Amino Acids (2008) 35: 345–353
DOI 10.1007/s00726-007-0616-y
Printed in The Netherlands

Amino Acids

# AAIndexLoc: predicting subcellular localization of proteins based on a new representation of sequences using amino acid indices

**E. Tantoso**[1] and **Kuo-Bin Li**[2]

[1] Bioinformatics Institute, Singapore
[2] Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei, Taiwan

Received September 28, 2007
Accepted October 4, 2007
Published online December 28, 2007; 

**Summary.** Identifying a protein's subcellular localization is an important step to understand its function. However, the involved experimental work is usually laborious, time consuming and costly. Computational prediction hence becomes valuable to reduce the inefficiency. Here we provide a method to predict protein subcellular localization by using amino acid composition and physicochemical properties. The method concatenates the information extracted from a protein's N-terminal, middle and full sequence. Each part is represented by amino acid composition, weighted amino acid composition, five-level grouping composition and five-level dipeptide composition. We divided our dataset into training and testing set. The training set is used to determine the best performing amino acid index by using five-fold cross validation, whereas the testing set acts as the independent dataset to evaluate the performance of our model. With the novel representation method, we achieve an accuracy of approximately 75% on independent dataset. We conclude that this new representation indeed performs well and is able to extract the protein sequence information. We have developed a web server for predicting protein subcellular localization. The web server is available at http://aaindexloc.bii.a-star.edu.sg.

**Keywords:** Subcellular localization – Support vector machine – Amino acid indices

## Introduction

Many novel bioinformatics applications rely on accurate prediction of protein's subcellular localization. The filtering of putative protein-protein interactions is one of the examples where false predictions are to be removed provided the two interacting partners are involved in different subcellular compartments (Mahdavi and Lin, 2007). Another example is the identification of serum biomarkers. Here selecting genes and gene products that possess identical protein localization may serve as one of the criteria (Klee et al., 2006). However, experimentally identifying the localization of a protein is usually a laborious, time-consuming and costly process. Therefore, computational predictions are important to minimize the time and cost in experimental work. Many efforts have been made in this regard (Nakai and Kanehisa, 1992; Nakashima and Nishikawa, 1994; Cedano et al., 1997; Chou and Elrod, 1998; 1999a, b; Nakai and Horton, 1999; Yuan, 1999; Chou, 2000a, b, 2001; Emanuelsson et al., 2000; Murphy et al., 2000; Nakai, 2000; Feng, 2001, 2002; Feng and Zhang, 2001; Hua and Sun, 2001; Chou and Cai, 2002; Gardy et al., 2003; Pan et al., 2003; Park and Kanehisa, 2003; Zhou and Doctor, 2003; Huang and Li, 2004; Gao et al., 2005a, b; Garg et al., 2005; Lei and Dai, 2005; Matsuda et al., 2005; Xiao et al., 2005, 2006a; Chou and Shen, 2006c, 2007a; Guo et al., 2006a; Hoglund et al., 2006; Lee et al., 2006; Xiao et al., 2006a; Chou and Shen, 2007a, b, d; Shi et al., 2007; Zhang and Ding, 2007). A summary in this area was given in a recent review paper (Chou and Shen, 2007d).

The approach for predicting protein subcellular localization can be divided into two steps, i.e., the representation and the classification one. The representation step is the most challenging part to obtain high prediction accuracy. The step can be seen as a data mining process where, for a protein in a given localization, information embedded in the primary sequence is extracted so that a computer program can discriminate the protein from proteins in other localizations. There have been several ways to extract the information from protein sequences, such as using amino acid composition (Cedano et al., 1997), signal sequence (Nakai and Horton, 1999) or N-Terminal sequence (Emanuelsson et al., 2000), n-peptide composition (Yu et al., 2004), pseudo-amino acid composition

(Chou, 2001), functional domain composition (Cai et al., 2003), gene ontology (Chou and Cai, 2003), amino acid property (Feng and Zhang, 2001; Sarda et al., 2005) and homology (Bhasin and Raghava, 2004; Xie et al., 2005).

