Prediction of Membrane Protein Types by Incorporating Amphipathic Effects

Kuo-Chen Chou*,[†,‡,§] and Yu-Dong Cai[†,||,⊥]

Gordon Life Science Institute, San Diego, California 92130, Tianjin Institute of Bioinformatics and Drug Discovery (TIBDD), Tianjin, China, Bioinformatics Research Centre, Donghua University, Shanghai 200050, China, Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai, China, and Biomolecular Sciences Department, University of Manchester Institute of Science & Technology, P.O. Box 88, Manchester, M60 1QD, U.K.

Received October 17, 2004

According to their intramolecular arrangement and position in a cell, membrane proteins are generally classified into the following six types: (1) type I transmembrane, (2) type II transmembrane, (3) multipass transmembrane, (4) lipid chain-anchored membrane, (5) GPI-anchored membrane, and (6) peripheral membrane. Situated in a heteropolar environment, these six types of membrane proteins must have quite different amphiphilic sequence-order patterns in order to stabilize their respective frameworks. To incorporate such a feature into the predictor, the amphiphilic pseudo amino acid composition has been formulated that contains a series of hydrophobic and hydrophilic correlation factors. The success rates thus obtained have been remarkably enhanced in identifying the types of membrane proteins, as demonstrated by the jackknife test and independent data set test, respectively.

INTRODUCTION

A cell is highly organized with many functional units or organelles. Most of these units are "enveloped" by one or more membranes, which are essential for the integrity and function of the cell. Actually, cell membranes are the structural basis for many important biological functions, such as maintaining the shape of a cell, regulating transport in and out of cell or organelle, allowing cell recognition, providing a stable site for binding and catalysis of enzymes, allowing selective receptivity and signal transduction, helping compartmentalize subcellular domains or micro-domains, regulating the fusion of the membrane with other membranes in the cell, allowing cell orientation or organelle motility, and providing a passageway across cell or subcellular domain for certain molecules. Although the basic structure of membranes is provided by the lipid bilayer, most of the specific functions are carried out by the membrane proteins (see, e.g., refs 1 and 2). To perform these functions, the membrane is specialized in that it contains specific proteins that enable it to perform its unique roles for that cell or organelle.

Membrane proteins can be reliably distinguished by using existing methods, as elaborated by many previous investigators.[3−5] The way that a membrane-bound protein is associated with the lipid bilayer usually reflects the function of the protein. The transmembrane proteins, for example, can function on both sides of membrane or transport molecules across it, whereas proteins that function on only one side of the lipid bilayer are often associated exclusively with either the lipid monolayer or a protein domain on that side. In view of this, it will certainly speed up the pace in determining the function of a newly found membrane protein if a fast and effective algorithm is available to predict its type. Particularly, the number of sequences entering into databanks has been rapidly increasing. For instance, the number of total sequence entries in SWISS-PROT[6] was only 3939 in 1986; recently, it jumped to 164 970 according to Release 45.3 (December 7, 2004) of SWISS-PROT (http://www.expasy.org/sprot/relotes/relstat.html), meaning that the number of total entries has been increased by more than 41 times! With the explosion of protein sequences in databanks, it is highly desirable to develop a fast and automated method to help determine the type of a newly found membrane protein.

In a previous study, the covariant discriminant algorithm[4] was introduced to predict the types of membrane proteins based on the amino acid composition. According to the conventional definition, the amino acid composition of a protein consists of 20 components, with each equal to the occurrence frequency of one of the 20 native amino acids in the protein.[7−9] Obviously, if using the conventional amino acid composition as the representation for a protein sample, all the sequence-order effects would be missed. For example, α-helix is the most important element in proteins that has the highest percentage of occupancy in comparison with all the other elements. Many helices are amphipathic (or amphiphilic), while others are not, depending on the environment where they are located. This kind of difference is important in stabilizing the structure and is reflected by the different sequence-order pattern in a helix.[10,11] Actually, the amphipathic character is correlated with the sequence of an entire protein as well. In view of this, it is anticipated that the prediction quality will certainly be significantly enhanced if this kind of sequence-order patterns can be effectively

* Corresponding author phone/fax: +858 455-0954; e-mail: kchou@san.rr.com.
[†] Gordon Life Science Institute.
[‡] TIBDD.
[§] Donghua University.
[||] Chinese Academy of Sciences.
[⊥] University of Manchester Institute of Science & Technology.

