Journal of Cellular Biochemistry 90:1250–1260 (2003)

Prediction and Classification of Protein Subcellular Location—Sequence-Order Effect and Pseudo Amino Acid Composition

Kuo-Chen Chou[1,2]* and Yu-Dong Cai[3]*

[1]Gordon Life Science Institute, San Diego, CA 92130
[2]Tianjin Institute of Bioinformatics and Drug Discovery, Tianjin, China
[3]Biomolecular Sciences Department, UMIST, Manchester, M60 1QD, United Kingdom

Abstract Given a protein sequence, how to identify its subcellular location? With the rapid increase in newly found protein sequences entering into databanks, the problem has become more and more important because the function of a protein is closely correlated with its localization. To practically deal with the challenge, a dataset has been established that allows the identification performed among the following 14 subcellular locations: (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. Compared with the datasets constructed by the previous investigators, the current one represents the largest in the scope of localizations covered, and hence many proteins which were totally out of picture in the previous treatments, can now be investigated. Meanwhile, to enhance the potential and flexibility in taking into account the sequence-order effect, the series-mode pseudo-amino-acid-composition has been introduced as a representation for a protein. High success rates are obtained by the re-substitution test, jackknife test, and independent dataset test, respectively. It is anticipated that the current automated method can be developed to a high throughput tool for practical usage in both basic research and pharmaceutical industry. J. Cell. Biochem. 90: 1250–1260, 2003. © 2003 Wiley-Liss, Inc.

Key words: augmented covariant discriminant algorithm; organelles; subcellular compartments; bioinformatics; high throughput tool; proteomics

A cell contains approximately 10^9 protein molecules. The human body consists of hundreds of cell types, all originating from the fertilized egg. During the embryonic and foetal periods, the number of cells increases dramatically. The cells mature and become specialized to form the various tissues and organs of the body. The human body hosts 10^{14} cells [Radford, 2003], or approximately 10^{23} protein molecules. It is quite interesting to see that the latter has the same order of magnitude as the Avogadro constant, isn't it? A cell consists of many different compartments, or organelles, each surrounded by a membrane. The organelles are specialized to carry out different tasks. For example, the mitochondrion functions as the "power plant," producing energy needed by the cell. The cell nucleus contains the genetic material (DNA), governing all functions of the cell. And the endoplasmic reticulum is, together with the ribosomes, responsible for synthesizing proteins.

According to the localization or compartment in a cell, proteins are generally classified into the following 14 categories: (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 (see Fig. 1 for a schematic illustration). Note that the cell wall, chloroplast, and vacuole proteins exist only in a plant cell, while the centriole proteins only in an animal cell. Given the sequence of a protein, how can we predict which subcellular location it belongs to? This is certainly a very

*Correspondence to: Yu-Dong Cai, Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester, M60 1QD, UK; or Kuo-Chen Chou.
E-mail: y.cai@umist.ac.uk or kchou@san.rr.com

