Front. Electr. Electron. Eng. China 2008, 3(4): 376–380
DOI 10.1007/s11460-008-0084-5

RESEARCH ARTICLE

Tong WANG, Hongbin SHEN, Lixiu YAO, Jie YANG, Kuochen CHOU

PCA for predicting quaternary structure of protein

Abstract The number and arrangement of subunits that form a protein are referred to as quaternary structure. Knowing the quaternary structure of an uncharacterized protein provides clues to finding its biological function and interaction process with other molecules in a biological system. With the explosion of protein sequences generated in the Post-Genomic Age, it is vital to develop an automated method to deal with such a challenge. To explore this problem, we adopted an approach based on the pseudo position-specific score matrix (Pse-PSSM) descriptor, proposed by Chou and Shen, representing a protein sample. The Pse-PSSM descriptor is advantageous in that it can combine the evolution information and sequence-correlated information. However, incorporating all these effects into a descriptor may cause 'high dimension disaster'. To overcome such a problem, the fusion approach was adopted by Chou and Shen. A completely different approach, linear dimensionality reduction algorithm principal component analysis (PCA) is introduced to extract key features from the high-dimensional Pse-PSSM space. The obtained dimension-reduced descriptor vector is a compact representation of the original high dimensional vector. The jackknife test results indicate that the dimensionality reduction approach is efficient in coping with complicated problems in biological systems, such as predicting the quaternary structure of proteins.

Keywords principal component analysis (PCA), quaternary structure of protein, pseudo position-specific score matrix (Pse-PSSM), dimension reduction method

1 Introduction

Proteins are at the center of the action in biological processes, and their function can only be understood based on the structure of their constituent polypeptide chains. Thus, protein structure plays a key role in cell biology, biochemistry and molecular biology. The structural hierarchy of proteins is defined at four levels: primary, secondary, tertiary, and quaternary.

Primary structure is defined by the amino acid sequence. Secondary structure is the local spatial arrangement of a polypeptide's backbone, without regard for the conformations of its side-chains. Tertiary structure refers to the three-dimensional structure of an entire polypeptide. The concept of quaternary structure was first put forward by Bernal in 1958 [1]. It refers to the non-covalent interaction of protein subunits to form oligomers. Oligomeric proteins are very common in nature. They can be divided further into two classes: homo-oligomers and hetero-oligomers. The former are composed of identical subunits while the latter are composed of non-identical subunits. The present study focuses on the homo-oligomers.

Oligomeric structure can also vary through arrangement of subunits. Thus, in the protein universe, there are many different classes of subunit construction, such as monomer, dimer, trimer, tetramer, and so forth (Fig. 1). Single subunit or polypeptide chain is called a monomer, two subunits a dimer, three a trimer, four a tetramer, etc. The Oligomeric proteins have more advantages than the monomers in terms of functional evolution of biomacromolecules [2]. It is easier for multi-subunit proteins to repair their defects by simply replacing the flawed subunit. Moreover, in many biological processes the quaternary structure is an interesting field in bioinformatics.

It is generally accepted that the amino acid sequence of most proteins contains all the information needed to fold the protein into its correct three-dimensional structure [4,5]. The quaternary structure of proteins, which is the association of tertiary structure subunits, depends on the existence of complementary 'patches' on their surfaces [6]. Therefore, the patches that are buried in the interfaces formed by the subunits play a vital role in both tertiary and quaternary structures. This suggests the possibility of predicting the quaternary structure from primary sequences [6].

Received June 12, 2008; accepted July 18, 2008

Tong WANG, Hongbin SHEN (✉), Lixiu YAO, Jie YANG
Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai 200240, China
E-mail: hbshen@sjtu.edu.cn