10.1021/ci049686v CCC: $30.25 © 2005 American Chemical Society
Published on Web 02/12/2005

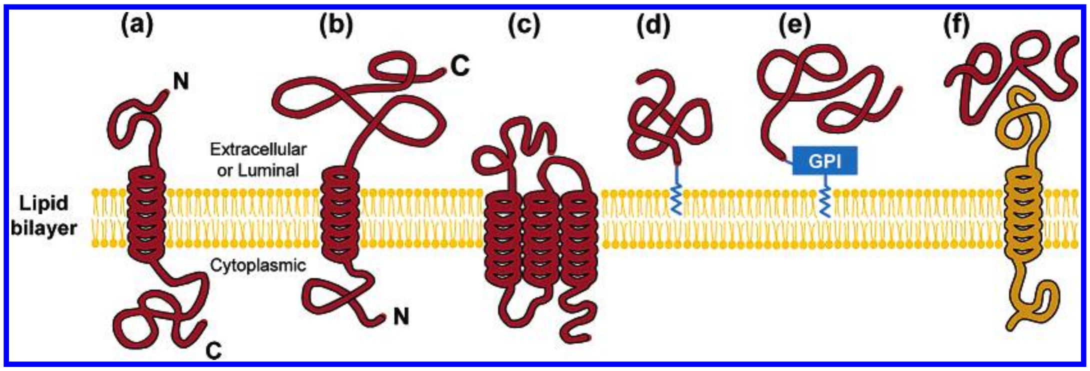

Figure 1. Schematic illustration to show the six types of membrane proteins: (1) type I transmembrane, (b) type II transmembrane, (c) multipass transmembrane, (d) lipid chain-anchored membrane, (e) GPI-anchored membrane, and (f) peripheral membrane. As shown in the figure, both type I and type II are of single-pass transmembrane. However, type I has a cytoplasmic C-terminus and an extracellular or luminal N-terminus for plasma membrane or organelle membrane, respectively, while the arrangement of N- and C-termini in type II is just the reverse.

incorporated into the predictor. Unfortunately, there is no way to reflect such features by the conventional amino acid composition. Thus, how to effectively incorporate the amphipathic effects into the predictor has become a critical challenge in order to improve the prediction quality. Besides, with more data available in databanks, the classification scheme should be gradually extended to cover more membrane protein types, as will be illustrated below.

SIX TYPES OF MEMBRANE PROTEINS

In this study, membrane proteins are classified into the following six types: (1) type I transmembrane, (2) type II transmembrane, (3) multipass transmembrane, (4) lipid chain-anchored membrane, (5) GPI-anchored membrane, and (6) peripheral membrane (Figure 1), where the last type was excluded in the previous studies[4,12] because it lacked the statistical significance data. The procedures to generate the training data set are basically the same as elaborated in a previous paper.[4] The data set was generated based on release 39.0 of SWISS-PROT.[6] To obtain a high-quality and well defined working data set, the data were screened strictly according to the following procedures. (1) Included were only those sequences with a clear description to explicitly indicate one of the above six types. (2) For protein sequences having the same name but from different species, only one was included to reduce the redundancy. (3) Sequences annotated by two or more types were not included because of lacking uniqueness. After the above screening procedures, we obtained a training data set of 2628 proteins, of which 372 are of type I transmembrane proteins, 151 of type II transmembrane proteins, 1903 of multipass transmembrane proteins, 104 of lipid chain-anchored membrane proteins, 68 of GPI-anchored membrane proteins, and 30 of peripheral membrane proteins. The accession numbers of the 2628 proteins in the training data set are given in Supporting Information A. Rather than the SWISS-PROT code as used in the previous paper,[4] the accession number is used in the Supporting Information because it is more stable for representing a unique protein sequence. It is instructive to conduct an analysis of the sequence identity for the proteins in each of the six membrane protein subsets. The sequence identity percentage between two protein sequences is defined as follows. Suppose one sequence is N_1 residues long and the other N_2 residues long ($N_1 \geq N_2$), and the maximum number of residues matched by sliding one sequence along the other is M. The sequence identity percentage between the two sequences is defined as $(M/N_1) \times 100\%$. The treatment for gaps is according to ref 13. The sequence matches performed between all members in each of the six subsets have indicated that the average sequence identity percentages for type I transmembrane protein subset, type II transmembrane protein subset, multipass transmembrane protein subset, lipid chain-anchored membrane protein subset, GPI-anchored membrane protein subset, and peripheral membrane protein subset are 7.97%, 7.94%, 8.31%, 7.94%, 7.92%, and 11.36%, respectively (cf. Supporting Information A). These numbers indicate that the majority of pairs in each of these subsets have very low sequence identity.

Likewise, an independent testing data set was also constructed. It contains 3160 proteins, of which 462 are of type I transmembrane proteins, 144 of type II transmembrane proteins, 2402 of multipass transmembrane proteins, 67 of lipid chain-anchored membrane proteins, 83 of GPI-anchored membrane proteins, and 2 of peripheral membrane proteins. The accession numbers of the 3160 proteins in the testing data set are given in Supporting Information B. None of the proteins in the testing data set occurs in the training data set. The corresponding average sequence identity percentages are 8.34%, 9.53%, 8.55%, 10.22%, 11.75, and 5.00%, respectively (cf. Supporting Information B), indicating that the sequence identity for majority of pairs in each of the six subsets in the independent data set is also very low.

