Protein Engineering, Design & Selection vol. 22 no. 11 pp. 699–705, 2009
Published online September 22, 2009 doi:10.1093/protein/gzp057

GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis

Wei-Zhong Lin[1], Xuan Xiao[1,3] and Kuo-Chen Chou[2]

[1]Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333001, China and [2]Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, CA 92130, USA

[3]To whom correspondence should be addressed.
E-mail: xiaoxuan0326@yahoo.com.cn

G-protein-coupled receptors (GPCRs) play fundamental roles in regulating various physiological processes as well as the activity of virtually all cells. Different GPCR families are responsible for different functions. With the avalanche of protein sequences generated in the postgenomic age, it is highly desired to develop an automated method to address the two problems: given the sequence of a query protein, can we identify whether it is a GPCR? If it is, what family class does it belong to? Here, a two-layer ensemble classifier called GPCR-GIA was proposed by introducing a novel scale called 'grey incident degree'. The overall success rate by GPCR-GIA in identifying GPCR and non-GPCR was about 95%, and that in identifying the GPCRs among their nine family classes was about 80%. These rates were obtained by the jackknife cross-validation tests on the stringent benchmark data sets where none of the proteins has ≥50% pairwise sequence identity to any other in a same class. Moreover, a user-friendly web-server was established at http://218.65.61.89:8080/bioinfo/GPCR-GIA. For user's convenience, a step-by-step guide on how to use the GPCR-GIA web server is provided. Generally speaking, one can get the desired two-level results in around 10 s for a query protein sequence of 300–400 amino acids; the longer the sequence is, the more time that is needed.

Keywords: ensemble classifier/fusion/K nearest neighbor algorithm/pseudo amino acid composition/web server

Introduction

G-protein-coupled receptors (GPCRs) are seven-helix transmembrane proteins that provide a molecular link between extracellular signals and intracellular reactions ranging from cell–cell communication processes to physiological responses (Heuss and Gerber, 2000; Milligan and White, 2001; Hall and Lefkowitz, 2002; Chou, 2005a). They are among the largest and most diverse protein families in mammalian genomes. Owing to their close relevance to a variety of diseases, such as cancer, diabetes, neurodegenerative, inflammatory and respiratory disorders, GPCRs are of utmost interest in drug development: over half of all prescription drugs currently on the market act by targeting these receptors directly or indirectly.

Many efforts have been invested in studying GPCR by both academic institutions and pharmaceutical industries. However, as membrane proteins, GPCRs are very difficult to crystallize and most of them will not dissolve in normal solvents. Accordingly, so far, very few crystal GPCR structures have been determined. Although the recently developed state-of-the-art NMR technique is a very powerful tool in determining the three-dimensional structures of membrane proteins (Oxenoid and Chou, 2005; Call *et al.*, 2006; Douglas *et al.*, 2007; Schnell and Chou, 2008), it is time-consuming and costly. Although some membrane protein structures can be derived with homology approaches (Chou, 2004), the number of templates for transmembrane proteins is very limited. In contrast, more than thousand GPCR sequences are known, and much more are expected to come in the near future. In view of this, it would be very useful to develop a computational method which can predict the classification of the families and subfamilies of GPCRs based on their primary sequences.

In a pioneer study (Chou and Elrod, 2002), Chou and Elrod attempted to identify the subfamily classes of the rhodopsin-like GPCR family by using the covariant-discriminant algorithm (Chou and Elrod, 1999). With more data available later, the study was extended to identify the main family classes of GPCRs (Chou, 2005b) with a similar approach. Stimulated by the encouraged results, some follow-up studies were conducted by using various different approaches as reported in Bhasin and Raghava (2005), Gao and Wang (2006) and Wen *et al.* (2007).

Although considerable progresses have been achieved during the past 6 years in this area, further studies are needed due to the following reasons. First, the data sets constructed to train the existing predictors cover very limited GPCR family classes. With the development of protein databases, more classes should be included to enhance the coverage scope for practical usage. Secondly, the reported success rates were derived based on a benchmark data set without being rigorously screened by a clear data-culling operation to avoid redundancy and homologous bias, and hence those reported success rates therein might be overestimated. As is well known, the more the family classes covered, the lower the odds are in getting a correct prediction. Also, the more stringent the benchmark data set in excluding homologous sequences, the harder it becomes to get a high success rate for cross-validation test (Xiao *et al.*, 2005; Chou and Shen, 2007c; Chou and Shen, 2008). The present study was devoted to address these problems by developing a new GPCR predictor. Moreover, a user-friendly web server, called GPCR-GIA, was designed for the new predictor. For the convenience of most experimental scientists who wish to utilize the predictor to generate the desired data but feel difficult to follow the detailed mathematics and processes, a step-by-step guide on how to use the web server predictor was provided.

Materials

Protein sequences were collected from the Swiss-Prot database release 54.8 of 05 February 2008 at http://www.ebi.ac.uk/swissprot/ by the 'UniProt Power Search'. To construct a higher quality benchmark data set with a wider coverage scope and lower homology bias, the data were screened strictly according to the following criteria. (i) Included were those with clear experimental annotations as hit by one of the key words listed in Table I where the corresponding functional features are also given; sequences annotated with ambiguous or uncertain terms, such as 'potential', 'probable', 'probably', 'maybe' or 'by similarity', were excluded. (ii) Sequences annotated with 'fragment' were excluded; also, sequences with less than 50 amino acid (AA) residues were removed because they might just be fragments. (iii) To reduce homology bias, a redundancy cutoff was operated to winnow those sequences which have $\geq$50% pairwise sequence identity to any others in a same family class. However, such a redundancy cutoff procedure was waived for those classes containing less than 20 sequences; otherwise, the samples left would be too few to have any statistical significance.

After strictly following the above procedures, we finally obtained 780 GPCRs, which are distributed among the nine GPCR family classes (Table II). Accordingly, the data set, $\mathbb{S}$, thus obtained is a union of the nine subsets as formulated below:

$$\mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \mathbb{S}_3 \cup \mathbb{S}_4 \cup \mathbb{S}_5 \cup \mathbb{S}_6 \cup \mathbb{S}_7 \cup \mathbb{S}_8 \cup \mathbb{S}_9 \quad (1)$$

The accession number and sequence for each of the proteins in the nine subsets of the GPCR benchmark data sets are given in the Supplementary Information A available at *PEDS* online. Meanwhile, in order for training a statistical predictor to distinguish GPCR proteins from non-GPCR proteins, a non-GPCR benchmark data set $\bar{\mathbb{S}}$ was also constructed by randomly collecting 780 non-GPCR proteins from the Swiss-Prot database at http://www.ebi.ac.uk/swissprot/ according to their annotations in the CC (comment or notes) and ID (identification) fields. The corresponding accession numbers and sequences are given in the Supplementary Information B available at *PEDS* online, in which none of the entries has $\geq$50% pairwise sequence identity to any other.

