Amino Acids (2007) 33: 69–74
DOI 10.1007/s00726-006-0475-y
Printed in The Netherlands

Amino Acids

Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition

J.-Y. Shi, S.-W. Zhang, Q. Pan, Y.-M. Cheng, and **J. Xie**

College of Automation, Northwestern Polytechnical University, Xi'an, China

Received October 6, 2006
Accepted November 5, 2006
Published online January 19, 2007; © Springer-Verlag 2007

Summary. As more and more genomes have been discovered in recent years, there is an urgent need to develop a reliable method to predict the subcellular localization for the explosion of newly found proteins. However, many well-known prediction methods based on amino acid composition have problems utilizing the sequence-order information. Here, based on the concept of Chou's pseudo amino acid composition (PseAA), a new feature extraction method, the multi-scale energy (MSE) approach, is introduced to incorporate the sequence-order information. First, a protein sequence was mapped to a digital signal using the amino acid index. Then, by wavelet transform, the mapped signal was broken down into several scales in which the energy factors were calculated and further formed into an MSE feature vector. Following this, combining this MSE feature vector with amino acid composition (AA), we constructed a series of MSEPseAA feature vectors to represent the protein subcellular localization sequences. Finally, according to a new kind of normalization approach, the MSEPseAA feature vectors were normalized to form the improved MSEPseAA vectors, named as IEPseAA. Using the technique of IEPseAA, C-support vector machine (C-SVM) and three multi-class SVMs strategies, quite promising results were obtained, indicating that MSE is quite effective in reflecting the sequence-order effects and might become a useful tool for predicting the other attributes of proteins as well.

Keywords: Multi-scale energy – Wavelet transform – Support vector machines – Chou's pseudo amino acid composition – Protein subcellular localizations

1. Introduction

One of the big challenges in the biological field concerns structure and function classification in terms of protein sequences. The function of a protein is closely correlated with its subcellular localization (Chou, 2000; Chou and Cai, 2002). During the last decade, many theoretical methods were developed in an attempt to predict the subcellular localization of a query protein based on its sequence information. In 1994, Nakashima and Nishikawa indicated that intracellular and extracellular proteins are significantly different in amino acid composition (AA) (Nakashima and Nishikawa, 1994). The subsequent studies showed that the AA is closely related to protein subcellular localization (see, e.g., Chou and Elrod, 1999). However, under the following conditions, two sequences are completely different in function and localization but they have a very similar AA, so similar that if the prediction was only based on the AA, both of the proteins would be predicted as belonging to the same region of the cell. One of the reasons is that the AA method does not consider the effects of sequence order and sequence length. To solve this problem, Chou introduced a concept of the pseudo amino acid composition (PseAA) to partially incorporate the sequence-order effect of a protein (Chou, 2001, 2005). Stimulated by this concept, a series of powerful prediction algorithms and novel approaches have been developed to predict protein subcellular localization (Chou and Cai, 2002; Cui et al., 2004; Gao et al., 2005; Pan et al., 2003; Xiao et al., 2005a, b; Liu et al., 2005; Shen and Chou, 2005a, b; Chou and Shen, 2006a, b, c; Zhang et al., 2006). Here we attempt to develop a different PseAA, the so-called multi-scale energy (MSE) approach, and combine it with AA to represent protein sequence. MSE can effectively reflect the sequence-order effect. With this new combined feature, multi-class SVM is applied to predict protein subcellular localization.

2. Materials and methods

2.1 Database

The database used here was constructed by Chou, and it includes a training dataset and an independent testing dataset (Chou, 2001). The training dataset consists of 2191 protein sequences, made up as follows: 145 chloroplast, 571 cytoplasm, 34 cytoskeleton, 49 endoplasmic reticulum, 24 extracellular, 25 Golgi apparatus, 37 lysosome, 84 mitochondria, 272 nucleus, 27 peroxisome, 699 plasma membrane and 24 vacuole proteins. The independent dataset consists of 2494 protein sequences, made up as follows: 112 chloroplast, 761 cytoplasm, 19 cytoskeleton, 106 endoplasmic reticulum, 95 extracellular, 4 Golgi apparatus, 31 lysosome, 163 mitochondria, 418 nucleuse, 23 peroxisome, and 762 plasma membrane proteins.