Received 23 July 2003; Accepted 29 August 2003

DOI 10.1002/jcb.10719

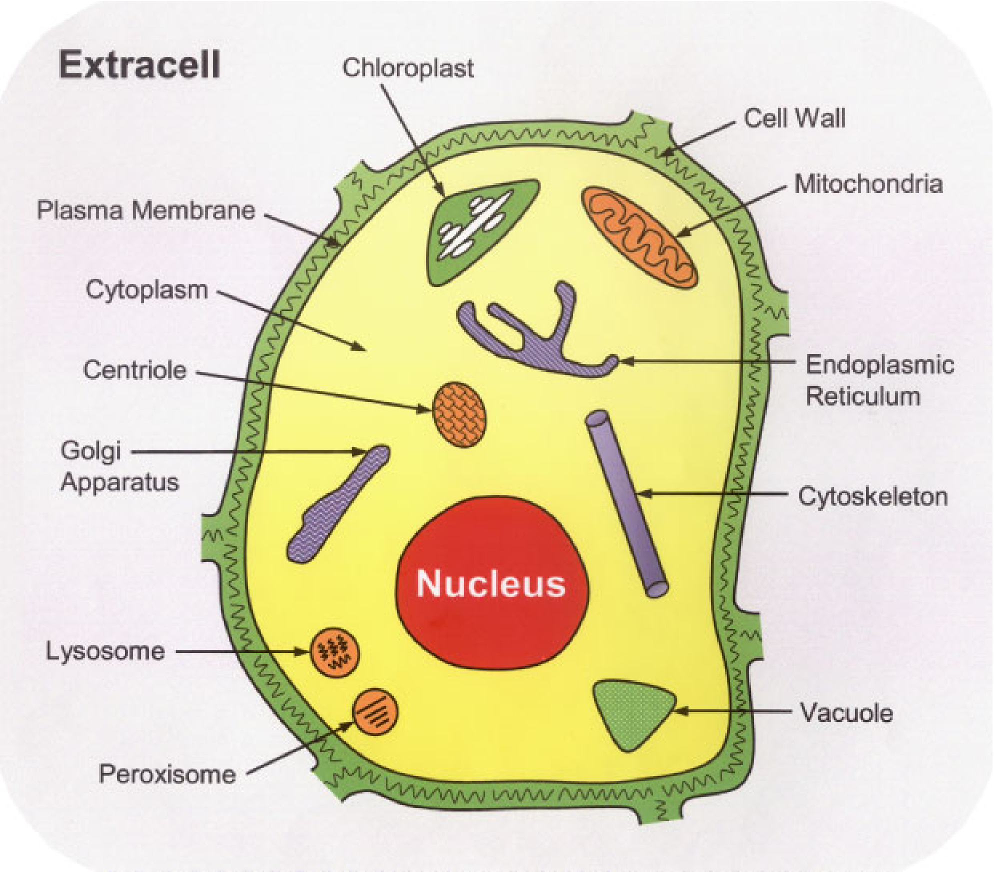

Fig. 1. Schematic illustration to show the 14 subcellular locations of proteins: (1) cell wall, (2) centriole, (3) chloroplast, (4) cytoplasm, (5) cytoskeleton, (6) endoplasmic reticulum, (7) extracellular, (8) Golgi apparatus, (9) lysosome, (10) mitochondria, (11) nucleus, (12) peroxisome, (13) plasma 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. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

important problem since the localization of a protein is closely correlated with its biological function. Although the information about protein subcellular location can be determined by conducting various experiments, that is both time-consuming and costly. Owing to the fact that the number of sequences entering into databanks has been rapidly increasing, for example, in 1986 the total sequence entries in SWISS-PROT [Bairoch and Apweiler, 2000] was only 3,939 while the number was increased to 59,021 in 1996 and 101,602 in 2000, the problem has become an urgent challenge to scientists. Particularly, it is anticipated that many more new protein sequences will be derived soon because of the recent success of the human genome project, which has provided an enormous amount of genomic information in the form of 3 billion base pairs, assembled into tens of thousands of genes. Therefore, the challenge will become even more urgent and critical. Actually, many efforts have been made trying to develop some computational methods for quickly predicting the subcellular locations of proteins [Nakai and Kanehisa, 1992; Nakashima and Nishikawa, 1994; Cedano et al., 1997; Claros et al., 1997; Reinhardt and Hubbard, 1998; Chou and Elrod, 1999b; Chou, 2001; Chou and Cai, 2002; Pan et al., 2003; Zhou and Doctor, 2003]. Of these methods, some [Claros et al., 1997; Nakai and Kanehisa, 1992] are based on the N-terminal sorting signals. Their merit is having a clear biological implication [Chou, 2002b]. However, as pointed out by Reinhardt and Hubbard [1998], “in large genome analysis projects genes are usually automatically assigned and these assignments are often unreliable for the 5′-regions.” “This can lead to leader sequences being missing or only partially included, thereby causing problems for prediction algorithms depending on them." Therefore, most of the existing algorithms are based on the information derived from entire protein sequences rather than their signal peptides only. However, because of the difficulty due to the extreme variance in sequence order and length, the majority of these algorithms are based on the amino-acid-composition alone. According to the classical definition, the amino acid composition consists of 20 components each representing the occurrence frequency of one of the 20 native amino acids in a given protein, and hence a protein is represented by a 20-D (dimensional) vector. Obviously, if using the classical amino acid composition alone to represent a protein, all the sequence-order and sequence-length effects would be missed out and the prediction method underlain with such a basis must bear a considerable intrinsic limitation. It is in only a few recent papers [Chou, 2000a, 2001; Cai et al., 2002; Pan et al., 2003] that some partial sequence-order effects were incorporated through a novel concept, the so-called pseudo-amino acid composition [Chou, 2001]. Meanwhile, a completely different approach, the so-called functional domain composition [Chou and Cai, 2002] was proposed that incorporated the functional type information. The introduction of the functional domain composition represents an important progress in directly relating the localization of proteins with their function. However, owing to the fact that the current functional domain database [Murvai et al., 2001] is far from complete yet, some proteins cannot be properly defined in terms of the functional domain composition, leading to some setback in practical application. Also, none of the aforementioned methods includes the training data for predicting proteins located in the cell wall and centriole. In view of this, we shall first construct a new training dataset that includes the cell wall and centriole proteins as well, followed by defining a new pseudo amino acid composition to take into account the sequence-order effects for predicting the attributes of proteins among their 14 localizations (Fig. 1).

MATERIALS AND METHODS

Construction of Working Datasets

Two working datasets, i.e., a training dataset and an independent testing dataset, were constructed based on release 40.0 of SWISS-PROT database [Bairoch and Apweiler, 2000] published on 08-Mar-2002 by following the same screening procedures and criteria as described by Chou and Elrod [1999b] to avoid inclusion of any localization-ambiguous or redundant sequences.

The training dataset thus obtained consists of 14 subsets and 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. The code for each of the 3,799 proteins in the training dataset is given in the Online Supplemental Material A, where the accession number rather than the SWISS-PROT name is used because the accession number is more stable for representing a unique protein sequence.

The independent testing dataset consists of 14 subsets and 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) 752 of mitochondria, (11) 914 of nucleus, (12) 84 of peroxisome, (13) 24 of plasma membrane, and (14) 17 of vacuole. The code for each of the 4,498 proteins in the training dataset is given in the Online Supplemental Material B.

