Molecular BioSystems

Cite this: *Mol. BioSyst.*, 2011, **7**, 911–919

www.rsc.org/molecularbiosystems

PAPER

GPCR-2L: predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions†

Xuan Xiao,*[ab] Pu Wang[a] and Kuo-Chen Chou*[b]

Received 20th August 2010, Accepted 18th November 2010
DOI: 10.1039/c0mb00170h

G protein-coupled receptors (GPCRs) are among the most frequent targets of therapeutic drugs. With the avalanche of newly generated protein sequences in the post genomic age, to expedite the process of drug discovery, it is highly desirable to develop an automated method to rapidly identify GPCRs and their types. A new predictor was developed by hybridizing two different modes of pseudo-amino acid composition (PseAAC): the functional domain PseAAC and the low-frequency Fourier spectrum PseAAC. The new predictor is called **GPCR-2L**, where "2L" means that it is a two-layer predictor: the 1st layer prediction engine is to identify a query protein as GPCR or not; if it is, the prediction will be automatically continued to further identify it as belonging to one of the following six types: (1) rhodopsin-like (Class A), (2) secretin-like (Class B), (3) metabotropic glutamate/pheromone (Class C), (4) fungal pheromone (Class D), (5) cAMP receptor (Class E), or (6) frizzled/smoothened family (Class F). The overall success rate of **GPCR-2L** in identifying proteins as GPCRs or non-GPCRs is over 97.2%, while identifying GPCRs among their six types is over 97.8%. Such high success rates were derived by the rigorous jackknife cross-validation on a stringent benchmark dataset, in which none of the included proteins had $\geq 40\%$ pairwise sequence identity to any other protein in a same subset. As a user-friendly web-server, **GPCR-2L** is freely accessible to the public at http://icpr.jci.edu.cn/bioinfo/GPCR-2L, by which one can obtain the 2-level results in about 20 s for a query protein sequence of 500 amino acids. The longer the sequence is, the more time it may usually need. The high success rates reported here indicate that it is a quite effective approach to identify GPCRs and their types with the functional domain information and the low-frequency Fourier spectrum analysis. It is anticipated that **GPCR-2L** may become a useful tool for both basic research and drug development in the areas related to GPCRs.

Introduction

G protein-coupled receptors (GPCRs) belong to the largest family of cell surface receptors. Known also as G protein-linked receptors (GPLRs), serpentine receptors, seven-transmembrane domain receptors, and 7TM (transmembrane) receptors (Fig. 1), GPCRs are comprised of a large protein family of transmembrane receptors. They have the function of activating the signal transduction pathways across the cell membranes that are indispensable for cellular responses. The GPCR-related pathways are the targets of hundreds of drugs, including neuroleptics, antihistamines, antidepressants, and antihypertensives, actually representing targets for 50% of

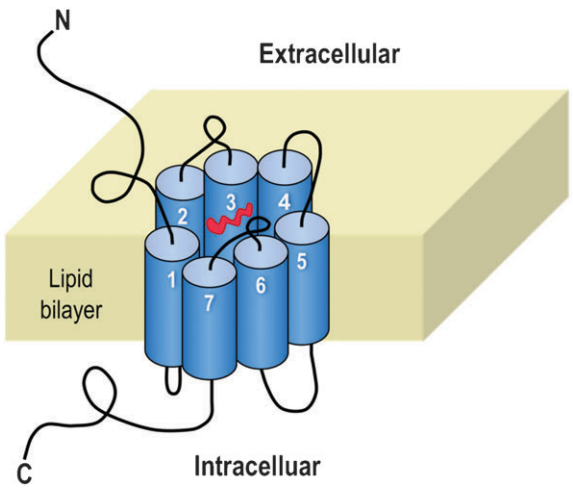

Fig. 1 GPCRs with the 7TM trademark. The cylinders represent the helices, which are connected by alternating cytoplasmic and extracellular hydrophilic loops. The 7TM helix bundle has a central pore on its extra-cellular surface. The red entity located in the central pore represents a ligand messenger.

[a] *Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China. E-mail: xxiao@gordonlifescience.org*
[b] *Gordon Life Science Institute, 13784 Torrey Del Mar Drive, San Diego, California 92130, USA. E-mail: kcchou@gordonlifescience.org*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c0mb00170h

marketed drugs.[1] GPCRs also mediate the actions of certain medications used to treat disorders as diverse as cardiovascular disease, drug dependency, and mental illness.[2]

The major role of GPCRs is to transduce extra-cellular signals across the cell membrane *via* guanine-binding proteins (G-proteins) with high specificity and sensitivity. GPCRs regulate many basic physicochemical processes contained in a cellular signalling network, such as smell, taste, vision, secretion, neurotransmission, metabolism, cellular differentiation and growth, inflammatory and immune response.[3–5]

A lot of effort has been invested in studying GPCRs by both academic institutions and pharmaceutical industries. Unfortunately, so far the functions of many GPCRs are still unknown and determining their ligands and signalling pathways by experimental approach is both time-consuming and costly. This is mainly due to the following methodological challenges unique to the membrane proteins: (1) they are difficult to over-express and purify; (2) they are difficult to crystallize because of their hydrophobicity; (3) that in turn makes it difficult to obtain good X ray diffraction data. Accordingly, so far very few crystal GPCR structures have been determined. Although the recently developed high-resolution NMR spectroscopy is a very powerful technique in determining the 3D (three-dimensional) structures of membrane proteins (see, *e.g.*, ref. 6–14), it is time-consuming and costly. Also, although using structural bioinformatics tools can often quickly acquire the desired 3D structures for drug design (see, *e.g.*, ref. 15–22), it fails to work for most GPCRs because very few of them have sufficiently high sequence similarity with existing structure-known proteins, a prerequisite condition for developing a reasonable starting structure *via* structural bioinformatics.[23]

Facing the avalanche of protein sequence data generated in the post genomic age, it is highly desirable to develop computational methods that can rapidly and effectively identify the functional families of GPCRs based on their primary sequences so as to provide useful information for classifying drug targets, a technique called "evolutionary pharmacology" for drug design.

Actually, in a pioneer study, the covariant-discriminant algorithm[24,25] was introduced to identify the 566 GPCRs within the rhodopsin-like family, classified into 7 subfamily classes.[26] Later, a similar approach was used to study the 1238 GPCRs classified into 3 main families.[27] Stimulated by the encouraging results, some follow-up studies were conducted as reported in ref. 28–32.

As pointed out in a recent review,[33] any prediction method developed in the Internet Age should provide a web-server to make it practically more useful. Among the aforementioned methods, only the one called GPCR-CA in ref. 31 has provided a web-server at http://icpr.jci.jx.cn/bioinfo/GPCR-CA that is freely accessible for the public to get the desired results. However, to some of GPCR types, the prediction success rates by GPCR-CA[31] still needs to be improved.

The present study was initiated in an attempt to develop a new method for predicting GPCRs and their types by not only providing a user-friendly web-server but also enhancing the prediction quality.

Materials and methods

To develop an effective statistical method for predicting GPCRs and their types, we need the following three things:[34] (i) a valid benchmark dataset; (ii) an effective mathematical formulation for the samples that can truly associate with the core feature of GPCRs; (iii) a powerful prediction algorithm (or engine). In this study, the aforementioned three necessities are to be realized as follows.

