Technical Report

Department of Computer Science
and Engineering
University of Minnesota
4-192 EECS Building
200 Union Street SE
Minneapolis, MN 55455-0159 USA

TR 03-007

Prediction of Protein Relative Solvent Accessibility with Support
Vector Machines and Long-range Interaction 3D Local Descriptor

Hyunsoo Kim and Haesun Park

January 31, 2003

Prediction of Protein Relative Solvent Accessibility with Support Vector Machines and Long-range Interaction 3D Local Descriptor

Hyunsoo Kim and Haesun Park *

Department of Computer Science and Engineering,
University of Minnesota,
200 Union Street S.E., 4-192 EE/CS Building,
Minneapolis, MN 55455, U.S.A.

January, 2003

corresponding author:

Haesun Park

E-mail: hpark@cs.umn.edu

Phone : 612-625-0041(Office), 612-625-4002(Dept. Office)

Fax : 612-625-0572

Keywords: protein structure prediction; solvent accessibility; support vector machines; PSSM; directed acyclic graph scheme; long range interaction

---

*This work was supported in part by the National Science Foundation grant CCR-0204109. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

ABSTRACT

The prediction of protein relative solvent accessibility gives us helpful information for the prediction of tertiary structure of a protein. The SVMpsi method which uses support vector machines (SVMs) and the position specific scoring matrix (PSSM) generated from PSI-BLAST has been applied to achieve better prediction accuracy of the relative solvent accessibility. We have introduced a three dimensional local descriptor which contains information about the expected remote contacts by the long-range interaction matrix as well as neighbor sequences. Moreover, we applied feature weights to kernels in support vector machines in order to consider the degree of significance that depends on the distance from the specific amino acid. Relative solvent accessibility based on a two state-model, for 25%, 16%, 5%, and 0% accessibility are predicted at 78.7%, 80.7%, 82.4%, and 87.4% accuracy respectively. Three state prediction results provide a 64.5% accuracy with 9%;36% threshold. The support vector machine approach has successfully been applied for solvent accessibility prediction by considering long-range interaction and handling unbalanced data.

INTRODUCTION

The task of predicting protein structure from the sequence is important since the function of a protein is closely related to its structure, which is difficult to determined experimentally. There are largely two classes of methods in protein structure prediction. The first class includes threading and comparative modelling, which rely on a prior knowledge of similarity among sequence and known structures. The second class, called de novo or ab-initio methods, predicts the protein structure from sequence alone without relying on the similarity or known structure. Currently it is difficult to get high resolution 3D structure from ab-initio methods for studying the docking of macro-molecules, predicting protein-partner, designing and improving ligands, and protein-protein interaction.¹

For the knowledge based model, the protein secondary structure prediction^{2, 3, 4, 5, 6, 7, 8} has been studied as an intermediate step to predict tertiary structure of proteins, especially in the case that the sequence similarity is lower than 30% because the secondary structure is more conserved than the protein sequence. The solvent accessibility predication has also been studied based on the neural network approach^{9, 10, 11, 12, 13} with training by the conjugate gradient descent algorithm, i.e. back-propagation, Bayesian method¹⁴, or information theory¹⁵.

Though the prediction of solvent accessibility is less accurate than that of secondary structure from the homology approach since it is less conserved than the secondary structure,¹⁰ there has been much effort spent to improve the prediction accuracy to obtain important information regarding a buried or exposed residue for constructing tertiary structure from the sequence. For example, the prediction of the secondary structure and solvent accessibility can be aligned to known 3D structure to detect putative remote homologues for threading . The predictions can also be used as additional constraints in ab-initio methods.

In this work, we have introduced three dimensional local descriptor and used the SVMpsi⁸ method, including feature weights, to improve prediction of relative solvent accessibility. We applied a directed acyclic graph (DAG) scheme¹⁶ for the 3-class classification problem to avoid one-versus-rest classification, which has higher complexity than one-versus-one classification.

METHODS

Relative Solvent Accessibility

Amino acid solvent accessibility is the degree to which a residue in a protein is accessible to a solvent molecule. The relative solvent accessibility can be calculated from

$$RelAcc_i = 100 * Acc_i / MaxAcc_i \quad (1)$$

where Acc_i for i th residue is the solvent accessibility (given in Å) calculated from coordinates by the DSSP program¹⁷. The number of water molecules around residue can be approximated by

$Acc_i/10$. The $MaxAcc_i$ is the maximum accessibility for i th residue.¹⁰ The $MaxAcc_i$ were given for ambiguous (B, Z) or undetermined (X) residue as well as 20 normal amino acids.

