Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

iCDI-PseFpt: Identify the channel–drug interaction in cellular networking with PseAAC and molecular fingerprints

Xuan Xiao [a,b,d,*], Jian-Liang Min [a], Pu Wang [a], Kuo-Chen Chou [c,d]

[a] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China
[b] Information School, Zhe-Jiang Textile & Fashion College, Ning-Bo 315211, China
[c] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia
[d] Gordon Life Science Institute, 53 South Cottage Road, Belmont, MA 02478, United States

HIGHLIGHTS

- Interaction between ion-channels and drugs.
- Interactions within the cellular networking.
- Significance for drug development.

ARTICLE INFO

Article history:
Received 19 February 2013
Received in revised form
26 July 2013
Accepted 14 August 2013
Available online 26 August 2013

Keywords:
Ion channels
Gray model
Molecular fingerprints
Fuzzy *K*-nearest neighbor algorithm
Pseudo amino acid composition

ABSTRACT

Many crucial functions in life, such as heartbeat, sensory transduction and central nervous system response, are controlled by cell signalings via various ion channels. Therefore, ion channels have become an excellent drug target, and study of ion channel–drug interaction networks is an important topic for drug development. However, it is both time-consuming and costly to determine whether a drug and a protein ion channel are interacting with each other in a cellular network by means of experimental techniques. Although some computational methods were developed in this regard based on the knowledge of the 3D (three-dimensional) structure of protein, unfortunately their usage is quite limited because the 3D structures for most protein ion channels are still unknown. With the avalanche of protein sequences generated in the post-genomic age, it is highly desirable to develop the sequence-based computational method to address this problem. To take up the challenge, we developed a new predictor called **iCDI-PseFpt**, in which the protein ion-channel sample is formulated by the PseAAC (pseudo amino acid composition) generated with the gray model theory, the drug compound by the 2D molecular fingerprint, and the operation engine is the fuzzy *K*-nearest neighbor algorithm. The overall success rate achieved by **iCDI-PseFpt** via the jackknife cross-validation was 87.27%, which is remarkably higher than that by any of the existing predictors in this area. As a user-friendly web-server, **iCDI-PseFpt** is freely accessible to the public at the website http://www.jci-bioinfo.cn/iCDI-PseFpt/. Furthermore, 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 the paper just for its integrity. It has not escaped our notice that the current approach can also be used to study other drug–target interaction networks.

1. Introduction

Ion channels represent a class of membrane spanning protein pores that mediate the flux of ions in a variety of cell types (Green, 1999). The pore-forming ion channel subunits are proteins. For example, the M2 proton channel is formed by four helices (Schnell and Chou, 2008) (Fig. 1a), while the p7 channel from Hepatitis C virus formed by six helices (OuYang et al., 2013) (Fig. 1b). Ion channels mediate and regulate crucial functions via controlling cell signaling in organ and cellular physiology, including the heart beat, sensory transduction and central nervous system function, and there are over 300 types of ion channels in a living cell (Gabashvili et al., 2007). Proper function of ion channels is crucial for all living cells. Ion channel dysfunction may lead to a number of diseases, the so-called channelopathies, such as

[*] Corresponding author at: Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333403, China. Tel.: +86 138 7980 9729.
E-mail addresses: xxiao@gordonlifescience.org, jdzxiaoxuan@163.com (X. Xiao), minjianliang@aliyun.com (J.-L. Min), wp3751@163.com (P. Wang). kcchou@gordonlifescience.org (K.-C. Chou).

0022-5193/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtbi.2013.08.013

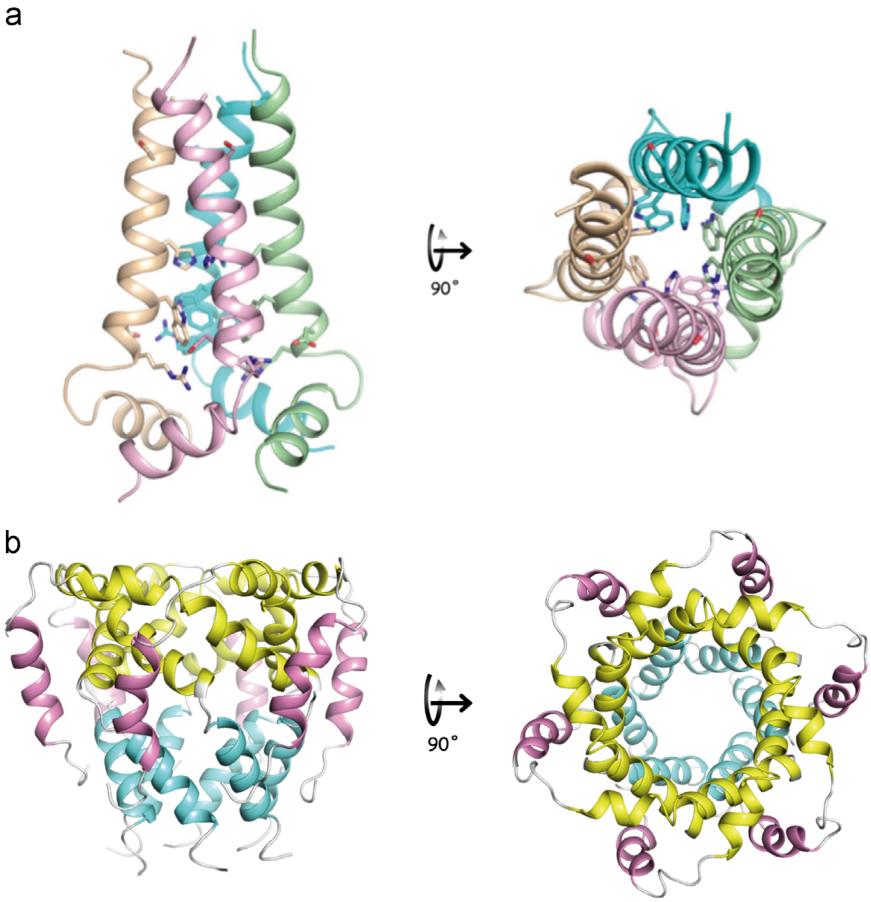


Fig. 1. Ribbon representation of the NMR structure as viewed from the side (left) and from the top (right) for (a) the M2 proton channel (Schnell and Chou, 2008), and (b) the p7 channel from Hepatitis C virus (OuYang et al., 2013). Reproduced from Schnell and Chou (2008) and OuYang et al. (2013) with permission.

epilepsy, arrhythmia, and type II diabetes. This kind of diseases is primarily treated with the drugs that modulate ion channels (Kaczorowski et al., 2008). Therefore, ion channels are excellent drug targets. Identification of drug–target interactions is an essential step during the drug discovery process, which is actually the most important task for the new medicine development (Knowles and Gromo, 2003). Many efforts in this regard have been made to discover new drugs in the past few years (Anderson, 2003; Burbidge et al., 2001; Lipinski et al., 2001). The most common methods are molecular docking simulations (Chou et al., 2003; Rarey et al., 1996), literature text mining (Zhu et al., 2005), and combining chemical structure, genomic sequence, and protein 3D (three-dimensional) structure information (Yamanishi et al., 2008). Actually, various structure-based drug design methods have become widely and increasingly used for drug development (Anderson, 2003; Chou, 2004a; Greer et al., 1994; Schames et al., 2004; Yuan et al., 2013).

