Residue-couple Model for Protein Subcellular Loaclization Prediction

Article

2 authors, including:

Zhirong Sun
Tsinghua University
95 PUBLICATIONS **3,445** CITATIONS

SEE PROFILE

Residue-couple Model for Protein Subcellular Loaclization Prediction

Jian Guo [1, 2] and Zhirong Sun [2,*]

1. Department of Mathematical Science, Tsinghua University, Beijing100084, People's Republic of China.

2. Institute of Bioinformatics, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, People's Republic of China.

* To whom correspondence should be address

email: sunzhr@mail.tsinghua.edu.cn

guojian1999@mails.tsinghua.edu.cn

ABSTRACT

As a key functional characteristic of protein, subcellular localization performs an important role in genome analysis. Therefore, an automatic, reliable and efficient prediction system for protein subcellular localization is needed for large-scale genome analysis. In this paper, we construct a new model (residue-couple model) and use support vector machine under this model frame for subcelluar localization. In addition of the traditional amino acid composition model, residue-couple model incorporates the effect of sequence order. The total accuracy of prediction reached up to 92.0% for prokaryotic protein sequences and 86.9% for eukaryotic protein sequences under 5-fold cross validation, which represents a significant improvement compared with the precede methods. We also prove that our model is robust to the errors of N-terminal in sequences.

Keywords: Subcellular localization; Residue-couple model; N-terminal sequence; Support vector machine;

INTRODUCTION

A great amount of genome sequences have been produced through high throughout experiments. The next step is to analysis these genome and protein sequences begging for finding new gene functions and key regulatory pathways. As a hot topic in genome science, genome function annotation including the assignment of a function for a potential gene in the raw sequence is a vitally important work in genome research. Subcellular location is a key function characteristic of potential gene which produce this protein because the function of a protein is closely correlated with its subcellular location. As a result, the knowledge of protein subcellular location plays a very important role in gene function prediction, which is useful in cell biology, molecular biology, pharmacology and medical science. Subcelluar localization analysis based on experiment is time consuming and costly. With the rapidly increasing number of sequences in database, it is highly necessary to develop an accurate, reliable and efficient system for protein subcellular prediction automatically.

Several efforts have been made in this regard. Up to now there are mainly two categories of methods for subcellular localization prediction. One is mainly based on the existence of sorting signals in N-terminal sequences (Nakai, 2000), which include signal peptides, mitochondrial targeting peptides and chloroplast transit peptides (Nielsen et al, 1997, 1999). Emanuelsson et al proposed an integrated prediction system with artificial network based on individual sorting signal predictions. This system can be use to find cleavage sites in sorting signals and can simulate the real sorting process to a certain extent. Nevertheless, The prediction accuracy of those methods based on sorting signals is highly correlated with the quality of protein N-terminal sequence assignment. Unfortunately, it is usually unreliable that annotate the N-terminal using known gene identification methods (Frishman et al., 1999). As a result, the prediction accuracy and reliability will decrease when signals are missing or only partially included.

The other category of methods is mainly based on the amino acid composition of protein sequences in different subcellular localizations. This approach is first suggested by Nakashima and Nishikwa (1994). They found that the intracellular and the extra cellular proteins could be discriminated with high accuracy by amino acid composition only. From then on, different statistical methods and machine learning methods have been used based on amino acid composition of protein sequences to improve prediction accuracy. Cedano et al (1997) adopted a statistical method with Mahalanobis distance for prediction. Reinhardt and Hubbard (1998) predicted subcellular locations with neural networks and reached the accuracy 66% for eukaryotic sequences and 81% for prokaryotic sequences. Chou et al (1999) proposed the covariant discriminant algorithm on the same prokaryotic dataset as Reinhardt et al. and achieved a total accuracy of 87%. Hua and Sun (2001) constructed a prediction system using support vector machine (SVM)——a new machine learning method based on the statistical learning theory——on the same prokaryotic and eukaryotic datasets. The prediction accuracy of Hua et al has reached up to 91.4% for prokaryotic proteins and 79.4% for eukaryotic proteins. Nevertheless, it will lose fairly information when protein sequences are decomposed into the amino acid composition. To overcome this fault, several methods appeared to combine the information of the amino acid composition with the information related to other biological knowledge. Nakai et al constructed an expert system based on both sorting signal and amino acid composition (1992, 1997). Chou (2000) and Feng (2001) cooperated the hydrophobicity index of residues pair into the prediction system and used Bayes Discriminate Function as a prediction tool. Yuan (1999) used a markov model which use the information not only from amino acid composition but also from sequence-order information.

