doi:10.1006/jmbi.2001.4580 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (2001) 308, 397–407

JMB

A Novel Method of Protein Secondary Structure Prediction with High Segment Overlap Measure: Support Vector Machine Approach

Sujun Hua and Zhirong Sun*

Institute of Bioinformatics Tsinghua University State key Laboratory of Biemembrane and Membrane Biotechnology Department of Biological Sciences and Biotechnology Beijing 100084, China

*Corresponding author

We have introduced a new method of protein secondary structure prediction which is based on the theory of support vector machine (SVM). SVM represents a new approach to supervised pattern classification which has been successfully applied to a wide range of pattern recognition problems, including object recognition, speaker identification, gene function prediction with microarray expression profile, etc. In these cases, the performance of SVM either matches or is significantly better than that of traditional machine learning approaches, including neural networks.

The first use of the SVM approach to predict protein secondary structure is described here. Unlike the previous studies, we first constructed several binary classifiers, then assembled a tertiary classifier for three secondary structure states (helix, sheet and coil) based on these binary classifiers. The SVM method achieved a good performance of segment overlap accuracy SOV = 76.2 % through sevenfold cross validation on a database of 513 non-homologous protein chains with multiple sequence alignments, which out-performs existing methods. Meanwhile three-state overall per-residue accuracy Q_3 achieved 73.5 %, which is at least comparable to existing single prediction methods. Furthermore a useful ''reliability index'' for the predictions was developed. In addition, SVM has many attractive features, including effective avoidance of overfitting, the ability to handle large feature spaces, information condensing of the given data set, etc. The SVM method is conveniently applied to many other pattern classification tasks in biology.

Keywords: protein structure prediction; protein secondary structure; support vector machine; supervised learning; the tertiary classifier

Introduction

The protein sequence-structure gap is widening rapidly. The number of known protein sequences[1] is exploding as a result of genome and other sequencing projects. The increasing number of protein sequences is much greater than the increasing number of known protein structures.[2] Therefore, computational predictive tools for protein structures are badly needed to narrow the widening gap.

Three-dimensional protein structures still can not be accurately predicted directly from sequences. An intermediate but useful step is to predict the protein secondary structure, which is a way to simplify the prediction problem by projecting the very complicated 3D structure onto one dimension, i.e. onto a string of secondary structural assignments for each residue.

In the field of machine learning, secondary structure prediction can be analyzed as a typical pattern recognition or classification problem where the class (secondary structure) of a given instance is predicted based on its sequence features. The usual goal of secondary structure prediction is to classify a pattern of adjacent residues as helix (H), sheet (E) or coil (C, the remaining part). The principle idea underlying almost all prediction methods is that

Abbreviations used: I.I.D., independent and identically distributed; MLP, multiple layer perceptron; NNs, neural networks; OSH, optimal separating hyperplane; RBF, radial basis function; RI, reliability index; SRM, structural risk minimization; SOV, segment overlap; SVM, support vector machine; SVs, support vectors.

E-mail address of the corresponding author: sunzhr@mail.tsinghua.edu.cn

0022-2836/01/020397–11 $35.00/0

the segments of consecutive residues prefer certain secondary structures.

Up to now, several machine learning approaches have successfully predicted protein secondary structures. Many different feedforward neural networks (NNs) have been used.[3-6] The basic architecture used in the early work of Qian & Sejnowski (1988) was a fully connected multiple layer perceptron (MLP) with back propagation (BP) algorithm. It achieved a performance of $Q_3 = 64.3\,\%$, with the Matthews correlation coefficients $C_H = 0.41$ for helices, $C_E = 0.31$ for sheets and $C_C = 0.41$ for coils on their data set.

The early secondary structure prediction methods using local information of a single sequence shared three major shortcomings: (i) the three-state pre-residue accuracy (Q_3) was not high (about 65 %); (ii) sheets were predicted at levels of 28-48 %, i.e. only slightly better than random; (iii) the predicted secondary structure segments were only half as long as the observed segments on average. The PHD method[4] overcame these shortcomings to a certain extend. PHD is a three-level NNs including a sequence-to-structure net (first level), a structure-to-structure net (second level) and a jury decision (third level). Some machine learning techniques such as early stopping, balanced training, etc. have been used. The prediction accuracy of PHD has been improved significantly by using evolutionary information contained in multiple sequence alignments. It achieved a performance of $Q_3 = 70.8\,\%$, $C_H = 0.60$, $C_E = 0.52$ and $C_C = 0.51$ on one non-homologous data set of 126 protein chains (the RS126 set) through cross validation and the prediction reliability index was introduced.

After PHD, further NN architectural and machine learning refinements have been used. Riss & Krogh[5] carefully designed the NN architecture to reduce the overfitting problem. For instance, adaptive encoding of the input amino acid residues by the weight-sharing technique was used to reduce the number of free parameters. In combination with multiple alignments, this method reached an overall accuracy $Q_3 = 71.3\,\%$, and correlation coefficients $C_H = 0.59$, $C_E = 0.50$ and $C_C = 0.41$ on the RS126 set. Chandonia & Karplus[7] presented an alternative cascaded NNs and a new method for decoding the outputs of the prediction network.

Prediction accuracy can be improved by combining more than one prediction method.[8,9] Recently, Cuff & Barton[9] have combined several widely used prediction methods PHD, DSC,[10] NNSSP[11] and PREDATOR.[12] Among these methods, PHD provided the most accurate predictions.