2.2 Discrete wavelet transform

Wavelet transform is based on the idea of mapping a signal onto a set of basis functions. A set of wavelet basis functions can be generated by scaling and shifting the mother wavelet, according to the following formula:

$$\Psi_{a,b} = \frac{1}{\sqrt{a}}\Psi\left(\frac{x-b}{a}\right) \tag{1}$$

where a is a positive real number and b is a real number, which indicate the scale and the time shift of a basis function, respectively.

If we discretize the scale and the shift parameters to integer values, namely $a = 2^p$, $b = 2^p \cdot q$, then we can get discrete wavelet basis function $\Psi_{p,q}$ defined as the following formula:

$$\Psi_{p,q} = 2^{\frac{-p}{2}}\Psi(2^{-p}x - q) \tag{2}$$

Here, $p = 1, 2, \ldots$ and $q = 0, 1, 2, \ldots$. The wavelet coefficients of the signal $f(x)$ are obtained by following formula:

$$w_{p,q} = \langle f(x), \Psi_{p,q}(x)\rangle \tag{3}$$

Now, we get the discrete wavelet transform (DWT). Actually, instead of computing the inner product defined in formulation (3), there are several computationally efficient algorithms to implement DWT. Here, Mallat's fast algorithm is used (Mallat, 1999). The basic idea of the fast algorithm is to represent the mother wavelet as a set of high pass and low pass filter banks. The signal is passed through the filter banks and decimated by a factor of 2. The outputs of the low pass filter are wavelet approximation coefficients, and those of the high pass filter are wavelet detail coefficients. The filtering process mentioned above is just called discrete wavelet decomposition, and we will use it to extract feature information of protein subcellular localization in the following section.

2.3 Multi-scale energy feature

According to AA, the protein sequence p_k can be characterized as a 20-D feature vector:

$$AA_k = [c_1^k, \ldots, c_i^k, \ldots, c_{20}^k], \quad k = 1, \ldots, N \tag{4}$$

Here, $c_i^k = n^i/L_k$ is the normalized occurrence frequency of amino acid α_i, n^i is the count of α_i appearing in sequence p_k,

$$\alpha_i = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\},$$

L_k is the length of the sequence p_k.

It is not sufficient to characterize a specific protein sequence only on the basis of AA; the position information of α_i in the protein sequence and the correlation information between amino acids should also be considered. For example, suppose we have two protein sequences p_1: AAADDD and p_2: DDAA. According to the AA method, both of feature vectors are represented as the following:

$$AA_1 = [0.5, 0, 0.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]^T$$

$$AA_2 = [0.5, 0, 0.5, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]^T$$

Obviously, both of AA_1 and AA_2 feature vectors are the same. We cannot distinguish protein p_1 from protein p_2 only based on the AA method. So we must develop other methods to distinguish between proteins. The method of PseAA (Chou, 2001) is one of these methods. According to the concept of PseAA, AA is always incorporated with some factors reflecting sequence-order effects.

Since the sequence of a protein is a series of English letters, it is difficult to apply the digital signal processing (DSP) method directly. In order to apply DSP, each protein sequence should first be coded into a digital signal; that is, the sequence of English letters should be translated into a numerical sequence. As is well known, the hydrophilicity value of amino acid is one of the most important physicochemical properties which play a key role in the protein folding as well as function, particularly for subcellular localization. Here, we choose the index HOPT810101 from the amino acid index database (Kawashima et al., 1999) to map the residues to the corresponding numerical value. Hence, such a coded protein sequence can be treated as a digital signal and further processed by all existing tools of DSP, such as wavelet transform.