AMPHIPATHIC EFFECTS AND PSEUDO AMINO ACID COMPOSITION

Membrane proteins are located in a heteropolar environment. Different types of membrane proteins might have different amphipathic sequence patterns that would make them to optimally match their respective microenvironments. Accordingly, a logical procedure to enhance the success rate in discriminating membrane protein types is to incorporate the amphipathic feature into the predictor. This can be realized in terms of the amphipathic pseudo amino acid composition,[14] as formulated below.

Suppose a protein **P** with a sequence of L amino acid residues:

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

where R_1 represents the residue at sequence position 1, R_2

represents the residue at position 2, and so forth. Since the amphipathic feature of a protein is mainly reflected by the hydrophobicity and hydrophilicity of its constituent amino acids, their indexes will be used to formulate the sequence-order correlated factors (Figure 2) through the following equations:

$$\begin{cases} \tau_1 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1} H^1_{i,i+1} \\ \tau_2 = \dfrac{1}{L-1}\displaystyle\sum_{i=1}^{L-1} H^2_{i,i+1} \\ \tau_3 = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2} H^1_{i,i+2} \\ \tau_4 = \dfrac{1}{L-2}\displaystyle\sum_{i=1}^{L-2} H^2_{i,i+2} \qquad , (\lambda < L) \\ \cdots\cdots\cdots\cdots \\ \tau_{2\lambda-1} = \dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda} H^1_{i,i+\lambda} \\ \tau_{2\lambda} = \dfrac{1}{L-\lambda}\displaystyle\sum_{i=1}^{L-\lambda} H^2_{i,i+\lambda} \end{cases} \tag{2}$$

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

$$\begin{cases} H^1_{i,j} = H^1(\mathrm{R}_i)\cdot H^1(\mathrm{R}_j) \\ H^2_{i,j} = H^2(\mathrm{R}_i)\cdot H^2(\mathrm{R}_j) \end{cases} \tag{3}$$

where $H^1(\mathrm{R}_i)$ and $H^2(\mathrm{R}_i)$ are respectively the hydrophobicity and hydrophilicity values for the ith ($i = 1, 2, ..., L$) amino acid in eq 1, and the dot (·) means the multiplication sign. Note that before substituting the values of hydrophobicity and hydrophilicity into eq 3, they were all subjected to a standard conversion as described by the following equation:

$$\begin{cases} H^1(\mathrm{R}_i) = \dfrac{H_0{}^1(\mathrm{R}_i) - \displaystyle\sum_{k=1}^{20}\frac{H^1_0(\mathscr{R}_k)}{20}}{\sqrt{\dfrac{\displaystyle\sum_{u=1}^{20}\left[H^1_0(\mathscr{R}_u) - \sum_{k=1}^{20}\frac{H^1_0(\mathscr{R}_k)}{20}\right]^2}{20}}} \\ H^2(\mathrm{R}_i) = \dfrac{H^2_0(\mathrm{R}_i) - \displaystyle\sum_{k=1}^{20}\frac{H^2_0(\mathscr{R}_k)}{20}}{\sqrt{\dfrac{\displaystyle\sum_{u=1}^{20}\left[H^2_0(\mathscr{R}_u) - \sum_{k=1}^{20}\frac{H^2_0(\mathscr{R}_k)}{20}\right]^2}{20}}} \end{cases} \tag{4}$$

where we use $\mathscr{R}_i$ ($i = 1, 2, ..., 20$) 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 symbols H^1_0 and H^2_0 represent the original hydrophobicity and hydrophilicity values for the amino acid in the follow-up brackets, and their values are taken from refs 31 and 15, respectively. The advantage to use the converted hydrophobicity and hydrophilicity values obtained via eq 4 is that they will have a zero mean value over the 20 native amino acids and will remain unchanged if going through the same conversion procedure again.

After incorporating the sequence-order correlated factors from eq 2 into the classical 20D (dimensional) amino acid composition, we obtain a pseudo amino acid composition with ($20 + 2\lambda$) components. In other words, the representation for a protein sample **P** is now formulated as

$$\mathbf{P} = \begin{bmatrix} p_1 \\ \vdots \\ p_{20} \\ p_{20+1} \\ \vdots \\ p_{20+\lambda} \\ p_{20+\lambda+1} \\ \vdots \\ p_{20+2\lambda} \end{bmatrix} \tag{5}$$

where

$$p_u = \begin{cases} \dfrac{f_u}{\displaystyle\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda}\tau_j}, & (1 \le u \le 20) \\ \dfrac{w\tau_u}{\displaystyle\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda}\tau_j}, & (20 + 1 \le u \le 20 + 2\lambda) \end{cases} \tag{6}$$