Amino acid composition was originally used to represent protein samples for predicting protein structural class (Chou and Zhang, 1994, 1995; Zhou, 1998), indicating that there is some correlation between AA composition of a protein and its attributes (Chou, 2000c, 2002). Since then, such a descriptor has been widely used to predict protein subcellular localization (see, e.g., Cedano et al., 1997; Chou and Elrod, 1999b; Hua and Sun, 2001; Zhou and Doctor, 2003; Jin et al., 2005). The AA composition does not contain any sequence order information. To avoid completely losing the sequence order information, the pseudo amino acid (PseAA) composition was introduced (Chou, 2001, 2005). Since the introduction of PseAA composition, it has been adopted to improve the prediction quality of various protein attributes by many investigators (Pan et al., 2003; Wang et al., 2004; Chou and Cai, 2005; Gao et al., 2005b; Liu et al., 2005a, b; Shen and Chou, 2005a, b; Du and Li, 2006; Mondal et al., 2006; Shen and Chou, 2006; Shen et al., 2006; Wang et al., 2006; Xiao et al., 2006a, b; Zhang et al., 2006a, b; Chen and Li, 2007; Ding et al., 2007; Kurgan et al., 2007; Lin and Li, 2007a, b; Mundra et al., 2007; Pu et al., 2007; Shen and Chou, 2007c; Shen et al., 2007; Shi et al., 2007; Zhang and Ding, 2007; Zhou et al., 2007). Because PseAA composition has been widely used, recently a webserver called PseAA was established at http://chou.med.harvard.edu/bioinf/PseAA/, by which users can generate various different kinds of PseAA compositions for a given protein sequence. ESLpred (Bhasin and Raghava, 2004) has used amino acid composition, dipeptide composition, physico-chemical properties and PSI-BLAST profiles to predict protein subcellular localization. An alternative method to extract protein localization information is to use signal sequences. TargetP (Emanuelsson et al., 2000) used the N-terminal sequence information only and was shown to be able to discriminate the protein in four locations, i.e., mitochondrion, chloroplast, secretory pathway and others. However, in the case where the signal region is located at regions other than the N-terminus, there is a risk of information loss if only the N-terminal sequence is used. As a result, Matsuda et al. (2005) introduced a representation method that uses different parts of a protein's sequence to predict its subcellular localization, i.e., N-terminus, middle and C-Terminus. Recently, a novel software, MultiLoc (Hoglund et al., 2006), also incorporated amino acid composition, N-terminus sequence and sequence motifs to represent a protein. MultiLoc has been shown to be an accurate protein localization predictor.

The second step for the localization prediction problem is the classification step. Once the protein is represented with an appropriate encoding scheme, the remaining work is to propose a robust classifier to predict the subcellular localization. Many classification approaches have been proposed lately, such as neural network (Reinhardt and Hubbard, 1998), support vector machine (Hua and Sun, 2001; Chou and Cai, 2002; Park and Kanehisa, 2003; Hoglund et al., 2006), Markov chain model (Yuan, 1999), covariant discriminant algorithm (Chou and Elrod, 1998, 1999a, b), fuzzy k-NN method (Huang and Li, 2004) and FDOD function (Jin et al., 2005).

AAIndexLoc is a new representation method for prediction of protein subcellular localization. We hypothesize that the physicochemical properties of amino acids play an important role in determining a protein's function and therefore might be used to predict the protein's localization. Given the 55 amino acid indices collected by ProtScale (http://www.expasy.org/cgi-bin/protscale.pl), we attempted to determine the optimum amino acid index to characterize a specific subcellular localization.We introduced the weighted amino acid (AA) composition, five-group-AA composition and five-group dipeptide composition as the new encoding scheme to represent the protein sequences. The rational of the weighted AA composition is that some amino acids may be more important in terms of protein translocation even if their frequency is relatively low. Therefore, the weighted AA composition provides a way to increase the contribution from the rare but critical amino acid residues. In addition to the weighted AA composition, we also categorize amino acids into five groups by using *k*-means clustering and calculate the group composition of a protein. Proteins in a common cellular location may share amino acids with similar physicochemical properties. Group composition is meant to extract such information. We have also considered the five-level dipeptide composition which might detect some features about the appearance of consecutive amino acids with certain properties. On top of that, to avoid losing global information, we divided protein sequences into three parts, i.e., the N-terminus, middle and C-Terminus.

Information from the N-terminus, middle and full-length protein is used to create input features for support vector machine (SVM) classifier. To test our approach, the localization data are divided into the training and the independent testing set. The training set is used to find the best performing AA index for individual localization by using the five-fold cross validation method; then the best performing model is used to predict the protein's

localization on the independent testing set. Our results show that accurate prediction of a protein's subcellular localization can be obtained using both the local and global information of a protein sequence.

## Materials and methods

### *Datasets*

Dataset created by MultiLoc (Hoglund et al., 2006) were used in our experiments. The MultiLoc datasets are categorized into animal (nine locations), fungal (nine locations) and plant (ten locations). Table 2 shows the number of sequences in each location. Note that MultiLoc data are not formed by three separate sets of animal, plants and fungal sequences. Instead, there is only one set of cytoplasmic sequences containing 1411 sequences. In Table 2, for example, one may find that all three versions of the datasets, the animal, the plant and the fungal, share the 1411 cytoplasmic sequences, but only plant version has the 449 chloroplast sequences.

### *Support vector machine*

Support vector machine (SVM), first introduced by Vapnik in 1995 (Vapnik, 1995), is a learning algorithm for pattern recognition and regression. SVM has recently gained a lot of attention in biology (Brown et al., 2000; Hua and Sun, 2001; Lee and Lee, 2003; Ward et al., 2003), particularly for classification purposes. In those applications, an SVM classifier is trained with a set of positively and negatively labelled samples. Once trained, the classifier can be used to classify an unlabeled sample into the positive or the negative class. In principle, SVM maps the input vector into a high dimensional space and constructs an optimal hyperplane with the maximum margin of separation between the hyperplane and the nearest data points of each class in the space.

In building the SVM classifiers for protein subcellular localization, each localization site is corresponded to a class. We used the scheme called One-vs-All SVM. For example, to predict the mitochondria protein, the set of mitochondria proteins are used as the positive samples and the proteins in all the other compartments are used as the negative samples.

SVMlight is a software package that contains an implementation of SVM in C language. The package is available from http://svmlight.joachims.org and was used throughout the work. The radial basis function (RBF) kernel was adopted to train the SVM model.