Table I. Keywords used to search the Swiss-Prot database for G-protein coupled receptors and their functional features (Matsunami *et al.*, 2000; Hill *et al.*, 2002; Otaki and Yamamoto, 2003)

Class index	Keyword	Feature
1	G-protein coupled receptor 1 family	Rhodopsin-like receptor
2	G-protein coupled receptor 2 family	Peptide hormones receptor
3	G-protein coupled receptor 3 family	Glutamate and calcium receptor
4	G-protein coupled receptor 4 family	Fungal mating pheromone receptor
5	G-protein coupled receptor 5 family	Cyclic AMP receptor
6	G-protein coupled receptor 6 family	Odorant receptors in drosophila
7	G-protein coupled receptor 7 family	Gustatory receptor of drosophila
8	G-protein coupled receptor Fz/Smo family	Frizzled/smoothened family
9	G-protein coupled receptor T2R family	T2R family in mammals

Table II. Breakdown of the G-protein coupled receptors obtained by following steps (1)–(3) and Eq. (3) in Materials

Subset index	GPCRs family	The number of protein sequence		
		Original data set $\mathbb{S}^0$	Learning data set $\mathbb{S}^\mathrm{L}$	Testing data set $\mathbb{S}^\mathrm{T}$
1	Rhodopsin-like receptor	540	144	396
2	Peptide hormones receptor	75	60	15
3	Glutamate and calcium receptor	25	20	5
4	Fungal mating pheromone receptor	12	11	1
5	Cyclic AMP receptor	4	4	0
6	Odorant receptors in drosophila	56	45	11
7	Gustatory receptor of drosophila	21	17	4
8	Frizzled/Smoothened family	16	14	2
9	T2R family in mammals	31	25	6
Overall		780	340	440

On the basis of data set $\mathbb{S}$ [see Eq. (1) and Table II], two working data sets, i.e. a learning data set $\mathbb{S}^\mathrm{L}$ and an independent testing data set $\mathbb{S}^\mathrm{T}$, are constructed. In order to fully use the data in $\mathbb{S}$ and meanwhile guarantee that $\mathbb{S}^\mathrm{L}$ and $\mathbb{S}^\mathrm{T}$ be completely independent of each other, the following condition is imposed:

$$\mathbb{S}^\mathrm{L} \cup \mathbb{S}^\mathrm{T} = \mathbb{S} \ \text{ and } \ \mathbb{S}^\mathrm{L} \cap \mathbb{S}^\mathrm{T} = \varnothing \qquad (2)$$

where $\cup$, $\cap$ and $\varnothing$ represent the symbols for 'union', 'intersection', and 'empty set' in the set theory, respectively. To avoid the situation that the numbers of proteins in some subsets of the learning data set $\mathbb{S}^\mathrm{L}$ might overwhelm those of the others, the 'bracket percentage distribution' procedure (Chou and Shen, 2006) was used to randomly assign the protein samples to the corresponding subsets of $\mathbb{S}^\mathrm{L}$ and $\mathbb{S}^\mathrm{T}$, as formulated below:

$$\begin{cases} n_i^\mathrm{L} = 100 + \mathrm{INT}\{(n_i^0 - 100) \times 0.1\}, \ \text{if } n_t^0 \geq 100 \\ n_i^\mathrm{L} = \mathrm{INT}\{n_i^0 \times 0.8\}, \ \text{if } 20 \leq n_i^0 < 100 \\ n_i^\mathrm{L} = \mathrm{INT}\{n_i^0 \times 0.9\}, \ \text{if } 10 \leq n_i^0 < 20 \\ n_i^\mathrm{L} = n_i^0, \ \text{if } n_i^0 < 10 \\ n_i^\mathrm{T} = n_i^0 - n_i^\mathrm{L}. \end{cases} \qquad (3)$$

where n_i^0, n_i^L and n_i^T, are the numbers of protein samples in the *i*th subset of the original data set $\mathbb{S}$, learning data set $\mathbb{S}^\mathrm{L}$ and testing data set $\mathbb{S}^\mathrm{T}$, respectively, and the symbol INT is the 'integer truncation operator' meaning to take the integer part for the number in the brackets right after it. The numbers of proteins thus obtained for the nine GPCR family classes in the learning data set $\mathbb{S}^\mathrm{L}$ and testing data set $\mathbb{S}^\mathrm{T}$ are given in Table II. The accession numbers and sequences for the corresponding proteins in the learning and testing data sets are given in the Supplementary Information C and Supplementary Information D available at *PEDS* online, respectively.

Method

Pseudo amino acid composition (PseAAC)

To develop a statistical method for predicting the attributes of proteins, an indispensable thing is to find an effective formulation to represent the protein samples concerned. Two kinds of models are often used to formulate protein samples. One is the sequential model, and the other the discrete model. In the sequential model, the sample of a protein is represented by its AA sequence, and the sequence similarity search-based tools such as BLAST (Altschul *et al.*, 1997) are used to perform prediction. However, this kind of approach failed to work when a query protein did not have significant homology to character-known proteins. Thus, various discrete models were introduced by representing the sample of a protein with a set of discrete numbers. The early-stage discrete model was to represent the sample of a protein with its AA composition or AAC [see, e.g. (Nakashima *et al.*, 1986)]. However, in the AAC model, all the sequence-order information is lost. To avoid totally miss the sequence-order information, the pseudo amino acid (PseAA) composition or PseAAC was introduced (Chou, 2001). The PseAAC approach can be used to formulate a protein sample with a discrete model yet without completely losing its sequence-order information, and have been widely used by investigators to study various problems in proteins and protein-related systems, such as protein structural class (Chen *et al.*, 2006a, b; Lin and Li, 2007b; Xiao *et al.*, 2006, 2008a, b; Zhang *et al.*, 2008b), protein secondary structure content (Chen *et al.*, 2009), protein fold pattern (Shen and Chou, 2006, 2009), protein quaternary structure attribute (Chou and Cai, 2003; Xiao *et al.*, 2009b), classification of AAs (Georgiou *et al.*, 2009), GPCR type (Qiu *et al.*, 2009), protein subcellular localization (Pan *et al.*, 2003; Xiao *et al.*, 2005; Chen and Li, 2007a; Chou and Shen, 2007a, 2008; Li and Li, 2008), protein subnuclear localization (Shen and Chou, 2005a; Mundra *et al.*, 2007), apoptosis protein subcellular localization (Chen and Li, 2007a, b; Ding and Zhang, 2008; Jiang *et al.*, 2008; Lin *et al.*, 2009), protein submitochondria localization (Du and Li, 2006; Zeng *et al.*, 2009), membrane protein type (Liu *et al.*, 2005; Shen and Chou, 2005b; Shen *et al.*, 2006; Wang *et al.*, 2006; Chou and Shen, 2007b; Lin, 2008), enzyme functional classification (Chou and Cai, 2004; Cai *et al.*, 2005; Chou, 2005c; Shen and Chou, 2007a; Zhou *et al.*, 2007; Ding *et al.*, 2009), cofactors of oxidoreductases (Zhang and Fang, 2008), lipase type (Zhang *et al.*, 2008a), conotoxin superfamily classification (Mondal *et al.*, 2006; Lin and Li, 2007a), protein–protein interactions (Chou and Cai, 2006), signal peptide (Chou and Shen, 2007d; Shen and Chou, 2007b) and other protein-related systems (Gonzalez-Diaz *et al.*, 2007, 2008a, b).