The support vector machine (SVM) method has been recently proposed by Vapnik and his co-workers[13-15] as a very effectively method for general purpose pattern recognition. Intuitively, SVM learns the boundary between samples belonging to two classes by mapping the input samples into a high dimensional space, and seeking a separating hyperplane in this space (see Figure 1). The separating hyperplane is chosen in such a way as to maximize its distance from the closest training samples (a quantity referred to as margin). The hyperplane is called the optimal separating hyperplane (OSH). SVM as a supervised machine learning technology is attractive because it has an extremely well developed learning theory, statistical learning theory.[14,15] The SVM approach is not only well-founded theoretically, but also superior in practical applications. SVM has been successfully applied to a wide range of pattern recognition problems, including isolated handwritten digit recognition,[13,16] object recognition,[17] speaker identification,[18] text categorization,[19] etc. In most of these cases, the performance of SVM either matches or is significantly better than that of traditional machine learning approaches, including NNs. SVM has a number of interesting properties, including effective avoidance of overfitting, the ability to handle large feature spaces, information condensing of the given data set, etc. A

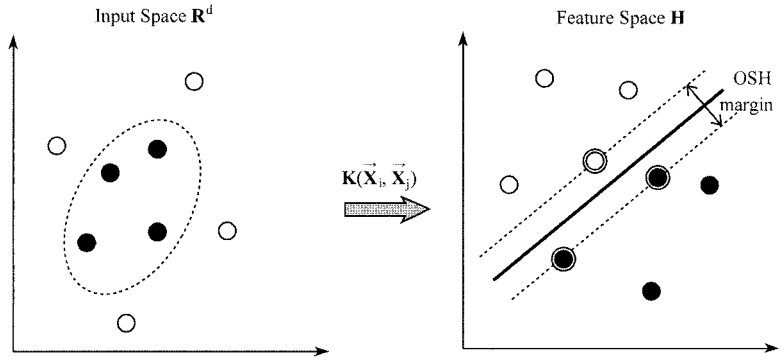

Figure 1. A separating hyperplane in the feature space corresponding to a non-linear boundary in the input space. Two classes denoted by circles and disks, respectively, are linear non-separable in the input space $\mathbf{R}^d$. SVM constructs the optimal separating hyperplane (OSH) (continuous line) which maximizes the margin between two classes by mapping the input space $\mathbf{R}^d$ into a high dimensional space, the feature space $\mathbf{H}$. The mapping is determined by a kernel function $K(\vec{x}_i, \vec{x}_j)$. Support vectors are identified with an extra circle.

brief introduction is given in the Supplementary Material.

Here, we describe the first use of the SVM approach to predict protein secondary structures. We will show that the SVM method can achieve a good performance of segment overlap measure SOV = 76.2 % which is a more realistic assessment of prediction quality meanwhile three-state overall per-residue accuracy Q_3 achieves 73.5 % which is comparable to the existing single prediction method, including PHD. It is possible to obtain combined prediction system with higher accuracy if the SVM method is combined with other methods.

Results

Parameter optimization of the prediction system

It is much simpler to construct one SVM binary classifier than NN. In the case of NN, an appropriate structure of NN is dependent on the skill of the designer. However, here we only need select a kernel function and regularization parameter C to train the SVM. The detail of the selection procedure can be seen in the Supplementary Material.

Here, we adopt the radial basis function (RBF, see equation (1)):

$$K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \exp(-r|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j|^2) \quad (1)$$

with the parameter $\gamma = 0.10$ and the regularization parameter $C = 1.50$ to construct the SVM classifiers.

Determination of the optimal window length for each binary classifier

The dependence of the testing accuracy on the size of the input residue window in the local coding scheme was tested for each binary classifier. A proper window length can lead to a good performance because a too short residue segment will lose some important classification information while a too long segment will decrease signal-to-noise ratio.

Unlike previous approaches which amount to directly constructing tertiary classifiers, we first construct binary classifiers. More details can be seen in Materials and Methods section. The optimal window length may vary for different binary classifiers (see Table 1). The optimal window length (l^*) for each binary classifier was determined on the RS126 set. The optimal window length based on the SVM method is slightly smaller than about 13, which was the optimum found in previous NN approaches.[3,4]

The results in Table 1 indicate that the optimal window length is related to the average length of the secondary structure segments. In general, longer mean secondary structure segments require larger optimal window lengths.

Table 1 also shows that the prediction accuracy of each binary classifier is not too sensitive to the window length. Using window lengths in the interval $[l^* - 2, l^* + 2]$, the variation of the prediction accuracy is quite small, less than 1.0 %.

Accuracy of each binary classifier

After determining the proper kernel function and its parameter, the regularization parameter C and the optimal window length (l^*), we constructed six SVM binary classifiers named H/∼H, E/∼E, C/∼C, H/E, E/C and C/H. Sevenfold cross validation tests were done on both the RS126 set and the CB513 set. The prediction accuracies of each binary classifier on both sets are shown in Table 2.

Table 2 shows a general increase in accuracy of between 1.63 % (E/C) and 2.66 % (H/∼H) with the larger set size from the RS126 set to the CB513 set. Our results are consistent with previous studies which have shown that an increase in the size of the database of known structures can improve the secondary structure prediction. The increase in accuracy as the set size increases may be due to the accidental addition of more easily predicted sequences to the set or better predictive patterns

Table 1. Dependence of testing accuracy on window length for each binary classifier

Binary classifier	Window length (l)							
	5	7	9	11	13	15	17	l^*
H/∼H	77.55	79.36	80.28	80.36	79.74	79.63	79.36	11
E/∼E	80.89	81.22	81.25	80.82	80.14	79.73	79.15	9
C/∼C	71.19	71.20	71.12	69.77	68.82	68.29	67.10	7
H/E	74.96	76.57	76.87	76.23	76.24	73.31	72.02	9
E/C	76.69	76.14	75.96	75.16	73.76	73.20	72.10	5
C/H	76.36	77.30	77.63	76.82	76.31	75.66	74.66	9

The l^* value is the optimal window length for each binary classifier. All the results are on the RS126 set. Combined results of sevenfold cross validation are shown.

learned by the classifiers trained on more sequences.

