BIOINFORMATICS

Vol. 17 no. 8 2001
Pages 721–728

Support vector machine approach for protein subcellular localization prediction

*Sujun Hua and Zhirong Sun**

Institute of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, People's Republic of China

Received on December 12, 2000; revised on March 28, 2001; accepted on April 24, 2001

ABSTRACT

Motivation: Subcellular localization is a key functional characteristic of proteins. A fully automatic and reliable prediction system for protein subcellular localization is needed, especially for the analysis of large-scale genome sequences.

Results: In this paper, Support Vector Machine has been introduced to predict the subcellular localization of proteins from their amino acid compositions. The total prediction accuracies reach 91.4% for three subcellular locations in prokaryotic organisms and 79.4% for four locations in eukaryotic organisms. Predictions by our approach are robust to errors in the protein N-terminal sequences. This new approach provides superior prediction performance compared with existing algorithms based on amino acid composition and can be a complementary method to other existing methods based on sorting signals.

Availability: A web server implementing the prediction method is available at http://www.bioinfo.tsinghua.edu.cn/SubLoc/.

Contact: sunzhr@mail.tsinghua.edu.cn; huasj00@mails.tsinghua.edu.cn

Supplementary information: Supplementary material is available at http://www.bioinfo.tsinghua.edu.cn/SubLoc/.

INTRODUCTION

High throughput genome sequencing projects are producing an enormous amount of raw sequence data. All this raw sequence data begs for methods that are able to catalog and synthesize the information into biological knowledge. Genome function annotation including the assignment of a function for a potential gene in the raw sequence is now the hot topic in bioinformatics. Subcellular localization is a key functional characteristic of potential gene products such as proteins (Eisenhaber and Bork, 1998). Therefore, a fully automatic and reliable prediction system for protein subcellular localization would be very useful.

Several attempts have been made to predict protein subcellular localization. Most of these prediction methods can be classified into two categories: one is based on the recognition of protein N-terminal sorting signals and the other is based on amino acid composition (Nakai, 2000). von Heijne and colleagues have worked extensively on identifying individual sorting signals, e.g. signal peptides, mitochondrial targeting peptides and chloroplast transit peptides (Nielsen *et al.*, 1997, 1999; von Heijne *et al.*, 1997). More recently, they proposed an integrated prediction system for subcellular localization using neural networks based on individual sorting signal predictions (Emanuelsson *et al.*, 2000). One advantage of their method is that it can recognize cleavage sites in the sorting signals and can mimic the real sorting process to a certain extent. The reliability of methods based on sorting signals is strongly dependent on the quality of the gene 5′-region or protein N-terminal sequence assignment. However, the assignments of 5′-regions are usually not reliable using known gene identification methods (Frishman *et al.*, 1999). Therefore, subcellular localization prediction methods which depend on sorting signals will be inaccurate when the signals are missing or only partially included. In addition, the known signals are not general enough to cover the resident proteins in each organelle (Nakai, 2000).

Other efforts are concentrated on the deviations of amino acid composition with different subcellular localizations. Nakashima and Nishikawa (1994) have shown that intracellular and extracellular proteins differ significantly in their amino acid composition. Andrade *et al.* (1998) indicated that the localizations correlate better with the surface composition due to evolutionary adaptation of proteins to different physio-chemical environments in each subcellular location. Cedano *et al.* (1997) proposed a statistical method using the Mahalanobis distance but did not obtain satisfying results. Reinhardt and Hubbard (1998) constructed a prediction system using supervised neural networks. They dealt with prokaryotic and eukaryotic sequences separately to obtain a total accuracy

*To whom correspondence should be addressed.

of 81% for three subcellular locations in prokaryotic sequences and 66% for four locations in eukaryotic sequences. Chou and Elrod (1999) proposed a covariant discriminant algorithm to achieve a total accuracy of 87% by the jackknife test on the same prokaryotic sequences used by Reinhardt and Hubbard. Nakai and colleagues developed an integrated expert system using both sorting signal knowledge and amino acid composition information (Nakai and Kanehisa, 1991, 1992; Nakai and Horton, 1997). Yuan (1999) used Markov chain models to achieve 89% accuracy for prokaryotic sequences and 73% for eukaryotic sequences on the same dataset used by Reinhardt and Hubbard.