According to the concept of PseAAC, a protein sequence containing L AAs can be formulated as

$$\mathbf{P} = [p_1,\ p_2, \ldots, p_{20},\ p_{20+1}, \ldots, p_{20+\lambda}]^{\mathbf{T}}, \quad (\lambda < L) \tag{4}$$

where the $20 + \lambda$ components are given by

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda}\tau_k}, & (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{k=1}^{\lambda}\tau_k}, & (20+1 \le u \le 20+\lambda) \end{cases} \tag{5}$$

where w is the weight factor which was set at 0.05 in (Chou, 2001), and τ_k the kth tier correlation factor that reflects the sequence order correlation between all the kth most contiguous residues along the protein sequence (Chou, 2001), as formulated by

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} \mathrm{J}_{i,i+k}, \quad k < L \tag{6}$$

with

$$\mathrm{J}_{i,i+k} = \frac{1}{3}\Big\{[H_1(\mathrm{R}_{i+k}) - H_1(\mathrm{R}_i)]^2 + [H_2(\mathrm{R}_{i+k}) - H_2(\mathrm{R}_i)]^2 + [M(\mathrm{R}_{i+k}) - M(\mathrm{R}_i)]^2\Big\} \tag{7}$$

where H_1 (R_i), H_2 (R_i) and M (R_i) are, respectively, the hydrophobicity value, hydrophilicity value and side-chain mass for the AA residue R_i and so forth. Note that before substituting the values of hydrophobicity, hydrophilicity and side-chain mass into Eq. (7), they are all subjected to a *standard conversion* as described by the following equation:

$$\begin{cases} H_1(\mathrm{R}_i) = \dfrac{H_1^0(\mathrm{R}_i) - < H_1^0 >}{\mathrm{SD}(H_1^0)} \\ H_2(\mathrm{R}_i) = \dfrac{H_2^0(\mathrm{R}_i) - < H_2^0 >}{\mathrm{SD}(H_2^0)} \\ M(\mathrm{R}_i) = \dfrac{M^0(\mathrm{R}_i) - < M^0 >}{\mathrm{SD}(M^0)} \end{cases} \tag{8}$$

where the symbols H_1^0 (R_i) and H_2^0 (R_i) are the original hydrophobicity and hydrophilicity values for R_i that can be obtained from Tanford (1962) and Hopp and Woods (1981), respectively, and M^0 (R_i) the mass of the side chain for R_i that can be found from any biochemistry text book. In Eq. (8), the symbol $< >$ means taking the average of the quantity therein over 20 native AAs, and SD means the corresponding standard deviation. The converted values obtained by Eq. (8) will have a zero mean value over the 20 native AAs, and will remain unchanged if going through the same conversion procedure again. As we can see, the first 20 components in Eq. (4), i.e. p_1, $p_2, \ldots$, p_{20}, are associated with the conventional AA composition of **P**, while the remaining components p_{20+1}, $p_{20+2}, \ldots$, $p_{20+\lambda}$ are the λ correlation factors that reflect the first tier, second tier, ..., and the λ-th tier sequence order correlation patterns (Chou, 2001). It is these additional λ factors that approximately incorporate the sequence-order effects. Note that λ is a parameter of integer [see Eq. (4)] and that choosing a different integer for λ will lead to a dimension-different PseAAC, as will be further discussed later. Note that the PseAAC derived from Eqs. (4–8) and utilized in the current study is just one of many different PseAAC modes. Since the concept of PseAAC has been widely used in dealing with various protein-related problems, recently a web-server called 'PseAAC' (Shen and Chou, 2008) was established at http://chou.med.harvard.edu/bioinf/PseAAC/ by which the user can generate 63 different modes of PseAAC.

K nearest neighbor (KNN) classifier with grey incidence analysis (GIA)

The K-nearest neighbor (KNN) classifier is quite popular in pattern recognition community owing to its good performance and simple-to-use feature. According to the KNN rule (Cover and Hart, 1967; Keller *et al.*, 1985; Denoeux, 1995), also named the 'voting KNN rule', the query protein should be assigned to the subset represented by a majority of its KNNs. There are many different definitions to measure the 'nearness' for the KNN classifier, such as Euclidean distance, Hamming distance (Mardia *et al.*, 1979) and Mahalanobis distance (Mahalanobis, 1936; Pillai, 1985).

Here, we shall introduce a novel scale, the so-called grey incidence degree or grey incidence analysis, to measure the 'nearness' for the KNN classifier. In 1982, Deng (1985) proposed a grey system theory to study the uncertainty of a system. According to the concept of the theory, if the information of a system investigated is fully known, it is called a 'white system'; if completely unknown, a 'black system'; if partially known, a 'grey system'. The approach is particularly useful in coping with complicated systems with uncertainty or insufficient training data [see, e.g. (Xiao *et al.*, 2008a)]. As one of the major components of the grey systems theory, the grey incidence degree can be formulated as follows (Liu *et al.*, 2006).