1. Benchmark datasets

Protein sequences were collected from the G protein-coupled receptor database (GPCRDB release 10.0) at http://www.gpcr.org/7tm_old/.[35] The GPCRDB is a molecular class-specific information system that collects, combines, validates, and disseminates heterogeneous data on GPCRs, and is updated automatically once every 4–5 months according to the Swiss-Prot and TrEMBL Data Banks.

According to their binding with different ligand types, GPCRs are classified into the following six main families (Fig. 2): (i) rhodopsin-like (Class A), (ii) secretin-like (Class B), (iii) metabotropic glutamate/pheromone (Class C), (iv) fungal pheromone (Class D), (v) cAMP receptor (Class E), and (vi) frizzled/smoothened family (Class F). To guarantee the quality, the data were screened strictly according to the following criteria: (i) sequences with less than 50 amino acid residues were removed because they might just be fragments; (ii) those protein sequences which contained irregular amino acid characters such as "B", "X" or "Z" were removed; (iii) those protein sequences which were annotated with more than one type were removed because they lacked uniqueness; (iv) to avoid any homology bias, a redundancy cutoff was imposed with the program CD-HIT[36] to winnow those sequences which had $\geq 40\%$ pairwise sequence identity to any other in a same subset (GPCR type) except for Class E (cAMP receptor), because it contained only 10 GPCR protein sequences and would have otherwise resulted in too few samples to have any statistical significance. Also, the proteins with codes Q54J91, Q54YI3, and Q55717 were removed because they have exactly the same sequences as P4907, P35352, and P13773, respectively.

After strictly following the above procedures (i)–(iv), we finally obtained a GPCR benchmark dataset $\mathbb{S}^{\mathrm{GPCR}}$ that contained 367 protein sequences, of which 236 belonged to Class A, 39 to Class B, 44 to Class C, 23 to Class D, 7 to Class E, and 18 to Class F (Table 1), which can be formulated as follows:

$$\mathbb{S}^{\mathrm{GPCR}} = \mathbb{S}_{\mathrm{A}}^{\mathrm{GPCR}} \cup \mathbb{S}_{\mathrm{B}}^{\mathrm{GPCR}} \cup \mathbb{S}_{\mathrm{C}}^{\mathrm{GPCR}} \cup \mathbb{S}_{\mathrm{D}}^{\mathrm{GPCR}} \cup \mathbb{S}_{\mathrm{E}}^{\mathrm{GPCR}} \cup \mathbb{S}_{\mathrm{F}}^{\mathrm{GPCR}} \quad (1)$$

where $\mathbb{S}_{\mathrm{A}}^{\mathrm{GPCR}}$ is the subset containing Class A GPCRs only, $\mathbb{S}_{\mathrm{B}}^{\mathrm{GPCR}}$ the subset containing Class B GPCRs only, and so forth; $\cup$ is the symbol for union in the set theory. Meanwhile, in order to train a statistical predictor with the ability to distinguish GPCR proteins from non-GPCR proteins, a non-GPCR benchmark dataset $\mathbb{S}^{\text{non-GPCR}}$ was also constructed by randomly collecting 1101 non-GPCR proteins from the UniProtKB at http://www.uniprot.org/ according to their annotations in the "Keyword" field. None of the proteins

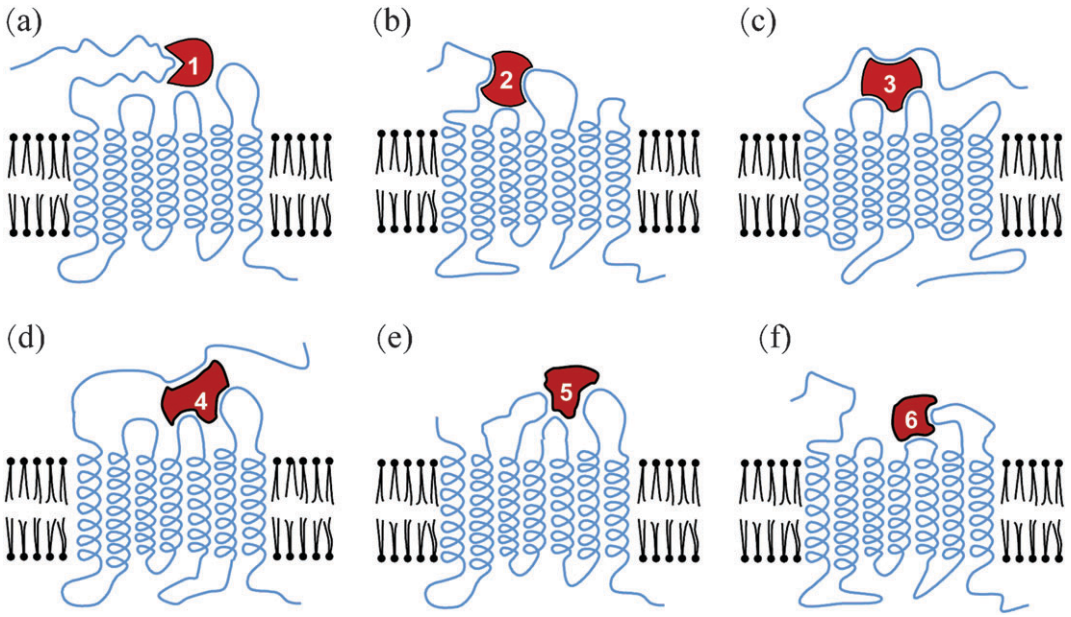

Fig. 2 Schematic drawing to show the six main function-different types of GPCRs. The six types are: (a) rhodopsin-like or class A; (b) secretin-like or class B; (c) metabotropic glutamate/pheromone or class C; (d) fungal pheromone or class D; (e) cAMP receptor or class E; and (f) frizzled/smoothened family or class F.

Table 1 Breakdown of the numbers of GPCRs and their six types as well as non-GPCRs for the benchmark dataset $\mathbb{S}$ used in this study

Protein attribute	Subtype	Subset	Number of sequences[a]
GPCR	Class A	$\mathbb{S}_A^{GPCR}$	236
	Class B	$\mathbb{S}_B^{GPCR}$	39
	Class C	$\mathbb{S}_C^{GPCR}$	44
	Class D	$\mathbb{S}_D^{GPCR}$	23
	Class E	$\mathbb{S}_E^{GPCR}$	7
	Class F	$\mathbb{S}_F^{GPCR}$	18
Non-GPCR		$\mathbb{S}^{non\text{-}GPCR}$	1101

[a] For the detailed sequences, see ESI S1.†

included in $\mathbb{S}^{non\text{-}GPCR}$ had $\geq 40\%$ pairwise sequence identity to any other by using the same screen procedure[36] as used in establishing $\mathbb{S}^{GPCR}$ of eqn (1). It is instructive to point out that the number of proteins in $\mathbb{S}^{non\text{-}GPCR}$ is significantly greater than that in $\mathbb{S}^{GPCR}$ to reflect the reality that in the protein universe most proteins are of non-GPCR.

