# A NOVEL METHOD FOR PROTEIN SUBCELLULAR LOCALIZATION: COMBINING RESIDUE-COUPLE MODEL AND SVM

JIAN GUO

*Department of Mathematical Science, Tsinghua University,*
*Beijing 100084, China*

YUANLIE LIN

*Department of Mathematical Science, Tsinghua University,*
*Beijing 100084, China*

ZHIRONG SUN

*Institute of Bioinformatics, Tsinghua University,*
*Beijing 100084, China*

Subcellular localization performs an important role in genome analysis as a key functional characteristic of proteins. Therefore, an automatic, reliable and efficient prediction system for protein subcellular localization is needed for large-scale genome analysis. This paper describes a new residue-couple model using a support vector machine to predict the subcellular localization of proteins. This new approach provides better predictions than existing methods. The total prediction accuracies on Reinhardt and Hubbard's dataset reach 92.0% for prokaryotic protein sequences and 86.9% for eukaryotic protein sequences with 5-fold cross validation. For a new dataset with 8304 proteins located in 8 subcellular locations, the total accuracy achieves 88.9%. The model shows robust against N-terminal errors in the sequences. A web server is developed based on the method which was used to predict some new proteins.

## 1 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 molecular biology. Subcellular localization is a key functional characteristic of potential gene products such as proteins. However, experimental subcelluar localization analysis is time-consuming and can not be performed on genome scale proteins. With the rapidly increasing number of sequences in databases, an accurate, reliable and efficient system is needed to automate the prediction of protein subcellular locations.

Three primary types of methods have been used to predict the protein subcellular location in the previous published papers. One is based on the existence of sorting signals in N-terminal sequences (Nakai, 2000) including signal peptides, mitochondrial targeting peptides and chloroplast transit peptides (Nielsen et al, 1997, 1999). Emanuelsson et al. proposed an integrated prediction system using an artificial neural network based on individual sorting signal predictions. This system could be use to find cleavage sites in sorting signals and simulate the real sorting process to a certain extent. Nevertheless, the prediction accuracy of the methods based on sorting signals is highly dependent on the quality of the protein N-terminal sequence assignment. Unfortunately, it is usually unreliable to annotate the N-terminal using known gene identification methods (Frishman, 1999). As a result, the prediction accuracy and reliability decrease when signals are missing or are only partially included.

The second type of methods is mainly based on the amino acid composition of protein sequences in different subcellular locations. This approach was first suggested by Nakashima & Nishikwa. They found that the intracellular and the extracellular proteins could be accurately discriminated only by amino acid composition. Different statistical methods and machine learning methods have been used 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 accuracy levels of 66% for eukaryotic sequences and 81% for prokaryotic sequences. Chou et al. (1999) proposed a covariant discriminant algorithm using the same prokaryotic dataset as Reinhardt et al. and achieved a total accuracy of 87%. Hua & Sun (2001) constructed a prediction system using a support vector machine (SVM), a new machine learning method based on the statistical learning theory, using the same prokaryotic and eukaryotic datasets. The prediction accuracy of Hua and Sun's method was as high as 91.4% for prokaryotic proteins and 79.4% for eukaryotic proteins. However, in those models, the protein sequences were decomposed into animo acid compositions, which results in a great mount of information loss. To overcome this fault, several methods were introduced to combine the information of the amino acid composition with the information related to other biological data. Nakai et al. constructed an expert system based on sorting signals and amino acid composition (Nakai et al, 1992, 1997). Chou (2001) and Feng and Zhang (2001) added the hydrophobicity index of residue pairs into the prediction system and used the Bayes Discriminate Function as a prediction tool. Yuan (1999) used the Markov model, which considered the information not only from amino acid composition but also from sequence-order.

The third approach is to do a similarity search on the sequence, extract a text from homologs and use a classifier on the text features. Nair and Rost (2002) analyzed the relation between sequence similarity and identity in subcellular localization and construct the webserver LOCkey.

This paper presents a novel approach combining the residue-couple model and the SVM for subcellular localization prediction. Residue-couples contain information of the amino acid composition and the order of the amino acids in the protein sequences. The information is important for subcellular localization. These residue-couples were used to train the SVM classifiers. By using a 5-fold cross validation test, the overall prediction accuracies reach 86.9% for eukaryotic proteins and 92.1% for prokaryotic proteins. The results show that the prediction accuracy is significantly improved with the novel approach. To test the prediction on a real protein, a putative gene sequence was selected from GeneBank. The prediction results are consistent with experimental data.

