

iGPCR-Drug: A Web Server for Predicting Interaction between GPCRs and Drugs in Cellular Networking

Xuan Xiao^{1,2,4*}, Jian-Liang Min¹, Pu Wang¹, Kuo-Chen Chou^{3,4}

1 Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, China, **2** Information School, ZheJiang Textile and Fashion College, NingBo, China, **3** Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia, **4** Gordon Life Science Institute, Belmont, Massachusetts, United States of America

Abstract

Involved in many diseases such as cancer, diabetes, neurodegenerative, inflammatory and respiratory disorders, G-protein-coupled receptors (GPCRs) are among the most frequent targets of therapeutic drugs. It is time-consuming and expensive to determine whether a drug and a GPCR are to interact with each other in a cellular network purely by means of experimental techniques. Although some computational methods were developed in this regard based on the knowledge of the 3D (dimensional) structure of protein, unfortunately their usage is quite limited because the 3D structures for most GPCRs are still unknown. To overcome the situation, a sequence-based classifier, called "iGPCR-drug", was developed to predict the interactions between GPCRs and drugs in cellular networking. In the predictor, the drug compound is formulated by a 2D (dimensional) fingerprint via a 256D vector, GPCR by the PseAAC (pseudo amino acid composition) generated with the grey model theory, and the prediction engine is operated by the fuzzy K-nearest neighbour algorithm. Moreover, a user-friendly web-server for iGPCR-drug was established at <http://www.jci-bioinfo.cn/iGPCR-Drug/>. For the convenience of most experimental scientists, a step-by-step guide is provided on how to use the web-server to get the desired results without the need to follow the complicated math equations presented in this paper just for its integrity. The overall success rate achieved by iGPCR-drug via the jackknife test was 85.5%, which is remarkably higher than the rate by the existing peer method developed in 2010 although no web server was ever established for it. It is anticipated that iGPCR-Drug may become a useful high throughput tool for both basic research and drug development, and that the approach presented here can also be extended to study other drug – target interaction networks.

Citation: Xiao X, Min J-L, Wang P, Chou K-C (2013) iGPCR-Drug: A Web Server for Predicting Interaction between GPCRs and Drugs in Cellular Networking. PLoS ONE 8(8): e72234. doi:10.1371/journal.pone.0072234

Editor: Seema Singh, University of South Alabama Mitchell Cancer Institute, United States of America

Received: April 25, 2013; **Accepted:** July 8, 2013; **Published:** August 27, 2013

Copyright: © 2013 Xiao et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by the grants from the National Natural Science Foundation of China (60961003 and 31260273), the Key Project of Chinese Ministry of Education (210116), the Province National Natural Science Foundation of JiangXi (2010GZS0122, 20114BAB211013 and 20122BAB201020), the Department of Education of JiangXi Province (GJJ12490), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (20120BDH80023), and the JiangXi Provincial Foundation for Leaders of Disciplines in Science (20113BCB22008). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: xxiao@gordonlifescience.org

Introduction

G-protein-coupled receptors (GPCRs), also known as G protein-linked receptors (GPLR), serpentine receptor, seven-transmembrane domain receptors, and 7 TM (transmembrane), form the largest family of cell surface receptors. GPCRs share a common global topology that consists of seven transmembrane alpha helices, intracellular C-terminal, an extracellular N-terminal, three intracellular loops and three extracellular loops (**Fig. 1**).

GPCR-associated proteins may play at least the following four distinct roles in receptor signaling: **(1)** directly mediate receptor signaling, as in the case of G proteins; **(2)** regulate receptor signaling through controlling receptor localization and/or trafficking; **(3)** act as a scaffold, physically linking the receptor to various effectors; **(4)** act as an allosteric modulator of receptor conformation, altering receptor pharmacology and/or other aspects of receptor function [1,2,3]. Involved in many diseases such as cancer, diabetes, neurodegenerative, inflammatory and respiratory disorders, GPCRs are among the most frequent targets of therapeutic drugs [4]. Over half of all prescription drugs currently on the market are actually acting by targeting GPCRs

directly or indirectly [5,6]. A lot of efforts have been invested for studying GPCRs in both academic institutions and pharmaceutical industries.

Identification of drug-target interactions is an essential step in the drug discovery process, which is the most important task for the new medicine development [7]. The methods commonly used in this regard are docking simulations [8,9], literature text mining [10], as well as a combination of chemical structure, genomic sequence, and 3D (three-dimensional) structure information, among others [11]. Obviously, an experimental 3D structure of a target protein is the key for identifying the drug-protein interaction; if it is not available, the common approach is to create a homology model of the target protein based on the experimental structure of a related protein [12,13,14]. However, the above methods need further development due to the following reasons. (1) None of these methods has provided a web-server for the public usage, and hence their practical application value is quite limited. (2) The prediction quality needs to be improved with the state-of-the-art machine learning techniques and updated training datasets. (3) GPCRs belong to membrane proteins, which

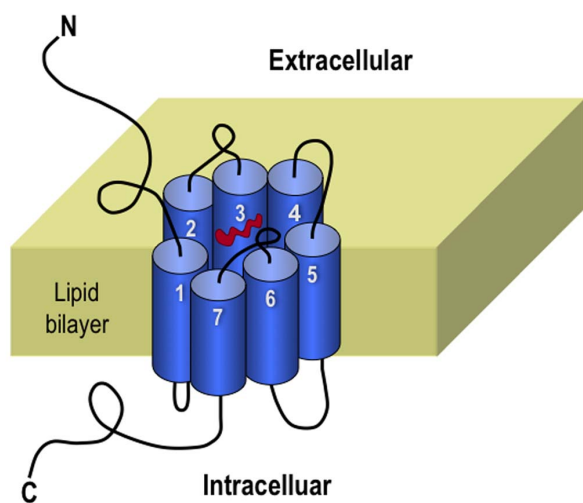


Figure 1. Schematic drawing of a GPCR. It consists of seven transmembrane alpha helices, intracellular C-terminal, an extracellular N-terminal, three intracellular loops and three extracellular loops. Reproduced from [4] with permission. doi:10.1371/journal.pone.0072234.g001

are very difficult to crystallize and most of them will not dissolve in normal solvents. Although a series of recent reports [15,16,17,18,19,20,21] have demonstrated that NMR is indeed a very powerful tool in determining the 3D structures of membrane proteins, it is time-consuming and costly. Also, although using various structural bioinformatics tools [12], particularly the homologous technique [22,23,24,25,26], can help acquire the structural and functional information of membrane proteins, unfortunately the number of templates for membrane proteins is quite limited. Therefore, it would be very useful to develop a computational method for predicting the interactions between drugs and GPCRs in cellular networking based on their sequence-derived features before a drug candidate was synthesized, so as to reduce the unnecessary waste of time and money [27]. And this is the goal of the current study.

According to a recent comprehensive review [28] and demonstrated by a series of recent publications (see, e.g., [29,30,31]), to establish a really useful statistical predictor for a protein system, we need to consider the following procedures: (i) construct or select a valid benchmark dataset to train and test the predictor; (ii) formulate the protein samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be predicted; (iii) introduce or develop a powerful algorithm (or engine) to operate the prediction; (iv) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (v) establish a user-friendly web-server for the predictor that is accessible to the public. Below, let us describe how to deal with these steps.