The 367 GPCR sequences classified into six types and 1101 non-GPCR sequences are given in ESI S1.†

2. Sample formulation or descriptor

There are many different manners to formulate protein sequence samples for statistical prediction. However, they can be basically classified into the following two models: the sequential model and discrete model.[34] The most straightforward sequential model for a protein sample is its entire amino acid sequence. Its advantage is that it is able to contain the most complete information of the protein. For this kind of sequential model, the sequence-similarity-search-based tools, such as BLAST,[37,38] are usually utilized for prediction. However, such an approach would fail to work when the query protein did not have significant homology to proteins with known characteristics, particularly for the current benchmark dataset in which none of proteins has $\geq 40\%$ pairwise sequence identity to any other in a same subset. To deal with the sequence-diversifying situation, various discrete models have been proposed by catching the cores of the prediction targets. The simplest discrete model used to represent a protein sample is its amino acid (AA) composition or AAC, which was widely used for predicting various protein attributes (see, *e.g.*, ref. 39–54). However, the prediction quality might be considerably limited by using the AAC-discrete model since all the sequence-order information would be totally lost accordingly. To avoid complete loss of the sequence-order information, the pseudo amino acid (PseAA) composition or PseAAC was proposed.[55] According to its original idea, the PseAAC is actually formulated by a set of discrete numbers, as long as it is different from the classical AAC, and it is derived from a protein sequence that is able to harbor some of its sequence order and pattern information, or able to reflect some physicochemical and biochemical properties of the constituent amino acids.[55,56] For a brief introduction about PseAAC, visit the Wikipedia web-page at http://en.wikipedia.org/wiki/Pseudo_amino_acid_composition. The concept of PseAAC has provided a very flexible mathematical frame for incorporating various effects into the protein sample formulation, and hence has been widely used to deal with many protein-related problems and sequence-related systems (see, *e.g.*, ref. 30,57–76). As summarized in ref. 34, so far there are about 16 different PseAAC modes that have been used to formulate the samples of proteins for predicting their attributes. These modes each have their own advantages and disadvantages In this study, we formulate the protein samples by hybridizing the functional domain mode and the low-frequency Fourier spectrum mode, as described below.

2.1 FunD (functional domain) PseAAC mode. The FunD is the core of a protein that is the structural basis of its function.

Accordingly, in determining the 3D (three-dimensional) structure of a protein by experiments (see, *e.g.*, ref. 9,10) or developing a 3D structure model for a drug-targeted protein (see, *e.g.*, ref. 16,17,23,77,78), the first priority always focuses on its FunD. Based on the 2005 FunDs in the SBASE-A database,[79] the FunD descriptor was originally introduced for predicting protein subcellular localization[80] and membrane protein type.[81] Since then, several new protein FunD databases have been established. They are COG,[82] KOG,[82] Pfam,[83] SMART,[84] and CDD.[85] In this study, the SMART database[84] was adopted to formulate the FunD composition for a protein sample as follows. The current SMART database contains 12 810 domain entries consisting of the following five categories: SMART, Pfam, coiled coil region, signal peptide, and transmembrane (see ESI S2†).

The relation of these domains with a given protein sequence can be found by running a program called "Sequence SMART" at http://smart.embl.de/. For instance, for the protein with accession number P28222, it was found to contain 7 "transmembrane" domains (res. 51–73, res. 86–108, res. 123–145, res. 165–187, res. 207–229, res. 315–337, res. 347–369) and one "low complexity" domain (res. 4–15). Thus, with each of the 12 810 domain sequences as a vector-base, a given protein sample **P** can be defined as a vector in a 12810-D space; *i.e.*,

$$\mathbf{P}_{\mathrm{FunD}} = [n_{\mathrm{D}(1)}\, n_{\mathrm{D}(2)} \cdots n_{\mathrm{D}(k)} \cdots n_{\mathrm{D}(12810)}]^{\mathbf{T}} \tag{2}$$

where **T** the transpose operator, and

$$n_{\mathrm{D}(k)} = \text{number of the occurrence for the } k\text{th domain in } \mathbf{P} \tag{3}$$

Take the aforementioned protein P28222 as an example: the number of the occurrence for the "transmembrane" domains is 7, and that for the "low complexity" domain is 1, while that for each of all the other domains is zero.

Thus, rather than the classic AAC discrete mode as defined in a 20-D space (see, *e.g.*, ref. 45), the protein **P** is now corresponding to a 12810-D vector with its components defined, respectively, by the sequence patterns of the 12 810 functional domains in the SMART database.[84] By doing so, not only are many sequence pattern features naturally incorporated into the formulation (eqn (2)) for protein samples but also considerable function-related information as well.

2.2 Low-frequency Fourier spectrum PseAAC mode. If none of the domains in the SMART database[84] were found in eqn (3), *i.e.*, $n_{\mathrm{D}(k)} = 0$ for ($k = 1,2,\ldots, 12\,810$), the protein sample $\mathbf{P}_{\mathrm{FunD}}$ as defined in eqn (2) would correspond to a nought vector and become meaningless. Under such a circumstance, we should instead use the low-frequency Fourier PseAAC mode to represent the protein sample as a complement. It was originally introduced for predicting membrane protein types,[86] as formulated below.

For a protein **P** with L amino acid residues, suppose $H(\mathrm{R}_1)$is the hydrophilic value of its 1st residue R_1, $H(\mathrm{R}_2)$ that of its 2nd residue R_2, and so forth. The hydrophilic values for the 20 native amino acids were taken from ref. 87. With these hydrophilic values along its sequence, the protein can be converted to a digital signal, from which we can generate $2L$ discrete Fourier spectrum numbers as given below:

$$\{F_1,F_2,\ldots,F_L, \Phi_1,\Phi_2,\ldots,\Phi_L\} \tag{4}$$

where the amplitude components F_k and phase components Φ_k ($k = 1,2,\ldots, L$) are defined by the following discrete Fourier spectrum transform formula:

$$\sum_{\ell=1}^{L} H(\mathrm{R}_i) \exp\left[-i\left(\frac{2\pi\ell}{L}\right)k\right] = F_k \exp(i\Phi_k), \quad (k = 1,2,\ldots,L) \tag{5}$$

where i represents the imaginary number.

The $2L$ Fourier spectrum numbers contain a substantial amount of information about the digit signal,[88] and hence can also be used to reflect the pattern of a protein sequence. Furthermore, in the L phase components $\{\Phi_1,\Phi_2,\ldots,\Phi_L\}$, the high-frequency components are noisier and hence only the low-frequency components are more important. This is just like the case of protein internal motions where the low-frequency (or Terahertz frequency) components are functionally more important (see, *e.g.*, ref. 89–91 as well as the web-sites at http://en.wikipedia.org/wiki/Low-frequency_Collective_Motion and http://homepages.sover.net/~bell/newFrontierpics.htm). Accordingly, we only need to consider the 1st λ phase components as well as their corresponding amplitudes, *i.e.*,

$$\{F_1,F_2,\ldots F_\lambda,\Phi_1,\Phi_2,\ldots,\Phi_\lambda\}, (\lambda < L) \tag{6}$$

After incorporating the above components into the classical 20-D ACC, we obtain the low-frequency Fourier spectrum PseAAC mode for representing the protein **P**; *i.e.*

$$\mathbf{P}_{\mathrm{Fourier}}^{\text{low-freq}} = [p_1 \cdots p_{20}\, p_{20+1} \cdots p_{20+\lambda}\, p_{20+\lambda+1} \cdots p_{20+2\lambda}]^{\mathbf{T}} \tag{7}$$

where

$$p_k = \begin{cases} \dfrac{f_k}{\sum_{i=1}^{20} f_i + w\sum_{i=1}^{\lambda}(F_i+\Phi_i)}, & (1 \le k \le 20) \\ \dfrac{wF_{k-20}}{\sum_{i=1}^{20} f_i + w\sum_{i=1}^{\lambda}(F_i+\Phi_i)}, & (20+1 \le k \le 20+\lambda) \\ \dfrac{w\Phi_{k-20-\lambda}}{\sum_{i=1}^{20} f_i + w\sum_{i=1}^{\lambda}(F_i+\Phi_i)}, & (20+\lambda+1 \le k \le 20+2\lambda) \end{cases} \tag{8}$$

where f_i ($i = 1, 2,\ldots,20$) are the occurrence frequencies of the 20 native amino acids in a protein, w is the weight factor, and λ the threshold for the low-frequency passing filter.[88] In the current study, we chose $w = 1/100$ to make the results of eqn (8) within the range easier to be handled (w can be of course assigned with other values, but this would not make a significant difference to the final results), and we took $\lambda = 5$ for the optimal result in this study.

