Journal of Cellular Biochemistry 99:517–527 (2006)

Predicting Protein Subcellular Location by Fusing Multiple Classifiers

Kuo-Chen Chou[1,2]* and Hong-Bin Shen[2]

[1]Gordon Life Science Institute, 13784 Torrey Del Mar, San Diego, California 92130
[2]Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China

Abstract One of the fundamental goals in cell biology and proteomics is to identify the functions of proteins in the context of compartments that organize them in the cellular environment. Knowledge of subcellular locations of proteins can provide key hints for revealing their functions and understanding how they interact with each other in cellular networking. Unfortunately, it is both time-consuming and expensive to determine the localization of an uncharacterized protein in a living cell purely based on experiments. With the avalanche of newly found protein sequences emerging in the post genomic era, we are facing a critical challenge, that is, how to develop an automated method to fast and reliably identify their subcellular locations so as to be able to timely use them for basic research and drug discovery. In view of this, an ensemble classifier was developed by the approach of fusing many basic individual classifiers through a voting system. Each of these basic classifiers was trained in a different dimension of the amphiphilic pseudo amino acid composition (Chou [2005] Bioinformatics 21: 10–19). As a demonstration, predictions were performed with the fusion classifier for proteins among the following 14 localizations: (1) cell wall, (2) centriole, (3) chloroplast, (4) cytoplasm, (5) cytoskeleton, (6) endoplasmic reticulum, (7) extracellular, (8) Golgi apparatus, (9) lysosome, (10) mitochondria, (11) nucleus, (12) peroxisome, (13) plasma membrane, and (14) vacuole. The overall success rates thus obtained via the resubstitution test, jackknife test, and independent dataset test were all significantly higher than those by the existing classifiers. It is anticipated that the novel ensemble classifier may also become a very useful vehicle in classifying other attributes of proteins according to their sequences, such as membrane protein type, enzyme family/sub-family, G-protein coupled receptor (GPCR) type, and structural class, among many others. The fusion ensemble classifier will be available at www.pami.sjtu.edu.cn/people/hbshen. J. Cell. Biochem. 99: 517–527, 2006.

Key words: cellular networking; subcellular compartment; ensemble classifier; fusion; voting; covariant discriminant algorithm; amphiphilic pseudo amino acid composition

The human body hosts 10^{14} cells [Radford, 2003]. A cell contains approximately 10^9 protein molecules that are located in many different compartments, or organelles (Fig. 1). Cell membrane functions as a boundary layer to contain the cytoplasm, while cell wall provides protection from physical injury. The cytoplasm, a jelly-like material, fills the cell and serves as a "molecular soup" in which all of the cell's organelles are suspended. The function of the cytoplasm and the organelles which sit in it, are critical to the cell's survival. The organelles are specialized to carry out different tasks. For instance, functioning as the "brain" of eukaryotic cells, nucleus houses the deoxyribonucleic acid (DNA), which stores genetic information. Chloroplast is the site of photosynthesis. Vacuole stores water and various chemicals. Centriole forms spindle fibers to separate chromosomes during cell division. Endoplasmic reticulum transports chemicals between cells and within cells. Golgi apparatus modifies chemicals to make them functional. Mitochondrion is the site of cellular respiration, that is, the release of chemical energy from food. Lysosome breaks large molecules into small molecules by inserting a molecule of water into the chemical bond. Peroxisome breaks down excess fatty acids and hydrogen peroxide (H_2O_2),

This article contains supplementary material, which may be viewed at the Journal of Cellular Biochemistry website at http://www.interscience.wiley.com/jpages/0730-2312/suppmat/index.html.

*Correspondence to: Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar, San Diego, CA 92130. E-mail: kchou@san.rr.com

Received 24 October 2005; Accepted 7 December 2005
DOI 10.1002/jcb.20879

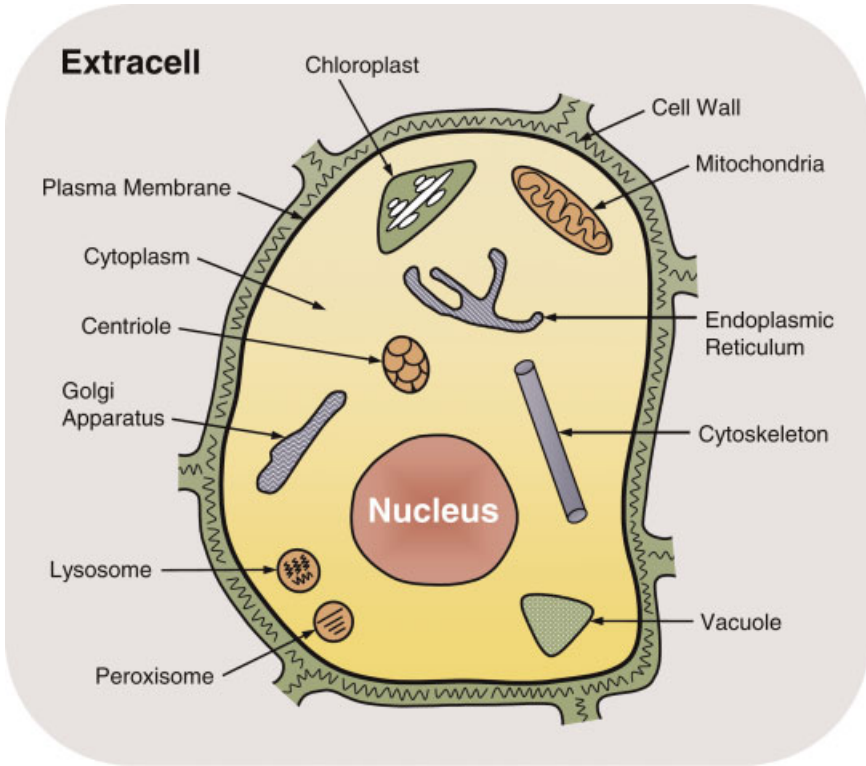

Fig. 1. Schematic illustration to show the 14 subcellular locations of proteins: (1) cell wall, (2) centriole, (3) chloroplast, (4) cytoplasm, (5) cytoskeleton, (6) endoplasmic reticulum, (7) extracellular, (8) Golgi apparatus, (9) lysosome, (10) mitochondria, (11) nucleus, (12) peroxisome, (13) plasma membrane, and (14) vacuole. Note that the cell wall, chloroplast, and vacuole proteins exist only in a plant cell, while the centriole proteins only in an animal cell. Reproduced from Chou and Cai [2003c] with permission. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

