African Journal of Biotechnology Vol. 10(74), pp. 16963-16968, 23 November, 2011
Available online at http://www.academicjournals.org/AJB
DOI: 10.5897/AJB11.429
ISSN 1684–5315 © 2011 Academic Journals

Full Length Research Paper

A novel ensemble and composite approach for classifying proteins based on Chou's pseudo amino acid composition

Jie Lin[1], Yan Wang[1]* and Xu Xu[1]

Department of Information Management and Information System, College of Economics and Management, Tong Ji University, Shanghai 216000, China.

Accepted 11 November, 2011

For the fact that the location of proteins gave some details about the function of a protein whose location was uncertain, protein classification was regarded as a very important task in the field of biological data mining. However, the success of a human genome project led to a protein sequence explosion. There is a great need to develop a computational method for fast and reliable prediction of the locations of proteins according to their primary sequences. In this paper, we used the composite classifier system that was formed by a set of k-nearest neighbor (K-NN) classifiers, each of which was defined in a different pseudo amino composition vector. In the pseudo amino composition vector space, protein can be presented by Pseudo amino acid composition. The location of a queried protein is determined by the outcome of choice made among these constituent individual classifiers. It is shown through the outcome that the classifier outperformed the single classifier widely used in biological literature. So the composite classifier can be employed as a robust method to predict protein location in the field of biological data mining.

Key words: Composite classifier system, biological data mining, atomic classifiers, pseudo amino acid composition.

INTRODUCTION

Since location plays a crucial role in protein function, prediction of the location of proteins remains very important in the field of protein biology. Given the sequence of a protein, how can its cellular location be determined? Subsequently, in this paper, we used one composite classifier system to predict it.

The recent success of the human genome project led to a protein sequence explosion. In 1986, the SWISS-PROT databank contained only 3939 protein sequence entries (Bairoch et al., 2000), but now, it has 522019 entries according to version 2010_11 released as of November 02, 2010, meaning that the number of protein sequences has increased by about 132 times in 24 years. Facing many difficulties in affording enough time and money to perform suitable functional tests, researchers are challenged to design computational methods for predicting structure and function. In this paper, proteins are divided into the following 5 types (Cedano et al., 1997): (1) integral membrane, (2) anchored membrane proteins (with transmembrane amino acid stretch), (3) extracellular proteins, (4) intracellular proteins (non nuclear), and (5) nuclear proteins. In some previous studies, Nakashima et al. (1994) reported the discrimination between intracellular and extracellular proteins by amino acid composition and residue-pair frequencies; whereas Chou and Elrod (1999) developed the covariant discriminate algorithm, which is a combination of the "Mahalanobis distance", and the invariance principle for treating a degenerate vector space that is cited in the following literature as "Chou's invariance theorem" (Chou, 2001; Mardia, 1977; Pillai, 1985; Matthews, 1975; Chou et al., 1995). Some of those existing prediction methods are based on the conventional amino acid composition [3] (Nakashima et al., 1994; Pillai,

*Corresponding author. E-mail: wangyan@hpu.edu.cn. Tel: +86-391-3987257. Fax: +86-391-3987257.

1985), while others are based on the Pseudo amino acid composition (Chou et al., 1999; Chou, 2005). The last algorithm was incorporated into the sequence order information; so, it can be called the pseudo amino acid composition. However, it was proposed by Chou et al. (1999). The advantages of Pseudo amino composition are: (1) it can incorporate a considerable amount of sequence order information; and (2) it has the same format as the amino acid composition, so that some algorithms used in amino acid composition can be applied in pseudo amino acid composition.

According to a recent comprehensive review by Chou (2011), to develop a useful predictor for protein systems, the following things must often be considered: (1) protein sample formulation, (2) operating algorithm (or engine), (3) benchmark dataset construction or selection, (4) anticipated accuracy, and (5) web-server establishment. Subsequently in this work, we would introduce the pseudo amino acid composition and composite classifier system. Moreover, some of these key procedures are described thus.

