PROTEINS: Structure, Function, and Genetics 43:246–255 (2001)

# Prediction of Protein Cellular Attributes Using Pseudo-Amino Acid Composition

**Kuo-Chen Chou***
*Computer-Aided Drug Discovery, Pharmacia, Kalamazoo, Michigan*

***ABSTRACT*** **The cellular attributes of a protein, such as which compartment of a cell it belongs to and how it is associated with the lipid bilayer of an organelle, are closely correlated with its biological functions. The success of human genome project and the rapid increase in the number of protein sequences entering into data bank have stimulated a challenging frontier: How to develop a fast and accurate method to predict the cellular attributes of a protein based on its amino acid sequence? The existing algorithms for predicting these attributes were all based on the amino acid composition in which no sequence order effect was taken into account. To improve the prediction quality, it is necessary to incorporate such an effect. However, the number of possible patterns for protein sequences is extremely large, which has posed a formidable difficulty for realizing this goal. To deal with such a difficulty, the pseudo-amino acid composition is introduced. It is a combination of a set of discrete sequence correlation factors and the 20 components of the conventional amino acid composition. A remarkable improvement in prediction quality has been observed by using the pseudo-amino acid composition. The success rates of prediction thus obtained are so far the highest for the same classification schemes and same data sets. It has not escaped from our notice that the concept of pseudo-amino acid composition as well as its mathematical framework and biochemical implication may also have a notable impact on improving the prediction quality of other protein features. Proteins 2001;43:246–255.** © 2001 Wiley-Liss, Inc.

**Key words: proteomics; protein cellular locations; membrane protein types; sequence-order-correlated factors; covariant-discriminant algorithm**

## INTRODUCTION

The success of the human genome project has stimulated the emergence of a new and far more challenging frontier: proteomics. Although proteins are generated according to their DNA code, they are far more complex and varied than DNA. To understand the molecular underpinnings of life, it is indispensable to study individual proteins, their complexes and cellular networks. This is what proteomics is all about. Actually, proteomics is the science of the cellular protein universe that will hold a key position in the new biology and medicine reshaped by the monumental achievement of the sequencing of the human genome. Every biochemical reaction essential to life depends on the marvelous functions of proteins in one way or another. For example, proteins can serve as the following: the beams and rafters of the cell; the glue that binds the body together; the enzymes that build up and break down our energy reserves; the "circuits" that power movement and thought; the hormones that course through our veins; the "guided missiles" that target infections; and much more. On the other hand, the function of a protein is closely correlated with its cellular attributes, such as in which subcellular location it resides, whether a membrane or not a membrane protein, and if so, what type of membrane protein it is.[1,2] The knowledge of protein cellular attributes plays a vitally important role in molecular biology, cell biology, pharmacology, and medical science.

Although the cellular attributes of a protein can be determined by conducting various experiments, it is time-consuming and costly to acquire this kind of knowledge solely by experiments. Because the number of sequences entering into databanks has been rapidly increasing, the challenge in expediting the determination procedure for protein cellular attributes has become critical and urgent. As shown in Table I, the number of sequence entries in SWISS-PROT[3] was 3,939 in 1986, in 1990 was 18,374, but already 80,000 in 1999. Even so, this is still a tiny fraction of all the proteins found in nature, estimated as about 2 million in number, depending how you count. In view of this, it is highly desirable to develop a theoretical method for quickly and accurately predicting the cellular attributes of proteins.

Actually, many efforts have been made in this regard.[4–9] It should be pointed out that the prediction algorithms by these authors were all based on the amino acid composition alone. Although this is a reasonable approximate approach and did yield some encouraging results as discussed in a recent review,[10] the prediction quality will be certainly improved if sequence order information can also be incorporated into the prediction algorithm. However, this is a very difficult task, as exemplified below. For a protein of only 50 residues, the number of different sequence order combinations would be $20^{50} \approx 1.1259 \times 10^{65}$. Actually, as shown in Table I, the average protein length is

*Correspondence to: Kuo-Chen Chou, Computer-Aided Drug Discovery, Pharmacia, MI 49007-4940. E-mail: kuo-chen.chou@am.pnu.com

Received 30 September 2000; Accepted 4 January 2001

Published online 00 Month 2001

**TABLE I. Growth of Protein Sequences in SWISS-PROT Data Bank**

| Release | Date | No. of sequence entries | No. of amino acids | Average length per sequence[a] |
|---|---|---|---|---|
| 2.0 | 09/86 | 3,939 | 900,163 | 229 |
| 5.0 | 09/87 | 5,205 | 1,327,683 | 236 |
| 9.0 | 11/88 | 8,702 | 2,498,140 | 287 |
| 12.0 | 10/89 | 12,305 | 3,797,482 | 309 |
| 16.0 | 11/90 | 18,364 | 5,986,949 | 326 |
| 20.0 | 11/91 | 22,654 | 7,500,130 | 331 |
| 24.0 | 12/92 | 28,154 | 9,545,427 | 339 |
| 27.0 | 10/93 | 33,329 | 11,484,420 | 345 |
| 30.0 | 10/94 | 40,292 | 14,147,368 | 351 |
| 32.0 | 11/95 | 49,340 | 17,385,503 | 352 |
| 34.0 | 10/96 | 59,021 | 21,210,389 | 359 |
| 35.0 | 11/97 | 69,113 | 25,083,768 | 363 |
| 37.0 | 12/98 | 77,977 | 28,268,293 | 363 |
| 38.0 | 07/99 | 80,000 | 29,085,965 | 364 |

[a]The average length per sequence is defined as the total number of amino acids divided by the total number of sequences. The quotient is rounded to an integer.

much longer than 50, and the database has longer proteins each year. In 1999 the average length increased to 364 residues. The number of different combinations for a protein of 364 residues will be $20^{364} \gg 1.1259 \times 10^{65}$! For such a huge number, it is impractical to construct a training data set to statistically cover all possible samples based on the current protein data. Furthermore, protein sequence lengths vary widely. This has posed an additional difficulty for including sequence order information, in both constructing a training data set and formulating an algorithm. Confronted with such a grim situation, how can we take into account the sequence order effect to improve the prediction quality? If it is not feasible to completely include all sequence order patterns, is there an approximate approach to partially count the sequence order effect? If so, what is the approach and how does it improve the prediction quality? The present study was initiated in an attempt to address these problems. First, let us introduce a new concept, the so-called "pseudo-amino acid composition," as described below.

## THE PSEUDO-AMINO ACID COMPOSITION

As illustrated above, owing to the huge number of possible sequence order patterns, it is very difficult to directly incorporate the sequence order effect into a statistical prediction algorithm. However, if we can indirectly take this effect into account through a set of discrete numbers, the problem will become much easier. That is why we try to introduce the pseudo-amino acid composition.