a potentially dangerous product of fatty-acid oxidation. The cytoskeleton is responsible for establishing cell shape, providing mechanical strength, locomotion, and intracellular transport of organelles. Most of these functions are performed by the proteins in a cell. Accordingly, the significance to identify the subcellular localization of an uncharacterized protein has become self-evident.

Although the information about protein subcellular localization can be determined by conducting various experiments, that is both time-consuming and costly. Particularly, the number of newly found protein sequences has increased explosively in the post genomic era. For instance, in 1986 Swiss-Prot [Bairoch and Apweiler, 2000] contained only 3,939 protein sequence entries, but now the number has jumped to 201,594 according to the version 48.6 released on December 6, 2005, implying that the number of protein sequences has increased by more than 50 times in less than two decades. Facing such an overwhelming number of newly found protein sequences, it is both challenging and urgently needed to develop an automated method for fast and reliably annotating the subcellular attributes of uncharacterized proteins. The knowledge thus obtained can help us timely utilize these newly found protein sequences for both basic research and drug discovery [Chou, 2004].

Actually, many efforts have been made in this regard [Nakashima and Nishikawa, 1994; Cedano et al., 1997; Nakai and Horton, 1999; Yuan, 1999; Chou and Elrod, 1999a,b; Nakai, 2000; Chou, 2001a; Chou and Cai, 2002; Pan et al., 2003; Park and Kanehisa, 2003; Zhou and Doctor, 2003; Gao et al., 2005; Garg et al., 2005; Shen and Chou, 2005b; Xiao et al., 2005c,d]. However, all these prediction methods were established based on a single classifier derived from a single training process regardless of whether the operation was performed with the covariant discriminant algorithm, or support vector machine (SVM), or neural network. Obviously, using a single classifier to deal with complicated protein sequences with extreme variation in both sequence order and length will

certainly limit the optimal result. The present study was initiated in an attempt to introduce the ensemble classifier by fusing many individual classifiers. The advantage by doing so is in reducing the variance caused by the peculiarities of a single training process so as to better grasp the overall expressive feature for conducting classification.

METHOD

In order to better reflect the sequence order and length effect, rather than the conventional amino acid composition, we adopt the pseudo amino acid composition [Chou, 2001a, 2005a] to represent the sample of a protein via a discrete model. However, instead of a fixed dimension, the pseudo amino acid composition adopted here is with a series of various dimensions. Below, let us first give a brief introduction about this.

Amphiphilic Pseudo Amino Acid Composition

Given a protein $\mathbf{P}$ with L amino acid residues

$$\mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\ldots\mathrm{R}_L \tag{1}$$

where R_1 represents the residue at the sequence position 1, R_2 at position 2, and so forth, its amphiphilic pseudo amino acid can be generally formulated as (see Appendix A):

$$\mathbf{P} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{20} \\ \vdots \\ p_\Lambda \end{bmatrix} \tag{2}$$

where the first 20 elements $p_1, p_2, \ldots, p_{20}$ are associated with the 20 components in the conventional amino acid composition (see Appendix A), the elements from p_{20+1} to p_Λ are Λ correlation factors through which the sequence-order effects can be indirectly reflected (see Fig. A1 of Appendix A). When $\Lambda = 20$, the pseudo amino acid composition is reduced to the conventional amino acid composition. The elements in Equation 2 can be easily calculated by following the procedures given in Appendix A. Although the dimension of pseudo amino acid composition (Λ) is allowed to vary, it is limited by certain conditions, depending on what kind of mode is used. For the amphiphilic alternative-mode adopted here, $\Lambda \leq 20 + 2(L-1)$ (see Appendix A).

Covariant Discriminant Classifier

Given a dataset, S, of N proteins classified into M cellular attributes, we can generally formulate it in terms of the union of M subsets; that is,

$$S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup \ldots \cup S_M \tag{3}$$

where each subset S_m $(m = 1, 2, \ldots, M)$ is composed of proteins with the same cellular attribute and its size (the number of proteins therein) is N_m. Obviously, we have $N = N_1 + N_2 + \ldots + N_M$. According to Equation 2, the kth protein in the subset S_m is formulated by

$$\mathbf{P}_m^k = \begin{bmatrix} p_{m,1}^k \\ p_{m,2}^k \\ \vdots \\ p_{m,20}^k \\ \vdots \\ p_{m,\Lambda}^k \end{bmatrix} \tag{4}$$

The standard vector for the subset S_m is defined by

$$\bar{\mathbf{P}}_m = \begin{bmatrix} \bar{p}_{m,1} \\ \bar{p}_{m,2} \\ \vdots \\ \bar{p}_{m.20} \\ \vdots \\ \bar{p}_{m,\Lambda} \end{bmatrix} \tag{5}$$

where

$$\bar{p}_{m,i} = \frac{1}{N_m}\sum_{k=1}^{N_m} p_{m,i}^k, \quad (i = 1, 2, \ldots, \Lambda) \tag{6}$$

