Prediction of Protein Solvent Accessibility Using Support Vector Machines

Zheng Yuan,[1,2*] Kevin Burrage,[2] and John S. Mattick[1]
[1]*Institute for Molecular Bioscience and ARC Special Centre for Functional and Applied Genomics, The University of Queensland, Brisbane, Australia*
[2]*Department of Mathematics, The University of Queensland, Brisbane, Australia*

ABSTRACT **A Support Vector Machine learning system has been trained to predict protein solvent accessibility from the primary structure. Different kernel functions and sliding window sizes have been explored to find how they affect the prediction performance. Using a cut-off threshold of 15% that splits the dataset evenly (an equal number of exposed and buried residues), this method was able to achieve a prediction accuracy of 70.1% for single sequence input and 73.9% for multiple alignment sequence input, respectively. The prediction of three and more states of solvent accessibility was also studied and compared with other methods. The prediction accuracies are better than, or comparable to, those obtained by other methods such as neural networks, Bayesian classification, multiple linear regression, and information theory. In addition, our results further suggest that this system may be combined with other prediction methods to achieve more reliable results, and that the Support Vector Machine method is a very useful tool for biological sequence analysis. Proteins 2002;48:566–570.** © 2002 Wiley-Liss, Inc.

Key words: protein structure prediction; machine learning; solvent accessibility; computer simulation

INTRODUCTION

The important issue in structural genetics is to solve the structures of gene products generated from the Human Genome Project and, further, to determine their functions. Compared with experimental techniques, computational methods are quick, automatic, and applicable for the analysis of a large amount of data, but the accuracies of current prediction methods are not good enough to identify the conformation of functional sites. Because the active sites of a protein are always located on its surface, accurately predicting the surface residues can be regarded as an important step toward determining its function. It has been observed that the distribution of surface residues of a protein is correlated with its subcellular environments and, consequently, using the information of surface residues has made an improvement in the prediction of protein subcellular location.[1] Prediction of protein surface residues from primary sequence has been intensively studied in the area of protein structure analysis.[2–9] Various methods developed in the last few years, although based on different databases and computational techniques, have reported prediction accuracies of approximately 70% for single sequence and 75% for multiple alignment sequences. One way to improve the prediction performance is to combine different prediction methods instead of relaying only on one method,[10,11] whereas another way is to modify the methods of defining surface residues of proteins.[9] In this study, the Support Vector Machine (SVM) method has been used as a new approach and compared with other methods.

The SVM method, recently developed by Vapnik and his collaborators,[12,13] is based on the structural risk minimization principle from computational learning theory. SVM maps the samples to a high-dimensional feature space, then constructs an optimal separating hyperplane that separates two classes (it can also be extended to multiclass problems). The hyperplane output by the SVM is given as an expansion on a small number of training points known as support vectors. The support vectors are always closest to the hyperplane and correspond to those points that are hardest to classify. SVM methods have been introduced to solve biological pattern recognition problems such as microarray data analysis,[14,15] protein fold recognition,[16] prediction of protein–protein interaction,[17] prediction of protein secondary structure,[18] and cancer diagnosis.[19] In this study, we applied SVM methods to the problem of protein solvent accessibility prediction. The SVM approach can achieve results better than or comparable to those of extant methods, such as neural networks,[3] Bayesian classification,[4] multiple linear regression,[8] and information theory.[9] Our results further suggest that the SVM approach can be combined with other methods to improve the prediction accuracy.

METHODS

Database and Prediction Accuracy Measurement

The non-redundant set of 531 protein domains selected by Cuff and Barton[11] was taken as the basis for training

Grant sponsor: Australia Research Council.

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

Received 18 March 2002; Accepted 12 April 2002

and testing of SVM. The residue solvent accessibility can be obtained by referring to their correspondent secondary structures files (DSSP-defined files) generated by the method of Kabsch and Sander.[20] Because this database was large enough for the training and testing, when we extracted the solvent accessibility value from the files, we simply excluded those with an inconsecutive residue number, which may contain mutation or deletion on protein sequence. However, the short-chain effect is not eliminated because several domains included in this database are less than 30 amino acids long. Thus, a total of 421 protein domains that remained were randomly divided into 10 groups with each one having 42 or 43 proteins. The relative solvent accessibility was calculated by normalizing the value over the maximum solvent accessibility values of amino acids. Residues were classified into two states (buried/exposed) by a cut-off threshold, three states (buried/partially buried/exposed) by two cut-off thresholds or more states.