It is instructive to conduct a sequence identity analysis for the proteins studied here. The sequence identity between two protein sequences is defined as follows. Suppose the maximum number of residues matched by sliding one sequence along the other is M, and the alignment length is L, the sequence identity between the two sequences is defined as M/L. The treatment for gaps is according to CLUSTALW [Thompson et al., 1994]. The average sequence identity obtained by the sequence match operation for each of the 14 subsets in the training dataset is given in the Online Supplemental Material A, and that in the testing dataset given in the Online Supplemental Material B. Furthermore, the similar sequence match operation was also performed for the dataset by combining the training and testing datasets. It was found that the average sequence identities for the 14 subsets in the combined dataset are consecutively 0.1020, 0.1946, 0.0691, 0.0655, 0.1024, 0.1005, 0.0596, 0.0769, 0.0806, 0.0698, 0.0652, 0.0856, 0.1222, and 0.1252. From these data, we can see that most sequences in a same subset have quite low sequence identity not only for the training and testing datasets, but also for their combination, a clear indication of exclusion of redundant and homologous sequences.

Pseudo Amino Acid Composition

In order to improve the quality of statistical prediction for protein subcellular location, one of the most important steps is how to give an effective representation for a protein. According to common sense, an effective representation should include as much information a protein has as possible. Obviously, the entire protein sequence contains of course the most complete information. However, if using the entire sequence of a protein as its representation to formulate a statistical prediction algorithm, one would face the difficulty to deal with almost an infinity of sample patterns, as elaborated by Chou [2001]. To formulate a feasible statistical prediction algorithm, a protein must be expressed in terms of a set of discrete numbers. The earlier approach in this regard was to represent proteins according their amino acid composition, which has substantially reduced the number of samples and made the statistical treatment become tractable. Such an approach was widely used to predict protein structural class [Nakashima et al., 1986; Chou, 1989, 1995; Chou and Zhang, 1994; Bahar et al., 1997; Chou et al., 1998; Zhou, 1998], predict protein secondary structure content [Krigbaum and Knutton, 1973; Muskal and Kim, 1992; Zhang et al., 1996; Chou, 1999b], to predict protein subcellular location [Nakashima and Nishikawa, 1994; Cedano et al., 1997; Reinhardt and Hubbard, 1998; Chou and Elrod, 1999b; Chou, 2000b], to predict GPCR types [Chou and Elrod, 2002; Elrod and Chou, 2002], and to predict enzyme family classes [Chou and Elrod, 2003]. However, as mentioned above, if using the 20-D 1classical amino acid composition to represent a protein, all the sequence-order and sequence-length effects would be missed out and the prediction method underlain with such a basis must bear a considerable intrinsic limitation. Here we are actually confronted with such a dilemma that, if wishing to include the complete information of an entire protein chain, the prediction would become impracticable; if wishing to make the prediction feasible, some of its information must be dropped. In view of this, can we find a compromise scenario where a protein is still represented by a set of discrete numbers which, however, are also able to contain as much of the sequence-order effects as possible? The introduction of the pseudo amino acid composition [Chou, 2001] is a promising effort in this regard that has made a remarkable improvement in predicting protein subcellular location. Unlike the classical amino acid composition that consists of only 20 components, the pseudo amino acid composition consists of $20 + \lambda$ discrete numbers, where the first 20 are the same as the 20 components in the amino acid composition and the remainders represent λ different ranks of sequence-order correlation factors [Chou, 2002a]. Here we would like to propose a different approach to define the pseudo amino acid composition as formulated below.

Suppose a protein X with a sequence of L amino acid residues:

$$\mathrm{R}_1\ \mathrm{R}_2\ \mathrm{R}_3\ \mathrm{R}_4\ \mathrm{R}_5\ \mathrm{R}_6\ \mathrm{R}_7 \cdots\cdots \mathrm{R}_\mathrm{L} \qquad (1)$$

where R_1 represents the residue at sequence position 1, R_2 the residue at position 2, and so forth. The pseudo amino acid composition for the protein is generated by merging two sets of sequence-order-correlated factors into the conventional amino acid composition. The first set is called delta function set, and its λ sequence-order-correlated factors are given by (Fig. 2):

$$\begin{cases} \delta_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} \Delta_{i,i+1} \\ \delta_2 = \frac{1}{L-2}\sum_{i=1}^{L-2} \Delta_{i,i+2} \\ \delta_3 = \frac{1}{L-3}\sum_{i=1}^{L-3} \Delta_{i,i+3} \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ \delta_\lambda = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} \Delta_{i,i+\lambda} \end{cases} \quad (\lambda < L) \qquad (2)$$

where $\Delta_{i,j}$ is a delta function defined by

$$\Delta_{i,j} = \Delta(\mathrm{R}_i, \mathrm{R}_j) = \begin{cases} 1, & \text{if } \mathrm{R}_i = \mathrm{R}_j \\ 0, & \text{otherwise} \end{cases} \qquad (3)$$

The second set is called hydrophobicity set, and its μ sequence-order-correlated factors are given by (Fig. 3):