### *Input features*

#### Amino acid (AA) composition

Let the sequence of a protein $P = x_1x_2\cdots x_N$ where $x_p \in S, p = 1,2,3,\cdots,N$. $S$ is the set of the 20 amino acid (AA) alphabets, $S = \{y_1, y_2, y_3, \cdots, y_{20}\}$. AA composition is defined as the percentage or fraction of amino acid $y_i$ in $P$, where $i = 1,2,3,\cdots,20$. Therefore, the composition $C(y_i)$ for amino acid $y_i$ is

$$C(y_i) = \frac{\text{num}(y_i)}{N} \times 100\% \tag{1}$$

where $\text{num}(y_i)$ is the number of amino acid $y_i$ in protein $P$, $N$ is the length of protein P.

#### Weighted AA composition

Let A denote a specific amino acid index for the 20 amino acids, $A = \{a_1, a_2, \cdots, a_{20}\}$ where $a_i$ is the index value for the amino acid $y_i$. The weighted AA composition for amino acid $y_i$ is defined as:

$$W(y_i) = C(y_i) \times a_i \tag{2}$$

#### Five-level grouping composition

Five-level grouping composition means that the amino acids are classified into five groups based on their amino acid index values, i.e., the highest (T), high (H), medium (M), low (L), and lowest (B) properties. After that, the composition of each group is calculated. The method used for grouping is $k$-means clustering ($k = 5$).

Let $G_m$ denotes the set of amino acids in group $m$, $G_m = \{g_1, g_2, \cdots, g_{N_m}\}$ where $N_m$ is the number of amino acids in $G_m$. The composition of $G_m$ is

$$CM(G_m) = \sum_{j=1}^{N_m} C(g_j) \tag{3}$$

#### Five-level dipeptide composition

As explained in the five-level grouping method, the 20 amino acids are classified into five groups, the highest (T), high (H), medium (M), low (L) and lowest (B) groups. The five-level dipeptide composition is defined as the composition of the occurrence of two consecutive groups, for example: TT, TH, TM, TL, TB, HT, HH, etc. There are 25 combinations of two consecutive group altogether.

#### Features vector

A protein sequence is divided into three parts, i.e., the N-terminus, the middle and the C-terminus. The feature vectors consist of the information from the N-terminus, middle and the full-length protein. We ignore the C-terminus because it does not give significant improvement to the prediction of protein localizations.

Let L = length of the protein P, $L_N$ = length of the N-terminus, $L_M$ = length of the middle part of the protein, and $L_C$ = length of the C-terminus. In this work, the length of N-Terminus and C-terminus is fixed while the length of middle part is varied depending on the length of protein. To determine the length of middle sequence, we have three conditions, i.e.:

1. If $L > L_N + L_C$, then $L_M = L - L_N - L_C$. Check $L_M$: if $L_M < 40$, then set $L_M = L/3$
2. If $L > L_N$ but $L < L_N + L_C$ then $L_M = L/3$
3. If $L \leq L_N$ then $L_N = L$, $L_M = L/3$

The length of N-terminus and C-terminus is determined computationally. We learned that the length of N-terminus is optimum at length 30 and the length of C-terminus is 10. However, for chloroplast localization, the length of N-terminus is different from that of other localizations. Eventually we set the length of N-terminus to 100 when training chloroplast sequences, since it is believed that the chloroplast targeting signal is at least 100 amino acids (Keegstra and Cline, 1999).

The feature vector representing a protein consists of the following features: AA composition, weighted AA composition, five-level grouping composition and five-level dipeptide composition from the N-terminus, the middle part and the full-length protein. There are 20 input features for AA composition. Weighted AA composition has 20 input features; five-level grouping composition has five input features and five-level dipeptide composition has 25 input features. Each region of a protein is thus represented by 70 input features altogether. To represent a protein, we have $70 * 3 = 210$ input features, i.e., the length of the feature vector is 210.

### *Performance measurement*

We use an independent dataset to test the SVM model we have built. The model is trained using 60% of the dataset by using five-fold cross validation and is tested on 40% of the remaining dataset which has not been touched. It should be noted that, however, among the independent dataset test, sub-sampling (e.g., 5 or 10-fold cross-validation) test, and jackknife test, which are often used for examining the accuracy of a statistical

prediction method, the jackknife test was deemed the most rigorous and objective (Chou and Zhang, 1995) as demonstrated by a incisive analysis in a comprehensive recent review (Chou and Shen, 2007d) and has been widely adopted by investigators to test the power of various prediction methods (Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003; Gao et al., 2005b; Wang et al., 2005; Xiao et al., 2005; Chen et al., 2006a, b; Chou and Shen, 2006a, b; Du and Li, 2006; Guo et al., 2006b; Kedarisetti et al., 2006; Mondal et al., 2006; Niu et al., 2006; Sun and Huang, 2006; Xiao et al., 2006a; Zhang et al., 2006a; Chen et al., 2007; Chou and Shen, 2007a, b, c, e; Ding et al., 2007; Lin and Li, 2007a, Liu et al., 2007; Shen and Chou, 2007b, c, d; Shen et al., 2007; Shi et al., 2007; Wen et al., 2007; Zhang and Ding, 2007; Zhou et al., 2007).