Methodology

1. Benchmark Dataset

The benchmark dataset \mathcal{S} can be formulated as

$$\mathcal{S} = \mathcal{S}^+ \cup \mathcal{S}^- \quad (1)$$

where \mathcal{S}^+ is the positive subset that consists of the interactive GPCR-drug pairs only, while \mathcal{S}^- the negative subset that contains of the non-interactive GPCR-drug pairs only, and the symbol \cup

represents the union in the set theory. Here, the “interactive” pair means the pair whose two counterparts are interacted with each other in the drug-target networks as defined in the KEGG database at <http://www.kegg.jp/kegg/>; while the “non-interactive” pair means that its two counterparts are not interacted with each other in the drug-target networks. The positive dataset \mathcal{S}^+ contains 620 GPCR-drug pairs, which were taken from [32]. The negative dataset \mathcal{S}^- contains 1,240 non-interactive GPCR-drug pairs, which were derived according to the following procedures as done in [32]: (i) separating each of the pairs in \mathcal{S}^+ into single drug and GPCR; (ii) re-coupling each of the single drugs with each of the single GPCRs into pairs in a way that none of them occurred in \mathcal{S}^+ ; (iii) randomly picking the pairs thus formed until they reached the number two times as many as the pairs in \mathcal{S}^+ . The 620 interactive GPCR-drug pairs and 1,240 non-interactive GPCR-drug pairs are given in Supporting Information S1. All the detailed information for the compounds or drugs listed there can be found in the KEGG database via their codes.

2. Sample Representation

Since each of the samples in the current network system contains a GPCR (protein) and a drug, a combination of the following two approaches were adopted to represent the GPCR–drug pair samples.

(a) Representing drugs with 2D molecular fingerprints. Although the number of drugs is extremely large, most of them are small organic molecules and are composed of some fixed small structures [33]. The identification of small molecules or structures can be used to detect the drug-target interactions [34]. Molecular fingerprints are bit-string representations of molecular structure and properties [35]. It should be pointed out that there are many types of structural representation that have been suggested for the description of drug molecules, including physicochemical properties [36], chemical graphs [37], topological indices [38], 3D pharmacophore patterns and molecular fields. In the current study, let us use the simple and generally adopted 2D molecular fingerprints to represent drug molecules, as described below.

First, for each of the drugs concerned, we can obtain a MOL file from the KEGG database [39] via its code that contains the detailed information of chemical structure. Second, we can convert the MOL file format into its 2D molecular fingerprint file format by using a chemical toolbox software called OpenBabel [40], which can be downloaded from the website at <http://openbabel.org/>. The current version of OpenBabel can generate four types of fingerprints: FP2, FP3, FP4 and MACCS. In the current study, we used the FP2 fingerprint format. It is a path-based fingerprint that identifies small molecule fragments based on all linear and ring substructures and maps them onto a bit-string using a hash function (somewhat similar to the Daylight fingerprints [41,42]). It is a length of 256-bit hexadecimal string or a 256-bit vector, whose component values are an integer between 0 and 15. Let us suppose V_1 is the 1st value of the 256-bit vector, V_2 that of the 2nd value, and so forth. Thus, the 256-bit vector can be converted to a digit signal. In order to find the inwardness of the drug fingerprint values, we implement the discrete Fourier transform, with the frequency-domain values given by

$$X_k = \sum_{l=1}^{256} V_l \exp \left[-j \left(\frac{2\pi l}{256} \right) k \right], \quad (k = 1, 2, \dots, 256) \quad (2)$$

where j represents the imaginary unit and X_k is a complex number

whose complex modulus or amplitude is given by

$$A_k = [\text{real}^2(X_k) + \text{imag}^2(X_k)]^{1/2} \quad (3)$$

where $\text{real}(X_k)$ is the real part of X_k and $\text{imag}(X_k)$ the corresponding image part. Thus we can generate the discrete Fourier spectrum as given by

$$\{A_1, A_2, \dots, A_{256}\} \quad (4)$$

The Fourier spectrum numbers contain substantial information about the digit signal, and hence can also be used to reflect certain characters of a drug. Thus, a drug compound \mathbf{d} now can be formulated as a 256-D (dimensional) vector given by

$$\mathbf{d} = [A_1 \ A_1 \ \dots \ A_j \ \dots \ A_{256}]^T \quad (5)$$

where A_j has the same meaning as in **Eq. 4**, and \mathbf{T} is the matrix transpose operator.

The 256-D vector thus obtained for each of the drug codes listed in Supporting Information S1 are given in Supporting Information S2.

(b) Representing GPCR sequences with grey model pseudo amino acid composition. The sequences of the GPCRs involved in this study are given in Supporting Information S3. Now the problem is how to effectively represent these receptor sequences for the current study. Generally speaking, there are two kinds of approaches to formulate protein sequences: the sequential model and the non-sequential or discrete model [43]. The most typical sequential representation for a protein sample with L residues is its entire amino acid sequence, as can be formulated as

$$\mathbf{P} = R_1 R_2 R_3 R_4 R_5 R_6 R_7 \dots R_L \quad (6)$$

where R_1 represents the 1st residue of the protein sequence \mathbf{P} , R_2 the 2nd residue, and so forth. A protein thus formulated can contain its most complete information. This is an obvious advantage of the sequential representation. To get the desired results, the sequence-similarity-search-based tools, such as BLAST [44,45], are usually utilized to conduct the prediction. However, this kind of approach failed to work when the query protein did not have significant homology to proteins of known characters. Thus, various non-sequential representation models were proposed. The simplest non-sequential model for a protein was based on its amino acid composition (AAC), as defined by

$$\mathbf{P} = [f_1 \ f_2 \ \dots \ f_{20}]^T \quad (7)$$

where $f_u (u=1,2,\dots,20)$ are the normalized occurrence frequencies of the 20 native amino acids [46,47] in the protein \mathbf{P} , and \mathbf{T} has the same meaning as in Eq. 5. The AAC-discrete model was widely used for identifying various attributes of proteins. However, as can be seen from Eq. 7, all the sequence order effects were lost by using the AAC-discrete model. This is its main shortcoming. To avoid completely losing the sequence-order information, the pseudo amino acid composition was proposed [48] to replace the simple amino acid composition (AAC) for representing the sample of a protein. Since the concept of PseAAC (also called ‘‘Chou’s PseAAC’’ [49]) was proposed in 2001 [48], it has been widely used to study various attributes of proteins, such as discriminating outer membrane proteins [50],

identifying antibacterial peptides [51], identifying allergenic proteins [52], predicting metalloproteinase family [53], predicting protein structural class [54], identifying bacterial virulent proteins [55], predicting supersecondary structure [56], predicting protein subcellular location [57,58,59,60], predicting membrane protein types [61,62], identifying GPCRs and their types [63], identifying protein quaternary structural attributes [64], predicting protein submitochondria locations [65], identifying risk type of human papillomaviruses [66], identifying cyclin proteins [67], predicting GABA(A) receptor proteins [68], classifying amino acids [69], predicting cysteine S-nitrosylation sites in proteins [30], among many others (see a long list of papers cited in the References section of [28]). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides [29,31], as well as other biological samples (see, e.g., [70,71]). Because it has been widely and increasingly used, recently two powerful soft-wares called ‘‘PseAAC-Builder’’ [72] and ‘‘propy’’ [73] were established for generating various special Chou’s pseudo-amino acid compositions, in addition to the web-server PseAAC [74] built in 2008. According to a recent review [28], the general form of PseAAC for a protein \mathbf{P} is formulated by

$$\mathbf{P} = [\psi_1 \ \psi_2 \ \dots \ \psi_u \ \dots \ \psi_\Omega]^T \quad (8)$$

where the subscript Ω is an integer, and its value as well as the components $\psi_u (u=1,2,\dots,\Omega)$ will depend on how to extract the desired information from the amino acid sequence of \mathbf{P} (cf. **Eq. 6**). Below, let us describe how to extract useful information from the benchmark dataset \mathbb{S} and Supporting Information S3 to define the GPCR samples concerned via Eq. 8.