Suppose $\mathbf{P}^{\mathrm{q}}$ is a query protein, and $\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N\}$ the N proteins in a benchmark data set classified into M subsets; i.e.

$$\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_{\mathrm{N}}\} \in \bigcup_{m=1}^{M} S_m \tag{9}$$

where $\in$ and $\cup$ are the symbols in the set theory, meaning 'member of' and union, respectively. Each subset S_m ($m =$ 1,2,..., M) is composed of proteins with the same type and its size (the number of proteins therein) is n_m. Thus, according to Eq. (4), the query protein $\mathbf{P}^{\mathrm{q}}$ and the ith protein in the training data set $\{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_N\}$ can be expressed as

$$\mathbf{P}^{\mathrm{q}} = \left[p_1^{\mathrm{q}}, p_2^{\mathrm{q}}, \ldots, p_{20}^{\mathrm{q}}, p_{20+1}^{\mathrm{q}}, \ldots, p_{20+\lambda}^{\mathrm{q}}\right]^{\mathbf{T}} \tag{10}$$

and

$$\mathbf{P}_i = \left[p_1^i, p_2^i, \ldots, p_{20}^i, p_{20+1}^i, \ldots, p_{20+\lambda}^i\right]^{\mathbf{T}} \tag{11}$$

The grey relational coefficient, $\gamma\ (p_{\mathrm{k}}^{\mathrm{q}}, p_k^i)$, between p_k^{q} and p_k^i, is defined as

$$\gamma\left(p_k^{\mathrm{q}}, p_k^i\right) = \frac{\Delta_{\mathrm{Min}} + \xi\Delta_{\mathrm{Max}}}{\Delta_k^{\mathrm{q},i} + \xi\Delta_{\mathrm{Max}}} \tag{12}$$

where

$$\Delta_k^{\mathrm{q},i} = \left|p_k^{\mathrm{q}} - p_k^i\right| \tag{13}$$

$$\Delta_{\mathrm{Max}} = \underset{i,k}{\mathrm{Max}}\left|p_k^{\mathrm{q}} - p_k^i\right|, \quad (j = 1, 2, \ldots, N;\ k = 1, 2, \ldots, 20 + \lambda) \tag{14}$$

$$\Delta_{\mathrm{Min}} = \underset{i,k}{\mathrm{Min}}\left|p_k^{\mathrm{q}} - p_k^i\right|, \quad (j = 1, 2, \ldots, N;\ k = 1, 2, \ldots, 20 + \lambda) \tag{15}$$

$$\xi = \text{distinguishing coefficient}, \in [0, 1] \tag{16}$$

The grey incidence degree, $\Gamma(\mathbf{P}^{\mathrm{q}}, \mathbf{P}_i)$, between the query protein $\mathbf{P}^{\mathrm{q}}$ and the ith protein $\mathbf{P}^i$ in the training data set is actually a weighted sum of the grey relational coefficients and can be formulated as

$$\Gamma(\mathbf{P}^{\mathrm{q}}, \mathbf{P}_i) = \sum_{k}^{20+\lambda} w_k \gamma\left(p_{k0}^{\mathrm{q}},\ p_{\mathrm{k}i}^i\right) \tag{17}$$

where the weighting factor, w_k, must satisfy $\sum_{k=1}^{20+\lambda} w_k = 1$. In this study, the distinguishing coefficient ξ is equal to 0.5 and the grey incidence degree is performed in equal-weighted mode, i.e. $w_k = 1/(20 + \lambda)\ \ (k = 1, 2, \ldots, 20 + \lambda)$.

The grey incidence degree $\Gamma(\mathbf{P}^{\mathrm{q}}, \mathbf{P}_i)$ as defined in Eq. (17) stands for the level of similarity between the query protein $\mathbf{P}^{\mathrm{q}}$ and the ith protein $\mathbf{P}^i$ in the training data set. As we can see from Eqs. (12–16), when $\mathbf{P}^{\mathrm{q}} \equiv \mathbf{P}_i$, we have $\Gamma(\mathbf{P}^{\mathrm{q}}, \mathbf{P}_i) =$ 1, meaning that the two proteins have perfect or 100% similarity.

Fusion of different λ in PseAAC and K in KNN classifier

As mentioned earlier, the PseAAC discrete model contains a parameter λ, which is associated with the number of components in a protein representation [Eq. (4), (10 or (11)]. Generally speaking, the larger the λ, the more components the PseAAC contains, and hence the more information the representation bears. However, λ must be smaller than the number of the AAs in the protein concerned [cf. Eq. (4)]. Also, it will reduce the cluster-tolerant capacity (Chou and Shen, 2007c) if the PseAAC contains too many components, so as to lower down the success rate of cross validation. Accordingly, for a given training data set, there is an optimal number for λ. It would be time-consuming and tedious to find the optimal λ by changing its value and doing tests one-by-one.

Likewise, the result identified by the KNN classifier mentioned in the last section may depend on the parameter K, the number of the nearest neighbors to the query protein $\mathbf{P}^{\mathrm{q}}$. In other words, for a given training data set, there is an optimal value for K as well.

It would be much more tedious and time-consuming to determine the optimal values for two uncertain parameters. To solve the problem, let us adopt the two-D (dimensional) fusion approach, as described below.

As mentioned in the Materials section, the shortest protein sequence considered is 50 AAs and hence we can set the maximum value for λ is 49 [Eq. (4)]. Also, for the current benchmark data set, when K > 9 the success rate by KNN classifier would remarkably decrease. Therefore, the results for the query protein $\mathbf{P}^{\mathrm{q}}$ identified with different λ of PseAAC and different K of KNN can be generally expressed by

$$C_{\lambda,\mathrm{K}}\left(\mathbf{P}^{\mathrm{q}}\right) \in \bigcup_{m=1}^{M} S_m(\lambda = 0,1,2,\ldots,49; \quad \mathrm{K} = 1,2,\ldots,9) \tag{18}$$

The voting score for the query protein $\mathbf{P}^{\mathrm{q}}$ belonging to the mth subset $S_m \in \mathbb{S}$ is defined by

$$Y_m(\mathbf{P}^{\mathrm{q}}) = \sum_{\lambda=0}^{49} \sum_{\mathrm{K}=1}^{9} w_{\lambda,\mathrm{K}} \Delta\left[C_{\lambda,\mathrm{K}}\left(\mathbf{P}^{\mathrm{q}}\right), S_m\right], (m = 1,2,\ \ldots,M) \tag{19}$$

where $w_{\lambda,\mathrm{K}}$ is the weight and was set at 1 for simplicity, the delta function in Eq. (15) is given by

$$\Delta\left[C_{\lambda,\mathrm{K}}\left(\mathbf{P}^{\mathrm{q}}\right), S_m\right] = \begin{cases} 1, & \text{if} C_{\lambda,\mathrm{K}}\left(\mathbf{P}^{\mathrm{q}}\right) \in S_m \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

thus the query protein $\mathbf{P}^{\mathrm{q}}$ is predicted belonging to the subset with which its score of Eq. (19) is the highest; i.e. the query protein $\mathbf{P}^{\mathrm{q}}$ is classified as the μth subset if

$$\mu = \mathbf{argmax}_m\left\{Y_m\left(\mathbf{P}^{\mathrm{q}}\right)\right\}, \quad (m = 1,2,\ \ldots,M) \tag{21}$$

where μ is the argument of m that maximize the score function Y_m of Eq. (19). If there are two and more arguments leading to a same maximum value, the query protein will be randomly assigned to one of the subcellular locations associated with these arguments although this kind of tie case rarely happens.