The sensitivity, specificity, Matthews' correlation coefficient (MCC) (Matthews, 1975) and overall accuracy for each localization site is calculated to evaluate the prediction performance. The formula for each measurement is given below:

$$\text{Sensitivity } (i) = \frac{\text{TP}(i)}{\text{TP}(i) + \text{FN}(i)} \tag{4}$$

$$\text{Specificity } (i) = \frac{\text{TP}(i)}{\text{TP}(i) + \text{FP}(i)} \tag{5}$$

$$\text{MCC } (i) = \frac{(\text{TP}(i) \times \text{TN}(i)) - (\text{FP}(i) \times \text{FN}(i))}{\sqrt{(\text{TP}(i) + \text{FN}(i))(\text{TP}(i) + \text{FP}(i))(\text{TN}(i) + \text{FP}(i))(\text{TN}(i) + \text{FN}(i))}} \tag{6}$$

$$\text{Accuracy} = \frac{1}{Z} \times \sum_{i=1}^{Z} \text{TP}(i) \tag{7}$$

where, TP = true positive, TN = true negative, FP = false positive, FN = false negative, MCC = Matthews' correlation coefficient, $i$ = localization $i$, and Z = number of localizations.

*Experimental setup*

Protein sequences are represented as described previously. The dataset is randomly divided into the training and the independent testing set. The training dataset contains 60% of the dataset, while the testing dataset contains 40% of the dataset.

The training set is used to choose the best model for predicting protein subcellular localization using five-fold cross validation. The best model is then used to predict the localization of the independent dataset (testing set) and the performance is calculated as described above (Performance Measurement section). Figure 1 illustrates the workflow of the experiments.

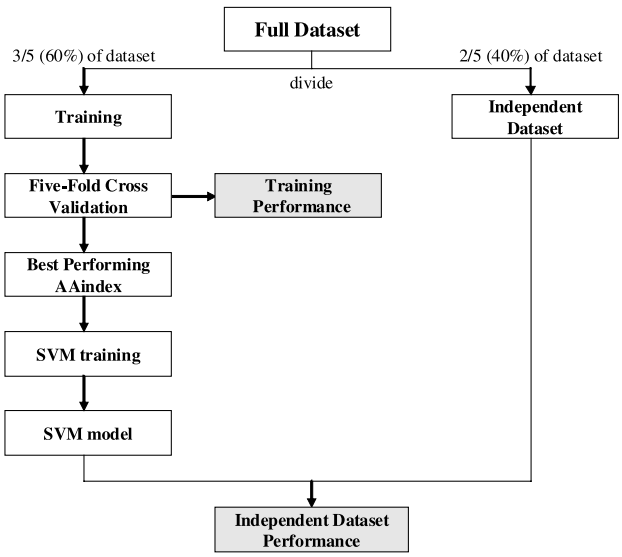


**Fig. 1.** Experimental setup

*Prediction system*

We have provided a web server to predict the subcellular localization for unknown proteins. It is available at http://aaindexloc.bii.a-star.edu.sg. The query protein sequence is first encoded by the best performing AA index for each individual localization. Then the encoded sequence is sent to the best classifier for each localization. The localization whose classifier gives the highest score will be assigned to query sequence as the predicted localization.

## Results

### *Choosing the best performing amino acid index*

We extracted 55 amino acid indices from the ProtScale. For each AA index, proteins were encoded and trained with SVM using the training set. The training performance measurements were calculated using five-fold cross validation. This process was repeated for all the AA indices in the ProtScale list. The best performing AA index is then reported for that particular localization (see Table 1). The selection of the best performing AA index is mainly based on MCC values. However, in addition to MCC, for localizations with fewer sequences, sensitivity needs to be considered as well. This is to ensure that the prediction does not bias towards high specificity but low sensitivity performance.

### *Improvement by using information from different parts of proteins*

Table 2 presents the improvement in performance using information from different parts of the protein sequences (as shown under the results of full-length-only and AAIndexLoc). Full-length-only is the method that used only the full-length protein sequence information to predict protein subcellular localization. On the other hand, AAIndexLoc used the information from the N-terminus, the middle part of protein sequences, and the full-length protein sequences to predict protein's localization. The results indicate that proteins contain information in different parts of their sequences and hence using information from different parts of proteins can increase the prediction accuracy of their localization. The question is how to determine the optimum number of amino acids for each of the three parts. Based on our computational analysis, the optimal