We used two kinds of class definitions: (1) buried (B) and exposed (E), (2) buried (B), intermediate (I), and exposed (E). For two-state definitions, we chose the threshold of the relative solvent accessibility 25%, 16%, 5%, 0%. For the three-state (buried, intermediate, exposed) description of relative accessibility, we selected the same thresholds as Rost and Saner¹⁰.

$$\begin{aligned} \text{Buried(B):} & \quad RelAcc < 9\% \\ \text{Intermediate(I):} & \quad 9 \leq RelAcc < 36\% \\ \text{Exposed(E):} & \quad RelAcc \geq 36\% \end{aligned}$$

The thresholds for the two-state definition and the three-state definition were chosen to compare with previous work and to find the dependency of thresholds for the prediction accuracy of the relative solvent accessibility.

3D Local Descriptor Coding Scheme

A local descriptor which represents the local environment of sequences by sliding window coding scheme^{18, 2} can be enhanced by embedding the long-range interaction in order to consider the three dimensional local environment. The three dimensional local descriptor represents the environment of the specific residue not only in the sequence, but also in the three dimensional space.

There are four kinds of significant driving forces that cause remote residues to contact. The first is the disulfide covalent bond which makes nearest neighbor contacts. The most predominant linkage by disulfide bonds among the secondary structural elements is the coil-coil linkage.¹⁹ The structure of coil region is relatively important because functionally important residues which are involved in a key protein-protein interaction usually lie on the coil region. The second is the salt bridge. The oppositely charged residues between (Asp, Glu) and (Lys, Arg) tends to form a salt bridge. The third is the hydrophobic interactions among (Phe, Ile, Leu, Val), especially homopairs between them are the most favorable hydrophobic interactions. The fourth is remote hydrogen bond formation, which frequently appear since it is a major force that forms the beta sheet.

The most probable remote contact sequence block in the whole sequence with respect to the current local environment can be found by the long-range interaction matrix²⁰. The matrix represents relative frequencies of long range interaction for each amino acid pair. It was obtained from statistical analysis of the accumulation of long range interactions that two residues are separated by at least 10 residues in the sequence and at least one of their atomic distances is less than the sum of the van der Waals radii of the two atoms plus 1.0Å.²⁰ The remote contact expectation score between given fragment o and expected remote contact fragment e is

$$E_r = \sum_{i=1}^w P(o(i), e(i)) \quad (2)$$

where $o(i)$ is the i th amino acid in the window fragment, $e(i)$ is i th amino acid in the candidate fragment, w is the window size, and $P(a, b)$ is the matrix component of the relative frequency of long-range interactions between two amino acids a and b . The sequences that stabilize proteins by building remote contacts tend to be more buried than the average accessibility of the rest of the sequences. However, quite high accessibility can also be found in the stabilization sequences since the remote contact can be driven by a salt bridge between high polar residues. In the folded protein structure, the hydrophilic side chains tend to contact polar solvent, but the hydrophobic side-chains tend to minimize the contact with the polar solvent.²¹ The weighted hydrophobicity for current the window can be written as:

$$H_c = \sum_{i=1}^w h(i) \exp(-|i - ((w + 1)/2)|^2/100), \quad (3)$$

where $h(i)$ is the hydrophobicity for i th amino acid. After identifying the most probable remote contact w sequences, we can also calculate a weighted hydrophobicity for remote contact H_r .

The final position-specific scoring matrix (PSSM) from PSI-BLAST²² against SWISS-PROT database²³ (after three iterations) is used as an input to SVMs. The matrix has $20 \times M$ elements, where M is the length of the target sequence, and each element represents the log-likelihood of that particular residue substitution at that position in the template. The final position-specific scoring matrix from PSI-BLAST against the SWALL²³ non-redundant protein sequence database is used. We applied PFILT^{25, 26} to mask out regions of low complexity sequences, the coiled coil region, and transmembrane spans. For PSI-BLAST, the E-value threshold for inclusion of 0.001 and three iterations were applied to search the non-redundant sequence database. The profile matrix elements in the range $[-7, 7]$ are scaled to the $[0, 1]$ range.