However, the prerequisite for conducting the structure-based drug design is the knowledge of the 3D structure of the target protein. Unfortunately, the crystal or reliable 3D structures for many drug-related proteins are often not available. It is both time-consuming and expensive to determine their crystal structures. Particularly, most ion channels are membrane proteins, which are very difficult to crystallize. Although the recently developed state-of-the-art NMR technique is a very powerful tool in determining the 3D structures of membrane proteins and channels (see, e.g., (Berardi et al., 2011; Call et al., 2006; Douglas et al., 2007; OuYang et al., 2013; Oxenoid and Chou, 2005; Schnell and Chou, 2008; Wang et al., 2009)), it is also time-consuming and costly. With the avalanche of protein sequence data generated in the post-genomic age, to timely get the information of 3D structures of targeted proteins, one of the feasible approaches is to resort to the homology modeling (see, e.g., (Chou, 2004b; Chou, 2004c; Chou, 2004d; Chou, 2004e; Chou, 2005b; Chou et al., 1997; Chou et al., 2000; Hillisch et al., 2004; Jorgensen, 2004)). Unfortunately, such an approach fails to work for those targeted proteins that do not have sufficiently high sequence similarity with any of the 3D known protein structures in the PDB bank, an indispensable template or condition for developing a reliable structure via the structural bioinformatics approach (Chou, 2004a). In view of this, it would greatly speed up the pace of drug development and save a lot of time and money (Sirois et al., 2005) if a sequence-based computational method can be developed in this regard.

Actually, in a pioneering work, Yamanishi et al. (2008) proposed a computational method to identify the interaction between drugs and target proteins from the integration of chemical and genomic spaces. Two years later, He et al., (2010) also proposed a method to do the same based on functional groups and biological features. However, none of these methods has provided a web-server and hence their usage is quite limited.

The present study was initiated in an attempt to develop a new and more powerful predictor for identifying the ion channel–drug interactions based on the sequences of protein channels and the 2D molecular fingerprints of drugs. Particularly, a user-friendly web-server has been established that is freely accessible to the public to maximize its usage.

According to a recent review (Chou, 2011) and demonstrated by a series of recent publications (Chou et al., 2011; Chou et al., 2012; Lin et al., 2011; Wang et al., 2011; Wu et al., 2012; Xiao et al., 2011a; Xiao et al., 2011d; Xiao et al., 2011e), to establish a really useful statistical predictor for a biomedical 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 statistical samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the attribute to be identified; (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; and (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.

2. Materials and methods

2.1. Benchmark dataset

In this study, the benchmark dataset set $\mathbb{S}$ can be formulated as

$$\mathbb{S} = \mathbb{S}^{+} \cup \mathbb{S}^{-} \tag{1}$$

where the positive subset $\mathbb{S}^{+}$ consists of the interactive ion channel–drug pairs only, while the negative subset $\mathbb{S}^{-}$ contains the non-interactive ion channel–drug pairs only, and 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 $\mathbb{S}^{+}$ contains 1372 ion channel–drug pairs, which were taken from He et al. (2010). The negative dataset $\mathbb{S}^{-}$ contains 2744 non-interactive ion channel–drug pairs, which were derived according to the following the procedures as done in He et al. (2010): (i) separating each of the pairs in $\mathbb{S}^{+}$ into single drug and ion-channel; (ii) re-coupling each of the single drugs with each of the single ion-channels into pairs in a way that none of them occurred in $\mathbb{S}^{+}$; (iii) randomly picking the pairs thus formed until they reached the number two times as many as the pairs in $\mathbb{S}^{+}$. The 1327 interactive ion channel–drug pairs and 2744 non-interactive ion channel–drug pairs are given in Online Supporting Information S1.

2.2. Sample formulation

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

2.2.1. Representing ion channel protein sequences with pseudo amino acid composition

The sequences of the protein channels involved in this study are given in Online Supporting Information S2. Now the problem is how to effectively represent these protein 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 (Chou and Shen, 2007). The most typical sequential representation for a protein sample with L residues is its entire amino acid sequence, as can be formulated as

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\mathrm{R}_3\mathrm{R}_4\mathrm{R}_5\mathrm{R}_6\mathrm{R}_7\cdots\mathrm{R}_L \tag{2}$$

where R_1 represents the 1st residue of the protein $\mathbf{P}$, 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 (Altschul, 1997; Wootton and Federhen, 1993), 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

$$\mathbf{P} = \begin{bmatrix} f_1 & f_2 & \cdots & f_{20} \end{bmatrix}^{\mathbf{T}} \tag{3}$$

where $f_u (u = 1, 2, \cdots, 20)$ are the normalized occurrence frequencies of the 20 native amino acids (Chou, 1995b; Nakashima et al., 1986) in the protein $\mathbf{P}$, and $\mathbf{T}$ the transposing operator. The AAC-discrete model was widely used for identifying various attributes of proteins. However, as can be seen from Eq. (3), 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 (Chou, 2001d; Chou, 2005a), or Chou's PseAAC (Cao et al., 2013; Lin and Lapointe, 2013), was proposed to replace the simple amino acid composition (AAC) for representing the sample of a protein. Since the concept of PseAAC was proposed in 2001(Chou, 2001d), it has been widely used to study various attributes of proteins, such as identifying bacterial virulent proteins (Nanni et al., 2012), predicting supersecondary structure (Zou et al., 2011), predicting protein subcellular location (Kandaswamy et al., 2010; Mei, 2012; Zhang et al., 2008), predicting membrane protein types (Chen and Li, 2013), discriminating outer membrane proteins (Hayat and Khan, 2012), identifying antibacterial peptides (Khosravian et al., 2013), identifying allergenic proteins (Mohabatkar et al., 2013), predicting metalloproteinase family (Mohammad Beigi et al., 2011), predicting protein structural class (Sahu and Panda, 2010), identifying GPCRs and their types (Zia Ur and Khan, 2012), identifying protein quaternary structural attributes (Sun et al., 2012), predicting protein submitochondria locations (Nanni and Lumini, 2008), identifying risk type of human papillomaviruses (Esmaeili et al., 2010), identifying cyclin proteins (Mohabatkar, 2010), predicting GABA(A) receptor proteins (Mohabatkar et al., 2011), classifying amino acids (Georgiou et al., 2009), predicting cysteine S-nitrosylation sites in proteins (Xu et al., 2013), among many others (see a long list of papers cited in the References section of Chou, (2011)). Recently, the concept of PseAAC was further extended to represent the feature vectors of DNA and nucleotides (Chen et al., 2013; Chen et al., 2012), as well as other biological samples (see, e.g., (Huang et al., 2012; Li et al., 2012)). Because it has been widely and increasingly used, recently two powerful soft-wares, called ‘PseAAC-Builder’ (Du et al., 2012) and ‘propy’ (Cao et al., 2013), were established for generating various special Chou's pseudo-amino acid compositions, in addition to the web-server ‘PseAAC’ (Shen and Chou, 2008) built in 2008. According to a recent review (Chou, 2011), the general form of PseAAC for a protein $\mathbf{P}$ is formulated by