## 2 Method and database

### 2.1 Database

The database generated by Reinhardt and Hubbard (1998), a commonly used subcellular localization dataset, was first used to test our new model. The sequences in this database were extracted from SWISSPORT 33.0 and the subcellubar location of each protein has been annotated. The set of sequences was filtered, keeping only those which appeared to be complete and those which appeared to have reliable location annotations. Transmembrane proteins were excluded because some reliable prediction methods for these proteins are already in existence (Rost et al 1996). Plant sequences were also removed to ensure a sufficient difference of the composition. The finally filtered dataset included 997 prokaryotic proteins (688 cytoplasm, 107 extracellular and 202 periplasmic proteins) and 2427 eukaryotic proteins (684 cytoplasm, 325 extracellular, 321 mitochondrial, and 1097 nuclear proteins).

A new much larger dataset, SL8304, was also constructed to further test the algorithm. The new database included 8304 eukaryotic proteins in 8 subcellular locations with 1019 chloroplast proteins, 2387 cytoskeleton proteins, 595 extracellular proteins, 211 Golgi proteins, 133 lysosomal proteins, 644 mitochondrial proteins, 3199 nuclear proteins and 116 peroxisomal proteins. All the proteins in this dataset were selected from SWISSPORT release 41 using the same selection rule as Reinhardt and Hubbard's dataset.

### 2.2 Classifier and support vector machine

The support vector machine (SVM) is a new machine learning method, which has been used for many kinds of pattern recognition problems. The principle of the SVM method is to transform the samples into a high dimension Hilbert space and seek a separating hyperplane in this space. The separating hyperplane, which is called the optimal separating hyperplane (OSH), is chosen in such a way as to maximize its distance from the closest training samples. As a supervised machine learning technology, SVM is well founded theoretically on Statistical Learning Theory. The SVM usually outperforms other traditional machine learning technologies, including the neural network and the k-nearest neighbor classifier. In recent years, SVM have been also used in bioinformatics. Hua & Sun (2001) first applied SVM to predict protein secondary structure and protein subcellular localization. More detailed descriptions of the SVM method can be found in Vapnik's publications (Vapnik, 1995, 1998).

There are several parameters in the SVM, including the kernel function and the regularization parameter C. The inner product in the feature space is called a kernel function. The present study adopted the widely used radial basis function (RBF):

The basic SVM algorithm is designed for binary classification problems only. Nevertheless, there are several methods to extend the SVM for classifying multi-class proteins. This paper used the “one-against-one” strategy. For a k-classification problem, the “one-against-one” strategy constructs k*(k-1) classifiers with each one trained with the data from two different classes. The final decision is based on a voting strategy, i.e., the test sample is classified into the class chosen by the most binary classifiers. The software toolbox used to implement the SVM in this paper was LIBSVM by Chih-Chung Chang and Chih-Jen Lin. The software toolbox can be downloaded from: http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

### 2.3 Residue-Couple model

The traditional subcellular location prediction model is primarily based on the amino acid composition model. However the amino acid composition model alone ignores a certain amount of information of the protein sequence. Unfortunately, the information about the sequence order effect can not be easily incorporated 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 utilizing the sequence order effect indirectly.

The model denotes a protein sequence as a series of letters:

$$R_1R_2R_3R_4R_5R_6R_7 \cdots\cdots R_L$$

where $R_l$ represents the amino acid in location $l(l=1,2,\ldots,L)$. The “residue-couple” is defined as follows:

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

where $H_{i,j}(n, n+k)$ =1 if the amino acid in location n is i and the one in location n+k is j; otherwise $H_{i,j}(n, n+k)$ =0 (Figure 1). The values 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 represents the frequency with which a mode of continuous residue pairs is observed in a protein sequence. $X_{i,j}^{2}$ is called the 2nd-rank residue-couple that represents the frequency with which the coupled 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 with which the coupled mode $(i,\_,\_,j)$ is observed in a protein sequence, and so forth. There are, therefore, 20×20=400 residue-couples in each rank (Figure 1).