Actually, $\bar{\mathbf{P}}_m$ as defined above can be deemed as a standard protein for the subset S_m. The similarity between proteins $\mathbf{P}$ (Eq. 2) and $\bar{\mathbf{P}}_m$ (Eq. 5) is defined by the following covariant discriminant function:

$$\mathfrak{M}(\mathbf{P}, \bar{\mathbf{P}}_m) = D_{\mathrm{Mar}}^2(\mathbf{P}, \bar{\mathbf{P}}_m) + \ln|\mathbf{C}_m|, \quad (m = 1, 2, \ldots, M) \tag{7}$$

where

$$D_{\mathrm{Mar}}^2(\mathbf{P}, \bar{\mathbf{P}}_m) = (\mathbf{P} - \bar{\mathbf{P}}_m)^{\mathrm{T}}\mathbf{C}_m^{-1}(\mathbf{P} - \bar{\mathbf{P}}_m) \tag{8}$$

is the squared Mahalanobis distance [Mahalanobis, 1936; Pillai, 1985; Chou and Zhang, 1994;

Chou, 1995] between $\mathbf{P}$ and $\bar{\mathbf{P}}_m$, $\mathbf{T}$ is the transpose operator, and

$$\mathbf{C}_m = \begin{bmatrix} c^m_{1,1} & c^m_{1,2} & \cdots & c^m_{1,\Lambda} \\ c^m_{2,1} & c^m_{2,2} & \cdots & c^m_{2,\Lambda} \\ \vdots & \vdots & \ddots & \vdots \\ c^m_{\Lambda,1} & c^m_{\Lambda,2} & \cdots & c^m_{\Lambda,\Lambda} \end{bmatrix} \quad (9)$$

is the covariance matrix for the subset S_m and its $\Lambda \times \Lambda$ elements are given by

$$c^m_{i,j} = \frac{1}{N_m - 1} \sum_{k=1}^{N_m} (p^k_{m,i} - \bar{p}_{m,i})(p^k_{m,j} - \bar{p}_{m,j}), \quad (i,j = 1,2,\ldots,\Lambda) \quad (10)$$

and $|\mathbf{C}_m|$ is the determinant of the matrix $\mathbf{C}_m$. The smaller the value of $M(\mathbf{P}, \bar{\mathbf{P}}_m)$, the greater the similarity between $\mathbf{P}$ and $\bar{\mathbf{P}}_m$. Therefore, the classifier can be formulated as follows:

$$\mathfrak{M}(\mathbf{P}, \bar{\mathbf{P}}_\mu) = \mathrm{Min}\{\mathfrak{M}(\mathbf{P}, \bar{\mathbf{P}}_1), \mathfrak{M}(\mathbf{P}, \bar{\mathbf{P}}_2), \ldots, \mathfrak{M}(\mathbf{P}, \bar{\mathbf{P}}_M)\} \quad (11)$$

where the operator Min means taking the least one among those in the brackets, and the subscript μ (=1, 2, 3, ..., or M) is the very subset which the query protein $\mathbf{P}$ belongs to.

Fusion of Individual Classifiers

As we can see from Equations 7–11, the classifier is closely associated with Λ, the dimension of the pseudo amino acid composition. Therefore, even for exactly the same training dataset, using different value of Λ will yield different result. Suppose

$$\{\Lambda\} = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_\Omega\} \quad (12)$$

represents a set of possible numbers for the dimensions of pseudo amino acid composition, then we have a set of corresponding classifiers as formulated by

$$\{\mathrm{CD}(\Lambda)\} = \{\mathrm{CD}(\Lambda_1), \mathrm{CD}(\Lambda_2), \ldots, \mathrm{CD}(\Lambda_\Omega)\} \quad (13)$$

where CD (Λ_1) is the covariant discriminant classifier trained in the Λ_1 dimensional space, CD (Λ_2) is the one in the Λ_2 dimensional space, and so forth. The ensemble classifier formed by fusing such a set of individual classifiers is formulated by

$$\mathbb{C} = \mathrm{CD}(\Lambda_1) \oplus \mathrm{CD}(\Lambda_2) \oplus \ldots \oplus \mathrm{CD}(\Lambda_\Omega) \quad (14)$$

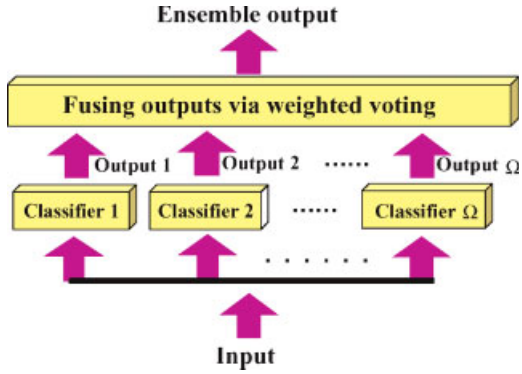

Fig. 2. Flowchart to show how the ensemble classifier $\mathbb{C}$ (Eq. 14) is formed by fusing Ω individual classifiers: CD (Λ_1), CD (Λ_2), ..., and CD (Λ_Ω), where $\Omega = 22$ and Λ_1, Λ_2, ..., Λ_{22} were equal to 20, 22, ..., 62, respectively, in this study. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

where the symbol $\oplus$ denotes the fusing operator, and $\mathbb{C}$ the ensemble classifier formed by fusing CD (Λ_1), CD (Λ_2), ..., and CD (Λ_Ω) according to the flowchart of Figure 2. In this study, Ω was set at 22, and Λ_1, Λ_2,...,Λ_{22} at 20, 22,..., 62, respectively; that is, $\Lambda_i = 20 + (i-1) \times 2$ with $(i = 1, 2, \ldots, 22)$, meaning that CD (Λ_1) was trained with the first 20 components of Equation 2, CD (Λ_2) trained with the first 22 components, and so forth.