Kuochen CHOU
Gordon Life Science Institute, San Diego, CA 92130, USA

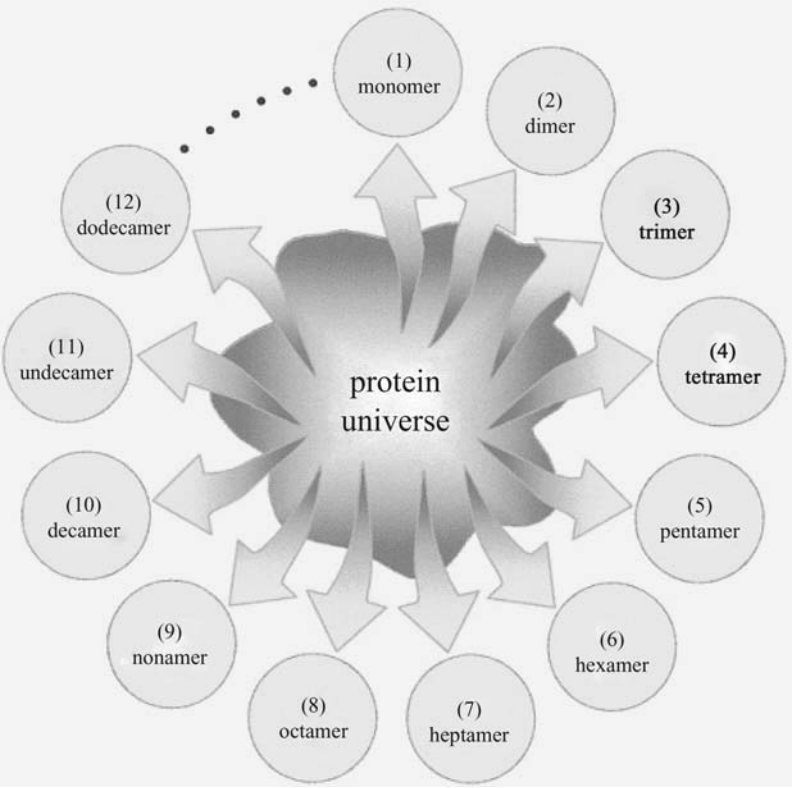

Fig. 1 Schematic drawing to illustrate that different polypeptide chains may form various oligomers. Note: Reproduced from Chou [3] with permission

Given a polypeptide chain, will it form a dimer, trimer or any other oligomer, or exist only as a monomer? Efforts have been made in developing computational tools to predict protein quaternary structure based on its sequence. In a pioneering study, Chou and Elrod [3] introduced the covariant discriminant algorithm to predict the quaternary structure of proteins based on the pseudo amino acid (PseAA) composition [7]. By using the PseAA composition, a protein sequence can be expressed by a discrete model yet without completely losing its sequence-order information. A web-server called PseAAC [8] was established at http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/ or http://chou.med.harvard.edu/bioinf/PseAAC/, by which users can generate various kinds of PseAA composition as desired.

Recently, a new discrete descriptor for a protein sequence, called the pseudo position-specific score matrix (Pse-PSSM) [9], is proposed by incorporating its evolution and sequence-order information. However, the Pse-PSSM descriptor would correspond to a very high dimensional vector if all possible coupling ranks were incorporated into one descriptor. This may cause many problems in statistical prediction, such as the 'dimension disaster', over-fitting and redundancy. To overcome such a difficulty, the approach by fusing many descriptors of different coupling ranks into one classifier, i.e., the so-called ensemble classifier, was utilized by Chou and Shen [9].

The present study was initiated in an attempt to propose a different approach, i.e., the principal component analysis (PCA), to extract the essential features from the vector space to avoid the high-dimensional difficulty. The result thus obtained is quite encouraging, indicating that the PCA approach can also be effectively used to deal with other complicated biological systems.

2 Materials and methods

2.1 Dataset

Protein sequences were collected from the Swiss-Prot database at http://www.ebi.ac.uk/swissprot/ (version 55.0 released on 26-February-2008). To collect as much desired information as possible and meanwhile ensure high-quality for the benchmark dataset, the data were screened strictly according to the following criteria and order:

1) Sequences annotated with 'fragment' were excluded. Also, sequences with less than 50 amino acid residues were excluded because they might just be fragments.

2) Sequences annotated with ambiguous or uncertain terms, such as 'potential', 'probable', 'probably', 'maybe', or 'by similarity', were removed from further consideration.

3) For the sequences kept after the above screening procedures and all that have clear experimental annotations, those annotated with 'subunit' were extracted.

4) Eight different quaternary structures were found. To reduce homology bias, a redundancy cutoff was operated by the PISCES [10] program to select sequences that have $> 60\%$ sequence identity to any other in the same quaternary structure type. Finally, we obtained a dataset containing 3726 sequences, of which 1073 are monomers, 1681 are homodimers, 224 are homotrimers, 549 are homotetramers, 17 are homopentamers, 115 are homohexamers, 46 are homooctamers, and 21 are homododecamers.