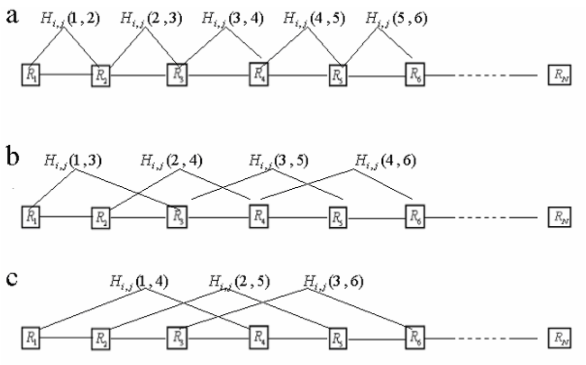

**Figure 1:** A schematic drawing to show the residue-couple with different rank: (a) the 1st-rank: the coupling mode between all the two consecutive residues. (b) the 2nd-rank: the coupling mode between two residues with only one amino acid between them. (c) the 3th-rank: the coupling mode between two residues with right two amino acid between them.

For each protein sequence, all the residue-couples were combined into a vector $x$ , that is, the first 400 components of $x$ were the 400 1st-rank residue-couples and the following 400 components are the 400 2nd-rank residue-couples, and so forth. Therefore, the final vector has a dimension of 400×m. The value of m is called the "coupling-degree", representing the total rank of residue-couples. This model contains the information for both the amino acid composition and the order effect of the protein sequence. Each protein sequence was analyzed in this way to obtain a set of 400×m dimension vectors (each vector corresponds to one vector). The set of vectors was used as the input vectors to the support vector machine for training and prediction (Figure 2).

### 2.4 Cross-validation and model selection

The current work used a 5-fold cross validation for testing because of our limited computational power. In the k-fold cross validation, the entire sample set was randomly divided into k equally sized subsets. One subset at a time was used as the test set and the other k-1 subsets were used to train the SVM. The final prediction results were generated by combining the results of each subset in turn.

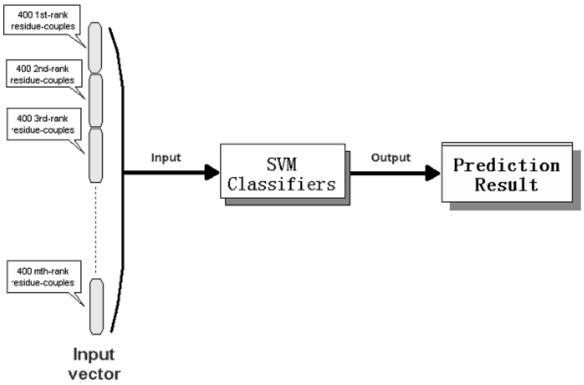
400 1st-rank residue-couples
400 2nd-rank residue-couples
400 3rd-rank residue-couples
400 mth-rank residue-couples
Input vector
Input
SVM Classifiers
Output
Prediction Result

**Figure 2:** The prediction process of our method. The input vector of SVM is a number of 400×m dimension vectors.

### 2.5 Prediction result assessments

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

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

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

$$Total\_Accuracy = \frac{\sum_{i=1}^{k} M_{\mathrm{i\,i}}}{N} \qquad (2)$$

$$Accuracy(i) = \frac{M_{ii}}{obs(i)} = \frac{M_{ii}}{\sum_{j=1}^{k} M_{ij}} \qquad (3)$$

Matthew's Correlation Coefficient (MCC) was defined as follows:

$$MMC_i = \frac{p_i \cdot n_i - u_i \cdot o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(n_i + u_i)(n_i + o_i)}} \qquad (4)$$

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

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

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

## 3 Results

### 3.1 Prediction Accuracy

Table 1: The prediction accuracy of the current method for eukaryotic proteins with different input vectors coupling-degrees. The results were based on the 5-fold cross-validation test. Table 1 shows that the total accuracy reached 86.9%. when the coupling-degree was equal to 6 and the kernel parameter $\gamma$ was 20. The accuracies in different subcellular locations are also listed in Table 1.

| | | Coupling-degree | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| Total Acc (%) | | 80.4 | 85.5 | 86.5 | 85.9 | 86.5 | ***86.9*** | 86.7 | 86.6 |
| Acc | Cyto | 79.5 | 84.2 | 85.0 | 84.5 | 86.0 | 85.8 | 85.7 | ***86.4*** |
| | Extra | 79.7 | 80.0 | 84.3 | 81.0 | 85.0 | ***85.9*** | 83.4 | 82.8 |
| | Mito | 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 |
| $\gamma$ | | 100 | 100 | 50 | 50 | 20 | 20 | 20 | 20 |