In this paper, we develop a new model based on residue-couple model and SVM for subcellular localization prediction. Residue-couples contain the information not only about amino acid composition, but also about the order of amino acids in protein sequences, which seems also important for subcellular localization. We used these residue-couples to train SVM classifiers for prediction. Consequently, the overall prediction accuracy under a 5-fold cross validation test is as high as 86.9% for eukaryotic proteins and 92.1% for prokaryotic proteins. The result represents a further step towards the practical prediction of protein subcellular localization.

METHOD AND DATABASE

Database

In this paper, we choose the database generated by Reinhardt and Hubbard (1998) to test our new model. It is a common used subcellular localization prediction dataset. The sequences in this database are extracted from SWISSPORT 33.0 and subcellubar location of each protein has been annotated. This set of sequences was filtered, only keeping those appeared complete and those had what appeared to be reliable location annotations. Transmembrane proteins are also excluded because some reliable prediction methods for these proteins have already existed (Rost, B. et al, 1996). Plant sequences are also removed for the sufficient difference of the composition. Finally, the filtered dataset include 997 prokaryotic proteins (688 cytoplasmic, 107 extracellular and 202 periplasmic proteins) and 2427 eukaryotic proteins (684 cytoplasmic, 325 extracellular, 321 mitochondrial and 1097 nuclear proteins).

Classifier and Support Vector Machine

SVM is a new statistical learning algorithm for pattern prediction and regression analysis. We only introduce the basic idea of SVM here. Any reader who wants to learn more details about SVM can read Vapnik's publications (1995, 1998).

Let's consider a two-class classification problem first. Assuming that there is a set of samples which can be shown as a number of input vectors with corresponding labels $y_i \in \{-1, +1\}$ (i=1,2,.....,N). We divide these samples into

two sets, one set is used for training and the other set is used for testing. Our goal is to construct a binary classifier base on the training set and to classify the test set as correctly as possible.

SVM is a margin classifier, which defines a boundary that maximizes the margin between samples in two classes. In the case that the samples cannot be separated in the original feature space, the SVM maps the input vectors into a space with high dimension (the mapped high dimensional space is called kernel space and the original space is called feature space) to construct a decision boundary that maximize the margin. Generally, the performance of the map from feature space to kernel space is difficult to be described with formula, therefore, we use the inner product of two samples point in kernel space instead of writing the performance of Φ. The inner product is called kernel function and denoted as $K(\vec{x_i}, \vec{x_j})$. The three common used kernel functions were defined as follows:

$$K(\vec{x_i}, \vec{x_j}) = (<\vec{x_i}, \vec{x_j}> +1)^d \tag{1}$$

$$K(\vec{x_i}, \vec{x_j}) = \exp(-\gamma \| \vec{x_i} - \vec{x_j} \|^2) \tag{2}$$

$$K(\vec{x_i}, \vec{x_j}) = \tanh(\upsilon <\vec{x_i}, \vec{x_j}> -\Theta) \tag{3}$$

Equation (1) is called the polynomial function of degree d. Equation (2) is the Radial Basic Function (RBF) kernel with parameter γ. Equation (3) is the muti-layer perceptron kernel with parameter γ and Θ. In this paper, we used RBF kernel only because a great number of experiments have proved that RBF kernel performed better than other two kernels in most of the experiments.

Utilizing the kernel function, the SVM binary classification problem can be convert to a Quadratic Programming problem (QP):

$$Maximize \qquad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K(\vec{\mathbf{x}_i}, \vec{\mathbf{x}_j})$$

$$subject \quad to \qquad 0 \le \alpha_i \le C \tag{4}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0 \qquad i = 1, 2, \ldots, N$$

Solve the QP problem to achieve the α_i (i=1,2,……,N) . Then we use the decision function for prediction:

$$f(\vec{\mathbf{x}}) = \text{sgn}(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\vec{\mathbf{x}}, \vec{\mathbf{x}_i}) + b) \tag{5}$$

Muticlass problem and Implement of SVM

Protein subcellular localization prediction is a typical multiclass classification problem, where the class number is 3 for prokaryotic proteins and 4 for eukaryotic proteins. A simple idea to solve mutliclass problem is to divide it into a number of binary classification problems. There are several strategies to implement the idea, including the common used "one-against-rest" and "one-against-one".

