Amino Acids (2007) 32: 483–488
DOI 10.1007/s00726-006-0439-2
Printed in The Netherlands

Amino Acids

Using ensemble classifier to identify membrane protein types

H.-B. Shen[1] and **K.-C. Chou**[1,2]

[1] Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China
[2] Gordon Life Science Institute, San Diego, CA, U.S.A.

Received June 13, 2006
Accepted June 16, 2006
Published online October 12, 2006; © Springer-Verlag 2006

Summary. Predicting membrane protein type is both an important and challenging topic in current molecular and cellular biology. This is because knowledge of membrane protein type often provides useful clues for determining, or sheds light upon, the function of an uncharacterized membrane protein. With the explosion of newly-found protein sequences in the post-genomic era, it is in a great demand to develop a computational method for fast and reliably identifying the types of membrane proteins according to their primary sequences. In this paper, a novel classifier, the so-called "ensemble classifier", was introduced. It is formed by fusing a set of nearest neighbor (NN) classifiers, each of which is defined in a different pseudo amino acid composition space. The type for a query protein is determined by the outcome of voting among these constituent individual classifiers. It was demonstrated through the self-consistency test, jackknife test, and independent dataset test that the ensemble classifier outperformed other existing classifiers widely used in biological literatures. It is anticipated that the idea of ensemble classifier can also be used to improve the prediction quality in classifying other attributes of proteins according to their sequences.

Keywords: Type-I – Type-II – Multi-pass transmembrane – Lipid-chain-anchored – GPI-anchored – Pseudo-amino acid composition – Ensemble classifier – Fusion – Voting

Introduction

The recent success of human genome project led to a protein sequence explosion. For instance, in 1986 the SWISS-PROT databank (Bairoch and Apweiler, 2000) contained only 3939 protein sequence entries, but now it contains 222,289 entries according to version 50.0 released as of May 30, 2006, meaning that the number of sequences has increased by about 56 times in 20 years. The avalanche of protein sequences generated in the post-genomic era challenges the speed and ability to timely characterize and annotate newly-found proteins. Membrane proteins are a very important part of proteins that can be distinguished from non-membrane proteins by using existing methods, as elaborated by previous investigators (see, e.g., Chou and Elrod, 1999; Rost et al., 1995). Membrane proteins are generally classified into the following 5 types: (1) Type I, (2) Type II, (3) multipass transmembrane, (4) lipid-chain-anchored, and (5) GPI-anchored (Fig. 1). Because the type of a membrane protein is closely correlated with its function (Alberts et al., 1994; Chou and Elrod, 1999; Lodish et al., 1995), it is very useful to develop an automated method for fast and accurately identifying the type of a given membrane protein. In a pioneer study, Chou and Elrod (1999) developed the covariant discriminant algorithm, which is a combination of the "Mahalanobis distance" (Chou and Zhang, 1994; Mahalanobis, 1936; Pillai, 1985) and the invariance principle for treating a degenerate vector space (Chou, 1995) that is cited in literatures as "Chou's invariance theorem" (see, e.g., Pan et al., 2003; Zhou and Doctor, 2003), to predict the type for a given membrane protein. Subsequently, varieties of prediction algorithms were proposed (Cai et al., 2003; Chou, 2001; Feng, 2001, 2002; Pan et al., 2003). Most of the existing prediction methods fall into two categories: one is based on the conventional AA (amino acid) composition (Chou and Zhang, 1993; Chou, 1989), and the other based on the PseAA (pseudo-amino acid) composition (Chou, 2001, 2005). The conventional AA composition did not include any sequence order effects, and hence the quality of prediction based on it would be limited no matter which algorithm was adopted. To cope with such a situation, the concept of "pseudo-amino acid composition" was proposed by Chou (2001). The advantages of PseAA composition are: (1) it can incorporate a considerable amount of sequence-order effects as well as