The predictor thus formed is called GPCR-GIA, where GIA stands for grey incidence analysis.

Results and discussion

Now let us demonstrate the prediction quality by using GPCR-GIA predictor. In statistical prediction, the independent data set test, sub-sampling test and jackknife test are often used in literatures for examining the accuracy of a predictor (Chou and Zhang, 1995). However, as elucidated in Chou and Shen (2008) and demonstrated by Eq. (50) of Chou and Shen (2007c), among the three cross-validation methods, the jackknife test is deemed the most objective that can always yield a unique result for a given benchmark data set, and hence has been increasingly and widely used by investigators to examine the accuracy of various predictors [see, e.g. (Zhou, 1998; Zhou and Assa-Munt, 2001; Zhou and Doctor, 2003; Zhou and Cai, 2006; Ding and Zhang, 2008; Li and Li, 2008; Lin, 2008; Lin *et al.*, 2008; Zhang *et al.*, 2008a; Chen *et al.*, 2009; Chou and Shen, 2009; Ding *et al.*, 2009; Georgiou *et al.*, 2009; Shen *et al.*, 2009; Xiao *et al.*, 2009a; Zeng *et al.*, 2009)].

The jackknife cross-validation results obtained by GPCR-GIA in identifying the proteins as GPCR or non-GPCR are given in Table III, from which we can see that the overall success rate was 95.4%. The corresponding results in identifying the GPCR proteins among their nine family classes are given in Table IV, from which we can see that the overall success rate was 79.7%.

Table III. Success rates obtained with GPCR-GIA by the jackknife cross-validation test in identifying proteins as GPCR or non-GPCR

Protein type	Number of proteins	Number of correct prediction	Success rate (%)
GPCR	780[a]	760	96.7
Non-GPCR	780[b]	729	93.5
Overall	1560	1489	95.4

[a]The sequences of the 780 GPCR proteins are given in the Supporting Information A available at *PEDS* Online.
[b]The sequences of the 780 non-GPCR proteins are given in the Supporting Information B available at *PEDS* Online.

Table IV. Success rates obtained with the GPCR-GIA predictor by the jackknife test and independent test in identifying nine main GPCR families

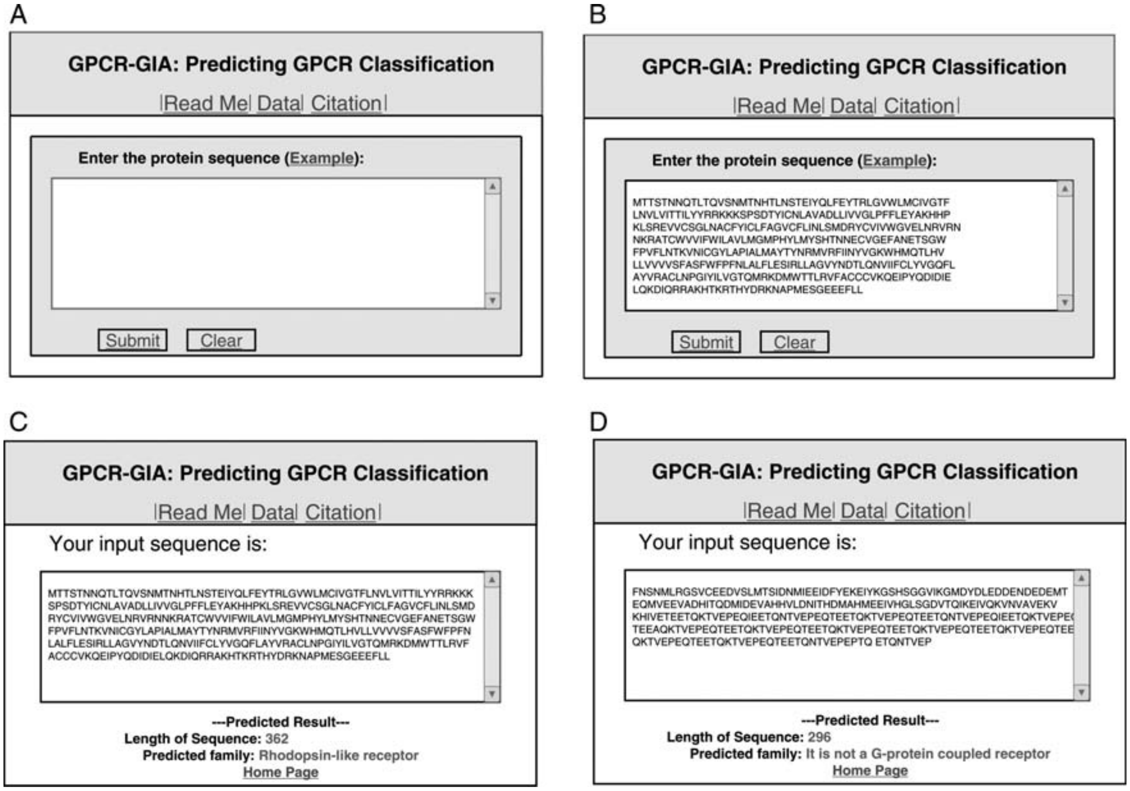
Family index	Functional feature	Jackknife[a]	Independent[b]
1	Rhodopsin-like receptor	123/144 = 85.4%	340/396 = 85.9%
2	Peptide hormones receptor	53/60 = 88.3%	14/15 = 93.3%
3	Glutamate and calcium receptor	10/20 = 50.0%	4/5 = 80%
4	Fungal mating pheromone receptor	5/11 = 45.5%	0/1 = 0.0%
5	Cyclic AMP receptor	3/4 = 75.0%	N/A
6	Odorant receptors in drosophila	36/45 = 80.0%	11/11 = 100.0%
7	Gustatory receptor of drosophila	7/17 = 41.2%	0/4 = 0.0%
8	Frizzled/smoothened family	10/14 = 71.4%	2/2 = 100.0%
9	T2R family in mammals	24/25 = 96.0%	4/6 = 66.7%
Overall		271/340 = 79.7%	375/440 = 85.2%

[a]Performed on the data set $\mathbb{S}^{\mathrm{L}}$ (cf. the Supporting Information C available at *PEDS*, online).
[b]Used the GPCR-GIA predictor trained by the data set $\mathbb{S}^{\mathrm{L}}$ to predict the proteins in the data set $\mathbb{S}^{\mathrm{T}}$ (cf. the Supporting Information D available at *PEDS* online).