Table 2 shows that the accuracy of the E/∼E classifier is slightly higher than that of the H/∼H classifier, but this does not indicate that prediction of sheet is easier than that of helix. The higher accuracy of the E/∼E is mainly attributed to the much smaller content of sheet.

Accuracy of the tertiary classifier

The prediction accuracies of tertiary classifiers on both data sets are shown in Table 3 and Table 4.

Several standard performance measures were used to assess the prediction accuracy.

Table 3 and Table 4 both show that ''SVM_MAX_D'' has the best performance among the single tertiary classifiers. Method for handling multiclass cases like SVM_MAX_D are widely used in other pattern recognition problems.[20]

Studies have shown that jury techniques combining several NNs[4] or other prediction methods[8,9] can be more accurate than methods based on a single prediction. The results in Table 3 and 4 also prove this with about 0.5 % increase in accuracy through jury decision.

Comparing the results in Table 4 with the results in Table 3 also shows a general increase in accuracy as the set size increases.

Table 4 shows that the final prediction system, SVM_JURY achieves a performance of three-state overall per-residue accuracy $Q_3 = 73.5\,\%$, correlation coefficients $C_H = 0.64$ for helices, $C_E = 0.52$ for sheets, $C_C = 0.51$ for coils and Segment Overlap accuracy $SOV = 76.2\,\%$ through sevenfold cross validation on a database of 513 non-homologous protein chains, with multiple sequence alignments.

Assigning a ''reliability index'' to the prediction

When using machine learning approaches for the prediction of the secondary structure of a new sequence, it is important to know the prediction reliability. In the PHD method, the difference between the maximal and the second largest output unit has been used to derive a ''reliability index'' (RI) which is given for each residue along with the prediction.[4] In practice, the reliability index offers an excellent tool for focusing on some key regions having high prediction accuracy.

As with the PHD method, a more intuitive reliability index can be derived using the information from the outputs of the binary classifier (H/∼H, E/∼E and C/∼C). Intuitively? If a sample is predicted to have large positive distance to the OSH the sample has large probability of belonging to the positive class. For example, if a residue is predicted to be helix and the output of classifier H/∼H (the distance to the OSH) is positive and relatively large, the residue has large probability of being helix.

The reliability index is defined as:

$$RI = \begin{cases} 0 & \text{if} \quad distance(I) > 0.2 \\ INTEGER(distance(I)/0.2) & \text{if} \quad 0.2 \leqslant distance(I) > 1.8 \\ 9 & \text{if} \quad distance(I) > 1.8 \end{cases} \tag{2}$$

For a residue predicted to be in state I (I = H, E or C), *distance*(*I*) means the output of the related classifier among H/∼H, E/∼E and C/∼C, i.e. the distance of the sample to the OSH. In general, the absolute value of *distance*(*I*) is in the interval [0, 2]. RI is an integer value from 0 to 9 with $RI = 9$ corresponding to a rather reliable prediction.

The distance to the OSH is proven to supply an effective measure for prediction reliability in Figure 2, which shows that the prediction is more reliable as the distance increases. The reliability of the prediction for residues can be related to *RI*. The curve in Figure 2 answers the question how reliable the prediction for all residues labeled with the particular index is. For instance, the expected accuracy for a residue with $RI = 4$ is 80.4 % with 12 % of all residues having $RI = 4$.

Discussion

Comparison with results of other methods

An objective comparison between SVM method and most of the existing widely used methods has been made. Here, we will emphasize ''one objective comparison''. The comparison will be objective only if the results generated by different methods are based on the same data set (including the same type of alignment profiles), the same secondary structure definition (including the same reduction method) and the same accuracy assessment otherwise the comparison is unfair. A comparison between the SVM method and PHD method which is one of the most accurate and reliable secondary structure prediction methods based on NNs can be seen in Table 5. It shows that the SVM method yields good result with segment overlap measure (SOV). SOV has been recently proposed[21,22] to assess the quality of a prediction in a more realistic manner. This is done by taking into account the

Table 2. Accuracy of each binary classifier on both data sets

Binary classifier	l^*	Accuracy on the RS126 set (%)	Accuracy on the CB513 set (%)
H/∼H	11	80.36	83.02
E/∼E	9	81.25	83.39
C/∼C	7	73.20	75.52
H/E	9	80.87	83.08
E/C	5	76.69	78.32
C/H	9	77.63	79.97

The l^* value is the optimal window length for each binary classifier. Combined results of sevenfold cross validation are shown.

type and position of secondary structure segment, the natural variation of segment boundaries among families of homologous proteins and the ambiguity at the end of each segment. On the RS126 set, the SOV with the SVM method (74.6 %) is 1.1 % higher than that of PHD (73.5 %) and much higher than that of DSC (71.1 %),[10] NNSSP (72.7 %)[11] and PREDATOR (70.3 %).[12] On the CB513 set, the SOV with the SVM method achieves 76.2 % and this is the first time, to the best of our knowledge, that a single prediction method predicts SOV with such high accuracy. Meanwhile, three-state per-residue accuracy Q_3 is comparable to the existing method, including the single best predictor PHD among DSC, PHD, NNSSP and PREDATOR reported in Cuff & Barton,[9] the refined NN proposed by Riis & Krogh[5] with $Q_3 = 71.3$ % on the RS126 set, the bidirectional recurrent NNs more recently proposed by Baldi *et al.*[6] with 72.0 % on the RS126 set, etc. In a word, the comparisons show that the SVM method achieved a good performance of SOV = 76.2 % on the CB513 set which is significantly higher than existing methods and three-state overall per-residue accuracy Q_3 achieved 73.5 % which is slightly higher than the above other methods on the same data set.