where f_i ($i = 1, 2, ..., 20$) are the normalized occurrence frequencies of the 20 amino acids in the protein **P**, τ_j is the j-tier sequence-correlation factor computed according to eq 2, and w is the weight factor. In the current study, we chose $w = 0.5$ to make the data within the range easier to be handled (w can be of course assigned with other values, but this would not have a big impact to the final results). As we can see from eqs 5−6, the first 20 components reflect the effect of the classical amino acid composition, while the components from $20 + 1$ to $20 + 2\lambda$ reflect the amphipathic sequence-order pattern. A set of such $20 + 2\lambda$ components is called the “amphipathic pseudo amino acid composition”.

PREDICTION ALGORITHMS

As we can see from eq 5, 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 can be applied to the amphipathic pseudo amino acid composition by a straightforward augmentation procedure as formulated below.

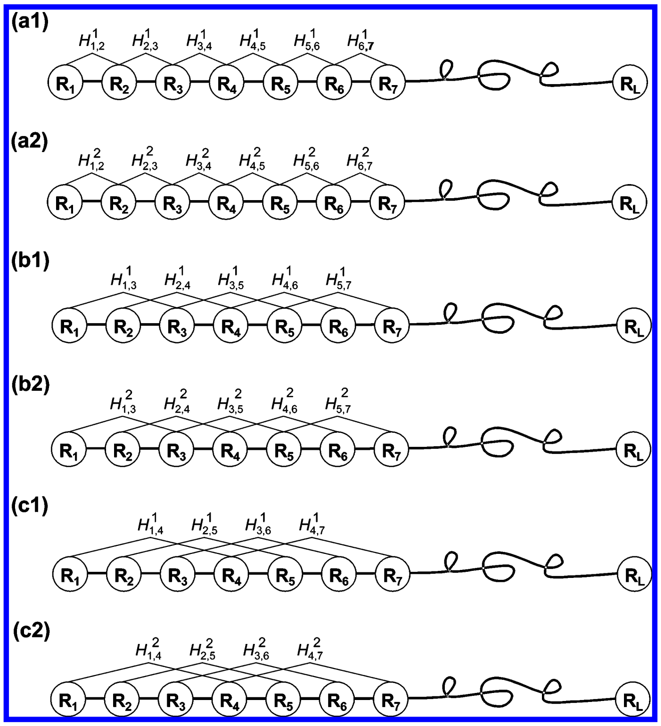
(a1)
$H^1_{1,2}$ $H^1_{2,3}$ $H^1_{3,4}$ $H^1_{4,5}$ $H^1_{5,6}$ $H^1_{6,7}$
(a2)
$H^2_{1,2}$ $H^2_{2,3}$ $H^2_{3,4}$ $H^2_{4,5}$ $H^2_{5,6}$ $H^2_{6,7}$
(b1)
$H^1_{1,3}$ $H^1_{2,4}$ $H^1_{3,5}$ $H^1_{4,6}$ $H^1_{5,7}$
(b2)
$H^2_{1,3}$ $H^2_{2,4}$ $H^2_{3,5}$ $H^2_{4,6}$ $H^2_{5,7}$
(c1)
$H^1_{1,4}$ $H^1_{2,5}$ $H^1_{3,6}$ $H^1_{4,7}$
(c2)
$H^2_{1,4}$ $H^2_{2,5}$ $H^2_{3,6}$ $H^2_{4,7}$
R_1 R_2 R_3 R_4 R_5 R_6 R_7 R_L

Figure 2. Schematic drawing to show (a1 and a2) the 1st rank, (b1 and b2) the 2nd rank, and (c1 and c2) the 3rd rank sequence-order coupling mode along a protein sequence through a hydrophobicity/hydrophilicity correlation function, where $H^1_{i,j}$ and $H^2_{i,j}$ are given by eq 3. Panels a1 and a2 reflect the coupling mode between all the most contiguous residues, panels b1 and b2 reflect that between all the 2nd most contiguous residues, and panels c1 and c2 reflect that between all the 3rd most contiguous residues.