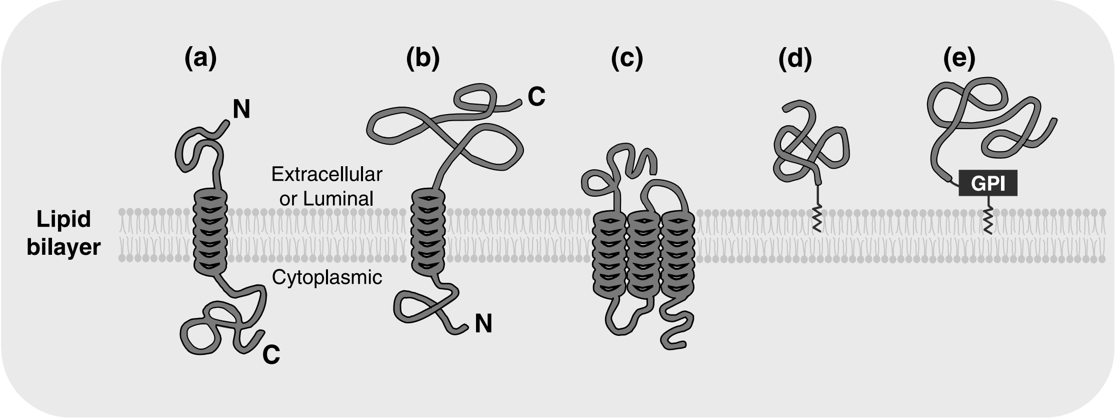

Fig. 1. Schematic drawing showing the following five types of membrane proteins: **a** Type-I transmembrane, **b** Type-II transmembrane, **c** multipass transmembrane, **d** lipid-chain anchored membrane, **e** GPI-anchored membrane. As shown here, although both Type-I and Type-II membrane proteins 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, while the arrangement of N- and C-termini in Type-II membrane proteins is just reverse. No such distinction was drawn between the extracellular (or luminal) and cytoplasmic sides for the other three types in the current classification scheme. Reproduced from Chou (2001) with permission

all the information in the AA composition, and (2) it has the same format as the AA composition except containing more components so that many of the existing analytical prediction algorithms can be straightforwardly applied to deal with it.

In the present study, based on the PseAA composition, a novel classifier, the so-called "ensemble classifier", is proposed for predicting the membrane protein type. Below, let us first give a brief introduction about the PseAA composition.

Representation of PseAA composition

According to the classical definition, the AA composition of a protein is defined by 20 discrete numbers with each representing the occurrence frequency of one of the 20 native amino acids in the protein. Thus, in terms of amino acid composition, a protein **P** can be expressed by a point or a vector in a 20D (dimensional) space as formulated by previous investigators (Chou and Zhang, 1993, 1994; Chou, 1989; Nakashima et al., 1986):

$$\mathbf{P} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_{20} \end{bmatrix} \tag{1}$$

where $p_1, p_2, \ldots, p_{20}$ are the occurrence frequencies of the 20 native amino acids in the protein **P**. However, if using the 20D AA composition to represent a protein, all its sequence-order and sequence-length effects will be totally lost. In view of this, rather than the conventional AA composition, Chou (2001) proposed representing a protein sample by its PseAA composition, which is defined in a $(20 + \lambda)$D space as formulated below:

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

where the first 20 components are the same as those in the conventional AA composition, and p_{20+1} is the 1$^{\text{st}}$ pseudo amino acid component related to the 1$^{\text{st}}$ rank of sequence-order correlation (see Fig. 1 of Chou, 2001), p_{20+2} is the 2$^{\text{nd}}$ pseudo amino acid component related to the 2$^{\text{nd}}$ rank of sequence-order correlation, and do forth. Given a protein sequence, each of these pseudo amino acid components $p_{20+1}, \ldots, p_{20+\lambda}$ can be easily computed according to Eqs. (2)–(6) of Chou (2001). Therefore, λ in Eq. (2) actually also represents the number of the pseudo amino acid components concerned. For instance, $\lambda = 38$ means taking the first 38 ranks of sequence-order correlations into consideration. Thus, according to Eq. (2), a protein sample is represented by a $(20 + \lambda)\text{D} = 58\text{D}$ vector.