In addition, DSSP,[23] the most widely used secondary structure definition method, provides an eight-state assignment of secondary structure. However, prediction methods are normally trained and assessed for only three states (H, E, C), so the eight states must be reduced to three. More detail can be seen in Material section. The comparisons above obey the rule that two prediction methods are compared only if they adopt the same eight- to three-state reduction method. A detail analysis of the effect on accuracy of applying different reduction methods can be seen in Cuff & Barton.[9] They found that prediction methods appear to

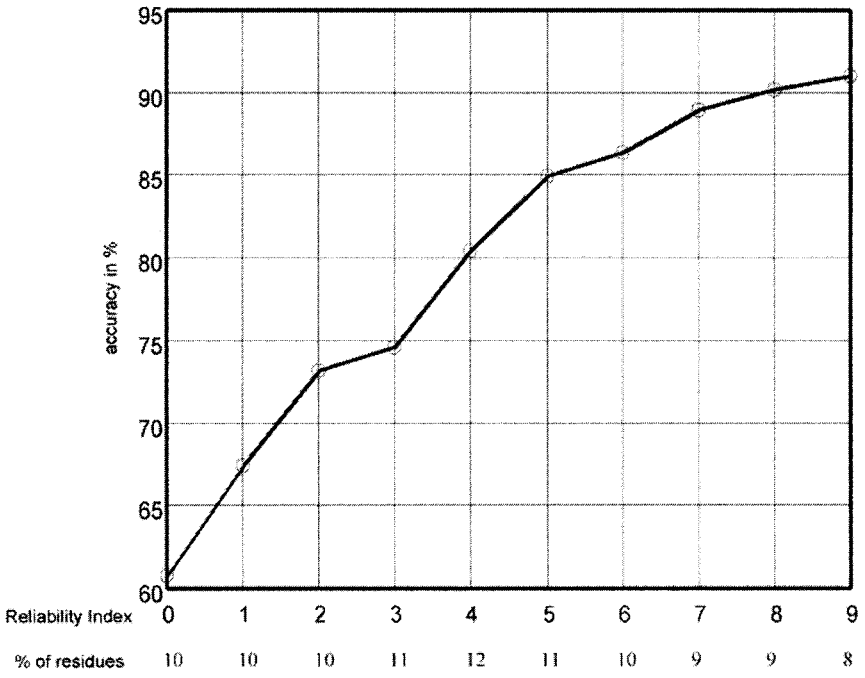

Figure 2. Expected prediction accuracy for residues with a reliability index equal to a given value. The accuracy of all residues with reliability index $RI = n$, $n = 0,1,\ldots,9$ is given. The fraction of residues that predicted with $RI = n$ are also given. For example, 12 % of all residues have $RI = 4$ and 80.4 % of these are correctly predicted.

Table 3. Accuracy of tertiary classifiers on the RS126 set

Type of tertiary classifier	Q_3 (%)	Q_H (%)	Q_E (%)	Q_C (%)	Q_H^{pre} (%)	Q_E^{pre} (%)	Q_C^{pre} (%)	C_H	C_E	C_C	SOV (%)
SVM_MAX_D	71.1	72.0	56.1	77.2	75.1	65.2	71.2	0.61	0.51	0.49	73.2
SVM_TREE1	68.0	72.0	48.5	74.0	74.0	59.0	69.0	0.60	0.43	0.44	69.5
SVM_TREE2	67.5	71.0	56.2	70.3	72.0	64.2	69.0	0.58	0.50	0.42	69.1
SVM_TREE3	66.8	71.0	52.7	76.0	70.0	60.5	72.0	0.56	0.46	0.46	68.6
SVM_VOTE	69.2	71.2	53.7	75.2	73.0	64.0	70.2	0.59	0.48	0.46	70.6
SVM_NN	70.5	72.0	55.9	76.0	74.3	64.6	70.2	0.61	0.50	0.47	72.5
SVM_JURY	71.6	72.5	58.1	77.0	76.0	65.0	70.2	0.62	0.52	0.51	74.6

More details for each tertiary classifier are given in the Methods section. Combined results of sevenfold cross validation are shown.

improve in accuracy with comparison to reduction method 2 (see Materials and Methods), when one uses another reduction method (E as E, H as H, rest to coil C including EE and HHHH). The apparent improvement of Q_3 is between 2.2 % and 4.9 %. Thus an objective comparison can not be made unless the same reduction method is used.

Furthermore, the SVM method has its unique merits and it can successfully avoid many problems which other machine learning approaches often encounter. For example, structures of NNs (especially the size of the hidden layer) are hard to determine; gradient based training algorithms only guarantee finding local minima; too many model parameters have to be optimized, overfitting problems are hard to avoid, etc. The effectiveness of SVM in overcoming these problems has proved that SVM is a promising method in practice.