**Table 1.** Best performing AAindex for each localization

| Dataset | Localization | AAindex | Ref. |
|---|---|---|---|
| MultiLoc (Animal) | Cytop | Proportion of residues 95% buried (in 12 proteins) | Chothia (1976) |
| | ER | Conformational parameter for beta-sheet | Deleage and Roux (1987) |
| | Extr | Optimized matching hydrophobicity | Sweet and Eisenberg (1983) |
| | Golgi | Hydrophilicity | Hopp and Woods (1981) |
| | Lyso | Optimized matching hydrophobicity | Sweet and Eisenberg (1983) |
| | Mito | Conformational parameter for beta-turn | Deleage and Roux (1987) |
| | Nuc | Hydrophobicity scale based on free energy of transfer | Guy (1985) |
| | Pero | Average area buried on transfer from standard state to folded protein | Rose et al. (1985) |
| | Plas | Hydrophilicity | Hopp and Woods (1981) |
| MultiLoc (Fungal) | Cytop | Hydrophilicity scale derived from HPLC peptide retention times | Parker et al. (1986) |
| | ER | Average flexibility index | Bhaskaran and Ponnuswamy (1988) |
| | Extr | Hydrophilicity | Hopp and Woods (1981) |
| | Golgi | Hydrophilicity | Hopp and Woods (1981) |
| | Mito | Free energy of transfer from inside to outside of a globular protein | Janin (1979) |
| | Nuc | Conformational preference for parallel beta strand | Lifson and Sander (1979) |
| | Pero | Refractivity | Jones (1975) |
| | Plas | Hydrophobicity scale based on free energy of transfer | Guy (1985) |
| | Vacu | Molar fraction (%) of 2001 buried residues | Janin (1979) |
| MultiLoc (Plant) | Chlo | Optimized matching hydrophobicity | Sweet and Eisenberg (1983) |
| | Cytop | Free energy of transfer from inside to outside of a globular protein | Janin (1979) |
| | ER | Molecular weight of each amino acid | Most textbooks |
| | Extr | Hydrophilicity | Hopp and Woods (1981) |
| | Golgi | Conformational preference for parallel beta strand | Lifson and Sander (1979) |
| | Mito | Free energy of transfer from inside to outside of a globular protein | Janin (1979) |
| | Nuc | Conformational parameter for beta-sheet (computed from 29 proteins) | Chou and Fasman (1978) |
| | Pero | Refractivity | Jones (1975) |
| | Plas | Hydrophobicity scale based on free energy of transfer | Guy (1985) |
| | Vacu | Molar fraction (%) of 2001 buried residues | Janin (1979) |

Reference comes from ProtScale

length of the N-terminus sequence is 30 amino acids, except for the prediction of chloroplast localization where the optimal length is 100. The length of C-terminal sequence is set to 10 amino acids.

### *Performance comparison*

We compared the result from AAIndexLoc with MultiLoc (see Table 2). Note that the performance of AAIndexLoc is based on the independent testing dataset, which comprises 40% of all the sequences in MultiLoc data.

## Discussion

Amino acid composition has been used to predict protein subcellular localization since Cedano et al. (1997) showed that there exists a statistically significant relationship between amino acid composition and cellular localization of proteins. However, using amino acid composition alone could misclassify proteins which have similar amino acid compositions but are located in different cellular compartment. In this study, in addition to amino acid composition, we explored the possibility of incorporating features based on amino acid's physicochemical properties into a computer classifier. These features include weighted AA composition, five-level grouping composition and five-level dipeptide composition. The weighted AA composition is proposed to capture the information contributed by rare but critical amino acid residues. The five-level grouping composition is designed to increase the composition bias for proteins in different localizations as the component amino acid residues with similar physicochemical property are now considered as one single residue. The five-level dipeptide composition might detect some features about the occurrences of consecutive amino acids with certain properties. We have shown that better prediction accuracy can be achieved by incorporating all those features.

From the result in Table 1, it shows that the best performing AAindex for each localization is dominated by hydrophobicity and hydrophilicity types of properties. It indicates that hydrophobicity and hydrophilicity are the two important properties in protein translocation. Moreover, protein's secondary structure may also play an im-