Prediction accuracy is defined as the number of correctly predicted cases over the total number of cases. Tenfold cross-validation is adopted here by choosing one group in turn as the testing group, whereas the proteins in the other groups are merged, making the training group. Therefore, the prediction accuracy is finally given as the average of 10-fold cross-validation. Matthews coefficients[21] were used to reflect the correlation between predicted and observed results.

SVM Implementation and Sequence Coding

SVM was implemented using SVM^{light} (http://www-ai.cs.uni-dortmund.de/SOFTWARE/SVM_LIGHT).[22] Dot, polynomial, and radial basis functions are examined in this report.

The application of SVMs to a multi-class problem (e.g., three or more states of solvent accessibility) was extended by using one-versus-others technique.[16] When protein solvent accessibility is classified into m states by m-1 cut-off thresholds, m SVMs are needed to predict m states. For each state, a SVM is trained on the samples of this state as positive and all samples of other states as negative. Because 10-fold cross-validation is used here to test the prediction accuracy, 10 m SVMs are trained for an m-state problem. In the testing procedure, the state of a residue is simply predicted as the one with maximum SVM output value.

For sequence coding, we adopted the same sliding-window coding schemes as used in the references.[3,18] Each residue can be represented by a 21-dimensional vector. The first 20 units stand for 20 types of amino acids; the last one represents the break or uncommon amino acid. When the window size n is chosen, a residue can be represented by a vector of $21 \times n$ units. The N and C terminal breaks are added to make a complete vector for those positions. The evolution information is considered to improve the prediction accuracy by using multiple alignment protein sequences. The elements of the 21-dimensional vector are the occurring frequencies of 20 amino acids for each residue position.

TABLE I. Prediction Results for Different Kernel Functions and Parameters[†]

Kernel function	Parameters	Accuracy (%)
Dot[a]		69.0
Radial basis[b]	$\gamma = 0.1$	70.8
Polynomial[c]	$\alpha = 1, \beta = 1, d = 2$	70.5
	$\alpha = 1, \beta = 1, d = 3$	70.8
	$\alpha = 1, \beta = 1, d = 4$	70.9
	$\alpha = 0.1, \beta = 0, d = 2$	70.0

[†]For details, refer to References 12 and 22.
[a]Kernel function is $s_i \cdot x$.
[b]Kernel function is $\exp(-\gamma\|s_i - x\|^2)$.
[c]Kernel function is $(\alpha s_i \cdot x + \beta)^d$.

RESULTS AND DISCUSSION

Effect of Different Kernel Functions

Because the power of SVM comes from the kernel representation that allows a nonlinear mapping of the input space to a higher dimensional feature space, the choice of a proper kernel function is an important issue for SVM training. Under many circumstances, using the function that can map the data to some other Euclidean space and using an appropriate decision function can give a better classification. The decision function can be expressed as:

$$f(x) = \sum_{i=1}^{N_s} \alpha_i K(s_i, x) + b \quad (1)$$

where a_i is the multiplier, s_i is a support vector, b is the bias, and x is the vector that represents a certain residue. $K(s_i, x)$ is the kernel function that can take different forms. Here, we use the commonly used kernel functions, including dot, polynomial, and radial basis functions, illustrated in Table I. For a two-state (buried/exposed) problem, a residue is predicted to be exposed if $f(x) > 0$; otherwise, it is predicted to be buried. The parameters can be obtained by training SVMs on the training samples. In the training procedure, after a certain kernel function is chosen, the regularization parameter C needs to be tuned. For the dot function, C was set at 0.07; for the polynomial functions, it was set at $0.005 \sim 0.01$, whereas it was 0.5 for the radial basis function.