The essence of pseudo-amino acid composition is, on one hand, to include the main feature of amino acid composition, but on the other, to include information beyond amino acid composition. The conventional amino acid composition contains 20 components, or discrete numbers, each reflecting the occurrence frequency of one of the 20 native

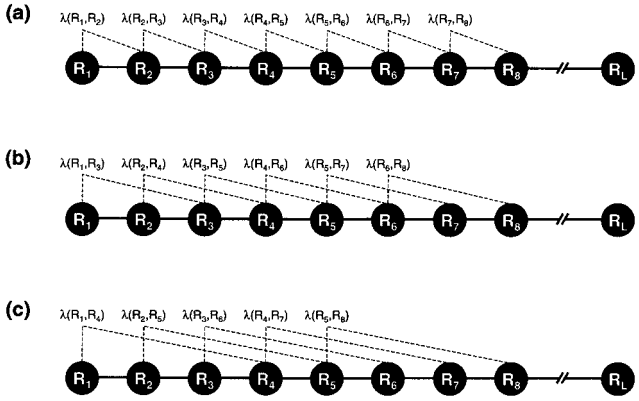


Fig. 1. A schematic drawing to show (a) the first-tier, (b) the second-tier, and (3) the third-tier sequence order correlation mode along a protein sequence. Panel (**a**) reflects the correlation mode between all the most contiguous residues, panel (**b**) that between all the secon-most contiguous residues, and panel (**c**) that between all the third-most contiguous residues.

amino acids in a protein. For the pseudo-amino acid composition, however, there are some other elements in addition to the 20 components. It is through these additional discrete numbers that the sequence order effect of a protein is approximately reflected and improvements are made, as will be shown below.

Consider a protein chain of $L$ amino acid residues:

$$R_1R_2R_3R_4R_5R_6R_7 \cdots R_L \tag{1}$$

effect can be approximately reflected with a set of sequence order-correlated factors as defined below:

$$\begin{cases} \theta_1 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1} \Theta(R_i, R_{i+1}) \\ \theta_2 = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2} \Theta(R_i, R_{i+2}) \\ \theta_3 = \dfrac{1}{L-3}\displaystyle\sum_{i=1}^{L-3} \Theta(R_i, R_{i+3}), \quad (\lambda < L) \\ \cdots \\ \theta_\lambda = \dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda} \Theta(R_i, R_{i+\lambda}) \end{cases} \tag{2}$$

where $\theta_1$ is called the first-tier correlation factor that reflects the sequence order correlation between all the most contiguous residues along a protein chain (Fig. 1a), $\theta_2$ the second-tier correlation factor that reflects the sequence order correlation between all the second most contiguous residues (Fig.1b), $\theta_3$ the third-tier correlation factor that reflects the sequence order correlation between all the 3rd most contiguous residues (Fig.1c), and so forth. In Eq. 2 the correlation function is given by

$$\Theta(R_i, R_j) = \frac{1}{3}\{[H_1(R_j) - H_1(R_i)]^2 + [H_2(R_j) - H_2(R_i)]^2 + [M(R_j) - M(R_i)]^2\} \tag{3}$$

where $H_1(R_i)$, $H_2(R_i)$, and $M(R_i)$ are, respectively, the hydrophobicity value, hydrophilicity value, and side-chain mass of the amino acid $R_{i;}$ and $H_1(R_j)$, $H_2(R_j)$, and $M(R_j)$ the corresponding values for the amino acid $R_j$. Note that before substituting the values of hydophobicity, hydrphilicity, and side-chain mass into Eq. 3, they were all subjected to a *standard conversion* as described by the following equation:

$$\begin{cases} H_1(i) = \dfrac{H_1^0(i) - \sum_{i=1}^{20} \dfrac{H_1^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20} \left[ H_1^0(i) - \sum_{i=1}^{20} \dfrac{H_1^0(i)}{20} \right]^2}{20}}} \\ H_2(i) = \dfrac{H_2^0(i) - \sum_{i=1}^{20} \dfrac{H_2^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20} \left[ H_2^0(i) - \sum_{i=1}^{20} \dfrac{H_2^0(i)}{20} \right]^2}{20}}} \\ M(i) = \dfrac{M^0(i) - \sum_{i=1}^{20} \dfrac{M^0(i)}{20}}{\sqrt{\dfrac{\sum_{i=1}^{20} \left[ M^0(i) - \sum_{i=1}^{20} \dfrac{M^0(i)}{20} \right]^2}{20}}} \end{cases} \quad (4)$$

where $H_1^0(i)$ is the original hydrophobicity value of the $i$th amino acid that was taken from Tanford,[11] $H_2^0(i)$ the corresponding original hydrophilicity value taken from Hopp and Woods,[12] and $M^0(i)$ the mass of the $i$th amino acid side chain that can be obtained from any biochemistry text book. Without loss of generality, 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. The data obtained by such a standard conversion (Eq. 4) will have a zero mean value and will remain unchanged if going through the same conversion procedure again.

As we can see from Figure 1, the sequence order effect of a protein can be, to some extent, reflected through a set of sequence-correlation factors $\theta_{1,}$ $\theta_{2,}$ $\theta_3$,…, $\theta_\lambda$, as defined by Eq. 2. Now let us augment the formulation of amino acid composition to include such a set of discrete numbers. To realize this, instead of using a 20-D (dimensional) vector defined by 20 components,[13] we use a (20 + λ)-D vector defined by 20 + λ discrete numbers to represent a protein **X**; i.e.,

$$\mathbf{X} = \begin{bmatrix} x_1 \\ \vdots \\ x_{20} \\ x_{20+1} \\ \vdots \\ x_{20+\lambda} \end{bmatrix}, \quad (5)$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j}, & (1 \le u \le 20) \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j}, & (20+1 \le u \le 20+\lambda) \end{cases} \quad (6)$$

where $f_i$ is the normalized occurrence frequency of the 20 amino acids in the protein **X**, $\theta_j$ is the $j$-tier sequence correlation factor computed according to Eqs. 2–4 for the protein **X**, and $w$ is the weight factor for the sequence order effect. In the current study, we chose $w = 0.05$. As we can see from Eqs. 5 and 6, the first 20 components reflect the effect of the amino acid composition, whereas the components from 20 + 1 to 20 + λ reflect the effect of sequence order. A set of such 20 + λ components as formulated by Eqs. 5 and 6 is called the pseudo-amino acid composition for protein **X**. It has the following three advantages: (a) It contains more sequence order effects not only than the 20-D conventional amino acid composition[13] but also than the 210-D pair-coupled amino acid composition[14] and the 400-D first-order coupled amino acid-composition,[15] as reflected by a series of sequence correlation factors with different tiers of correlation (see Fig.1 and Eq. 2). (b) These factors are defined by a correlation function that allows users to introduce any other biochemical quantities (in addition to the hydrophobicity, hydrophilicity, and side-chain mass as explicitly expressed in Eq. 3) to obtain the optimal results for various cases concerned. (c) The pseudo-amino acid composition has the same formulation as the conventional one except containing more components (discrete numbers); accordingly, all the existing prediction algorithms based on the conventional amino acid composition can be straightforwardly extended to cover the pseudo-amino acid composition, as described below.