**Table 2.** Performance of AAIndexLoc compared with MultiLoc for animal, fungal and plant dataset

| | MultiLoc | | | Full-length-only (Independent dataset) | | | AAIndexLoc (Independent dataset) | | |
|---|---|---|---|---|---|---|---|---|---|
| | Se | Sp | MCC | Se | Sp | MCC | Se | Sp | MCC |
| *Animal dataset* | | | | | | | | | |
| Cytop (1411) | 0.67 | 0.85 | 0.68 | 0.79 | 0.45 | 0.40 | 0.87 | 0.73 | 0.72 |
| Er (198) | 0.68 | 0.56 | 0.60 | 0.15 | 0.80 | 0.34 | 0.44 | 0.59 | 0.49 |
| Extr (843) | 0.79 | 0.83 | 0.77 | 0.61 | 0.61 | 0.54 | 0.78 | 0.75 | 0.72 |
| Golgi (150) | 0.71 | 0.43 | 0.53 | 0.00 | 0.00 | 0.00 | 0.60 | 0.69 | 0.63 |
| Lyso (103) | 0.69 | 0.36 | 0.48 | 0.00 | 0.00 | 0.00 | 0.33 | 0.64 | 0.45 |
| Mito (510) | 0.88 | 0.82 | 0.83 | 0.57 | 0.46 | 0.46 | 0.83 | 0.77 | 0.78 |
| Nuc (837) | 0.82 | 0.73 | 0.73 | 0.77 | 0.60 | 0.61 | 0.68 | 0.78 | 0.68 |
| Pero (157) | 0.71 | 0.31 | 0.44 | 0.00 | 0.00 | 0.00 | 0.25 | 0.70 | 0.41 |
| Plas (1238) | 0.73 | 0.90 | 0.76 | 0.88 | 0.63 | 0.66 | 0.75 | 0.80 | 0.71 |
| | Overall accuracy = 74.6% | | | Overall accuracy = 67.6% | | | Overall accuracy = 74.5% | | |
| *Fungal dataset* | | | | | | | | | |
| Cytop (1411) | 0.68 | 0.85 | 0.69 | 0.77 | 0.46 | 0.40 | 0.85 | 0.68 | 0.66 |
| Er (198) | 0.71 | 0.59 | 0.63 | 0.21 | 0.71 | 0.38 | 0.30 | 0.83 | 0.49 |
| Extr (843) | 0.73 | 0.81 | 0.73 | 0.55 | 0.58 | 0.49 | 0.79 | 0.75 | 0.73 |
| Golgi (150) | 0.71 | 0.53 | 0.60 | 0.00 | 0.00 | 0.00 | 0.58 | 0.73 | 0.64 |
| Mito (510) | 0.88 | 0.82 | 0.83 | 0.65 | 0.50 | 0.52 | 0.86 | 0.72 | 0.76 |
| Nuc (837) | 0.81 | 0.74 | 0.73 | 0.76 | 0.57 | 0.59 | 0.74 | 0.76 | 0.70 |
| Pero (157) | 0.68 | 0.30 | 0.43 | 0.00 | 0.00 | 0.00 | 0.36 | 0.41 | 0.37 |
| Plas (1238) | 0.76 | 0.89 | 0.78 | 0.88 | 0.63 | 0.65 | 0.84 | 0.79 | 0.76 |
| Vacu (63) | 0.76 | 0.24 | 0.42 | 0.00 | 0.00 | 0.00 | 0.22 | 0.35 | 0.27 |
| | Overall accuracy = 74.9% | | | Overall accuracy = 67.5% | | | Overall accuracy = 77.3% | | |
| *Plant dataset* | | | | | | | | | |
| Chlo (449) | 0.88 | 0.85 | 0.85 | 0.59 | 0.58 | 0.55 | 0.80 | 0.82 | 0.79 |
| Cytop (1411) | 0.68 | 0.85 | 0.70 | 0.75 | 0.45 | 0.40 | 0.83 | 0.72 | 0.69 |
| Er (198) | 0.72 | 0.54 | 0.61 | 0.02 | 0.67 | 0.11 | 0.47 | 0.58 | 0.51 |
| Extr (843) | 0.68 | 0.81 | 0.70 | 0.52 | 0.57 | 0.47 | 0.78 | 0.76 | 0.73 |
| Golgi (150) | 0.75 | 0.41 | 0.54 | 0.00 | 0.00 | 0.00 | 0.65 | 0.56 | 0.59 |
| Mito (510) | 0.85 | 0.81 | 0.81 | 0.58 | 0.46 | 0.46 | 0.78 | 0.76 | 0.75 |
| Nuc (837) | 0.82 | 0.75 | 0.75 | 0.73 | 0.59 | 0.59 | 0.70 | 0.74 | 0.67 |
| Pero (157) | 0.71 | 0.34 | 0.47 | 0.00 | 0.00 | 0.00 | 0.23 | 0.60 | 0.36 |
| Plas (1238) | 0.74 | 0.89 | 0.77 | 0.86 | 0.62 | 0.64 | 0.84 | 0.77 | 0.75 |
| Vacu (63) | 0.70 | 0.20 | 0.36 | 0.00 | 0.00 | 0.00 | 0.19 | 0.29 | 0.23 |
| | Overall accuracy = 74.6% | | | Overall accuracy = 63.8% | | | Overall accuracy = 76.0% | | |

The number of sequences in each localization is given inside the parentheses. Best performing AAindex is chosen from 60% of the whole dataset and the remaining 40% independent dataset is used to validate the performance of AAIndexLoc as well as the best performing index

portant role in determining its localization. For example, some targeting sequences have the tendency to form a particular secondary structure such as alpha helix or beta sheet (Endo et al., 1989; Hammen et al., 1994). Furthermore, Clausmeyer et al. showed that there exists a high specificity and evolutionary conservation within the signal sequences. The conservation occurs at the level of the biochemical properties of the amino acids (Clausmeyer et al., 1993).

Prediction of protein subcellular localization has been focused on using the full-length protein sequences or considering only the targeting sequences, such as TargetP. However, using the full length protein sequence might overlook some of the local information and hence shows a poor performance in predicting certain localization such as mitochondria. Using target sequence information especially the N-terminal sequence may increase the prediction performance for mitochondria localization. Nevertheless, there are cases which we do not know where the targeting sequences are and therefore are unable to use this information. In this study we found that splitting protein sequences into three parts can effectively remedy this problem.

Table 2 shows the improvement in prediction performance using different parts of proteins as compared to using the full-length protein sequence alone (shown under

the results of full-length-only and AAIndexLoc). We found a considerable improvement of the prediction accuracy using different parts of the proteins. This suggests that different regions of a protein sequence may have different contribution in protein's translocation. This observation can thus be used to improve the performance of protein subcellular localization prediction. Matsuda et al. (2005) have also used the different parts of protein sequences, i.e., N-terminus, middle and C-terminus. Opposed to their work, we do not use the C-terminus sequences because the addition of C-terminus information into the SVM training does not improve the prediction performance. In fact, we noticed that N-terminus sequence plays the most important role in determining the mitochondria localization. Although incorporating N-terminus sequence indeed improves the prediction performance for proteins in mitochondria, the improvement was not seen in predicting other localizations.

