Prediction of protein accessible surface areas by support vector regression

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · NOVEMBER 2004

Impact Factor: 2.63 · DOI: 10.1002/prot.20234 · Source: PubMed

2 AUTHORS:

Zheng Yuan
University of Queensland
22 PUBLICATIONS **918** CITATIONS
SEE PROFILE

Bixing Huang
Queensland Health
34 PUBLICATIONS **613** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Bioinformatics 57:558–564 (2004)

Prediction of Protein Accessible Surface Areas by Support Vector Regression

Zheng Yuan[1]* **and Bixing Huang**[2]
[1]*Institute for Molecular Bioscience and ARC Centre in Bioinformatics, University of Queensland, St. Lucia, Australia*
[2]*Public Health Microbiology Laboratory, Queensland Health Scientific Services, Coopers Plains, Australia*

ABSTRACT **A novel support vector regression (SVR) approach is proposed to predict protein accessible surface areas (ASAs) from their primary structures. In this work, we predict the real values of ASA in squared angstroms for residues instead of relative solvent accessibility. Based on protein residues, the mean and median absolute errors are 26.0 Å^2 and 18.87 Å^2, respectively. The correlation coefficient between the predicted and observed ASAs is 0.66. Cysteine is the best predicted amino acid (mean absolute error is 13.8 Å^2 and median absolute error is 8.37 Å^2), while arginine is the least predicted amino acid (mean absolute error is 42.7 Å^2 and median absolute error is 36.31 Å^2). Our work suggests that the SVR approach can be directly applied to the ASA prediction where data preclassification has been used. Proteins 2004;57:558–564.**

Key words: protein structure prediction; machine learning; accessible surface area; solvent accessibility; support vector; protein sequence analysis

INTRODUCTION

An important approach toward predicting the structure of a protein is to predict structural properties such as secondary structure and solvent accessibility. Solvent accessibility reflects the degree to which a residue interacts with the solvent molecules and is a valuable reporter on the folding state of a protein. Because active sites of proteins are often located at their surface, the prediction of exposed residues is important for understanding and predicting the relationship between the structure and function of a protein. Furthermore, accurate prediction of solvent accessibility can aid the prediction of other structural properties such as protein secondary structure.

Prediction of protein solvent accessibility may use different ways to code protein sequences. The coding can be based on a single sequence or a group of homologous sequences (multiple sequence alignment). The prediction methods based on simple back-propagation neural networks use multiple sequence alignment information[1,2] or only single-sequence information.[3,4] The multiple sequence alignment information is also used in other methods, such as recurrent neural networks,[5] support vector machines,[6] and probability profiles.[7] The methods of Bayesian theory,[8] information theory,[9,10] and multiple linear regression[11] are based only on single-sequence input. In addition to using protein sequence information, a recently developed method also took into account long-range interaction extracted from protein structures.[12] All the above methods use solvent states defined by different thresholds. This has the following drawback: If 2 or 3 states are used, the accuracy of prediction will decrease. In addition, the arbitrary choice of cutoff thresholds makes it difficult to compare results obtained from different methods. To overcome these shortcomings, a method has been recently proposed to predict consecutive values of relative solvent accessibility.[13]

In previous studies, solvent accessibility was defined as relative solvent accessibility (RSA). The absolute value of solvent accessibility for an amino acid is its accessible surface area (ASA) in a protein structure. The RSA is obtained by normalizing the ASA value over the maximum value of exposed surface area obtained for either (1) each amino acid[1] or for (2) an extended tripeptide conformation of Ala-X-Ala or Gly-X-Gly.[14] The basis for taking this step was that different amino acids have different propensities for being on protein surfaces (different scales). Amino acids have different ASA distributions with largely different mean and median values.[15] However, it is difficult to compare the prediction performance of various methods, if the comparison is based on RSA and different normalizing values are used. Furthermore, it is difficult to compare the prediction accuracy for different amino acid types if the same RSA threshold is chosen. For example, in the extended Ala-X-Ala conformation, ASAs for glycine (Gly) and tryptophan (Trp) are 70.27 Å^2 and 209.57 Å^2, respectively.[14] If residues are classified as buried or exposed states with the same RSA cutoff threshold, different ASA cutoff thresholds are to be adopted according to different amino acids. The RSA threshold 20% is equal to an ASA threshold of 14.1 Å^2 for Gly and 41.9 Å^2 for Trp. This means that Gly is regarded as exposed if its ASA is greater than 14.1 Å^2, while Trp is regarded as buried if its ASA is

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsor: Australia Research Council.