## THE PREDICTION ALGORITHMS

In this section we shall show how the existing prediction algorithms are reformulated in terms of the pseudo-amino acid composition.

Suppose there are $N$ proteins forming a set $S$, which is the union of $m$ subsets; i.e.,

$$S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup \cdots \cup S_m \quad (7)$$

Each subset is composed of proteins with the same cellular attribute. Its size is given by $n_\xi$ ($\xi = 1, 2, 3, \ldots, m$), where $n_\xi$ represents the number of proteins in the subset $S_\xi$. Obviously, $N = n_1 + n_2 + \ldots + n_m$. The key to the incorporation of the sequence order effect is to replace the conventional amino acid composition by the pseudo-amino acid composition as formulated in Eq. 5, where a protein is represented by a vector or a point in a (20 + λ)-D space, rather than a 20-D space as used by the previous investigators.[5,8,13,16,17] According to such a baseline, the $k$th protein in the subset $S_\xi$ should now be represented by

$$\mathbf{X}_k^{\xi} = \begin{bmatrix} x_{k,1}^{\xi} \\ x_{k,2}^{\xi} \\ \vdots \\ x_{k,20+\lambda}^{\xi} \end{bmatrix}, (k = 1, 2,. . ., n_{\xi}; \quad \xi = 1, 2,. . ., m), \tag{8}$$

where $x_{k,u}^{\xi}$ (u = 1, 2,. . ., 20 + λ) has the same meaning as $x_u$ of Eq. 5 but is associated with $\mathbf{X}_k^{\xi}$ instead of $\mathbf{X}$. The *standard vector* for the subset $S_{\xi}$ is defined by

$$\bar{\mathbf{X}}^{\xi} = \begin{bmatrix} \bar{x}_1^{\xi} \\ \bar{x}_2^{\xi} \\ \vdots \\ \bar{x}_{20+\lambda}^{\xi} \end{bmatrix}, \quad (\xi = 1, 2, 3,. . ., m) \tag{9}$$

where

$$\bar{x}_i^{\xi} = \frac{1}{n_{\xi}} \sum_{i=1}^{n_{\xi}} x_{k,i}^{\xi}, \quad (i = 1, 2,. . ., 20 + \lambda). \tag{10}$$

Actually $\bar{\mathbf{X}}^{\xi}$ as defined above can be viewed to represent a standard protein (a pseudo-protein) for the subset $S_{\xi}$.

Suppose $\mathbf{X}$ is the protein whose cellular attribute is to be predicted. It can be either one of the $N$ proteins in the set $S$, or a protein outside of it. Its pseudo-amino acid composition has been given by Eq. 5. Now the problem is how to effectively define the similarity between the query protein $\mathbf{X}$ and the *standard vectors* $\bar{\mathbf{X}}^{\xi}$. Algorithms with various measures were proposed, as follows.

### The Least Hamming Distance Algorithm[16,17]

The algorithm was originally proposed by Chou[16] to predict protein structural class based on the 20-D amino acid composition. The hypothesis was that the similarity of any two proteins could be measured by their Hamming distance or city-block metric.[18] Now, based on the pseudo-amino acid composition, the Hamming distance between $\mathbf{X}$ and $\bar{\mathbf{X}}^{\xi}$ should be

$$D_{\mathrm{H}}(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = \sum_{i=1}^{20+\lambda} |x_i - \bar{x}_i^{\xi}|, \quad (\xi = 1, 2, 3,. . ., m) \tag{11}$$

Thus, the prediction rule was given by

$$D_{\mathrm{H}}(\mathbf{X}, \bar{\mathbf{X}}^{\mu}) = \mathrm{Min}\{D_{\mathrm{H}}(\mathbf{X}, \bar{\mathbf{X}}^{1}), D_{\mathrm{H}}(\mathbf{X}, \bar{\mathbf{X}}^{2}), D_{\mathrm{H}}(\mathbf{X}, \bar{\mathbf{X}}^{3}),. . ., D_{\mathrm{H}}(\mathbf{X}, \bar{\mathbf{X}}^{m})\} \tag{12}$$

where μ can be 1, 2, 3,…, or $m$, and the operator **Min** means taking the least one among those in the brackets, and the superscript μ is the cellular attribute predicted for the query protein $\mathbf{X}$. If there is a tie, μ is not uniquely determined, but that rarely occurs.

### The Least Euclidean Distance Algorithm[13]

Rather than Hamming distance, Nakashima et al.[13] used the square Euclidean distance as a scale to measure the similarity between two proteins. Thus, instead of Eqs. 11 and 12, the similarity between $\mathbf{X}$ and $\bar{\mathbf{X}}^{\xi}$ should be defined by

$$D_{\mathrm{E}}^2(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = \sum_{1}^{20+\lambda} (x_i - \bar{x}_i^{\xi})^2, \quad (\xi = 1, 2, 3,. . ., m) \tag{13}$$

and the prediction rule given by

$$D_{\mathrm{E}}^2(\mathbf{X}, \bar{\mathbf{X}}^{\mu}) = \mathbf{Min}\{D_{\mathrm{E}}^2(\mathbf{X}, \bar{\mathbf{X}}^{1}), D_{\mathrm{E}}^2(\mathbf{X}, \bar{\mathbf{X}}^{2}), D_{\mathrm{E}}^2(\mathbf{X}, \bar{\mathbf{X}}^{3}),. . ., D_{\mathrm{E}}^2(\mathbf{X}, \bar{\mathbf{X}}^{m})\}. \tag{14}$$

### The ProtLock Algorithm[5]

The scale to measure the similarity between proteins $\mathbf{X}$ and $\bar{\mathbf{X}}^{\xi}$ in the ProtLock algorithm[5] should be modified as

$$D_{\mathrm{P}}^2(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = (\mathbf{X} - \bar{\mathbf{X}}^{\xi})^{\mathbf{T}} \mathbf{C}^{-1} (\mathbf{X} - \bar{\mathbf{X}}^{\xi}), \quad (\xi = 1, 2, 3,. . ., m) \tag{15}$$

where $\mathbf{C}$ is a matrix defined by

$$\mathbf{C} = \begin{bmatrix} c_{1,1} & c_{1,2} & \cdots & c_{1,20+\lambda} \\ c_{2,1} & c_{2,1} & \cdots & c_{2,20+\lambda} \\ \vdots & \vdots & \ddots & \vdots \\ c_{20+\lambda,1} & c_{20+\lambda,2} & \cdots & c_{20+\lambda,20+\lambda} \end{bmatrix}, \tag{16}$$

the superscript $\mathbf{T}$ is the transposition operator, and $\mathbf{C}^{-1}$ the inverse matrix of $\mathbf{C}$. The matrix elements $c_{i,j}$ in Eq. 16 are given by

$$c_{i,j} = \sum_{\xi=1}^{m} \sum_{k=1}^{n_{\xi}} [x_{k,i}^{\xi} - \bar{x}_i][x_{k,j}^{\xi} - \bar{x}_j], \quad (i, j = 1, 2, \ldots, 20 + \lambda), \tag{17}$$

where

$$x_i = \frac{1}{N} \sum_{\xi=1}^{m} \sum_{k=1}^{n_{\xi}} x_{k,i}^{\xi} = \frac{1}{N} \sum_{\xi=1}^{m} n_{\xi} \bar{x}_i^{\xi}, \quad (i = 1, 2,. . ., 20 + \lambda). \tag{18}$$