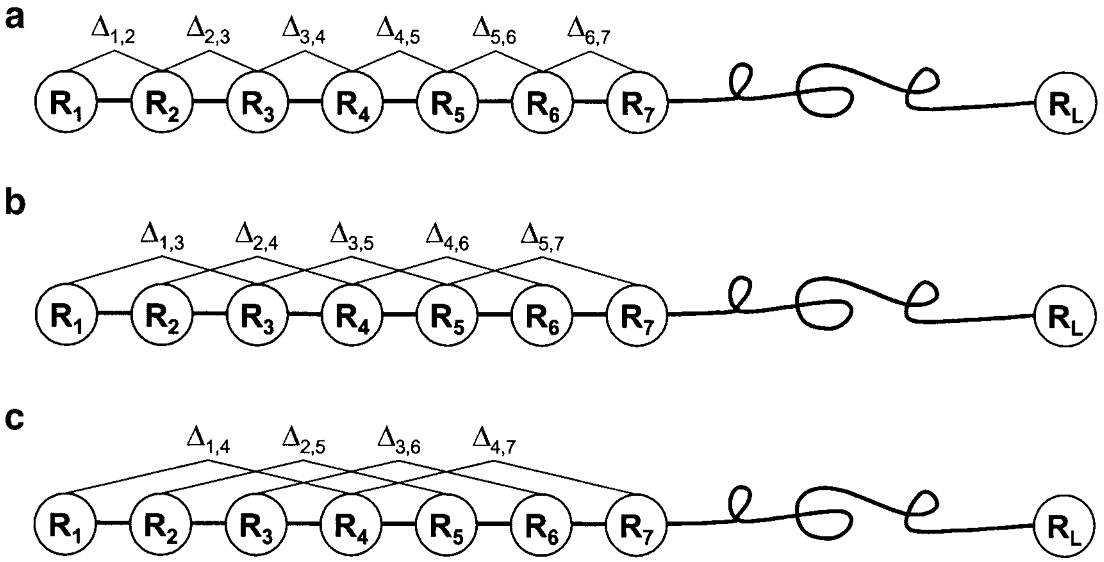

Fig. 2. A schematic drawing to show (**a**) the 1st-rank, (**b**) the 2nd-rank, and (**c**) the 3rd-rank sequence-order-coupling mode along a protein sequence through a delta-function, where $\Delta_{i,j}$ is given by Equation (3). Panel (a) reflects the coupling mode between all the most contiguous residues, panel (b) that between all the 2nd most contiguous residues, and panel (c) that between all the 3rd most contiguous residues.

$$\begin{cases} h_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} H_{i,i+1} \\ h_2 = \frac{1}{L-2}\sum_{i=1}^{L-2} H_{i,i+2} \\ h_3 = \frac{1}{L-3}\sum_{i=1}^{L-3} H_{i,i+3}, \qquad (\mu < L) \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ h_\mu = \frac{1}{L-\mu}\sum_{i=1}^{L-\mu} H_{i,i+\mu} \end{cases} \tag{4}$$

where $H_{i,j}$ is a hydrophobicity correlation function given by

$$H_{i,j} = H(\mathrm{R}_i) \cdot H(\mathrm{R}_j) \tag{5}$$

where $H(\mathrm{R}_i)$ and $H(\mathrm{R}_j)$ are the hydrophobicity values for R_i and R_j, respectively, and the dot $(\cdot)$ means the multiplication sign. Note that before substituting the values of hydophobicity into Equation (5), they were all subjected to a *Standard Conversion* as described by the following equation:

$$H(i) = \frac{H^0(i) - \sum_{i=1}^{20}\frac{H^0(i)}{20}}{\sqrt{\frac{\sum_{i=1}^{20}\left[H^0(i)-\sum_{i=1}^{20}\frac{H^0(i)}{20}\right]^2}{20}}} \tag{6}$$

where we use the numerical indices 1, 2, 3, ..., 20 to represent the 20 native amino acids according to the alphabetical order of their single-letter codes: A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y. And $H_1^0(i)$ is the original hydrophobicity value of the ith amino acid that was taken from Tanford [1962]. The 20 converted hydrophobicity values obtained by Equation 6 will have a zero mean value, and will remain unchanged if going through the same conversion procedure again.

After merging the sequence-order-correlated factors from Equations (2) and (4) into the classical 20-D amino acid composition, we obtain a pseudo amino acid composition with $20 + \lambda + \mu$ components. In other words, the representation for protein **X** is now formulated as

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_{20} \\ x_{20+1} \\ \vdots \\ x_{20+\lambda} \\ x_{20+\lambda+1} \\ \vdots \\ x_{20+\lambda+\mu} \end{bmatrix}, \tag{7}$$