The process of how the ensemble classifier $\mathbb{C}$ works is as follows. Suppose the predicted classification results for the query protein $\mathbf{P}$ by $\Omega = 22$ individual classifiers are $Q_1, Q_2, \ldots, Q_\Omega$, respectively; that is,

$$\{Q_1, Q_2, \ldots, Q_\Omega\} \in \{S_1, S_2, \ldots, S_M\} \quad (15)$$

and the voting score for the protein $\mathbf{P}$ belonging to the jth subset is defined by

$$Y_j = \sum_{i=1}^{\Omega=22} w_i \Delta(Q_i, S_j), \quad (j = 1, 2, \ldots, M) \quad (16)$$

where w_i is the weighted factor, which can be assigned according to some rule to optimize the predicted results. For simplicity, let us just set $w_i = 1$ in this study. And the delta function in Equation 16 is given by

$$\Delta(Q_i, S_j) = \begin{cases} 1, & \text{if } Q_i \in S_j \\ 0, & \text{otherwise} \end{cases} \quad (17)$$

thus the query protein $\mathbf{P}$ is predicted belonging to the class with which its score of Equation 16 is the highest; that is, suppose

$$Y_\mu = \mathrm{Max}\{Y_1, Y_2, \ldots, Y_M\} \quad (18)$$

where the operator Max means taking the maximum one among those in the brackets,

and the subscript μ is the very attribute predicted for the query protein **P**. If there is a tie, the query protein may not be uniquely determined and will be randomly assigned among those with a tie, but cases like that rarely occur.

RESULTS AND DISCUSSION

To demonstrate the power of the ensemble classifier, the same training and testing datasets investigated by Chou and Cai [2003c] were used. Each of the two datasets covers 14 subcellular locations. The training dataset contains 3,799 proteins, of which (1) 71 are of cell wall, (2) 65 of centriole, (3) 316 of chloroplast, (4) 1,113 of cytoplasm, (5) 249 of cytoskeleton, (6) 289 of endoplasmic reticulum, (7) 393 of extracell, (8) 90 Golgi apparatus, (9) 123 of lysosome, (10) 389 of mitochondria, (11) 399 of nucleus, (12) 147 of peroxisome, (13) 69 of plasma membrane, and (14) 86 of vacuole (Fig. 1). The independent testing dataset contains 4,498 proteins, of which (1) 35 are of cell wall, (2) 4 of centriole, (3) 855 of chloroplast, (4) 186 of cytoplasm, (5) 131 of cytoskeleton, (6) 136 of endoplasmic reticulum, (7) 1,252 of extracell, (8) 41 Golgi apparatus, (9) 57 of lysosome, (10) 762 of mitochondria, (11) 914 of nucleus, (12) 84 of peroxisome, (13) 24 of plasma membrane, and (14) 17 of vacuole. The codes of these proteins in the training and independent testing datasets can be found at http://www.interscience.wiley.com/jpages/0730-2312/suppmat. Compared with the two datasets, most of the existing datasets cover much less locations. For instance, the datasets investigated by Nakashima and Nishikawa [1994] only covered two locations, those by Cedano et al. [1997] five locations, those by Yuan [1999] three or four locations, and those by Garg et al. [2005] four locations.

The demonstration was conducted by the three most typical approaches in statistical prediction [Chou and Zhang, 1995]; that is, the re-substitution test, independent dataset test, and jackknife test, as reported below.

Re-Substitution Test

The so-called re-substitution test is an examination for the self-consistency of a classifier. When the re-substitution test is performed for the current study, the subcellular location of each protein in the data set is in turn identified using the rule parameters derived from the same data set, the so-called training dataset. The success rate thus obtained for predicting the 14 subcellular locations of the 3,799 proteins is summarized in Table I, from which we can see that 3,280 proteins were correctly predicted for their subcellular locations, and only 519 proteins incorrectly predicted. The overall success rate was 86.4%. However, during the process of the re-substitution test, the rule parameters derived from the training data set include the information of the query protein later plugged back in the test. This will certainly underestimate the error and enhance the success rate because the same proteins are used to derive the rule parameters and to test themselves. Accordingly, the success rate thus obtained represents an optimistic estimation [Chou and Maggiora, 1998; Chou et al., 1998;

TABLE I. Overall Success Rates for the 14 Subcellular Locations (Fig. 1) of Proteins by Different Classifiers and Test Methods

		Test method		
Classifier	Input form	Resubstitution	Jackknife	Independent datset
ProtLock [Cedano et al., 1997]	Amino acid composition	$\frac{1,655}{3,799} = 43.6\%$	$\frac{1,614}{3,799} = 42.5\%$	$\frac{1,829}{4,498} = 40.7\%$
Covariant discriminant [Chou and Elrod, 1999b]	Amino acid composition	$\frac{2,580}{3,799} = 67.9\%$	$\frac{2,339}{3,799} = 61.6\%$	$\frac{2,751}{4,498} = 61.2\%$
Augmented covariant discriminant [Chou, 2001a]	Pseudo amino acid composition[a]	$\frac{3,245}{3,799} = 85.4\%$	$\frac{2,574}{3,799} = 67.8\%$	$\frac{3,246}{4,498} = 72.2\%$
Ensemble	Pseudo amino acid composition[b]	$\frac{3,280}{3,799} = 86.4\%$	$\frac{2,666}{3,799} = 70.2\%$	$\frac{3,331}{4,498} = 74.1\%$

[a]The series-mode [Chou and Cai, 2003c] was used to calculate the pseudo amino acid composition with $\Lambda = 20 + \lambda + \mu = 20 + 13 + 13 = 46$.
[b]The amphiphilic mode (Appendix A) was used to calculate the pseudo amino acid composition with $\{\Lambda\} = \{\Lambda_1, \Lambda_2, \ldots, \Lambda_{22}\} = \{20, 22, \ldots, 62\}$ (cf. Eq. 12).