For a k-class problem, "one-against-one" strategy construct k•(k-1) classifiers and each one trains the data from two different classes. The final decision was based on the voting strategy, i.e., the test sample will be classified into the class chose by most binary classifiers. For the one-against-rest strategy, k classifiers are constructed and the ith classifier is trained with all of the examples in the ith class with positive labels, and all the other examples are

denoted with negative labels. Each test sample will be classified into the class that maximizes the distance between the test sample and the binary boundary.

The software which we used to implement SVM in this paper is osm_svm3.00 made by junshui ma et al. It is a matlab SVM toolbox for both classification and regression problem. The core part of this toolbox is based on LIBSVM2.33 (a widely used SVM software) and it is fast enough to handle with large-scale practice data. The multi-classification strategy used in osu_svm3.00 is based on "one-against-one" scheme. In this paper, we write a matlab program to use the APIs provided by osu_svm3.00 toolbox.

Residue-Couple Model

The traditional subcellular localization prediction model is mainly based on amino acid composition model. Nevertheless, it is obvious that amino acid composition model ignores amount of information of protein sequence. Unfortunately, it is difficult to directly incorporate the information of sequence order effect into a pattern recognition model for prediction because of the huge number of possible sequence order patterns (Chou 2001). However, inspired by Chou's Quasi-Sequence-Order model and Yuan's markov chain model, we developed a new model that can utilize the sequence order effect indirectly.

Denote a protein sequences as a serial of letters:

$$R_1 R_2 R_3 R_4 R_5 R_6 R_7 \cdots\cdots R_N$$

where R_i represents the amino acid in residue i. Define "residue-couple" as follows:

$$X_{i,j}^1 = \frac{1}{N-1} \sum_{n=1}^{N-1} H_{i,j}(n, n+1)$$

$$X_{i,j}^2 = \frac{1}{N-2} \sum_{n=1}^{N-2} H_{i,j}(n, n+2)$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$X_{i,j}^k = \frac{1}{N-k} \sum_{n=1}^{N-k} H_{i,j}(n, n+k) \tag{6}$$

$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$

$$X_{i,j}^m = \frac{1}{N-m} \sum_{n=1}^{N-m} H_{i,j}(n, n+m), \qquad \text{m<N, I=1,2,}\cdots\text{,20 and j}=\text{1,2,}\cdots\text{,20}$$

where $H_{i,j}(n, n+k)$ =1, if the amino acid in residues n is i and the one in residues n+k is j; otherwise $H_{i,j}(n, n+k)$ =0.(see Figure 1) The value scope of i and j range from 1 to 20, representing the 20 different amino acids (briefly denoted as A, C, D, E, F, G, H, I, J, K, L, M, N, P, Q, R, S, T, V, W, Y). $X_{i,j}^1$ ($i, j = 1,2,\cdots,20$) is called the 1st-rank residue-couple that represent the frequency that a mode of continuous residue pairs observed in a protein sequence. $X_{i,j}^2$ is called the 2nd-rank residue-couple that represents the frequency that the couple mode $(i, \_, j)$ is observed in a protein sequence ("_" represents any type of amino acid). $X_{i,j}^k$ is called the kth-rank

residue-couple that represents the frequency that the couple mode ($i, \_, \_, j$) is observed in a protein sequence, and so forth. It is obvious that there were $20 \times 20 = 400$ residue-couples in each rank (see Figure 1).

For each protein sequence, we combined all residue-couples into a vector $\vec{x}$, that is, the first 400 components of $\vec{x}$ is the 400 1st-rank residue-couples and the following 400 components is the 400 2nd-rank residue-couples, and so forth. Therefore, the final vector has $400 \times m$ dimension. The value of m is called "coupling-degree", representing the total rank of residue-couples. This model contains the information of both amino acid composition and the order effect of protein sequence. We handled each protein sequence as above and got a set of $400 \times m$ dimension vectors (each vector corresponds to one vector). The set of vectors was used as the input vectors of the support vector machine for training and prediction (see Figure 2).