Meanwhile, predictions were also performed with the GPCR-GIA trained by the data set $\mathbb{S}^{\mathrm{L}}$ (see the Supporting Information C available at PEDS online) on the independent testing data set $\mathbb{S}^{\mathrm{T}}$ (see the Supporting Information D available at PEDS online). As shown in Table IV, the overall success rate was 85.2%. However, it should be pointed out that the independent data set test performed here was just for a demonstration of practical application. Because the selection of independent data set often bears some sort of arbitrariness (Chou and Zhang, 1995), the jackknife test is deemed more objective than the independent data set test. Therefore, the power of a predictor should be measured by the success rate of jackknife test (Chou and Shen, 2008).

A step-by-step guide of using GPCR-GIA

Here, let us provide a step-by-step guide on how to use the web server to get the desired results.



GPCR-GIA: Predicting GPCR Classification

Enter the protein sequence (Example):

MTTSTNNQTLTQVSNMTNHTLNSTEIYQLFEYTRLGVWLMCIVGTF
LNVLVITTILYYRRKKKSPSDTYICNLAVADLLIVVGLPFFLEYAKHHP
KLSREVVCSGLNACFYICLFAGVCFLINLSMDRYCVIVWGVELNRVRN
NKRATCWVVIFWILAVLMGMPHYLMYSHTNNECVGEFANETSGW
FPVFLNTKVNICGYLAPIALMAYTYNRMVRFIINYVGKWHMQTLHV
LLVVVVSFASFWFPFNLALFLESIRLLAGVYNDTLQNVIIFCLYVGQFL
AYVRACLNPGIYILVGTQMRKDMWTTLRVFACCCVKQEIPYQDIDIE
LQKDIQRRAKHTKRTHYDRKNAPMESGEEEFLL

Your input sequence is:

MTTSTNNQTLTQVSNMTNHTLNSTEIYQLFEYTRLGVWLMCIVGTFLNVLVITTILYYRRKKK
SPSDTYICNLAVADLLIVVGLPFFLEYAKHHPKLSREVVCSGLNACFYICLFAGVCFLINLSMD
RYCVIVWGVELNRVRNNKRATCWVVIFWILAVLMGMPHYLMYSHTNNECVGEFANETSGW
FPVFLNTKVNICGYLAPIALMAYTYNRMVRFIINYVGKWHMQTLHVLLVVVVSFASFWFPFN
LALFLESIRLLAGVYNDTLQNVIIFCLYVGQFLAYVRACLNPGIYILVGTQMRKDMWTTLRVF
ACCCVKQEIPYQDIDIELQKDIQRRAKHTKRTHYDRKNAPMESGEEEFLL

---Predicted Result---
Length of Sequence: 362
Predicted family: Rhodopsin-like receptor

Your input sequence is:

FNSNMLRGSVCEEDVSLMTSIDNMIEEIDFYEKEIYKGSHSGGVIKGMDYDLEDDENDEDEMT
EQMVEEVADHITQDMIDEVAHHVLDNITHDMAHMEEIVHGLSGDVTQIKEIVQKVNVAVEKV
KHIVETEETQKTVEPEQIEETQNTVEPEQTEETQKTVEPEQTEETQNTVEPEQIEETQKTVEPE
TEEAQKTVEPEQTEETQKTVEPEQTEETQKTVEPEQTEETQKTVEPEQTEETQKTVEPEQTEE
QKTVEPEQTEETQKTVEPEQTEETQNTVEPEPTQ ETQNTVEP

---Predicted Result---
Length of Sequence: 296
Predicted family: It is not a G-protein coupled receptor

Figure 1. A semi-screenshot to show (**A**) the top page of the web server GPCR-GIA at http://218.65.61.89:8080/bioinfo/GPCR-GIA, (**B**) the sequence of a query protein entered into the input box of the GPCR-GIA web server, (**C**) the output predicted by the GPCR-GIA for the input taken from Example 1 of the Example window and (**D**) the output predicted by the GPCR-GIA for the input taken from Example 2 of the Example window. See the text in 'A step-by-step guide of using GPCR-GIA for further explanation.

Step 1. Open the web page http://218.65.61.89:8080/bioinfo/GPCR-GIA and you will see the top page of the predictor on your computer screen, as shown in Figure 1A. Click on the Read Me button to see a brief introduction about GPCR-GIA predictor and the caveat in using it.

Step 2. Either type or copy and paste the query protein sequence into the input box at the center of Figure 1A. The input sequence should be in the single-letter AA code, as shown by clicking on the Example button right above the input box.

Step 3. Click on the Submit button to see the predicted result. For example, if you use the sequence of Example 1 in the Example window, the input screen should look like the illustration in Figure 1B. After clicking the Submit button, you will see the predicted result of 'Rhodopsin-like receptor', as shown on the output screen of Figure 1C. It takes less than 10 s for a protein sequence of 300 AAs before the predicted results appear on your computer screen; the longer the sequence is, the more time that is needed. If you use the sequence of Example 2 in the Example window; after clicking the Submit button, you will see that 'It is not a G-protein coupled receptor', as shown on the output screen of Figure 1D.

Step 4. Click on the Citation button to find the relevant papers that document the detailed development and algorithm of GPCR-GIA.

Step 5. Click on the Data button to download the benchmark data sets that were used to train and test the GPCR-GIA predictor.

Caveat. To obtain the predicted result with the expected success rate, the entire sequence of the query protein rather than its fragment should be used as an input. A sequence with less than 50 AA residues is generally deemed as a fragment.

Conclusions

GPCR-GIA was developed for identifying GPCR proteins and their family classes solely based on the sequence information. GPCR-GIA was established on the following two cornerstones: one is the combination of the PseAAC with the grey incidence analysis (GIA), and the other is the fusion of many individual basic classifiers. For the convenience of most experimental biologists, a very user-friendly web server has been designed for GPCR-GIA that is freely accessible to the public at http://218.65.61.89:8080/bioinfo/GPCR-GIA, by which one can easily get the desired results without the need to understand the mathematical details. The web server will be periodically updated to expand the coverage scope by including new entries of GPCR proteins and reflect the continuous development of GPCR-GIA.