MATERIALS AND METHODS

Pseudo amino acid composition

As we all know, the conventional amino acids (AA) composition did not include any sequence order effects. Instead of using the conventional 20-D amino acid composition to represent the sample of a protein, Prof. Kuo-Chen Chou proposed the pseudo amino acid composition (PseAAC) in order for it to be included in the sequence-order information (Chou et al., 1999; Chou, 2001; Chou, 2005; Chou et al., 2009). PseAAC allows users to generate various kinds of pseudo amino acid composition for a given protein sequence. The conventional amino acid composition contains 20 components or discrete numbers, each reflecting the occurrence frequency of one of the 20 native amino acids in a protein. For the pseudo amino acids composition, there are, however, sequence order effects in addition to the 20 components. Equations (1) and (2) show the difference between the amino acid composition and the pseudo amino acid composition.

A protein X is represented by a vector in 20D (dimensional) spaces as defined by previous investigators (Chou et al., 1993, 1994; Chou, 1989; Nakashima et al., 1986; Mahalanobis, 1936). It contains 20 components, or discrete numbers, each reflecting the occurrence frequency of one of the 20 native amino acids in a protein:

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_{20} \end{pmatrix} \quad (1)$$

Where, x_i is the normalized occurrence frequency of the 20 amino acids in protein *X*.

In 2001, Chou proposed the pseudo amino acid composition as shown in (2). In Equation (2), instead of using a 20D(dimensional) vector defined by 20 components, we used a (20+ λ)D vector defined by 20+ λ discrete numbers to represent protein *X*, where x_i has the same meaning as in amino acid composition, whereas the additional components from 20+1 to 20+ λ reflect the effect of sequence order. Here, x_{20+1} is the 1st pseudo amino acid component related to the 1st rank of sequence order correlation (Figure 2), x_{20+2} is the 2nd pseudo amino acid component related to the 2nd rank of sequence order correlation, and so forth. As such, they were called pseudo amino acid components (for a brief introduction about Chou's pseudo amino acid composition, visit the Wikipedia web-page at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition). In addition to this component in (1), a protein can be represented as follows:

$$X = \begin{pmatrix} x_i \\ x_2 \\ \vdots \\ x_{20} \\ x_{20+1} \\ \vdots \\ x_{20+\lambda} \end{pmatrix} \quad (2)$$

Given a protein sequence, the pseudo amino acid component can be computed according to Equations (2) to (6) (Chou, 2001).

KNN classifier

With the KNN classifier, when K >1, the attribute of the query protein P will be determined via the choice made by a majority of its K nearest neighbors, as can be described as follows. Suppose (p_1,p_2, … p_n) (N>=K) are the n proteins in training dataset, the query protein P will be predicted to belong to the ith class, if the most neighbors in K of it belong to *ith* class (Figure 1).

In Figure 1, where the query protein P is represented by the character q with a filled circle, proteins belonging to subset (category 1) are represented by the open circle with number 1, proteins that do not are represented by the open circle with number 2, and so forth. When K=1, the query protein is predicted to belong to category 2 as its nearest protein does; when K=3, the query protein is predicted to belong to category 3 because two of its three nearest proteins belong to that category; and when K=9, the query protein is predicted to belong to category 2 again because the majority of its nine nearest proteins belong to category 2.

Composite classifier system

Now, we shall introduce the composite classifier system in order to deal with it on the basis of pseudo amino acid composition. The framework of the composite classifier system was established by combining lots of atomic classifiers together in order to reduce the variance caused by the peculiarities of a single training set and hence be able to generate a more expressive concept in classification than a single classifier. In this paper, we used a set of k-nearest neighbor (K-NN) classifiers which is trained by different dataset generated by different λ.