First, let us represent the protein sequence by a series of real numbers. Listed in **Table 1** are the ten different kinds of physicochemical properties usually used for identifying protein attributes [75]. For the current study, however, it was found through many preliminary tests that when the 10th physicochemical property (i.e., the ‘‘mean polarity’’) was used, the best prediction quality was observed. This is quite consistent with the observations that polar amino acids play an important role in membrane protein receptors [12,17]. Accordingly, the 20 numerical values of the mean polarity in **Table 1** were used to encode the 20 native amino acids in a GPCR sequence. Note that to ensure that each of these numerical codes was a positive number as required by the Grey model used later, during the encoding process, each of the mean polarity values in **Table 1** was added by a number of 1.20. Thus, for a given GPCR sequence with L amino acids (cf. **Eq. 6**), we can convert it into a series of L real number as formulated by

$$\mathbf{P} = (r_1 \ r_2 \ \dots \ r_L) \quad (9)$$

where r_1 is the mean polarity value for the 1st amino acid residue in the GPCR protein \mathbf{P} , e.g., if the 1st residue is A, then we have $r_1 = (-0.06 + 1.20) = 1.14$; r_2 is the mean polarity value for the 2nd amino acid residue plus 1.20; and so forth. Now, we can use the grey system model to extract the useful information of \mathbf{P} via **Eq. 8** to formulate its PseAAC.

According to the grey system theory [76], 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 model developed based on such a theory is called ‘‘grey model’’, which is a kind of nonlinear and dynamic model formulated by a differential equation. The grey model is particularly useful for solving complicated problems [77] that are

Table 1. Ten physicochemical property codes for each of the 20 native amino acids^a.

Amino acid	Ten physicochemical property codes ^b									
	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
A	0.62	-0.50	15	2.35	9.87	6.11	91.50	89.09	27.5	-0.06
C	0.29	-1.00	47	1.71	10.78	5.02	117.7	121.2	44.6	1.36
D	-0.90	3.00	59	1.88	9.60	2.98	124.5	133.1	40.0	-0.80
E	-0.74	3.00	73	2.19	9.67	3.08	155.1	147.1	62.0	-0.77
F	1.19	22.50	91	2.58	9.24	5.91	203.4	165.2	115.5	1.27
G	0.48	0.00	1	2.34	9.60	6.06	66.40	75.07	0.0	-0.41
H	-0.40	20.50	82	1.78	8.97	7.64	167.3	155.2	79.0	0.49
I	1.38	21.80	57	2.32	9.76	6.04	168.8	131.2	93.5	1.31
K	-1.50	3.00	73	2.20	8.90	9.47	171.3	146.2	100.0	-1.18
L	1.06	21.80	57	2.36	9.60	6.04	167.9	131.2	93.5	1.21
M	0.64	21.30	75	2.28	9.21	5.74	170.8	149.2	94.1	1.27
N	-0.78	0.20	58	2.18	9.09	10.76	135.2	132.1	58.7	-0.48
P	0.12	0.00	42	1.99	10.60	6.30	129.3	115.1	41.9	0.00
Q	-0.85	0.20	72	2.17	9.13	5.65	161.1	146.2	80.7	-0.73
R	-2.53	3.00	101	2.18	9.09	10.76	202.0	174.2	105	-0.84
S	-0.18	0.30	31	2.21	9.15	5.68	99.10	105.1	29.3	-0.50
T	-0.05	20.40	45	2.15	9.12	5.60	122.1	119.1	51.3	-0.27
V	1.08	21.50	43	2.29	9.74	6.02	141.7	117.2	71.5	1.09
W	0.81	23.40	130	2.38	9.39	5.88	237.6	204.2	145.5	0.88
Y	0.26	22.30	107	2.20	9.11	5.63	203.6	181.2	117.3	0.33

^aThe numerical codes of the physicochemical properties can be obtained from the text biochemistry book (e.g., [101]) and the papers [102,103].

^bThe 1st physicochemical property is for "hydrophobicity", 2nd for "hydrophilicity", 3rd for "side-chain mass", 4th for "pK1 (C^α-COOH)", 5th for "pK2 (NH3)", 6th for "PI (25°C)", 7th for "average buried volume", 8th for "molecular weight", 9th for "side-chain volume", and 10th for "mean polarity".

doi:10.1371/journal.pone.0072234.t001

lack of sufficient information, or need to process uncertain information and reduce random effects of acquired data. In the grey system theory, an important and generally used model is called GM(1,1) [76]. By following the similar procedures as described in [78,79,80,81], **Eq. 8** would become a feature vector with dimension $\Omega = 22$ and each of its components defined by

$$\psi_u = \begin{cases} \frac{f_u}{\sum_{i=1}^{20} f_i + \sum_{k=1}^2 w_k a_k}, & 1 \leq u \leq 20 \\ \frac{w_{u-20} a_{u-20}}{\sum_{i=1}^{20} f_i + \sum_{k=1}^2 w_k a_k}, & 21 \leq u \leq 22 \end{cases} \quad (10)$$

where f_u has the same meaning as **Eq. 7**, $w_k (k = 1, 2)$ is the weight factor (in this study we choose $w_1 = 10^5$ and $w_2 = 10^2$ to get the best results), and a_1 and a_2 are given by

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{Y} \quad (11)$$

where

$$\mathbf{B} = \begin{bmatrix} -\frac{r_2}{\ln\left(\frac{r_1+r_2}{r_1}\right)} & 1 \\ -\frac{r_3}{\ln\left(\frac{r_1+r_2+r_3}{r_1+r_2}\right)} & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ -\frac{r_L}{\ln\left(\frac{\sum_{k=1}^L r_k}{\sum_{k=1}^{L-1} r_k}\right)} & 1 \end{bmatrix} \quad (12)$$

and

$$\mathbf{Y} = \begin{bmatrix} r_2 \\ r_3 \\ \vdots \\ r_L \end{bmatrix} \quad (13)$$

(c) Representing GPCR-drug pairs. Now the pair between a GPCR protein **P** and a drug compound **D** can be formulated by combing **Eq. 5** and **Eq. 8**, as given by

$$\mathbf{G} = \mathbf{P} \oplus w_3 \mathbf{d} = [\psi_1 \ \cdots \ \psi_{22} \ w_3 A_1 \ \cdots \ w_3 A_{256}]^T \quad (14)$$

where \mathbf{G} represents the GPCR-drug pair, \oplus the orthogonal sum [43], w_3 the weight factor that was chosen as $1/700$ in this study to get the best results, and $\psi_u (u = 1, 2, \dots, 22)$ are given in **Eq. 10**.

3. Fuzzy K -Nearest Neighbor Algorithm

The fuzzy K -Nearest Neighbour (k NN) classification method [82] is quite popular in the pattern recognition community owing to its good performance and ease of use. It is particularly effective in dealing with complicated biological systems, such as identifying nuclear receptor subfamilies [83], characterizing the structure of fast-folding proteins [84], classifying G protein-coupled receptors [85], predicting protein quaternary structural attributes [86], predicting protein structural classes [87,88,89,90], identifying membrane protein types [91], and so forth. The rationale of the fuzzy method is based on the fact that it is impossible to define a feature vector that can contain all the entire information of a complicated system. Therefore, it is logically more reasonable to treat this kind of object as a fuzzy system. Below, let us give a brief introduction how to use the fuzzy k NN approach to identify the interactions between the GPCR and the drug compounds in the network concerned.