$$\mathbf{P} = [\Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_u \quad \cdots \quad \Psi_\Omega]^{\mathbf{T}} \tag{4}$$

where the subscript Ω is an integer, and its value as well as the components $\Psi_u (u = 1, 2, \cdots, \Omega)$ will depend on how to extract the desired information from the amino acid sequence of $\mathbf{P}$ (cf. Eq. (2)). Below, let us describe how to extract useful information from the benchmark dataset $\mathbb{S}$ to define the protein samples concerned via Eq. (4).

As is well known, a protein sequence is composed of 20 different types of native amino acids denoted by A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y (see the Online Supporting Information S2). 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 (Xiao and Chou, 2007). 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

Table 1
The numerical values of the ten physicochemical properties of the 20 native amino acids[a].

Amino acid	1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th
	Hydrophobicity	Hydrophilicity	Side-chain mass	pK1 (C^a-COOH)	pK2 (NH3)	PI (25°C)	Average buried volume	Molecular weight	Side-chain volume	Mean polarity
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

[a] The numerical values of the physicochemical properties can be obtained from the text biochemistry book (e.g., (Voet and Voet, 1995)) and the papers (Hopp and Woods, 1981; Tanford, 1962).

the observations (Chou, 2004e; Schnell and Chou, 2008) that polar amino acids play an important role in protein channels. Accordingly, the 20 numerical values of the mean polarity in Table 1 were used to encode the 20 native amino acids in a protein sequence. Note that to ensure that each of these numerical codes was a positive number as required by the Gray model used later, during the encoding process, each of the mean polarity values in Table 1 was added by 1.20. Thus, for a given protein sequence with L amino acids (cf. Eq. (2)), we can convert it into a series of L real numbers as described by

$$\mathbf{P} = (\, r_1 \quad r_2 \quad \cdots \quad r_L \,) \tag{5}$$

where r_1 is the mean polarity value for the 1st amino acid residue in the 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 gray system model to extract the useful information of $\mathbf{P}$ via Eq. (5) to formulate its PseAAC. According to the gray system theory (Deng, 1989), 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 “gray system”. The model developed based on such a theory is called “gray model”, which is a kind of nonlinear and dynamic model formulated by a differential equation. The gray model is particularly useful for solving complicated problems (Lin et al., 2013) that are lack of sufficient information, or need to process uncertain information and reduce random effects of acquired data. In the gray system theory, an important and generally used model is called GM(1,1) (Deng, 1989). By following the similar procedures as described in Lin et al. (2009), Lin et al. (2011), Lin et al. (2012) and Xiao et al. (2008), Eq. (4) would become a feature vector with dimension $\Omega = 22$ and each of its components defined by

$$\psi_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + \sum_{k=1}^{2} w_k a_k}, & 1 \le u \le 20 \\ \dfrac{w_{u-20} a_{u-20}}{\sum_i^{20} f_{i=1} + \sum_{k=1}^{2} w_k a_k}, & 21 \le u \le 22 \end{cases} \tag{6}$$

where f_u has the same meaning as Eq. (3), w_k $(k = 1, 2)$ is the weight factor (in this study we choose $w_1 = w_2 = 10^2$ to get the best results), and a_1 and a_2 are given by

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{Y} \tag{7}$$

where

$$\mathbf{B} = \begin{bmatrix} -\dfrac{r_2}{\ln\left(\frac{r_1 + r_2}{r_1}\right)} & 1 \\ -\dfrac{r_3}{\ln\left(\frac{r_1 + r_2 + r_3}{r_1 + r_2}\right)} & 1 \\ \vdots & \vdots \\ \vdots & \vdots \\ -\dfrac{r_L}{\ln\left(\frac{\sum_{k=1}^{L} r_k}{\sum_{k=1}^{L-1} r_k}\right)} & 1 \end{bmatrix} \tag{8}$$

and

$$\mathbf{Y} = \begin{bmatrix} r_2 \\ r_3 \\ \vdots \\ r_L \end{bmatrix} \tag{9}$$

2.2.2. Representing drug compounds with the 2D molecular fingerprint

Many molecular descriptors have been designed to capture the characteristics of structural information of a drug molecule (Filimonov et al., 1999; Xue and Bajorath, 2000). Among them, molecular fingerprints could provide the most detailed description of molecular structures, were often used in similarity search, virtual screening and QSAR/QSPR research (Casanola-Martin et al., 2010; Ewing et al., 2006; Nisius and Bajorath, 2009; Willett, 2011). Molecular fingerprints are a way of encoding the structure of a molecule, derived from the presence or absence of particular substructural fragments of drugs (Nisius and Bajorath, 2009; Owen et al., 2011; Tan et al., 2008). The most common type of fingerprint is a series of binary digits (bits) that represent the presence or absence of particular substructures in the drug molecule concerned. Below, let us describe how to use molecular fingerprints to represent drug compounds.

Although the number of drugs is extremely large, most of them are small organic molecules and are composed of some fixed small structures (Finn et al., 1998; Wallach and Lilien, 2009). The identification of small molecules or structures can be used to detect the target–drug interactions with positive influence (Vogt et al., 2007). For molecular fingerprints, which are bit-string representations of molecular structure and properties, the dimensionality of encoded descriptors has been studied intensely with the following conclusions: 2D fingerprints are powerful molecular descriptors, by which one can successfully recognize the active compounds even by their strings and atom numbers (Eckert and Bajorath, 2007). Many types of properties or characters have been proposed to represent drug molecules, including physicochemical properties (Laurent et al., 2006), chemical graphs (Gregori-Puigjane et al., 2011), topological indices (Ren, 2002), 3D pharmacophore patterns (Wang et al., 1994) and molecular fields (Tehan et al., 2001). In this study, let us use the simple but generally adopted 2D molecular fingerprints to represent drug molecules, as elaborated below.