Materials and methods: ensemble classifier

In the present study, we shall introduce the ensemble classifier to deal with the problem on the basis of PseAA composition. The framework of ensemble classifier system was established by combining numerous basic classifiers together in order to reduce the variance caused by the peculiarities of a single training set and hence be able to learn a more expressive concept in classification than a single classifier. The prediction system is constituted by a set of nearest neighbor (NN) classifiers (Cover and Hart, 1967), with each trained based on the dataset generated with different λ of Eq. (2). If $\lambda = \lambda_i$, the individual classifier thus obtained is denoted by $\mathrm{NN}(\lambda_i)$. Suppose

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_\Gamma\} \tag{3}$$

represents a set of possible numbers for λ of Eq. (2), then we have a set of corresponding classifiers $\mathrm{NN}(\lambda_1), \mathrm{NN}(\lambda_2), \ldots, \mathrm{NN}(\lambda_\Gamma)$, respectively. The ensemble classifier formed by combining such a set of individual classifiers is formulated by

$$\mathbb{NN} = \mathrm{NN}(\lambda_1) \forall \mathrm{NN}(\lambda_2) \forall \cdots \forall \mathrm{NN}(\lambda_\Gamma) \tag{4}$$

where $\mathbb{NN}$ denotes the ensemble classifier; $\mathrm{NN}(\lambda_1)$ the individual nearest neighbor classifier trained by proteins based on $(20 + \lambda_1)$ components (see Eq. (2)), $\mathrm{NN}(\lambda_2)$ the classifier based on $(20 + \lambda_2)$ components, and so forth; the symbol $\forall$ denotes the combination operator. When $\lambda_1 = \lambda_2 = \cdots = \lambda_\Gamma = 0$, the ensemble classifier $\mathbb{NN}$ is degenerated into a single classifier based on the conventional 20D AA composition only without any sequence-order effects (see Eq. (1)). The prediction result is determined according to the voting scores of all the constituent classifiers: $\mathrm{NN}(\lambda_1), \mathrm{NN}(\lambda_2), \ldots, \mathrm{NN}(\lambda_\Gamma)$. The ensemble classifier thus formed can better reflect the sequence-order effects and reduce the variance caused by the peculiarities of some individual subsets.

A flowchart to show how the ensemble classifier works is given in Fig. 2, from which we can see that the final output of the ensemble classifier $\mathbb{NN}$ is actually a "fusion" of the outputs produced by a set of basic classifiers: $\mathrm{NN}(\lambda_1), \mathrm{NN}(\lambda_2), \ldots, \mathrm{NN}(\lambda_\Gamma)$. The outcome of the fusion is a voting result among the constituent individual classifiers operated independently with different $\lambda_i (i = 1, 2, \ldots, \Gamma)$, respectively.

Let us consider a problem of classifying entities into μ classes, as formulated by

$$\mathbb{C} = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mu\} \tag{5}$$

The available information is assumed to consist in a training dataset S of N proteins:

$$S = \{(\mathbf{P}_1, \xi_1), (\mathbf{P}_2, \xi_2), \ldots, (\mathbf{P}_N, \xi_N)\} \tag{6}$$

where ξ_1 is the class label for protein $\mathbf{P}_1$, ξ_2 the class label for protein $\mathbf{P}_2$, and so forth. The class labels $\{\xi_1, \xi_2, \ldots, \xi_N\}$ in Eq. (6) take the values from $\mathbb{C}$ of Eq. (5); i.e.,

$$\{\xi_1, \xi_2, \ldots, \xi_N\} \in \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mu\} \tag{7}$$

With a given set of values for $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_\Gamma\}$, we can generate Γ different pseudo amino acid datasets according to Eq. (2) (Chou, 2001), and hence Γ classifiers: $\mathrm{NN}(\lambda_1), \mathrm{NN}(\lambda_2), \ldots, \mathrm{NN}(\lambda_\Gamma)$. Thus, the predicted result by the ensemble of the Γ classifiers (cf. Eqs. (3) and (4)) may be formulated as follows: Suppose $\mathbf{P}$ is a query protein whose classification predicted by the Γ individual classifiers are $\mathrm{Q}_1, \mathrm{Q}_2, \ldots, \mathrm{Q}_\Gamma$, respectively; i.e.,