2.2 Pseudo position-specific score matrix descriptor

Given a query protein sequence $\boldsymbol{P}$, to predict its quaternary structure type, the first important thing we need to do is use a proper descriptor to represent it. The descriptor not only contains as much information of the sequence as possible, but also can be handled by powerful prediction algorithms. The Pse-PSSM descriptor recently introduced by Chou and Shen [9] is a good one in this regard. Below, let us give a brief introduction about the Pse-PSSM descriptor. According to Ref. [9], a protein sequence containing L amino acids can be represented by a 40-D (dimensional) vector, i.e.,

$$\boldsymbol{P}^{\xi}_{\text{Pse-PSSM}} = \left[\begin{array}{cccccccc} \bar{\mathbb{E}}_1 & \bar{\mathbb{E}}_2 & \cdots & \bar{\mathbb{E}}_{20} & G^{\xi}_1 & G^{\xi}_2 & \cdots & G^{\xi}_{20} \end{array}\right]^{\mathrm{T}}, \quad \xi = 0,\ 1,\ 2,\ \text{or}\ L-1, \tag{1}$$

where T is the transpose operator, and

$$\bar{\mathbb{E}}_j = \frac{1}{L}\sum_{i=1}^{L} \mathbb{E}_{i\to j}, \quad j = 1, 2, \ldots, 20, \tag{2}$$

where L is the number of amino acids in the protein, $\mathbb{E}_{i\to j}$ represents the score of the amino acid residue in the ith position of the protein sequence that is being changed to amino acid type j during the evolution process. Here, the numerical codes 1, 2,..., 20 are used to denote the 20 native amino acid types according to the alphabetical order of their single character codes. The PSSM scores of $\mathbb{E}_{i\to j}$ in Eq. (2) were generated using PSI-BLAST [11] to search the Swiss-Prot database through three iterations with 0.001 as the E-value cut-off for multiple sequence alignment against the sequence of the protein concerned, followed by a standardization procedure given below:

$$\mathbb{E}_{i\to j} = \frac{\mathbb{E}^0_{i\to j} - \dfrac{1}{20}\sum\limits_{k=1}^{20}\mathbb{E}^0_{i\to k}}{\sqrt{\dfrac{1}{20}\sum\limits_{u=1}^{20}\left(\mathbb{E}^0_{i\to u} - \dfrac{1}{20}\sum\limits_{k=1}^{20}\mathbb{E}^0_{i\to k}\right)^2}}, \quad i = 1,2,\ldots,L, \quad j = 1,2,\ldots,20, \tag{3}$$

where $\mathbb{E}^0_{i\to j}$ represents the original scores directly created by PSI-BLAST and are generally positive or negative integers. The standardized scores will have a zero mean value over the 20 amino acids and will remain unchanged if it goes through the same conversion procedure again. The positive score means that the corresponding mutation occurs more frequently in the alignment than expected by chance, while the negative one means just the opposite.

The components $G^{\xi}_1, G^{\xi}_2, \ldots, G^{\xi}_{20}$ in Eq. (1) are given by

$$G^{\xi}_j = \frac{1}{L-\xi}\sum_{i=1}^{L-\xi}\left[\mathbb{E}_{i\to j} - \mathbb{E}_{(i+\xi)\to j}\right]^2, \quad j = 1,2,\ldots,20, \quad \xi < L, \tag{4}$$

meaning that G^1_j is the correlation factor by coupling the most contiguous PSSM scores along the protein chain for the amino acid type j, that G^2_j is obtained by coupling the second-most contiguous PSSM scores, etc. Because the length of the shortest protein sequence in our dataset is $L = 50$, the value allowed for ξ in Eqs. (1) and (4) must be smaller than 50; when $\xi = 0$, G^{ξ}_j becomes a naught element and Eq. (1) is degenerated to a 20-D vector. Thus, according to the Pse-PSSM descriptor, as shown in Eq. (1), a protein can be represented by one 20-D vector ($\xi = 0$) and 49 different 40-D vectors, each of which corresponds to a different ξ (1,2,...,49). To avoid the over-fitting problem and reduce the cluster-tolerance capacity [12], instead of combining the 50 individual vectors ($\xi = 0,1,2,\ldots,49$) into one ($20 + 40 \times 49$) = 1980-D vector, Chou and Shen [9] introduced an ensemble classifier by fusing the results obtained based on each of the individual vector descriptors through a voting system.