Cai, 2001; Zhou and Assa-Munt, 2001]. Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of a predictor, especially for its algorithm part. A prediction algorithm certainly cannot be deemed as a good one if its self-consistency is poor. In other words, the re-substitution test is necessary but not sufficient for evaluating a predictor. As a complement, a cross-validation test for an independent testing data set is needed because it can reflect the effectiveness of a predictor in practical application. This is important especially for checking the validity of a training database: whether it contains sufficient information to reflect all the important features concerned so as to yield a high success rate in application.

Independent Dataset Test

As a showcase for practical application, predictions were also performed for the aforementioned 4,498 proteins in the independent dataset based on the rule-parameters derived from the 3,799 proteins in the training dataset. The overall success rate thus obtained for the 4,498 proteins is given in Table I as well.

Jackknife Test

As is well known, the independent data set test, sub-sampling test, and jackknife test are the three methods often used for cross-validation in statistical prediction. Among these three, however, the jackknife test is deemed as the most rigorous and objective one, as discussed by a comprehensive review [Chou and Zhang, 1995]. Therefore, jackknife test has been used by more and more investigators [Zhou, 1998; Yuan, 1999; Feng, 2001; Hua and Sun, 2001; Zhou and Assa-Munt, 2001; Chou, 2001b; Luo et al., 2002; Pan et al., 2003; Zhou and Doctor, 2003; Liu et al., 2005b; Wang et al., 2005; Shen and Chou, 2005a,b; Shen et al., 2005a,b; Xiao et al., 2005a,c] in examining the power of various predictors. During jackknifing, each protein in the dataset is in turn singled out as a tested protein and all the rule-parameters are calculated based on the remaining proteins. In other words, the subcellular location of each protein is identified by the rule parameters derived using all the other proteins except the one being identified. During the process of jackknifing both the training data set and testing data set are actually open, and a protein will in turn move from one to the other. The overall jackknife success rate thus obtained for the 3,799 proteins in the training dataset is also given in Table I.

Furthermore, to facilitate comparison, listed in Table I are also the results predicted by various other methods on the same datasets. Meanwhile, to show the advantage of the ensemble classifier, the overall jackknife success rate obtained by each of the individual classifiers is listed in Table II. From the two tables, the following can be observed. (1) The current ensemble classifier remarkably outperformed the other classifiers in all the three test methods, indicating that the ensemble classifier is indeed a very powerful one. (2) Among the three test methods, the success rate obtained by re-substitution is the highest (86.4%), that by the independent dataset test is the next (74.1%), and that by the jackknife test is the least (70.2%). This is fully consistent

TABLE II. The Jackknife Success Rate Obtained by Each of the 22 Individual Classifiers (cf. Eqs. 13 and 14)

Classifier[a]	Dimension[b]	Success rate[c]	Classifier[a]	Dimension[b]	Success rate[c]
CD (Λ_1)	20	61.6%	CD (Λ_{12})	42	67.9%
CD (Λ_2)	22	63.8%	CD (Λ_{13})	44	68.0%
CD (Λ_3)	24	65.5%	CD (Λ_{14})	46	67.9%
CD (Λ_4)	26	66.1%	CD (Λ_{15})	48	68.1%
CD (Λ_5)	28	67.4%	CD (Λ_{16})	50	67.9%
CD (Λ_6)	30	67.9%	CD (Λ_{17})	52	67.7%
CD (Λ_7)	32	68.0%	CD (Λ_{18})	54	67.7%
CD (Λ_8)	34	67.8%	CD (Λ_{19})	56	67.3%
CD (Λ_9)	36	67.9%	CD (Λ_{20})	58	66.8%
CD (Λ_{10})	38	67.9%	CD (Λ_{21})	60	66.5%
CD (Λ_{11})	40	68.0%	CD (Λ_{22})	62	66.7%

[a]Individual basic classifier CD (Λ_i) ($i = 1, 2, \ldots, \Omega$) (see Eq. 13).
[b]The dimension of the amphiphilic pseudo amino acid composition considered here was given by $\Lambda_i = [20 + 2(i - 1)]$ ($i = 1, 2, \ldots, 22$), on which each of the individual classifier CD (Λ_i) was operated.
[c]The overall jackknife success rate was derived from the training dataset of 3,799 proteins taken from [Chou and Cai, 2003c].

with what was expected because the jackknife test is the most stringent examination method among these three. A similar trend can also be seen for the results by the other classifiers. (3) The overall jackknife success rate (70.2%) by the ensemble classifier (Table I) is higher than any of the 22 jackknife success rates obtained by each of the 22 individual classifiers (Table II), clearly indicating that the ensemble classifier formed by fusing many basic classifiers is more powerful than each of their individuals.

The goal of this study is not to determine the possible upper limit of the success rate in predicting protein subcellular location, but to propose a novel approach by fusing many individual classifiers each based on different dimensions of pseudo amino acid compositions that might help to open a new avenue to further increase our ability to deal with this very complicated and difficult problem. It should be realized that it is too premature to construct a complete or quasi-complete training dataset based on the knowledge available so far. Without a complete or quasi-complete training dataset, any attempt to determine such an upper limit would be unjustified, and the result thus obtained might be misleading no matter how powerful the classifier is.

It should be pointed out that some proteins may occur in several different subcellular locations, that is, bear the feature of "multiplex locations." Also, some proteins are known to be shuttled from one subcellular compartment to another, and back again. For this kind of multiplex locations case or dynamic case, a different approach to train the predictor and count the success rates is needed as elaborated in [Cai and Chou, 2004; Chou and Cai, 2005a].

CONCLUSION

Classifiers that are established on the pseudo amino acid composition can incorporate a considerable amount of sequence order effects of proteins, and hence perform much better in predicting their subcellular locations than those based on the conventional amino acid composition. However, there are many choices in selecting the dimension of the pseudo amino acid composition, and each different choice may lead to a different outcome. It is both time-consuming and tedious to determine the optimal one. The current ensemble classifier formed by fusing many such single classifiers can automatically solve the problem, leading to much higher success prediction rates. The fusion ensemble classifier may also be used to predict many other attributes of proteins according to their sequences, and become a powerful tool in proteomics and bioinformatics.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their constructive suggestions, which were very helpful for strengthening the presentation of this article.