And the prediction rule should be given by

$$D_{\mathrm{P}}^2(\mathbf{X}, \bar{\mathbf{X}}^{\mu}) = \mathbf{Min}\{D_{\mathrm{P}}^2(\mathbf{X}, \bar{\mathbf{X}}^{1}), D_{\mathrm{P}}^2(\mathbf{X}, \bar{\mathbf{X}}^{2}), D_{\mathrm{P}}^2(\mathbf{X}, \bar{\mathbf{X}}^{3}),. . ., D_{\mathrm{P}}^2(\mathbf{X}, \bar{\mathbf{X}}^{m})\}. \tag{19}$$

### The Covariant Discriminant Algorithm[7,8]

Instead of geometrical distance, in the covariant discriminant algorithm a function was used as a scale to measure the similarity between proteins $\mathbf{X}$ and $\bar{\mathbf{X}}^{\xi}$. The smaller the value of the function, the greater the similarity between the two proteins. Based on the pseudo-amino acid composition, the function, which was called the covariant discriminant function, should be expressed as

$$F(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = D_{\mathrm{M}}^2(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) + \ln|\mathbf{C}^{\xi}|, \quad (\xi = 1, 2, 3,. . ., m) \tag{20}$$

where

$$D_{\mathrm{M}}^2(\mathbf{X}, \bar{\mathbf{X}}^{\xi}) = (\mathbf{X} - \bar{\mathbf{X}}^{\xi})^{\mathbf{T}} \mathbf{C}_{\xi}^{-1} (\mathbf{X} - \bar{\mathbf{X}}^{\xi}) \tag{21}$$

is the squared Mahalanobis distance[19–21] between $\mathbf{X}$ and $\bar{\mathbf{X}}^{\xi}$, and

$$\mathbf{C}_{\xi} = \begin{bmatrix} c^{\xi}_{1,1} & c^{\xi}_{1,2} & \cdots & c^{\xi}_{1,20+\lambda} \\ c^{\xi}_{2,1} & c^{\xi}_{2,1} & \cdots & c^{\xi}_{2,20+\lambda} \\ \vdots & \vdots & \ddots & \vdots \\ c^{\xi}_{20+\lambda,1} & c^{\xi}_{20+\lambda,2} & \cdots & c^{\xi}_{20+\lambda,20+\lambda} \end{bmatrix} \quad (22)$$

is the covariance matrix for the subset $S_{\xi}$ and its elements are given by

$$c^{\xi}_{i,j} = \frac{1}{n_{\xi} - 1} \sum_{k=1}^{n_{\xi}} [x^{\xi}_{k,i} - \bar{x}^{\xi}_{i}][x^{\xi}_{k,j} - \bar{x}^{\xi}_{j}], \quad (i, j = 1, 2, \ldots, 20 + \lambda), \quad (23)$$

and $|\mathbf{C}^{\xi}|$ is the determinant of the matrix $\mathbf{C}^{\xi}$. Likewise, the prediction rule should be expressed by

$$F(\mathbf{X}, \bar{\mathbf{X}}^{\mu}) = \mathbf{Min}\{F(\mathbf{X}, \bar{\mathbf{X}}^{1}), F(\mathbf{X}, \bar{\mathbf{X}}^{2}), F(\mathbf{X}, \bar{\mathbf{X}}^{3}), \ldots, F(\mathbf{X}, \bar{\mathbf{X}}^{m})\}. \quad (24)$$

Note that the sum of the 20 + λ components in Eq. 6 is equal to 1 (imposed by the normalization condition); i.e., of the 20 + λ components, only 20 + λ−1 are independent. Accordingly, the covariance matrix $\mathbf{C}^{\xi}$ as defined by Eq. 23 must be a singular matrix.[21] This implies that the Mahalanobis distance given by Eq. 21 and the corresponding covariant discriminant function by Eq. 20 would be divergent and be undefined. To overcome such a difficulty, let us take the following dimension-reducing procedure.[21] Instead of defining a protein in a (20 + λ)-D space, let us define it in a (20 + λ−1)-D space by leaving out one of its pseudo-amino acid components. The remaining components thus obtained would be completely independent and hence the corresponding covariance matrix $\mathbf{C}^{\xi}$ no longer singular. In such a (20 + λ−1)-D space, the Mahalanobis distance (Eq. 21) and covariant discriminant function (Eq. 20) can be defined without the difficulty of divergence. Furthermore, according to the *invariance theorem* given in Appendix A of Chou,[21] the values of the Mahalanobis distance and covariant discriminant function will remain the same regardless of which one of the 20 + λ components is left out. Accordingly, the values of the Mahalanobis distance and covariant discriminant function can be uniquely defined through such a dimension-reducing procedure. The same procedure can also be used to solve the divergence problem occurring in Eq. 15 of the ProtLock algorithm.

## RESULTS AND DISCUSSION

To show the improvement of prediction quality by introducing the pseudo-amino acid composition, tests were conducted for three different classification schemes of protein cellular attributes, i.e., 12 protein subcellular locations, 5 membrane protein types, and 9 membrane protein locations. For the data in each of these classification schemes, predictions were performed by each of the aforementioned algorithms based on the amino acid composition and the pseudo-amino acid composition.

As we can see from Eqs. 5 and 6, the first 20 components in the (20 + λ)-D space represent the contribution from the amino acid composition, and the last λ components (from 20 + 1 to 20 + λ) represent the contribution from the sequence order effect. The greater the number λ, the more the sequence order effect that is incorporated. Accordingly, with an increase in λ, the rate of correct prediction by the self-consistency test will be generally enhanced. Note that the number of λ does have an upper limit, i.e., it must be smaller than the number of amino acid residues of the shortest protein chain in the data set concerned (see Fig.1 and Eq. 2). In addition, because of the information loss during the jackknifing process, the rate of correct prediction by the jackknife test does not always monotonically increase with λ. Because jackknife tests are thought to be one of the most effective and objective methods for cross-validation in statistics,[18,22] the optimal value for λ should be the one that results in the best overall jackknife-tested rate. It should be pointed out that the optimal value of λ for one algorithm is not necessarily the same for another. However, because it has been shown that the covariant discriminant algorithm yields the best prediction quality,[8–10,23–25] for simplification the optimal value of λ determined by the covariant discriminant algorithm was used here for all the other algorithms. Actually, compared with the remarkable improvement thus obtained, the slight inaccuracy due to using the same optimal value for λ is very trivial.

