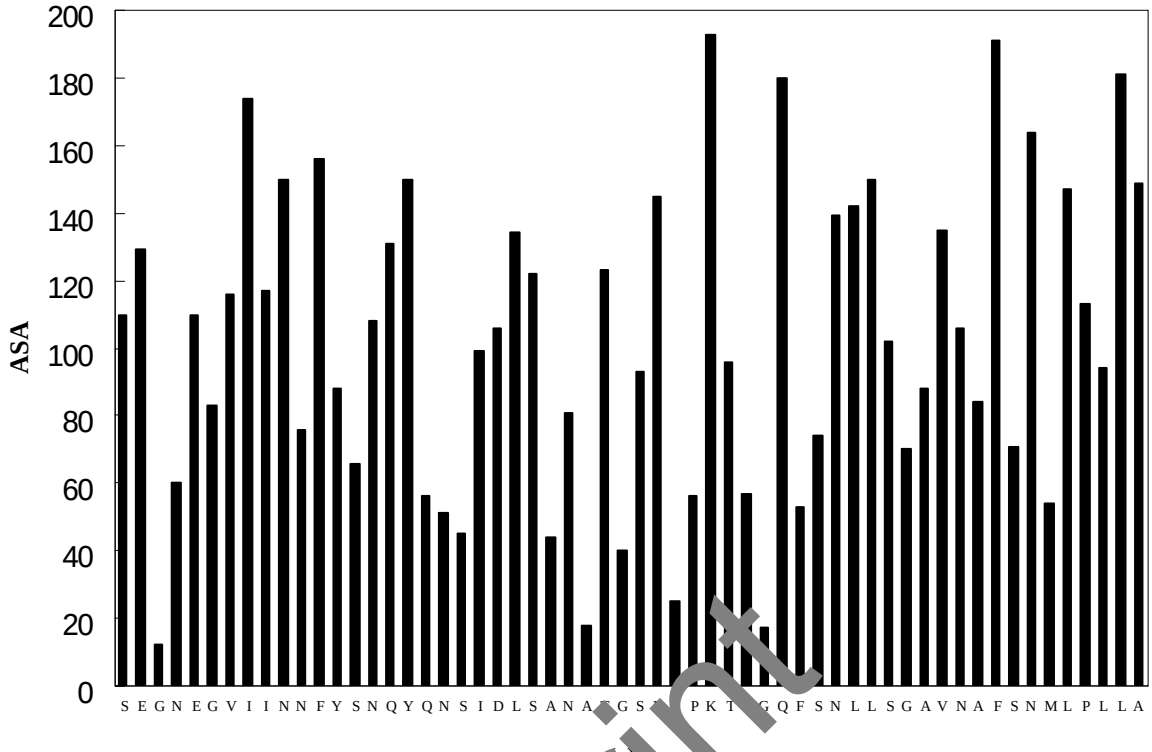


(a)

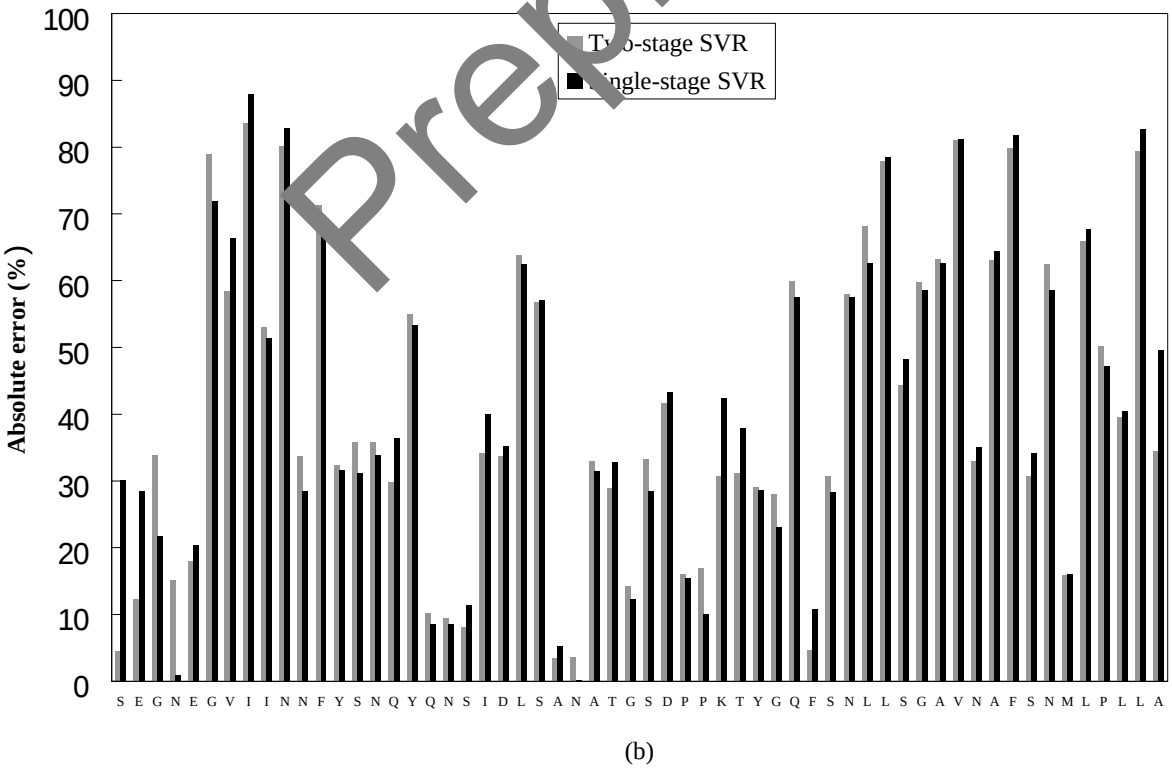


(b)

Figure 4: (a) Observed ASA values for the protein (PDB code 1TND:B) with the lowest mean absolute error, 10.1%, in the prediction by using two-stage SVR, and (b) the absolute errors for predicted values of single-stage SVR and two-stage SVR methods. The single-stage SVR method resulted in a mean absolute error of 11.0% for this protein.



(a)



(b)

Figure 5: (a) Observed ASA values of the protein (PDB code 2MEV:4) with the largest mean absolute error, 40.7%, in prediction by using two-stage SVR method, and (b) the absolute errors for predicted values of single-stage SVR and two-stage SVR methods. The single-stage SVR method resulted in a mean absolute error of 41.5% for this protein.