Projecting the signal onto a set of wavelet basis functions with various scales, the fine-scale and large-scale information of a protein hydrophilicity signal can be simultaneously investigated. Here, the wavelet basis function used is symlet wavelet. The features extracted from the wavelet-based multi-resolution information (Pittner and Kamarthi, 1999) can distinguish between different types of protein signals effectively. Consequently, sequence p_k can be characterized as an $(m+1)$-D MSE feature vector:

$$\text{MSE}_k = [d_1^k, \ldots, d_j^k, \ldots, d_m^k, a_m^k] \tag{5}$$

Here, m is the coarsest scale of decomposition, d_j^k is the root mean square energy of the wavelet detail coefficients in the corresponding jth scale, and a_m^k is the root mean square energy of the wavelet approximation coefficients in the scale m. The energy factors d_j^k and a_m^k are defined as:

$$d_j^k = \sqrt{\frac{1}{N_j}\sum_{n=0}^{N_j-1}[u_j^k(n)]^2}, \quad a_m^k = \sqrt{\frac{1}{N_m}\sum_{n=0}^{N_m-1}[v_m^k(n)]^2}, \quad j = 1, 2, \ldots, m \tag{6}$$

Here, N_j is the number of the wavelet detail coefficients, N_m is the number of the wavelet approximation coefficients, $u_j^k(n)$ is the nth detail coefficient in the corresponding jth scale, and $v_m^k(n)$ is the nth approximation coefficient in the scale m. For the protein sequence p_k with length L_k, m equals $INT(\log_2(L_k))$.

2.4 Improved pseudo amino acid composition

Obviously, MSE contains the approximate and detailed information of the protein signal, which reflects sequence-order effects. Combining MSE with AA, we can construct the following $(20+m+1)$-D feature vector X_k to represent sequence p_k.

$$X_k = [c_1^k, \ldots, c_i^k, \ldots, c_{20}^k, \lambda_1^k, \ldots, \lambda_j^k, \ldots, \lambda_m^k, \lambda_{m+1}^k]^T \tag{7}$$

Here, $\lambda_j^k = d_j^k, \lambda_{m+1}^k = a_m^k, j = 1, \ldots, m$. Based on the property of the dataset we are using, we choose $m = 11$.

According to Chou's PseAA, protein sequence p_k can be represented as following formula:

$$X_k = [x_1^k, \ldots, x_i^k, \ldots, x_{20+m+1}^k]^T$$
$$x_i^k = \begin{cases} \dfrac{c_i^k}{\sum_{j=1}^{20} c_j^k + \sum_{n=1}^{m+1} w_n\lambda_n^k}, & 1 \leq i \leq 20 \\ \dfrac{w_{i-20}\lambda_{i-20}^k}{\sum_{j=1}^{20} c_j^k + \sum_{n=1}^{m+1} w_n\lambda_n^k}, & 20+1 \leq i \leq 20+m+1 \end{cases} \tag{8}$$

Here, w_n is the weight factor of the nth MSE component.

In order to get better prediction results, different weights of MSE components should be obtained by certain optimal processes or some experimental approaches. These methods can work well for limited factors. When more and more components are integrated into PseAA, it is very difficult to optimize their weights, and it will even become an NP-Complete problem.

To integrate more components into PseAA, we introduce another normalization approach, described as follows:

$$\phi(x, i) = \frac{x_i - V^i_{\min}}{V^i_{\max} - V^i_{\min}} \tag{9}$$

Here, x_i is the ith attribute of the feature vector x, $V^i_{\max}$ is the maximum and $V^i_{\min}$ is the minimum of the ith attribute of the feature vector.

With such mapping functions $\{\phi(x, i)\}$, all attributes of training or testing samples can be scaled from their original ranges into [0, 1].