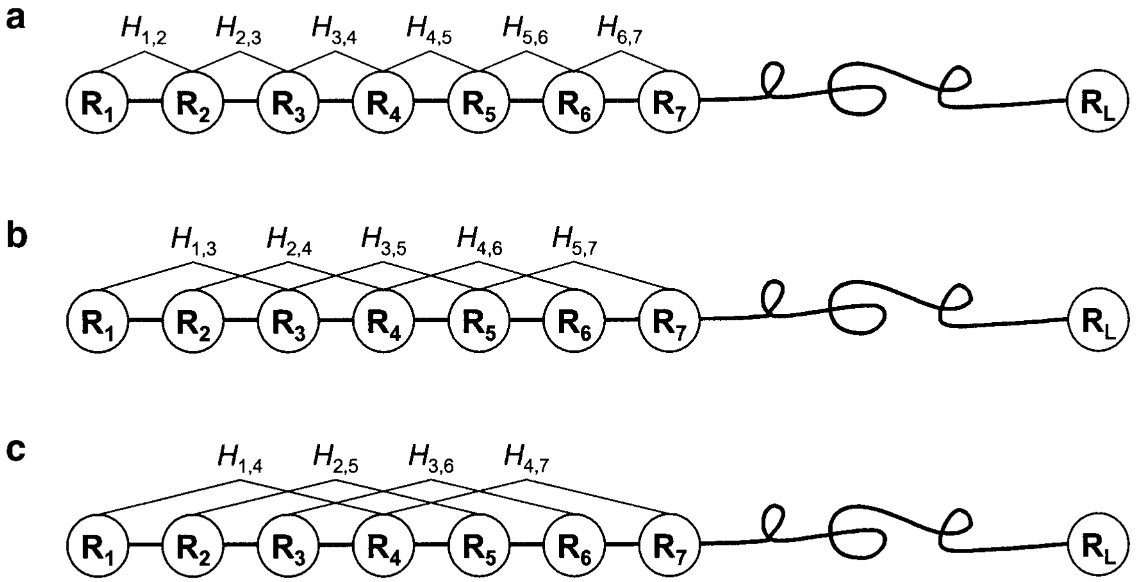
a
$H_{1,2}$ $H_{2,3}$ $H_{3,4}$ $H_{4,5}$ $H_{5,6}$ $H_{6,7}$
R_1 R_2 R_3 R_4 R_5 R_6 R_7 R_L
b
$H_{1,3}$ $H_{2,4}$ $H_{3,5}$ $H_{4,6}$ $H_{5,7}$
R_1 R_2 R_3 R_4 R_5 R_6 R_7 R_L
c
$H_{1,4}$ $H_{2,5}$ $H_{3,6}$ $H_{4,7}$
R_1 R_2 R_3 R_4 R_5 R_6 R_7 R_L

Fig. 3. A schematic drawing to show (**a**) the 1st-rank, (**b**) the 2nd-rank, and (**c**) the 3rd-rank sequence-order-coupling mode along a protein sequence through a hydrophobicity correlation function, where $H_{i,j}$ is given by Equation (5). See legend to Figure 2 for further explanation.

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w_1 \sum_{j=1}^{\lambda} \delta_j + w_2 \sum_{k=1}^{\mu} h_k}, & (1 \le u \le 20) \\ \dfrac{w_1 \delta_{u-20}}{\sum_{i=1}^{20} f_i + w_1 \sum_{j=1}^{\lambda} \delta_j + w_2 \sum_{k=1}^{\mu} h_k}, & (20 + 1 \le u \le 20 + \lambda) \\ \dfrac{w_2 h_{u-20-\lambda}}{\sum_{i=1}^{20} f_i + w_1 \sum_{j=1}^{\lambda} \delta_j + w_2 \sum_{k=1}^{\mu} h_k}, & (20 + \lambda + 1 \le u \le 20 + \lambda + \mu) \end{cases} \tag{8}$$

where f_i is the normalized occurrence frequency of the 20 amino acids in the protein **X**, δ_j the j-tier sequence-correlation factor computed according to Equation (2), h_k the k-tier sequence-correlation factor computed according to Equation (4), and w_1 and w_2 are the weight factors. In the current study, we chose $w_1 = w_2 = 0.5$ to make the data within the range easier to be handled (w_1 and w_2 can be of course assigned with other values, but this would not have a big impact to the final results). As we can see from Equations (7–8), the first 20 components reflect the effect of the classical amino-acid-composition, while the components from $20 + 1$ to $20 + \lambda + \mu$ reflect the effect of sequence order. A set of such $20 + \lambda + \mu$ components is called the series-mode pseudo-amino-acid-composition. It has the following three advantages: (a) It contains more sequence-order effects not only than the conventional 20-D amino-acid-composition [Nakashima et al., 1986], but also than the 210-D pair-coupled amino-acid-composition [Chou, 1999b] and the 400-D 1st-order coupled amino-acid-composition [Liu and Chou, 1999], as reflected by a series of sequence-correlation factors with different tiers and modes of correlation (see Figs. 2 and 3 and Eqs. (2) and (4)). (b) Compared with the previous approach [Chou, 2001] where the pseudo amino acid components were generated by combining various biochemical quantities through a parallel mode, the series mode is adopted in the current approach and hence the situation of mutual cancellation among themselves in counting the sequence-order effects would be less likely to happen. (c) The introduction of delta-function (Eq. 3) has not only made the computation quite simple but also very effective in counting the sequence-order effects.