*Correspondence to: Zheng Yuan, Institute for Molecular Bioscience, University of Queensland, St. Lucia 4072, Australia.
E-mail: z.yuan@imb.uq.edu.au

Received 5 January 2004; Accepted 14 May 2004

Published online 29 July 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20234

greater than 14.1 Å^2 but less than 41.9 Å^2. This definition is not applicable when we explore a protein surface as a whole, in particular, protein-binding sites containing a set of amino acids. In this case, it is more meaningful to know their total ASA than their individual relative accessibility. To solve this problem, one approach is to use ASA values directly. In the previous protein solvent accessibility predictions, the overall prediction accuracy is used as an important index to reflect the prediction performances. Approximately 70–75% of residues could be correctly predicted if the RSA cutoff threshold was set at the range of 20–25%. However, this index is not adequate. In this study, we regard the goal of predicting protein solvent accessibility as the prediction of protein ASA, because ASA can directly reflect the degree to which residues are in contact with the solvent molecules.

Here, we propose a new approach to predict protein ASA values based on support vector regression (SVR). A well-prepared data set is used to test this approach. The prediction results of this method have been compared with those of the neural network method[13] and support vector machine (SVM) classification.[6] Our results show that the new approach can predict the ASA values, with a mean absolute error of 26.0 Å^2 and a correlation coefficient of 0.66. Moreover, adoption of different amino acid scales is an important step for achieving higher prediction accuracy. This approach can be used to predict the exposed and buried residues for a novel protein and be helpful in analyzing its structural and functional properties.

METHODS

Support Vector Regression (SVR)

The goal of the regression formulation is to estimate an unknown continuous-valued function based on a finite number of samples. In this study, we try to find a regression formula based on training samples presented by M observations. Each observation consists of a pair: a vector x characterizing a residue in the sequence and its associated ASA value y. The principle of SVR is formulated[16,17] to estimate the following function by a linear regression:

$$f(x) = (w,x) + b \tag{1}$$

Here, (w,x) means the inner product of "weights" (w) and x, and b is the "bias". To generalize the support vector algorithm to the regression case, an analogue of the soft margin is constructed in the space of the target values y by using Vapnik's ϵ-insensitive loss function described by

$$L_\epsilon[y - f(x)] = \begin{cases} 0 & if \quad |y - f(x)| \le \epsilon \\ |y - f(x)| - \epsilon, & \text{otherwise,} \end{cases} \tag{2}$$

in which ϵ is the tolerance to error and only those deviations larger than ϵ are considered as errors. To estimate the function $f(x)$, we minimize the norm and the regularized empirical risk function:

$$\frac{1}{2}\|w\|^2 + \frac{C}{M}\sum_{i=1}^{M} L_\epsilon[y_i - f(x_i)], \tag{3}$$

where C is a regularization constant determining the trade-off between training errors and model complexity. This problem can be transformed into the constrained convex optimization problem by employing slack variable ξ and ξ^*

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{M}(\xi_i + \xi_i^*)$$

$$\text{subject to} \quad \begin{cases} f(x_i) - y_i \le \epsilon + \xi_i \\ y_i - f(x_i) \le \epsilon + \xi_i^* \\ \xi_i, \xi_i^* \ge 0 \text{ for } i = 1, \ldots, M. \end{cases} \tag{4}$$