For example, suppose that the training samples contain the following three feature vectors:

$$Trn(1) = [0.1, 0.5, 1, 10, 100, 1000]$$
$$Trn(2) = [0.2, 0.8, 3, 15, 90, 1090]$$
$$Trn(3) = [0.5, 0.3, 4, 8, 120, 800]$$

The testing feature vector is $x = [0.3, 0.6, 2, 11, 100, 900]$. When the mapping is done, the feature vectors will be changed into the following vectors, respectively:

$$Trn'_1 = [0.00, 0.40, 0.00, 0.29, 0.33, 0.69]$$
$$Trn'_2 = [0.25, 1.00, 0.67, 1.00, 0.00, 1.00]$$
$$Trn'_3 = [1.00, 0.00, 1.00, 0.00, 1.00, 0.00]$$
$$x' = [0.50, 0.60, 0.33, 0.43, 0.33, 0.34]$$

Under this situation, we have no need to calculate all weights of components which are mapped into the same range. Naturally, the weight is chosen as 1.

This improved PseAA has two interesting properties: it can firstly avoid the need to calculate all weights and is able to prevent the attributes of greater numeric ranges from dominating those of smaller numeric ranges during the process of calculation. Conveniently, we use the symbol IEPseAA to represent this Improved Pseudo Amino Acid Composition method.

2.5 Weighted SVM

For some classification problems, such as the prediction of protein subcellular localization, numbers of data in different classes are imbalanced, and the results naturally tend to favor the majority class. Hence, some approaches of using different penalty parameters in the SVM formulation have been proposed to address the so-called class imbalance problem (Osuna et al., 1997). By introducing different penalty parameters, C-SVM (Vapnik, 1998) becomes

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C_+ \sum_{y_i=1} \xi_i + C_- \sum_{y_i=1} \xi_i$$
$$\text{subject to } y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i$$
$$\xi_i \geq 0, \quad i = 1, \ldots, N \tag{10}$$

Its dual is:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T Q \boldsymbol{\alpha} + \mathbf{e}^T\boldsymbol{\alpha}$$
$$\text{subject to } 0 \leq \alpha_i \leq C_+, \quad \text{if } y_i = 1$$
$$0 \leq \alpha_i \leq C_-, \quad \text{if } y_i = -1$$
$$\mathbf{y}^T\boldsymbol{\alpha} = 0 \tag{11}$$

Here, $C_+ = w_+ \cdot C$, $C_- = w_- \cdot C$ and C is the penalty parameter which is same for all classes in the original C-SVM. For the Ω-class problem, w_+ or w_- becomes w_ω which is the weight of ωth class and defined as follows: $w_\omega = N/N_\omega, \quad \omega = 1, \ldots, \Omega$. N_ω is the sample number of class ω in the training set and N is the total sample number in the training set.

2.6 Multi-class SVM

SVM has been proved to be a fruitful learning machine, especially for classification (Vapnik, 1998). It was originally designed for binary classification. We can construct Ω-class SVMs to solve the Ω-class classification problem based on the binary class SVM, which is an ongoing research issue.

There are mainly two kinds of approaches for multi-class SVM. One directly processes all data in one optimization formulation (Crammer and Singer, 2001). The other decomposes multi-class into a series of binary SVMs, including the "One-Versus-Rest" (OVR) strategy (Vapnik, 1998), the "One-Versus-One" (OVO) strategy (Kreßel, 1999), and the "Directed Acyclic Graph" (DAG) strategy (Platt et al., 2000). Extensive experiments have shown that OVR, OVO and DAG are practical (Hsu and Lin, 2002; Rifin and Klautau, 2004).

OVR is probably the earliest approach for multi-class SVM. For the Ω-class problem, it constructs Ω binary SVMs. The ith binary SVM is trained with all examples. The positive samples are taken from the ith class and negative samples are taken from the other classes. For a given test sample, all Ω binary SVMs are evaluated, and the test sample is labeled as the class with the largest value of the decision functions.

OVO constructs $\Omega(\Omega - 1)/2$ binary SVMs. Each binary SVM is trained with the examples from two different classes. During the evaluation, each of the $\Omega(\Omega - 1)/2$ SVMs casts one vote for its most favored class, and finally the class with the most votes wins (Kreßel, 1999).