This paper introduces a new prediction method for protein subcellular localization based on amino acid composition. This method, called Support Vector Machine (SVM), was recently proposed by Vapnik and co-workers (Cortes and Vapnik, 1995; Vapnik, 1995, 1998) as a very effective method for general purpose supervised pattern recognition. The SVM approach is not only well founded theoretically because it is based on extremely well developed machine learning theory, Statistical Learning Theory (Vapnik, 1995, 1998), but is also superior in practical applications. The SVM method has been successfully applied to isolated handwritten digit recognition (Cortes and Vapnik, 1995; Scholkopf *et al.*, 1995), object recognition (Roobaert and Hulle, 1999), text categorization (Drucker *et al.*, 1999), microarray data analysis (Brown *et al.*, 2000), protein secondary structure prediction (Hua and Sun, 2001), etc. Here, we construct a prediction system for subcellular localization called SubLoc based on the SVM method. The results show that the prediction accuracy is significantly improved with this novel method and the method is very robust to errors in the protein N-terminal sequence.

MATERIALS AND METHODS

Data set

The dataset used to examine the effectiveness of the new prediction method was generated by Reinhardt and Hubbard (1998). The sequences in this dataset were extracted from SWISSPROT release 33.0 and included only those essential sequences which appeared complete and had reliable localization annotations coming directly from experiments. No transmembrane proteins were included as they could be quite reliably predicted by some known methods (Rost *et al.*, 1996; Hirokawa *et al.*, 1998; Lio and Vannucci, 2000). Redundancy was reduced such that none had >90% sequence identity to any other in the set. Finally, as shown in Table 1, the dataset included 997 prokaryotic sequences which were classified into three location categories (cytoplasmic, periplasmic and extracellular) and 2427 eukaryotic sequences belonging to four location categories (nuclear, cytoplasmic, mitochondrial and extracellular).

Table 1. Number of sequences within each subcellular localization category of the dataset (Reinhardt and Hubbard, 1998)

Species	Subcellular localization	Number of sequences
Prokaryotic	Cytoplasmic	688
	Periplasmic	202
	Extracellular	107
Eukaryotic	Nuclear	1097
	Cytoplasmic	684
	Mitochondrial	321
	Extracellular	325

Support vector machine

Here we briefly describe the basic ideas behind SVM for pattern recognition, especially for the two-class classification problem, and refer readers to Vapnik (1995, 1998) for a full description of the technique.

For a two-class classification problem, assume that we have a set of samples, i.e. a series of input vectors $\vec{\mathbf{x}}_i \in \mathbb{R}^d (i = 1, 2, \ldots, N)$ with corresponding labels $y_i \in \{+1, -1\} (i = 1, 2, \ldots, N)$. Here, $+1$ and -1 indicate the two classes. To predict protein subcellular localization, the input vector dimension is 20 and each input vector unit stands for one amino acid. The goal is to construct a binary classifier or derive a decision function from the available samples which has a small probability of misclassifying a future sample.

SVM implements the following idea: it maps the input vectors $\vec{\mathbf{x}} \in \mathbb{R}^d$ into a high dimensional feature space $\Phi(\vec{\mathbf{x}}) \in \mathbb{H}$ and constructs an Optimal Separating Hyperplane (OSH), which maximizes the margin, the distance between the hyperplane and the nearest data points of each class in the space $\mathbb{H}$ (see Figure 1). Different mappings construct different SVMs. The mapping $\Phi(\cdot)$ is performed by a kernel function $K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j)$ which defines an inner product in the space $\mathbb{H}$.

The decision function implemented by SVM can be written as:

$$f(\vec{\mathbf{x}}) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\vec{\mathbf{x}}, \vec{\mathbf{x}}_i) + b\right) \qquad (1)$$

where the coefficients α_i are obtained by solving the following convex Quadratic Programming (QP) problem:

$$\begin{aligned} \text{Maximize} \quad & \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) \\ \text{subject to} \quad & 0 \leqslant \alpha_i \leqslant C \qquad (2) \\ & \sum_{i=1}^{N} \alpha_i y_i = 0 \quad i = 1, 2, \ldots, N. \end{aligned}$$