Below, we will introduce a different approach to cope with the problem caused by the high dimensional descriptor.

2.3 PCA algorithm

Since the Pse-PSSM descriptor is high-dimensional, the operations on this representation are computationally expensive. Therefore, subspace learning (dimensionality reduction) to detect reduced intrinsic dimensionality in the high-dimensional feature space is necessary for predicting the quaternary structure of proteins.

The PCA is one of the most popular linear subspace methods that can extract useful features with low computational complexity [13–17]. It is based on computation of low-dimensional representation of a high-dimensional dataset that maximizes the total scatter, which is optimal in reconstruction, as depicted below.

Given a dataset of $\boldsymbol{X} = [\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N]$ in an m-dimensional real space R^m, where $\vec{x}_i$, $i = 1,2,\ldots,N$, belongs to C classes denoted as $[\Phi_1, \Phi_2, \ldots, \Phi_C]$. The objective of the PCA is to find a transformation matrix $\boldsymbol{A}$ to map $\boldsymbol{X}$ into

$\boldsymbol{Y}$ in a new d-dimensional real space R^d (where $d < m$), i.e.,

$$\boldsymbol{Y} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{X}, \tag{5}$$

where $\boldsymbol{Y} = [\vec{y}_1, \vec{y}_2, \ldots, \vec{y}_N]$ and $\vec{y}_i$, $i = 1,2,\ldots,N$, are corresponding vectors in the new space. To obtain an optimized transformation matrix $\boldsymbol{A}$, we need to maximize the following objective function:

$$\begin{aligned}
\boldsymbol{A}_{\mathrm{opt}} &= \arg\max_{\boldsymbol{A}} \sum_{i=1}^{N} \left\| \overrightarrow{y_i} - \overrightarrow{\overline{y}} \right\|^2 \\
&= \arg\max_{\boldsymbol{A}} \sum_{i=1}^{N} \left\| \boldsymbol{A}^{\mathrm{T}} \left(\overrightarrow{x_i} - \overrightarrow{\overline{x}} \right) \right\|^2 \\
&= \arg\max_{\boldsymbol{A}} \sum_{i=1}^{N} \mathrm{tr}\left[\boldsymbol{A}^{\mathrm{T}} \left(\overrightarrow{x_i} - \overrightarrow{\overline{x}} \right) \left(\overrightarrow{x_i} - \overrightarrow{\overline{x}} \right)^{\mathrm{T}} \boldsymbol{A} \right] \\
&= \arg\max_{\boldsymbol{A}} \mathrm{tr}\left(\boldsymbol{A}^{\mathrm{T}} \boldsymbol{S}_{\mathrm{T}} \boldsymbol{A} \right),
\end{aligned} \tag{6}$$

where $\overrightarrow{\overline{y}} = (1/N)\sum \overrightarrow{y}_i$, $\overrightarrow{\overline{x}} = (1/N)\sum \overrightarrow{x}_i$ and $\boldsymbol{S}_{\mathrm{T}}$ is the total scatter matrix with $\boldsymbol{S}_{\mathrm{T}} = \sum_{i=1}^{N} \left(\overrightarrow{x_i} - \overrightarrow{\overline{x}} \right) \left(\overrightarrow{x_i} - \overrightarrow{\overline{x}} \right)^{\mathrm{T}}$. The optimal transformation matrix of the PCA is the set of m-dimensional eigenvectors of $\boldsymbol{S}_{\mathrm{T}}$ corresponding to the d largest eigenvalues.

3 Results and discussion

In statistical prediction, the sub-sampling test and the jackknife test are two cross-validation methods often used for examining the accuracy of a predictor. However, as demonstrated by Eq. (50) in a recent comprehensive review [18], the sub-sampling (e.g., 5-fold cross-validation) test cannot avoid arbitrariness even for a simple benchmark dataset. Accordingly, the jackknife test has been widely adopted by investigators [19–24] to test the power of predictors. Therefore, we also used the jackknife test to examine the performance of the PCA in predicting the quaternary structure of proteins.