DAG has the same training process as the OVO strategy, but it has a different evaluation process. During the evaluation, DAG uses the directed acyclic graph architecture to make a decision (Platt et al., 2000). The idea of DAG is easily implemented. Let $T = 1, 2, \ldots, \Omega$, which is a list of class labels. When a test sample is given, DAG first evaluates this sample with the binary SVM, which corresponds to the first and last element in list T. If the classifier prefers one of the two classes, then the other one will be eliminated from the list. After each testing, one class label will be excluded. Through $\Omega - 1$ binary SVM evaluating, the last label remaining in the list will be the answer.

Here, the SVM software we used is LIBSVM, which can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/ for academic research (Hsu and Lin, 2002). The RBF kernel is applied in all the following experiments.

2.7 System assessment

The jackknife test has been considered to be one of the most objective test methods in examining the power of a prediction method, as illustrated in a comprehensive review article (Chou and Zhang, 1995). It has been adopted by more and more leading investigators to test the powers of various predictors (see, e.g., Chou, 1995; Chou and Cai, 2004; Gao et al., 2005, 2006; Liu et al., 2005; Shen and Chou, 2005a, b, 2006; Shen et al., 2005, 2006; Sun and Huang, 2006; Wen et al., 2006; Xiao et al., 2005a, 2006b, c; Zhang et al., 2006; Zhou, 1998). During the process of the jackknife test, each protein in the training dataset is singled out in turn as a test sample, and the remaining proteins are used as training samples. For an independent test, proteins in the training dataset are used to train the rule parameters, and those in the independent test dataset are used as test samples. The prediction for an independent test dataset is just for a demonstration of practical application.

To assess the quality of the test, the total prediction accuracy and the prediction accuracy of the each protein subcellular localization can be defined as:

$$\text{Total accuracy} = \frac{1}{N}\sum_{\omega=1}^{\Omega} p(\omega) \tag{12}$$

$$\text{accuracy}(\omega) = \frac{p(\omega)}{obs(\omega)} \tag{13}$$

Here, N is the total number of proteins, Ω is the sum of the classes, $obs(\omega)$ is the number of proteins observed in class ω and $p(\omega)$ is the number of proteins correctly predicted in class ω.

3. Results and discussion

3.1 Results of different prediction methods

The results of our IEPseAA approach using the C-SVM and OVO multi-class SVM classification strategy are shown in Table 1. With the same dataset constructed by Chou, the results of other five methods (Chou, 2001; Pan et al., 2003; Xiao et al., 2005a, b; Gao et al., 2005) are also listed in Table 1.

From Table 1, we can see that the total accuracy of IEPseAA in the jackknife and independent tests is 80.3 and 87.0%, respectively. These degrees of accuracy are remarkably higher than those of the other methods listed in Table 1. For example, in the jackknife test, the total accuracy of IEPseAA is 12.6 and 6.7% higher than that of Pan's method (Pan et al., 2003) and Xiao's method (Xiao et al., 2005b), respectively. These results show that IEPseAA is effective and helpful for the prediction of protein subcellular localization. MSE can extract more sequence-order information.

3.2 Comparison of three multi-class SVM strategies

In order to make the comparison among three multi-class SVMs (OVR, OVO and DAG), based on LIBSVM with some modification of its source codes, we have also adopted DAG and OVR strategies to predict protein subcellular localization with C-SVM. The results are shown in Table 2.

Table 1. Total accuracy (%) of IEPseAA and other methods on Chou's dataset (2001) by the jackknife and independent tests

Method	Jackknife test	Independent test
Chou's (Chou, 2001)	1600/2191 = 73.0%	2017/2494 = 80.9%
Pan's (Pan et al., 2003)	1483/2191 = 67.7%	1842/2494 = 73.9%
Xiao's (Xiao et al., 2005a)	1590/2191 = 72.6%	1865/2494 = 74.8%
Xiao's (Xiao et al., 2005b)	1612/2191 = 73.6%	1990/2494 = 79.8%
Gao's (Gao et al., 2005)	1532/2191 = 69.9%	–
IEPseAA	1759/2191 = 80.3%	2170/2494 = 87.0%