Suppose there are N proteins forming a set S, which is the union of 6 subsets; i.e.Each subset is composed of

$$S = S_1 \cup S_2 \cup S_3 \cup S_4 \cup S_5 \cup S_6 \tag{7}$$

proteins with the same type. The number of proteins in subset S_1 is n_1, the number of proteins in subset S_2 is n_2, and so forth. Obviously, we have $N = n_1 + n_2 + n_3 + n_4 + n_5 + n_6$. The key to the incorporation of the amphipathic 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 $(20 + 2\lambda)$D space, rather than a 20D space as used by the previous investigators.[4,7−9,16] According to such a baseline, the kth protein in the subset S_m should now be represented by

$$\mathbf{P}_k^m = \begin{bmatrix} p_{k,1}^m \\ \vdots \\ p_{k,20}^m \\ p_{k,20+1}^m \\ \vdots \\ p_{k,20+\lambda}^m \\ p_{k,20+\lambda+1}^m \\ \vdots \\ p_{k,20+2\lambda}^m \end{bmatrix} \quad (k = 1, 2, ..., n_m;\ m = 1, 2, ..., 6) \tag{8}$$

where $p_{k,u}^m$ $(u = 1, 2, ..., 20 + 2\lambda)$ has the same meaning as p_u of eq 5 but is associated with $\mathbf{P}_k^m$ instead of $\mathbf{P}$. The standard vector for the subset S_m is defined by

$$\mathbf{P}_k^m = \begin{bmatrix} p_{k,1}^m \\ \vdots \\ p_{k,20}^m \\ p_{k,20+1}^m \\ \vdots \\ p_{k,20+\lambda}^m \\ p_{k,20+\lambda+1}^m \\ \vdots \\ p_{k,20+2\lambda}^m \end{bmatrix} \quad (k = 1, 2, ..., n_m;\ m = 1, 2, ..., 6) \tag{8}$$

where

$$\bar{p}_i^m = \frac{1}{n_m}\sum_{k=1}^{n_m} p_{k,i}^m,\ (i = 1, 2, ..., 20 + 2\lambda) \tag{10}$$

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

Suppose that $\mathbf{P}$ is a query protein whose type is to be predicted. It can be either one of the N proteins in the set S or a protein outside of it. Its amphipathic pseudo amino acid composition has been given by eq 5. Now the problem is how to effectively define the similarity between the query protein $\mathbf{P}$ and the standard vector $\bar{\mathbf{P}}^m$. Algorithms with various criteria were proposed, as can be briefly described as follows.

Least Hamming Distance Algorithm.[8] The algorithm was originally proposed by Chou in the Second Chemical Congress of the North American Continent, Las Vegas (1980) for predicting protein structural class based on the 20D amino acid composition. The hypothesis was that the similarity of any two proteins could be measured by their Hamming distance or city-block metric.[17] The smaller the distance between the two proteins, the higher their similarity is. Now based on the amphipathic pseudo amino acid composition, instead of 20 components, the Hamming distance between $\mathbf{P}$ and $\bar{\mathbf{P}}^m$ should involve $20 + 2\lambda$ components; i.e.

$$D_{\mathrm{H}}(\mathbf{P}, \bar{\mathbf{P}}^m) = \sum_{i=1}^{20+2\lambda} |p_i - \bar{p}_i^m| \quad (m = 1, 2, 3, ..., 6) \tag{11}$$

Thus, the prediction rule was given by

$$D_{\mathrm{H}}(\mathbf{P}, \bar{\mathbf{P}}^\mu) = \mathbf{Min}\{D_{\mathrm{H}}(\mathbf{P}, \bar{\mathbf{P}}^1), D_{\mathrm{H}}(\mathbf{P}, \bar{\mathbf{P}}^2), ..., D_{\mathrm{H}}(\mathbf{P}, \bar{\mathbf{P}}^6)\} \tag{12}$$

where μ can be 1, 2, 3, ..., or 6, and the operator **Min** means taking the least one among those in the brackets, and the superscript μ is the type predicted for the query membrane protein $\mathbf{P}$. If there is a tie, μ is not uniquely determined, but that rarely occurs.

Least Euclidean Distance Algorithm.[7] Rather than Hamming distance, Nakashima et al. [7] 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 $\mathbf{P}$ and $\bar{\mathbf{P}}^m$ should be defined by

$$D^2_{\mathrm{E}}(\mathbf{P}, \bar{\mathbf{P}}^m) = \sum_{i=1}^{20+2\lambda} (p_i - \bar{p}_i^m)^2 \quad (m = 1, 2, 3, ..., 6) \quad (13)$$

and the prediction rule given by

$$D^2_{\mathrm{E}}(\mathbf{P}, \bar{\mathbf{P}}^\mu) = \mathbf{Min}\{D^2_{\mathrm{E}}(\mathbf{P}, \bar{\mathbf{P}}^1), D^2_{\mathrm{E}}(\mathbf{P}, \bar{\mathbf{P}}^2), ..., D^2_{\mathrm{E}}(\mathbf{P}, \bar{\mathbf{P}}^6)\} \quad (14)$$