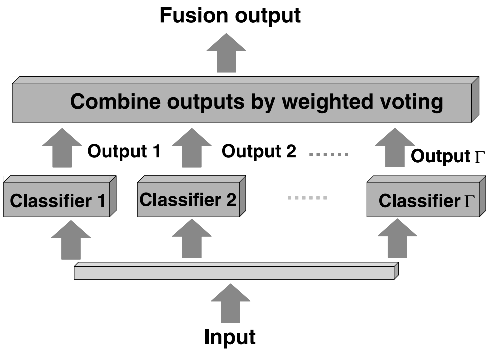

Fig. 2. Flowchart showing how the ensemble classifier $\mathbb{NN}$ (Eq. (4)) is formed by fusing Γ individual classifiers; $\Gamma = 40$ in the current study as indicated by Eq. (12)

$$\{\mathrm{Q}_1, \mathrm{Q}_2, \ldots, \mathrm{Q}_\Gamma\} \in \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mu\} \tag{8}$$

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

$$Y_j = \sum_{i=1}^{\Gamma} \delta(\mathrm{Q}_i, \mathrm{C}_j), \quad (j = 1, 2, \ldots, \mu) \tag{9}$$

where the delta function is given by

$$\delta(\mathrm{Q}_i, \mathrm{C}_j) = \begin{cases} 1, & \text{if } \mathrm{Q}_i \in \mathrm{C}_j \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

then the query protein $\mathbf{P}$ is predicted to the class with which its score of Eq. (9) is the highest; i.e., suppose

$$Y_\Phi = \mathbf{Max}\{Y_1, Y_2, \ldots, Y_\mu\} \tag{11}$$

where the operator **Max** means taking the maximum one among those in the brackets, and the subscript Φ is the very class predicted for the query protein $\mathbf{P}$. If there is a tie for the voting results, the query protein will be randomly assigned to one of the locations associated with the tie case. However, cases like that rarely happened in this study.

Results and discussion

The same training and testing datasets originally constructed by Chou and Elrod (1999) were used for the current study. The training dataset contains 2059 membrane protein sequences, of which 435 are Type-1 transmembrane proteins, 152 Type-2 transmembrane proteins, 1311 multipass transmembrane proteins, 51 lipid-chain anchored transmembrane proteins, and 110 GPI-anchored transmembrane proteins (Fig. 1). The independent testing dataset contains 2625 proteins, of which 478 are Type-1 transmembrane proteins, 180 Type-2 transmembrane proteins, 1867 multipass transmembrane proteins, 14 lipid-chain anchored transmembrane proteins, and 86 GPI-anchored transmembrane proteins.

Suppose

$$\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_\Gamma\} = \{1, 2, \ldots, 40\} \tag{12}$$

which means that the current ensemble classifier is constituted by 40 basic NN classifiers: the 1^{st} one, $\mathrm{NN}(\lambda_1)$, is defined in $20 + \lambda_1 = 21$ pseudo amino acid space, the 2^{nd} one, $\mathrm{NN}(\lambda_2)$, defined in $20 + \lambda_2 = 22$ pseudo amino acid space, and so forth. Also, for the NN classifier, there are many options for the similarity metric, such as Euclidean distance metric (Nakashima et al., 1986), Hamming distance metric (Chou, 1989), and Mahalanobis distance metric (Chou, 1995; Chou and Zhang, 1994; Mahalanobis, 1936). Here, we adopt the Euclidean distance.

The demonstration was conducted by the independent dataset test and jackknife test. In the independent dataset test, the rule parameters were derived from the proteins only in the training dataset, and the prediction was made for proteins in an independent dataset. In the jackknife test, each protein in the training dataset was singled out in turn as a ‘‘test protein’’ and all the rule parameters were calculated from the remaining $N-1$ proteins. In other words, the type of each protein was predicted by the rules derived using all the other proteins except the one that was being predicted. During the jackknifing process, both the training and testing dataset were actually open, and a protein was in turn moving from one to the other. Because the selection of independent dataset often bears some sort of arbitrariness, the jackknife test is deemed more objective than the independent dataset test. Actually, jackknife tests are thought one of the most rigorous and objective methods for cross-validation in statistics (see Chou and Zhang (1995) for a comprehensive review and Mardia et al. (1979) for the mathematical principles), and have been used by more and more investigators (Feng, 2001, 2002; Gao et al., 2005; Guo et al., 2006; Luo et al., 2002; Sun and Huang, 2006; Wang et al., 2004, 2005; Xiao et al., 2005a, b, 2006a, b; Zhang et al., 2006; Zhou and Assa-Munt, 2001; Zhou and Cai, 2006; Zhou and Doctor, 2003) in examining the power of various prediction methods.