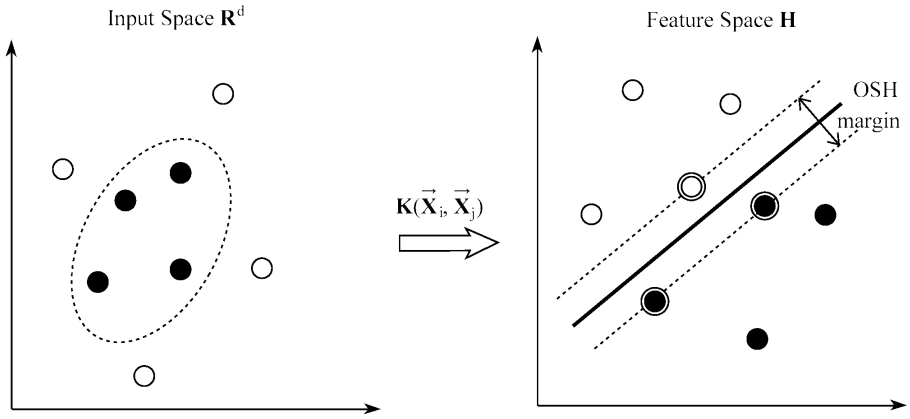

Fig. 1. A separating hyperplane in the feature space corresponding to a non-linear boundary in the input space. Two classes denoted by circles and disks are linear non-separable in the input space $\mathbb{R}^d$. SVM constructs *the Optimal Separating Hyperplane* (OSH) (the solid line) which maximizes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

In the equation (2), C is a regularization parameter which controls the trade off between margin and misclassification error. These $\vec{\mathbf{x}}_j$ are called *Support Vectors* only if the corresponding $\alpha_i > 0$.

Several typical kernel functions are

$$K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = (\vec{\mathbf{x}}_i \bullet \vec{\mathbf{x}}_j + 1)^d, \quad (3)$$

$$K(\vec{\mathbf{x}}_i, \vec{\mathbf{x}}_j) = \exp(-\gamma \|\vec{\mathbf{x}}_i - \vec{\mathbf{x}}_j\|^2), \quad (4)$$

Equation (3) is the *polynomial kernel function* of degree d which will revert to the linear function when $d = 1$. Equation (4) is the *Radial Basic Function* (RBF) kernel with one parameter γ.

For a given dataset, only the kernel function and the regularization parameter C are selected to specify one SVM. SVM has many attractive features. For instance, the solution of the QP problem is globally optimized while with neural networks the gradient based training algorithms only guarantee finding a local minima. In addition, SVM can handle large feature spaces, can effectively avoid overfitting by controlling the margin, can automatically identify a small subset made up of informative points, i.e. the Support Vectors, etc.

Design and implementation of the prediction system

Protein subcellular localization prediction is a multi-class classification problem. Here, the class number is equal to 3 for prokaryotic sequences and 4 for eukaryotic sequences. A simple strategy to handle the multi-class classification is to reduce the multi-classification to a series of binary classifications. For a k-class classification, k SVMs are constructed. The ith SVM will be trained with all of the samples in the ith class with positive labels and all other samples with negative labels. We refer to SVMs trained in this way as 1-v-r SVMs (short for one-versus-rest). Finally one unknown sample is classified into the class that corresponds to the 1-v-r SVM with the highest output value.

This method was used to construct a prediction system (i.e. one 3-class classifier for prokaryotic sequences and one 4-class classifier for eukaryotic sequences) for protein subcellular localization. The prediction system is named SubLoc and is available at http://www.bioinfo.tsinghua.edu.cn/SubLoc/.

The software used to implement SVM was *SVM*$^{\text{light}}$ by Joachims (1999) which can be freely downloaded from http://ais.gmd.de/~thorsten/svm_light/ for academic use. The core optimization method for solving the QP problem was based on the 'LOQO' algorithm (Vanderbei, 1994). In this work, training a binary SVM usually takes less than 10 min on a PC running at 500 MHz. The algorithm spends less time on the classification of unknown samples because we only need to calculate the inner products between the unknown samples and a small subset made up of the Support Vectors. SVM is, consequently, an efficient classifier.