ProtLock Algorithm.[16] Likewise, the scale to measure the similarity between proteins **P** and $\bar{\mathbf{P}}^m$ in the ProtLock algorithm[16] should be modified as

$$D^2_{\mathrm{P}}(\mathbf{P}, \bar{\mathbf{P}}^m) = (\mathbf{P} - \bar{\mathbf{P}}^m)^{\mathbf{T}}\mathbf{C}^{-1}(\mathbf{P} - \bar{\mathbf{P}}^m) \quad (m = 1, 2, 3, ..., 6) \quad (15)$$

where **C** is a matrix defined by

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

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

$$c_{i,j} = \sum_{m=1}^{6}\sum_{k=1}^{n_m}[p^m_{k,i} - \bar{p}_i][p^m_{k,j} - \bar{p}_j] \quad (i, j = 1, 2, ..., 20 + 2\lambda) \quad (17)$$

where

$$\bar{p}_i = \frac{1}{N}\sum_{m=1}^{6}\sum_{k=1}^{n_m} p^m_{k,i} = \frac{1}{N}\sum_{m=1}^{6} n_m\, \bar{p}_i^m \quad (i = 1, 2, ..., 20 + 2\lambda) \quad (18)$$

And the prediction rule should be given by

$$D^2_{\mathrm{P}}(\mathbf{P}, \bar{\mathbf{P}}^\mu) = \mathbf{Min}\{D^2_{\mathrm{P}}(\mathbf{P}, \bar{\mathbf{P}}^1), D^2_{\mathrm{P}}(\mathbf{P}, \bar{\mathbf{P}}^2), ..., D^2_{\mathrm{P}}(\mathbf{P}, \bar{\mathbf{P}}^6)\} \quad (19)$$

Covariant Discriminant Algorithm.[4,18] In the covariant discriminant algorithm, rather than the geometrical distances as mentioned above, a function was used as a scale to measure the similarity between proteins **P** and $\bar{\mathbf{P}}^m$. The smaller the value of the function, the higher the similarity between the two proteins is. Based on the amphipathic pseudo amino acid composition, the function, which was called the "covariant discriminant function", should be expressed as

$$F(\mathbf{P}, \bar{\mathbf{P}}^m) = D^2_{\mathrm{M}}(\mathbf{P}, \bar{\mathbf{P}}^m) + \ln|\mathbf{C}_m| \quad (m = 1, 2, 3, ..., 6) \quad (20)$$

where

$$D^2_{\mathrm{M}}(\mathbf{P}, \bar{\mathbf{P}}^m) = (\mathbf{P} - \bar{\mathbf{P}}^m)^{\mathbf{T}}\mathbf{C}_m^{-1}(\mathbf{P} - \bar{\mathbf{P}}^m) \quad (21)$$

is the squared Mahalanobis distance[19−22] between **P** and $\bar{\mathbf{P}}^m$, and

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

is the covariance matrix for the subset S_m and its elements in given by

$$c^m_{i,j} = \frac{1}{n_\xi - 1}\sum_{k=1}^{n_m}[p^{\mathrm{m}}_{k,i} - \bar{p}^m_i][p^{\mathrm{m}}_{k,j} - \bar{\mathrm{p}}^m_j] \quad (i, j = 1, 2, ..., 20 + 2\lambda) \quad (23)$$

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

$$F(\mathbf{P}, \bar{\mathbf{P}}^\mu) = \mathbf{Min}\{F(\mathbf{P}, \bar{\mathbf{P}}^1), F(\mathbf{P}, \bar{\mathbf{P}}^2), F(\mathbf{P}, \bar{\mathbf{P}}^3), ..., F(\mathbf{P}, \bar{\mathbf{P}}^6)\} \quad (24)$$

where **Min** and μ, which also occur in eqs 14 and 19, have already been explicitly defined in eq 12.

Although the number of subsets formulated above is set at $m = 6$ (corresponding to six different types of membrane proteins), the formulation is valid for any number of subsets. For example, when a case studied involves a classification of 12 categories (subsets), all the user has to do is just to replace 6 with 12 in the above equations. It is instructive to point out, however, that the sum of the $20 + 2\lambda$ components in eq 6 is equal to 1 (imposed by the normalization condition); i.e., of the $20 + 2\lambda$ components, only $20 + 2\lambda - 1$ are independent. Accordingly, the covariance matrix $\mathbf{C}_m$ 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 meaningless. To overcome such a difficulty, we can adopt the following dimension-reducing procedure.[22] Instead of defining a protein in a $(20 + 2\lambda)$D space, let us define it in a $(20 + 2\lambda - 1)$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 $\mathbf{C}_m$ would no longer be singular. In such a $(20 + 2\lambda - 1)$D space, the Mahalanobis distance (eq 21) as well as the covariant discriminant function (eq 20) can exist without the divergence difficulty at all. Furthermore, according to the invariance theorem given in the Appendix A of Chou,[22] the values of the Mahalanobis distance and the covariant discriminant function will remain the same regardless of which one of the $20 + 2\lambda$ components is left out. Accordingly, the values of both the Mahalanobis distance and the 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.[16]

RESULTS AND DISCUSSION

The newly constructed data sets as given in the Supporting Information A and B will serve as the training and independent testing data sets, respectively. Both consist of six subsets corresponding to six membrane protein types (Figure 1).