For the current benchmark dataset of the 1468 protein sequences as given in ESI S1,† 1459 sequences were found to contain the FunD segments in the SMART database, and hence could be meaningfully defined in the 12810-D FunD space (eqn (2)); only 9 protein sequences did not contain any

FunD segments and hence were defined by the low-frequency Fourier spectrum PseAAC mode (eqn (7)).

Below, let us deal with the third necessity, *i.e.*, the prediction engine.

3. Fuzzy *K*-nearest neighbor classifier

The prediction engine adopted in this study was based on the fuzzy *K*-nearest neighbor (*K*-NN) rule. According to the *K*-NN rule,[92] the query protein should be assigned to the subset represented by the majority of its *K*-nearest neighbors. Recently, various classifiers based on the *K*-NN rule have been successfully used to predict protein subcellular localization,[93,94] membrane proteins and their types,[95] proteases and their types,[96] as well as many other protein attributes.[33]

Fuzzy *K*-NN classifier[97] is a special variation of the *K*-NN classifier. Instead of simply assigning the label of class based on a voting from the nearest neighbors, it attempts to estimate the membership values according to the fuzzy principle to indicate to what degree the query sample belongs to the classes concerned. Since it is impossible to contain the complete information of a protein sequence by a discrete model, the fuzzy principle would be particularly useful in dealing with these kinds of problems.[98–100]

According to the fuzzy *K*-NN algorithm,[97] for a system consisting of *N* proteins $\{\mathbf{P}_1,\mathbf{P}_2,\ldots\mathbf{P}_N\}$ classified into *M* classes $\{1,2,\ldots,M\}$, the fuzzy membership value of a query protein **P** belonging to the *i*th class is given by:

$$\mu_i(\mathbf{P}) = \frac{\sum_{j=1}^{K} \mu_i(\mathbf{P}_j)\mathrm{d}(\mathbf{P},\mathbf{P}_j)^{-2/(\varphi-1)}}{\sum_{j=1}^{K} \mathrm{d}(\mathbf{P},\mathbf{P}_j)^{-2/(\varphi-1)}} \qquad (9)$$

where *K* is the number of the nearest neighbors counted; $\mu_i(\mathbf{P}_j)$ is the fuzzy membership value of the protein $\mathbf{P}_j$ to the *i*th class (it is set to 1 if the d($\mathbf{P}$, $\mathbf{P}_j$) is the distance between the query protein $\mathbf{P}_j$ in the training dataset; and $\phi(>1)$ is the fuzzy coefficient for determining how heavily the distance is weighted when calculating each nearest neighbor's contribution to the membership value. Various metrics can be chosen for d($\mathbf{P}$,$\mathbf{P}_j$), such as Euclidean distance, Hamming distance,[101] and Mahalanobis distance.[45] In this paper, the Euclidean metric was used. The values of ϕ and *K* will be mentioned later. After calculating all the memberships for a query protein, it is assigned to the class with which it has the highest membership value; *i.e.*, the predicted class for the query protein **P** should be

$$\Omega = \mathbf{argmax}_i\{\mu_i(\mathbf{P})\} \qquad (10)$$

where Ω is the argument of *i* that maximizes $\mu_i(\mathbf{P})$.

Now we have all the three necessities established for predicting GPCRs and their types. The predictor thus established is called **GPCR-2L**, where "2L" means the prediction consists of two layers. The 1st layer is to identify a query protein as GPCR or not; if it is a GPCR, the 2nd layer will be automatically continued to further identify the GPCR among the following six types: (1) Class A Rhodopsin like, (2) Class B Secretin like, (3) Class C Metabotropic glutamate/pheromone, (4) Class D Fungal pheromone, (5) Class E cAMP receptors, and (6) Class F Frizzled/Smoothened family. To provide an intuitive picture, a flowchart to show the process of how the **GPCR-2L** classifier works is given in Fig. 3.

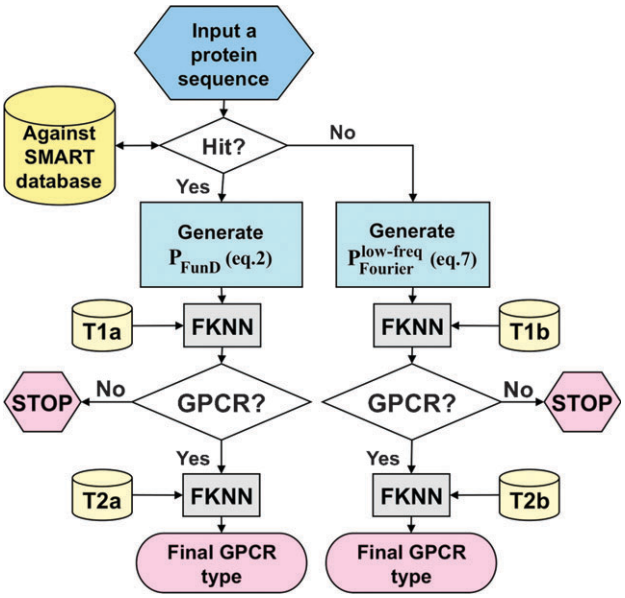

Fig. 3 Flowchart to show the operation process of the GPCR-2L predictor. **T1** represents the training dataset taken from the ESI S1† when it is classified into GPCRs and non-GPCRs; **T2** the training dataset for GPCRs when they are classified into six different types; **a** represents the corresponding training samples when formulated by $\mathbf{P}_{\mathrm{FunD}}$ of eqn (2); **b** represents the corresponding training samples when formulated by $\mathbf{P}^{\mathrm{Low\text{-}freq}}_{\mathrm{Fourier}}$ of eqn (7); FKNN represents the fuzzy *K*-NN classifier.

It is instructive to point out that the following self-consistency principle should be followed in utilizing **GPCR-2L**. Regardless of which kind of PseAAC mode is adopted for protein samples, the query proteins and the proteins used to train the prediction engine must be defined in the same PseAAC mode. For instance, if a query protein is defined in the 12810-D PseAAC space of eqn (2), then the prediction should be carried out based on all those proteins in the training set that can be defined in the exactly same 12810-D PseAAC space as well. However, if the query protein in the 12810-D PseAAC space is a naught vector and, hence, must be defined instead in the (20 + 2λ)-D PseAAC space of eqn (7), then all the proteins in the training dataset must also be formulated in the same (20 + 2λ)-D PseAAC space to train the prediction engine.