Suppose $\mathbb{S}(N) = \{\mathbf{G}_1, \mathbf{G}_2, \dots, \mathbf{G}_N\}$ is a set of vectors representing N GPCR-drug pairs in a training set classified into two classes $\{C^+, C^-\}$, where C^+ denotes the interactive pair class while C^- the non-interactive pair class; $\mathbb{S}^*(\mathbf{G}) = \{\mathbf{G}_1^*, \mathbf{G}_2^*, \dots, \mathbf{G}_K^*\} \subset \mathbb{S}(N)$ is the subset of the K nearest neighbor pairs to the query pair \mathbf{G} . Thus, the fuzzy membership value for the query pair \mathbf{G} in the two classes of $\mathbb{S}(N)$ is given by [92]

$$\begin{cases} \mu^+(\mathbf{G}) = \frac{\sum_{j=1}^K \mu^+(\mathbf{G}_j^*) d(\mathbf{G}, \mathbf{G}_j^*)^{-2/(\varphi-1)}}{\sum_{j=1}^K d(\mathbf{G}, \mathbf{G}_j^*)^{-2/(\varphi-1)}} \\ \mu^-(\mathbf{G}) = \frac{\sum_{j=1}^K \mu^-(\mathbf{G}_j^*) d(\mathbf{G}, \mathbf{G}_j^*)^{-2/(\varphi-1)}}{\sum_{j=1}^K d(\mathbf{G}, \mathbf{G}_j^*)^{-2/(\varphi-1)}} \end{cases} \quad (15)$$

where K is the number of the nearest neighbors counted for the query pair \mathbf{G} ; $\mu^+(\mathbf{G}_j^*)$ and $\mu^-(\mathbf{G}_j^*)$, the fuzzy membership values of the training sample \mathbf{G}_j^* to the class C^+ and C^- , respectively, as will be further defined below; $d(\mathbf{G}, \mathbf{G}_j^*)$, the Euclidean distance between \mathbf{G} and its j th nearest pair \mathbf{G}_j^* in the training dataset $\mathbb{S}(N)$; $\varphi (> 1)$, the fuzzy coefficient for determining how heavily the distance is weighted when calculating each nearest neighbor's contribution to the membership value. Note that the parameters K and φ will affect the computation result of **Eq. 15**, and they will be optimized by a grid-search as will be described later. Also, various other metrics can be chosen for $d(\mathbf{G}, \mathbf{G}_j^*)$, such as Hamming distance [93] and Mahalanobis distance [94,95].

The quantitative definitions for the aforementioned $\mu^+(\mathbf{G}_j^*)$ and $\mu^-(\mathbf{G}_j^*)$ in **Eq. 15** are given by

$$\mu^+(\mathbf{G}_j^*) = \begin{cases} 1, & \text{if } \mathbf{G}_j^* \in C^+ \\ 0, & \text{otherwise} \end{cases} \quad \mu^-(\mathbf{G}_j^*) = \begin{cases} 1, & \text{if } \mathbf{G}_j^* \in C^- \\ 0, & \text{otherwise} \end{cases} \quad (16)$$

Substituting the results obtained by **Eq. 16** into **Eq. 15**, it follows that if $\mu^+(\mathbf{G}) > \mu^-(\mathbf{G})$ then the query pair \mathbf{G} is an interactive couple; otherwise, non-interactive. In other words, the outcome can be formulated as

$$\mathbf{G} \in \begin{cases} C^+, & \text{if } \mu^+(\mathbf{G}) > \mu^-(\mathbf{G}) \\ C^-, & \text{otherwise} \end{cases} \quad (17)$$

The predictor thus established is called **iGPCR-Drug**. To provide an intuitive overall picture of how the classifier works, a flowchart is provided in **Fig. 2** to show its operation process.

Results and Discussion

1. Metrics for Performance Evaluation

To provide a more intuitive and easier-to-understand method to measure the prediction quality, here the criteria proposed in [96] was adopted. According to those criteria, the rates of correct predictions for the interactive GPCR-drug pairs in dataset \mathbb{S}^+ and the non-interactive GPCR-drug pairs in dataset \mathbb{S}^- are respectively defined by (cf. **Eq. 1**).

$$\begin{cases} \Lambda^+ = \frac{N^+ - N_+^+}{N^+}, & \text{for the interactive GPCR-drug pairs} \\ \Lambda^- = \frac{N^- - N_+^-}{N^-}, & \text{for the non-interactive GPCR-drug pairs} \end{cases} \quad (18)$$

where N^+ is the total number of the interactive GPCR-drug pairs investigated while N_+^+ the number of the interactive GPCR-drug pairs incorrectly predicted as the non-interactive GPCR-drug pairs; N^- the total number of the non-interactive GPCR-drug pairs investigated while N_+^- the number of the non-interactive GPCR-drug pairs incorrectly predicted as the interactive GPCR-drug pairs. The overall success prediction rate is given by [97]

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{N_+^+ + N_+^-}{N^+ + N^-} \quad (19)$$

It is obvious from **Eqs. 18–19** that, if and only if none of the interactive GPCR-drug pairs and the non-interactive GPCR-drug pairs are mispredicted, i.e., $N_+^+ = N_+^- = 0$ and $\Lambda^+ = \Lambda^- = 1$, we have the overall success rate $\Lambda = 1$. Otherwise, the overall success rate would be smaller than 1.

On the other hand, it is interesting to point out that the following equation set is often used in literatures for examining the performance quality of a predictor.

$$\begin{cases} \text{Sn} = \frac{\text{TP}}{\text{TP} + \text{FN}} \\ \text{Sp} = \frac{\text{TN}}{\text{TN} + \text{FP}} \\ \text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \\ \text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \end{cases} \quad (20)$$

where TP represents the true positive; TN, the true negative; FP, the false positive; FN, the false negative; Sn, the sensitivity; Sp, the specificity; Acc, the accuracy; MCC, the Mathew's correlation coefficient.

Obviously, the relations between the symbols in **Eq. 18** or **Eq. 19** and those in **Eq. 20** are given by

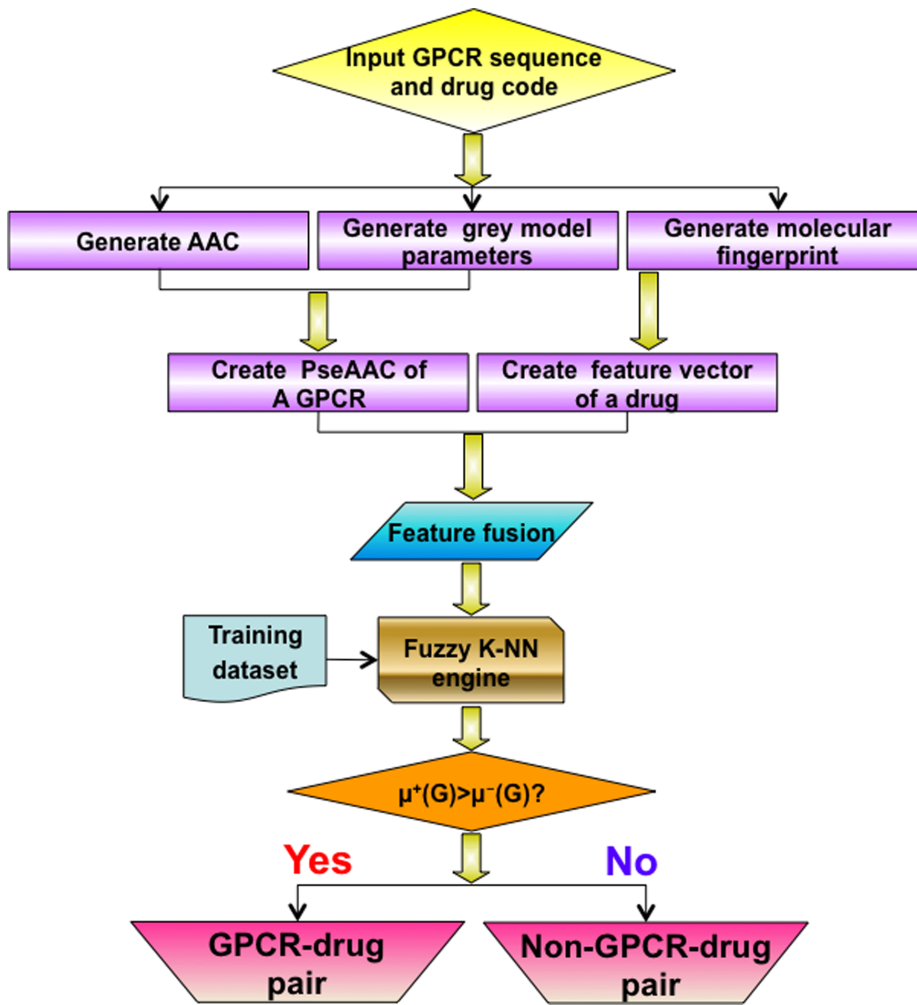


Figure 2. A flowchart to show the operation process of the iGPCR-Drug predictor. See the text for further explanation. doi:10.1371/journal.pone.0072234.g002