The performance of ensemble classifier is evaluated by two methods: i.e. accuracy and Matthew correlation coefficients (MCC) (Matthews, 1975), which are defined as follows. Suppose that $i(1,2,\ldots,5)$ denote the 5 membrane protein types as illustrated in Fig. 1, respectively, m_i is the number of proteins observed as type i, and $\psi_{i,j}(i,j=1,2,\ldots,5)$ represents the number of proteins predicted as type j for those observed as type i. Thus we have

$$\text{Accuracy}_i = \frac{p_i}{m_i} \quad (i=1,2,\ldots,5) \tag{13}$$

$$\text{MCC}_i = \frac{p_i n_i - u_i o_i}{\sqrt{(p_i+u_i)(p_i+o_i)(n_i+u_i)(n_i+o_i)}} \tag{14}$$

where

$$\begin{cases} p_i = \psi_{i,i} \\ n_i = \sum_{j\neq i}^{5}\sum_{k\neq i}^{5}\psi_{j,k} \\ o_i = \sum_{j\neq i}^{5}\psi_{j,i} \\ u_i = \sum_{j\neq i}^{5}\psi_{i,j} \end{cases} \tag{15}$$

In other words, p_i is the number of correct predictions for type i; n_i the number of correct predictions not of type i; o_i the number of over-predictions for type i; and u_i the number of under-predictions.

The overall success rates obtained by the ensemble classifier for the jackknife test and independent dataset test are shown in Table 1, where, for facilitating comparison, the success rates by the other algorithms are also

Table 1. Overall success rates obtained by different classifiers and test methods for the 2059 proteins in the training dataset and 2625 proteins in the independent testing dataset[a]

Classifier	Input form	Test method	
		Jackknife[b]	Independent dataset[c]
Least hamming distance (Chou, 1989)	Amino-acid composition	$\frac{1279}{2059}=62.1\%$	$\frac{1751}{2625}=66.7\%$
Least Euclidean distance (Nakashima et al., 1986)	Amino-acid composition	$\frac{1293}{2059}=62.8\%$	$\frac{1816}{2625}=69.2\%$
ProtLock (Cedano et al., 1997)	Amino-acid composition	$\frac{1348}{2059}=65.5\%$	$\frac{1674}{2625}=63.8\%$
Covariant-discriminant (Chou and Elrod, 1999)	Amino-acid composition	$\frac{1573}{2059}=76.4\%$	$\frac{2085}{2625}=79.4\%$
Augmented covariant discriminant (Chou, 2001)	Pseudo-amino acid composition	$\frac{1665}{2059}=80.0\%$	$\frac{2298}{2625}=87.5\%$
Ensemble classifier[d]	Pseudo-amino acid composition	$\frac{1767}{2059}=85.8\%$	$\frac{2540}{2625}=96.8\%$

[a] The datasets investigated here were taken from Chou and Elrod (1999)

[b] Conducted for the 2059 membrane proteins in the training dataset classified into 5 different types (Fig. 2)

[c] Predicted for the 2625 independent membrane proteins based on the rule parameters derived from the 2059 membrane proteins in the training dataset

[d] The ensemble classifier $\mathbb{N}\mathbb{N}$ consists of 40 basic individual classifiers with $\{\lambda_1,\lambda_2,\ldots,\lambda_\Gamma\}=\{1,2,\ldots,40\}$, respectively (cf. Eqs. (3), (4), and (12))

Table 2. The detailed success rates and their Matthew correlation coefficients (Eq. (14)) for each of the five membrane types obtained by the jackknife test and independent dataset test[a]

Type of membrane protein	Jackknife		Independent dataset	
	Accuracy (%)	MCC	Accuracy (%)	MCC
Type-1	81.2	0.737	96.0	0.950
Type-2	44.7	0.527	79.4	0.862
Multipass	95.8	0.800	99.0	0.934
Lipid-chain anchored	47.1	0.654	57.1	0.675
GPI anchored	60.0	0.640	90.7	0.915

[a] See footnotes of Table 1 for further explanation

listed. It can be seen from Table 1 that the success rates by the current ensemble classifier approach for both jackknife test and independent dataset test are remarkably higher than those by the others. Moreover, the detailed success rates and their Matthew correlation coefficients for each of the five membrane types by both tests are given in Table 2.