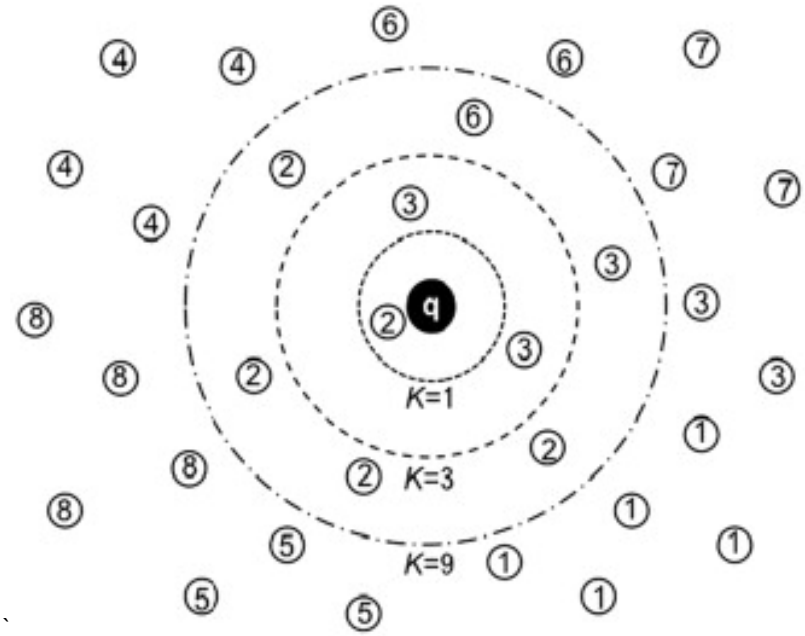

Figure 1. An illustration showing how the KNN classifier depends on the selection of parameter K in identifying the attribute category of a query protein. Reproduced from Chou (2011) with permission.

In this study, we used the composite classifier system frame defined by Dr. Shen Hong bin in 2007. We changed only his atomic classifier-NN classifier into K-NN classifier because we divided the proteins' location into 5 styles. The Dr. Shen's definition of the composite classifier system can be described as follows:

Suppose $\wedge = \{\lambda_1, \lambda_2, \ldots \lambda_\Gamma\}$ represents a set of possible numbers for λ, then we can get a set of corresponding classifiers K-NN (λ_1), ... K-NN(λ_Γ) respectively, that is, the atomic classifier K-NN (λ_1) trained by proteins based on (20+ λ_1) components, K-NN(λ_2) based on (20+ λ_2) components, and so forth. For the K-NN classifier, we adopted the Euclidean distance and k=3. The final classifier that was integrated by such a set of individual classifiers can be introduced as:

$$CoMNN = K-NN(\lambda_1) \forall K-NN(\lambda_2) \cdots K-NN(\lambda_\Gamma) \quad (3)$$

Where, CoMNN is the integrated classifier that can be described by Figure 3. The symbol $\forall$ represents the combination operator:

$$C = \{C_1, C_2, \cdots, C_\mu\} \quad (4)$$

We can use the S represents N proteins in a training dataset S:

$$S = \{(P_i, C_j)\} \quad i \in (1, \cdots, N), j \in (1, \cdots \mu) \quad (5)$$

Chou has proved that the greater the number of λ, the more the sequence order effect that is been incorporated, but it must be smaller than the number of amino acid residues of the shortest protein chain in the data set concerned. In this paper, the shortest chain residue is 8. However, if we give the value of λ as 8, we can generate 8 different pseudo amino acid datasets according to (2); consequently, we can get 8 classifier K-NN (λ_1), ... K-NN(λ_8). Suppose P is a query protein whose classification is predicted by the 8 atomic classifiers as $Q_1, Q_{2,} \cdots, Q_8$, respectively; the following equations can be realized thus:

$$\{Q_1, Q_{2,} \cdots, Q_8\} \in C \quad (6)$$

and the final score for protein P belonging to the jth class is defined by:

$$Y_j = \sum_{i=1}^{8} \delta(Q_i, C_j) \quad j \in (1,2,3,4,5) \quad (7)$$

Where, the delta function is defined by:

$$\delta(Q_i, C_j) = \begin{cases} 1 & \text{if } Q_i \in C_j \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

For protein P, we can choose Y_α as the very class and Y_α as defined in Equation (9):