As discussed before, according to the Pse-PSSM descriptor (Eq. (1)), a protein can be represented by a combination of one 20-D vector $\boldsymbol{P}^0_{\text{Pse-PSSM}}$ and 49 different 40-D vectors $\boldsymbol{P}^{\xi}_{\text{Pse-PSSM}}$, $\xi = 1,2,\ldots,49$. However, of the 49 different 40-D vectors, the first 20 components in $\boldsymbol{P}^{\xi}_{\text{Pse-PSSM}}$ and the 20 components in $\boldsymbol{P}^0_{\text{Pse-PSSM}}$ are actually the same as the 20 components in Eq. (2). Accordingly, in merging the 50 vectors into one as denoted by $\boldsymbol{P}_{\text{Pse-PSSM}}$, the 20 components only need to be used once. Thus, $\boldsymbol{P}_{\text{Pse-PSSM}}$ is actually corresponding to a $20 \times 50 = 1000$-D vector.

By using the PCA algorithm formulated in Sect. 2 to extract the most important features from the 1000-D vector of $\boldsymbol{P}_{\text{Pse-PSSM}}$, as denoted by $\boldsymbol{P}^*_{\text{Pse-PSSM}}$. Figure 2 shows the relations between the number of the principal components and data variance for the dataset consisting of 8 quaternary structures of proteins. As can be seen from Fig. 2, the higher the kept variance is, the higher the dimension of $\boldsymbol{P}^*_{\text{Pse-PSSM}}$ will be. To preserve energy and at the same time extract the most important features, 95% of the total data variance is applied in the PCA, which yields the 128-D $\boldsymbol{P}^*_{\text{Pse-PSSM}}$. Subsequently, the K-nearest neighbor (KNN) algorithm [25–27] was used to predict the quaternary structure types based on the 128-D vector descriptor.

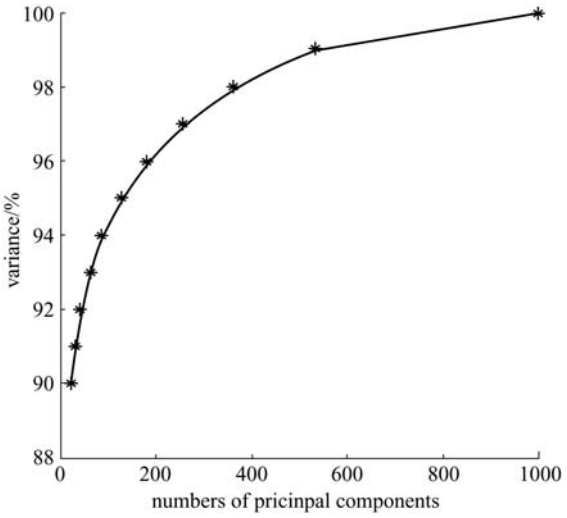

Fig. 2 Relation between principal components and data variance for 8 quaternary structures of proteins

The detailed jackknife success rate for each of the 8 quaternary structures by using 128-D $\boldsymbol{P}^*_{\text{Pse-PSSM}}$ and 1000-D $\boldsymbol{P}_{\text{Pse-PSSM}}$ as the protein descriptors are shown in Table 1. It can be observed that the success rates by using the PCA features extraction algorithm are higher than those without using the PCA algorithm. It is important to point out that in most cases the performance of the KNN algorithm depends on the selection of the number of the nearest neighbors k and it is found that $k = 1$ gives the best prediction accuracy in this study.

4 Conclusions

To classify complicated biological systems, the high-dimensional vectors problem has to be resolved. The PCA algorithm is adopted to extract key information from the high-dimensional space and reduce the original high-dimensional vector to a lower-dimensional one. The PCA is used to effectively find important information from the high dimensional data space. Its application to quaternary structures of proteins prediction just demonstrates its advantages. Also, the PCA approach can be used to deal with other complicated biological systems.

Table 1 Jackknife success rates for each of 8 quaternary structures of proteins obtained by using original Pse-PSSM 1000-D descriptor $\boldsymbol{P}_{\text{Pse-PSSM}}$ and dimension-reduced 128-D descriptor $\boldsymbol{P}^{*}_{\text{Pse-PSSM}}$ to represent protein samples