Appendix A

Amphiphilic Pseudo Amino Acid Composition

For reader's convenience, here let us give a brief introduction of the amphiphilic pseudo amino acid composition. For more information about various modes of pseudo amino acid composition and their applications, refer to [Chou, 2001a, 2005a,b; Cai and Chou, 2003, 2005, 2006; Pan et al., 2003; Chou and Cai, 2005, 2006, 2004, 2005b, 2006a,b; Wang et al., 2004, 2005; Cai et al., 2005; Gao et al., 2005; Liu et al., 2005a; Shen and Chou, 2005a,b; Shen et al., 2005a; Xiao et al., 2005b,c, 2006a,b].

For a protein with a sequence generally formulated by Equation 1 of the text, its amino acid composition is given by [Chou and Zhang, 1994; Chou, 1995]

$$\mathbf{P}_{\mathrm{AA}} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{20} \end{bmatrix} \tag{A1}$$

where f_1 is the occurrence frequency of amino acid A in the protein, f_2 that of amino acid C, and so forth. Here, without loss of generality, the single codes of the 20 native amino acids are used according to their alphabetical order. Also, the 20 occurrence frequencies are normalized to 100. As we can see from Equation A1, if a protein is represented by such a set of discrete numbers, all its sequence information would be lost. To keep its representation with a discrete mode but without completely losing its sequence-order information, we can define a pseudo amino acid composition by merging a series of sequence-order-correlated factors into the conventional amino acid composition. As is well known, the

hydrophobicity and hydrophilicity play a very important role to the folding of a protein as well as its microenvironment and interior packing (see, e.g., [Chou et al., 1984, 1986, 1990]). For instance, many helices in proteins are amphiphilic that is formed by the hydrophobic and hydrophilic amino acids according to a special order along the helix chain, as illustrated by the "wenxiang" diagram [Chou et al., 1997]. Therefore, these two indices may be one of the optimal choices to reflect the sequence order effects. In view of this, the sequence-order effects can be indirectly and partially, but quite effectively, reflected through the following equations (see Fig. A1):

$$\begin{cases} \tau_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^1 \\ \tau_2 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1}^2 \\ \tau_3 = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2}^1 \\ \tau_4 = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2}^2, (\lambda < L) \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \tau_{2\lambda-1} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^1 \\ \tau_{2\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^2 \end{cases} \quad \text{(A2)}$$

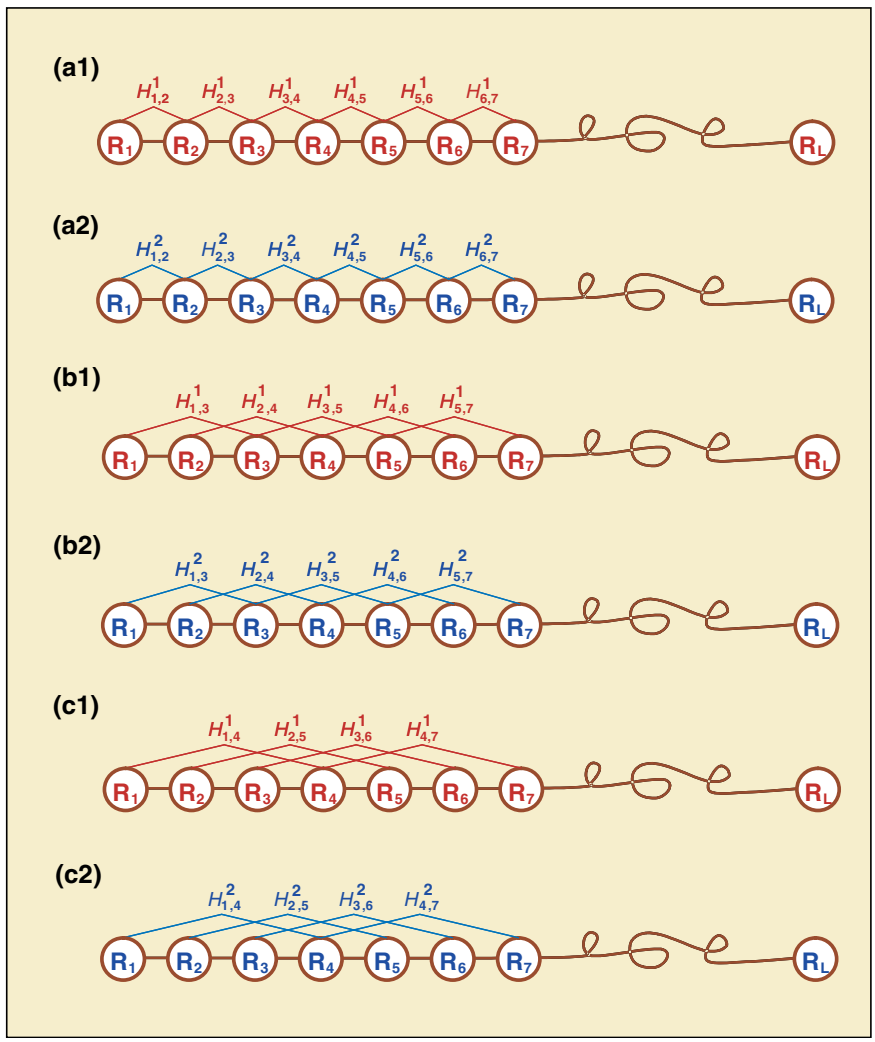

Fig. A1. A schematic drawing to show the amphiphilic correlation along a protein chain, where the values of $H_{i,j}^1$ and $H_{i,j}^2$ are given by Equations A3 and A4 and Table AI. The correlation via hydrophobicity is shown in red, while the correlation via hydrophilicity in blue. **Panel a1/a2** reflects the coupling mode between all the most contiguous residues, **panel b1/b2** that between all the second most contiguous residues, and **panel c1/c2** that between all the third most contiguous residues. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