Results and discussion

To examine the performance quality of a statistical predictor, three cross-validation test methods are often used, *i.e.*, the independent dataset test, subsampling (such as 5-, 7- or 10-fold division) test, and jackknife test.[101] Of these three, however, the jackknife test is deemed the most objective without arbitrariness as elucidated in ref. 102 and demonstrated by eqn (1) of ref. 103 or eqn (50) of ref. 104 Therefore, the jackknife cross-validation has been increasingly adopted to examine the accuracy of various predictors (see, *e.g.*, ref. 60,62,64,67–70,73,74,105–107). Accordingly, in this

study, we also used the jackknife cross-validation to examine the prediction quality of **GPCR-2L**. In the jackknife test, all the proteins in the benchmark dataset will be singled out one-by-one and tested by the predictor trained by the remaining protein samples. Accordingly, during the jackknifing process, both the training dataset and testing dataset are actually open, and each protein sample will be in turn moved between the two. The jackknife test can exclude the "memory" effect. Also, the arbitrariness problem, as elaborated in ref. 103 and 104 for the independent dataset test and subsampling test, can be avoided because the outcome obtained by the jackknife cross-validation is always unique for a given benchmark dataset.

The values of φ and K used in eqn (9) were determined by optimizing the overall jackknife success rate through a 2-D search. It was found that, when the query protein and the protein samples in the training dataset were defined in the 12810-D PseAAC space, the optimal values for φ and K in eqn (9) were 1.21 and 5, respectively (see Fig. 4); when the query protein and the protein samples in the training dataset were defined in the $(20 + 2\lambda)$-D PseAAC space, the optimal values for φ and K in eqn (9) were 1.51 and 3, respectively.

The results thus obtained using **GPCR-2L** on the benchmark dataset (*cf.* ESI S1†) are given in Tables 2 and 3, from which we can see that the overall success rate in identifying proteins as GPCRs or non-GPCRs is 97.28% while identifying GPCRs among their six main types is 97.82%. In contrast, if no FunD formulation (eqn (2)) was used to represent the protein samples whatsoever, and all the protein samples in the benchmark dataset were represented by the low-frequency Fourier spectrum PseAAC formulation (eqn (7)), the corresponding overall success rates would drop down to 87.53% and 73.46%, respectively, clearly indicating the advantage of using the FunD formulation to represent the protein samples for identifying GPCRs and their types.

To evaluate the stability for the results predicted by **GPCR-2L**, we also calculate Matthew's correlation coefficient (MCC) index

Table 2 Success rates in identifying GPCR and non-GPCR by the jackknife test

Protein attribute	Number of proteins	Number of correct prediction	Success rate (%)	MCC
GPCR	367	360	98.09	0.93
Non-GPCR	1101	1068	97.00	0.93
Overall	1468	1428	97.28	

Table 3 Success rates in identifying the six types of GPCRs by the jackknife test

GPCR type	Number of GPCRs	Number of correct prediction	Success rate (%)	MCC
Class A	236	236	100.00	0.96
Class B	39	36	92.31	0.94
Class C	44	43	97.73	0.99
Class D	23	19	82.61	0.88
Class E	7	7	100.00	1.00
Class F	18	18	100.00	1.00
Overall	367	359	97.82	

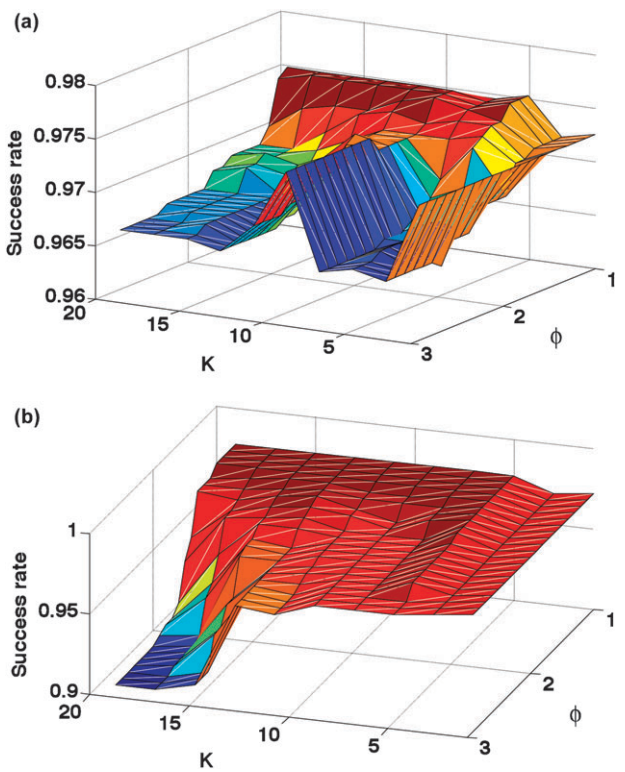

Fig. 4 3D graph to show the variation of the jackknife success rates with the different parameters of φ and K. The values of φ and K used in eqn (9) were determined by optimizing the overall jackknife success rate through a 2-D search. Here the query protein and the protein samples in the training dataset were defined in the 12810-D PseAAC space. Panels (a) and (b) indicate the results for the 1st and 2nd level predictions, respectively.

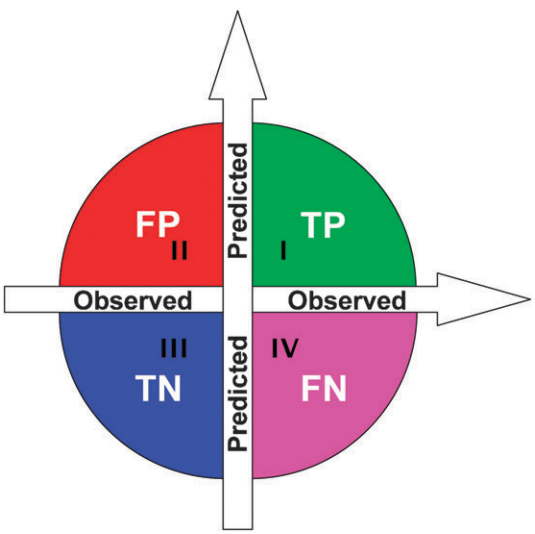

Fig. 5 Distribution of predicted results in four quadrants. (I) TP, the true positive quadrant (green) for correct prediction of the positive dataset, (II) FP, the false positive quadrant (red) for incorrect prediction of the negative dataset; (III) TN, the true negative quadrant (blue) for correct prediction of the negative dataset; and (IV) FN, the false negative quadrant (pink) for incorrect prediction of the positive dataset.