Finally, there leaves room for further improvement of our prediction method. On one hand, it has been shown that better multiple sequence alignment profiles yield better prediction.[21] Jones[24] used the alignment profile generated by PSI-BLAST to design a set of NNs and his PSIPRED method achieved an overall per-residue accuracy $Q_3 = 76.5$ % and SOV = 73.5 % on his own data set including 187 protein chains. An entirely objective comparison between the SVM method and PSIPRED has not been carried out at this time for the lack of Jones's data set. A rough comparison between the two methods shows that Q_3 of the SVM method is about 3 % lower than that of PSIPRED on different data sets however SOV of the SVM method is about 3 % higher than that of PSIPRED. As mentioned above, SOV is used to assess the quality of prediction in a more realistic manner. Comparing PSIPRED with PHD method, both methods adopted the similar cascaded NNs and the mainly difference is the input of the first-level NN. The PSI-BLAST profiles were used in PSIPRED. These profiles adopted by Jones have some basic advantages: more distant sequences are found; the probability of each residue at a specific position is computed using a more rigorous statistical approach; and each sequence is properly weighted with respect to the amount of information it carries. These profiles contain much more useful information than our training profiles which are based on the older HSSP profile approach. It indicates that the improvements of PSIPRED seem to result in mainly from the use of PSI-BLAST generated profiles. Therefore it is quite possible to achieve significant improvements by incorporating PSI-BLAST profiles in the SVM approach. Further work is going on to construct a prediction system based on the latest version of RCSB Protein Data Bank (PDB) and the PSI-BLAST alignment profiles. It will be expected to achieve higher accuracy on a larger data set and new type of alignment profiles. On the other hand, it has been shown that prediction accuracy can be improved by combing more than one prediction method.[8,9] It is possible to obtain higher prediction accuracy when our single method is combined with other good single method, for example, PHD, PSIPRED, etc.

Table 4. Accuracy of tertiary classifier on the CB513 set

Type of tertiary classifier	Q_3 (%)	Q_H (%)	Q_E (%)	Q_C (%)	Q_H^{pre} (%)	Q_E^{pre} (%)	Q_C^{pre} (%)	C_H	C_E	C_C	SOV (%)
SVM_MAX_D	72.9	74.8	58.6	79.0	78.2	67.1	68.5	0.64	0.53	0.51	75.4
SVM_TREE1	68.9	73.5	54.0	73.1	79.0	63.0	66.2	0.64	0.47	0.45	72.1
SVM_TREE2	68.2	72.0	61.0	69.0	77.0	67.2	64.1	0.62	0.54	0.40	71.4
SVM_TREE3	67.5	69.5	46.6	77.0	75.1	63.0	69.2	0.58	0.43	0.51	70.8
SVM_VOTE	70.7	73.0	56.2	76.6	76.5	65.0	67.1	0.62	0.50	0.48	73.2
SVM_NN	72.0	74.7	57.7	77.4	78.0	66.7	69.0	0.64	0.52	0.51	75.0
SVM_JURY	73.5	75.2	60.3	79.5	79.1	67.3	70.2	0.64	0.52	0.51	76.2

More details for each tertiary classifier are given in the Methods section. Combined results of sevenfold cross validation are shown.

Table 5. Comparison with the results of the PHD, the NN based approach

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	I_H	I_E	I_C
PHD[1]	73.5	70.8	72	66	72	73	60	-	0.60	0.52	0.51	9.3	5.0	-
SVM[1]	74.6	71.2	73	58	75	77	66	69	0.61	0.51	0.52	9.2	4.5	7.6
(observed)												9.0	5.1	5.9
PHD[2]	–	72.1	70	62	79	77	64	72	0.63	0.53	0.52	10.3	5.0	7.2
(observed)												9.3	5.3	5.9
SVM[2]	76.2	73.5	75	60	79	79	67	70	0.65	0.53	0.54	9.2	4.8	7.8
(observed)												9.4	5.4	5.9

PHD[1], SVM[1]: Results obtained on the RS126 set using the eight- to three-state reduction method 1. (Except for this, all other results in this paper are based on reduction method 2.) The results of PHD[1] are from Rost & Sander (1993)[4] and Rost *et al.* (1994).[21]
PHD[2]: Results obtained on another non-homologous data set which contains 250 protein chains (Rost & Sander, 1994).[4]
SVM[2]: Results obtained on the CB513 set.
– The result can not be obtained from the papers.
The SVM and PHD methods both use jury decision. Combined results of sevenfold cross validation are shown.

Condensing information by SVM

One attractive property of SVM is that SVM condenses information in the training samples to provide a sparse representation using a very small number of samples, support vectors (SVs). More details can be seen in the Supplementary Material. The SVs characterize the solution to the problem in the following sense: if all the other training samples are removed and the SVM are retrained, then the solution would be unchanged. It is believed that all the information about classification in the training samples can be represented by these SVs. In a typical case, the number of SVs is small compared to the total number of training samples which enables the SVM to efficiently classify new samples, since the majority of the training samples can be safely ignored. This is a crucial property when analyzing large data sets containing many uninformative patterns, as is the case in many data mining problems since the SVM can effectively remove the uninformative patterns in the data set. This property will be especially useful in the field of bioinformatics as a mass of experimental data explodes.

For secondary structure prediction, the ratio of SVs to all training samples is shown in Table 6 for each binary classifier. The ratio for each classifier is about 50 %, which means that half of the training samples could represent all of the classification information. With many other classification problems, a ratio of 50 % is relatively higher which indicates that the protein secondary structure classification problem is relatively difficult.

Table 6. Ratio of number of SVs to all training samples

Binary classifier	Ratio (%) (SVs/all samples)
H/∼H	50.46
E/∼E	43.92
C/∼C	59.02
H/E	50.27
E/C	53.16
C/H	52.62

The results are for the CB513 set. Combined results of sevenfold cross validation are shown.

The influence of noise and outliers

In practical problems, the influence of noise and outliers is inevitable. If the window length is not appropriate, the signal-to-noise ratio will decrease. All the discussions about SVM are based on the basic precondition that all samples in the training set are independent and identically distributed (I.I.D.) according to the unknown dependency to be estimated.[14] A training set which is polluted by noise, i.e, training samples are not I.I.D., influences the prediction accuracy. In addition, if we try to correctly classify the outliers, the classifier might be quite complex but with poor generalization performance.