We found that the middle part of a protein sequence plays an important role in determining localizations except mitochondria. As a result, we decided to include the N-terminal and the middle sequences. To avoid losing information on the full-length sequence, we also incorporate the information for the full-length protein sequence.

The optimal length of the N-terminal sequence is set to 30 residues except for the chloroplast localization. We found that the performance for chloroplast localization using 30 residues in the N-terminus resulted in low prediction accuracies. Since the length of chloroplast targeting peptides is believed to be at most 100 residues (Keegstra and Cline, 1999), we set the length of the N-terminal residues for chloroplast to be 100 and obtained a substantial increase in prediction accuracy of chloroplast localization. The length of 100 residues in N-terminus does not apply to other localization such as mitochondria. This observation shows that care should be taken when choosing the length of the N-terminus for localization prediction.

## Conclusions

We have introduced a novel bioinformatics application, AAIndexLoc, for predicting protein subcellular localization by using amino acid properties. Results show that the incorporation of amino acid's physicochemical properties indeed improves the prediction performance. We also found that both the local (N-terminal and middle sequence) and the global sequence information contribute to accurate prediction of protein subcellular localization. We have set up a web server for the prediction of protein subcellular localization. It is available at http://aaindexloc.bii.a-star.edu.sg.

## Acknowledgements

The authors would like to thank Mr. Stephen Wong and Ms. Tan Yang-Hwee of Bioinformatics Institute, Singapore for the help in cluster computers. We are also grateful to Mr. Danny Chuon from the Web Services Team of Bioinformatics Institute, Singapore for providing the help in setting up the web services. Finally, we would like to thank Mr. Alfred Song of Bioinformatics Institute, Singapore in providing some helps during the process.

## References

Bhasin M, Raghava GP (2004) ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. Nucleic Acids Res 32: W414–W419

Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 97: 262–267

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257–3263

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

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243: 444–448

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357: 116–121

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423–428

Chen YL, Li QZ (2007) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245: 775–783

Chou KC (2000a) Prediction of protein structural classes and subcellular locations. Curr Protein Pept Sci 1: 171–208

Chou KC (2000b) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278: 477–483

Chou KC (2000c) Review: prediction of protein structural classes and subcellular locations. Curr Protein Peptide Sci 1: 171–208

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43: 246–255

Chou KC (2002) A new branch of proteomics: prediction of protein cellular attributes. In: Weinrer PW, Lu Q (eds) Gene cloning and expression technologies. Eaton Publishing, Westborough MA, pp 57–70

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769

Chou KC, Cai YD (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun 311: 743–747

Chou KC, Cai YD (2005) Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inf Model 45: 407–413

Chou KC, Elrod DW (1998) Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun 252: 63–68

Chou KC, Elrod DW (1999a) Prediction of membrane protein types and subcellular locations. Proteins 34: 137–153

Chou KC, Elrod DW (1999b) Protein subcellular location prediction. Protein Eng 12: 107–118

Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006b) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-Nearest Neighbor classifiers. J Proteome Res 5: 1888–1897

Chou KC, Shen HB (2006c) Predicting protein subcellular location by fusing multiple classifiers. J Cell Biochem 99: 517–527

Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6: 1728–1734

Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. J Cell Biochem 100: 665–678

Chou KC, Shen HB (2007c) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360: 339–345

Chou KC, Shen HB (2007d) Recent progress in protein subcellular location prediction. Anal Biochem 370: 1–16

Chou KC, Shen HB (2007e) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357: 633–640

Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269: 22014–22020

Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Clausmeyer S, Klosgen RB, Herrmann RG (1993) Protein import into chloroplasts. The hydrophilic lumenal proteins exhibit unexpected import and sorting specificities in spite of structurally conserved transit peptides. J Biol Chem 268: 13869–13876

Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept Lett 14: 811–815

Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7: 518

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

Endo T, Shimada I, Roise D, Inagaki F (1989) N-terminal half of a mitochondrial presequence peptide takes a helical conformation when bound to dodecylphosphocholine micelles: a proton nuclear magnetic resonance study. J Biochem (Tokyo) 106: 396–400

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499

Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biol 2: 291–303

Feng ZP, Zhang CT (2001) Prediction of the subcellular location of prokaryotic proteins based on the hydrophobicity index of amino acids. Int J Biol Macromol 28: 255–261

Gao QB, Wang ZZ, Yan C, Du YH (2005a) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579: 3444–3448

Gao Y, Shao S, Xiao X, Ding Y, Huang Y, Huang Z, Chou KC (2005b) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K, Brinkman FS (2003) PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. Nucleic Acids Res 31: 3613–3617

Garg A, Bhasin M, Raghava GP (2005) Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 280: 14427–14432

Guo J, Lin Y, Liu X (2006a) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6: 5099–5105

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006b) Classifying G protein-coupled receptors and nuclear receptors on the basis of protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Hammen PK, Gorenstein DG, Weiner H (1994) Structure of the signal sequences for two mitochondrial matrix proteins that are not proteolytically processed upon import. Biochemistry 33: 8610–8617