Cross-validation and model selection

In this paper we used 5-fold cross validation for test because of our limited computational power. In the process of k-fold cross validation, the entire sample set was divided into k equally sized subsets randomly. In each turn, one subset was used as the test set and the other k-1 subsets were used for training SVM. The final prediction result will combine the results in each turn.

Prediction Result Assessments

The total prediction accuracy, the prediction accuracy in each location and the Matthew's Correlation Coefficient (MCC) were used to assess of the prediction result.

Denote M_{ij} as the number of proteins observed in location i and predicted in location j, then the total number of proteins observed in state i is $obs_i = \sum_{j=1}^{k} M_{ij}$, where k is the number class. The total number of proteins predicted in state i is $pre_i = \sum_{j=1}^{k} M_{ji}$.

The total prediction accuracy and the prediction accuracy in location i are defined as follows:

$$Total\_Accuracy = \frac{\sum_{i=1}^{k} M_{ii}}{N} \tag{7}$$

$$Accuracy(i) = \frac{M_{ii}}{obs(i)} = \frac{M_{ii}}{\sum_{j=1}^{k} M_{ij}} \tag{8}$$

Matthew's Correlation Coefficient is defined as follows:

$$MMC_i = \frac{p_i \cdot n_i - u_i \cdot o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(n_i + u_i)(n_i + o_i)}} \tag{9}$$

$$p_i = M_{ii} \qquad n_i = \sum_{j \neq i}^{3} \sum_{k \neq i}^{3} M_{jk}$$

$$o_i = \sum_{j \neq i}^{3} M_{ji} \qquad u_i = \sum_{j \neq i}^{3} M_{ij}$$

where p_i is the number of correctly predicted sequences in location i, n_i is the number of correctly predicted sequences do not in location i, u_i is the number of under-predicted sequences and o_i is the number of over-predicted sequences.

RESULT

Prediction Accuracy

Table 1 shows the prediction accuracy of our method for euokaryotic proteins with different coupling-degrees of input vectors. The result was based on a 5-fold cross-validation test. From Table 1 we can find that the accuracy was improved with the increase of coupling-degree and achieved the total accuracy 86.9% when the value of coupling-degree was equal to 6 and the kernel parameter γ is 20. As a matter of fact the total accuracy had a great improvement while the coupling-degree was more than 2. The accuracies of different subcellular locations were also typed in the table.

Table 2 shows the accuracies of our method for prokaryotic proteins with different compling-degrees of input vectors. Only five different compling-degrees were typed in the table as we find that the result changed slight with the increasing of compling-degrees.

Comparison with other methods

The prediction result of our method is compared with the result of other subcellular localization prediction methods. For the eukaryotic sequences, the neural network method (Reihardt and Hubbard,1998), the Markov model (Yuan,1999) and the SVM method (Hua, 2001) are compared with the residue-couple model introduced in this paper (see Table 5). The results show that the total accuracy of the residue-couple model is 20.9% higher than that of the neural network method and 7.5 higher than that of the SVM method. For cytoplasmic and nuclear sequences, the prediction accuracies were 30.8% and 22% higher than the neural network method and 8.9% and 6.8% higher than the SVM method. The result of our model is obviously higher than that of Hua's method, although both of us use the same classifier——support vector mahine. This fairly reflects that the residue-couple model mined more useful information from protein sequences than amino acid composition model, especially for cyotoplasmic and mitochondrial sequences (8.9% and 8.7 higher than Hua's work).

Both the residue-couple model and the markov model (Yuan, 1999) used the sequence order information for prediction. The total accuracy of our work was 13.9% higher than that of the markov model. The accuracy for extracellular and nuclear is 23.7% and 20.1% higher than those of the markov model method, although the accuracy for mitochondrial is 3.8% lower (nevertheless, the MCC of residue-couple model for mitochondrial is 0.72, much higher than that of markov model). Although the two methods are both based on the residue order information, the powerful classification capability of SVM helps us to achieve higher accuracies.

The results of the MCC in different methods are also shown in Table 5. The MCC of each subcellular location using residue-couple model is higher than all other corresponding one in the Table 5.