Prediction system assessment

The prediction quality was examined using the jackknife test, an objective and rigorous testing procedure. In the jackknife test, each protein was singled out in turn as a test protein with the remaining proteins used to train SVM. The total prediction accuracy, the prediction accuracy and Matthew's Correlation Coefficient (MCC) (Matthews, 1975) for each location calculated for assessment of the prediction system are given by

Table 2. Prediction accuracies for prokaryotic sequences with different type of kernel functions

Location	Linear		Polynomial*		RBF	
	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	98.1	0.83	97.5	0.86	97.5	0.86
Periplasmic	66.8	0.68	78.7	0.78	78.2	0.78
Extracellular	74.8	0.76	75.7	0.77	76.6	0.77
Total accuracy	89.3	–	91.4	–	91.4	–

Linear: polynomial kernel with $d = 1$; Polynomial*: polynomial kernel with $d = 9$ which is finally used in our prediction system; RBF: RBF kernel with $C = 1000$ was used for each SVM. The results were given by the jackknife test.

Table 3. Prediction accuracies for eukaryotic sequences with different type of kernel functions

Location	Polynomial		RBF*	
	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	78.4	0.63	76.9	0.64
Extracellular	70.2	0.71	80.0	0.78
Mitochondrial	46.1	0.53	56.7	0.58
Nuclear	88.0	0.72	87.4	0.75
Total accuracy	77.3	–	79.4	–

Polynomial: polynomial kernel with $d = 9$; RBF*: RBF kernel with $\gamma = 16.0$ which is finally used in our prediction system. $C = 500$ was used for each SVM. The results were given by the jackknife test.

$$\text{total accuracy} = \frac{\sum_{i=1}^{k} p^{(i)}}{N}, \tag{5}$$

$$\text{accuracy}\,(i) = \frac{p(i)}{\text{obs}(i)}, \tag{6}$$

$$\text{MCC}(i) = \frac{p(i)n(i) - u(i)o(i)}{\sqrt{(p(i)+u(i))(p(i)+o(i))(n(i)+u(i))(n(i)+o(i))}}. \tag{7}$$

Here, N is the total number of sequences, k is the class number, obs(i) is the number of sequences observed in location i, $p(i)$ is the number of correctly predicted sequences of location i, $n(i)$ is the number of correctly predicted sequences not of location i, $u(i)$ is the number of under-predicted sequences and $o(i)$ is the number of over-predicted sequences.

RESULTS

SubLoc prediction accuracy

The prediction accuracies by jackknife tests for prokaryotic sequences are shown in Table 2. The total accuracy predicted by the current method reached 89.3% with the simplest linear kernel function. This indicates that the prokaryotic samples can be well separated by a proper linear hyperplane in the input space. The accuracy could be improved by using the more complex non-linear kernel function. The total accuracy was improved to 91.4% using the RBF kernel with $\gamma = 5.0$ or the polynomial kernel function of degree 9. The details of prediction accuracies for each test protein by the jackknife test are given in the supplementary material at http://www.bioinfo.tsinghua.edu.cn/SubLoc/.

Table 3 shows the results for the eukaryotic sequences. The training procedure did not converge when a linear kernel was used which suggested that no hyperplane in the input space can clearly separate the eukaryotic samples. However, a proper non-linear kernel did work. Using the polynomial kernel function of degree 9, the total prediction accuracy was 77.3% and could be further improved to 79.4% using the RBF kernel with $\gamma = 16.0$.

Tests have been done with various kernel function parameters and value of the regularization parameter C. For the limited computational power, we use the results by 5-fold cross validation to select the appropriate parameters. The details of dataset partition for the cross validation and the prediction accuracies with different parameters by the cross validation can be seen at http://www.bioinfo.tsinghua.edu.cn/SubLoc/. The results by the cross validation we obtained were very close to the results by the jackknife test. Finally, the prediction system used the polynomial kernel function of degree 9 for prokaryotic sequences with $C = 1000$ and RBF kernel with $\gamma = 16.0$ for eukaryotic sequences with $C = 500$.

Comparison with other prediction methods

The SVM method predictions were compared with other prediction methods. The Reinhardt and Hubbard dataset was also tested with the neural network method (Reinhardt and Hubbard, 1998) and the covariant discriminant algorithm (Chou and Elrod, 1999). These two methods and the SVM method are all based on amino acid composition alone. The results for prokaryotic and eukaryotic sequences are summarized in Tables 4 and 5, respectively. The results of the covariant discrimination, the Markov model and the SVM method were obtained by the jackknife test while the neural network method results were with 6-fold cross validation.