$$Y_\alpha = \text{Max}\{Y_1, \cdots, Y_\mu\} \quad (9)$$

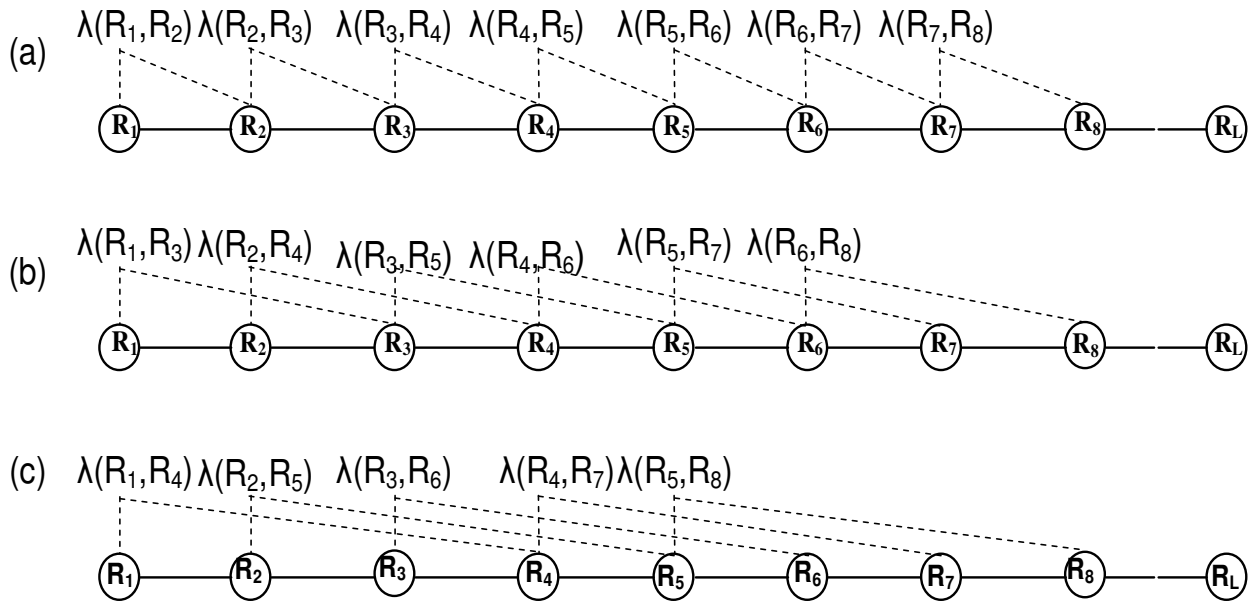

Figure 2. A schematic drawing to show (a) the first-tier, (b) the second-tier and (c) the third-tier sequence order correlation mode along a protein sequence. Reproduced from Chou (2001) with permission.

RESULTS

In the study written by Cedano et al. (1997), proteins' locations were generally classified into the following 5 types: (1) integral membrane proteins, (2) anchored membrane proteins, (3) extracellular proteins, (4) intracellular proteins (non nuclear), and (5) nuclear proteins. The corresponding 5 characters (M, A, E, I, N) represent these location types, respectively. In this paper, we still used the definition; although, the same training and testing dataset that we used originally was constructed by Cedano et al. (1997). In training dataset, every kind of protein has 200 sequences that have been reported in the SWISS-PROT. Another non-homologous 200 sequences were also abstracted from SWISS-PROT, and would be the testing dataset.

For a fair comparison, the same data studied by Cedano et al. (1997) were adopted here. However, because of the change and obsolescence of code names, some proteins' sequences could no longer be retrieved from the SWISS-PROT database. Of the 1000 protein originally used by Cedano sequences, 980 protein sequences were retrieved. They formed the training data set, which consisted of 196 A proteins, 193 E proteins, 197 I protein, 200 M protein and 194 N proteins. For the same reason of testing the dataset used by Cedano, 189 proteins were retrieved from SWISS-PROT database.

The performance of the composite classifier was evaluated by two methods: that is, accuracy and Matthew correlation coefficients (MCC) (Matthews BW. 1975). Suppose that *i*(1,2,3,4,5) denotes the 5 proteins' locations, respectively; m_i is the number of proteins observed as location *i*, and $\phi_{i,j}(i,j=1,2,\cdots,5)$ represents the number of proteins that were predicted to be having type *j* for those observed as type *i*. Thus, we have:

$$Accuracy_i = \frac{a_i}{m_i} \qquad (i=1,2,\cdots,5) \tag{10}$$

$$MCC_i = \frac{a_i n_i - u_i o_i}{\sqrt{(a_i+u_i)(a_i+o_i)(n_i+u_i)(n_i+o_i)}} \tag{11}$$

where

$$\begin{cases} a_i = \phi_{i,i} \\ n_i = \sum_{j\neq i}^{5}\sum_{k\neq i}^{5}\phi_{j,k} \\ o_i = \sum_{j\neq i}^{5}\phi_{j,i} \\ u_i = \sum_{j\neq i}^{5}\phi_{i,j} \end{cases} \tag{12}$$

To facilitate comparison, the accuracy and MCC of ProtLoc, NN classifier (the special case of KNN when k=1) and the composite classifier are shown in Table 1.