To solve the optimization problem, Lagrange multipliers are added to the condition equations, and the above problem can be written as its dual form:

$$\text{maximize} \quad -\epsilon\sum_{i=1}^{M}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{M}(\alpha_i - \alpha_i^*)y_i - \frac{1}{2}\sum_{i,j=1}^{M}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)(x_i,x_j)$$

$$\text{subject to} \quad 0 \le \alpha_i, \alpha_i^* \le C \text{ and } \sum_{i=1}^{M}(\alpha_i - \alpha_i^*) = 0, \tag{5}$$

where α_i and $\alpha_i^*$are Lagrange multipliers to be solved. Only the nonzero values of Lagrange multipliers are useful in predicting the regression line, and their corresponding samples are known as support vectors. All points located in the ϵ-tube have Lagrange multipliers as zero, thereby not contributing to the regression. Extension of SVR to nonlinear functions is realized by introducing transformation function $\Phi(x)$ to map the data point in the input space to a higher dimensional feature space. Therefore, (x_i,x_j) in Eq. (5) may be replaced by $\Phi(x_i) \cdot \Phi(x_j)$, which is defined as kernel function $k(x_i,x_j)$. $k(x,x')$ can be polynomial kernel

$$k(x,x') = (x \cdot x' + 1)^n \tag{6}$$

or radial basis function (RBF)

$$k(x,x') = \exp(-\gamma\|x - x'\|^2) \tag{7}$$

In addition, w in Eq. (1) can be given as

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\Phi(x_i), \tag{8}$$

and Eq. (1) can be rewritten as

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)k(x_i,x) + b, \tag{9}$$

where l is the number of support vectors. Once the support vectors, their Lagrange multipliers, and the bias b are determined from the training samples, Eq. (9) can be used to predict the ASA values of a novel protein. The above algorithm was implemented by SVM_light.[18]

TABLE I. Prediction Results Based on Different Models Examined by 3-Fold Cross-Validation

Control Parameters			Absolute Error (Å^2)		
Normalization	γ	C	Mean	Median	Correlation
Method 1	0.1	0.7	29.4	23.66	0.62
	0.01	2.0	29.0	23.66	0.62
	0.01	5.0	29.8	22.94	0.60
Method 2	0.1	0.7	28.9	22.85	0.62
	0.01	2.0	27.2	21.37	0.66
	0.01	5.0	26.0	18.87	0.66

Sequence Coding and ASA Normalization

The vector x representing a residue is extracted by a sliding window coding scheme using single sequences as input.[1] The window size is set at 15 residues. The ASA value is computed for the central residue. In the sliding window, each residue is coded by a 21-dimensional vector, representing the 20 types of amino acids plus one unit for the breaks and uncommon amino acids. Therefore, a residue is represented by a (21 * 15) = 315-dimensional vector.

The absolute ASA value for each residue is obtained using DSSP (Dictionary of Protein Secondary Structure).[19] We normalize ASA values using two different normalization methods. In the first normalization method (referred to as "method 1"), all values are divided by 317, the maximum ASA value observed in the data set used (see Results and Discussion section). Thus, the normalized values are within the range [0, 1]. In the second normalization method (referred to as "method 2"), the ASA values are divided by the corresponding value for the extended Ala-X-Ala conformation of the different amino acid types. To allow the comparison of our method with previous methods, we use the same values of Ala-X-Ala as given in Ahmad et al.[13] It is worth noting that the normalization step can simplify the handling of data, as the ASA values of different amino acids are at the same scale ("method 2"). However, the final predicted ASA values in squared angstroms were those transformed from normalized values. The SVR algorithm is trained on the normalized ASA data and, therefore, the predicted value is still regarded as the normalized values when one makes a prediction. Although two different normalizing methods are used, the predicted results can be transformed back to their corresponding ASA absolute values. Therefore, the performances of the two normalization methods can be compared based on real ASA values.

Database and Prediction Accuracy Measurement

Protein chains of 1277 were selected using PDBREPRDB[20] from the Protein Data Bank (PDB).[21] All proteins are not shorter than 60 amino acids in length, and the pairwise identity is not more than 25%. Protein structures solved by X-ray crystallography are with resolution $\leq$ 2.0 Å^2 and R-factor $\leq$ 0.2. The names of protein chains are given in Table V (see supplementary material). To perform 3-fold cross-validation tests, we randomly divided this data set into 3 groups, each containing a roughly equal number of protein sequences. One group in turn was chosen as the testing set, while the proteins in other groups were merged to form a training set.

