PROTEINS: Structure, Function, and Bioinformatics 54:738–743 (2004)

A Novel Method for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles

Jian Guo,[1,2†] Hu Chen,[1†] Zhirong Sun[1*], and Yuanlie Lin[2]
[1]*Institute of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing, China*
[2]*Department of Mathematical Sciences, Tsinghua University, Beijing, China*

ABSTRACT **A high-performance method was developed for protein secondary structure prediction based on the dual-layer support vector machine (SVM) and position-specific scoring matrices (PSSMs). SVM is a new machine learning technology that has been successfully applied in solving problems in the field of bioinformatics. The SVM's performance is usually better than that of traditional machine learning approaches. The performance was further improved by combining PSSM profiles with the SVM analysis. The PSSMs were generated from PSI-BLAST profiles, which contain important evolution information. The final prediction results were generated from the second SVM layer output. On the CB513 data set, the three-state overall per-residue accuracy, Q3, reached 75.2%, while segment overlap (SOV) accuracy increased to 80.0%. On the CB396 data set, the Q3 of our method reached 74.0% and the SOV reached 78.1%. A web server utilizing the method has been constructed and is available at http://www.bioinfo.tsinghua.edu.cn/pmsvm. Proteins 2004;54:738–743.** © 2004 Wiley-Liss, Inc.

Key words: protein structure prediction; protein secondary structure; support vector machine; position-specific scoring matrices; PSI-BLAST

INTRODUCTION

A large number of genome sequences have been produced in high-throughput experiments. The next step is to analyze these genome and protein sequences to find new gene functions.[1] The prediction of protein structure and function from amino acid sequences is one of the most important problems in molecular biology. This problem is becoming more pressing as the number of known protein sequences is explored as a result of genome and other sequencing projects, and the protein sequence–structure gap is widening rapidly.[2,3] Therefore, computational tools to predict protein structures are badly needed to narrow the widening gap. Although the prediction of three-dimensional (3D) protein structures is the ultimate goal, the structure still cannot be accurately predicted directly from sequences. An intermediate but useful step is to predict the protein secondary structure, which provides some knowledge and simplifies the complicated 3D structure prediction problem.

The fundamental elements of the secondary structure of proteins are α-helices, β-sheets, coils, and turns. Some methods have been developed for defining various protein secondary structure elements from the atomic coordinates in the Protein Data Bank (PDB), such as DSSP,[4] STRIDE,[5] and DEFINE.[6] According to DSSP, 8 types of protein secondary structure elements were classified and denoted by letters: H (α-helix), E (extended β-strand), G (3_{10} helix), I (π-helix), B (isolated β-strand), T (turn), S (bend) and "_" (coil). The 8 classes are usually reduced to three states, helix (H), sheet (E), and coil (C) by different reduction methods.[7] Thus, the secondary structure prediction can be analyzed as a typical three-state pattern recognition or classification problem, where the secondary structure class of a given amino acid residue in a protein is predicted based on its sequence features.

Since the 1970s, many methods have been developed for predicting protein secondary structures. Early works usually relied on the single-residue statistics in various secondary structural elements, for example, the Chou–Fasman method[8] and the Garnier–Osguthorpe–Robson (GOR I) method.[9] Nearly 20 years later, a significant improvement was made in the PHD method,[10] which is a three-level neural network including some machine learning techniques. After the PHD method, many further neural networks and machine learning refinements were developed.[11–13] Several machine learning approaches have successfully predicted protein secondary structures, and prediction accuracies were further improved. In 2001, Hua and Sun[14] introduced a new method, support vector machine (SVM), which is based on statistical learning theory (SLT). The SVM method achieved good segment overlap accuracy, SOV = 76.2%, and good three-state overall per-residue accuracy, Q_3 = 73.5%.[14]

Here, we describe an improved dual-layer SVM combined with a position-specific scoring matrix (PSSM) gener-

†These two authors contributed equally in this work.

Grant sponsor: Fondational Science Research Grant of Tsinghua University (JC2001043); 863 Projects (2002AA234041); 973 Project (2003CB715903); NSFC (90303017).