quaternary structures of proteins	original Pse-PSSM 1000-D descriptor $\boldsymbol{P}_{\text{Pse-PSSM}}$/%	dimension-reduced 128-D descriptor $\boldsymbol{P}^{*}_{\text{Pse-PSSM}}$/%
monomer	$\frac{664}{1073}=61.88$	$\frac{602}{1073}=56.1$
homodimer	$\frac{1104}{1681}=65.68$	$\frac{1122}{1681}=66.75$
homotrimer	$\frac{96}{224}=42.86$	$\frac{132}{224}=58.93$
homotetramer	$\frac{264}{549}=48.09$	$\frac{329}{549}=59.93$
homopentamer	$\frac{3}{17}=17.65$	$\frac{4}{17}=23.53$
homohexamer	$\frac{24}{115}=20.87$	$\frac{44}{115}=38.26$
homooctamer	$\frac{17}{46}=36.96$	$\frac{24}{46}=52.17$
homododecamer	$\frac{11}{21}=52.38$	$\frac{11}{21}=52.38$
overall	$\frac{2183}{3726}=58.59$	$\frac{2268}{3726}=60.87$

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 60704047).

References

1. Klotz I M, Langerman N R, Darnall D W. Quaternary structure of proteins. Annual Review of Biochemistry, 1970, 39: 25–62
2. Price N C. Assembly of multi-subunit structures. In: Pain R H, ed. Mechanisms of Protein Folding. New York: Oxford University Press, 1994, 160–193
3. Chou K C, Cai Y D. Predicting protein quaternary structure by pseudo-amino acid composition. Proteins, 2003, 53(2): 282–289
4. Anfinsen C B. Principles that govern the folding of protein chains. Science, 1973, 181(96): 223–230
5. Anfinsen C B, Haber E, Sela M, et al. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proceedings of the National Academy of Sciences of the United States of America, 1961, 47: 1309–1314
6. Garian R. Prediction of quaternary structure from primary structure. Bioinformatics, 2001, 17(6): 551–556
7. Chou K C. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins, 2001, 43(3): 246–255
8. Shen H B, Chou K C. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Analytical Biochemistry, 2008, 373(2): 386–388
9. Chou K C, Shen H B. MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochemical and Biophysical Research Communications, 2007, 360(2): 339–345
10. Wang G, Dunbrack R L Jr. PISCES: a protein sequence culling server. Bioinformatics, 2003, 19(12): 1589–1591
11. Schäffer A A, Aravind L, Madden T L, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Research, 2001, 29(14): 2994–3005
12. Chou K C. A key driving force in determination of protein structural classes. Biochemical and Biophysical Research Communications, 1999, 264(1): 216–224
13. Malinowski E R, Howery D G. Factor Analysis in Chemistry. New York: John Wiley, 1980
14. Deming S N. Chemometrics: an overview. Clinical Chemistry, 1986, 32(9): 1702–1706
15. Du Q S, Jiang Z Q, He W Z, et al. Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. Journal of Biomolecular Structure and Dynamics, 2006, 23(6): 635–640
16. Wen Y, Lu Y, Shi P F. Handwritten Bangla numeral recognition system and its application to postal automation. Pattern Recognition, 2007, 40(1): 99–107
17. Liang Z Z, Zhang D, Shi P F. The theoretical analysis of GLRAM and its applications. Pattern Recognition, 2007, 40(3): 1032–1041
18. Chou K C, Shen H B. Recent progress in protein subcellular location prediction. Analytical Biochemistry, 2007, 370(1): 1–16
19. Du P, He T, Li Y. Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features. Biochemical and Biophysical Research Communications, 2007, 358(1): 336–341
20. Du P, Li Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics, 2006, 7: 518
21. Huang Y, Cai J, Ji L, et al. Classifying G-protein coupled receptors with bagging classification tree. Computation Biology and Chemistry, 2004, 28(4): 275–280
22. Wang M, Yang J, Liu G P, et al. Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. Protein Engineering Design and Selection, 2004, 17(6): 509–516
23. Shen H B, Chou K C. Ensemble classifier for protein fold pattern recognition. Bioinformatics, 2006, 22(14): 1717–1722
24. Wang S Q, Yang J, Chou K C. Using stacked generalization to predict membrane protein types based on pseudo-amino acid composition. Journal of Theoretical Biology, 2006, 242(4): 941–946
25. Chou K C, Shen H B. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. Journal of Proteome Research, 2006, 5(8): 1888–1897
26. Denoeux T. A k-nearest neighbor classification rule based on Dempster-Shafer theory. IEEE Transactions on Systems, Man, and Cybernetics, 1995, 25(5): 804–813
27. Keller J M, Gray M R, Givens J A Jr. A fuzzy K-nearest neighbor algorithm. IEEE Transactions on Systems, Man, and Cybernetics, 1985, 15(4): 580–585