To measure the performance of SVR in this application, the absolute error (AE) and Pearson's correlation coefficient between predicted and observed ASA values were calculated. To compare our results with those obtained previously using predefined solvent states, we selected a variety of thresholds and calculated 2-state classification accuracies. The accuracy is defined as the percentage of correctly predicted residues among total residues.

RESULTS AND DISCUSSION

Predicting ASA Values of Protein Residues

Among the 256,715 residues in the 1277 protein chains selected, the largest ASA value was 317 Å^2. This value was used in method 1 (see Methods section) to normalize ASA values. To train the algorithm, the value of ϵ was set as 0.01 (Eq. 2) and the RBF kernel (Eq. 7) was selected. Different values of γ and C were tried, and the results are given in Table I. Because all predicted ASA values should be in the range [0, 317], we assigned all the values less than zero as zero and those greater than 317 as 317. This step can improve prediction performance slightly. Using normalization method 2 and setting $\gamma = 0.01$ and $C = 5.0$, we obtained the best prediction results. Figure 1 shows the distribution of AEs for all residues. This distribution is far from normal. To best describe this skewed distribution, we use not only the mean value but also the median value. The mean and median AEs of this skewed distribution were 26.0 Å^2 and 18.87 Å^2, respectively. The correlation coefficient between the predicted and observed ASA value was 0.66. These results show that the second normalization method is better than the first. We presume that this is because method 2 takes into account the properties of various amino acids when normalizing the data, while method 1 treats different amino acids as the same. Therefore, normalization method 2 was used in all subsequent computer simulations. The performance of other kernel functions was inferior to radial basis function. For example, using the polynomial kernel function $k(x,x') = (x \cdot x' + 1)^2$, the mean and median values of AEs were 29.3 Å^2 and 22.36 Å^2, respectively. The correlation coefficient was 0.60.

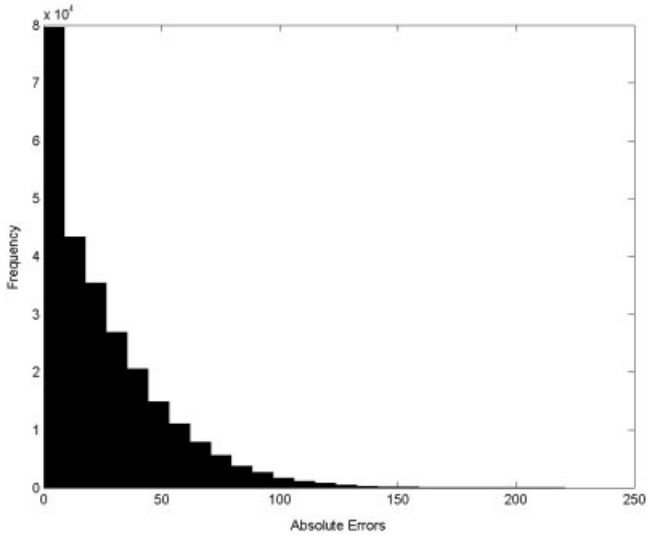

Fig. 1. Distribution of absolute errors for protein residues. Mean and median values are 26.0 Å^2 and 18.87 Å^2, respectively.

The mean AEs for individual proteins are related to their lengths. The error distribution is shown in Figure 2. Given an error cutoff of 30 Å^2, 82.6% of long sequences (>150 amino acids) and 53.1% of short sequences (≤150 amino acids) have been predicted with errors lower than this threshold. These results suggest that predictions for small proteins are less accurate.

Listed in Table II are mean and median values for AEs of the amino acids. Cys (C) is predicted with the least errors (mean value is 13.8 Å^2 and median value is 8.37 Å^2), while Arg (R) is predicted with the greatest errors (mean 42.7 Å^2 and median 36.31 Å^2).