### Twelve Protein Subcellular Locations

To facilitate comparison, the same data set studied by Chou and Elrod[8] was adopted here. However, because the change of code names, the sequences for some proteins could no longer be retrieved from the SWISS-PROT databank.[3] Of the 2,319 proteins originally listed in Appendix A of Chou and Elrod,[8] 2,191 protein sequences were retrieved. They form the data set $S^{12}$, which consists of 145 chloroplast proteins, 571 cytoplasm, 34 cytoskeleton, 49 endoplasmic reticulum, 224 extracellular, 25 Golgi apparatus, 37 lysosome, 84 mitochondria, 272 nucleus proteins, 27 peroxisome, 699 plasma membrane, and 24 vacuole. For the convenience of those readers who are not trained as a cellular biologist, a schematic illustration is given in Figure 2 to show the 12 different subcellular locations of proteins. For the dataset $S^{12}$, it was found that the optimal number for λ was 8. When λ = 8, the overall rates of correct prediction by different algorithms using the pseudo-amino acid composition are given in Table II. To facilitate comparison, also listed there are the results without using the pseudo-amino acid composition. Furthermore, to test the consistency, an independent data set $\bar{S}^{12}$ was constructed, which was also adopted from Chou and Elrod.[8] However, for the above reason, of the 2,591 independent proteins originally studied by Chou and Elrod,[8] only 2,494 protein sequences were retrieved. They are 112 chloroplast proteins, 761 cytoplasm, 19 cytoskeleton, 106 endoplasmic reticulum, 95 extracellular, 4 Golgi apparatus, 31 lyso-

some, 163 mitochondria, 418 nucleus proteins, 23 peroxisome, and 762 plasma membrane. None of these proteins occurs in the dataset $S^{12}$. The predicted results for the 2,494 independent proteins in $\overline{S}^{12}$ by various methods using the rule parameters derived from the 2,191 proteins in $S^{12}$ are also given in Table II.

### Five Membrane Protein Types

Cell membranes are crucial to the life of a cell. Although the basic structure of biological membranes is provided by the lipid bilayer, most of the specific functions are performed by membrane proteins. The way that a membrane-bound protein is associated with the lipid bilayer usually reflects the function of the protein.[1,2] In the literature, the definitions for the category of membrane proteins and their types are not unique. In this article, membrane proteins are categorized into the following five discriminative types: (a) type I transmembrane protein (Fig. 3a); (b) type II transmembrane protein (Fig. 3b); (c) multipass transmembrane proteins (Fig. 3c); (d) lipid-chain anchored membrane proteins (Fig. 3d); and (e) GPI-anchored membrane proteins (Fig. 3e). The peripheral membrane proteins were not included because they do not have a unique sequence feature that can be used to discriminate from nonmembrane proteins as transmembrane proteins[26] and anchored membrane proteins do.[27,28] In addition, so far the available peripheral membrane protein sequences are too few to be statistically significant. To show the impact by using the pseudo-amino acid composition on the prediction of membrane protein types, we use the same data set as used by Chou and Elrod.[9] It contains 2,059 membrane protein sequences, of which 435 are type I transmembrane proteins, 152 type II transmembrane proteins, 1,311 multipass transmembrane proteins, 51 lipid chain-anchored membrane proteins, and 110 GPI anchored membrane proteins. The names of the 2,059 membrane proteins, classified into five groups, were given in Table 1 of Chou and Elrod.[9] Here let us use $S^5$ to represent the data set of the 2,059 membrane proteins. For such a data set, it was found that the optimal number for $\lambda$ was 20. When $\lambda = 20$, the overall rates of correct prediction by different algorithms, with and without using the pseudo-amino acid composition, are given in Table III. Moreover, as a demonstration of practical application, predictions were also conducted for proteins in an independent dataset $\overline{S}^5$ based on the rule parameters derived from $S^5$. The independent data set $\overline{S}^5$ was also taken from Chou and Elrod,[9] which contains 2,625 membrane proteins, of which 478 are type I transmembrane proteins, 180 type II transmembrane



Fig. 2. Schematic illustration to show the 12 subcellular locations of proteins: chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracell, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, plasma membrane, and vacuole. Note that the vacuole and chloroplast proteins exist only in a plant. Reproduced from Chou KC, Elrod DW, Protein subcellular location prediction, Protein Eng 1999;12:107–118; by permission of Oxford University Press.

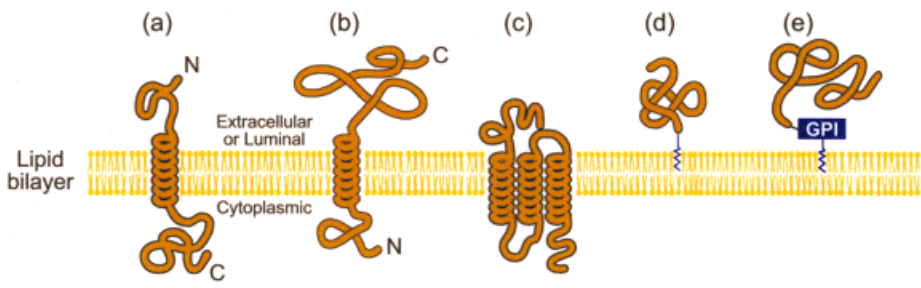


Fig. 3. Schematic drawing showing the following five types of membrane proteins: (a) type I transmembrane, (b) type II transmembrane, (c) multipass transmembrane, (d) lipid-chain anchored membrane, and (e) GPI-anchored membrane. As shown from the figure, although both type I and type II membrane proteins are of single-pass transmembrane, type I has a cytoplasmic C-terminus and an extracellular or luminal N-terminus for plasma membrane or organelle membrane, respectively, whereas the arrangement of N- and C-termini in type II membrane proteins is just reversed. No such distinction was drawn between the extracellular (or luminal) and cytoplasmic sides for the other three types in the current classification scheme. Adapted from Figures 1 and 2 by Chou KC, Elrod DW, Prediction of membrane protein types and subcellular locations. Prot Struct Funct Genet 1999;34:137–153, with permission granted by Wiley Publishers.