For the prokaryotic sequences, the comparison results were showed in Table 6. The total accuracy of the residue-couple model was about 11% higher than neural network and 5.5% higher than covariant discriminant algorithm. The accuracy for cytoplasmic sequences reached up to 99%, although the total accuracy has no evident

improvement compared with the Hua's method (only 0.6% higher than Hua's method).

Robustness to errors in the N-terminal sequence

Our model is much more robust to errors in protein-terminal sequence than those methods based on sorting signals. In order to prove this conjecture, we removed N-terminal segments with length of 10, 20, 30 and 40 amino acids from the protein sequences, and retrained the SVM classifiers with the remained part of each sequence. The results for eukaryotic sequences and for prokaryotic sequences are shown in Table 3 and Table 4. From the tables we can find that the total accuracy reduced only 3.2% for eukaryotic sequences and 1.1% for prokaryotic sequences even though 40 residues in the N-terminal are removed.

DISCUSSION AND FURTHER WORK

The method based on the residue-couple model and SVM classifier performed powerful in subcellular localization prediction. Compared with other methods, the advantage of our method is much more evident for eukaryotic protein sequences than for prokaryotic sequences. The total accuracy of our method exceeded only 0.6% for prokaryotic protein sequences in comparison with that of Hua's method (see Table 6). However, this index improves 7.5% for eukaryotic proteins (see Table 5). As a matter of fact, the prokaryotic proteins have been classified with high accuracy even use linear classifiers based on amino acid composition only (the total accuracy achieved 89.3% with linear kernel SVM, Hua et al, 2001). This result probably reflects that the amino acid composition is the key characteristic of prokaryotic proteins, which have relative simple sequences structure and relative simple biology function. However, eukaryotic protein sequences seem much more complex than prokaryotic sequences and amino acid composition doesn't contain enough information to prediction protein location. Therefore, for eukaryotic proteins, the accuracies of the methods based on amino acid composition model perform obviously lower than out method based on residue-couple, which not only reflects the information from both amino acid composition, but also reflects the information of sequence order.

In recent further, we will centralize on three aspects to improve our work. One is to combine other complementary methods. Mitochondiral proteins was still not well predicted (65.4%), although the accuracy has been higher than that of all other prediction methods (Table 5) except the markov model. 19% of the proteins that belong to mitochondiral were incorrectly classified into the location of cytoplasmic. Similar conclusion was also reported by Hua (2001). This means that it is difficult to discriminate the proteins in cytoplasmic and mitochondiral only based on residue-couple information. Due to the relative high prediction accuracy for mitochondiral proteins using markov model (69%), a logical procedure was to combine the markov model and the residue-couple model throughout some proper strategies, which is right one of the plans in our future research. Combination methods base on sorting methods is also under consideration.

The second aspect in future work is to incorporate other informative features, including gene expression profile (Drawid and Gerstein,2000;Murphy et al,2000) and regulatory pathway information. Some information fusion technologies, such as the meta learning methods may be used for combination of information from different dataset and different types of format.

The third aspect is to improve the SVM classifiers, including how to select more proper kernels, how to speed up our prediction system and how to filter noise and outliers. Several papers have introduce new methods to the solving the noise and outliers problems (Zhang,X., 1999). Some new SVM software such as Herosvm can speed up the running process for hundreds of times. We are also attempting to combine active learning stratrgy with SVM for further improvement.

CONCLUSION

In this paper, we developed a new method for subcellular localization based on the residue-couple model, which does not only contain the information of amino acid composition, but also reflected the effect of the residue order in sequences. The high accuracies for both protaryotic (92.0%) and eutarkoytic sequences (86.9%) proved that our method performed well compared with other methods for subcellular location prediction. Furthermore, our method was robust to the errors in the N-terminal of sequences. In conclusion, we have constructed a more powerful system for subcellular localization prediction and it would be a useful tool for large-scale protein function analysis.

ACKNOWLEDGEMENT

REFERENCES

Andrade,M.A., O'Donoghue,S.I. and Rost,B. (1998) Adaption of protein surfaces to subcellular location. *J. Mol. Boil.*, **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. Boil.*, **266**, 594-600.