Table 1. Overall Rates of Correct Prediction for the Six Types of Membrane Proteins (Figure 1) by Different Predictors and Test Methods

algorithm	input form	test method: jackknife[a]	test method: independent data set[b]
least Hamming distance[8]	amino acid composition[c]	1703/2628 = 64.80%	2025/3160 = 64.08%
	amphipathic pseudo amino acid composition[d]	1947/2628 = 74.09%	2350/3160 = 74.37%
least Euclidean distance[7]	amino acid composition[c]	1795/2628 = 68.30%	2113/3160 = 66.87%
	amphipathic pseudo amino acid composition[d]	1960/2628 = 74.58%	2378/3160 = 75.25%
ProtLock[16]	amino acid composition[c]	1659/2628 = 63.13%	2156/3160 = 68.23%
	amphipathic pseudo amino acid composition[d]	2040/2628 = 77.63%	2620/3160 = 82.91%
covariant−discriminant[4]	amino acid composition[c]	1949/2628 = 74.16%	2572/3160 = 81.39%
	amphipathic pseudo amino acid composition[d]	2264/2628 = **86.15**%	2863/3160 = **90.60**%

[a] Conducted for the 2628 membrane proteins classified into 6 different types (Figure 1) in the training data set as given in the Supporting Information A. [b] Conducted based on the rule parameters derived from the 2628 proteins in the training data set for the 3160 proteins in the independent data set as given in the Supporting Information B. [c] The dimension of the conventional amino acid composition is 20, where no sequence-order effects whatsoever are incorporated. [d] The dimension of the amphipathic pseudo amino acid composition for the current study is 60, where the sequence-order effects are incorporated through $\lambda = 20$ hydrophobic correlation factors and $\lambda = 20$ hydrophilic correlation factors (see eq 5), with weight factor $w = 0.5$ (see eq 6).

Since the sequence-order effects are incorporated through the amphipathic pseudo amino acid components (eq 5), a question is naturally raised: how many pseudo amino acid components should be used, or what numbers should be assigned for λ during the prediction? Actually, the value of λ was determined through an optimal process by maximizing the success rate from the jackknife test. This is because among the independent data set test, subsampling test, and jackknife test that are 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[23] for a comprehensive discussion about this and Mardia et al.[17] for the underlying mathematical principle. Therefore, the jackknife test has been used by more and more investigators[24−30] in examining the power of various prediction methods. The optimal value thus obtained for the current training data set was $\lambda = 20$, meaning that the amphipathic pseudo amino acid composition contains $\lambda = 20$ hydrophobic correlation factors and $\lambda = 20$ hydrophilic correlation factors and that any protein in this study should be represented by a $(20 + 40)$D = 60D vector (see eq 5 and Figure 2).

Jackknife Test. As mentioned above, jackknife test is the key for examining a prediction method.[17,23] During jackknifing, each protein in the data set 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 type of each protein is identified by the rule parameters derived using all the other proteins except the one that 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 rates thus obtained for the 2628 proteins in the Supporting Information A by different approaches are given in Table 1.

Independent Data Set Test. Moreover, as a demonstration of practical application, predictions with different approaches were also conducted for the 3160 proteins in the independent data set (Supporting Information B) based on the rule parameters derived from the 2628 proteins in the training data set (Supporting Information A). The corresponding results are also given in Table 1.

From Table 1 we can see the following. (1) Compared with the results obtained by using the conventional amino acid composition to represent the sample of a protein, the success rates obtained by using the amphipathic pseudo amino acid composition are remarkably enhanced for all the four different operation algorithms; i.e., the least Hamming distance algorithm,[8] the least Euclidean distance algorithm,[7] the ProtLock algorithm,[16] and the covariant discriminant algorithm.[4] This is fully consistent with what is expected because some important sequence-order effects have been incorporated into each of the four algorithms through the amphipathic pseudo amino acid components (eq 5) that are particularly effective in dealing membrane proteins. **(2)** Among the four algorithms, the covariant discriminant algorithm yields the highest success rates by using the amphipathic pseudo amino acid composition: the overall success rate by the jackknife test is 86.15%, and that by the independent data set test is 90.60%.

CONCLUSION

Dwelling in a heteropolar environment, different types of membrane proteins will have different amphipathic sequence-order patterns. By incorporating such a feature into a predictor, the success rate in identifying the type of a membrane protein can be significantly enhanced. An effective approach to realize this is through the amphipathic pseudo amino acid composition, which contains a series of hydrophobic and hydrophilic correlation factors, as formulated in eqs 2−5 and Figure 2.