Protein residues have been classified into two states by a cut-off threshold of 20%, and the size of sliding-window has been chosen as 15. Prediction accuracies are given by averaging over the results of the 10-fold cross-validation test shown in Table I. It can be observed that, except for the dot function, the choices of other kernel functions did not make a significant difference to the prediction results. Using the dot kernel function gives a slightly lower accuracy, but the learning and testing time is quite short. On the contrary, although the higher-order polynomial function and radial basis function can give slightly better prediction performance, they also need a much longer training and testing time. Ignoring the marginal difference generated by the nonlinear kernel functions, we chose a certain function [i.e., $K(s_i,x) = (s_i \times x + 1)^2$] as the basic

TABLE II. Prediction Accuracies for Different Window Sizes

Window size	Accuracy (%)	Standard deviation of accuracy (%)
11	70.3	0.9
13	70.3	0.8
15	70.5	0.8
17	70.5	0.8
19	70.5	0.7
21	70.5	0.7

kernel function for the following computer simulations because it can retain the prediction accuracy and takes shorter learning and testing time.

Effect of Different Window Sizes

Using the kernel function $K(s_i,x) = (s_i \times x + 1)^2$ and a surface residue classification threshold of 20%, we chose different window sizes to observe the prediction results. Increasing the window size can provide more local information. It is reasonable to expect that prediction accuracy would increase with the enlargement of the window size. However, from Table II, we find that window size has a very limited effect on prediction accuracy. The standard deviations of prediction accuracies for different window sizes are all less than 1%. Therefore, we selected 15 for the window size for the following computer simulations.

Prediction by Multiple Sequence Alignment With Different Cut-Off Thresholds

The buried or exposed state of a residue is always defined according to different cut-off thresholds. As shown in Figure 1, thresholds from 3 to 40% have been used to explore the prediction accuracy. With the threshold set at 15%, which can approximately split the database evenly, the prediction accuracy is 70.1%. This percentage of accuracy can properly reflect the prediction capability for single-sequence input, because other cut-off thresholds may artificially elevate the percentage of accuracy resulting from an uneven splitting of the database.[4]

It was well known that including the evolutionary information could improve the prediction accuracy of secondary structure or solvent accessibility by about 3–5%.[3,11] The prediction accuracies are listed in Table III. With the same threshold, the results are compared according to different input information: single sequences and multiple sequences. Large improvements can be obtained when cut-off thresholds of 15 and 20% are chosen and the correlation coefficient between predicted and observed results are 0.47 and 0.46, respectively.

Residues can be classified into three states by two cut-off thresholds (e.g., 4 and 16%). We set up three SVMs and each of them represents an accessibility state. The training procedure for each SVM is the same as that for the buried/exposed problem, assigning the samples of the state as positive and samples of other states as negative. Every residue is run against each SVM and it is predicted to the state with maximum SVM output value. The

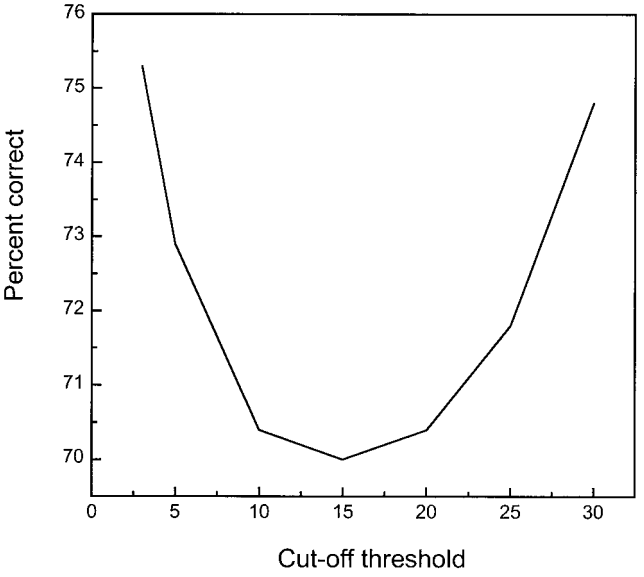

Fig. 1. Prediction accuracies versus cut-off thresholds. The cut-off threshold set at 15% equally split the dataset (an equal number of exposed and buried residues) and gave a prediction accuracy of 70.1%.