$$\begin{cases} TP = N^+ - N_{-}^+ \\ TN = N^- - N_{+}^- \\ FP = N_{+}^- \\ FN = N_{-}^+ \end{cases} \quad (21)$$

Substituting **Eq. 21** into **Eq. 20** and also noting **Eqs. 18–19**, we obtain.

$$\begin{cases} Sn = \Lambda^+ = 1 - \frac{N_{-}^+}{N^+} \\ Sp = \Lambda^- = 1 - \frac{N_{+}^-}{N^-} \\ Acc = \Lambda = 1 - \frac{N_{-}^+ + N_{+}^-}{N^+ + N^-} \\ MCC = \frac{1 - \left(\frac{N_{-}^+ + N_{+}^-}{N^+ + N^-}\right)}{\sqrt{\left(1 + \frac{N_{-}^+ - N_{+}^-}{N^+}\right) \left(1 + \frac{N_{+}^- - N_{-}^+}{N^-}\right)}} \end{cases} \quad (22)$$

Now we can easily see: when $N_{-}^+ = 0$ meaning none of the interactive GPCR-drug pairs was mispredicted to be a non-interactive GPCR-drug pair, we have the sensitivity $Sn = 1$; while $N_{+}^- = N^+$ meaning that all the interactive GPCR-drug pairs were mispredicted to be the non-interactive GPCR-drug pairs, we have the sensitivity $Sn = 0$. Likewise, when $N_{+}^- = 0$ meaning none of the non-interactive GPCR-drug pairs was mispredicted, we have the specificity $Sp = 1$; while $N_{-}^+ = N^-$ meaning all the non-interactive GPCR-drug pairs were incorrectly predicted as interactive GPCR-drug pairs, we have the specificity $Sp = 0$. When $N_{-}^+ = N_{+}^- = 0$ meaning that none of the interactive GPCR-drug pairs in the dataset \mathbb{S}^+ and none of the non-interactive GPCR-drug pairs in \mathbb{S}^- was incorrectly predicted, we have the overall accuracy $Acc = \Lambda = 1$; while $N_{-}^+ = N^+$ and $N_{+}^- = N^-$ meaning that all the interactive GPCR-drug pairs in the dataset \mathbb{S}^+ and all the non-interactive GPCR-drug pairs in \mathbb{S}^- were mispredicted, we have the overall accuracy $Acc = \Lambda = 0$. The MCC correlation coefficient is usually used for measuring the quality of binary (two-class) classifications. When $N_{-}^+ = N_{+}^- = 0$ meaning that none of the interactive GPCR-drug pairs in the dataset \mathbb{S}^+ and none of the non-interactive GPCR-drug pairs in \mathbb{S}^- was mispredicted, we have $MCC = 1$; when $N_{-}^+ = N^+/2$ and $N_{+}^- = N^-/2$ we have $MCC = 0$ meaning no better than random prediction; when

$N_{+}^{-} = N^{+}$ and $N_{-}^{+} = N^{-}$ we have $MCC = -1$ meaning total disagreement between prediction and observation. As we can see from the above discussion, it is much more intuitive and easier-to-understand when using **Eq. 22** to examine a predictor for its sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient.

2. Cross-Validation

How to properly examine the prediction quality is a key for developing a new predictor and estimating its potential application value. Generally speaking, to avoid the “memory effect” [43] of the resubstitution test in which a same dataset was used to train and test a predictor, the following three cross-validation methods are often used to examine a predictor for its effectiveness in practical application: independent dataset test, subsampling or K -fold (such as 5-fold, 7-fold, or 10-fold) test, and jackknife test [93]. However, as elaborated by a penetrating analysis in [98], considerable arbitrariness exists in the independent dataset test. Also, as demonstrated by Eqs. 28–30 in [28], the subsampling test (or K -fold cross validation) cannot avoid arbitrariness either. Only the jackknife test is the least arbitrary that can always yield a unique result for a given benchmark dataset. Therefore, the jackknife test has been widely recognized and increasingly adopted by investigators to examine the quality of various predictors (see, e.g., [51,52,99]). In view of this, the success rate by the jackknife test was also used to optimize the two uncertain parameters K and φ in **Eq. 15**. The result thus obtained is shown in **Fig. 3**, from which we obtain when $K=10$ and $\varphi=1.8$ the **iGPCR-Drug** predictor reaches its optimized status.

The success rates thus obtained by the jackknife test in identifying interactive GPCR-drug pairs or non-interactive GPCR-drug pairs are given in **Table 2**, from which we can see that the overall success rate by **iGPCR-Drug** on the benchmark dataset \mathbb{S} was about 85.5%. In contrast, the corresponding success rate obtained by He et al. [32] in using six biochemical and physicochemical features to formulate GPCR-Drug samples was only 78.49%. The remarkable improved success rate indicates that introducing 2D molecular fingerprints to represent drug samples and using the grey PseAAC to represent GPCR samples are really a promising approach for studying the interactions of GPCRs and drugs in cellular network, where the former can catch the essence of the drug sample whereas the latter can catch the essence of the GPCR sample.

It is instructive to point out that, compared with the existing sequence-based methods, although the current approach could get better results because of introducing the 2D molecular fingerprints to represent drug samples and using grey PseAAC to represent the GPCR samples, it is still a sequence-based or “sequence-derived” approach, and hence could not avoid some limitation. Particularly,

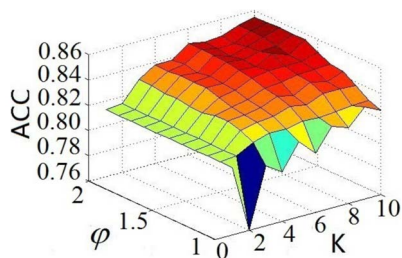


Figure 3. A 3D graph to show how to optimize the two parameters K and φ for the **iGPCR-Drug** predictor. doi:10.1371/journal.pone.0072234.g003

Table 2. The jackknife success rates obtained **iGPCR-Drug** in identifying interactive GPCR-drug pairs and non-interactive GPCR-drug pairs for the benchmark dataset \mathbb{S} (cf. Supporting Information S1).

Performance evaluation (cf. Eq. 10 or 22)	iGPCR-Drug ^a	Method by He et al. ^b
$\Lambda^+ \text{ Sn or}$	$\frac{496}{620} = 80.00\%$	N/A
$\Lambda^- \text{ Sp or}$	$\frac{1095}{1240} = 88.30\%$	N/A
$\Lambda \text{ Acc or}$	$\frac{1591}{1860} = 85.5\%$	78.49%
MCC	67.75%	N/A

^aThe parameters used: $w_1 = 10^5$ and $w_2 = 10^2$ (cf. **Eq. 10**), $w_3 = 1/700$ (cf. **Eq. 14**), and $K = 10$ and $\varphi = 1.8$ (cf. **Eq. 15**).