Table 2: The accuracies of the method for prokaryotic proteins with different input vectors coupling-degrees. Only five different coupling-degrees are shown in the table since the result changed little with increasing coupling-degree.

| | | Coupling-degree | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 |
| Total Acc | | 90.7 | 91.3 | 91.2 | 91.5 | ***92.0*** |
| Acc | Cyto | ***99.1*** | 98.4 | 98.1 | 99.0 | 99.0 |
| | Peri | 70.1 | 76.6 | ***78.5*** | 73.8 | 77.6 |
| | Extra | 72.8 | 74.8 | 74.8 | 75.3 | ***75.7*** |
| $\gamma$ | | 100 | 100 | 100 | 100 | 100 |

## 3.2 Prediction result and comparison with other methods

Table 3: The comparisons of different prediction method for the eukaryotic sequences. 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.

| Location | ANN | Markov model | | Amino acid composition +SVM | | Residue-couple model +SVM | |
|---|---|---|---|---|---|---|---|
| | Acc ( %) | Acc (%) | MCC | Acc (%) | MCC | Acc (%) | MCC |
| Cyto | 55 | 78.1 | 0.60 | 76.9 | 0.64 | ***85.8*** | ***0.77*** |
| Extra | 75 | 62.2 | 0.63 | 80.0 | 0.78 | ***85.6*** | ***0.89*** |
| Mito | 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 Acc | 66 | 73.0 | -- | 79.4 | -- | ***86.9*** | -- |

Table 4 : The comparisons of different methods for the prokaryotic sequences. 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).

| Location | ANN | Covariant discriminant | Markov model | | Amino acid composition +SVM | | Residue-couple model +SVM | |
|---|---|---|---|---|---|---|---|---|
| | Acc (%) | Acc (%) | Acc (%) | MCC | Acc (%) | MCC | Acc MCC (%) | |
| Cyto | 80 | 91.6 | 93.6 | 0.83 | 97.5 | 0.86 | ***99.0*** | ***0.89*** |
| Extra | 77 | 80.4 | 77.6 | 0.77 | 75.7 | 0.77 | 77.6 | ***0.79*** |
| Peri | ***85*** | 72.7 | 79.7 | 0.69 | 78.7 | 0.78 | 75.7 | 0.78 |
| Total Acc | 81 | 86.5 | 89.1 | -- | 91.4 | -- | ***92.0*** | -- |

The prediction result from this method was also compared with that of other subcellular localization methods. For eukaryotic sequences, the residue-couple model is compared with the neural network method (Reinhardt & Hubbard, 1998), the Markov model (Yuan, 1999) and Hua and Sun's simple SVM method (Hua & Sun, 2001) in Table 5. The results showed that the total accuracy of the residue-couple model was 20.9% higher than that of the neural network method and 7.5% higher than that of the SVM method. For cytoplasm 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 prediction accuracy of this model was obviously higher than that of Hua & Sun's SVM method, even though it used the same support vector machine classification algorithm. This clearly reflects that residue-couple model was able to mine more useful information from the protein sequences than the amino acid composition model, especially for cytoplasm and mitochondrial sequences (8.9% and 8.7 higher than Hua and Sun's work).

Both the residue-couple model and the Markov model used sequence order information for the predictions. The total accuracy of the residue-couple model was 13.9% higher than that of the Markov model. The accuracy for extra-cellular and nuclear proteins was 23.7% and 20.1% higher than those of the Markov model method, although the accuracy for mitochondrial proteins was 3.8% lower (nevertheless, the MCC of the residue-couple model for mitochondrial was 0.72, much higher than that of the Markov model). Although both methods were based on residue order information, the powerful classification capability of SVM allowed the new method to achieve greater accuracies.

The MCC results for the different methods are also listed in Table 3. The MCC of each subcellular location using the residue-couple model was higher than all the other models, as shown in Table 3.

For the prokaryotic sequences, the results are compared in Table 4. The total accuracy of the residue-couple model was about 11% higher than that of the neural network method and 5.5% higher than that of the covariant discriminant algorithm. The accuracy for cytoplasm sequences reached 99%, although the total accuracy had no significant improvement compared with Hua & Sun's method.

For the new data with 8304 proteins and 8 subcellular locations, the total accuracy achieved 88.9%. The accuracy and the MCC for each subcellular location are listed in Table 5.