Table 4 Comparison between **GPCR-2L** and **GPCR-AKNN**[108] on an independent testing dataset (ESI S5†) that contains 2192 proteins, none of which occur in the dataset of ESI S1† used to train **GPCR-2L**. The detailed predicted results by the two predictors are given in ESI S6,† where for facilitating comparison, the corresponding experimental observed results are also listed

GPCR families	GPCR-AKNN	GPCR-2L
Rhodopsin like	$\frac{1545}{1619} = 95.43\%$	$\frac{1610}{1619} = 99.44\%$
Secretin like	$\frac{210}{267} = 78.65$	$\frac{250}{267} = 93.63\%$
Metabotropic glutamate/pheromone	$\frac{147}{160} = 91.87\%$	$\frac{149}{160} = 93.13\%$
Fungal pheromone	$\frac{33}{37} = 89.19\%$	$\frac{37}{37} = 100\%$
Frizzled/smoothened family	$\frac{108}{109} = 99.08\%$	$\frac{103}{109} = 94.50\%$
Overall	$\frac{2043}{2192} = \mathbf{93.20}\%$	$\frac{2149}{2192} = \mathbf{98.04}\%$

according to the following equation:

$$\mathrm{MCC} = \frac{(\mathrm{TP})(\mathrm{TN}) - (\mathrm{FP})(\mathrm{FN})}{\sqrt{[\mathrm{TP} + \mathrm{FP}][\mathrm{TP} + \mathrm{FN}][\mathrm{TN} + \mathrm{FP}][\mathrm{TN} + \mathrm{FN}]}} \quad (11)$$

where TP represents the true positive; TN, the true negative; FP, the false positive; and FN, the false negative (see Fig. 5). The MCC values thus obtained are also given in Tables 2 and 3, from which we can see that **GPCR-2L** not only can yield very high success rates in predicting GPCRs and their types, but the results predicted by it are also very stable.

Meanwhile, it was found that during the above prediction 63 different function domains were found from the 367 GPCR proteins in the $\mathbb{S}^{\mathrm{GPCR}}$ dataset, and that 1187 different functional domains were found from the 1101 non-GPCR proteins in the $\mathbb{S}^{\text{non-GPCR}}$ dataset. The names of the 367 and 1187 functional domains are listed in S3 and S4 of the ESI, respectively.†

As a demonstration to show the comparison of the current **GPCR-2L** with the other predictors in this area, an independent testing dataset was constructed that contained 2192 GPCR sequences (ESI S5†) classified into 5 families: (i) rhodopsin like, (ii) secretin like, (iii) metabotropic glutamate/ pheromone, (iv) fungal pheromone, and (v) frizzled/ smoothened family. To avoid bias, none of the proteins included in the testing dataset occur in ESI S1,† the dataset used to train **GPCR-2L**. The reason why the independent dataset only covers 5 (rather than 6) families is because the current GPCRDB contains very few protein sequences for the cAMP GPCR family and that all its seven members have already been assigned to the training dataset with no members left for the independent test. Also, since most of the existing predictors for GPCR either do not provide web-servers, or their web-servers are not working, here let us just compare **GPCR-2L** with GPCR-AKNN.[108] The detailed predicted results by GPCR-AKNN[108] and **GPCR-2L** on each of the 2192 proteins in the independent testing dataset (ESI S5†) are given in ESI S6† where, for facilitating comparison, the corresponding experimental observed results are also given. The overall success rates obtained by the two predictors are summarized in Table 4, from which we can see that **GPCR-2L** outperformed **GPCR-AKNN**[108] by about 5%.

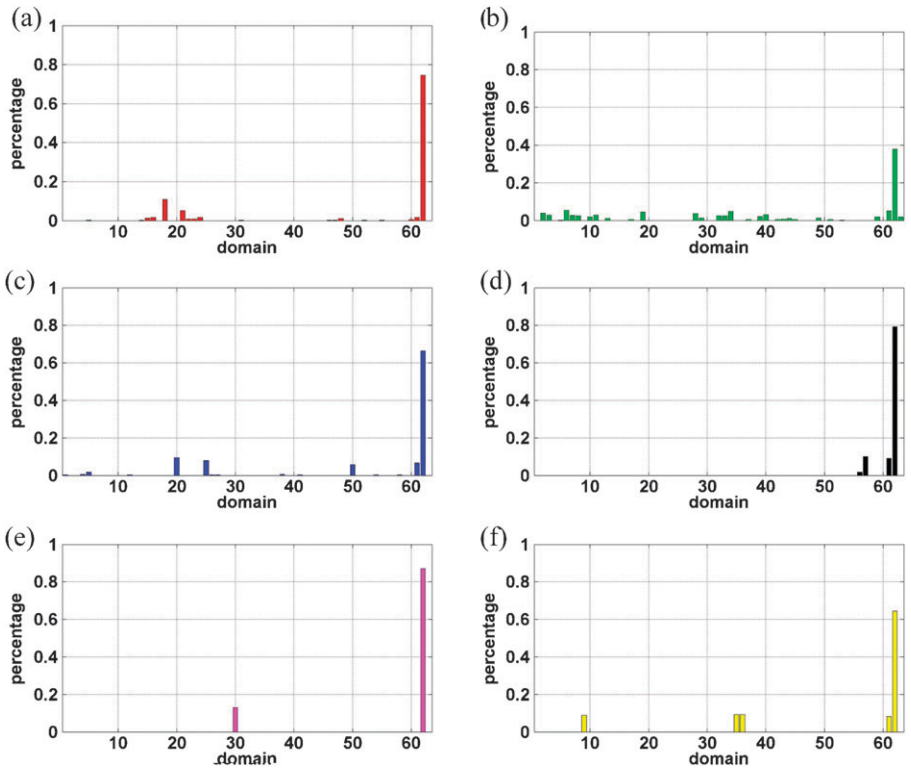

Fig. 6 Distribution of functional domains in different types of GPCRs. The samples of GPCRs were taken from the 367 GPCRs in ESI S1.† The horizontal axis is for the 63 functional domains (see ESI S3†) sorted according to the lexicographic order; the vertical axis is for their occurrence percentages in different GPCR types. Panels (a), (b), (c), (d), (e), and (f) are for GPCR types of rhodopsin-like (Class A), secretin-like (Class B), metabotropic glutamate/ pheromone (Class C), fungal pheromone, cAMP receptor (Class E), and frizzled/smoothened family (Class F), respectively.

It is instructive to point out that the reason why the current approach can achieve such high success rates in identifying GPCRs and their types is that the descriptor or formulation used in this study to represent the protein samples has truly grasped the core of the target. This is one of the indispensable prerequisites for establishing a successful statistical predictor, as elaborated in a recent comprehensive review.[34] Owing to the sample descriptors of eqn (2) and (7), proteins with the same target class are highly clustered while proteins with different target classes are distinctly separated. For instance, this can be visually seen from Fig. 6, where the statistical distributions of different GPCR types with their domains are depicted. It can be seen from the Figure that GPCRs of the same type have the functional domains mainly concentrated in a quite narrow regions, and that GPCRs of different types have remarkably different functional domain distributions, which is particularly true as shown by the bars at the location 62 of the horizontal axis.

Acknowledgements

The authors wish to expresses their gratitude to the three anonymous reviewers, whose constructive comments were very helpful for improving the presentation of the paper. The work in this research was supported by the grants from the National Natural Science Foundation of China (no. 60961003), the Key Project of Chinese Ministry of Education (no.210116), the Province National Natural Science Foundation of Jiangxi(no.2009GZS0064), the Department of Education of JiangXi Province (No.GJJ09271), and the plan for training youth scientists (stars of Jing-Gang) of province Jiangxi.