^bSee ref. [32].

doi:10.1371/journal.pone.0072234.t002

it cannot be used to predict the binding site and binding energy between GPCR and drug. Only when the 3D structures for both the GPCR receptor and its drug ligand are known or well defined, can we try to predict their binding details via molecular docking (see, e.g., [9]). Nevertheless, before their 3D structures are available, the current sequence-derived approach can serve as a high throughput tool for predicting GPCR–drug interactions in cellular networking. This is particularly useful in conducting large-scale analysis for the avalanche of biological sequences generated in the post-genomic age.

Besides, to further validate the current predictor, we took 314 GPCR-drug pairs from the study by Yamanishi et al. [100] that had been confirmed by experiments as interactive pairs and none of them occurred in the current benchmark dataset used to train our predictor. It was observed that, of the 314 pairs in such an independent dataset, 271 were correctly identified by **iGPCR-Drug** as interactive pairs; i.e., the success rate was 86.33%, quite consistent with the above-mentioned jackknife success rate (85.55%) achieved by the predictor on the benchmark dataset \mathbb{S} (cf. **Eq. 1**).

To enhance the value of its practical applications, the web server for **iGPCR-Drug** has been established that can be freely accessible at <http://www.jci-bioinfo.cn/iGPCR-Drug/>. It is anticipated that the web server will become a useful high throughput tool for both basic research and drug development in the relevant areas, or at the very least play a complementary role to the existing method [32] for which so far no web-server whatsoever has been provided yet.

3. The Protocol or User Guide

For the convenience of the vast majority of biologists and pharmaceutical scientists, here let us provide a step-by-step guide to show how the users can easily get the desired result by means of the web server without the need to follow the complicated mathematical equations presented in this paper for the process of developing the predictor and its integrity.

Step 1. Open the web server at the site <http://www.jci-bioinfo.cn/iGPCR-Drug/> and you will see the top page of the predictor on your computer screen, as shown in **Fig. 4**. Click on the Read Me button to see a brief introduction about **iGPCR-Drug** predictor and the caveat when using it.

Step 2. Either type or copy/paste the query pairs into the input box at the center of **Fig. 4**. Each query pair consists of two parts: one is for the protein sequence, and the other for the drug. The GPCR sequence should be in FASTA format, while the drug

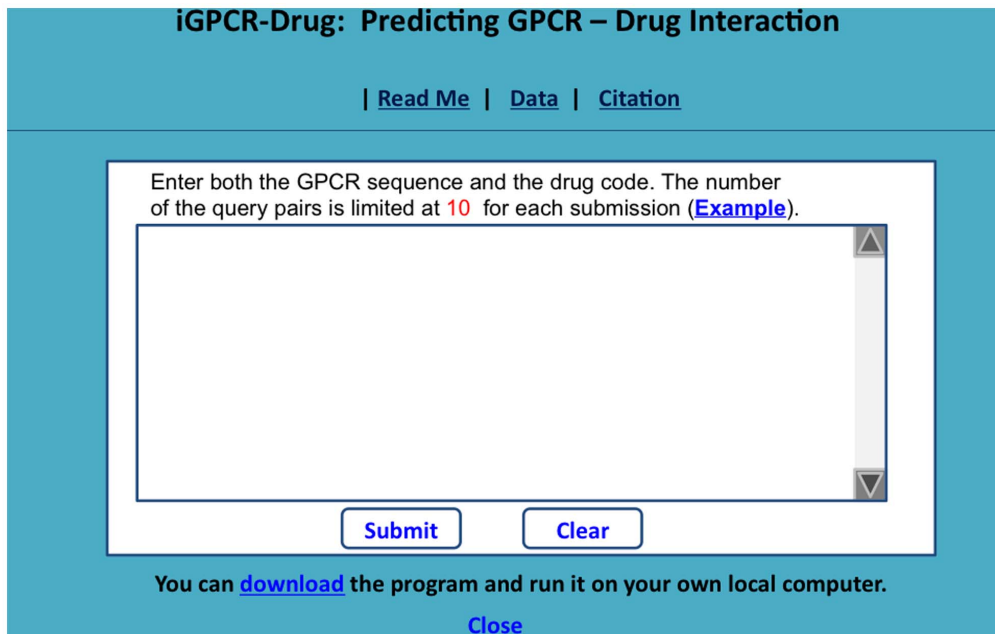


Figure 4. A semi-screenshot to show the top page of the iGPCR-Drug web-server. Its web-site address is at <http://www.jci-bioinfo.cn/iGPCR-Drug>. doi:10.1371/journal.pone.0072234.g004

in the KEGG code. Examples for the query pairs input can be seen 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 four query pairs in the [Example](#) window as the input, after clicking the [Submit](#) button, you will see on your screen that the “hsa:10161” GPCR and the “D00528” drug are an interactive pair, and that the “hsa:10800” GPCR and the “D00411” drug are also an interactive pair, but that the “hsa:1909” GPCR and the “D02566” drug are not an interactive pair, and that the “hsa:2913” GPCR and the “D01699” drug are not an interactive pair either. All these results are fully consistent with the experimental observations. It takes about 10 seconds before the results are shown on the screen.

Step 4. Click on the [Citation](#) button to find the relevant paper that documents the detailed development and algorithm of **iGPCR-Durg**.

Step 5. Click on the [Data](#) button to download the benchmark dataset used to train and test the **iGPCR-Durg** predictor.

Step 6. The program code is also available by clicking the button [download](#) on the lower panel of **Fig. 4**.

Supporting Information

Supporting Information S1 The benchmark dataset contains 1,860 GPCR-drug pair samples, of which 620 are interactive and

References

1. Heuss C, Gerber U (2000) G-protein-independent signaling by G-protein-coupled receptors. *Trends Neurosci* 23: 469–475.
2. Milligan G, White JH (2001) Protein-protein interactions at G-protein-coupled receptors. *Trends Pharmacol Sci* 22: 513–518.
3. Hall RA, Lefkowitz RJ (2002) Regulation of G protein-coupled receptor signaling by scaffold proteins. *Circ Res* 91: 672–680.
4. Chou KC (2005) Prediction of G-protein-coupled receptor classes. *Journal of Proteome Research* 4: 1413–1418.
5. Lundstrom K (2006) Latest development in drug discovery on G protein-coupled receptors. *Curr Protein Pept Sci* 7: 465–470.
6. Overington JP, Al-Lazikani B, Hopkins AL (2006) How many drug targets are there? *Nat Rev Drug Discov* 5: 993–996.

1,240 non-interactive. The codes listed here were from the KEGG database at <http://www.kegg.jp/kegg/>. (PDF)

Supporting Information S2 The fingerprints for the drug codes listed in Supporting Information S1. Each of these fingerprints is a 256-D vectors generated by the OpenBabel software downloaded from <http://openbabel.org/>. (PDF)

Supporting Information S3 The protein sequences for the GPCRs listed in Supporting Information S1. (PDF)

Acknowledgments

The authors wish to thank the two anonymous reviewers for their constructive comments, which were very helpful for strengthening the presentation of this paper.

Author Contributions

Conceived and designed the experiments: XX PW KCC. Performed the experiments: JLM PW. Analyzed the data: JLM PW KCC. Contributed reagents/materials/analysis tools: XX. Wrote the paper: XX KCC.

7. Knowles J, Gromo G (2003) A guide to drug discovery: Target selection in drug discovery. *Nat Rev Drug Discov* 2: 63–69.
8. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 261: 470–489.
9. Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: *ibid.*, 2003, Vol.310, 675). *Biochem Biophys Res Comm* 308: 148–151.
10. Zhu S, Okuno Y, Tsujimoto G, Mamitsuka H (2005) A probabilistic model for mining implicit ‘chemical compound-gene’ relations from literature. *Bioinformatics* 21 Suppl 2: ii245–251.

11. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 24: i232–240.
12. Chou KC (2004) Review: Structural bioinformatics and its impact to biomedical science. *Current Medicinal Chemistry* 11: 2105–2134.
13. Jorgensen WL (2004) The many roles of computation in drug discovery. *Science* 303: 1813–1818.
14. Hillisch A, Pineda LF, Hilgenfeld R (2004) Utility of homology models in the drug discovery process. *Drug Discov Today* 9: 659–669.
15. OuYang B, Xie S, Berardi MJ, Zhao XM, Dev J, et al. (2013) Unusual architecture of the p7 channel from hepatitis C virus. *Nature* 498: 521–525.
16. Schnell JR, Chou JJ (2008) Structure and mechanism of the M2 proton channel of influenza A virus. *Nature* 451: 591–595.
17. Berardi MJ, Shih WM, Harrison SC, Chou JJ (2011) Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching. *Nature* 476: 109–113.
18. Wang J, Pielak RM, McClintock MA, Chou JJ (2009) Solution structure and functional analysis of the influenza B proton channel. *Nature Structural and Molecular Biology* 16: 1267–1271.
19. Call ME, Wucherpfennig KW, Chou JJ (2010) The structural basis for intramembrane assembly of an activating immunoreceptor complex. *Nature Immunology* 11: 1023–1029.
20. Call ME, Schnell JR, Xu C, Lutz RA, Chou JJ, et al. (2006) The structure of the zeta/zeta transmembrane dimer reveals features essential for its assembly with the T cell receptor. *Cell* 127: 355–368.
21. Xu C, Gagnon E, Call ME, Schnell JR, Schwieters CD, et al. (2008) Regulation of T cell receptor activation by dynamic membrane binding of the CD3epsilon cytoplasmic tyrosine-based motif. *Cell* 135: 702–713.
22. Chou KC, Howe WJ (2002) Prediction of the tertiary structure of the beta-secretase zymogen. *Biochem Biophys Res Commun* 292: 702–708.
23. Chou KC (2004) Insights from modelling three-dimensional structures of the human potassium and sodium channels. *Journal of Proteome Research* 3: 856–861.
24. Chou KC (2005) Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. *Journal of Proteome Research* 4: 1681–1686.
25. Chou KC (2005) Insights from modeling the 3D structure of DNA-CBF3b complex. *Journal of Proteome Research* 4: 1657–1660.
26. Chou KC (2005) Modeling the tertiary structure of human cathepsin-E. *Biochem Biophys Res Commun* 331: 56–60.
27. Sirois S, Hatzakis G, Wei D, Du Q, Chou KC (2005) Assessment of chemical libraries for their druggability. *Comput Biol Chem* 29: 55–67.
28. Chou KC (2011) Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). *Journal of Theoretical Biology* 273: 236–247.
29. Chen W, Feng PM, Lin H, Chou KC (2013) iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Research* 41: e68; Available: <http://nar.oxfordjournals.org/content/41/6/e68.long>.
30. Xu Y, Ding J, Wu LY, Chou KC (2013) iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS ONE* 8: e55844.
31. Chen W, Lin H, Feng PM, Ding C, Zuo YC, et al. (2012) iNuc-PhysChem: A Sequence-Based Predictor for Identifying Nucleosomes via Physicochemical Properties. *PLoS ONE* 7: e47843.
32. He Z, Zhang J, Shi XH, Hu LL, Kong X, et al. (2010) Predicting drug-target interaction networks based on functional groups and biological features. *PLoS ONE* 5: e9603.
33. Finn P, Muggleton S, Page D, Srinivasan A (1998) Pharmacophore Discovery Using the Inductive Logic Programming System PROGOL. *Mach Learn* 30: 241–270.
34. Vogt I, Stumpfe D, Ahmed HE, Bajorath J (2007) Methods for computer-aided chemical biology. Part 2: Evaluation of compound selectivity using 2D molecular fingerprints. *Chem Biol Drug Des* 70: 195–205.
35. Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discov Today* 12: 225–233.
36. Laurent S, Elst LV, Muller RN (2006) Comparative study of the physicochemical properties of six clinical low molecular weight gadolinium contrast agents. *Contrast Media Mol Imaging* 1: 128–137.
37. Gregori-Puigjane E, Garriga-Sust R, Mestres J (2011) Indexing molecules with chemical graph identifiers. *J Comput Chem* 32: 2638–2646.
38. Ren B (2002) Application of novel atom-type AI topological indices to QSPR studies of alkanes. *Comput Chem* 26: 357–369.
39. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 34: D354–357.
40. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, et al. (2011) Open Babel: An open chemical toolbox. *J Cheminform* 3: 33.
41. Gillet VJ, Willett P, Bradshaw J (2003) Similarity searching using reduced graphs. *J Chem Inf Comput Sci* 43: 338–345.
42. Butina D (1999) Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets. *J Chem Inf Comput Sci* 39: 747–750.
43. Chou KC, Shen HB (2007) Review: Recent progresses in protein subcellular location prediction. *Analytical Biochemistry* 370: 1–16.
44. Altschul SF (1997) Evaluating the statistical significance of multiple distinct local alignments. In: Suhai S, editor. *Theoretical and Computational Methods in Genome Research*. New York: Plenum. 1–14.
45. Wootton JC, Federhen S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem* 17: 149–163.
46. Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. *J Biochem* 99: 152–162.
47. Chou KC, Zhang CT (1994) Predicting protein folding types by distance functions that make allowances for amino acid interactions. *Journal of Biological Chemistry* 269: 22014–22020.
48. Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. *PROTEINS: Structure, Function, and Genetics* (Erratum: *ibid*, 2001, Vol44, 60) 43: 246–255.
49. Lin SX, Lapointe J (2013) Theoretical and experimental biology in one –A symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers. *J Biomedical Science and Engineering (JBISE)* 6: 435–442; Available: <http://dx.doi.org/10.4236/jbise.2013.64054>.
50. Hayat M, Khan A (2012) Discriminating Outer Membrane Proteins with Fuzzy K-Nearest Neighbor Algorithms Based on the General Form of Chou's PseAAC. *Protein & Peptide Letters* 19: 411–421.
51. Khosraviyan M, Faramarzi FK, Beigi MM, Behbahani M, Mohabatkar H (2013) Predicting Antibacterial Peptides by the Concept of Chou's Pseudo-amino Acid Composition and Machine Learning Methods. *Protein & Peptide Letters* 20: 180–186.
52. Mohabatkar H, Beigi MM, Abdolahi K, Mohsenzadeh S (2013) Prediction of Allergenic Proteins by Means of the Concept of Chou's Pseudo Amino Acid Composition and a Machine Learning Approach. *Medicinal Chemistry* 9: 133–137.
53. Mohammad Beigi M, Behjati M, Mohabatkar H (2011) Prediction of metalloproteinase family based on the concept of Chou's pseudo amino acid composition using a machine learning approach. *Journal of Structural and Functional Genomics* 12: 191–197.
54. Sahu SS, Panda G (2010) A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction. *Computational Biology and Chemistry* 34: 320–327.
55. Nanni L, Lumini A, Gupta D, Garg A (2012) Identifying Bacterial Virulent Proteins by Fusing a Set of Classifiers Based on Variants of Chou's Pseudo Amino Acid Composition and on Evolutionary Information. *IEEE/ACM Trans Comput Biol Bioinform* 9: 467–475.
56. Zou D, He Z, He J, Xia Y (2011) Superssecondary structure prediction using Chou's pseudo amino acid composition. *Journal of Computational Chemistry* 32: 271–278.
57. Wan S, Mak MW, Kung SY (2013) GOASVM: A subcellular location predictor by incorporating term-frequency gene ontology into the general form of Chou's pseudo-amino acid composition. *Journal of Theoretical Biology* 323: 40–48.
58. Chang TH, Wu LC, Lee TY, Chen SP, Huang HD, et al. (2013) EuLoc: a web-server for accurately predict protein subcellular localization in eukaryotes by incorporating various features of sequence segments into the general form of Chou's PseAAC. *Journal of Computer-Aided Molecular Design* 27: 91–103.
59. Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2008) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. *Amino Acids* 34: 565–572.
60. Kandaswamy KK, Pugalenthi G, Moller S, Hartmann E, Kalies KU, et al. (2010) Prediction of Apoptosis Protein Locations with Genetic Algorithms and Support Vector Machines Through a New Mode of Pseudo Amino Acid Composition. *Protein and Peptide Letters* 17: 1473–1479.
61. Chen YK, Li KB (2013) Predicting membrane protein types by incorporating protein topology, domains, signal peptides, and physicochemical properties into the general form of Chou's pseudo amino acid composition. *Journal of Theoretical Biology* 318: 1–12.
62. Huang C, Yuan JQ (2013) A Multilabel Model Based on Chou's Pseudo-Amino Acid Composition for Identifying Membrane Proteins with Both Single and Multiple Functional Types. *J Membr Biol* 246: 327–334.
63. Zia Ur R, Khan A (2012) Identifying GPCRs and their Types with Chou's Pseudo Amino Acid Composition: An Approach from Multi-scale Energy Representation and Position Specific Scoring Matrix. *Protein & Peptide Letters* 19: 890–903.
64. Sun XY, Shi SP, Qiu JD, Suo SB, Huang SY, et al. (2012) Identifying protein quaternary structural attributes by incorporating physicochemical properties into the general form of Chou's PseAAC via discrete wavelet transform. *Molecular BioSystems* 8: 3178–3184.
65. Nanni L, Lumini A (2008) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino Acids* 34: 653–660.
66. Esmacili M, Mohabatkar H, Mohsenzadeh S (2010) Using the concept of Chou's pseudo amino acid composition for risk type prediction of human papillomaviruses. *Journal of Theoretical Biology* 263: 203–209.
67. Mohabatkar H (2010) Prediction of cyclin proteins using Chou's pseudo amino acid composition. *Protein & Peptide Letters* 17: 1207–1214.