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

Chou.K.C. (2000) Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Oder Effect. *Biochem. biophys. res. commun*. **278**, 477-483

Drawid,A., and Gerstein,M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the Yeast genome. *J. Mol. Boil*. **301**, 1059-1075.

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. Boil.*, **300**, 1005-1016.

Feng Z. and Zhang C.T. (2001) Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. *Int. j. biol. macromol.* **28,** 225-261.

Feng Z. (2001) Prediction of the Subcellular Location of Prokaryotic Proteins Based on a New Representation of the Amino Acid Composition. *Bioploymers*, **58**, 491-449.

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

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. Boil*. **308**, 397-407

Hua,S.J. and Sun,Z.R (2001) Support vector machine approach for protein subcellular localization prediction. *Bioinformatics*, **17**, 721-728.

Murphy, R.F, Boland, M.V. and Velliste, M. (2000) Towards a system 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. and Kanehisa,M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells.

Genomics, **14**, 897-911.

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

Nakai,K. (2000) Protein sorting signals and prediction of subcellular localization. *Advances in Protein Chemistry*, **54**, 277-344.

Nakashima,H., and Nishikawa,K. (1994) Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J. Mol. Boil.*, **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 Sys.*, **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. *Nucl. Acids Res.,* **26**, 2230-2236.

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

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

Vapnik,V. (1998) *Statistical Learning Theory.* John Wiley and Sons, Inc., New York.

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.

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

Zhang,Z.(1999) Using class-center vectors to build support vector machines. *Proceeding of the 1999 IEEE Signal Processing Society Workshop*, 3-11, IEEE Press, NJ.

Table 1: Prediction accuracies for eukaryotic sequences with different coupling-degrees

		Coupling-degree							
		1	2	3	4	5	6	7	8
Total accuracy (%)		80.4	85.5	86.5	85.9	86.5	***86.9***	86.7	86.6
Accuracy	Cytoplasmic	79.5	84.2	85.0	84.5	86.0	85.8	85.7	***86.4***
	Extracellular	79.7	80.0	84.3	81.0	85.0	***85.9***	83.4	82.8
	Mitochondrial	54.2	60.8	64.5	58.9	63.9	***65.4***	63.0	62.0
	Nuclear	88.7	95.1	94.6	***96.1***	93.9	94.2	95.3	95.0
γ		100	100	50	50	20	20	20	20

γ is the paremeter in equation (2). We test a number of different values of y in each coupling-degree with 5-fold cross validation and choose the best one. In the training process we use Radial Basic Function only.

Table 2 : Prediction accuracies for prokaryotic sequences with different coupling-degree

		Coupling-degree				
		1	2	3	4	5
Total accuracy		90.7	91.3	91.2	91.5	_**92.0**_
Accuracy	Cytoplasmic	_**99.1**_	98.4	98.1	99.0	99.0
	Periplasmic	70.1	76.6	_**78.5**_	73.8	77.6
	Extracellular	72.8	74.8	74.8	75.3	_75.7_
γ		100	100	100	100	100

γ is the paremeter in equation (2). We test a number of different values of y in each coupling-degree with 5-fold cross validation and choose the best one. In the training process we use Radial Basic Function only.

Table 3 : performance comparisons for the eukaryotic protein sequences with one segment of N-terminal sequences removed

	Accuracy (%)					MCC			
	Total	Cyto	Extra	Mito	Nuclear	Cyto	Extra	Mito	Nuclear
COMPLETE	86.9	85.8	85.9	65.4	97.2	0.77	0.89	0.72	0.85
CUT-10	85.7	84.2	83.7	62.3	94.1	0.76	0.88	0.69	0.84
CUT-20	85.0	83.5	82.5	58.6	94.4	0.75	0.86	0.66	0.83
CUT-30	84.4	83.2	81.9	56.1	94.3	0.75	0.86	0.63	0.83
CUT-40	83.7	82.5	80.3	53.0	94.4	0.73	0.86	0.61	0.82

COMPLETE: Prediction with complete sequences; CUT-10: Prediction for the rest part of sequences when 10 N-terminal amino acids were excluded; CUT-20, CUT-30, CUT-40 have similar meanings.