*Correspondence to: Zhirong Sun, Institute of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 10084, China. E-mail: sunzhr@mail.tsinghua.edu.cn

Received 18 June 2003; Accepted 3 September 2003

ated from PSI-BLAST. The combined method, which is referred to as PMSVM, provides a good SOV of 80.0% and Q_3 of 76.2%, which is nearly 3% higher than the simple SVM method's SOV and Q_3.[14] The method is also compared with existing prediction methods. The results show that our method more effectively predicts secondary structures.

MATERIALS AND METHODS

Data Set

Two data sets are frequently used in protein secondary structure predictions to test algorithms. One is the RS126 data set, which include 126 protein chains and was developed by Rost and Sander.[10] The other data set, which is called CB513, is much larger. It was constructed by Cuff and Barton,[7] and contains 513 protein chains. Almost all sequences in the RS126 set are included in the CB513 set. Both are nonhomologous, but the homology measurement of CB513 is more strict than in the RS126 set. Removal of protein chains contained in both the RS126 set and the CB513 set gives another data set, which include 396 protein sequences and is named the CB396 set. RS126 was mostly used to develop early prediction methods with CB513 set and CB396, now widely used. The CB513 and CB396 sets were used to compare the present algorithm with other prediction method.

The Definition of Protein Secondary Structure

The automatic assignments of secondary structure to experimentally determined 3D structures are usually performed using DSSP,[4] STRIDE,[5] and DEFINE.[6] This work exclusively used the DSSP assignments, which distinguish the secondary structure into 8 categories: H (α-helix), G (3_{10} helix), I (π-helix), E (extended β-strand), B (isolated β-strand), T (turn), S (bend), and coil ("_"). The 8 structure classes were reduced into 3 classes. There are four main methods to perform the reduction process. (1) DSSP: H, G to H; E, B to E; all other states to C; (2) DSSP: H to H; E to E; all other states to C; (3) DSSP: H, G, I to H; E to E; all other states to C; and (4) DSSP: H, G to H; E to E; all other states to C.

In this article, definition (1) was adopted, because it is considered to be the strictest definition, which usually results in lower prediction accuracy than other definitions.

PSI-BLAST Profiles

This work used multiple-sequence alignment profiles generated from the PSI-BLAST[15] program for each protein chain in the CB513 and CB396 sets. First, we obtained a database, which contained all known databases: all nonredundant GenBank translations, PDB, SwissPort, PIR databank, and PRF databank. Then the low-complexity regions, transmembrane regions, and coiled-coil segments were removed from the database. A program named *pfilt* was used to remove these regions.[16] Then, encoded BLAST data bank files were generated from filtered FASTA files. Finally, the PSI-BLAST program was used to query each protein in the CB513 and CB396 sets against the filtered NR database to generate PSSM profiles. These profiles were scaled to the required 0–1 range using the standard logistic function

$$f(x) = \frac{1}{1 + \exp(-x)},$$

where x is the raw profile matrix value. These profiles were then used as the input information to the first-layer SVM.

Support Vector Machine

The SVM is a new machine learning method that developed rapidly and has been widely used in many kinds of pattern recognition problems. The basic method of SVM is to transform the samples into a high-dimension Hilbert space and to seek a separating hyperplane in this space. The separating hyperplane, which is called the optimal separating hyperplane (OSH), is chosen in such a way as to maximize its distance from the closest training samples. As a supervised machine learning technology, SVM is well-founded theoretically on statistical learning theory. SVM has been successfully applied to many fields of pattern recognition, including object recognition,[17] speaker identification,[18] and text categorization.[19] The SVM usually outperforms other machine learning technologies, including Neural Networks and K-Nearest Neighbor classifiers. In recent years, the SVM has been used in bioinformatics, including gene expression profile classification, detection of remote protein homologies and recognition of translation initiation sites. Hua and Sun[14] used a single-layer SVM to analyze protein secondary structure with excellent prediction results (in this article, this method is called the simple SVM). More details about SVM can be found in Vapnik's publications.[20,21]

Here, we describe a dual-layer SVM system used to predict secondary structure. The dual-layer SVM system combined with the PSI-BLAST profiles provides more accurate prediction than Hua and Sun's[14] simple SVM prediction system.