Among the independent dataset tests, sub-sampling (for example, 5 or 10-fold cross-validation) test and jackknife test were often used for examining the accuracy of a statistical prediction method (Chou et al., 1995). The jackknife test was deemed the least arbitrary that can always yield a unique result for a given benchmark dataset, as elucidated in Chou et al. (2008, 2010a, b, c)

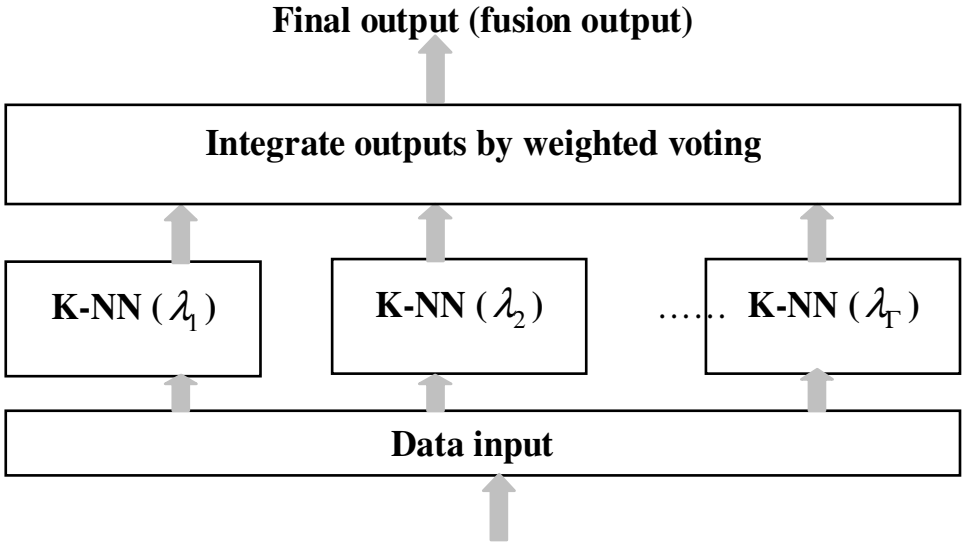

Figure 3. Flowchart showing how the composite classifier system called CoMNN is integrated by atomic classifiers.

Table 1. The detailed success rates and their Matthew correlation.

Type	ProtLoc[a]		K-NN (k=1)[b]		CoMNN[b]	
	Accuracy (%)	MCC	Accuracy (%)	MCC	Accuracy (%)	MCC
M	92.5	0.897	85.0	0.800	90.0	0.927
A	67.0	0.388	58.3	0.243	70.8	0.340
E	70.9	0.669	64.5	0.739	74.2	0.867
I	70.3	0.658	64.8	0.737	81.5	0.813
N	75.0	0.798	77.5	0.724	82.5	0.882[a]

[a] The method used by Cendao in 1997; [b] the same testing dataset used by Cendao in 1997. K-NN, k-nearest neighbor.

and demonstrated by Equations 28 to 32 of Chou (2011). Therefore, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods (Chen et al., 2009; Ding et al., 2009; Gu et al., 2010; Lin et al., 2008; Mohabatkar, 2010; Qiu et al., 2009; Sahu et al., 2010; Zeng et al., 2009; Zhou et al., 2007; Chou, 2010a, b, c; Xiao et al., 2011; Zou et al., 2011). So, in this paper, we chose jackknife methods as our cross-validation methods.

DISCUSSION

The overall success rates obtained by the composite classifier system are shown in Table 1, while the success rate of other algorithms used by Cedano are also shown in Table 1. It can be seen from Table 1 that the success rates by the current composite classifier system approach are remarkably higher than the method used by Cedano.