Table 2. Total accuracy (%) of DAG, OVR and OVO with IEPseAA and C-SVM

Multi-class SVM strategy	Jackknife test	Independent test
DAG	1760/2191 = 80.3%	2167/2494 = 86.9%
OVR	1764/2191 = 80.5%	2163/2494 = 86.7%
OVO	1759/2191 = 80.3%	2170/2494 = 87.0%

Table 3. The number of support vectors and the consumed time of DAG, OVR and OVO with IEPseAA and C-SVM

Strategy	Training			Independent test	
	SV	Max	Min	Max	Min
DAG	1573	2.141	2.109	1.578	1.562
OVR	1686	4.875	4.843	1.785	1.766
OVO	1573	2.047	2.000	1.672	1.563

Table 2 shows that DAG, OVR and OVO have very similar classification accuracy, and the difference among them may be mainly focused on the number of support vectors, the training time and the testing time. To validate this guess further, we ran training and independent tests 10 times in the computer with Pentium IV 2.0G CPU and 256 Mb memories, respectively. The number of support vectors (SV) and their maximum (Max) and minimum (Min) time are listed in Table 3.

Although OVR only requires Ω binary SVMs, each binary SVM is optimized on all the N training examples. OVO or DAG has $\Omega(\Omega-1)/2$ binary SVMs to train; however, the total training time of OVO or DAG is still less than that of OVR. The reason is that the individual binary SVM of OVO or DAG is only trained on the examples of two classes. In our experiments, OVR has a heavy training computational burden, whose training time is almost 2.5 times that of OVO or DAG.

During the testing, we found that the testing time was almost proportional to the number of support vectors, and OVR takes much more time than OVO and DAG. In addition, DAG is a little faster than OVO in testing time, but it occupies a little bit more memory than OVO because it needs extra data structure to index the set of these binary SVMs.

As described above, except for the training time, DAG, OVO and OVR are very similar in terms of their other performance. Hence, we think that DAG and OVO may be more suitable for predicting protein subcellular localization.

3.3 Effect of weighted SVM on the results

As is well known, C-SVM is often biased toward the class with more examples in an imbalanced dataset. Chou's protein subcellular localization dataset is an imbalanced database. In order to deal with this problem, the training with weight factor is used to improve the prediction accuracy with a small sample size.

Table 4. Prediction accuracy (%) of C-SVM and weighted SVM with OVO strategy and the same IEPseAA

Localization	Jackknife test		Independent test	
	C-SVM	Weighted SVM	C-SVM	Weighted SVM
Chloroplast	63.4	70.3	69.6	74.1
Cytoplasm	90.2	85.5	90.0	86.6
Cytoskeleton	44.1	47.1	94.7	94.7
Endoplasmic reticulum	40.8	42.9	84.9	89.6
Extracellular	67.4	66.5	75.8	78.9
Golgi apparatus	20.0	28.0	50.0	50.0
Lysosome	48.6	48.6	80.6	80.6
Mitochondria	37.0	35.7	57.1	61.4
Nucleus	86.0	85.7	81.3	82.3
Peroxisome	7.4	22.2	43.5	69.6
Plasma membrane	96.1	95.1	99.3	98.3
Vacuole	16.7	25.0	–	–
Total accuracy	80.3	79.8	87.0	86.8

Here, we apply weighted SVM to Chou's dataset. The prediction accuracies of 12-class subcellular localizations are listed in Table 4.

From Table 4, we can see that: 1) With weighted SVM, the accuracies of most classes which have small number of samples can be improved. For example, the accuracies of peroxisome, vacuole and Golgi apparatus of weighted SVM, which have 27, 24 and 25 training examples, are 22.2, 25 and 28%, respectively in the jackknife test, and these figures are 14.8, 8.3 and 8.0% higher than those of C-SVM, respectively; 2) Compared with C-SVM, the accuracies of the classes which have more samples will fall only a little with weighted SVM. For example, the accuracy of plasma membrane (whose training sample number is 699) and cytoplasm (whose training sample number is 571) with weighted SVM are 1 and 4.7% lower than that with C-SVM, respectively, in the jackknife test. The reason may be that the weighted factor w_ω selected is not suitable. The question of how to select a suitable weighted factor w_ω is our next research work.