Coding Scheme

As with Hua and Sun's work,[14] the present analysis used the classical local coding scheme of the protein sequences with a sliding window. PSI-BLAST matrix with n rows and 20 columns can be defined for single sequence with n residues. For the first layer in the prediction system, each residue is coded as a 21-dimensional vector, where the first 20 elements of the vector are the corresponding elements in PSI-BLAST matrix. For the second layer, the vector corresponding to a residue has 4 elements, where the first 3 elements represent the 3 secondary structures (H, E, C). The last unit was added in order to allow a window to extend over the N- and the C-terminus. If the window length is l, the dimension of the feature vector is $21*l$ for the first layer and $4*l$ for the second layer.

Prediction System Structure

A dual-layer SVM structure was used in the prediction system (see Fig. 1). The first layer is an SVM classifier that classifies each residue of each sequence into the 3 secondary structure classes (H, E, or C). The one-against-rest

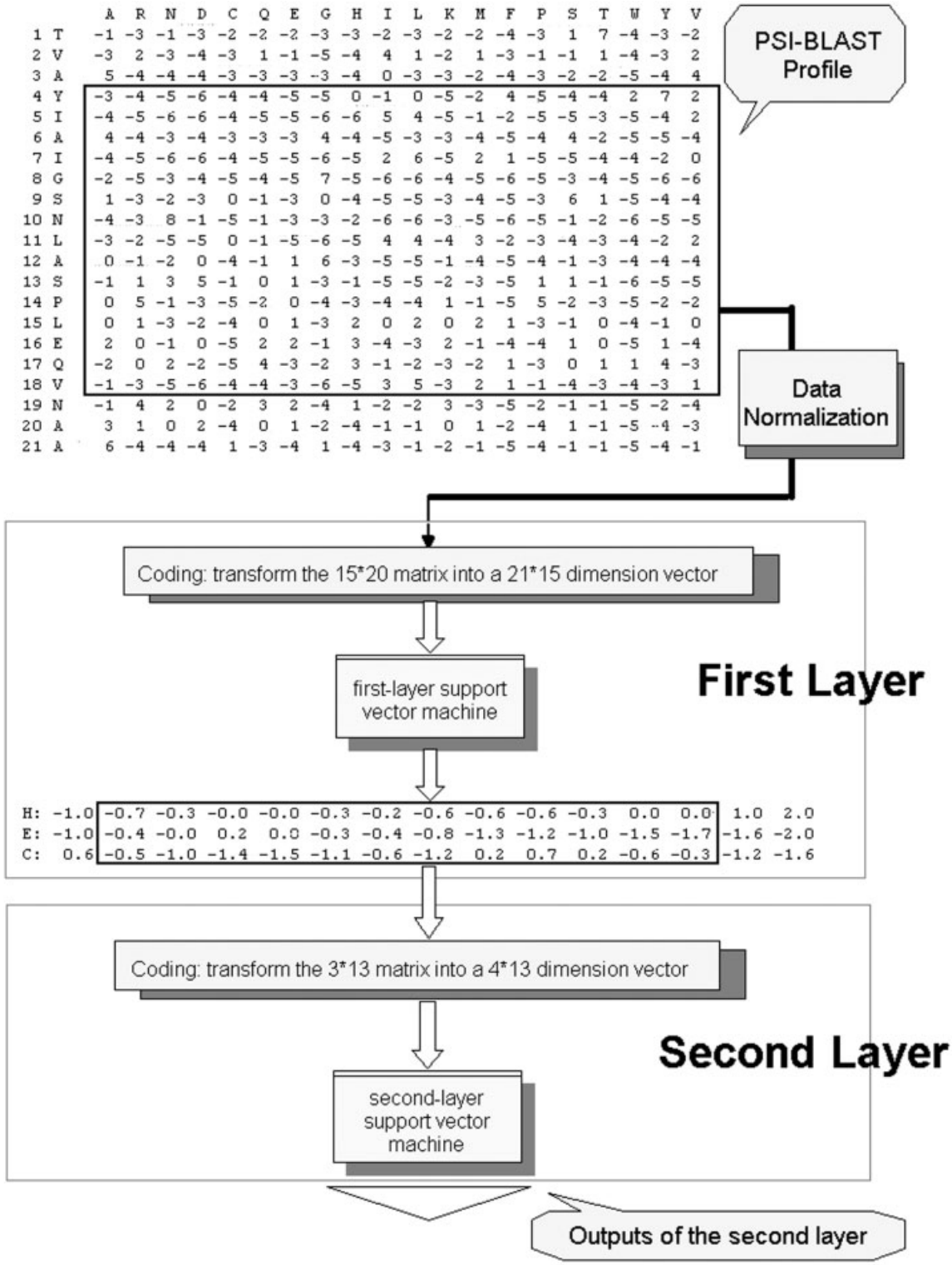