To understand where the difficulties in prediction lie, we selected two prediction examples for further analysis (Fig. 3). As an example of a well-predicted protein, Figure 3(A) shows the predicted and observed ASA values for a contractile protein (PDB ID: 1F3C). As an example for a poorly predicted protein, Figure 3(B) shows ASA values for Myxoma virus protein (PDB ID: 1JJG). Protein 1F3C [Fig. 3(A)] is well predicted, with a mean AE of 15.1 Å^2. The observed (solid line) and predicted (dashed line) values are in high consensus. Most of the positions of ASA maxima are well predicted. 1JJG is poorly predicted, with a mean AE of 44.6 Å^2. The worst predicted region is from position 21 to position 31, where the peak values are most strongly underpredicted.

In particular, it appears that residues with high ASA values are consistently underpredicted. In order to verify the hypothesis, we classified the residues into 3 groups: buried (ASA = 0 Å^2), moderately exposed (0 Å^2 < ASA < 120 Å^2), and highly exposed (ASA ≥ 120 Å^2). They cover 13%, 77%, and 10% of the total residues, respectively. The mean AEs for the 3 groups, with median values in parentheses, are 17.1(12.4) Å^2, 23.5(17.7) Å^2, and 58.0(54.9) Å^2, respectively. It is clear that the largest errors are from the predictions of highly exposed residues. Indeed, more than 95% of the highly exposed residues have predicted values less than their observed ones. Underprediction of the ASA values of highly exposed residues is one of the systematic errors for this method. This phenomenon may be attributed to the unbalanced distribution of data points in the data set. Larger ASA values always occur with a lower number of residues. Due to the relatively small number of highly exposed residues in the data set, they cannot be well learned. To solve this problem, a weight may be given to the exposed residues when we train the SVR algorithm.

Comparison of SVR With Other Methods

The work most closely related to the present study is the recently developed neural network method, which predicted real values of RSA.[13] This neural network method was used to examine a number of data sets, with sizes varying from 126 protein chains to 502 protein domains ("CB-502 data set"). The mean AEs were between 18.8% and 19.4% for these data sets, irrespective of their sizes.[13] The reported correlation coefficients were between 0.4718 and 0.4870. To compare with the method, we used SVR to predict the real values of RSA on the CB-502 data set, using the same testing procedure. Six-round tests were performed, and the results are the average.[13] Using the parameters $\epsilon = 0.01$, $\gamma = 0.01$, and $C = 1.0$, SVR achieved a mean AE of 18.5% and a correlation coefficient of 0.520. In contrast, the neural networks gave a mean AE of 18.8% and a correlation coefficient of 0.482. We show all results in Table III. Hence, SVR achieved comparable or slightly better results. Based on the CB-502 data set, more than 25,000 support vectors were used for prediction. When our large data set (1277 protein chains; see Table V, supplementary materials) was used, an improvement was observed. The mean AE decreased to 17.0% and the correlation coefficient increased to 0.617. SVM parameters were set as $\epsilon = 0.01$, $\gamma = 0.01$, and $C = 5$, and more than 160,000 data points were taken as support vectors. This improvement is attributed to the utility of a large data set. Since more data carry more knowledge, a classifier that was trained on a larger data set can learn the rules more accurately. However, nearly 130,000 support vectors are responsible for a relatively minor improvement, suggesting that the method may reach its limit if prediction is based solely on protein sequences.

We also examined our approach on prediction of solvent states. In previous studies, solvent states were defined as exposed and buried by one threshold or as exposed, buried, and partially buried by two thresholds. Based on the new, large data set, we applied SVMs to the 2-class classification problem—exposed or buried states of a residue.[6] The performance was measured by the percentage of correctly predicted residues among total residues (accuracy). In the formulation of the ASA prediction problem as a classification problem, we preclassified the data into buried and exposed states, trained SVMs, and then computed the prediction accuracy. In the formulation as a regression problem, we trained the SVMs, classified the prediction results based on different cutoff thresholds, and then computed the prediction accuracy. Unlike previous methods based on RSA, we defined the states based on ASA values. Different models have been examined by 3-fold