References

1 T. Gudermann, *J. Mol. Med.*, 1995, **73**, 51.
2 B. L. Roth, D. L. Willins and W. K. Kroeze, *Drug Alcohol Depend.*, 1998, **51**, 73–85.
3 J. M. Baldwin, *Curr. Opin. Cell Biol.*, 1994, **6**, 180–190.
4 R. J. Lefkowitz, *Nat. Cell Biol.*, 2000, **2**, E133.
5 K. C. Chou and D. W. Elrod, *J. Proteome Res.*, 2002, **1**, 429–433.
6 K. Oxenoid and J. J. Chou, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 10870–10875.
7 M. E. Call, J. R. Schnell, C. Xu, R. A. Lutz, J. J. Chou and K. W. Wucherpfennig, *Cell*, 2006, **127**, 355–368.
8 S. M. Douglas, J. J. Chou and W. M. Shih, *Proc. Natl. Acad. Sci. U. S. A.*, 2007, **104**, 6644–6648.
9 J. R. Schnell and J. J. Chou, *Nature*, 2008, **451**, 591–595.
10 J. Wang, R. M. Pielak, M. A. McClintock and J. J. Chou, *Nat. Struct. Mol. Biol.*, 2009, **16**, 1267–1271.
11 R. M. Pielak, R. Jason, J. R. Schnell and J. J. Chou, *Proc. Natl. Acad. Sci. U. S. A.*, 2009, **106**, 7379–7384.
12 M. E. Call, K. W. Wucherpfennig and J. J. Chou, *Nat. Immunol.*, 2010, **11**, 1023–1029.
13 R. M. Pielak and J. J. Chou, *Biochim. Biophys. Acta*, 2010, DOI: 10.1016/j.bbamem.2010.1004.1015.
14 R. M. Pielak and J. J. Chou, *Biochem. Biophys. Res. Commun.*, 2010, **401**, 58–63.
15 K. C. Chou, D. Q. Wei and W. Z. Zhong, *Biochem. Biophys. Res. Commun.*, 2003, **308**, 148–151.
16 K. C. Chou, *Biochem. Biophys. Res. Commun.*, 2004, **316**, 636–642.
17 K. C. Chou, *Biochem. Biophys. Res. Commun.*, 2004, **319**, 433–438.
18 K. C. Chou, *J. Proteome Res.*, 2004, **3**, 1284–1288.
19 D. Q. Wei, Q. S. Du, H. Sun and K. C. Chou, *Biochem. Biophys. Res. Commun.*, 2006, **344**, 1048–1055.
20 S. Q. Wang, Q. S. Du and K. C. Chou, *Biochem. Biophys. Res. Commun.*, 2007, **354**, 634–640.
21 J. F. Wang and K. C. Chou, *Curr. Drug Metab.*, 2010, **11**, 342–346.
22 S. Q. Wang, X. C. Cheng, W. L. Dong, R. L. Wang and K. C. Chou, *Biochem. Biophys. Res. Commun.*, 2010, **401**, 188–191.
23 K. C. Chou, *Curr. Med. Chem.*, 2004, **11**, 2105–2134.
24 K. C. Chou and D. W. Elrod, *Protein Eng.*, 1999, **12**, 107–118.
25 K. C. Chou and D. W. Elrod, *Proteins: Struct., Funct., Genet.*, 1999, **34**, 137–153.
26 D. W. Elrod and K. C. Chou, *Protein Eng.*, 2002, **15**, 713–715.
27 K. C. Chou, *J. Proteome Res.*, 2005, **4**, 1413–1418.
28 Q. B. Gao and Z. Z. Wang, *Protein Eng., Des. Sel.*, 2006, **19**, 511–516.
29 Z. Wen, M. Li, Y. Li, Y. Guo and K. Wang, *Amino Acids*, 2007, **32**, 277–283.
30 J. D. Qiu, J. H. Huang, R. P. Liang and X. Q. Lu, *Anal. Biochem.*, 2009, **390**, 68–73.
31 X. Xiao, P. Wang and K. C. Chou, *J. Comput. Chem.*, 2009, **30**, 1414–1423.
32 Q. Gu, Y. S. Ding and T. L. Zhang, *Protein Pept. Lett.*, 2010, **17**, 559–567.
33 K. C. Chou and H. B. Shen, *Nat. Sci.*, 2009, **2**, 63–92 (openly accessible at http://www.scirp.org/journal/NS/).
34 K. C. Chou, *Curr. Proteomics*, 2009, **6**, 262–274.
35 F. Horn, J. Weare, M. W. Beukers, S. Horsch, A. Bairoch, W. Chen, O. Edvardsen, F. Campagne and G. Vriend, *Nucleic Acids Res.*, 1998, **26**, 275–279.
36 W. Li and A. Godzik, *Bioinformatics*, 2006, **22**, 1658–1659.
37 S. F. Altschul, in *Theoretical and Computational Methods in Genome Research*, ed. S. Suhai, Plenum, New York, 1997, pp. 1–14.
38 J. C. Wootton and S. Federhen, *Comput. Chem.*, 1993, **17**, 149–163.
39 P. Klein and C. Delisi, *Biopolymers*, 1986, **25**, 1659–1672.
40 P. Klein, *Biochim. Biophys. Acta*, 1986, **874**, 205–215.
41 H. Nakashima, K. Nishikawa and T. Ooi, *J. Biochem.*, 1986, **99**, 152–162.
42 B. A. Metfessel, P. N. Saurugger, D. P. Connelly and S. T. Rich, *Protein Sci.*, 1993, **2**, 1171–1182.
43 K. C. Chou and C. T. Zhang, *J. Biol. Chem.*, 1994, **269**, 22014–22020.
44 H. Nakashima and K. Nishikawa, *J. Mol. Biol.*, 1994, **238**, 54–61.
45 K. C. Chou, *Proteins: Struct., Funct., Genet.*, 1995, **21**, 319–344.
46 J. Cedano, P. Aloy, J. A. P'erez-Pons and E. J. Querol, *J. Mol. Biol.*, 1997, **266**, 594–600.
47 G. P. Zhou, *J. Protein Chem.*, 1998, **17**, 729–738.
48 W. Liu and K. C. Chou, *J. Protein Chem.*, 1998, **17**, 209–217.
49 G. P. Zhou and N. Assa-Munt, *Proteins: Struct., Funct., Genet.*, 2001, **44**, 57–59.
50 G. P. Zhou and K. Doctor, *Proteins: Struct., Funct., Genet.*, 2003, **50**, 44–48.
51 K. Y. Feng, Y. D. Cai and K. C. Chou, *Biochem. Biophys. Res. Commun.*, 2005, **334**, 213–217.
52 Q. S. Du, Z. Q. Jiang, W. Z. He, D. P. Li and K. C. Chou, *J. Biomol. Struct. Dyn.*, 2006, **23**, 635–640.
53 B. Niu, Y. D. Cai, W. C. Lu, G. Y. Zheng and K. C. Chou, *Protein Pept. Lett.*, 2006, **13**, 489–492.
54 S. Jahandideh, P. Abdolmaleki, M. Jahandideh and E. B. Asadabadi, *Biophys. Chem.*, 2007, **128**, 87–93.
55 K. C. Chou, *Proteins: Struct., Funct., Genet.*, 2001, **43**, 246–255 (*Erratum*: *ibid.*, 2001, Vol. **44**, 60).
56 K. C. Chou, *Biochem. Biophys. Res. Commun.*, 2000, **278**, 477–483.
57 X. Xiao, S. Shao, Y. Ding, Z. Huang, Y. Huang and K. C. Chou, *Amino Acids*, 2005, **28**, 57–61.
58 X. Xiao, S. H. Shao, Y. S. Ding, Z. D. Huang and K. C. Chou, *Amino Acids*, 2006, **30**, 49–54.
59 X. Xiao, S. H. Shao, Z. D. Huang and K. C. Chou, *J. Comput. Chem.*, 2006, **27**, 478–482.
60 X. B. Zhou, C. Chen, Z. C. Li and X. Y. Zou, *J. Theor. Biol.*, 2007, **248**, 546–551.
61 Y. S. Ding and T. L. Zhang, *Pattern Recognit. Lett.*, 2008, **29**, 1887–1892.