Fig. 1. The dual-layer architecture of the PMSVM system. The system include three parts: the PSI-BLAST profile, the first layer, and the second layer. The profile is tranformed into a number of 21*15 demension vectors using the slide-window method. These vectors are input into the first-layer SVM. The outputs of the first-layer SVM are a number of 3D vectors representing the probability that the residue belongs to that class. Using the slide-window method, the outputs of the first-layer SVM are tranformed into a number of 4*13 dimensional vector, which are used as the inputs of the second-layer SVM. The final decisions are based on the outputs of the second-layer SVM.

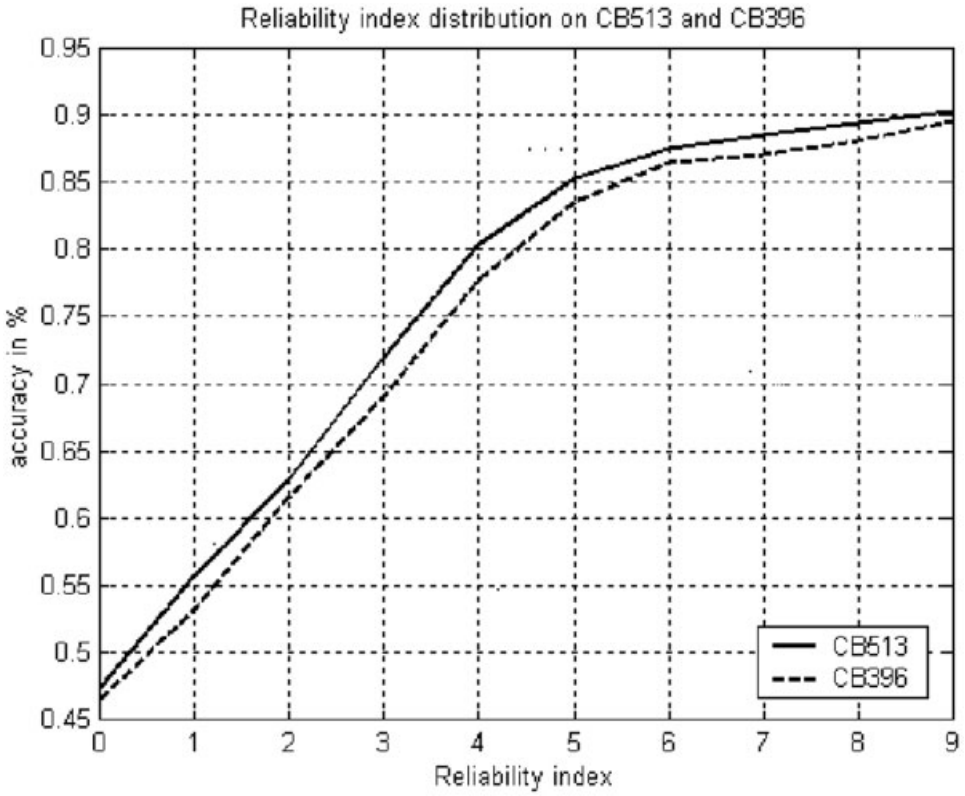

Fig. 2. The Q_3 distribution on different Reliability indices (from 0 to 9).

strategy was used for the multiclass classification, so there were three outputs for each residue. The outputs represent the probability that the residue belongs to that class. Since the consecutive patterns are correlated (e.g., a helix contains at least 4 consecutive patterns, and a sheet contains at least 3 consecutive patterns), the second-layer SVM classifier filtered successive outputs from the first layer. The target outputs of the second layer were the same as the first layer. As with the first-layer SVM, the second layer also uses the one-against-rest strategy, with each residue classified into the class with the largest output value.