Prediction Algorithms

As we can see from Equations (7), the pseudo-amino-acid-composition has the same formulation as the conventional one except containing more components. Therefore, all the existing prediction algorithms based on the conventional amino-acid-composition, such as the least Hamming distance algorithm [Chou, 1989], the ProtLock algorithm [Cedano et al., 1997], and the covariant discriminant algorithm [Chou and Elrod, 1999b] can be applied on the pseudo-amino-acid-composition by a straightforward augmentation procedure as illustrated in Chou [2001], and hence there is no need to repeat here. It is instructive, however, to point out that, since the normalization condition imposed by Equation (8), the $20+\lambda+\mu$ components of the pseudo amino acid composition are not independent. Therefore, a dimension-reduced operation by leaving out one of the components and making the rest completely independent is needed when using the augmented covariant discriminant algorithm; i.e., a protein should be defined in a $(20+\lambda+\mu-1)$-D space instead of $(20+\lambda+\mu)$-D space. Otherwise, a divergence difficulty will occur. However, which one of the $20+\lambda+\mu$ components should be removed? Anyone. The reason is that according to the "*Invariance Theorem*" given in Appendix A of Chou [1995], the values of the covariant discriminant function will remain the same regardless of which one of the $20+\lambda+\mu$ components is left out. The theorem can also be used to address similar problems occurring in other algorithms that involve covariance matrix [Cedano et al., 1997; Zhou, 1998; Zhou and Assa-Munt, 2001; Pan et al., 2003; Zhou and Doctor, 2003].

RESULTS AND DISCUSSION

The newly constructed datasets as given in the Online Supplemental Materials A and B will serve as the training and independent testing datasets, respectively. Both consist of 14 subsets corresponding to 14 subcellular locations (Fig. 1). Compared with the 12 subcellular location dataset [Chou and Elrod, 1999b] that is so far the largest in number of locations considered, the datasets used here cover even more localizations.

The demonstration was conducted by three most typical approaches in statistical prediction [Chou and Zhang, 1995]; i.e., the re-substitution test, jackknife test, and independent dataset test. Since the sequence-order effects are incorporated through the pseudo amino acid components (Eq. 7), a question is naturally raised: how many pseudo amino acid components should be used, or what numbers should be assigned for λ and μ during prediction? Actually, these numbers are determined through an optimal process by maximizing the success rate from the jackknife test. This is because among the independent dataset test, sub-sampling test and jackknife test often used for cross-validation in statistical prediction, the jackknife test is deemed as the most effective and objective one; see, e.g., Chou and Zhang [1995] for a comprehensive discussion about this, and Mardia et al. [1979] for the mathematical principle. The optimal values thus obtained for the current training dataset are $\lambda=\mu=13$, meaning that the pseudo amino acid composition contains 13 delta-function correlation factors and 13 hydrophobicity correlation factors and that any protein in this study should be represented by a 46-D vector (see Eq. 7).

Re-Substitution Test

The so-called re-substitution test is an examination for the self-consistency of a prediction method. When the re-substitution test is performed for the current study, the subcellular location of each protein in the dataset is in turn identified using the rule parameters derived from the same data set, the so-called training or learning dataset. The overall success rate thus obtained for predicting the 14 subcellular locations of the 3,799 proteins is given in Table I, from which we can see that, of the 3,799 proteins, 3,245 were correctly predicted for their subcellular locations, and only 554 proteins incorrectly predicted. The overall success rate is 85.4%, indicating a good self-consistency for such a complicated problem involving 14 subcellular locations. 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

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

Algorithm	Input form	Test method		
		Resubstitution[a]	Jackknife[a]	Independent dataset[b]
Least Hamming distance [Chou, 1989]	Amino acid composition[c]	$\frac{1428}{3799} = 37.6\%$	$\frac{1392}{3799} = 36.6\%$	$\frac{1371}{4498} = 30.5\%$
Least Euclidean distance [Nakashima et al., 1986]	Amino acid composition[c]	$\frac{1391}{3799} = 36.6\%$	$\frac{1361}{3799} = 35.8\%$	$\frac{1388}{4498} = 30.9\%$
ProtLock [Cedano et al., 1997]	Amino acid composition[c]	$\frac{1655}{3799} = 43.6\%$	$\frac{1614}{3799} = 42.5\%$	$\frac{1829}{4498} = 40.7\%$
Covariant discriminant [Chou and Elrod, 1999a]	Amino acid composition[c]	$\frac{2580}{3799} = 67.9\%$	$\frac{2339}{3799} = 61.6\%$	$\frac{2751}{4498} = 61.2\%$
Augmented covariant discriminant [Chou, 2000a]	Pseudo amino acid composition[d]	$\frac{3245}{3799} = 85.4\%$	$\frac{2574}{3799} = 67.8\%$	$\frac{3246}{4498} = 72.2\%$

[a] Conducted for the 3,799 proteins classified into 14 subcellular locations in the training dataset as given in the Online Supplemental Material A.

[b] Conducted based on the rule parameters derived from the 3,799 proteins in the training dataset for the 4,498 proteins in the independent dataset given in the Online Supplemental Material B.

[c] The dimension of the conventional amino acid composition is 20, where no sequence-order effects are incorporated.