There are 20 units for each residue from the PSSM. In order to allow a window to extend over the N-terminus and the C-terminus, an additional 21st unit (spacer) was added for each residue. Then, each input vector has $21 \times w$ components, where w is a sliding window size.

The H_c , E_r , and H_r values are appended to the original feature vector to build three dimensional local descriptors. Therefore, each input vector size will be $21 \times w + 3 \times w$. If the expected remote contact is not found because the expectation score is smaller than the threshold ($E_t = 1.2 \times w$), E_r and H_r are filled with zeros. The window is shifted residue by residue through the protein chain.

PHDacc¹⁰ consists of two different networks with window size 9 and 13 consecutive residues for jury decision. In Jnet⁶, a neural network with a sliding window of 17 residues for the first input and 19 for the second input were built. Both NETASA²⁷ and Manesh *et al.*'s method¹⁵ based on information theory used a window size of 17. We built input vectors considering consecutive 15 residues for predicting the central 8th residue after finding the optimal window length (See Table 2).

Support Vector Machines

There are many problems that the different classes cannot be linearly separable in the original

feature space. A support vector machine finds a nonlinear decision function in the input space by mapping the data into a linear separable higher dimensional feature space and separating it there by means of maximizing the geometric margin and minimizing the training error at the same time. The primal optimization problem can be written as:

$$\begin{aligned} \min_{\mathbf{w}, \xi_i} \quad & \frac{1}{2} \mathbf{w}^t \mathbf{w} + C \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & y_i [\mathbf{w}^t \mathbf{x}_i + b] \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad i = 1, \dots, n, \end{aligned} \quad (4)$$

where $\mathbf{x}_i$ represents an input vector, $y_i = \pm 1$ according to whether $\mathbf{x}_i$ is in the positive or negative class, n is the number of the training data, and C is a parameter that controls the trade-off between margin and classification error represented by slack variables ξ_i 's. The corresponding dual quadratic programming problem with an incorporation of a kernel function can be written as :

$$\begin{aligned} \max_{\alpha_i} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1, \dots, n. \end{aligned} \quad (5)$$

where α_i represents the influence of single i th training example limited by C and $K(\mathbf{x}_i, \mathbf{x}_j)$ is a kernel function to handle the non-linear separable case.

The support vector machines (SVMs) avoid overfitting and must find the unique minimum of a convex function for training given data set.^{28, 29, 30, 31} The decision boundary is represented sparsely as a linear combination of the training set examples.^{32, 33} Recently, SVMs have also been shown to perform well in multiple areas of biological analysis including protein secondary structure prediction,^{7, 8} protein subcellular localization prediction,³⁴ multi-class protein fold recognition,^{35, 36, 37} gene function from microarray expression data,³⁸ cancer tissue classification from microarray expression data,^{39, 40} gene selection for cancer classification,⁴¹ and protein-protein interaction problems.⁴² Also, SVMs are well suited to solve pattern recognition problems, for example, isolated handwritten digit recognition,⁴³ 3D object recognition,^{44, 45} speaker identification,⁴⁶ face detection,⁴⁷ and text categorization.^{48, 49}

Kernel Feature Weight Scheme

The solvent accessibility for a specific amino acid can be determined by the three dimensional local environment. We assumed that the contribution can be different since the closer amino acid may affect more solvent accessibility in the local environment. We scaled the feature values and could derive the modified kernel since the dual form of the SVMs only contains the input vectors in the kernel function.

$$K_m(\mathbf{x}_i, \mathbf{x}_j) = K(W\mathbf{x}_i, W\mathbf{x}_j) \quad (6)$$

where W is a diagonal matrix which contains weight factors. However, we scaled all input vectors ($\mathbf{x}' = W\mathbf{x}$) to avoid matrix vector multiplications whenever the kernel function is calculated. The 20 numerical values that are row elements of the PSSM for an amino acid were scaled by $\exp(-z^2/100) + 1.0$, where z is the sequential distance between the specific amino acid and the amino acid at the window center. The diagonal elements of W for the appended part for the remote contact residues were set to 1. The scale function was designed to cover the range of integers $z \in [-7, 7]$ for optimal window length 15.