TABLE III. Prediction Accuracies From Different Input Information (Single Sequences and Multiple Sequences)

	Threshold (%)			
Input information	15	20	25	30
Single sequences	70.1	70.4	71.8	74.8
Multiple sequences	73.9	73.8	74.6	76.2

TABLE IV. Prediction Accuracies for Three and Four States of Solvent Accessibility Using Support Vector Machines

		Accuracy (%)	
States	Threshold (%)	Single sequences	Multiple sequences
Three states	4; 36	54.4 ± 0.8	58.1 ± 0.9
	9; 16	66.3 ± 0.9	68.4 ± 1.4
	9; 36	55.2 ± 1.0	57.1 ± 1.3
Four states	9; 16; 36	50.8 ± 1.0	52.5 ± 1.3

prediction results for three or four states problems with one-versus-others scheme are listed in Table IV. Prediction accuracies were averaged over 10 groups and expressed as the mean ± standard deviation. Multiple alignment sequence input can gain 2–4% prediction performance improvement.

Comparison With Other Methods

Because we used a different database (421 protein domains) from those used by other authors, a direct comparison is not valid. To compare the methods on the same dataset, the dataset previously used by Rost and Sander[3] to train and test neural networks, is also applied here to train and test the SVM method. The 126 proteins were equally divided into seven groups for cross-validation

tests, and 16% was selected as the cut-off threshold. Under these conditions, the SVM method achieved an accuracy of 72.6% (standard deviation 2%), regardless of the selection of kernel functions. On the same dataset, the accuracy of the neural network approach was 70.0% with sevenfold cross-validation,[3] whereas the Bayesian method[4] was 71.1% by the Jack-knife test. On the same dataset, the SVM method gave the best prediction result compared with other different methods. As shown above, when SVM is applied to a large dataset (421 domains), it can only reached an accuracy of 70.1% (threshold 15%). One reason for this is that a smaller dataset often results in larger fluctuations of prediction results. This was also reflected by the larger standard deviation of 2%, in contrast with those in a larger dataset, which were always less than 1%. The multiple linear method[8] achieved an accuracy of 71.5% based on a different database and the Jack-knife test, when a threshold of 20% was chosen. It is worth noting that the information theory method[9] achieved higher prediction accuracies when it used a different definition of surface residues, even based on the same database. The authors of this method also found that, when their method was applied to DSSP-defined 215 proteins, the accuracies were less than 70.0% regardless of various thresholds. However, these methods can achieve comparable results because a 1–2% accuracy difference may be attributable to the different databases and test procedures these methods used.

As for the 126-protein dataset, when three states of solvent accessibility were defined by thresholds 9 and 36%, the prediction accuracies for neural networks and Bayesian classification were 52.4 and 54.2%, respectively. With the same dataset and same definition of states, SVM achieved an accuracy of 52.8% ($\pm$1.8%). But when applied to the large dataset, the accuracy was 55.2% ($\pm$1.0%). Because we used a different dataset from the information theory method, an astrict comparison shows that, using DSSP definition of the solvent accessibility, SVM can obtain better prediction performance. For example, when four states were adopted, the SVM method achieved an accuracy of 50.8%, whereas information theory obtained 39.3%.

The SVM approach has been shown to be better than other machine learning methods on some problems,[18,23,24] and the application of SVM on this problem can also achieve better results or at least comparable prediction accuracies with other methods. A significant lack of improvement by the SVM method in some parts of the problem is attributable to its complexity. The local sequence information is not accurate enough for determining the protein solvent accessibility. However, the SVM approach can be combined with other methods for the problem. When complicated features, such as the structure and long-distance residue correlation are considered, the SVM approach may be a suitable method for further application because of its advantages—computational efficiency, data adaptability, easy representation, etc.[17]

It is obvious that the accurate prediction of protein solvent accessibility is helpful to determining protein structure and function. Secondary structure has been regarded as an important feature for recognition of protein folds and identification of distantly related protein sequences.[25–30] The predicted secondary structure can assist in the identification of remote homologs in the absence of clear sequence homology. Protein solvent accessibility can also be a factor to consider for prediction of protein function.[29] Furthermore, there is a strong relationship between secondary structure and its environment. Because solvent accessibility has an important role in determining protein secondary structure,[31] proper consideration of it makes the prediction of secondary structure more effective. Accurate prediction of the surface residues and definition of the surface residue patches are still the basis for reliable prediction of protein functional sites.[32] Therefore, future work will focus on improving the prediction accuracy and using the predicted results to enhance the methods for predicting protein structure and function.