62 G. Y. Zhang and B. S. Fang, *J. Theor. Biol.*, 2008, **253**, 310–315.
63 X. Xiao, P. Wang and K. C. Chou, *J. Theor. Biol.*, 2008, **254**, 691–696.
64 Y. H. Zeng, Y. Z. Guo, R. Q. Xiao, L. Yang, L. Z. Yu and M. L. Li, *J. Theor. Biol.*, 2009, **259**, 366–372.
65 H. Lin, H. Wang, H. Ding, Y. L. Chen and Q. Z. Li, *Acta Biotheor.*, 2009, **57**, 321–330.
66 X. Xiao, P. Wang and K. C. Chou, *J. Appl. Crystallogr.*, 2009, **42**, 169–173.
67 H. Lin, H. Ding, F. B. Feng-Biao Guo, A. Y. Zhang and J. Huang, *Protein Pept. Lett.*, 2008, **15**, 739–744.
68 H. J. Lin, *J. Theor. Biol.*, 2008, **252**, 350–356.
69 F. M. Li and Q. Z. Li, *Protein Pept. Lett.*, 2008, **15**, 612–616.
70 X. Jiang, R. Wei, T. L. Zhang and Q. Gu, *Protein Pept. Lett.*, 2008, **15**, 392–396.
71 D. N. Georgiou, T. E. Karakasidis, J. J. Nieto and A. J. Torres, *J. Theor. Biol.*, 2009, **257**, 17–26.
72 Y. Fang, Y. Guo, Y. Feng and M. Li, *Amino Acids*, 2008, **34**, 103–109.
73 H. Ding, L. Luo and H. Lin, *Protein Pept. Lett.*, 2009, **16**, 351–355.
74 C. Chen, L. Chen, X. Zou and P. Cai, *Protein Pept. Lett.*, 2009, **16**, 27–31.
75 H. Gonzalez-Diaz, Y. Gonzalez-Diaz, L. Santana, F. M. Ubeira and E. Uriarte, *Proteomics*, 2008, **8**, 750–778.
76 Z. S. He, J. Zhang, X. H. Shi, L. L. Hu, X. G. Kong, Y. D. Cai and K. C. Chou, *PLoS One*, 2010, **5**, e9603.
77 K. C. Chou, *FEBS Lett.*, 1995, **363**, 123–126.
78 K. C. Chou, A. G. Tomasselli and R. L. Heinrikson, *FEBS Lett.*, 2000, **470**, 249–256.
79 J. Murvai, K. Vlahovicek, E. Barta and S. Pongor, *Nucleic Acids Res.*, 2001, **29**, 58–60.
80 K. C. Chou and Y. D. Cai, *J. Biol. Chem.*, 2002, **277**, 45765–45769.
81 Y. D. Cai, G. P. Zhou and K. C. Chou, *Biophys. J.*, 2003, **84**, 3257–3263.
82 R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D. A. Natale, *BMC Bioinformatics*, 2003, **4**, 41.
83 R. D. Finn, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer and A. Bateman, *Nucleic Acids Res.*, 2006, **34**, D247–251.
84 I. Letunic, R. R. Copley, B. Pils, S. Pinkert, J. Schultz and P. Bork, *Nucleic Acids Res.*, 2006, **34**, D257–260.
85 A. Marchler-Bauer, J. B. Anderson, M. K. Derbyshire, C. DeWeese-Scott, N. R. Gonzales, M. Gwadz, L. Hao, S. He, D. I. Hurwitz, J. D. Jackson, Z. Ke, D. Krylov, C. J. Lanczycki, C. A. Liebert, C. Liu, F. Lu, S. Lu, G. H. Marchler, M. Mullokandov, J. S. Song, N. Thanki, R. A. Yamashita, J. J. Yin, D. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2007, **35**, D237–240.
86 H. Liu, M. Wang and K. C. Chou, *Biochem. Biophys. Res. Commun.*, 2005, **336**, 737–739.
87 T. P. Hopp and K. R. Woods, *Proc. Natl. Acad. Sci. U. S. A.*, 1981, **78**, 3824–3828.
88 A. V. Oppenheim, A. S. Willsky and S. H. Nawab, *Signals and Systems*, Prentice Hall, New York, 1985.
89 K. C. Chou and N. Y. Chen, *Sci. Sin.*, 1977, **20**, 447–457.
90 K. C. Chou, *Biophys. Chem.*, 1988, **30**, 3–48.
91 K. C. Chou, *Trends Biochem. Sci.*, 1989, **14**, 212.
92 T. M. Cover and P. E. Hart, *IEEE Trans. Inf. Theory*, 1967, **IT–13**, 21–27.
93 K. C. Chou and H. B. Shen, *PLoS One*, 2010, **5**, e11335.
94 K. C. Chou and H. B. Shen, *PLoS One*, 2010, **5**, e9931; openly accessible at http://www.plosone.org/article/info%9933Adoi%9932F9910.1371%9932Fjournal.pone.0009931.
95 K. C. Chou and H. B. Shen, *Biochem. Biophys. Res. Commun.*, 2007, **360**, 339–345.
96 K. C. Chou and H. B. Shen, *Biochem. Biophys. Res. Commun.*, 2008, **376**, 321–325.
97 J. M. Keller, M. R. Gray and J. A. Givens, *IEEE Trans. Syst. Man. Cybern.*, 1985, **15**, 580–585.
98 C. T. Zhang, K. C. Chou and G. M. Maggiora, *Protein Eng.*, 1995, **8**, 425–435.
99 H. B. Shen, J. Yang and K. C. Chou, *J. Theor. Biol.*, 2006, **240**, 9–13.
100 Y. S. Ding, T. L. Zhang and K. C. Chou, *Protein Pept. Lett.*, 2007, **14**, 811–815.
101 K. C. Chou and C. T. Zhang, *Crit. Rev. Biochem. Mol. Biol.*, 1995, **30**, 275–349.
102 K. C. Chou and H. B. Shen, *Nat. Protoc.*, 2008, **3**, 153–162.
103 K. C. Chou and H. B. Shen, *Nat. Sci.*, 2010, **2**, 1090–1103 (openly accessible at http://www.scirp.org/journal/NS/).
104 K. C. Chou and H. B. Shen, *Anal. Biochem.*, 2007, **370**, 1–16.
105 M. Esmaeili, H. Mohabatkar and S. Mohsenzadeh, *J. Theor. Biol.*, 2010, **263**, 203–209.
106 G. Y. Zhang, H. C. Li, J. Q. Gao and B. S. Fang, *Protein Pept. Lett.*, 2008, **15**, 1132–1137.
107 L. Chen, T. Huang, X. H. Shi, Y. D. Cai and K. C. Chou, *Molecules*, 2010, **15**, 8177–8192 (Openly accessible at www.mdpi.com/journal/molecules).
108 X. Xiao and W. R. Qiu, *Interdiscip. Sci.: Comput. Life Sci.*, 2010, **2**, 180–184.