where $H_{i,j}^1$ and $H_{i,j}^2$ are the hydrophobicity and hydrophilicity correlation functions given by

$$\begin{cases} H_{i,j}^1 = 100w[h^1(\mathrm{R}_i)h^1(\mathrm{R}_j)] \\ H_{i,j}^2 = 100w[h^2(\mathrm{R}_i)h^2(\mathrm{R}_j)] \end{cases} \quad \text{(A3)}$$

where $h^1(R_i)$ and $h^2(R_i)$ are, respectively, the hydrophobicity and hydrophilicity values for the ith(ith $= 1,2,\ldots,L$) amino acid in Equation 1, and w is the weight factor. In the current study, we chose $w = 0.5$ to make the data within the range easier to be handled (w can be of course assigned with other values, but this would not have a big impact to the final results). In Equation A2 τ_1 and τ_2 are called the 1st-tier correlation factors that reflect the sequence-order correlation between all the most contiguous residues along a protein chain through hydrophobicity and hydrophilicity, respectively [Fig. A1 (a1,a2)], τ_3 and τ_4 are the corresponding second-tier correlation factors that reflect the sequence-order correlation between all the second most contiguous residues [Fig. A1 (b1,b2)], and so forth. Note that before substituting the values of hydrophobicity into Equation A3, they were all subjected to a *standard conversion* as described by the following equation:

$$\begin{cases} h_1(\mathrm{R}_i) = \dfrac{h_1^0(\mathrm{R}_i) - \langle h_1^0 \rangle}{\mathrm{SD}(h_1^0)} \\ h_2(\mathrm{R}_i) = \dfrac{h_2^0(\mathrm{R}_i) - \langle h_2^0 \rangle}{\mathrm{SD}(h_2^0)} \end{cases} \quad \text{(A4)}$$

where the symbols $h_1^0(\mathrm{R}_i)$ and $h_2^0(\mathrm{R}_i)$ represent the original hydrophobicity value [Tanford, 1962] and hydrophilicity value [Hopp and Woods, 1981] for amino acid R_i, respectively (Table AI); $\langle h_1^0 \rangle$ and $\langle h_2^0 \rangle$ their means over 20 native amino acids; $\mathrm{SD}(h_1^0)$ and $\mathrm{SD}(h_2^0)$ their standard deviations. The converted hydrophobicity and hydrophilicity values obtained by Equation A4 will have a zero mean value over the 20 native amino acids, and will remain unchanged if going through the same conversion procedure again. After merging the sequence-order-correlated factors from Equation A2 into the classical 20D (dimensional) amino acid composition (Eq. A1), we obtain a pseudo amino acid composition with $20 + 2\lambda$ components. In other words, the representation for the protein sequence of Equation 1 is now formulated as

$$\mathbf{P}_{\mathrm{PseAA}} = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{20} \\ \tau_1 \\ \tau_2 \\ \vdots \\ \tau_{2\lambda} \end{bmatrix}, \quad \text{(A5)}$$

which can be easily converted to Equation 2 by performing a normalization procedure according to the following equation:

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{2\lambda} \tau_j}, & (1 \le u \le 20) \\ \dfrac{\tau_u}{\sum_{i=1}^{20} f_i + \sum_{j=1}^{2\lambda} \tau_j}, & (20+1 \le u \le 20+2\lambda) \\ 20 + 2\lambda = \Lambda. \end{cases} \quad \text{(A6)}$$

For reader's convenience, the $\Lambda = 20 + 21 \times 2 = 62$ components of $\mathbf{P}_{\mathrm{PseAA}}$ in Equation A5 for each of the proteins in the training and testing datasets studied here are given in the Online Supplementary Materials A and B, respectively, from which the user can easily generate the normalized pseudo amino acid composition $\mathbf{P}$ of Equation 2 with any dimension of $\Lambda \le 62$ through Equation A6. For

TABLE AI. The Amino Acid Parameters Used for Deriving the Amphiphilic Pseudo Amino Acid Components (cf. Eq. A4)

Code	Hydrophobicity[a] $h_1{}^0$	Hydrophilicity[b] $h_2{}^0$
A	0.62	−0.5
C	0.29	−1.0
D	−0.90	3.0
E	−0.74	3.0
F	1.19	−2.5
G	0.48	0.0
H	−0.40	−0.5
I	1.38	−1.8
K	−1.50	3.0
L	1.06	−1.8
M	0.64	−1.3
N	−0.78	2.0
P	0.12	0.0
Q	−0.85	0.2
R	−2.53	3.0
S	−0.18	0.3
T	−0.05	−0.4
V	1.08	−1.5
W	0.81	−3.4
Y	0.26	−2.3

[a]The hydrophobicity values were taken from [Tanford, 1962].
[b]The hydrophilicity values were taken from [Hopp and Woods, 1981].

instance, to generate the $\mathbf{P}$ with $\Lambda = 20$, just read the first 20 data for each protein in the Online Supplementary Materials followed by substituting them into Equation A6; to generate the $\mathbf{P}$ with $\Lambda = 22$, just read the first 22 data followed by the same procedure; and so forth. Actually, suppose the length of the shortest protein sequence studied here is $L_{\min}$, by following the above procedures one can always generate the $\mathbf{P}$ with a dimension of $\Lambda \leq [20 + 2(L_{\min} - 1)]$(see Eq. A2 and Fig. A1). It should be pointed out that, according to the definition of the classical amino acid composition, all its components must be ≥ 0; it is not always true, however, for the pseudo amino acid composition: the components derived from the sequence correlation factors (cf. Eq. A2) may also be < 0. But this will not affect the existence of Equations 7 and 8 as proved in Appendix A of [Chou, 2005b].