Parameter Optimization for SVMs

When using SVMs, we need to select a kernel function and the regularization parameter C , and to construct tertiary classifiers using simple binary classifiers. From previous work, it was found that the Gaussian RBF (Radial Basis Function) kernel is appropriate for complex classification problems.

$$K(\mathbf{x}'_i, \mathbf{x}'_j) = \exp(-\gamma \|\mathbf{x}'_i - \mathbf{x}'_j\|^2) \quad (7)$$

The optimal separating hyperplane can be represented by the support vectors of which α is nonzero. Each support vector contributes one local Gaussian function centered at that data point.

The parameter γ and C can be selected from the optimization process. The optimized parameters are $\gamma = 0.01$, $C = 1.0$ for the 2-state model, and $\gamma = 0.01$, $C = 1.5$ for the 3-state model. We also tested linear kernels and polynomial kernels with different degrees d and $a = b = 1$.

$$K(\mathbf{x}'_i, \mathbf{x}'_j) = \mathbf{x}'_i \cdot \mathbf{x}'_j \quad (8)$$

$$K(\mathbf{x}'_i, \mathbf{x}'_j) = [a \mathbf{x}'_i \cdot \mathbf{x}'_j + b]^d \quad (9)$$

Table 1 shows that the RBF kernel produces the most accurate prediction for the solvent accessibility. Though the even degrees of polynomial kernels were also investigated, there was no special behavior compared with odd degrees. The value of the regulation parameter C for linear kernel and polynomial kernel were optimized by numerical experiments.

Handling Unbalanced Data

In binary classification problems, if the number of samples of one class is much larger than that of the other classes, the decision boundary tends to be determined to make more correct decision for the larger class in order to maximize total accuracy. For handling this unbalanced data, there are three kinds of approaches. The first is the discard approach that discards training points of

the larger size class to balance the number of training points of each class. Though this approach reduces the number of points to gain balance and lower complexity, it may eliminate points that contain critical information for classification. The second is the duplicate approach that duplicates the training points of the smaller size to achieve balance. The third method is the different penalty parameters method which uses different penalty parameters in the SVM formulation ³¹:

$$\begin{aligned}
\max_{\alpha_i} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\
s.t. \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq C_+, \text{ if } y_i = 1 \\
& 0 \leq \alpha_i \leq C_-, \text{ if } y_i = -1.
\end{aligned} \tag{10}$$

Using different penalty parameters (C_+ and C_-), we can resolve the situation that the recall value of the smaller class is too small to produce good prediction accuracy.

We used the third method to treat unbalanced data for most classifications. The duplicate method was used for fully buried residue classification since we could not control well when the difference of the number of classes was too large. The first method was adopted for the KP480 data set since there was not enough memory to store the whole KP480 sliding window data points.

Training and Testing Data Sets

We used 3 different data sets. The first data set (HMK24) consists of 19 training sequences and 5 test sequences. The training set contains the sequences of 1bp2, 1cpv, 1ctf, 1gcr, 1lz1, 1mbd, 1pcy, 1rn3, 1tpp, 2act, 2alp, 2apr, 2sga, 3dfr, 3tln, 4fxn, 451c, 5cpa, and 9pap. The test set contains the sequences of 1nxb, 1ubq, 2cpp, 2prk, and 2sns. The buried residues are defined as those with < 20% of relative solvent accessibility for the 2 state model and buried (0-5%), intermediate (5-40%), or exposed (> 40%) for the 3 state model to compare with previous work.⁹

The second data set (RS126) contains 126 proteins with less than 25% pairwise sequence identity. It has been used to study conservation and prediction of solvent accessibility in protein families.¹⁰

The third data set (KP480) were prepared from CB513 by removing proteins that were shorter than 30 residues and those from the result of PSI-BLAST that contained only a few sequences in the first iteration.⁸

Three dimensional coordinates of proteins were obtained from the PDB databank⁵⁰ and the solvent accessibility was calculated with the DSSP program of Kalbsch and Sander¹⁷. The relative solvent accessibility was calculated by Eqn. (1). The group (buried (B), exposed (E): buried (B), intermediate (I), exposed (E)) is determined by the corresponding thresholds. The second and third data sets are divided into 7 folds that have a similar number of proteins for cross validation tests. The first data set is studied without cross-validation testing.