**TABLE II. Overall Rates of Correct Prediction for the 12 Subcellular Locations of Proteins by Different Algorithms and Test Methods**

| Algorithm | Input form | Test method: Self-consistency[a] | Jackknife[a] | Independent data set[b] |
|---|---|---|---|---|
| Least Hamming distance (Chou PY, 1980; 1989) | Amino acid composition[c] | $\frac{1067}{2191} = 48.7\%$ | $\frac{1033}{2191} = 47.2\%$ | $\frac{1151}{2494} = 46.2\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1080}{2191} = 49.3\%$ | $\frac{1047}{2191} = 47.8\%$ | $\frac{1172}{2494} = 47.0\%$ |
| Least Euclidean distance (Nakashima et al., 1986) | Amino acid composition[c] | $\frac{1096}{2191} = 50.0\%$ | $\frac{1063}{2191} = 48.5\%$ | $\frac{1197}{2494} = 48.0\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1113}{2191} = 50.8\%$ | $\frac{1076}{2191} = 49.1\%$ | $\frac{1215}{2494} = 48.7\%$ |
| ProtLock (Cedano et al., 1997) | Amino acid composition[c] | $\frac{1023}{2191} = 46.7\%$ | $\frac{971}{2191} = 44.3\%$ | $\frac{1018}{2494} = 40.8\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1128}{2191} = 51.5\%$ | $\frac{1068}{2191} = 48.7\%$ | $\frac{1165}{2494} = 46.7\%$ |
| Covariant-discriminant (Chou and Elrod, 1999) | Amino acid composition[c] | $\frac{1751}{2191} = 79.9\%$ | $\frac{1492}{2191} = 68.1\%$ | $\frac{1888}{2494} = 75.7\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1880}{2191} = \mathbf{85.8}\%$ | $\frac{1600}{2191} = \mathbf{73.0}\%$ | $\frac{2017}{2494} = \mathbf{80.9}\%$ |

[a]Conducted for the 2,191 proteins in $S^{12}$ that are classified into 12 subcellular locations as described in the text and Figure 2.
[b]Conducted based on the rule parameters derived from the 2,191 proteins in $S^{12}$ for the 2,494 independent proteins in $\overline{S}^{12}$ (see text).
[c]The conventional amino acid composition consists of 20 components each representing the occurrence frequency of one of the 20 native amino acids in a protein.
[d]The pseudo-amino acid composition consists of $20 + \lambda$ components (Eq. 5). The optimal number for $\lambda$ is 8.

proteins, 1,867 multipass transmembrane proteins, 14 lipid chain-anchored membrane proteins, and 86 GPI-anchored membrane proteins. The predicted results thus obtained are also given in Table III.

### Nine Membrane Protein Locations

Associated with different locations, membrane proteins usually have different biological functions.[1,2] On the basis of their cellular locations, membrane proteins may be classified into nine discriminative categories: (a) chloroplast, (b) endoplasmic reticulum, (c) Golgi apparatus, (d) lysosome, (e) mitochondria, (f) nucleus, (g) peroxisome, (h) plasma membrane, and (i) vacuole. Such a classification covers almost all the organelles in an animal or plant cell that have a lipid bilayer for their membrane structure (see, e.g., Refs. 1 and 2). The data set for such a classification scheme was taken from Chou and Elrod[9] and is represented here by $S^9$. It contains 2,105 protein sequences, of which 55 are chloroplast membrane proteins, 64 endoplasmic reticulum membrane proteins, 44 Golgi membrane proteins, 21 lysosome membrane proteins, 154 mitochondria membrane proteins, 26 nucleus membrane proteins, 37 peroxisome membrane proteins, 1,680 plasma membrane proteins, and 24 vacuole membrane proteins. The names of the 2,105 membrane proteins, grouped into nine categories, were given in Table 2 of Chou and Elrod.[9] For the data set $S^9$, it was found that the optimal number for $\lambda$ was 21. When $\lambda = 21$, the overall rates of correct prediction by different algorithms, with and without the pseudo-amino acid composition, are given in Table IV. Likewise, just for a demonstration, predictions were also performed for the proteins in a corresponding independent data set $\overline{S}^9$. It was also taken from Chou and Elrod[9] and contains 2,698 protein sequences, of which 293 are chloroplast membrane proteins, 79 endoplasmic reticulum membrane proteins, 35 Golgi membrane proteins, 433 mitochondria membrane proteins, 1,841 plasma membrane proteins, and 17 vacuole membrane proteins. None of these proteins occurs in the dataset $S^9$. The predicted results thus obtained are summarized in Table IV.

## DISCUSSION

From Tables II–IV, the following can be observed.

(1) For the self-consistency and jackknife tests, the overall rates of correct prediction by the covariant discriminant algorithm using the pseudo-amino acid composition are remarkably higher than the corresponding rates by the same prediction algorithm but without using the pseudo-amino acid composition, and substantially higher than the corresponding rates by the other prediction algorithms. For example, as shown in Table III, the rates obtained by the covariant discriminant algorithm using the pseudo-amino acid composition for the membrane protein-type prediction are 90.9% (self-consistency) and 80.9% (jackknife), which are about 5–10% higher than the corresponding rates with the same prediction algorithm but without using the pseudo-amino acid composition, and 13–31% higher than those with the other prediction algorithms. In

**TABLE III. Overall Rates of Correct Prediction for the Five Membrane Protein Types by Different Algorithms and Test Methods**

| Algorithm | Input form | Test method: Self-consistency[a] | Jackknife[a] | Independent data set[b] |
|---|---|---|---|---|
| Least Hamming distance (Chou PY, 1980; 1989) | Amino acid composition[c] | $\frac{1293}{2059} = 62.8\%$ | $\frac{1279}{2059} = 62.1\%$ | $\frac{1751}{2625} = 66.7\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1236}{2059} = 60.0\%$ | $\frac{1221}{2059} = 59.3\%$ | $\frac{1657}{2625} = 63.1\%$ |
| Least Euclidean distance (Nakashima et al., 1986) | Amino acid composition[c] | $\frac{1307}{2059} = 63.5\%$ | $\frac{1293}{2059} = 62.8\%$ | $\frac{1816}{2625} = 69.2\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1258}{2059} = 61.1\%$ | $\frac{1249}{2059} = 60.7\%$ | $\frac{1776}{2625} = 67.7\%$ |
| ProtLock (Cedano et al., 1997) | Amino acid composition[c] | $\frac{1372}{2059} = 66.6\%$ | $\frac{1348}{2059} = 65.5\%$ | $\frac{1674}{2625} = 63.8\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1452}{2059} = 70.5\%$ | $\frac{1401}{2059} = 68.0\%$ | $\frac{1690}{2625} = 64.4\%$ |
| Covariant-discriminant (Chou and Elrod, 1999) | Amino acid composition[c] | $\frac{1670}{2059} = 81.1\%$ | $\frac{1573}{2059} = 76.4\%$ | $\frac{2085}{2625} = 79.4\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1872}{2059} = \mathbf{90.9}\%$ | $\frac{1665}{2059} = \mathbf{80.9}\%$ | $\frac{2298}{2625} = \mathbf{87.5}\%$ |