4. Conclusion

A novel feature extraction method, the multi-scale energy approach, which calculates the root mean square energy of the wavelet transform coefficients in different scales to reflect sequence-order effects, was proposed in this paper. Furthermore, a new kind of normalization approach by combining MSE with AA to construct the improved PseAA (IEPseAA) was formulated. On the basis of such a frame, multi-class SVMs were adopted to predict protein subcellular localization, and better results were obtained.

Compared with the other PseAA approaches, the current IEPseAA approach can more effectively reflect the sequence-order effects for predicting protein subcellular localization. Moreover, it is also indicated that the weighted SVM can solve the problem of C-SVM biasing toward the class with more samples in an imbalanced dataset.

Acknowledgements

This paper was supported in part by the National Natural Science Foundation of China (No. 60372085) and the Technological Innovation Foundation of Northwestern Polytechnical University (No. KC02).

References

Chou KC (1995) A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins Struct Funct Genet 21: 319–344

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

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

Chou KC (2005) Review: Progress in protein structural class prediction and its impact to bioinformatics and proteomics. Curr Protein Peptide Sci 6: 423–436

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

Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321: 1007–1009 (Corrigendum: ibid., 2005, Vol. 329, 1362)

Chou KC, Elrod DW (1999) Protein subcellular localization 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, Zhang CT (1995) Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Crammer K, Singer Y (2001) On the algorithmic implementation of multiclass kernel-based vector machines. J Mach Learn Res 2: 265–292

Cui Q, Jiang T, Liu B, Ma S (2004) Esub8: A novel tool to predict protein subcellular localizations in eukaryotic organisms. BMC Bioinform 5: 66–72

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

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

Hsu C, Lin CJ (2002) A comparison of methods for multi-class support vector machines. IEEE Trans Neural Networks 13: 415–425

Kawashima S, Ogata H, Kanehisa M (1999) AAIndex: amino acid index database. Nucleic Acids Res 27: 368–369

Kreßel UH (1999) Pairwise classification and support vector machines. In: Schölkopf B, Burges CJ, Smola AJ (eds) Advances in kernel methods: support vector learning. MIT Press, Cambridge, MA, pp 255–268

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

Mallat S (1999) A wavelet tour of signal processing. Academic Press, New York

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

Osuna E, Freund R, Girosi F (1997) Support vector machines: Training and applications. AI Memo 1602, MIT, Cambridge, MA

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

Pittner S, Kamarthi SV (1999) Feature extraction from wavelet coefficients for pattern recognition tasks. IEEE Trans Pattern Anal Mach Intell 21: 83–88

Platt J, Cristianini N, Shawe-Taylor J (2000) Large margin DAGs for multiclass classification. Adv Neural Inform Proc Syst 12: 547–553

Rifin R, Klautau A (2004) In defense of one-vs-all classification. J Mach Learn Res 5: 101–141

Shen HB, Chou KC (2005a) 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 (2005b) 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 (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722

Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334: 577–581

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

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

Vapnik V (1998) Statistical learning theory. Wiley, New York

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 (in press) (DOI: 10.1007/s00726-006-0341-y)

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2005a) Using cellular automata images and pseudo amino acid composition to predict protein subcellular localization. Amino Acids 30: 49–54

Xiao X, Shao SH, Ding YS, Huang ZD, Huang Y, Chou KC (2005b) Using complexity measure factor to predict protein subcellular localization. Amino Acids 28: 57–61

Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC (2005c) An application of gene comparative image for predicting the effect on replication ratio by HBV virus gene missense mutation. J Theor Biol 235: 555–565

Xiao X, Shao SH, Chou KC (2006a) A probability cellular automaton model for hepatitis B viral infections. Biochem Biophys Res Commun 342: 605–610

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

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction 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

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

Authors' address: S.-W. Zhang, College of Automation, Northwestern Polytechnical University, Xi'an 710072, China,
E-mail: Zhangsw@nwpu.edu.cn