Table 5: The prediction result of our new dataset with 8304 proteins and 8 subcellular locations. The results are based on 5-fold cross validation. We used the RBF kernel with parameters: $\gamma$ =21 and C=500.

| | chlop | cyto | extra | golgi | lyso | mito | nuclear | perox |
|---|---|---|---|---|---|---|---|---|
| Acc (%) | 91.4 | 90.1 | 82.4 | 68.7 | 91.0 | 72.2 | 93.4 | 74.1 |
| MCC | 0.92 | 0.83 | 0.85 | 0.82 | 0.94 | 0.78 | 0.86 | 0.83 |

### 3.3 Robustness against errors in the N-terminal sequence

The residue-couple model was also much more robust against errors in the protein terminal sequence than methods based on sorting signals. To show this, the samples were randomly divided into 5 equally sized subsets. One subset at a time was used as the testing set while the other 4 subsets were used to train the SVM. N-terminal segments with lengths of 10, 20, 30 and 40 amino acids were removed from the protein sequences in the testing set while keeping the full sequences of the proteins in the training set. Therefore, the SVM classifiers were trained on full sequences and tested on sequences with several missing N-terminal segments. The final prediction results were generated by combining the results of each subset in turn. The results for eukaryotic sequences and for prokaryotic sequences listed in Table 6 and Table 7 show that the total accuracy decreased only 3.2% for eukaryotic sequences and 1.1% for prokaryotic sequences even though 40 residues in the N-terminal were removed.

Table 6: performance comparisons for the eukaryotic protein sequences with one segment of N-terminal sequences removed. 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.

| | Accuracy (%) | | | | | **MCC** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total | Cyto | Extra | Mito | Nucl | Cyto | Extra | Mito | Nucl |
| Complete | 86.9 | 85.8 | 85.9 | 65.4 | 97.2 | 0.77 | 0.89 | 0.72 | 0.85 |
| CUT-10 | 85.2 | 85.2 | 81.5 | 59.8 | 93.6 | 0.75 | 0.87 | 0.68 | 0.83 |
| CUT-20 | 84.0 | 84.8 | 80.0 | 54.5 | 93.3 | 0.73 | 0.86 | 0.63 | 0.82 |
| CUT-30 | 83.1 | 83.5 | 80.0 | 50.5 | 93.4 | 0.72 | 0.85 | 0.60 | 0.82 |
| CUT-40 | 82.5 | 82.9 | 78.8 | 48.6 | 93.2 | 0.71 | 0.83 | 0.59 | 0.82 |

Table 7: performance comparisons for the prokaryotic protein sequences with one segment of N-terminal sequences removed. 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.

| | **Accuracy (%)** | | | | **MCC** | | |
|---|---|---|---|---|---|---|---|
| | Total | Cyto | Extra | Peri | Cyto | Extra | Peri |
| Complete | 92.0 | 98.7 | 77.6 | 76.7 | 0.90 | 0.79 | 0.77 |
| CUT-10 | 92.0 | 98.3 | 79.4 | 77.2 | 0.89 | 0.81 | 0.78 |
| CUT-20 | 91.5 | 98.4 | 78.5 | 74.8 | 0.88 | 0.80 | 0.76 |
| CUT-30 | 91.4 | 98.7 | 79.4 | 72.8 | 0.88 | 0.80 | 0.75 |
| CUT-40 | 90.8 | 98.0 | 78.5 | 72.8 | 0.87 | 0.79 | 0.73 |

## 4 Discussion and future work

The results showed that the residue-couple model successfully predicted subcellular locations. Compared with other methods, the prediction accuracy of the residue-couple model was much more evident for eukaryotic protein sequences than for prokaryotic sequences. The total accuracy of the method was only 0.6% than Hua & Sun' method (Table 3) for prokaryotic protein sequences. However, the accuracy was 7.5% better for eukaryotic proteins (Table 6). Note that the prokaryotic proteins have been classified with high accuracy even using linear classifiers based on amino acid composition only (the total accuracy reached 89.3% with a linear kernel SVM). This result probably reflects that their amino acid composition, which has relative simple sequence structure and biological function, is the key characteristic of prokaryotic proteins. However, eukaryotic protein sequences seem much more complex than prokaryotic sequences and their amino acid composition does not contain enough information to predict protein location. Therefore, for eukaryotic proteins, the accuracy of existing methods based on amino acid composition models are with significantly lower than the residue-couple model, which not only considers the from the amino acid composition information, but also the sequence order information.