Acknowledgements

We wish to express our gratitude to the anonymous reviewers whose constructive comments were very helpful for strengthening the presentation of this paper.

Funding

The work in this research was supported by the grants from the National Natural Science Foundation of China (No.

60661003), the department of education of JiangXi Province (No.GJJ09271), and the plan for training youth scientists (stars of Jing-Cang) of Jiangxi Province.

References

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res.*, **25**, 3389–3402.

Bhasin,M. and Raghava,G.P. (2005) *Nucleic Acids Res.*, **33**, W143–W147.

Cai,Y.D., Zhou,G.P. and Chou,K.C. (2005) *J. Theor. Biol.*, **234**, 145–149.

Call,M.E., Schnell,J.R., Xu,C., Lutz,R.A., Chou,J.J. and Wucherpfennig,K.W. (2006) *Cell*, **127**, 355–368.

Chen,Y.L. and Li,Q.Z. (2007a) *J. Theor. Biol.*, **248**, 377–381.

Chen,Y.L. and Li,Q.Z. (2007b) *J. Theor. Biol.*, **245**, 775–783.

Chen,C., Tian,Y.X., Zou,X.Y., Cai,P.X. and Mo,J.Y. (2006a) *J. Theor. Biol.*, **243**, 444–448.

Chen,C., Zhou,X., Tian,Y., Zou,X. and Cai,P. (2006b) *Anal. Biochem.*, **357**, 116–121.

Chen,C., Chen,L., Zou,X. and Cai,P. (2009) *Protein Pept. Lett.*, **16**, 27–31.

Chou,K.C. (2001) *PROTEINS Struct. Funct. Genet.*, **43**, 246–255 (Erratum: ibid., 2001, Vol.44, 60).

Chou,K.C. (2004) *Curr. Med. Chem.*, **11**, 2105–2134.

Chou,K.C. (2005a) *J. Proteome Res.*, **4**, 1681–1686.

Chou,K.C. (2005b) *J. Proteome Res.*, **4**, 1413–1418.

Chou,K.C. (2005c) *Bioinformatics*, **21**, 10–19.

Chou,K.C. and Cai,Y.D. (2003) *PROTEINS: Struct. Funct. Genet.*, **53**, 282–289.

Chou,K.C. and Cai,Y.D. (2004) *Protein Sci.*, **13**, 2857–2863.

Chou,K.C. and Cai,Y.D. (2006) *J. Proteome Res.*, **5**, 316–322.

Chou,K.C. and Elrod,D.W. (1999) *Protein Eng.*, **12**, 107–118.

Chou,K.C. and Elrod,D.W. (2002) *J. Proteome Res.*, **1**, 429–433.

Chou,K.C. and Shen,H.B. (2006) *J. Proteome Res.*, **5**, 1888–1897.

Chou,K.C. and Shen,H.B. (2007a) *J. Proteome Res.*, **6**, 1728–1734.

Chou,K.C. and Shen,H.B. (2007b) *Biochem. Biophys. Res. Commun.*, **360**, 339–345.

Chou,K.C. and Shen,H.B. (2007c) *Anal. Biochem.*, **370**, 1–16.

Chou,K.C. and Shen,H.B. (2007d) *Biochem. Biophys. Res. Commun.*, **357**, 633–640.

Chou,K.C. and Shen,H.B. (2008) *Nat. Protoc.*, **3**, 153–162.

Chou,K.C. and Shen,H.B. (2009) *Open Bioinformat. J.*, **3**, 31–50 (open accessible at http://www.bentham.org/open/tobioij/).

Chou,K.C. and Zhang,C.T. (1995) *Critic. Rev. Biochem. Mol. Biol.*, **30**, 275–349.

Cover,T.M. and Hart,P.E. (1967) *IEEE Trans. Info. Theory*, **IT-13**, 21–27.

Deng,J.L. (1985) *Sys. Control Lett.*, **1**, 288–294.

Denoeux,T. (1995) *IEEE Trans. Syst. Man Cybern.*, **25**., 804–813.

Ding,Y.S. and Zhang,T.L. (2008) *Pattern Recognit. Lett.*, **29**, 1887–1892.

Ding,H., Luo,L. and Lin,H. (2009) *Protein Pept. Lett.*, **16**, 351–355.

Douglas,S.M., Chou,J.J. and Shih,W.M. (2007) *Proc. Natl Acad. Sci. USA*, **104**, 6644–6648.

Du,P. and Li,Y. (2006) *BMC Bioinformat.*, **7**, 518.

Gao,Q.B. and Wang,Z.Z. (2006) *Protein Eng. Des. Sel.*, **19**, 511–516.

Georgiou,D.N., Karakasidis,T.E., Nieto,J.J. and Torres,A. (2009) *J. Theor. Biol.*, **257**, 17–26.

Gonzalez-Diaz,H., Vilar,S., Santana,L. and Uriarte,E. (2007) *Curr. Top. Med. Chem.*, **10**, 1015–1029.

Gonzalez-Diaz,H., Gonzalez-Diaz,Y., Santana,L., Ubeira,F.M. and Uriarte,E. (2008a) *Proteomics*, **8**, 750–778.

Gonzalez-Diaz,H., Prado-Prado,F. and Ubeira,F.M. (2008b) *Curr. Top. Med. Chem.*, **8**, 1676–1690.

Hall,R.A. and Lefkowitz,R.J. (2002) *Circ. Res.*, **91**, 672–680.

Heuss,C. and Gerber,U. (2000) *Trends Neurosci.*, **23**, 469–475.

Hill,C.A., Fox,A.N., Pitts,R.J., Kent,L.B., Tan,P.L., Chrystal,M.A., Cravchik,A., Collins,F.H., Robertson,H.M. and Zwiebel,L.J. (2002) *Science*, **298**, 176–178.

Hopp,T.P. and Woods,K.R. (1981) *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.

Jiang,X., Wei,R., Zhang,T.L. and Gu,Q. (2008) *Protein Pept. Lett.*, **15**, 392–396.

Keller,J.M., Gray,M.R. and Givens,J.A. (1985) *IEEE Trans. Syst. Man Cybern.*, **15**, 580–585.

Li,F.M. and Li,Q.Z. (2008) *Protein Pept. Lett.*, **15**, 612–616.