Final Prediction

We obtained one-versus-one classifiers (E/B) for 2 state (exposed/buried) relative solvent accessibility, and one-versus-rest classifiers (B/ $\sim$ B, I/ $\sim$ I, E/ $\sim$ E) and three one-versus-one classifiers (B/I, E/B, and I/E) for 3 state relative solvent accessibility from SVMs. We adopted a directed acyclic graph (DAG) scheme¹⁶ for which prediction results were as good as the jury results that used all 6 classifiers. The jury decision scheme suffers from handling unbalanced data and larger data points for one-versus-rest classifiers.⁸

If the test point is not buried ($\sim$ B) from E/B one-versus-one classifier, I/E classifier is applied, while if the point is not exposed ($\sim$ E) from E/B classifier, B/I classifier is applied to tell if it is buried or intermediate. Three different kinds of DAG schemes can be constructed, such as DAG1 (starts from B/I), DAG2 (starts from E/B), DAG3 (starts from I/E). We observed that the prediction results were almost the same in all cases.

RESULTS AND DISCUSSION

Holbrook *et al.*⁹ achieved 72.0% correct overall prediction for the testing set in the binary model at window size=11 and 54.0% prediction of solvent accessibility in the ternary model at window size=7 using 10 hidden nodes. We obtained 78.7% in the 2-state model and 62.4% in the 3-state model using window size=15 for the first data set. The improvement is 6.7% for the 2-state model and 8.4% for the 3-state model.

Manesh *et al.*¹⁵ reported 70.0% prediction accuracy for the 2-state model with threshold 9, and 58.1% for the 3-state model with thresholds 9;16 using information theory with a set of 215 protein sequences that is used by NETASA²⁷. We only discuss these values for fair comparison, although they also reported other results by their own method to calculate accessible surface area (ASA) instead of DSSP. PHDacc¹⁰ showed 86% of the completely buried sites were correctly predicted as having 0% relative accessibility. Jnet⁶ reported 75.0% prediction accuracy when the relative solvent accessibility threshold between buried and exposed is 25%, and 86.6% for fully buried residues using PSI-BLAST²² profiles. We cannot directly compare between Jnet and SVMpsi since Jnet used the CB480 data set which is slightly different from the KP480 data set. Recently, G.Pollastri *et al.*¹³ achieved 77.2% for the two classes by BRNN (Bidirectional Recurrent Neural Network) as well as PSI-BLAST profiles. They claimed that the improvement is due both to the larger training set and the BRNN architecture, which can capture long-range interactions.

Table 3 shows that the methods using PSI-BLAST (Jnet⁶, BRNN¹³, and SVMpsi⁸) are much better than the other methods. The results from the SVMpsi method were comparable or better than the other methods to predict relative solvent accessibility. Though the direct comparison of our results with BRNN is not possible due both to different training sets and different training sets size, the BRNN and SVMpsi, considering long-range interactions, produce relatively good prediction results.

We already showed that the SVMpsi method can achieve good performance for protein secondary

structure prediction by cross-validation tests and the fifth CASP (Critical Assessment of Structure Prediction) experiment in previous work.⁸ This was the first use of the support vector machines approach to predict relative solvent accessibility using novel three dimensional local descriptors that contain information of hydrophobicity for the current window, the possibility of remote contact, and hydrophobicity of the expected remote contact window. Though the profiles are just slightly modified by statistical knowledge for the remote contact in a protein sequence, the method achieves a very high degree of prediction accuracy. While the protein secondary structure tends to be determined by local sequence environment, the solvent accessibility is much more related to the tertiary interactions between residues far apart in the sequence, but close in three dimensional space.

ACKNOWLEDGEMENTS

The authors would like to thank University of Minnesota Supercomputing Institute(MSI) for intensive numerical computing. We would also thank Prof. Thorsten Joachims for making SVMlight software so widely available, James A. Cuff & Prof. Geoffrey J. Barton for providing the data set, and Prof. David T. Jones for PFILT software and his kind help.