CONCLUSION

In this report, we have applied the SVM approach to predicting protein solvent accessibility. The prediction accuracies were averaged over 10-fold cross-validation results, and their standard deviations are approximately 1% based on the large non-redundant database. We found that the selection of different kernel functions only led to a marginal difference to the prediction accuracies, and that the window size has only minor impact on the prediction accuracy. Because the goal of this report was to provide a new approach for protein solvent accessibility prediction, the results suggest that SVM is a successful one. However, to further improve the accuracy with only local information (e.g., using the sliding window technique) is a difficult task, because protein solvent accessibility is somewhat determined by information from the whole sequence and even the structure. The SVM approach can be selected as a method to combine with other methods for this problem.

ACKNOWLEDGMENTS

The authors thank Dr. Kevin Gates of the Department of Mathematics, The University of Queensland, for helpful discussions. We are grateful to one anonymous referee for the advice on enhancing the presentation of this article.

REFERENCES

1. Andrade MA, O'Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. J Mol Biol 1998;276:517–525.
2. Holbrook SR, Muskal SM, Kim SH. Predicting surface exposure of amino acids from protein sequence. Protein Eng 1990;3:659–665.
3. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
4. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
5. Pascarella S, De Persio R, Bossa F, Argos P. Easy method to predict solvent accessibility from multiple protein sequence alignments. Proteins 1998;32:190–199.
6. Mucchielli-Giorgi MH, Hazout S, Tuffery P. PredAcc: prediction of solvent accessibility. Bioinformatics 1999;15:176–177.
7. Carugo O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. Protein Eng 2000;13:607–609.

8. Li X, Pan X-M. New methods for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
9. Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AAM. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459.
10. Zhang X, Mesirov JP, Waltz DL. Hybrid system for protein secondary structure prediction. J Mol Biol 1992;225:1049–1063.
11. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34:508–519.
12. Vapnik V. The nature of statistical learning theory. New York: Spring-Verlag; 1995.
13. Cortes C, Vapnik V. Support vector networks. Mach Learn 1995;20:273–297.
14. Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. Bioinformatics 2000;16:906–914.
15. Chow ML, Moler EJ, Mian IS. Identifying marker genes in transcription profiling data using a mixture of feature relevance experts. Physiol Genomics 2001;5:99–111.
16. Ding CHQ, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. Bioinformatics 2001;17:349–358.
17. Bock JR, Gough DA. Predicting protein–protein interaction from primary structure. Bioinformatics 2001;17:455–460.
18. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 2001;308:397–407.
19. Ramaswamy S, Tamayo P, Rifkin R, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 2001;98:15149–15154.
20. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2557–2637.
21. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
22. 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. p 42–56.
23. Dreiseitl S, Ohno-Machado L, Kittler H, Vinterbo S, Billhardt H, Binder M. A comparison of machine learning methods for the diagnosis of pigmented skin lesions. J Biomed Inform 2001;34:28–36.
24. Hua S, Sun Z. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 2001;17:721–728.
25. Rice DW, Eisenberg D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J Mol Biol 1997;267:1026–1038.
26. Aurora R, Rose GD. Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparisons. Proc Natl Acad Sci USA 1998;95:2818–2823.
27. Jones DT, Tress M, Bryson K, Hadley C. Successful recognition of protein folds using threading methods biased by sequence similarity and predicted secondary structure. Proteins 1999;3:104–111.
28. Hargbo J, Elofsson A. Hidden Markov models that use predicted secondary structures for fold recognition. Proteins 1999;36:68–76.
29. Kelley LA, MacCallum RM, Sternberg MJE. Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 2000;299:499–520.
30. Geourjon C, Combet C, Blanchet C, Deléage G. Identification of related proteins with weak sequence identity using secondary structure information. Protein Sci 2001;10:788–797.
31. Macdonald JR Jr, Johnson WC. Environmental features are important in determining protein secondary structure. Protein Sci 2001;10:1172–1177.
32. Lichtarge O, Yamamoto KO, Cohen FE. Identification of functional surfaces of zinc binding domains of intracellular receptors. J Mol Biol 1997;274:325–337.