Training and Testing

Seven-fold cross-validation was used on the CB396 and CB513 data sets to test the method's efficiency. The whole data set was randomly divided into 7 subsets of equal size. In each validation, one subset was used for testing while the rest was used for training. Several parameters were regulated to optimize the training. This analysis used the radial basis function (RBF) kernel in both the first- and the second-layer SVM, where γ is a parameter to be determined. The analysis used the soft-margin SVM, so the regularization parameter C also needed to be regulated. γ_1 and C_1 were defined as the gamma parameter and the regularization parameter in the first-layer SVM, while γ_2 and C_2 were defined as the gamma parameter and the regularization parameter in the second-layer SVM. For the CB513 data set, $\gamma_1 = 0.05$, $C_1 = 2.3$, $\gamma_2 = 2.5$, and $C_2 = 2.0$; for the CB396 data set, $\gamma_1 = 0.05$, $C_1 = 2.0$, $\gamma_2 = 2.4$, and $C_2 = 2.5$.

$$K(x_i, x_j) = \exp(-\gamma(x_i - x_j)^2) \quad (1)$$

Reliability Index

The prediction reliability index (RI) was used to assess the effectiveness of the approaches for the prediction of the secondary structure of a new sequence. The RI offers an excellent tool for focusing on key regions having high prediction accuracy. There are different definitions of the RI. Here, we used a definition similar to that proposed by Rost and Sander[10]: RI = INTEGER [(maximal_output(I) − second_largest_output(I)]/0.5). If the value of RI > 9, then set RI = 9, so the value of RI is an integer between 0 and 9. The distribution of the prediction accuracy with different RIs is illustrated in Figure 2. The prediction accuracy of residues with higher RI values is much better than those with lower RI values. Therefore, the definition of RI reflects the prediction reliability.

RESULTS AND DISCUSSION

Several standard performance measures were used to assess prediction accuracy. The three-state overall per-residue accuracy (Q_3), the Matthew's correlation coefficients (C_H, C_E, C_C), and the SOV were used to evaluate the accuracy.[10,22,23] The per-residue accuracies for each type of secondary structure (Q_H, Q_E, Q_C, Q_H^{pre}, Q_E^{pre}, Q_C^{pre}) were also calculated. The PMSVM method was compared with

TABLE I. Comparison with the results of the PHD, the Simple SVM and our PMSVM

Method	SOV (%)	Q_3 (%)	Q_H (%)	Q_E (%)	Q_C (%)	Q_H^{pre} (%)	Q_E^{pre} (%)	Q_C^{pre} (%)	C_H	C_E	C_C
PHD1	73.5	70.8	72	66	72	73	60	—	0.6	0.52	0.51
SVM1	74.6	71.2	73	58	75	77	66	69	0.61	0.51	0.52
PHD2	—	72.1	70	62	79	77	64	72	0.63	0.53	0.52
SVM2	76.2	73.5	75	60	79	79	67	70	0.65	0.53	0.54
PMSVM1	80.0	75.2	80.4	71.5	72.8	79.4	66.4	76.4	0.71	0.61	0.61
PMSVM2	78.1	74.0	79.3	69.3	72	79.4	66.4	73.6	0.7	0.6	0.59

PHD, SVM1: Results obtained on the RS126 set.
PHD2: Results obtained on another data set which contains 250 protein chains (Rost and Sander).[22]
SVM2: Results obtained on CB513 set.
PMSVM1: Result obtained on CB513 set. First-layer SVM parameters: $\gamma = 0.05$, $C = 2.3$, Second-layer SVM parameters: $\gamma = 2.5$, $C = 2$.
PMSVM2: Result obtained on CB396 set. First-layer SVM parameters: $\gamma = 0.05$, $C = 2.0$, Second-layer SVM parameters: $\gamma = 2.5$, $C = 2$.