[a]Conducted for the 2,059 membrane proteins classified into five different types as described in the text and Figure 3.
[b]Conducted based on the rule parameters derived from the 2,059 membrane proteins for the 2,625 independent membrane proteins (see text).
[c]See footnote c of Table II.
[d]The pseudo-amino acid composition consists of 20 + λ components (Eq. 5). The optimal number for λ is 20.

addition, for the 12 protein subcellular location predictions (Table II), the self-consistency and jackknife rates by the covariant discriminant algorithm using the pseudo-amino acid composition are 85.8% and 73.0%, respectively, which are about 5–6% higher than those with the same prediction algorithm but without using the pseudo-amino acid composition, and 24–39% higher than those from the other prediction algorithms. This indicates that the prediction quality can be remarkably improved after taking into account the sequence order effect by means of the pseudo-amino acid composition. Moreover, when the number of the cellular attributes concerned was reduced, the prediction quality could be further enhanced. For example, when the number of subcellular locations was reduced from 12 ($S^{12}$) to 7 ($S^7$) by excluding the small subsets (see Table 1 of Chou and Elrod[8]), the corresponding self-consistency and jackknife rates were increased to 86.6% and 77.9%; when reduced to 5, the corresponding rates increased to 88.4% and 82.7%. This indicates that the prediction quality can be further improved if one can: (a) narrow down the scope of cellular attributes 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 and the prediction be conducted among 10 possible subcellular locations instead of 12); and (b) improve the training data of small subsets by adding into them more new proteins that have been found belonging to the cellular attributes defined by these subsets.

(2) Although a combination of the self-consistency test and jackknife test, as described above, is the most appropriate way to measure the power of a statistical prediction algorithm,[22] for a demonstration of practical application, the results obtained by the independent data set tests are also given in Tables II–IV. As mentioned above, the independent data set test rates listed in Table II were the predicted results for the data set $\overline{S}^{12}$ using the rule parameters derived from $S^{12}$, those in Table III were the predicted results for the data set $\overline{S}^5$ using the rule parameters derived from $\overline{S}^5$, and those in Table IV the predicted results for the data set $\overline{S}^9$ using the rule parameters derived from $S^9$. As shown in these tables, for the independent data set test, the rates of correct prediction by the covariant discriminant algorithm using the input of pseudo-amino acid composition are also remarkably higher than the corresponding rates by the other approaches, fully consistent with the results obtained by the self-consistency and jackknife tests.

(3) A similar improvement was observed as well for the ProtLock algorithm[5] by using the pseudo-amino acid composition. However, for the least Hamming distance algorithm[16,17] and the least Euclidean distance algorithm,[13] using the pseudo-amino acid composition only slightly improved the prediction quality (e.g., for the cases of the 12 protein subcellular location predictions and the 9 membrane protein location predictions, as shown in Tables II and IV, respectively), and sometimes the prediction quality even diminished (e.g., for the case of the 5 membrane protein type prediction, as shown in Table III). This is because the simple geometry distance algorithms, such as Hamming distance and Euclidean distance, have less power in discriminating different clusters for overlapping

**TABLE IV. Overall Rates of Correct Prediction for the Nine Locations of Membrane Proteins by Different Algorithms and Test Methods**

| Algorithm | Input form | Test method: Self-consistency[a] | Jackknife[a] | Independent data set[b] |
|---|---|---|---|---|
| Least Hamming distance (Chou PY, 1980; 1989) | Amino acid composition[c] | $\frac{808}{2105} = 38.4\%$ | $\frac{791}{2105} = 37.6\%$ | $\frac{993}{2698} = 36.8\%$ |
| | Pseudo-amino acid composition[d] | $\frac{868}{2105} = 41.2\%$ | $\frac{842}{2105} = 40.0\%$ | $\frac{1044}{2698} = 38.7\%$ |
| Least Euclidean distance (Nakashima et al., 1986) | Amino acid composition[c] | $\frac{865}{2105} = 41.1\%$ | $\frac{844}{2105} = 40.1\%$ | $\frac{991}{2698} = 36.7\%$ |
| | Pseudo-amino acid composition[d] | $\frac{884}{2105} = 42.0\%$ | $\frac{855}{2105} = 40.6\%$ | $\frac{1037}{2698} = 38.4\%$ |
| ProtLock (Cedano et al., 1997) | Amino acid composition[c] | $\frac{1006}{2105} = 47.8\%$ | $\frac{970}{2105} = 46.1\%$ | $\frac{1044}{2698} = 38.7\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1194}{2105} = 56.7\%$ | $\frac{1137}{2105} = 54.0\%$ | $\frac{1257}{2698} = 46.6\%$ |
| Covariant-discriminant (Chou and Elrod, 1999) | Amino acid composition[c] | $\frac{1569}{2105} = 74.5\%$ | $\frac{1387}{2105} = 65.9\%$ | $\frac{1810}{2698} = 67.1\%$ |
| | Pseudo-amino acid composition[d] | $\frac{1910}{2105} = 90.7\%$ | $\frac{1702}{2105} = 80.9\%$ | $\frac{2121}{2698} = 78.6\%$ |

[a]Conducted for the 2,105 membrane proteins classified into nine different cellular locations as described in the text.
[b]Conducted based on the rule parameters derived from the 2,105 membrane proteins for the 2,698 independent membrane proteins (see text).
[c]See footnote c of Table II.
[d]The pseudo-amino acid composition consists of 20 + λ components (Eq. 5). The optimal number for λ is 21.

distributions, as elucidated by Chou and Zhang.[29] And for these algorithms, the introduction of additional components as was done with the pseudo-amino acid composition might not necessarily lead to improvements.

(4) A question might be raised as asking whether the improvements are really originated from incorporating the sequence order effect or due to the limited numbers (and limited variety) of structures in some of the categories. This question can be addressed as follows. (a) To make a comparison with the previous algorithms based on a completely equivalent (i.e., "apple" to "apple") condition, we should use the same database. The data sets used here were taken from the previous publications,[8,9] where the data were generated by strictly following certain screening procedures to minimize the possibility of any two similar sequences occurring in a same category. In addition, the sequence matches performed between all members in each category of proteins thus obtained have indicated that most pairs have very low sequence identity (<20%). The average sequence identity in each category is smaller than 12%. The number of pairs having high sequence identity (>90%) is very small. For example, for the data set of 12 protein subcellular locations, the percentages of pairs having >90% sequence identity in the chloroplast, cytoplasm, cytoskeleton, endoplasmic reticulum, extracellular, Golgi apparatus, lysosome, mitochondria, nucleus, peroxisome, plasma membrane, and vacuole subsets are 0.12%, 0.04%, 0.036%, 0.034%, 0.02%, 0%, 0%, 0%, 0.01%, 0.057%, 0.01%, and 1.1%, respectively. Obviously, such a small amount of high-sequence identity proteins cannot be the origin of the remarkable gains; the significant improvements shown in Tables II–IV must arise from effectively counting sequence order effect by using the pseudo-amino acid composition. Actually, exactly the same data sets were used for computing the success rates by the algorithms using the pseudo-amino acid composition and those without using it, respectively. If there is any bias because of the existence of some small amount of high similar sequences, it would affect the results obtained by both approaches, but not solely those by using the pseudo-amino acid composition. (b) To further support the above argument, let us consider the data set constructed by Reinhardt and Hubbard.[6] Their data set from prokaryotic cells consists of 688 cytoplasmic proteins, 107 extracellular proteins, and 202 periplasmic proteins. Within each category, as described by them, "the sequence identity was calculated between all pairs and sequences were kept such that none had >90% sequence identity to any other." For such a data set, the success rate by subsampling cross validation using neural networks was 81%, and that by jackknife cross validation using the covariant-discriminant algorithm was 86.5%.[7] Both results were obtained based on the 20-D conventional amino acidcomposition. However, when the sequence order effect has been taken into account via the (20 + λ)-D pseudo-amino acid composition, the corresponding success rate obtained with the covariant-discriminant algorithm was 90.0% when λ = 13, indicating a remarkable improvement. (c) In statistical predictions, a reported success rate is meaningful only when it is associated with a given data set based on which the rate was derived. Therefore, applied on different data sets, a same prediction algorithm may yield different