ACKNOWLEDGMENT

We thank the four anonymous reviewers whose constructive comments are very helpful for strengthening the presentation of this paper.

Supporting Information Available: Training and independent testing data sets. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Lodish, H.; Baltimore, D.; Berk, A.; Zipursky, S. L.; Matsudaira, P.; Darnell, J. *Molecular Cell Biology*, 3rd ed.; Scientific American Books: New York, 1995; Chapter 3.

(2) Alberts, B.; Bray, D.; Lewis, J.; Raff, M.; Roberts, K.; Watson, J. D. *Molecular Biology of the Cell*, 3rd ed.; Garland Publishing: New York & London, 1994; Chapter 1.

(3) Rost, B.; Casadio, R.; Fariselli, P.; Sander, C. Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **1995**, *4*, 521−533.

(4) Chou, K. C.; Elrod, D. W. Prediction of membrane protein types and subcellular locations. *Proteins: Struct., Funct., Genet.* **1999**, *34*, 137−153.

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

(6) Bairoch, A.; Apweiler, R. The SWISS-PROT protein sequence data bank and its supplement TrEMBL. *Nucleic Acids Res.* **2000**, *25*, 31−36.

(7) 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.

(8) Chou, P. Y. In *Prediction of Protein Structure and the Principles of Protein Conformation*; Fasman, G. D., Ed.; Plenum Press: New York, 1989; pp 549−586.

(9) Chou, J. J.; Zhang, C. T. A joint prediction of the folding types of 1490 human proteins from their genetic codons. *J. Theor. Biol* **1993**, *161*, 251−262.

(10) Chou, K. C.; Zhang, C. T.; Maggiora, G. M. Disposition of amphiphilic helices in heteropolar environments. *Proteins: Struct., Funct., Genet.* **1997**, *28*, 99−108.

(11) Chou, J. J.; Li, S.; Bax, A. Study of conformational rearrangement and refinement of structural homology models by use of heteronuclear dipolar couplings. *J. Biomol. NMR* **2000**, *18*, 217−227.

(12) Cai, Y. D.; Zhou, G. P.; Chou, K. C. Support vector machines for predicting membrane protein types by using functional domain composition. *Biophys. J.* **2003**, *84*, 3257−3263.

(13) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673−4680.

(14) Chou, K. C. Prediction of protein cellular attributes using pseudo-amino-acid-composition. *Proteins: Struct., Funct., Genet.* **2001**, *43*, 246−255; Erratum: **2001**, *44*, 60).

(15) Hopp, T. P.; Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. U.S.A.* **1981**, *78*, 3824−3828.

(16) Cedano, J.; Aloy, P.; P'erez-Pons, J. A.; Querol, E. Relation between amino acid composition and cellular location of proteins. *J. Mol. Biol* **1997**, *266*, 594−600.

(17) Mardia, K. V.; Kent, J. T.; Bibby, J. M. *Multivariate Analysis*; Academic Press: London, 1979; Chapters 11−13, pp 322−381.

(18) Chou, K. C.; Elrod, D. W. Protein subcellular location prediction. *Protein Eng.* **1999**, *12*, 107−118.

(19) Mahalanobis, P. C. On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India* **1936**, *2*, 49−55.

(20) Pillai, K. C. S. In *Encyclopedia of Statistical Sciences*; Kotz, S., Johnson, N. L., Eds.; John Wiley & Sons: New York, 1985; Vol. 5, pp 176−181.

(21) Chou, K. C.; Zhang, C. T. Predicting protein folding types by distance functions that make allowances for amino acid interactions. *J. Biol. Chem.* **1994**, *269*, 22014−22020.

(22) Chou, K. C. 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.

(23) Chou, K. C.; Zhang, C. T. Review: Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **1995**, *30*, 275−349.

(24) Zhou, G. P. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **1998**, *17*, 729−738.

(25) Yuan, Z. Prediction of protein subcellular locations using Markov chain models. *FEBS Lett.* **1999**, *451*, 23−26.

(26) Feng, Z. P. Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. *Biopolymers* **2001**, *58*, 491−499.

(27) Hua, S.; Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **2001**, *17*, 721−728.

(28) Zhou, G. P.; Assa-Munt, N. Some insights into protein structural class prediction. *Proteins: Struct., Funct., Genet.* **2001**, *44*, 57−59.

(29) Pan, Y. X.; Zhang, Z. Z.; Guo, Z. M.; Feng, G. Y.; Huang, Z. D.; He, L. Application of pseudo amino acid composition for predicting protein subcellular location: stochastic signal processing approach. *J. Protein Chem.* **2003**, *22*, 395−402.

(30) Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins. *Proteins: Struct., Funct., Genet.* **2003**, *50*, 44−48.

(31) Tanford, C. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *J. Am. Chem. Soc.* **1962**, *84*, 4240−4274.

CI049686V