[d] The dimension of the pseudo amino acid composition for the current study is 46, where the sequence-order effects are incorporated through 13 delta-function correlation factors ($\lambda = 13$) and 13 hydrophobicity correlation factors ($\mu = 13$); see Equation (7).

some sort of optimistic estimation [Chou, 1995; Chou and Elrod, 1999b; Cai, 2001; Zhou and Assa-Munt, 2001; Pan et al., 2003]. Nevertheless, the re-substitution test is absolutely necessary because it reflects the self-consistency of an identification method, especially for its algorithm part. An identification 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 an identification method. As a complement, a cross-validation test for an independent testing data set is needed because it can reflect the effectiveness of an identification method 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.

Jackknife Test

As mentioned above, jackknife test is the key for examining a prediction method [Mardia et al., 1979; Chou and Zhang, 1995; Zhou and Assa-Munt, 2001]. 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 which is 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 success rate by jackknife test thus obtained for the 3,799 proteins is given in Table I as well.

Independent Dataset Test

Moreover, as a demonstration of practical application, predictions were also conducted for the 4,498 proteins in the independent dataset (Online Supplemental Materials B) based on the rule-parameters derived from the 3,799 proteins in the training dataset (Online Supplemental Materials A). The result thus obtained is also given in Table I, from which we can see that, of the 4,498 proteins, 3,246 were correctly predicted for their subcellular locations. The overall success rate is 72.2%.

Furthermore, to facilitate comparison, the results predicted by various other methods on the same datasets are also listed in Table I, stimulating the following discussions. (1) If the samples of proteins are completely randomly assigned among the 14 possible subsets, the success rate would generally be $1/14 \approx 7.1\%$, and the corresponding rate by the weighted random assignment would be $(71/3799)^2 + (65/3799)^2 + (316/3799)^2 + (1113/3799)^2 + (249/3799)^2 + (289/3799)^2 + (393/3799)^2 + (90/3799)^2 + (123/3799)^2 + (389/3799)^2 + (399/3799)^2 + (147/3799)^2 + (69/3799)^2 + (86/3799)^2 \approx 13.9\%$, provided that the number of proteins in each subcellular location as given in the Online Supplemental Materials A is used to represent the weight of each subset. Therefore, all the rates listed in Table 1 are significantly higher than the corresponding completely randomized rate and weighted randomized rate, implying that the amino acid composition does play an important (although not a unique) role for protein subcellular location. (2) No matter whether the re-substitution test, jackknife test or independent dataset test, the overall rates of correct prediction obtained by the current pseudo-amino acid composition approach, are significantly higher than those by the previous approaches: 31–49% higher than the simple geometry algorithms [Nakashima et al., 1986; Chou, 1989]; 25–42% higher than the ProtLock algorithm [Cedano et al., 1997]; and 6–18% higher than the covariant-discriminant algorithm [Chou and Elrod, 1999b]. This is fully consistent with what is expected because all the sequence-order effects are completely ignored in those approaches. (3) The success rates by jackknife test are decreased compared with those by the re-substitution test. The decrement is more remarkable for small subsets, such as centriole subset. This is because the cluster-tolerant capacity [Chou, 1999a] for small subsets is usually low. And hence the information loss resulting from jackknifing will have a greater impact upon the small subsets than the large ones. However, the overall success rate by the current approach can still reach 67.8%, which is significantly higher than the corresponding rates by the other approaches. It is also due to the information loss during jackknifing that the overall success rate is not always monotonously increased with the dimension of pseudo amino acid composition. Actually, different training dataset may have different optimal number of pseudo amino acid components to yield the highest overall jackknife success rate, as discussed earlier [Chou, 2001]. It is expected that the overall jackknife success rate can be further enhanced through improving the cluster-tolerant capacity of small subsets by adding into them more new proteins that have been found belonging to these subsets. (4) Narrowing the scope of localization will increase the success rate of prediction. For example, when the protein subcellular locations to be identified was reduced to the scope among chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracell, and nucleus, the corresponding success rate by jackknife test was increased from 67.8% to 77.3%. This indicates that the prediction quality can be substantially improved if one can narrow down the scope of subcellular location for a query protein according to its source and other relevant information (e.g., if a query protein is from an animal organism, one can safely exclude the chloroplast and vacuole subsets from consideration).

CONCLUSION

The development in statistical prediction of protein subcellular location generally consists of two cores: one is to construct a training dataset and the other is to formulate a prediction algorithm. The process in constructing training dataset from two subsets [Nakashima and Nishikawa, 1994], to five subsets [Cedano et al., 1997], to 12 subsets [Chou and Elrod, 1999b], and to the current 14 subsets reflects the advance in the first core. The second core can be further separated into two sub-cores: one is how to give a mathematical expression to effectively represent a protein and the other is how to find an operational equation to accurately perform the prediction. The process in expressing a protein from the 20-D amino-acid-composition space [Nakashima et al., 1986; Chou, 1989, 1995], to the $(20 + \lambda)$-D parallel-mode pseudo-amino-acid-composition space [Chou, 2001], and to the current $(20 + \lambda + \mu)$ series-mode pseudo-amino-acid-composition space reflects the progress of defining a protein in terms of different mathematical representations. The process in conducting prediction using the operation from the simple geometry distance algorithms [Nakashima et al., 1986; Chou, 1989], to the Mahalanobis distance algorithm [Cedano et al., 1997], to the covariant discriminant algorithm [Chou and Elrod, 1999b], and to the augmented covariant discriminant algorithm [Chou, 2000a] reflects the development by means of different mathematical operations.