As seen in Table 4, for prokaryotic sequences, the total accuracy of the SVM method is about 10% higher than that of the neural network method and about 5% higher than that of the covariant discriminant algorithm. The accuracy for cytoplasmic sequences reached 97.5% with the SVM method which is much higher than for the other methods. For eukaryotic sequences, the total accuracy was 13% higher than that of the neural network method (Table 5). The prediction accuracies for nuclear and cytoplasmic sequences were 15% and 22% higher than those of the

Table 4. Performance comparisons for the prokaryotic sequences. The neural network results were given by cross validation. The covariant discrimination, the Markov model and SVM method results were given by the jackknife test

Location	Neural network	Covariant discrimination	Markov model		SVM	
	Accuracy (%)	Accuracy (%)	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	80	91.6	93.6	0.83	97.5	0.86
Periplasmic	85	72.3	79.7	0.69	78.7	0.78
Extracellular	77	80.4	77.6	0.77	75.7	0.77
Total accuracy	81	86.5	89.1	–	91.4	–

Table 5. Performance comparisons for the eukaryotic sequences. The neural network results were given by cross validation. The Markov model and SVM method results were given by the jackknife test

Location	Neural network	Markov model		SVM	
	Accuracy (%)	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	55	78.1	0.60	76.9	0.64
Extracellular	75	62.2	0.63	80.0	0.78
Mitochondrial	61	69.2	0.53	56.7	0.58
Nuclear	72	74.1	0.68	87.4	0.75
Total accuracy	66	73.0	–	79.4	–

neural network method, although the accuracy for mitochondrial sequences was about 4% lower. These results indicate that the prediction accuracy can be significantly improved using the same classification information (amino acid composition) with a more powerful machine learning method.

The SVM method was also compared with the Markov chain model (Yuan, 1999), which was based on the full sequence information including the order information while the SVM method is based only on the amino acid composition. The total accuracies using the SVM method were 2.3% higher for prokaryotic sequences and 6.4% higher for eukaryotic sequences (Tables 4 and 5). For both the prokaryotic and eukaryotic sequences, the MCC of each subcellular location using the SVM method was higher than the corresponding one from Yuan's method.

Assigning a reliability index to the prediction

When using machine learning approaches for the prediction of protein subcellular localization, it is important to know the prediction reliability. For neural network methods, a Reliability Index (RI) is usually assigned according to the difference between the highest and the second-highest network output score (Rost and Sander, 1993; Reinhardt and Hubbard, 1998; Emanuelsson *et al.*, 2000). The simple idea is easily used with the SVM prediction system, i.e. assigning an RI according to the difference (noted as diff) between the highest and the second-highest output value of the 1-*v*-*r* SVMs in the multi-class classification. RI is defined as:

$$\mathrm{RI} = \begin{cases} \mathrm{INTEGER}\,(\mathrm{diff}) + 1 & \text{if} \quad 0 \leqslant \mathrm{diff} < 9.0 \\ 10 & \text{if} \quad \mathrm{diff} \geqslant 9.0. \end{cases} \tag{8}$$

The RI assignment is a useful indication of the level of certainty in the prediction for a particular sequence. Figures 2 and 3 show the statistical results for prokaryotic sequences. Similar curves were obtained for the eukaryotic case (data not shown). The expected prediction accuracy with RI equal to a given value and the fraction of sequences for each given RI were calculated (Figure 2). For example, the expected accuracy for a sequence with RI = 3 is 91% with 14% of all sequences having RI = 3. The average prediction accuracy was also calculated for RI above a given cut-off (Figure 3). About 78% of all sequences have RI $\geqslant$ 3 and of these sequences about 95.5% were correctly predicted by the SubLoc system.

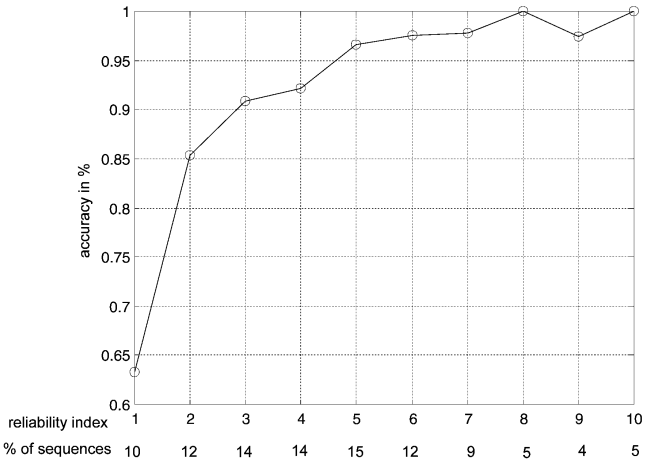

Fig. 2. Expected prediction accuracy with a reliability index equal to a given value. The fractions of sequences that are predicted with RI = n, $n = 1, 2, \ldots, 10$ are also given.