From the KEGG database at http://www.kegg.jp/kegg/, we can get a MOL file. The latter can be further converted into a 2D molecular fingerprint file via a software called OpenBabel (O'Boyle et al., 2011) downloaded from the website http://openbabel.org/. There are four types of fingerprints generated by OpenBabel: FP2, FP3, FP4 and MACCS. Of the four, we choose to use the FP2 fingerprint format, which is a path-based fingerprint. It can identify small molecule fragments based on all linear and ring substructures for the molecules with the length from 1–7 atoms (excluding the 1-atom substructures C and N) by mapping them onto a bit-string using a hash function (somewhat similar to the Daylight fingerprints (Butina, 1999; Gillet et al., 2003)). The drug representation thus obtained is a 256-bit hexadecimal string, as can be formulated by a 256-D vector given below

$$\mathbf{D} = [\, d_1 \quad d_2 \quad \cdots \quad d_v \quad \cdots \quad d_{256} \,]^{\mathbf{T}} \tag{10}$$

where $\mathbf{D}$ is a drug compound in the network system concerned, $d_v (v = 1, 2, \cdots, 256)$ represents its v-th component, and $\mathbf{T}$ is the transpose operator. For readers' convenience, the numerical values of the 256-D vectors for the drug compounds investigated in the current network system are given in the Online Supporting Information S3.

2.2.3. The ion channel–drug pair representation

Now the pair between a protein ion channel $\mathbf{P}$ and a drug compound $\mathbf{D}$ can be formulated by combing Eqs. (4) and (10), as given by

$$\mathbf{G} = \mathbf{P} \oplus w_3\mathbf{D} = [\, \psi_1 \quad \cdots \quad \psi_{22} \quad w_3 d_1 \quad \cdots \quad w_3 d_{256} \,]^{\mathbf{T}} \tag{11}$$

where $\mathbf{G}$ represents the ion channel–drug pair, $\oplus$ the orthogonal sum, w_3 the weight factor that was chosen as 1/300 in this study to get the best results, and ψ_u $(u = 1, 2, \cdots, 22)$ are given in Eq. (6).

2.3. Fuzzy K-nearest neighbor algorithm

The fuzzy K-Nearest Neighbor (KNN) classification method (Keller and Hunt, 1985) 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 (Xiao et al., 2012), characterizing the structure of fast-folding proteins (Roterman et al., 2011), classifying G protein-coupled receptors (Xiao et al., 2011c), predicting protein quaternary structural attributes (Xiao et al., 2011b), predicting protein structural classes (Ding et al., 2007; Maggiora et al., 1996; Shen et al., 2005; Zhang et al., 1995), identifying membrane protein types (Shen et al., 2006), 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 KNN approach to identify the interactions between the protein ion channels and the drug compounds in the network concerned. For simplification, hereafter, let us use the word “channel–drug pair” or just “pair” to represent “ion channel–drug pair” unless otherwise specifically indicated.

Suppose $\mathbb{S}(N) = \{\mathbf{G}_1, \mathbf{G}_2, \cdots, \mathbf{G}_N\}$ is a set of vectors representing N channel–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; $\mathbb{S}^*(\mathbf{G}) = \{\mathbf{G}_1^*, \mathbf{G}_2^*, \cdots, \mathbf{G}_K^*\} \subset \mathbb{S}(N)$ 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 Wang et al., (2011)

$$\begin{cases} \mu^+(\mathbf{G}) = \dfrac{\sum_{j=1}^{K} \mu^+(\mathbf{G}_j^*) d(\mathbf{G}, \mathbf{G}_j^*)^{-2/(\varphi-1)}}{\sum_{j=1}^{K} d(\mathbf{G}, \mathbf{G}_j^*)^{-2/(\varphi-1)}} \\ \mu^-(\mathbf{G}) = \dfrac{\sum_{j=1}^{K} \mu^-(\mathbf{G}_j^*) d(\mathbf{G}, \mathbf{G}_j^*)^{-2/(\varphi-1)}}{\sum_{j=1}^{K} d(\mathbf{G}, \mathbf{G}_j^*)^{-2/(\varphi-1)}} \end{cases} \tag{12}$$

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 jth 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. (12), 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 (Chou and Zhang, 1995) and Mahalanobis distance (Chou, 1995a; Mahalanobis, 1936).

The quantitative definitions for the aforementioned $\mu^+(\mathbf{G}_j^*)$ and $\mu^-(\mathbf{G}_j^*)$ in Eq. (12) are given by

$$\mu^+(\mathbf{G}_j^*) = \begin{cases} 1, & \text{if } \mathbf{G}_j^* \in C^+ \\ 0, & \text{otherwise} \end{cases} \qquad \mu^-(\mathbf{G}_j^*) = \begin{cases} 1, & \text{if } \mathbf{G}_j^* \in C^- \\ 0, & \text{otherwise} \end{cases} \tag{13}$$

Substituting the results obtained by Eq. (13) into Eq. (12), 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

$$\mathbf{G} \in \begin{cases} C^+, & \text{if } \mu^+(\mathbf{G}) > \mu^-(\mathbf{G}) \\ C^-, & \text{otherwise} \end{cases} \tag{14}$$

The predictor thus established is called **iCDI-PseFpt**, where 'i' means 'identify', 'CDI' means 'channel-drug interaction', and “PseFpt” means using the PseAAC of protein sequence and fingerprint of drug. To provide an intuitive overall picture, a flowchart is provided in Fig. 2 to show the process of how the classifier works in identifying ion channel–drug interactions.

2.4. Criteria for performance evaluation

To provide a more intuitive and easier-to-understand method to measure the prediction quality, here the criteria proposed in Chou (2001a) and Chou (2001b) was adopted. According to Chou's criteria, the rates of correct predictions for the interactive ion channel–drug pairs in dataset $\mathbb{S}^+$ and the non-interactive ion