The datasets constructed in this study have covered so far the largest area of protein localization. Therefore, many proteins which were totally out of the classification scheme before [Chou and Elrod, 1999b], such as those belonging to the localization of cell wall and centriole, can now be investigated.

One of the remarkable advantages for the series-mode pseudo-amino-acid-composition representation is that it allows containing two or more sets of factors generated by different sequence-coupled modes, and hence providing more potential and flexibility to incorporate the sequence-order effects. It has not escaped our notice that the series-mode pseudo-amino-acid-composition introduced here may become a very useful vehicle in proteomics and bioinformatics that can be used to improve the quality for predicting many other important characteristics and attributes of proteins as well, such as secondary structure contents [Zhang et al., 1996; Chou, 1999b; Liu and Chou, 1999], folding classes [Chou et al., 1998; Liu and Chou, 1998; Zhou, 1998], GPCR types [Chou and Elrod, 2002; Elrod and Chou, 2002], enzyme family classes [Chou and Elrod, 2003], and protein quaternary structure attributes [Chou and Cai, 2003].

ACKNOWLEDGMENTS

The authors thank the three anonymous referees whose comments were very helpful for strengthening the presentation of this paper.

REFERENCES

Bahar I, Atilgan AR, Jernigan RL, Erman B. 1997. Understanding the recognition of protein structural classes by amino acid composition. Proteins: Structure, Function, and Genetics 29:172–185.

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: Structure, Function, and Genetics 43:336–338.

Cai YD, Liu XJ, Xu XB, Chou KC. 2002. Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem 84:343–348.

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 PY. 1989. Prediction of protein structural classes from amino acid composition. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York: Plenum Press. pp 549–586.

Chou KC. 1995. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function, and Genetics 21: 319–344.

Chou KC. 1999a. A key driving force in determination of protein structural classes. Biochem Biophysical Res Commun 264:216–224.

Chou KC. 1999b. Using pair-coupled amino acid composition to predict protein secondary structure content. J Protein Chem 18:473–480.

Chou KC. 2000a. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochemical Biophysical Res Commun 278:477–483.

Chou KC. 2000b. 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: Structure, Function, and Genetics 43:246–255 (Erratum: Proteins: Struct Funct Genet 2001, Vol. 44:60).

Chou KC. 2002a. A new branch of proteomics: prediction of protein cellular attributes. Chapter 4. In: Weinrer PW, Lu Q, editors. Gene cloning and expression technologies. Westborough, MA: Eaton Publishing. pp 57–70.

Chou KC. 2002b. Prediction of protein signal sequences. Curr Protein Peptide Sci 3:615–622.

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. Predicting protein quaternary structure by pseudo amino acid composition. Proteins: Structure, Function, and Genetics 53:282–289.

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

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

Chou KC, Elrod DW. 2002. Bioinformatical analysis of G-protein-coupled receptors. J Proteome Res 1:429–433.

Chou KC, Elrod DW. 2003. Prediction of enzyme family classes. J Proteome Res 2:183–190.

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, Liu W, Maggiora GM, Zhang CT. 1998. Prediction and classification of domain structural classes. Proteins: Structure, Function, and Genetics 31:97–103.

Claros MG, Brunak S, von Heijne G. 1997. Prediction of N-terminal protein sorting signals. Curr Opin Struct Biol 7:394–398.

Elrod DW, Chou KC. 2002. A study on the correlation of G-protein-coupled receptor types with amino acid composition. Protein Eng 15:713–715.

Krigbaum WR, Knutton SP. 1973. Prediction of the amount of secondary structure in a globular protein from its amino acid composition. Proc Natl Acad Sci USA 70: 2809–2813.

Liu W, Chou KC. 1998. Prediction of protein structural classes by modified Mahalanobis discriminant algorithm. J Protein Chem 17:209–217.

Liu W, Chou KC. 1999. Protein secondary structural content prediction. Protein Eng 12:1041–1050.

Mardia KV, Kent JT, Bibby JM. 1979. Multivariate analysis. London, UK: Academic Press. pp 322, 381.

Murvai J, Vlahovicek K, Barta E, Pongor S. 2001. The SBASE protein domain library, release 8.0: A collection of annotated protein sequence segments. Nucleic Acids Res 29:58–60.

Muskal SM, Kim S-H. 1992. Predicting protein secondary structure content: A tandem neural network approach. J Mol Biol 225:713–727.

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.

Nakashima H, Nishikawa K, Ooi T. 1986. The folding type of a protein is relevant to the amino acid composition. J Biochem 99:152–162.

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.

Radford T. 2003. Metaphors and dreams. The Scientist 17: 24–26.

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

Tanford C. 1962. Contribution of hydrophobic interaction to the stability of the globular conformation of proteins. J Amer Chem Soc 84:4240–4274.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

Zhang CT, Zhang Z, He Z. 1996. Prediction of the secondary structure content of globular proteins based on structural classes. J Protein Chem 15:775–786.

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: Structure, Function, and Genetics 44:57–59.

Zhou GP, Doctor K. 2003. Subcellular location prediction of apoptosis proteins. Proteins: Structure, Function, and Genetics 50:44–48.