The influence of noise and outliers could be reduced using the method of central support vector machine (CSVM), which is an improved SVM.[25] Another idea is to retrain the SVM classifier only on the samples which were predicted to have relatively large distances to the OSH. Thus we discard a fraction of samples which are hard to predict because they are located near the OSH so the boundary between these two classes will become much clearer. The result should give simpler classifiers that have better performance.

More applications in biology

The protein secondary structure prediction problem is just one typical classification task in bioinformatics which has many other similar tasks. Here we have demonstrated that the SVM method is competitive with and even superior to other, more frequently used machine learning methods and SVM offers some advantages e.g. effective avoidance of overfitting, the ability to handle large feature spaces, information condensing of the

given data set, convenient machine capacity control, etc.

SVM can also be applied for many other tasks, for example, in biosequence analysis, such as the identification and classification of splice sites, promoters and other cis-acting elements in genomic DNA, in microarray data analysis, such as class prediction and gene function prediction.

While this article was being finalized, other papers appeared which had used the SVM method to functionally classify genes by using gene expression data from DNA microarray hybridization experiments,[26] to detect remote protein homologies (http://www.cse.ucsc.edu/research/compbio/discriminative/), to recognize translation initiation sites (TIS) (http://www.bioinfo.de/isb/gcb99/talks/zien/). The conclusions of those papers generally agree with ours that SVM as a discriminative supervised machine learning technology offers us a powerful tool in dealing with the pattern classification tasks in biology.

Materials and Methods

Data set

Two data sets were used to develop and test the algorithms. One is the data set of 126 protein chains proposed by Rost & Sander,[4] referred to as the RS126 set. Many current generation secondary structure prediction methods were developed and tested on this set. It is a non-homologous data set according to the definition of Rost & Sander.[4] They used percentage identity to measure the homology and defined non-homologous to mean that no two proteins in the set share more than 25% sequence identity over a length of more than 80 residues.

The other much larger data set was constructed by Cuff & Barton[9] and is referred to as the CB513 set since it has 513 protein chains. Almost all the sequences in the RS126 set are included in the CB513 set. It is also a non-homologous data set, i.e. an SD score of $\geqslant 5$ is regarded as homology. The SD score is a more stringent measure than the percentage identity. In fact, 11 pairs of proteins in the RS126 set are sequence similar when using the SD score instead of percentage identity. The CB513 set contains 16 chains of $\leqslant 30$ residues. It has been shown that very short chains in the set will slightly decrease the accuracy for hard definition of secondary structures.

Multiple sequence alignments

The prediction procedure introduced here can be easily extended to the use of multiple sequence alignments of proteins homologous with the target protein. The profiles of the multiple alignments for the RS126 set are taken from the HSSP database,[27] the same database as used by the PHD method. The objective was to compare the prediction accuracy of different prediction algorithms as objectively as possible. The RS126 set including the multiple sequence alignment profiles is available at http://www.sander.embl-heidelberg.de/hssp/. For the CB513 set, another automatic procedure proposed by Cuff & Barton[9] was used to generate multiple alignment profiles. The CB513 set including the automatically generated multiple alignment profiles is available at http://barton.ebi.ac.uk/.

Protein secondary structure definition

The automatic assignments of secondary structure to experimentally determined 3D structures is usually performed by DSSP,[23] STRIDE[28] or DEFINE.[29]

Different assignment methods influence the prediction accuracy to some extent, as discussed by Cuff & Barton.[9] Here, we concentrate exclusively on the DSSP assignments, which distinguish eight secondary structure classes: H(α-helix), G(3_{10}-helix), I(π-helix), E(β-strand), B(isolated β-bridge), T(turn), S(bend) and -(rest). We reduced the eight classes to three state, helix (H), sheet (E) and coil (C) using two different methods: (i) DSSP: H, G and I to H; E to E; all other states to C; (ii) DSSP: H and G to H; E and B to E; all other states to C. Reduction method 1 is the same as the PHD method.[4] We adopted it to objectively compare the prediction performance of PHD and SVM. The second reduction method is adopted because it is now widely used. Other reduction methods have been proposed and some effects of the different methods on prediction performance were discussed by Cuff & Barton.[9]

Prediction accuracy assessment

Multiple cross-validation trials are necessary to minimize variation in the results caused by a particular choice of training or test sets. A full jack-knife test is not feasible especially on the CB513 set due to the limited computational power. Therefore, the sevenfold cross validation was used on both RS126 set and the CB513 set. The RS126 set or the CB513 set was divided into seven subsets with each subset having similar size and similar content of each type of secondary structure. Take the CB513 set as one example. In fact, we tried several (>10) different random partitions of the CB513 set (six subsets have 73 protein chains and one subset has 75 chains). For each partition, we calculated the number of residues and the content of each secondary structure type (H, E and C) of each subset. The partition finally selected had the minimal bias. The procedure we did was used to avoid the selection of extremely biased parition which would give an inauthentic prediction accuracy. All results reported in this paper were obtained using the cross validation.