success rates. Unless a training data set is complete or quasi-complete, caution in using the current training data sets for practical application is dictated by the caveat that some protein cellular attributes might be mispredicted if they fall outside the "frame" defined by the current limited training data sets. (d) The goal of this study is not to determine the possible upper limit of the success rate for protein cellular attribute predictions, but to show that, for the same database, the success rate can be significantly improved after taking sequence order effect into account through the pseudo-amino acid composition. This is because it is too premature to construct a complete or quasi-complete training data set based on the protein sequences available so far. Without a complete or quasi-complete training data set, any attempt to determine such an upper limit would be unjustified, and the result thus obtained might be misleading no matter how powerful the prediction algorithm is.

## CONCLUSIONS

The existing algorithms for predicting the cellular attributes of proteins were all based on the amino acid-composition (see, e.g., Refs. 5, 6, and 8). This is because the extremely large number of sequence order patterns in proteins and their diverse lengths have made it very difficult to take into account the sequence order effect, in both the algorithm formulation and the training data construction. To tackle such a difficult problem, a set of discrete numbers was introduced to approximately reflect the sequence order effect. The pseudo-amino acid composition is a combination of such a set of sequence correlation factors and the 20 amino acid components. Thus, except for the difference in dimension, the pseudo-amino acid composition has the same mathematical framework as the conventional amino acid composition. Accordingly, all the existing prediction algorithms based on the amino acid composition can be straightforwardly extended to fit the pseudo-amino acid composition. By using the pseudo-amino acid composition, a remarkable improvement in prediction quality has been observed for both the covariant-discriminant algorithm[8] and the ProtLock algorithm.[5] The success rates with the covariant-discriminant algorithm using self-consistency and jackknife tests for the 12 protein subcellular location predictions were 85.8% and 73.0%, respectively, the corresponding rates for the 5 membrane protein type predictions 90.9% and 80.9%, and those for the 9 membrane protein location predictions 90.7% and 80.9%.

It is anticipated that the concept of pseudo-amino acid composition and its mathematical framework and biochemical implication may have a series of impacts on the other areas of proteins as well.

## ACKNOWLEDGMENTS

The author thanks Raymond B. Moeller, Cindy Brennan, Wendy Vanderheide, and Katie Crawford of Pharmacia's Graphic Services Group for their help of drawing the figures in this article, and Dr. Jinhe Li for illuminative discussions. The author also thanks the anonymous reviewer for the valuable comments in helping strengthen the presentation of this work.

## REFERENCES

1. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD. Molecular biology of the cell, 3rd ed. New York and London: Garland Publishing; 1994. Chap. 1.
2. Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J. Molecular cell biology 3rd ed. New York: Scientific American Books; 1995. Chap. 3.
3. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. Nucleic Acids Res. 1997;25:31–36.
4. Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. J Mol Biol 1994;238:54–61.
5. Cedano J, Aloy P, Perez-pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. J Mol Biol 1997;266:594–600.
6. Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. Nucleic Acids Res 1998;26: 2230–2236.
7. Chou KC, Elrod DW. Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun 1998;252:63–68.
8. Chou KC, Elrod DW. Protein subcellular location prediction. Protein Eng 1999;12:107–118.
9. Chou KC, Elrod DW. Prediction of membrane protein types and subcellular locations. Proteins Struct Funct Genet 1999;34:137–153.
10. Chou KC. Review: prediction of protein structural classes and subcellular locations. Curr Protein Peptide Sci 2000;1:171–208.
11. Tanford C. Contribution of hydrphobic interactions to the stability of the globular conformation of proteins. J Am Chem Soc 1962;84: 4240–4274.
12. Hopp TP, Woods KR. Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 1981;78:3824–3828.
13. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem 1986;99:152–162.
14. Chou KC. Using pair-coupled amino acid composition to predict protein secondary structure content. J Protein Chem 1999;18:473–480.
15. Liu WM, Chou KC. Prediction of protein secondary structure content. Protein Eng 1999;12:1041–1050.
16. Chou PY. Amino acid composition of four classes of proteins. in Abstracts of Papers, Part I, Second Chemical Congress of the North American Continent, Las Vegas, 1980.
17. Chou PY. 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; 1989, p 549–586.
18. Mardia KV, Kent JT, Bibby JM. Multivariate analysis, London: Academic Press; 1979. pp 322 and 381.
19. Mahalanobis PC. On the generalized distance in statistics. Proc Natl Inst Sci India 1936;2:49–55.
20. Pillai KCS. Mahalanobis $D^2$. In: Kotz S, Johnson NL, editors. Encyclopedia of statistical sciences, vol 5. New York: John Wiley & Sons; 1985. Pp 176–181. (This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics.)
21. Chou KC. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins Struct Funct Genet 1995;21:319–344.
22. Chou KC, Zhang CT. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
23. Chou KC, Liu W, Maggiora GM, Zhang CT. Prediction and classification of domain structural classes. Proteins Struct Funct Genet 1998;31:97–103.
24. Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem 1998;17:729–738.
25. Nakai K. Protein sorting signals and prediction of subcellular localization. Adv Protein Chem 2000;54:277–344.
26. Rost B, Casadio R, Fariselli P, Sander C. Transmembrane helices predicted at 95% accuracy. Protein Sci 1995;4:521–523.
27. Resh MD. Myristylation and palmitylation of Src family members: the fats of the matter. Cell 1994;76:411–413.
28. Casey PJ. Protein lipidation in cell signalling. Science 1995;268: 221–225.
29. Chou KC, Zhang CT. Predicting protein folding types by distance functions that make allowances for amino acid interaction. J Biol Chem 1994;269:22014–22020.