Hoglund A, Donnes P, Blum T, Adolph HW, Kohlbacher O (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs, and amino acid composition. Bioinformatics 22: 1158–1165

Hua S, Sun Z (2001) Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17: 721–728

Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20: 21–28

Jin L, Tang H, Fang W (2005) Prediction of protein subcellular locations using a new measure of information discrepancy. J Bioinform Comput Biol 3: 915–927

Kedarisetti KD, Kurgan L, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348: 981–988

Keegstra K, Cline K (1999) Protein import and routing systems of chloroplasts. Plant Cell 11: 557–570

Klee EW, Finlay JA, McDonald C, Attewell JR, Hebrink D, Dyer R, Love B, Vasmatzis G, Li TM, Beechem JM, Klee GG (2006) Bioinformatics methods prioritizing serum biomarker candidates. J Clin Chem 52: 2162–2164

Kurgan LA, Stach W, Ruan J (2007) Novel scales based on hydrophobicity indices for secondary protein structure. J Theor Biol 248: 354–366

Lee K, Kim DW, Na D, Lee KH, Lee D (2006) PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. Nucleic Acids Res 34: 4655–4666

Lee Y, Lee CK (2003) Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics 19: 1132–1139

Lei Z, Dai Y (2005) An SVM-based system for predicting protein subnuclear localizations. BMC Bioinformatics 6: 291

Lin H, Li QZ (2007a) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354: 548–551

Lin H, Li QZ (2007b) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28: 1463–1466

Liu DQ, Liu H, Shen HB, Yang J, Chou KC (2007) Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32: 493–496

Liu H, Wang M, Chou KC (2005a) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336: 737–739

Liu H, Yang J, Wang M, Xue L, Chou KC (2005b) Using fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J 24: 385–389

Mahdavi M, Lin Y-H (2007) False positive reduction in protein–protein interaction predictions using gene ontology annotations. BMC Bioinformatics 8: 262

Matsuda S, Vert JP, Saigo H, Ueda N, Toh H, Akutsu T (2005) A novel representation of protein sequences for prediction of subcellular location using support vector machines. Protein Sci 14: 2804–2813

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451

Mondal S, Bhavna R, Mohan Babu R, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243: 252–260

Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD (2007) Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. Pattern Recogn Lett 28: 1610–1615

Murphy RF, Boland MV, Velliste M (2000) Towards a systematics for protein subcelluar location: quantitative description of protein localization patterns and automated analysis of fluorescence microscope images. Proc Int Conf Intell Syst Mol Biol 8: 251–259

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

Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24: 34–36

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

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

Niu B, Cai YD, Lu WC, Li GZ, Chou KC (2006) Predicting protein structural class with AdaBoost Learner. Protein Pept Lett 13: 489–492

Pan YX, Zhang ZZ, Guo ZM, Feng GY, Huang ZD, He L (2003) Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. J Protein Chem 22: 395–402

Park KJ, Kanehisa M (2003) Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. Bioinformatics 19: 1656–1663

Pu X, Guo J, Leung H, Lin Y (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices. J Theor Biol 247: 259–265

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

Sarda D, Chua GH, Li KB, Krishnan A (2005) pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties. BMC Bioinformatics 6: 152

Shen H, Chou KC (2005a) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo-amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334: 288–292

Shen HB, Chou KC (2005b) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Commun 337: 752–756

Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722

Shen HB, Chou KC (2007a) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20: 39–46

Shen HB, Chou KC (2007b) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355: 1006–1011

Shen HB, Chou KC (2007c) Using ensemble classifier to identify membrane protein types. Amino Acids 32: 483–488

Shen HB, Chou KC (2007d) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85: 233–240

Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo-amino acid composition. J Theor Biol 240: 9–13

Shen HB, Yang J, Chou KC (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33: 57–67

Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33: 69–74

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

Vapnik V (1995) The nature of statistical learning theory. Springer-Verlag, New York

Wang M, Yang J, Chou KC (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids 28: 395–402

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. Protein Eng Des Sel 17: 509–516

Wang SQ, Yang J, Chou KC (2006) Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. J Theor Biol 242: 941–946

Ward JJ, McGuffin LJ, Buxton BF, Jones DT (2003) Secondary structure prediction with support vector machines. Bioinformatics 19: 1650–1655

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32: 277–283

Xiao X, Shao S, Ding Y, Huang Z, Chou KC (2006a) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006b) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xie D, Li A, Wang M, Fan Z, Feng H (2005) LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. Nucleic Acids Res 33: W105–W110

Yu CS, Lin CJ, Hwang JK (2004) Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions. Protein Sci 13: 1402–1406

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

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006a) Prediction of protein homo-oligomer types by pseudo amino acid composition: Approached with an improved feature extraction and Naive Bayes Feature Fusion. Amino Acids 30: 461–468

Zhang T, Ding Y, Chou KC (2006b) Prediction of protein subcellular location using hydrophobic patterns of amino acid sequence. Comput Biol Chem 30: 367–371

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33: 623–629

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17: 729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins 44: 57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins 50: 44–48

Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248: 546–551

**Authors' address:** Kuo-Bin Li, Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei 112, Taiwan,
Fax: +886 2 2820 2508, E-mail: kbli@ym.edu.tw