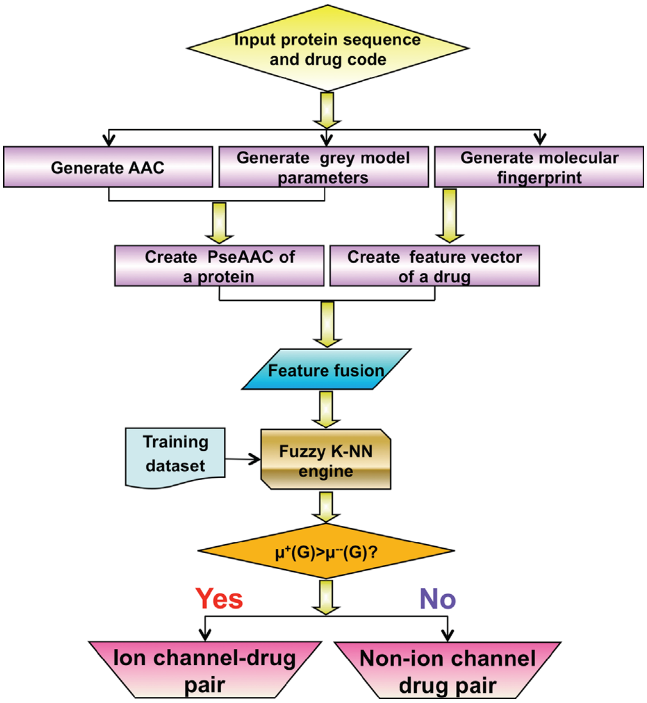

Input protein sequence and drug code
Generate AAC
Generate grey model parameters
Generate molecular fingerprint
Create PseAAC of a protein
Create feature vector of a drug
Feature fusion
Training dataset
Fuzzy K-NN engine
μ⁺(G)>μ⁻(G)?
Yes
No
Ion channel-drug pair
Non-ion channel drug pair

Fig. 2. A flowchart to show the operation process of the iCDI-PseFpt predictor. See the text for further explanation.

channel–drug pairs in dataset $\mathbb{S}^-$ are respectively defined by (cf. Eq. (1))

$$\begin{cases} \Lambda^+ = \frac{N^+ - N_-^+}{N^+}, & \text{for the interactive ion channel–drug pairs} \\ \Lambda^- = \frac{N^- - N_+^-}{N^-}, & \text{for the non–interactive ion channel–drug pairs} \end{cases} \tag{15}$$

where N^+ is the total number of the interactive ion channel–drug pairs investigated while N_-^+ the number of the interactive ion channel–drug pairs incorrectly predicted as the non-interactive ion channel–drug pairs; N^- the total number of the non-interactive ion channel–drug pairs investigated while N_+^- the number of the non-interactive ion channel–drug pairs incorrectly predicted as the interactive ion channel–drug pairs. The overall success prediction rate is given by Chou (2001c)

$$\Lambda = \frac{\Lambda^+ N^+ + \Lambda^- N^-}{N^+ + N^-} = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \tag{16}$$

It is obvious from Eqs. (15) and (16) that, if and only if none of the interactive ion channel–drug pairs and the non-interactive ion channel–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 instructive to point out that the following equation set is often used in literatures for examining the performance quality of a predictor

$$\begin{cases} \mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}} \\ \mathrm{Acc} = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \\ \mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}} \end{cases} \tag{17}$$

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.

The relations between the symbols in Eq. (16) and those in Eq. (17) are given by

$$\begin{cases} \mathrm{TP} = N^+ - N_-^+ \\ \mathrm{TN} = N^- - N_+^- \\ \mathrm{FP} = N_+^- \\ \mathrm{FN} = N_-^+ \end{cases} \tag{18}$$

Substituting Eq. (18) into Eq. (17) and also noting Eqs. (15) and (16), we obtain

$$\begin{cases} \mathrm{Sn} = \Lambda^+ = 1 - \frac{N_-^+}{N^+} \\ \mathrm{Sp} = \Lambda^- = 1 - \frac{N_+^-}{N^-} \\ \mathrm{Acc} = \Lambda = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} \\ \mathrm{MCC} = \frac{1 - (N_-^+/N^+ + N_+^-/N^-)}{\sqrt{[1 + (N_+^- - N_-^+)/N^+][1 + (N_-^+ - N_+^-)/N^-]}} \end{cases} \tag{19}$$

Obviously, when $N_-^+ = 0$ meaning none of the interactive ion channel–drug pairs was mispredicted to be a non-interactive ion channel–drug pair, we have the sensitivity Sn = 1; while $N_-^+ = N^+$ meaning that all the interactive ion channel–drug pairs were mispredicted to be the non-interactive ion channel–drug pairs, we have the sensitivity Sn = 0. Likewise, when $N_+^- = 0$ meaning none of the non-interactive ion channel–drug pairs was mispredicted, we have the specificity Sp = 1; while $N_+^- = N^-$ meaning all the non-interactive ion channel–drug pairs were incorrectly predicted as interactive ion channel–drug pairs, we have the specificity Sp = 0. When $N_-^+ = N_+^- = 0$ meaning that none of the interactive ion channel–drug pairs in the dataset $\mathbb{S}^+$ and none of the non-interactive ion channel–drug pairs in $\mathbb{S}^-$ was incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = \Lambda = 1$; while $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the interactive ion channel–drug pairs in the dataset $\mathbb{S}^+$ and all the non-interactive ion channel–drug pairs in $\mathbb{S}^-$ were mispredicted, we have the overall accuracy $\mathrm{Acc} = \Lambda = 0$. 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 ion channel–drug pairs in the dataset $\mathbb{S}^+$ and none of the non-interactive ion channel–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. (19) to examine a predictor for its sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient.

3. Results and discussion

3.1. Cross-validation

In statistical prediction, 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 (Chou and Zhang, 1995). Jackknife test is also called Leave-One-Out (LOO) test, during which each of the samples in a benchmark dataset was singled out one-by-one and tested by the predictor trained by the remaining samples. Among the above three cross-validation methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset as elaborated in Chou and Shen (2008) and demonstrated by Eqs. (28)–(30) in Chou (2011). Therefore, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (Chou et al., 2011;

Chou et al., 2012; Hayat and Khan, 2012; Khosravian et al., 2013; Mohabatkar et al., 2013; Nanni and Lumini, 2008; Sahu and Panda, 2010; Xiao and Chou, 2007). 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. (12). As shown in Fig. 3, it was a 2-D (dimensional) grid search. The search scope for the parameter K was within the range from 2 to 10 and its search interval was 1; while the scope for the parameter φ was within the range from 1.1 to 2 and its search interval was 0.1. Each of the ACC values obtained via the 2-D grid search was marked with a grid point. When the values of the two parameters were $K=4$ and $\varphi=1.9$, the **iCDI-PseFpt** predictor reached its optimal state.

The success rates by the jackknife test in identifying the channel–drug pairs are given in Table 2, from which we can see that the overall success rate achieved by **iCDI-PseFpt** on the benchmark dataset $\mathbb{S}$ was 87.27%, which is remarkably higher than 80.78%, the rate reported by He et al. (2010).

It is anticipated that the **iCDI-PseFpt** predictor will become a useful 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 (He et al., 2010) for which so far no web-server whatsoever has been provided yet.

3.2. Web server and user guide

To enhance the value of its practical applications, a web-server for **iCDI-PseFpt** was established at the website http://www.jci-bioinfo.cn/iCDI-PseFpt. Moreover, for the convenience of the vast majority of experimental 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 above mathematical equations for its development and integrity.