REFERENCES

1. Baker D, Sali A. Protein structure prediction and structural genomics. *Science* 2001;294:93–96.
2. Rost B, Sander C. Prediction of secondary structure at better than 70% accuracy. *J Mol Biol* 1993;232:584–599.
3. King RD, Sternberg MJE. Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci* 1996;5:2298–2310.
4. Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 1997;27:329–335.
5. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202.
6. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000;40:502–511.
7. Hua SJ, Sun ZR. A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach. *J Mol Biol* 2001;308:397–407.
8. Kim H, Park H. Protein secondary structure prediction based on an improved support vector machines approach. 2003. submitted to *Protein Eng*.
9. Holbrook SR, Muskalis SM, Kim SH. Predicting surface exposure of amino acids from protein sequence. *Protein Eng* 1990;3(8):659–665.

10. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins* 1994;20:216–226.
11. Pascarella S, Persio RD, Bossa F, Argos P. Easy method to predict solvent accessibility from multiple protein sequence alignments. *Proteins* 1999;32:190–199.
12. Li X, Pan XM. New method for accurate prediction of solvent accessibility from protein sequence. *Proteins* 2001;42:1–5.
13. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* 2002;47:142–153.
14. Thompson MJ, Goldstein RA. Predicting solvent accessibility: higher accuracy using bayesian statistics and optimized residue substitution classes. *Proteins* 1996;25:38–47.
15. Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AAM. Predicting of protein surface accessibility with information theory. *Proteins* 2001;42:452–459.
16. Heiler M. Optimization Criteria and Learning Algorithms for Large Margin Classifiers: Diploma Thesis, University of Mannheim.; 2002.
17. Kabsch W, Sander C. A dictionary of protein secondary structure. *Biopolymers* 1983;22:2577–2637.
18. Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* 1988;202:865–884.
19. Fiser A, Simon I. Predicting the oxidation state of cysteins by multiple sequence alignment. *Bioinformatics* 2000;3:251–256.
20. Dosztányi Z, Fiser A, Simon I. Stabilization centers in proteins: Identification, characterization and predictions. *J Mol Biol* 1997;272:597–612.
21. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH. Hydrophobicity of amino acid residues in globular proteins. *Science* 1985;229:834–838.
22. 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.
23. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res* 2000;28:45–48.
24. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci USA* 1992;89.
25. Jones DT, Taylor WR, Thornton JM. A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry* 1994;33:3038–3049.
26. Jones DT, Swindells MB. Getting the most from PSI-BLAST. *TRENDS in Biochemical Science* 2002;27:161–164.
27. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics* 2002;18(6):819–824.

28. Burges CJC, Schölkopf B. Improving the accuracy and speed of support vector learning machines In: Mozer, M., Jordan, M., Petsche, T., editors. *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press; 1997. p 375–381.
29. Burges CJC. A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery* 1998;2:121–167.
30. Cristianini N, Shawe-Taylor J. *Support Vector Machines and other kernel-based learning methods*: Cambridge University Press; 2000.
31. Osuna E, Freund R, Girosi F. Support vector machines: Training and applications. Tech. Rep. AI Memo 1602. MIT A.I. Lab. 1997.
32. Vapnik V. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag; 1995.
33. Vapnik V. *Statistical Learning Theory*. New York: John Wiley & Sons; 1998.
34. Hua SJ, Sun ZR. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* 2001;17:721–728.
35. Cai YD, Liu XJ, Xu XB, Zhou GP. Support vector machines for predicting protein structural class. *BMC Bioinformatics* 2001;2:3.
36. Ding CHQ, Dubchak I. Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics* 2001;17:349–358.
37. Leslie C, Eskin E, Noble W. Stafford the spectrum kernel: A string kernel for SVM protein classification. 2002. To appear in *Proceedings of the Pacific Symposium on Biocomputing*.
38. Michael PSB, William NG, David L, Nello C, Charles S, Manuel AJ, David H. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc Natl Acad Sci USA* 2000;97:262–267.
39. Mukherjee S, Tamayo P, Mesirov JP, Slonim D, Verri A, Poggio T. Support vector machine classification of microarray data. Tech. Rep. AI Memo 1677. Massachusetts Institute of Technology. 1999.
40. 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(10):906–914.
41. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. *Machine Learning* 2002;46:389–422.
42. Bock JR, Gough DA. Predicting protein-protein interactions from primary structure. *Bioinformatics* 2001;17:455–460.
43. LeCun Y, Jackel LD, Bottou L, Brunot A, Cortes C, Denker JS, Drucker H, Guyon I, Muller UA, Sackinger E, Simard P, Vapnik V. Comparison of learning algorithms for handwritten digit recognition In: Fogelman, F., Gallinari, P., editors. *International Conference on Artificial Neural Networks*; 1995. p 53–60.