Further studies, which will focus on three aspects to improve our work, are planned for the immediate futher,. One is to combine the residue-couple model with other complementary methods. Mitochondrial proteins are still not well predicted (65.4%), although the accuracy was higher than that of all other prediction methods (Table 3) except the Markov model. 19% of the mitochondrial proteins were incorrectly classified into the cytoplasm. A similar conclusion was also reported by Hua and Sun. This means that it is difficult to discriminate the proteins in cytoplasm and mitochondria based solely on residue-couple information. The relatively high prediction accuracy for mitochondrial proteins using the Markov model (69%) points to a combination of the Markov model and the residue-couple model as the next logical model to investigate. Future research will identify the proper strategies to combine these models. Combined methods based on sorting methods are also under consideration.

The second aspect of future work is to incorporate other types of data into the model, including gene expression profiles (Murphy et al, 2000, Nakai et al, 1997) and regulatory pathway information. Some information fusion technologies, such as meta learning methods, may be used to combine information from different datasets and different types of formats.

The third aspect is to improve the SVM classifiers, including finding ways to select better kernels, to speed up the prediction system, and to filter noise and outliers. Several papers have introduced new methods addressing the noise and outliers problems (Zhang, 1999). Some new SVM software tools such as Herosvm (Dong et al, 2002) significantly speed up the process. We are also attempting to combine an active learning strategy with the SVM method for further improvements.

## 5 Webserver and application

The residue-couple model has been integrated into a webserver system so as to provide a subcellular localization service. The server address is:
http://www.bioinfo.tsinghua.edu.cn/CoupleLoc

## 6 Conclusion

A residue-couple model was developed for subcellular localization, which not only considered the amino acid composition information, but also the residue order information. The high accuracies for both prokaryotic (92.0%) and eukaryotic sequences (86.9%) showed that the new method performed well compared with other methods for subcellular location prediction. Furthermore, the method was robust agianst the errors in the N-terminal of sequences, and one real test with an unknown protein sequence comfirmed the prediction accuracy. Therefore, the residue-couple model is a more powerful system for subcellular location prediction which will be a useful tool for large-scale protein function analysis.

## 7 Acknowledgement

The authors thank Hongxiu Ning for his kindly help. We thank Dr. A.Reinhardt for providing the dataset. We thank Mr. Chih-Chung Chang and Chih-Jen Lin for providing the LIBSVM software. We also thankMr. Jiashun Zheng, Mr. Tao Guo and Dr. Shuqun Liu for their helpful suggestions. This work was supported by Foudational Science Research Grant of Tsinghua University (No. JC2001043) the 863 projects (2002AA234041).

## 8 References

1. 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.
2. Chou,K.C. and Elord,D. (1999) Protein subcellular location prediction. Protein Eng., 12, 107-118.
3. Chou.K.C. (2001) Prediction of Protein Subcellular Locations by Incorporating Quasi-Sequence-Oder Effect. Biochem. biophys. res. commun. 278, 477-483
4. 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.
5. 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.
6. Frishman, D., Mironov, A. and Gelfand, M. (1999) Start of bacterial genes: estimating the reliability of computer prediction. Gene, 234, 257-265
7. Hua,S.J. and Sun,Z.R (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics, 17, 721-728.
8. Jian-xiong Dong, Ching Y. Suen and Adam Krzyzak. (2002) A fast parallel optimization for training support vector machine. Technical Report, CENPARMI, Concordia University.
9. Nair R. and Rost B. (2002) Seqence conserved for subcellular localization. Protein Science, 11: 2836-2847.
10. 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

11. 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.
12. Nakai,K. and Kanehisa,M. (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics, 14, 897-911.
13. 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.
14. Nakai,K. (2000) Protein sorting signals and prediction of subcellular localization. Advances in Protein Chemistry, 54, 277-344.
15. 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.
16. 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.
17. Reinhardt,A. and Hubbard,T. (1998) Using neural networks for prediction of the subcellular location of proteins. Nucl. Acids Res., 26, 2230-2236.
18. Rost, B. and Fariselli, P. and Casadio.R. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci., 5, 1704-1718
19. Vapnik,V. (1995) The Nature of Statistical Learning Theory. Springer-Verlag, New York.
20. Vapnik,V. (1998) Statistical Learning Theory. John Wiley and Sons, Inc., New York.
21. Yuan,Z. (1999) Prediction of protein subcellular locations using Markov chain models. FEBS Letters, 451, 23-26.
22. Zhang,Z.(1999) Using class-center vectors to build support vector machines. Proceeding of the 1999