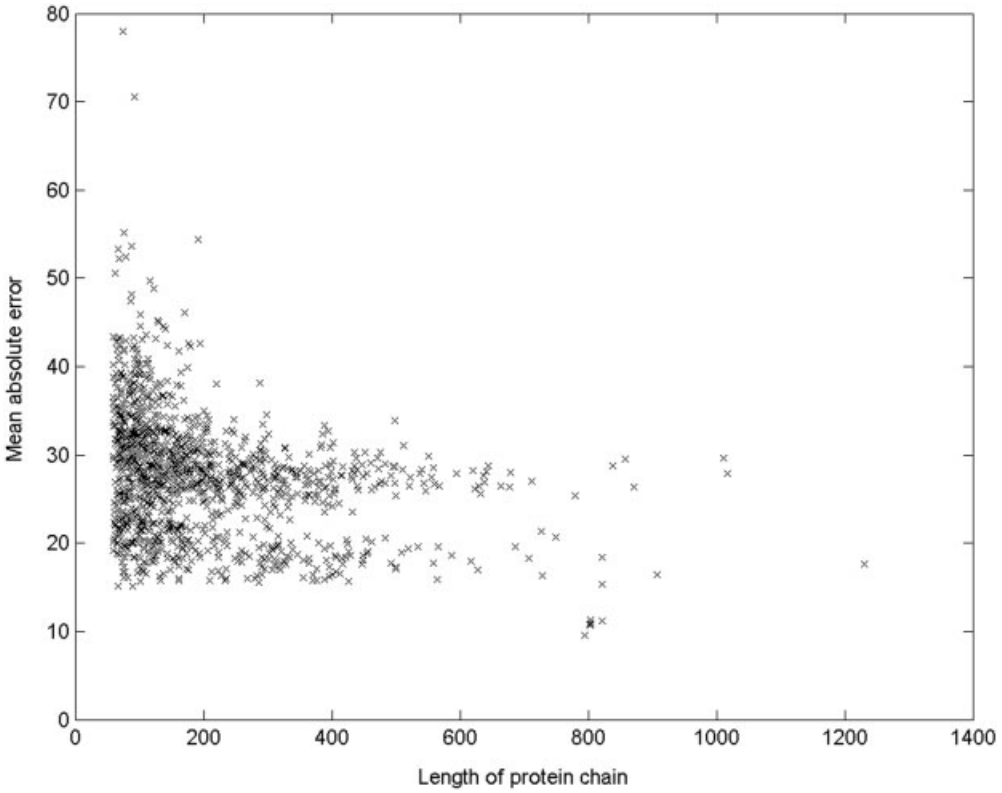

Fig. 2. Distribution of protein mean absolute errors according to protein lengths.

TABLE II. Mean and Median Absolute Errors of ASA (Å^2) for Different Amino Acid Types

Amino acid	Mean	Median
A	17.8	12.81
C	13.8	8.37
D	31.0	26.80
E	34.0	29.21
F	23.5	13.88
G	16.5	13.31
H	31.9	25.04
I	19.2	11.48
K	33.3	27.30
L	21.3	12.99
M	26.2	17.00
N	32.6	28.89
P	27.8	24.25
Q	35.1	30.48
R	42.7	36.31
S	24.6	20.78
T	25.9	21.30
V	18.2	11.15
W	29.4	19.39
Y	31.1	22.03

cross-validation. Nine cutoff thresholds from 5 Å^2 to 100 Å^2 were used, and the accuracies were calculated. The results are shown in Table IV. For each model, we examined a set of control parameters and reported the best results.