Conclusions

An ensemble of classifiers is a set of classifiers, whose individual classification decisions are combined in some way, typically by a weighted or unweighted voting, to classify new samples. In the current ensemble classifier, the individual classifiers are the NN classifiers (Cover and Hart, 1967) with each trained based on different PseAA composition spaces (Chou, 2001), as reflected by different λ, the number of the pseudo amino acid components (cf. Eqs. (2) and (3)). It is instructive to note that although the elemental classifier used here is the NN classifier, others such as the covariant discriminant classifier (Chou and Elrod, 1999) and SVM classifier (Cortes and Vapnik, 1995) can also be used to replace the NN classifier for forming different ensemble classifiers. Moreover, the elemental classifiers in an ensemble classifier can also be a mixture of individual classifiers with complete different types. It is anticipated that the approach of ensemble classifier as introduced here might have a series of positive impacts for improving the prediction quality for many other protein attributes as well.

Acknowledgement

H. B. Shen would like to express his gratitude to Dr. Yan Huang at Shanghai Institute of Technical Physics, Chinese Academy of Science, for valuable discussion.

References

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1994) Molecular biology of the cell, chapter 1, 3rd edn. Garland Publishing, New York London

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

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257–3263

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

Chou JJ, Zhang CT (1993) A joint prediction of the folding types of 1490 human proteins from their genetic codons. J Theor Biol 161: 251–262

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

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

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

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

Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. J Biol Chem 269: 22014–22020

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

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

Cortes C, Vapnik V (1995) Support vector networks. Mach Learn 20: 273–293

Cover TM, Hart PE (1967) Nearest neighbour pattern classification. IEEE Trans Inform Theory IT-13: 21–27

Feng ZP (2001) Prediction of the subcellular location of prokaryotic proteins based on a new representation of the amino acid composition. Biopolymers 58: 491–499

Feng ZP (2002) An overview on predicting the subcellular location of a protein. In Silico Biol 2: 291–303

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

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

Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J (1995) Mol Cell Biol, Chapter 3, 3rd ed. Scientific American Books, New York

Luo RY, Feng ZP, Liu JK (2002) Prediction of protein strctural class by amino acid and polypeptide composition. Eur J Biochem 269: 4219–4225

Mahalanobis PC (1936) On the generalized distance in statistics. Proc Natl Inst Sci India 2: 49–55

Mardia KV, Kent JT, Bibby JM (1979) Multivariate analysis: Chapter 11: Discriminant analysis; Chapter 12: Multivariate analysis of variance; Chapter 13: cluster analysis. Academic Press, London, pp 322–381

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405: 442–451

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

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

*Pillai KCS (1985) Mahalanobis D2. In: Kotz S, Johnson NL (eds) Encyclopedia of statistical sciences, Vol. 5. Wiley, New york, pp 176–181

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

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

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Design Select 17: 509–516

Wang M, Yang J, Xu ZJ, Chou KC (2005) SLLE for predicting membrane protein types. J Theor Biol 232: 7–15

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

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

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006a) Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location. Amino Acids 30: 49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006b) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30: 461–468

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins Struct Function Genet 44: 57–59

Zhou GP, Cai YD (2006) Predicting protease types by hybridizing gene ontology and pseudo amino acid composition. Proteins Struct Function Bioinform 63: 681–684

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins Struct Function Genet 50: 44–48

Authors' address: Kuo-Chen Chou, Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, U.S.A.,
Fax: +1-858-484 1018, E-mail: kchou@san.rr.com

* This reference also presents a brief biography of Mahalanobis who was a man of great originality and who made considerable contributions to statistics, New York.