Step 1. Open the web-server at the site http://www.jci-bioinfo.cn/iCDI-PseFpt 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 **iCDI-PseFpt** predictor and the caveat when using it.

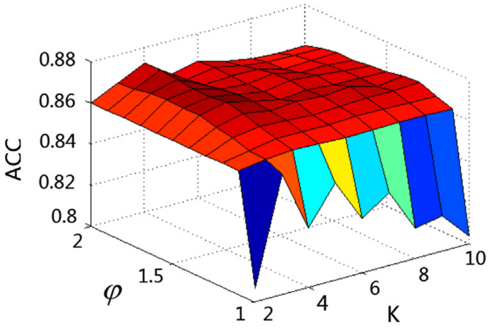


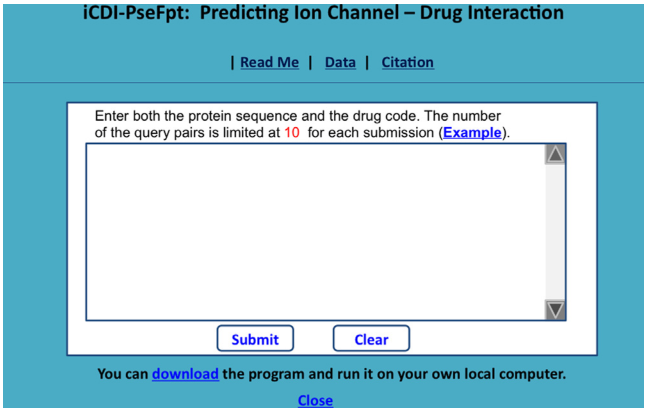
Fig. 3. A 3D graph to show how to optimize the two parameters K and φ for the iCDI-PseFpt predictor.

Table 2
The jackknife success rates obtained **iCDI-PseFpt** in identifying ion channel–drug pairs and non-ion channel drug pairs for the benchmark dataset $\mathbb{S}$ (cf. Online Supporting Information S1).

Performance evaluation (cf. Eq. (17) or (19))	iCDI-PseFpt[a] (%)	Method by He et al.[b] (%)
Sn or Λ^+	$1184/1372=86.30$	N/A
Sp or Λ^-	$2408/2744=87.76$	N/A
Acc or Λ	$3592/4116=87.27$	80.78
MCC	72.33%	N/A

[a] The parameters used: $w_1=w_2=10^2$ (cf. Eq. (6)) and $K=4$ and $\varphi=1.9$ (cf. Eq. (12)).

[b] See Ref. He et al. (2010).


iCDI-PseFpt: Predicting Ion Channel – Drug Interaction

| Read Me | Data | Citation

Enter both the protein sequence and the drug code. The number of the query pairs is limited at 10 for each submission (Example).

Submit Clear

You can download the program and run it on your own local computer.

Close

Fig. 4. A semi-screenshot to show the top page of the iCDI-SeqFpt web-server. Its web-site address is at http://www.jci-bioinfo.cn/iCDI-PseFpt.

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 protein sequence should be in FASTA format, while the drug 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 three query pairs in the Example window as the input, after clicking the Submit button, you will see on your screen that the "hsa:1134" channel and the "D05453" drug are an interactive pair, and that the "hsa:10369" channel and the "D01295" drug are also an interactive pair, but that the "hsa:6531" channel and the "D00474" drug are not. All these results are fully consistent with the experimental observations.

Step 4. Click on the Citation button to find the relevant paper that documents the detailed development and algorithm of **iCDI-SeqFpt**.

Step 5. Click on the Data button to download the benchmark dataset used to train and test the **iCDI-SeqFpt** predictor.

Step 6. The program is also available by clicking the button of "download" on the lower panel of Fig. 4.

Acknowledgments

The authors wish to thank the anonymous reviewer, whose constructive comments are very helpful for strengthening the presentation of this article. This work was supported by the Grants from the National Natural Science Foundation of China (Nos. 60961003 and 31260273), the Key Project of Chinese Ministry of Education (No. 210116), the Province National Natural Science Foundation of JiangXi (Nos. 2010GZS0122, 20114BAB211013 and 20122BAB201020), the Department of Education of JiangXi Province (GJJ12490), the LuoDi plan of the Department of Education of JiangXi Province (KJLD12083), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No. 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.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2013.08.013.

References

Altschul, S.F., 1997. Evaluating the statistical significance of multiple distinct local alignments. In: Suhai, S. (Ed.), Theoretical and Computational Methods in Genome Research. Plenum, New York, pp. 1–14.

Anderson, A.C., 2003. The process of structure-based drug design. Chemistry and Biology 10, 787–797.

Berardi, M.J., Shih, W.M., Harrison, S.C., Chou, J.J., 2011. Mitochondrial uncoupling protein 2 structure determined by NMR molecular fragment searching. Nature 476, 109–113.

Burbidge, R., Trotter, M., Buxton, B., Holden, S., 2001. Drug design by machine learning: support vector machines for pharmaceutical data analysis. Computers and Chemistry 26, 5–14.

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. Journal of Chemical Information and Computer Sciences 39, 747–750.

Call, M.E., Schnell, J.R., Xu, C., Lutz, R.A., Chou, J.J., Wucherpfennig, K.W., 2006. The structure of the zetazeta transmembrane dimer reveals features essential for its assembly with the T cell receptor. Cell 127, 355–368.

Cao, D.S., Xu, Q.S., Liang, Y.Z., 2013. Propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29, 960–962.

Casanola-Martin, G.M., Marrero-Ponce, Y., Khan, M.T., Khan, S.B., Torrens, F., Perez-Jimenez, F., Rescigno, A., Abad, C., 2010. Bond-based 2D quadratic fingerprints in QSAR studies: virtual and in vitro tyrosinase inhibitory activity elucidation. Chemical Biology and Drug Design 76, 538–545.

Chen, W., Feng, P.M., Lin, H., Chou, K.C., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Research 41, e68.

Chen, W., Lin, H., Feng, P.M., Ding, C., Zuo, Y.C., Chou, K.C., 2012. iNuc-PhysChem: a sequence-based predictor for identifying nucleosomes via physicochemical properties. PLoS One 7, e47843.

Chen, Y.K., Li, K.B., 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.

Chou, K.C., 1995a. Does the folding type of a protein depend on its amino acid composition? FEBS Letters 363, 127–131.

Chou, K.C., 1995b. A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space. Proteins: Structure, Function and Genetics 21, 319–344.

Chou, K.C., 2001a. Using subsite coupling to predict signal peptides. Protein Engineering 14, 75–79.

Chou, K.C., 2001b. Prediction of protein signal sequences and their cleavage sites. Proteins: Structure, Function and Genetics 42, 136–139.