Several standard performance measures were used to assess prediction accuracy. Three-state overall per-residue accuracy (Q_3), Matthew's correlation coefficients (C_H,C_E,C_C) and SOV were calculated to evaluate accuracy with the details are given in the previous paper.[4] The per-residue accuracy for each type of secondary structure (Q_H,Q_E,Q_C; Q_H^{pre},Q_E^{pre},Q_C^{pre}) was also calculated. We distinguish Q_I and Q_I^{pre}(here, I = H, E and C) as:

$$Q_I(\%) = \frac{\text{number of residues correctly predicted in state I}}{\text{number of residues observed in state I}} \times 100 \tag{3}$$

and

$$Q_I^{pre}(\%) = \frac{\text{number of residues correctly predicted in state I}}{\text{number of residues predicted in state I}} \times 100 \tag{4}$$

Protein secondary structure prediction is a typical tertiary classification problem. Almost all previous machine learning approaches sought to directly design the tertiary

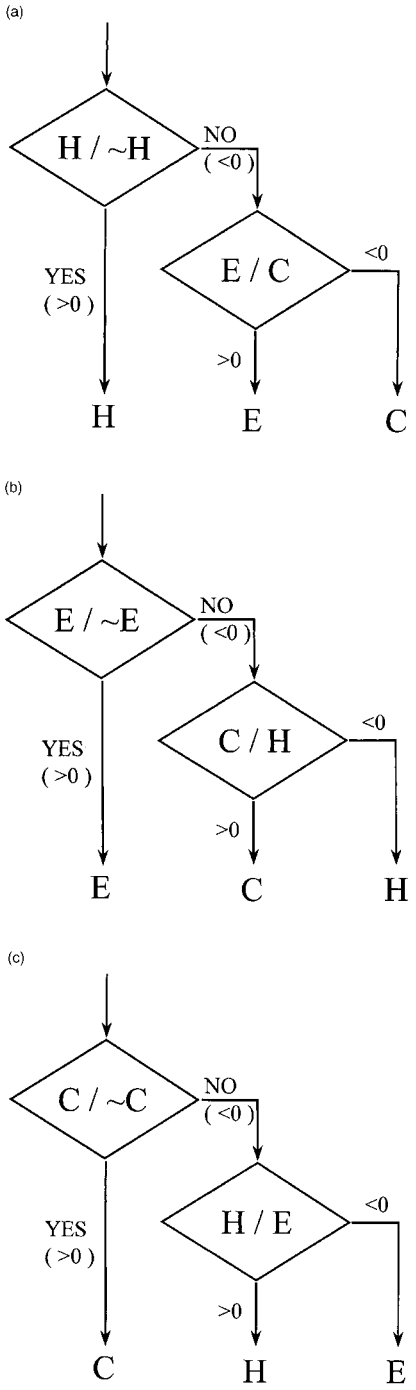

Figure 3. The structures of tertiary classifiers, (a) SVM_TREE1, (b) SVM_TREE2 and (c) SVM_TREE3. Each of them is made up of two cascaded binary classifiers. Take SVM_TREE1 as an example, a sample will be classified as helix (H) if the output of the first binary classifier, H/~H is larger than 0, otherwise the second classifier, E/C will be used. If the output of E/C is larger than 0, the sample will be classified as sheet (E), otherwise coil (C) will be assigned.

classifier. Unlike these previous studies, we first designed several binary classifiers, then assembled a tertiary classifier based on these binary classifiers.

Coding scheme

As with the neural network approach, we adopted the classical local coding scheme of the protein sequence with a sliding window.[3] For a single sequence, each residue is coded by the orthogonal binary vector (1,0,...,0) or (0,1,...,0), etc. The vector is 21-dimensional. Among the first twenty units of the vector, each unit stands for one type of amino acid. In order to allow a window to extend over the N terminus and the C terminus, the 21st unit has to be added for each residue. If the window length is l, the dimensionality of the feature vector (or the sample space) is $21 \cdot l$. After including the evolutionary information, the multiple sequence alignments are represented for the single sequence. The frequency of occurrence of each of the 20 amino acid residues at one position in the alignment is computed for each residue.

Constructing the binary classifiers

We first constructed several SVM binary classifiers including three one-versus-rest classifiers (say, ''one'': positive class, ''rest'': negative class) named H/~H, E/~E and C/~C, and three classifiers named H/E, E/C and C/H which distinguish the sample between each of two states. For example, the classifier H/E is constructed on the training samples having helices and sheets and it classifies the testing sample as helix or sheet.

Before constructing each binary classifier, we first selected the appropriate kernel function and its parameters, the regularization parameter C (see Supplementary Material) and the optimal window length. These selections were done *via* numerical computing.

The programs for constructing the SVM classifier were written in the C language. The SVM training procedure amounts to solving a convex quadratic programming problem (see Supplementary Material). The core optimization method used here was based on the ''LOQO'' algorithm.[30-32]

Tertiary classifier design

The goal of the machine learning approach to secondary structure prediction is to construct a tertiary classifier with good prediction performance. Thus, the next step is to design a tertiary classifier using the above trained binary classifier. Several design methods were used.

We combined the three one-versus-rest classifiers (H/~H, E/~E and C/~C) to handle the multiclass case. The class (H, E or C) for a testing sample was assigned as that corresponding to the largest positive distance to the OSH. We called the combined tertiary classifier ''SVM_MAX_D''.

Other tertiary classifier structures like decision trees are shown in Figure 3. They are referred to as ''SVM_TREE1'', ''SVM_TREE2'' and ''SVM_TREE3''. These classifiers were made up of two cascaded binary classifiers as shown in Figure 3.

Another structure of classifier was also designed. It combined all the six binary classifiers using a simple voting principle that the testing sample was predicted to be in state I if the largest number of the six binary classifiers classified it as state I. One exception was that the testing sample had two classifications in each state then it was considered to be a coil. This classifier was referred to as ''SVM_VOTE''.

Another classifier was combined with a NN and was named ''SVM_NN''. The outputs of the binary classifiers were the inputs to the NN so the input layer size is six. The NN had one hidden layer and used the back propagation (BP) algorithm. The hidden layer size was taken as twenty after various tests. The output layer of the network had three units. The variables (weights and bias) of the fully connected network were determined during the training procedure.