Models 1 and 2 are SVRs but use different methods of normalization. As we observed before, normalization method 2 outperformed method 1, even if we used a new index to measure the prediction accuracy. Neither model 1 nor model 3 takes into account the different scales of amino acids. Nonetheless, model 3 (SVM classification) performs at least comparably if not better than model 2. However, additional parameters may have an impact on their results. For example, different cost functions may be applied to the regression problem, and balance training may be used in classification problem. When a cutoff threshold of 34 Å^2 is used, the data are split into two groups with a roughly equal number of residues. At this threshold, model 1, model 2, and model 3 give the accuracies 70.2%, 75.0%, and 73.2%, respectively. The comparable results from classification and regression suggest that we can apply the regression problem directly here to avoid the arbitrary definition of states. Furthermore, the use of scales for different amino acids clearly improves the prediction results significantly. Of models 1 and 2, model 2 is always more accurate. Given a threshold of 34 Å^2, about 5% improvement can be achieved. Therefore, it is worthwhile to find an optimal set of normalizing values.

CONCLUSIONS

ASA values of residues in a protein sequence constitute its solvent profile, an important property of proteins. Many efforts have been made in the last few years to improve the prediction accuracy, but a great achievement seems to be

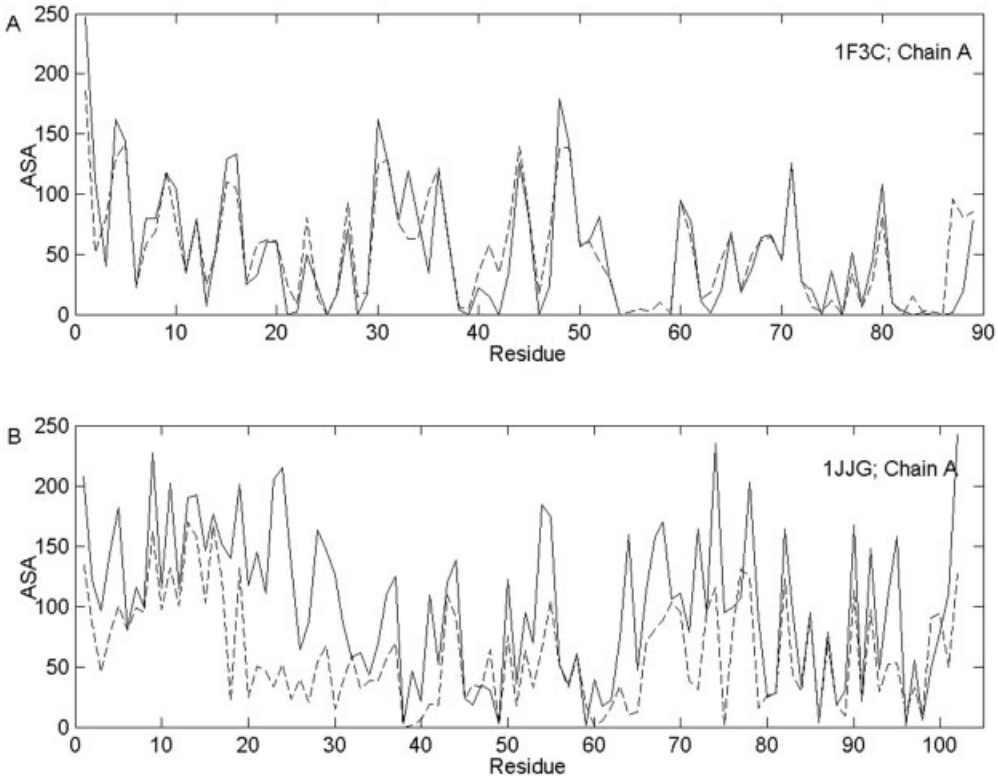

Fig. 3. Predicted solvent profile for protein 1F3C chain A and protein 1JJG chain A. Solid lines represent the observed ASA values and dashed lines represent the predicted values. (**A**) 1F3C chain A, with mean absolute error 15.1 Å^2; (**B**) 1JJG chain A, with mean absolute error 44.6 Å^2.

TABLE III. Prediction Results for Neural Network (NN) and Support Vector Regression (SVR) Based on Different Data Sets

Method	Data Set	Mean Absolute Error (%)	Correlation Coefficient
NN[a]	RS-126	19.4	0.477
	Carugo-338	19.0	0.490
	CB-502	18.8	0.482
SVR	CB-502	18.5	0.520
	YH-1277[b]	17.0	0.617

[a]All results for neural network method are from Ahmad et al.[13] Data set RS-126 was first used by Rost and Sander[1], data set Carugo-338 was given by Carugo,[9] and CB-502 is a subset of the data set given by Cuff and Barton.[2]
[b]YH-1277 is the data set of 1277 protein chains in this work.