Hua and Sun's simple SVM method and the famous PHD method. The results from the PMSVM method are very good. On the CB513 set, the SOV was 80.0%, nearly 4% higher than that of the simple SVM method (76.2%). The three-state per-residue accuracy Q_3 was 75.2%, which is nearly 2% higher than the simple SVM method (73.5%) and 3% higher than the PHD method. The results obtained on the CB396 set was slightly lower than the results on the CB513. Cuff and Barton[7] also found that many other methods have slightly lower accuracies with the CB396 set. More comparisons with other methods are shown in Table I.

The prediction accuracies using only the first-layer SVM have been computed. Although the value of Q_3 was nearly the same as with the dual-layer prediction method, the SOV was about 2% lower. The results reflect the fact that the second layer filters some noise from the first layer and improves the accuracies.

A web prediction system was developed using the PMSVM method and is available at http://www.bioinfo.tsinghua.edu.cn/pmsvm. This webpage was tested with several new protein sequences in the PDB with good results. The webserver was also used to predict some secondary structures of severe acute respiratory syndrome (SARS) proteins, which we hope will provide useful information to experimental biologists.

Further improvements of the prediction method will be made in future work. SVM is one of the best available machine learning methods, but it is still a passive learning method. In recent years, boosting methods and active learning methods have developed rapidly. Boosting is a general method for improving the accuracy of any given learning algorithm. The active learning method actively selects a subset of samples and trains the classification system on the subset to achieve more accurate prediction results. It is our hope that the combination of boosting or active learning with the SVM will achieve higher prediction accuracies. The second need is to further filter the noise and outliers in the prediction process. If the window length is not appropriate, or the training samples are not independent and identical, the noise-to-signal ratio will increase. The central SVM may help to reduce noise and outliers. Another idea is to use the wavelet transform method to filter the outputs of the first and second layers. Wavelet transforms filter signal noise and outliers; therefore, they should improve the prediction accuracy. The third aspect is to combine the information of other alignment profiles with the PSI-BLAST profile. Cuff and Barton's work showed that combining PSI-BLAST with HMMER2 profiles improved the predictions compared to using the PSI-BLAST profiles only. Therefore, practical strategies may be developed to fuse different information from different alignment profiles.

ACKNOWLEDGMENTS

Our thanks to J. A. Cuff and G. J. Barton for providing the CB513 data set, to D. T. Jones for providing the useful *pfilt* program, and to Thorsten Joachims for providing the SVM light program.

REFERENCES

1. Thorton JM. From genome to function. Science 2001;292:2095–2097.
2. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–48.
3. 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.
4. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577–2637.
5. Frishman D, Argos P. Knowledge-based secondary structure assignment. Proteins 1995;23:566–579.
6. Richards FM, Kundrot CE. Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. Proteins 1988;3:71–84.
7. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34:508–519.
8. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry 1974;13:211–215.
9. Garnier J, Osguthorpe DJ, Robson B. Analysis and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97–120.
10. Rost B, Sander C. Prediction of secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599. 220–223.
11. Riis SK, Krogh A. Improving prediction of protein secondary structure using structured neural networks and multiple sequence alignments. J Comput Biol 1996;3:163–183.
12. Baldi P, Brunak S, Frasconi P, Soda G, Pollastri G. Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 1999;15:937–946.
13. Chandonia JM, Karplus M. New methods for accurate prediction of protein secondary structure. Proteins 1999;35:293–306.
14. 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.
15. 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.
16. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.
17. Roobaert D, Hulle MM. View based 3D object recognition with support vector machines. In: Proceedings of the IEEE International Workshop on Neural Networks for Signal Processing IEEE Press: Wisconsin; 1999. p 77–84.
18. Schmidt M, Grish H. Speaker identification via support vector classifiers. In: Proceeding of the International Conference on Acoustics, Speech and Signal Processing. Long Beach, CA: IEEEE Press; 1996. p 105–108.
19. Drucker H, Wu D, Vapnik V. Support vector machines for spam categorization. IEEE Trans Neural Networ 1999;10:1048–1054.
20. Vapnik V. The nature of statistical learning theory. New York: Springer-Verlag; 1995.
21. Vapnik V. Statistical learning theory. New York: Wiley; 1998.
22. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. J Mol Biol 1994;235:13–26.
23. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of SOV, a segment based measure for protein secondary structure prediction assessment. Proteins 1999;34:220–223.