Lin,H. (2008) *J. Theor. Biol.*, **252**, 350–356.

Lin,H. and Li,Q.Z. (2007a) *Biochem. Biophys. Res. Commun.*, **354**, 548–551.

Lin,H. and Li,Q.Z. (2007b) *J. Comput. Chem.*, **28**, 1463–1466.

Liu,H., Wang,M. and Chou,K.C. (2005) *Biochem. Biophys. Res. Commun.*, **336**, 737–739.

Liu,S.F., Fang,Z.G. and Lin,Y. (2006) *Sci. Inq.*, **7**, 111–124.

Lin,H., Ding,H., Guo,F.B., Zhang,A.Y. and Huang,J. (2008) *Protein Pept. Lett.*, **15**, 739–744.

Lin,H., Wang,H., Ding,H., Chen,Y.L. and Li,Q.Z. (2009) *Acta Biotheor.*, **57**, 321–330.

Mahalanobis,P.C. (1936) *Proc. Natl Inst. Sci. India* **2**, 49–55.

Mardia,K.V., Kent,J.T. and Bibby,J.M. (1979) *Multivariate Analysis*, Academic Press, London, pp. 322–381.

Matsunami,H., Montmayeur,J.P. and Buck,L.B. (2000) *Nature*, **404**, 601–604.

Milligan,G. and White,J.H. (2001) *Trends Pharmacol. Sci.*, **22**, 513–518.

Mondal,S., Bhavna,R., Mohan Babu,R. and Ramakumar,S. (2006) *J. Theor. Biol.*, **243**, 252–260.

Mundra,P., Kumar,M., Kumar,K.K., Jayaraman,V.K. and Kulkarni,B.D. (2007) *Pattern Recognit. Lett.*, **28**, 1610–1615.

Nakashima,H., Nishikawa,K. and Ooi,T. (1986) *J. Biochem.*, **99**, 152–162.

Otaki,J.M. and Yamamoto,H. (2003) *J. Theor. Biol.*, **223**, 27–37.

Oxenoid,K. and Chou,J.J. (2005) *Proc. Natl Acad. Sci. USA*, **102**, 10870–10875.

Pan,Y.X., Zhang,Z.Z., Guo,Z.M., Feng,G.Y., Huang,Z.D. and He,L. (2003) *J. Protein Chem.*, **22**, 395–402.

Pillai,K.C.S. (1985) In Kotz,S. and Johnson,N.L. (eds), *Encyclopedia of Statistical Sciences*, Vol. **5**, John Wiley & Sons, pp. 176–181.

Qiu,J.D., Huang,J.H., Liang,R.P. and Lu,X.Q. (2009) *Anal. Biochem.*, **390**, 68–73.

Schnell,J.R. and Chou,J.J. (2008) *Nature*, **451**, 591–595.

Shen,H.B. and Chou,K.C. (2005a) *Biochem. Biophys. Res. Commun.*, **337**, 752–756.

Shen,H.B. and Chou,K.C. (2005b) *Biochem. Biophys. Res. Commun.*, **334**, 288–292.

Shen,H.B. and Chou,K.C. (2006) *Bioinformatics*, **22**, 1717–1722.

Shen,H.B. and Chou,K.C. (2007a) *Biochem. Biophys. Res. Commun.*, **364**, 53–59.

Shen,H.B. and Chou,K.C. (2007b) *Biochem. Biophys. Res. Commun.*, **363**, 297–303.

Shen,H.B. and Chou,K.C. (2008) *Anal. Biochem.*, **373**, 386–388.

Shen,H.B. and Chou,K.C. (2009) *J. Theor. Biol.*, **256**, 441–446.

Shen,H.B., Yang,J. and Chou,K.C. (2006) *J. Theor. Biol.*, **240**, 9–13.

Shen,H.B., Song,J.N. and Chou,K.C. (2009) *J. Biomed. Sci. Eng. (JBiSE)*, **2**, 136–143 (open accessible at http://www.srpublishing.org/journal/jbise/).

Tanford,C. (1962) *J. Am. Chem. Soc.*, **84**, 4240–4274.

Wang,S.Q., Yang,J. and Chou,K.C. (2006) *J. Theor. Biol.*, **242**, 941–946.

Wen,Z., Li,M., Li,Y., Guo,Y. and Wang,K. (2007) *Amino Acids*, **32**, 277–283.

Xiao,X., Shao,S., Ding,Y., Huang,Z., Huang,Y. and Chou,K.C. (2005) *Amino Acids*, **28**, 57–61.

Xiao,X., Shao,S.H., Huang,Z.D. and Chou,K.C. (2006) *J. Comput. Chem.*, **27**, 478–482.

Xiao,X., Lin,W.Z. and Chou,K.C. (2008a) *J. Comput. Chem.*, **29**, 2018–2024.

Xiao,X., Wang,P. and Chou,K.C. (2008b) *J. Theor. Biol.*, **254**, 691–696.

Xiao,X., Wang,P. and Chou,K.C. (2009a) *J. Comput. Chem.*, **30**, 1414–1423.

Xiao,X., Wang,P. and Chou,K.C. (2009b) *J. Appl. Crystallogr.*, **42**, 169–173.

Zeng,Y.H., Guo,Y.Z., Xiao,R.Q., Yang,L., Yu,L.Z. and Li,M.L. (2009) *J. Theor. Biol.*, **259**, 366–372.

Zhang,G.Y. and Fang,B.S. (2008) *J. Theor. Biol.*, **253**, 310–315.

Zhang,G.Y., Li,H.C. and Fang,B.S. (2008a) *Protein Pept Lett.*, **15**, 1132–1137.

Zhang,T.L., Ding,Y.S. and Chou,K.C. (2008b) *J. Theor. Biol.*, **250**, 186–193.

Zhou,G.P. (1998) *J. Protein Chem.*, **17**, 729–738.

Zhou,G.P. and Assa-Munt,N. (2001) *PROTEINS Struct. Funct. Genet.*, **44**, 57–59.

Zhou,G.P. and Cai,Y.D. (2006) *PROTEINS Struct. Funct. Bioinformat.*, **63**, 681–684.

Zhou,G.P. and Doctor,K. (2003) *PROTEINS Struct. Funct. Genet.*, **50**, 44–48.

Zhou,X.B., Chen,C., Li,Z.C. and Zou,X.Y. (2007) *J. Theor. Biol.*, **248**, 546–551.

Received July 6, 2009; revised August 27, 2009;
accepted August 27, 2009

Edited by Michael Deem