TABLE IV. Prediction Accuracy (%) for Support Vector Regression and Support Vector Machine Classification

Thresholds (Å^2)	Model 1	Model 2	Model 3
5	76.3	78.5	78.2
15	70.1	74.6	73.7
25	69.1	74.4	72.9
34	70.2	75.0	73.2
45	72.7	76.2	74.4
60	76.4	78.8	77.0
75	80.1	81.9	80.3
90	83.6	85.0	83.7
100	86.0	87.1	86.1

Model 1: support vector regression, normalization method 1, $\epsilon = 0.01$, $\gamma = 0.01$, and $C = 2.0$; Model 2: support vector regression, normalization method 2, $\epsilon = 0.01$, $\gamma = 0.01$, and $C = 5.0$; and Model 3: support vector classification, $\gamma = 0.1$, and $C = 1.0$.

very difficult.[2,6,7,9–11,13] In this study, we provide an SVR approach for prediction of protein ASA real values. Examined on a large data set, ASA values can be predicted with a mean AE of 26.0 Å^2 and the correlation coefficient 0.66. Formulated as 2-class problems, the prediction accuracy is greater than 74%. Probably, the accuracy for predicting protein solvent accessibility from sequence information has a limit (correlation coefficient about 0.7). In our view, therefore, addressing the question of how solvent accessibility predictions can be used to address biologically important functions of proteins is a more pressing problem than achieving mere improvements in prediction accuracy. For example, solvent accessibility may be helpful to solve specific biological problems, such as fold recognition, by incorporating the predicted ASA values into the substitution matrix[22] and protein subcellular localization predictions using the information of predicted surface residues.[23]

ACKNOWLEDGMENTS

Our thanks to the reviewers for their helpful suggestions. The time-consuming computer simulations were performed at the High Performance Computing Facility at the University of Queensland. We also thank Tim Bailey, John Mattick, Rohan Teasdale, and Lynn Fink for their helpful discussions.

REFERENCES

1. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
2. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
3. Holbrook SR, Muskal SM, Kim SH. Predicting surface exposure of amino-acids from protein-sequence. Protein Eng 1990;3:659–665.
4. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819–824.
5. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
6. Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins 2002;48:566–570.
7. Gianese G, Bossa F, Pascarella S. Improvement in prediction of solvent accessibility by probability profiles. Protein Eng 2003;16:987–992.
8. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
9. Carugo O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. Protein Eng 2000;13:607–609.
10. Naderi-Manesh H, Sadeghi M, Arab S, Moosavi-Movahedi AA. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459.
11. Li X, Pan XM. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
12. Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local discriptor. Proteins 2004;54:557–562.
13. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629–635.
14. Samanta U, Bahadur RP, Chakrabarti P. Quantifying the accessible surface area of protein residues in their local environment. Protein Eng 2002;15:659–667.
15. Lins L, Thomas A, Brasseur R. Analysis of accessible surface of residues in proteins. Protein Sci 2003;12:1406–1417.
16. Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag; 2000.
17. Smola A, Scholkopf B. A tutorial on support vector regression. NeuroCOLT Technical Report, NC-TR-1998-030, http://www.neurcolt.com; 1998.
18. Joachims T. Making large-scale SVM learning practical. In: Schölkopf B, Burges C, Smola A, editors. Advances in kernel methods—support vector learning. Cambridge, MA: MIT Press; 1999.
19. Kabsch W, Sander C. Dictionary of protein secondary structure—pattern-recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
20. Noguchi T, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. Nucleic Acids Res 2003;31:492–493.
21. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
22. Teodorescu O, Galor T, Pillardy J, Elber R. Enriching the sequence substitution matrix by structural information. Proteins 2004;54:41–48.
23. Andrade MA, O'Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. J Mol Biol 1998;276:517–525.