Robustness to errors in the N-terminal sequence

Some evidence has indicated that a method based on amino acid composition would be more robust to errors in the gene 5′-region annotation, i.e. the protein N-terminal sequence (Reinhardt and Hubbard, 1998) than methods based on sorting signals. Our results support this suggestion. We removed N-terminal segments which lengths of 10, 20, 30 and 40, respectively, from full protein

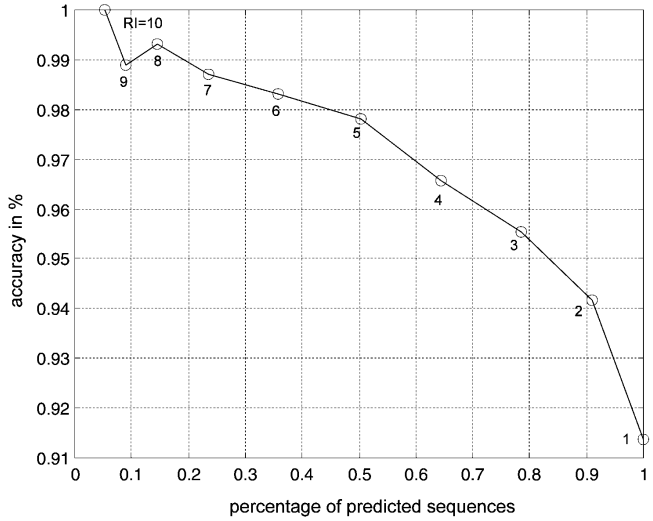

Fig. 3. Average prediction accuracy with a reliability index above a given cut-off. For example, about 75% of all sequences have RI $\geqslant 3$ and of these sequences about 95% are correctly predicted with the SubLoc system.

Table 6. Performance comparisons for the prokaryotic sequences with one segment of N-terminal sequence removed

	Accuracy (%)				MCC		
	Total	Cyto	Peri	Extra	Cyto	Peri	Extra
COMPLETE	91.3	97.8	76.2	77.6	0.85	0.77	0.78
CUT-10	91.5	90.6	77.3	78.6	0.86	0.78	0.78
CUT-20	90.6	96.5	77.2	77.6	0.85	0.75	0.76
CUT-30	91.1	97.0	77.8	78.5	0.86	0.76	0.77
CUT-40	90.1	96.4	74.8	78.5	0.84	0.73	0.77

COMPLETE: prediction performance for the complete sequences; CUT-10: prediction performance for the remaining sequence parts when 10 N-terminal amino acids were removed; CUT-20, CUT-30 and CUT-40 have similar meanings. Cyto, Peri and Extra are short for Cytoplasmic, Periplasmic and Extracellular, respectively.

sequences, then trained the SVM classifiers using the remaining parts of the sequences. Only the results of the 5-fold cross validation were given instead of the jackknife test because of the limited computational power. As mentioned before, the results by these two testing procedures are so close that the variations of the prediction accuracies with the removed segment lengths could be accurately reflected by the 5-fold cross validation results. The results for prokaryotic and eukaryotic sequences are summarized in Tables 6 and 7. The results indicate that the accuracies changed little for both the prokaryotic and eukaryotic cases. When even 40 amino acid segments were removed, the total accuracies were only reduced 1.2% for prokaryotic sequences and 3% for eukaryotic sequences. Predictions based on sorting signals would not be very reliable if this important information in the N-terminal sequence was missing.

DISCUSSION AND CONCLUSION

SVM information condensation

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, the Support Vectors (SVs). The SVs characterize the solution to the problem in the following sense: if all the other training samples are removed and the SVM is 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 quite small compared to the total number of training samples. This is a crucial property when analyzing large datasets containing many uninformative patterns which will be especially useful in the bioinformatics field as the mass of experimental data explodes. Table 8 shows the number of SVs for each binary classifier for the 977 prokaryotic sequences using the RBF kernel or the polynomial kernel. The results show that for this classification task, the ratio of SVs to all training samples is in the range of 13–30%.