Chou, K.C., 2001c. Prediction of signal peptides using scaled window. Peptides 22, 1973–1979.

Chou, K.C., 2001d. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins: Structure, Function and Genetics (43), 246–255. (Erratum: ibid., 2001, vol. 44, p. 60).

Chou, K.C., 2004a. Review: structural bioinformatics and its impact to biomedical science. Current Medicinal Chemistry 11, 2105–2134.

Chou, K.C., 2004b. Molecular therapeutic target for type-2 diabetes. Journal of Proteome Research 3, 1284–1288.

Chou, K.C., 2004c. Insights from modeling the 3D structure of the extracellular domain of alpha7 nicotinic acetylcholine receptor. Biochemical and Biophysical Research Communication 319, 433–438.

Chou, K.C., 2004d. Modeling extracellular domains of GABA-A receptors: subtypes 1, 2, 3, and 5. Biochemical and Biophysical Research Communications 316, 636–642.

Chou, K.C., 2004e. Insights from modeling three-dimensional structures of the human potassium and sodium channels. Journal of Proteome Research 3, 856–861.

Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.

Chou, K.C., 2005b. Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. Journal of Proteome Research 4, 1681–1686.

Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review). Journal of Theoretical Biology 273, 236–247.

Chou, K.C., Zhang, C.T., 1995. Prediction of protein structural classes. Critical Reviews in Biochemistry and Molecular Biology 30, 275–349.

Chou, K.C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Analytical Biochemistry 370, 1–16.

Chou, K.C., Shen, H.B., 2008. Cell-PLoc: a package of Web servers for predicting subcellular localization of proteins in various organisms (updated version: Cell-PLoc 2.0: an improved package of web-servers for predicting subcellular localization of proteins in various organisms. Natural Science 2010 (2), 1090–1103, http://dx.doi.org/10.4236/ns.2010.210136. (Nature Protocols 3, 153–162).

Chou, K.C., Jones, D., Heinrikson, R.L., 1997. Prediction of the tertiary structure and substrate binding site of caspase-8. FEBS Letters 419, 49–54.

Chou, K.C., Tomasselli, A.G., Heinrikson, R.L., 2000. Prediction of the tertiary structure of a caspase-9/inhibitor complex. FEBS Letters 470, 249–256.

Chou, K.C., Wei, D.Q., Zhong, W.Z., 2003. Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. Biochemical and Biophysical Research Communications 308, 148–151.

Chou, K.C., Wu, Z.C., Xiao, X., 2011. iLoc-Euk: a multi-label classifier for predicting the subcellular localization of singleplex and multiplex eukaryotic proteins. PLoS One 6, e18258.

Chou, K.C., Wu, Z.C., Xiao, X., 2012. iLoc-Hum: using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Molecular Biosystems 8, 629–641.

Deng, J.L., 1989. Introduction to Gray System Theory. The Journal of Gray System, 1–24.

Ding, Y.S., Zhang, T.L., Chou, K.C., 2007. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein and Peptide Letters 14, 811–815.

Douglas, S.M., Chou, J.J., Shih, W.M., 2007. DNA-nanotube-induced alignment of membrane proteins for NMR structure determination. Proceedings of the National Academy of Sciences of the United States of America 104, 6644–6648.

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.

Eckert, H., Bajorath, J., 2007. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discovery Today 12, 225–233.

Esmaeili, 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.

Ewing, T., Baber, J.C., Feher, M., 2006. Novel 2D fingerprints for ligand-based virtual screening. Journal of Chemical Information and Modeling 46, 2423–2431.

Filimonov, D., Poroikov, V., Borodina, Y., Gloriozova, T., 1999. Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descryptors. Journal of Chemical Information and Modeling 39, 666–670.

Finn, P., Muggleton, S., Page, D., Srinivasan, A., 1998. Pharmacophore discovery using the inductive logic programming system PROGOL. Machine Learning 30, 241–270.

Gabashvili, I.S., Sokolowski, B.H., Morton, C.C., Giersch, A.B., 2007. Ion channel gene expression in the inner ear. Journal of the Association for Research in Otolaryngology 8, 305–328.

Georgiou, D.N., Karakasidis, T.E., Nieto, J.J., 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.

Gillet, V.J., Willett, P., Bradshaw, J., 2003. Similarity searching using reduced graphs. Journal of Chemical Information and Computer Sciences 43, 338–345.

Green, W.N., 1999. Ion channel assembly: creating structures that function. Journal of General Physiology 113, 163–170.

Greer, J., Erickson, J.W., Baldwin, J.J., Varney, M.D., 1994. Application of the three-dimensional structures of protein target molecules in structure-based drug design. Journal of Medicinal Chemistry 37, 1035–1054.

Gregori-Puigjane, E., Garriga-Sust, R., Mestres, J., 2011. Indexing molecules with chemical graph identifiers. Journal of Computational Chemistry 32, 2638–2646.

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 and Peptide Letters 19, 411–421.

He, Z., Zhang, J., Shi, X.H., Hu, L.L., Kong, X., Cai, Y.D., Chou, K.C., 2010. Predicting drug-target interaction networks based on functional groups and biological features. PLoS One 5, e9603.

Hillisch, A., Pineda, L.F., Hilgenfeld, R., 2004. Utility of homology models in the drug discovery process. Drug Discovery Today 9, 659–669.

Hopp, T.P., Woods, K.R., 1981. Prediction of protein antigenic determinants from amino acid sequences. Proceedings of the National Academy of Sciences of the United States of America 78, 3824–3828.

Huang, T., Wang, J., Cai, Y.D., Yu, H., Chou, K.C., 2012. Hepatitis C virus network based classification of hepatocellular cirrhosis and carcinoma. PLoS One 7, e34460.

Jorgensen, W.L., 2004. The many roles of computation in drug discovery. Science 303, 1813–1818.

Kaczorowski, G., McManus, O., Priest, B., Garcia, M., 2008. Ion channels as drug targets: the next GPCRs. Journal of Cell Biology 131, 399–405.

Kandaswamy, K.K., Pugalenthi, G., Moller, S., Hartmann, E., Kalies, K.U., Suganthan, P.N., Martinetz, T., 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.

Keller, J.M., Hunt, D.J., 1985. Incorporating fuzzy membership functions into the perceptron algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence 7, 693–699.

Khosravian, M., Faramarzi, F.K., Beigi, M.M., Behbahani, M., Mohabatkar, H., 2013. Predicting antibacterial peptides by the concept of Chou's pseudo-amino acid composition and machine learning methods. Protein and Peptide Letters 20, 180–186.

Knowles, J., Gromo, G., 2003. A guide to drug discovery: target selection in drug discovery. Nature Reviews Drug Discovery 2, 63–69.