68. Mohabatkar H, Mohammad Beigi M, Esmacili A (2011) Prediction of GABA(A) receptor proteins using the concept of Chou's pseudo-amino acid composition and support vector machine. *Journal of Theoretical Biology* 281: 18–23.
69. Georgiou DN, Karakasisis TE, Nieto JJ, Torres A (2009) Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *Journal of Theoretical Biology* 257: 17–26.
70. Li BQ, Huang T, Liu L, Cai YD, Chou KC (2012) Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network. *PLoS ONE* 7: e33393.
71. Jiang Y, Huang T, Lei C, Gao YF, Cai YD, et al. (2013) Signal propagation in protein interaction network during colorectal cancer progression. *BioMed Research International*, 2013: 287019. Available: <http://dx.doi.org/10.1155/2013/287019>.
72. Du P, Wang X, Xu C, Gao Y (2012) PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. *Analytical Biochemistry* 425: 117–119.
73. Cao DS, Xu QS, Liang YZ (2013) propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* 29: 960–962.
74. Shen HB, Chou KC (2008) PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. *Analytical Biochemistry* 373: 386–388.
75. Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. *Protein & Peptide Letters* 14: 871–875.
76. Deng JL (1989) Introduction to Grey System Theory. *The Journal of Grey System*: 1–24.
77. Lin WZ, Fang JA, Xiao X, Chou KC (2013) iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. *Molecular BioSystems* 9: 634–644.
78. Lin WZ, Fang JA, Xiao X, Chou KC (2011) iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. *PLoS ONE* 6: e24756.
79. Lin WZ, Fang JA, Xiao X, Chou KC (2012) Predicting Secretory Proteins of Malaria Parasite by Incorporating Sequence Evolution Information into Pseudo Amino Acid Composition via Grey System Model. *PLoS One* 7: e49040.
80. Lin WZ, Xiao X, Chou KC (2009) GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with grey incidence analysis. *Protein Eng Des Sel* 22: 699–705.
81. Xiao X, Lin WZ, Chou KC (2008) Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. *Journal of Computational Chemistry* 29: 2018–2024.
82. Keller JM, Gray MR, Givens JA (1985) A fuzzy k-nearest neighbours algorithm. *IEEE Trans Syst Man Cybern* 15: 580–585.
83. Xiao X, Wang P, Chou KC (2012) iNR-PhysChem: A Sequence-Based Predictor for Identifying Nuclear Receptors and Their Subfamilies via Physical-Chemical Property Matrix. *PLoS ONE* 7: e30869.
84. Roterman I, Konieczny L, Jurkowski W, Prymula K, Banach M (2011) Two-intermediate model to characterize the structure of fast-folding proteins. *Journal of Theoretical Biology* 283: 60–70.
85. Xiao X, Wang P, Chou KC (2011) GPCR-2L: Predicting G protein-coupled receptors and their types by hybridizing two different modes of pseudo amino acid compositions. *Molecular Biosystems* 7: 911–919.
86. Xiao X, Wang P, Chou KC (2011) Quat-2L: a web-server for predicting protein quaternary structural attributes. *Molecular Diversity* 15: 149–155.
87. Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. *Protein & Peptide Letters* 14: 811–815.
88. Maggiora GM, Zhang CT, Chou KC, Elrod DW (1996) Combining fuzzy clustering and neural networks to predict protein structural classes. In: Devillers J, editor. *Neural Networks in QSAR and Drug Design*. London: Academic Press. 255–279.
89. Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. *Biochem Biophys Res Commun* 334: 577–581.
90. Zhang CT, Chou KC, Maggiora GM (1995) Predicting protein structural classes from amino acid composition: application of fuzzy clustering. *Protein Engineering* 8: 425–435.
91. Shen HB, Yang J, Chou KC (2006) Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. *Journal of Theoretical Biology* 240: 9–13.
92. Wang P, Xiao X, Chou KC (2011) NR-2L: A Two-Level Predictor for Identifying Nuclear Receptor Subfamilies Based on Sequence-Derived Features. *PLoS ONE* 6: e23505.
93. Chou KC, Zhang CT (1995) Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* 30: 275–349.
94. Mahalanobis PC (1936) On the generalized distance in statistics. *Proc Natl Inst Sci India* 2: 49–55.
95. Chou KC (1995) Does the folding type of a protein depend on its amino acid composition? *FEBS Letters*. 363: 127–131.
96. Chou KC (2001) Prediction of signal peptides using scaled window. *Peptides* 22: 1973–1979.
97. Chou KC (1995) A novel approach to predicting protein structural classes in a (20–1)-D amino acid composition space. *Proteins: Structure, Function & Genetics* 21: 319–344.
98. Chou KC, Shen HB (2010) Cell-PLoc 2.0: An improved package of web-servers for predicting subcellular localization of proteins in various organisms. *Natural Science* 2: 1090–1103. Available: <http://dx.doi.org/10.4236/ns.2010.210136>.
99. Chou KC (2013) Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Molecular Biosystems* 9: 1092–1100.
100. Yamanishi Y, Kotera M, Kanehisa M, Goto S (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 26: i246–254.
101. Voet D, Voet JG (1995) *Biochemistry*. New York: John Wiley & sons, Inc. 2nd ed. 5–6. 5–6 p.
102. Tanford C (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *Journal of American Chemical Society* 84: 4240–4274.
103. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. *Proc Natl Acad Sci USA* 78: 3824–3828.