REFERENCES

Bairoch A, Apweiler R. 2000. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res 25:31–36.

Cai YD. 2001. Is it a paradox or misinterpretation. Proteins: Struct Func Genet 43:336–338.

Cai YD, Chou KC. 2003. Nearest neighbour algorithm for predicting protein subcellular location by combining functional domain composition and pseudo-amino acid composition. Biochem Biophys Res Comm 305:407–411.

Cai YD, Chou KC. 2004. Predicting 22 protein localizations in budding yeast. Biochem Biophys Res Comm 323:425–428.

Cai YD, Chou KC. 2005. Predicting enzyme subclass by functional domain composition and pseudo amino acid composition. J Proteome Res 4:967–971.

Cai YD, Chou KC. 2006. Predicting membrane protein type by functional domain composition and pseudo amino acid composition. J Theor Biol 238:395–400.

Cai YD, Zhou GP, Chou KC. 2005. Predicting enzyme family classes by hybridizing gene product composition and pseudo-amino acid composition. J Theor Biol 234:145–149.

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

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

Chou KC. 2001a. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Struct Funct Genet (Erratum: ibid., 2001, Vol.44, 60) 43:246–255.

Chou KC. 2001b. Using subsite coupling to predict signal peptides. Protein Eng 14:75–79.

Chou KC. 2004. Review: Structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134.

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

Chou KC. 2005b. 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. 2003a. 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. 2003b. Predicting protein quaternary structure by pseudo amino acid composition. Proteins: Struct Funct Genet 53:282–289.

Chou KC, Cai YD. 2003c. Prediction and classification of protein subcellular location: Sequence-order effect and pseudo amino acid composition. J Cell Biochem (Addendum, ibid. 2004, 91, 1085) 90:1250–1260.

Chou KC, Cai YD. 2004. Predicting enzyme family class in a hybridization space. Protein Sci 13:2857–2863.

Chou KC, Cai YD. 2005a. Predicting protein localization in budding yeast. Bioinformatics 21:944–950.

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

Chou KC, Cai YD. 2006a. Prediction of protease types in a hybridization space. Biochem Biophys Res Comm 339:1015–1020.

Chou KC, Cai YD. 2006b. Predicting protein–protein interactions from sequences in a hybridization space. J Proteome Res 5:316–322.

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

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

Chou KC, Maggiora GM. 1998. Domain structural class prediction. Protein Eng 11:523–538.

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. Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30:275–349.

Chou KC, Nemethy G, Scheraga HA. 1984. Energetic approach to packing of a-helices: 2. General treatment of nonequivalent and nonregular helices. J Am ChemSoc 106:3161–3170.

Chou KC, Nemethy G, Rumsey S, Tuttle RW, Scheraga HA. 1986. Interactions between two beta-sheets: Energetics of beta/beta packing in proteins. J Mol Biol 188:641–649.

Chou KC, Nemethy G, Scheraga HA. 1990. Review: Energetics of interactions of regular structural elements in proteins. Acc Chem Res 23:134–141.

Chou KC, Zhang CT, Maggiora GM. 1997. Disposition of amphiphilic helices in heteropolar environments. Proteins: Struct Funct Genet 28:99–108.

Chou KC, Liu W, Maggiora GM, Zhang CT. 1998. Prediction and classification of domain structural classes. Proteins: Struct Funct Genet 31:97–103.

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.
Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC. 2005. Using pseudo amino acid composition to predict protein subcellular location: Approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28:373–376.
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.
Hopp TP, Woods KR. 1981. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78:3824–3828.
Hua S, Sun Z. 2001. Support vector machine approach for protein subcellular localization prediction. Bioinformatics 17:721–728.
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, Ling JG, Chou KC. 2005b. Prediction of protein signal sequences and their cleavage sites by statistical rulers. Biochem Biophys Res Comm 338:1005–1011.
Luo RY, Feng ZP, Liu JK. 2002. Prediction of protein strctural class by amino acid and polypeptide composition. Eur J Biochem 269:4219–4225.
Mahalanobis PC. 1936. On the generalized distance in statistics. Proc Natl Inst Sci India 2:49–55.
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.
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.
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 acid and amino acid pairs. Bioinformatics 19:1656–1663.
Pillai KCS. 1985. Mahalanobis D2. In: Kotz S, Johnson NL, editors. "Encyclopedia of Statistical Sciences." New York: John Wiley & Sons. This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics, pp 176–181.
Radford T. 2003. Metaphors and dreams. The Scientist 17:24–26.
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 Comm 337:752–756.
Shen HB, Yang J, Chou KC. 2005a. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. J Theor Biol: 10.1016/j.jtb.2005.08.016 (online ahead of print).
Shen HP, Yang J, Liu XJ, Chou KC. 2005b. Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334:577–581.
Tanford C. 1962. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J Am Chem Soc 84:4240–4274.
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 M, Yang J, Xu ZJ, Chou KC. 2005. SLLE for predicting membrane protein types. J Theor Biol 232: 7–15.
Xiao X, Shao S, Ding Y, Huang Z, Chen X, Chou KC. 2005a. 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 S, Ding Y, Huang Z, Chen X, Chou KC. 2005b. Using cellular automata to generate Image representation for biological sequences. Amino Acids 28:29–35.
Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC. 2005c. Using complexity measure factor to predict protein subcellular location. Amino Acids 28:57–61.
Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC. 2006a. Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular 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. Journal of Computational Chemistry 27:478–482.
Yuan Z. 1999. Prediction of protein subcellular locations using Markov chain models. FEBS Letters 451:23–26.
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: Struct Funct Genet 44:57–59.
Zhou GP, Doctor K. 2003. Subcellular location prediction of apoptosis proteins. Proteins: Struct Funct Genet 50:44–48.