Table 4: performance comparisons for the prokaryotic protein sequences with one segment of N-terminal sequences removed

	Accuracy (%)				MCC		
	Total	Cyto	Extra	Peri	Cyto	Extra	Peri
COMPLETE	92.0	99.0	77.6	75.4	0.89	0.79	0.78
CUT-10	91.8	98.8	78.5	74.8	0.88	0.80	0.78
CUT-20	91.3	98.8	77.6	72.8	0.88	0.79	0.75
CUT-30	91.6	98.6	78.5	74.8	0.88	0.79	0.76
CUT-40	90.9	98.4	76.6	72.8	0.87	0.78	0.74

COMPLETE: Prediction with complete sequences; CUT-10: Prediction for the rest part of sequences when 10 N-terminal amino acids were excluded; CUT-20, CUT-30, CUT-40 have similar meanings.

Table 5: The comparisons of different prediction method for the eukaryotic sequences

Location	Neural network	Markov model		Amino acid composition +SVM		Residue-couple model +SVM	
	Accuracy (%)	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	55	78.1	0.60	76.9	0.64	_**85.8**_	_**0.77**_
Extracellular	75	62.2	0.63	80.0	0.78	_**85.6**_	_**0.89**_
Mitochondria	61	_**69.2**_	0.53	56.7	0.58	65.4	_**0.72**_
Nuclear	72	74.1	0.68	87.4	0.75	_**94.2**_	_**0.85**_
Total accuracy	66	73.0	--	79.4	--	_**86.9**_	--

The result of neural network model and residue-couple model are given by cross validation. The Markov model and SVM result were given by the jackknife (leave one out cross validation).

Table 6 : The comparisons of different methods for the prokaryotic sequences

Location	Neural network	Covariant discrimination	Markov model		Amino acid composition +SVM		Residue-couple model+SVM	
	Accuracy (%)	Accuracy (%)	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
Cytoplasmic	80	91.6	93.6	0.83	97.5	0.86	_**99.0**_	_**0.89**_
Extracellular	77	80.4	77.6	0.77	75.7	0.77	77.6	_**0.79**_
Periplasmic	_**85**_	72.7	79.7	0.69	78.7	0.78	75.7	0.78
Total accuracy	81	86.5	89.1	--	91.4	--	_**92.0**_	--

The result of neural network model and residue-couple model are given by cross validation. The Markov model and SVM result were given by the jackknife (leave one out cross validation).

Figure Legend

Figure 1: A schematic drawing to show the residue-couple with different rank: (a) the 1st-rank: the coupling mode between all the two consecutive residues. (b) the 2nd-rank: the coupling mode between two residues with only one amino acid between them. (c) the 3th-rank: the coupling mode between two residues with right two amino acid between them.
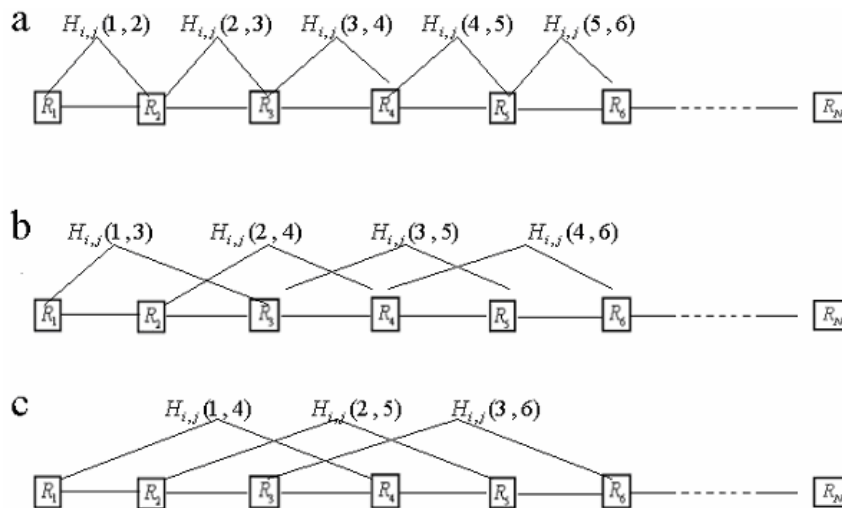
Figure 2: The prediction process of our method. The input vector of SVM is a number of $400 \times m$ dimension vectors.