Finally, as with the PHD method and others, we used the jury technique to combine all the results of the tertiary classifiers discussed above. The combined classifier was named ''SVM_JURY''.

Prediction ''reliability index''

The output information from the classifiers H/∼H, E/∼E and C/∼C was used to develop a position-specific RI for the predictions. The simple and intuitive RI offers more help on using the prediction results.

Acknowledgments

The authors would like to thank Professor Yanda Li for useful comments and discussion, thank J. A. Cuff and G. J. Barton for providing the CB513 data set and thank Professor R. J. Vanderbei for providing the programs of ''LOQO'' algorithm. This work is supported in part by the National Natural Science Grant (China) (No. 39980007) and partially by the National Key Foundational Research Grant (985) and TongFang Grant.

References

1. Bairoch, A. & Apweiler, R. (2000). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 2000. *Nucl. Acids Res.* **28**, 45-48.
2. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N. & Weissig, H. *et al*. (2000). The Protein Data Bank. *Nucl. Acids Res.* **28**, 235-242.
3. Qian, N. & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865-884.
4. Rost, B. & Sander, C. (1993). Prediction of secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584-599.
5. Riis, S. K. & Krogh, A. (1996). Improving prediction of protein secondary structure suing structured neural networks and multiple sequence alignments. *J. Comput. Biol.* **3**, 163-183.
6. Baldi, P., Brunak, S., Frasconi, P., Soda, G. & Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics,* **15**, 937-946.
7. Chandonia, J. M. & Karplus, M. (1999). New methods for accurate prediction of protein secondary structure. *Proteins: Struct. Funct. Genet.* **35**, 293-306.
8. Zhang, X., Mesirov, J. P. & Waltz, D. L. (1992). Hybrid system for protein secondary structure prediction. *J. Mol. Biol.* **225**, 10490-1063.
9. Cuff, J. A. & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Struct. Funct. Genet.* **34**, 508-519.
10. King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **5**, 2298-2310.
11. Salamov, A. A. & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* **247**, 11-15.
12. Frishman, D. & Argos, P. (1995). Knowledge-based secondary structure assignment. *Proteins: Struct. Funct. Genet.* **23**, 566-579.
13. Cortes, C. & Vapnik, V. (1995). Support vector networks. *Machine Learning,* **20**, 273-293.
14. Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.
15. Vapnik, V. (1998). *Statistical Learning Theory*, John Wiley and Sons, Inc., New York.
16. Scholkopf, B., Burges, C. & Vapnik, V. (1995). Extracting support data for a given task. In *Proceedings, First International Conference on Knowledge Discovery and Data Mining* (Fayyad, U. M. & Uthurusamy, R., eds), pp. 252-257, AAAI Press, Menlo Park, CA.
17. Roobaert, D. & Hulle, M. M. (1999). View-based 3D object recognition with Support Vector Machines, In *Proceedings of the IEEE Neural Networks for Signal Processing Workshop* (Hu, Y. H., Larsen, J., Wilson, E. & Douglas, S., eds), pp. 77-84, IEEE Press, NJ.
18. Schmidt, M. & Grish, H. (1996). Speaker identification *via* support vector classifiers. In *The Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 1996*, pp. 105-108, IEEE Press, Long Beach, CA.
19. Drucker, H., Wu, D. & Vapnik, V. (1999). Support vector machines for spam categorization. *IEEE Trans. on Neural Networks,* **10**, 1048-1054.
20. Boser, B. E., Guyon, I. M. & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on Computational Learning Theory*, ACM, Pittsburgh.
21. Rost, B., Sander, C. & Schneider, R. (1994). Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* **235**, 13-26.
22. Zemla, A., Venclovas, C., Fidelis, K. & Rost, B. (1999). A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins: Struct. Funct. Genet.* **34**, 220-223.
23. Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers,* **22**, 2577-2637.
24. Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195-202.
25. Zhang, X. (1999). Using class-center vectors to build support vector machines, *Proceeding of the 1999 IEEE Signal Processing Society Workshop* (Hu, Y. H., Larsen, J., Wilson, E. & Douglas, S., eds), pp. 3-11, IEEE Press, NJ.
26. Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W. & Furey, T. S. *et al*. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA,* **97**, 262-267.
27. Dodge, C., Schnerder, R. & Sander, C. (1998). The HSSP database of protein structure-sequence alignments and family profiles. *Nucl. Acids Res.* **26**, 313-315.

28. Frishman, D. & Argos, P. (1995). Knowledge-based secondary structure assignment. *Proteins: Struct. Funct. Genet.* **23**, 566-579.
29. Richards, F. M. & Kundrot, C. E. (1988). Identification of structural motifs from protein coordinate data: secondary structure and first-level supersecondary structure. *Proteins: Struct. Funct. Genet.* **3**, 71-84.
30. Vanderbei, R. J. (1994). Interior point methods: algorithms and formulations. *ORSA J. Comput.* **6**, 32-34.
31. Vanderbei, R. J. (1994). *LOQO: An Interior Point Code for Quadratic Programming*, Program in Statistics & Operations Research Technical report, Princeton University, NJ.
32. Joachims, T. (1999). Making large-scale SVM learning practical. In *Advances in Kernel Methods Vector Learning* (Schölkopf, B., Burges, C. & Smola, A., eds), pp. 42-56, MIT-Press, MA.

Edited by B. Holland

(Received 24 November 2000; received in revised form 21 February 2001; accepted 24 February 2001)

http://www.academicpress.com/jmb

Supplementary Material is abailable on IDEAL