SVM parameters selection

SVM still has a few tunable parameters which need to be determined. SVM training includes the selection of the proper kernel function parameters and the regularization parameter C. The selection of the kernel function parameters is very important because they implicitly define the structure of the high dimensional feature space where the maximal margin hyperplane is found. The regularization parameter C controls the complexity of the learning machine to a certain extent and influences the training speed. Although successful theoretical methods are not available for parameter selection, the accuracy of the subcellular localization prediction is not sensitive to this selection. The results in Tables 2 and 3 show that almost the same accuracies were obtained with different kernel types. Furthermore, large variations of the parameters including γ for the RBF kernel, degree d for the polynomial kernel and the regularization parameter C, had little influence on the classification performance (see the supplementary material). In addition, the results in Table 8 indicated that almost the same Support Vectors were used in SVMs with different kernels. This important phenomenon was previously observed by Vapnik (1995). If so, the set of SVs could be considered as a robust characteristic of the dataset.

Table 7. Performance comparisons for the eukaryotic sequences with one segment of N-terminal sequence removed

	Accuracy (%)					MCC			
	Total	Cyto	Extra	Mito	Nuclear	Cyto	Extra	Mito	Nuclear
COMPLETE	78.3	76.7	77.2	56.4	86.0	0.64	0.77	0.55	0.73
CUT-10	77.2	74.0	77.8	52.7	86.1	0.62	0.77	0.50	0.73
CUT-20	76.3	73.2	78.5	51.4	84.8	0.61	0.76	0.50	0.71
CUT-30	76.1	72.5	76.3	50.5	85.8	0.60	0.73	0.48	0.72
CUT-40	75.3	71.5	74.2	46.7	86.3	0.58	0.71	0.46	0.72

COMPLETE: prediction performance for the complete sequences; CUT-10: prediction performance for the remaining sequence parts when 10 N-terminal amino acids were removed; CUT-20, CUT-30 and CUT-40 have similar meanings. Cyto, Extra and Mito are short for Cytoplasmic, Extracellular and Mitochondrial, respectively.

Combining with other methods and incorporating other features

Several ways may improve the prediction performance. Single prediction methods have limitations. For instance, the methods based on sorting signals are sensitive to errors in the N-terminal sequence. The methods including the SubLoc system based on composition can not effectively classify sequences with similar amino acid compositions. The mitochondrial sequences were not well predicted by the SubLoc system (Table 3) while Yuan's method effectively predicted these sequences, possibly due to the similar amino acid compositions between the mitochondrial and cytoplasmic sequences. In addition, as pointed out by Nakai (2000), isoforms can not be well localized by the methods based on composition. Therefore, a combination of complementary methods may improve the accuracy.

Another strategy is to incorporate other informative features. The methods mentioned above all use classification information derived from protein sequences. More recently, other useful classification information for location has been investigated. Drawid and Gerstein (2000) have localized all the yeast proteins using a Bayesian system integrating features in the whole genome expression data. Murphy *et al.* (2000) analyzed the locations using information from fluorescence microscope images. As pointed out previously, SVM can easily deal with high dimensional data so the SVM method can easily incorporate other useful features which may improve the prediction accuracy.

In conclusion, a new method for protein subcellular localization prediction is presented. This new approach provides superior prediction performance compared with existing algorithms based on amino acid composition and can be a complementary method to other existing methods based on sorting signals. Furthermore, predictions by the SVM approach are robust to errors in gene 5′-region annotation. It is anticipated that the current prediction method would be a useful tool for the large-scale analysis of genome data.

Table 8. Number of the Support Vectors for various kernel functions. The total number of prokaryotic samples was 997. The kernel functions were RBF with $\gamma = 5.0$ and the polynomial function with degree $d = 9$

Binary classifier	RBF	Polynomial	Shared SVs	Union
Cyto/ $\sim$ Cyto	199	172	131	240
Peri/ $\sim$ Peri	303	237	192	348
Extra/ $\sim$ Extra	126	126	89	163

Cyto/ $\sim$ Cyto: the SVM trained with all of cytoplasmic sequences with positive labels and all other sequences with negative labels; Peri/ $\sim$ Peri and Extra/ $\sim$ Extra have similar meanings. Shared SVs: the number of shared SVs for both kernel functions; Union: the total number of SVs for both kernel functions.