44. Blanz V, Scholkopf B, Bulthoff H, Burges C, Vapnik VN, Vetter T. Comparison of view-based object recognition algorithms using realistic 3d models. Proc of ICANN'96. LNCS 1996;1112:251–256.
45. Pontil M, Verri A. Support vector machines for 3D object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 1998;20(6):637–646.
46. Schmidt M. Identifying speakers with support vector networks. 1996. Proceedings of Interface '96, Sydney, 1996.
47. Osuna E, Freund R, Girosi F. Training support vector machines: an application to face detection. 1997. CVPR'97, 1997.
48. Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of the European Conference on Machine Learning. Berlin: Springer; 1998. p 137–142.
49. Lodhi H, Cristianini N, Shawe-Taylor J, Watkins C. Text classification using string kernels. Advances in Neural Information Processing Systems 2000;13:563–569.
50. 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.

kernel function	E/B ¹	B/I ²	E/B ²	I/E ²
linear	75.83	68.98	80.38	67.04
polynomial (d=2)	76.68	68.91	81.28	67.88
polynomial (d=3)	76.73	68.98	81.35	67.52
polynomial (d=4)	76.67	68.88	81.40	67.40
RBF	76.84	69.08	81.69	68.08

Table 1: Prediction accuracy with different type of Kernel functions. The sliding window size 13 was used for E/B, 15 for the others. Results for E/B¹ was obtained with the threshold of 25 in case of the 2-state model. Results for B/I², E/B², and E/I² are obtained with thresholds 9;36 in case of the 3-state model. The results are on the RS126 with PSI-BLAST profiles and a L_1 norm soft margin SVM with the RBF kernel function using the corresponding optimized γ and C parameters. Combined results of 7-fold cross validation are shown.

classifier	$l = 11$	$l = 13$	$l = 15$	$l = 17$	$l = 19$	l^*
E/B ¹	76.82	76.84	76.75	76.60	76.60	13
B/I ²	69.11	69.06	69.08	69.16	69.12	17
E/B ²	81.42	81.55	81.69	81.57	81.53	15
I/E ²	67.94	68.24	68.08	68.02	68.03	13

Table 2: Dependency of the testing accuracy on the window length for each binary classifier. Results for E/B¹ are obtained with threshold 25 in case of 2-state model. Results for B/I², E/B², and E/I² are obtained with thresholds 9;36 in case of the 3-state model. The results are on the RS126 with PSI-BLAST profile and a L_1 soft margin SVM with the RBF kernel function using the corresponding optimized γ and C parameters. The l^* value is the optimal window length for each binary classifier. Combined results of 7-fold cross validation are shown.

method	state threshold(%)				
	2-state(25%)	2-state(16%)	2-state(5%)	2-state(0%)	3-state(9%,36%)
PHDacc	-	74.8	-	86.0	57.5
SVMpsi*	76.8	77.8	79.8	86.2	59.6
Jnet*	75.0	-	79.0	86.6	-
Jnet [†]	76.2	-	79.8	86.5	-
BRNN	77.2	-	81.2	86.5	-
NETASA	70.3	-	74.6	87.9	-
SVMpsi [†]	78.7	80.7	82.4	87.4	64.5

Table 3: Accuracy of the relative solvent accessibility for different thresholds. Jnet*, BRNN, and SVMpsi methods are based on the PSI-BLAST profiles. Results for SVMpsi* are obtained for the same data set, i.e. RS126, with PHDacc. Results for SVMpsi[†] are obtained from KP480 data set. Combined results of 7-fold cross validation are shown except that the result of BRNN are obtained from 3-fold cross validation. PHDacc results are from Rost and Sander¹⁰, Jnet results are from Cuff and Barton (Jnet*: PSI-BLAST profiles, Jnet[†]: combined PSI-BLAST and HMMER2 profiles)⁶, BRNN results are from Pollastri *et al.*¹³, and NETASA results are from Ahmad and Gromiha.²⁷