Although, the atomic classifier used here is K-NN classifier, in future researches, others such as decision tree classifier and SVM classifier can also be used to replace the K-NN classifier for integrating different composite classifier systems (Bing et al., 2006; Cortes and Vapnik, 1995). It is anticipated that the approach of composite classifier system as used here might have a series of positive impacts for bioinformatics. If we can extract the feature and get the correlation of features of text correctly, the method can also be used in classification for text.

User-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors (Chou et al., 2009). Thus, we shall make efforts in our

future work to provide a web-server for the method presented in this paper.

ACKNOWLEDGEMENTS

The authors acknowledged Professor Shen who provided them the web service for calculating PseAAC of protein. This work was supported by the National Science Foundation of China (No. 71071114), Doctoral Fund of Ministry of Education of China (No. 200802470009), Shanghai Leading Academic Discipline Project (No.B310) and Scholarship Award for Excellent Doctoral Student granted by Ministry of Education, China.

REFERENCES

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

Bing Niu, Yu-Dong Cai, Wen-Cong Lu, Guo-Zheng Li, Kuo-Chen Chou (2006). Predicting Proteins Structural Class with AdaBoost Learner. Protein Pept Lett. 13: 489-492.

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

Chen C, Chen L, Zou X, Cai P (2009). Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine. Protein Pept Lett. 16: 27-31.

Chou K C (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. J. Theor. Biol. 273: 236-247

Chou K C (2009). Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Current Proteomics. 6: 262-274. Chou KC (2005). Review: Progress in protein structural class predictionand its impact to bioinformatics and proteomics. Curr Protein Pept Sci. 6: 423-436.

Chou KC (2001). Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Function Genet. 44 (43): 246-255.

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

Chou K C, Shen H B (2010)a. A new method for predicting the subcellular localization of eukaryotic proteins with both single and multiple sites: Euk-mPLoc 2.0. PLoS ONE. *5*:e9931.

Chou K C, Shen H B (2010)b. Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization. PLoS ONE. 5: e11335.

Chou K C, Shen H B (2010)c. Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. Natural Science. 2: 1090-1103.

Chou K C, Shen H B (2009). Review: recent advances in developing web-servers for predicting protein attributes. Natural Science. 2: 63-92.

Chou K C, Shen H B (2008). Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. Nature Protocols. 3: 153-162.

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

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

Chou J J, 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 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. PlenumPress. 549–586.

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

Ding H, Luo L, Lin H (2009). Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition. Protein Pept Lett. 16: 351-355.

Gu Q, Ding Y S, Zhang T L (2010). Prediction of G-Protein-Coupled Receptor Classes in Low Homology Using Chou's Pseudo Amino Acid Composition with Approximate Entropy and Hydrophobicity Patterns. Protein Pept Lett. 17: 559-567.

H B Shen (2007). "Using ensemble classifier to identify membrane protein types. Amino Acids. 32: 483-448.

Lin H (2008). The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition. J.Theor. Biol. 252:350-356.

Mahalanobis PC (1936). On the generalized distance in statistics. Proc NatlInst Sci India. 2: 49-55.

Mardia K V (1977). Mahalanobis distances and angles.In Multivariate Analysis IV. pp.176-181.

Nakashima H, Nishikawa K (1994). Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pairfrequencies. J. Mol. Biol. 238: 54-61.

Mohabatkar H (2010). Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein Pept Lett. 17: 1207-1214.

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.

Pillai K C S (1985). Mahalanobis D2.In Encyclopedia of Statistical Sciences (Kotz, S. & Johnson, N. L., eds), John Wiley and Sons. 5: 176-181.

Qiu J D, Huang J H, Liang R P, Lu X Q (2009). Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform. Anal. Biochem. 390: 68-73.

Sahu S S, Panda G (2010). A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. Comput Biol Chem. 34: 320-327.

X, Wang P, Chou K C (2011). GPCR-2L: Predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. Mol Biosyst. *7*: 911-919.

Zeng Y H, Guo Y Z, Xiao R Q, Yang L, Yu L Z, Li M L (2009).Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach. J. Theor. Biol. 259: 366-372.

Zhou X B, Chen C, Li Z C, Zou X Y (2007). Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J. Theor. Biol. 248: 546-551. Xiao

Zou D, He Z, He J, Xia Y (2011). Supersecondary structure prediction using Chou's pseudo amino acid composition. J.Comput.Chem. 32: 271-278.