ACKNOWLEDGEMENTS

The authors would like to thank Dr A.Reinhardt (Wellcome Trust Genome Campus, Hinxton, UK) for providing the dataset. This work was supported by a National Natural Science Grant (China) (No. 39980007) and partially by a National Key Foundational Research Grant (985) and a TongFang Grant.

REFERENCES

Andrade,M.A., O'Donoghue,S.I. and Rost,B. (1998) Adaption of protein surfaces to subcellular location. *J. Mol. Biol.*, **276**, 517–525.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C.W., Furey,T.S., Ares,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Cedano,J., Aloy,P., Perez-Pons,J.A. and Querol,E. (1997) Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol.*, **266**, 594–600.

Chou,K.C. and Elrod,D. (1999) Protein subcellular location prediction. *Protein Eng.*, **12**, 107–118.

Cortes,C. and Vapnik,V. (1995) Support vector networks. *Mach. Learn.*, **20**, 273–293.

Drawid,A. and Gerstein,M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing pro-

teins: comprehensive application to the yeast genome. *J. Mol. Biol.*, **301**, 1059–1075.

Drucker,H., Wu,D. and Vapnik,V. (1999) Support vector machines for spam categorization. *IEEE Trans. Neural Netw.*, **10**, 1048–1054.

Eisenhaber,F. and Bork,P. (1998) Wanted: subcellular localization of proteins based on sequence. *Trans. Cell Biol.*, **8**, 169–170.

Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.*, **300**, 1005–1016.

Frishman,D., Mironov,A. and Gelfand,M. (1999) Starts of bacterial genes: estimating the reliability of computer predictions. *Gene*, **234**, 257–265.

von Heijne,G., Nielsen,H., Engelbrecht,J. and Brunak,S. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.*, **10**, 1–6.

Hirokawa,T., Boon-Chieng,S. and Shigeki,M. (1998) SOSUI: classification and secondary structure prediction system for membrane proteins. *Bioinformatics*, **14**, 378–379.

Hua,S.J. and Sun,Z.R. (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J. Mol. Biol.*, in press.

Joachims,T. (1999) Making large-scale SVM learning practical. In Scholkopf,B., Burges,C. and Smola,A. (eds), *Advances in Kernel Methods-Support Vector Learning*. MIT Press, Cambridge, MA, pp. 42–56.

Lio,P. and Vannucci,M. (2000) Wavelet change-point prediction of transmembrane proteins. *Bioinformatics*, **16**, 376–382.

Matthews,B.W. (1975) Comparison of predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Murphy,R.F., Boland,M.V. and Velliste,M. (2000) Towards a systematics for protein subcellular location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 251–259.

Nakai,K. (2000) Protein sorting signals and prediction of subcellular localization. *Adv. Protein Chem.*, **54**, 277–344.

Nakai,K. and Horton,P. (1997) Better prediction of protein cellular localization sites with the *k* nearest neighbors classifier. *Intell. Syst. Mol. Biol.*, **5**, 147–152.

Nakai,K. and Kanehisa,M. (1991) Expert system for predicting protein localization sites in Gram-negative bacteria. *Proteins: Struct. Funct. Genet.*, **11**, 95–110.

Nakai,K. and Kanehisa,M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.

Nakashima,H. and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Biol.*, **238**, 54–61.

Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal perptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.

Nielsen,H., Brunak,S. and von Heijne,G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng.*, **12**, 3–9.

Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res.*, **26**, 2230–2236.

Roobaert,D. and Hulle,M.M. (1999) View-based 3D object recognition with support vector machines. In Hu,Y.H., Larsen,J., Wilson,E. and Douglas,S. (eds), *Proceedings of the IEEE Neural Networks for Signal Processing Workshop*. IEEE Press, Totowa, NJ, pp. 77–84.

Rost,B. and Sander,C. (1993) Prediction of secondary structure at better than 70% accuracy. *J. Mol. Biol.*, **232**, 584–599.

Rost,B., Fariselli,P. and Casadio,R. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Sci.*, **5**, 1704–1718.

Scholkopf,B., Burges,C. and Vapnik,V. (1995) Extracting support data for a given task. In Fayyad,U.M. and Uthurusamy,R. (eds), *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Menlo Park, CA, pp. 252–257.

Vanderbei,R.J. (1994) Interior point methods: algorithms and formulations. *ORSA J. Comput.*, **6**, 32–34.

Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.

Yuan,Z. (1999) Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.*, **451**, 23–26.