Laurent, S., Elst, L.V., Muller, R.N., 2006. Comparative study of the physicochemical properties of six clinical low molecular weight gadolinium contrast agents. Contrast Media and Molecular Imaging 1, 128–137.

Li, B.Q., Huang, T., Liu, L., Cai, Y.D., Chou, K.C., 2012. Identification of colorectal cancer related genes with mRMR and shortest path in protein–protein interaction network. PLoS One 7, e33393.

Lin, S.X., Lapointe, J., 2013. Theoretical and experimental biology in one. Journal of Biomedical Science and Engineering (JBiSE) 6, 435–442, http://dx.doi.org/10.4236/jbise.2013.64054.

Lin, W.Z., Xiao, X., Chou, K.C., 2009. GPCR-GIA: a web-server for identifying G-protein coupled receptors and their families with gray incidence analysis. Protein Engineering Design and Selection 22, 699–705.

Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2011. iDNA-Prot: identification of DNA binding proteins using random forest with gray model. PLoS One 6, e24756.

Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2012. Predicting secretory proteins of malaria parasite by incorporating sequence evolution information into pseudo amino acid composition via gray system model. PLoS One 7, e49040.

Lin, W.Z., Fang, J.A., Xiao, X., Chou, K.C., 2013. iLoc-Animal: a multi-label learning classifier for predicting subcellular localization of animal proteins. Molecular BioSystems 9, 634–644.

Lipinski, C.A., Lombardo, F., Dominy, B.W., Feeney, P.J., 2001. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced Drug Delivery Reviews 46 (1-3), 3–26.

Maggiora, G.M., Zhang, C.T., Chou, K.C., Elrod, D.W., 1996. Combining fuzzy clustering and neural networks to predict protein structural classes. In: Devillers, J. (Ed.), Neural Networks in QSAR and Drug Design. Academic Press, London, pp. 255–279.

Mahalanobis, P.C., 1936. On the generalized distance in statistics. Proceedings of the National Institute of Sciences of India 2, 49–55.

Mei, S., 2012. Predicting plant protein subcellular multi-localization by Chou's PseAAC formulation based multi-label homolog knowledge transfer learning. Journal of Theoretical Biology 310, 80–87.

Mohabatkar, H., 2010. Prediction of cyclin proteins using Chou's pseudo amino acid composition. Protein and Peptide Letters 17, 1207–1214.

Mohabatkar, H., Mohammad Beigi, M., Esmaeili, 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.

Mohabatkar, H., Beigi, M.M., 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.

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.

Nakashima, H., Nishikawa, K., Ooi, T., 1986. The folding type of a protein is relevant to the amino acid composition. Journal of Biochemistry 99, 152–162.

Nanni, L., Lumini, A., 2008. Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids 34, 653–660.

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 Transactions on Computational Biology and Bioinformatics 9, 467–475.

Nisius, B., Bajorath, J., 2009. Molecular fingerprint recombination: generating hybrid fingerprints for similarity searching from different fingerprint types. ChemMedChem 4, 1859–1863.

O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., Hutchison, G.R., 2011. Open Babel: an open chemical toolbox. Journal of Cheminformatics 3, 33.

OuYang, B., Xie, S., Berardi, M.J., Zhao, X.M., Dev, J., Yu, W., Sun, B., Chou, J.J., 2013. Unusual architecture of the p7 channel fromhepatitis C virus. Nature 498, 521–525.

Owen, J.R., Nabney, I.T., Medina-Franco, J.L., Lopez-Vallejo, F., 2011. Visualization of molecular fingerprints. Journal of Chemical Information and Modeling 51, 1552–1563.

Oxenoid, K., Chou, J.J., 2005. The structure of phospholamban pentamer reveals a channel-like architecture in membranes. Proceedings of the National Academy of Sciences of the United States of America 102, 10870–10875.

Rarey, M., Kramer, B., Lengauer, T., Klebe, G., 1996. A fast flexible docking method using an incremental construction algorithm. Journal of Molecular Biology 261, 470–489.

Ren, B., 2002. Application of novel atom-type AI topological indices to QSPR studies of alkanes. Computers and Chemistry 26, 357–369.

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.

Sahu, S.S., 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.

Schames, J.R., Henchman, R.H., Siegel, J.S., Sotriffer, C.A., Ni, H., McCammon, J.A., 2004. Discovery of a novel binding trench in HIV integrase. Journal of Medicinal Chemistry 47, 1879–1881.

Schnell, J.R., Chou, J.J., 2008. Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451, 591–595.

Shen, H.B., Chou, K.C., 2008. PseAAC: a flexible web-server for generating various kinds of protein pseudo amino acid composition. Analytical Biochemistry 373, 386–388.

Shen, H.B., Yang, J., Chou, K.C., 2006. Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. Journal of Theoretical Biology 240, 9–13.

Shen, H.B., Yang, J., Liu, X.J., Chou, K.C., 2005. Using supervised fuzzy clustering to predict protein structural classes. Biochemical and Biophysical Research Communications 334, 577–581.

Sirois, S., Hatzakis, G.E., Wei, D.Q., Du, Q.S., Chou, K.C., 2005. Assessment of chemical libraries for their druggability. Computational Biology and Chemistry 29, 55–67.

Sun, X.Y., Shi, S.P., Qiu, J.D., Suo, S.B., Huang, S.Y., Liang, R.P., 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.

Tan, L., Lounkine, E., Bajorath, J., 2008. Similarity searching using fingerprints of molecular fragments involved in protein–ligand interactions. Journal of Chemical Information and Modeling 48, 2308–2312.

Tanford, C., 1962. Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. Journal of American Chemical Society 84, 4240–4274.

Tehan, B.G., Lloyd, E.J., Wong, M.G., 2001. Molecular field analysis of clozapine analogs in the development of a pharmacophore model of antipsychotic drug action. Journal of Molecular Graphics and Modeling 19 (417–426, 468).

Voet, D., Voet, J.G., 1995. Biochemistry, 2nd ed. John Eiley & sons, Inc, New York, pp. 5–6.

Vogt, I., Stumpfe, D., Ahmed, H.E., Bajorath, J., 2007. Methods for computer-aided chemical biology. Part 2: evaluation of compound selectivity using 2D molecular fingerprints. Chemical Biology and Drug Design 70, 195–205.

Wallach, I., Lilien, R., 2009. The protein-small-molecule database, a non-redundant structural resource for the analysis of protein–ligand binding. Bioinformatics 25, 615–620.

Wang, J., Pielak, R.M., McClintock, M.A., Chou, J.J., 2009. Solution structure and functional analysis of the influenza B proton channel. Nature Structural and Molecular Biology 16, 1267–1271.

Wang, P., Xiao, X., Chou, K.C., 2011. NR-2L: a Two-level predictor for identifying nuclear receptor subfamilies based on sequence-derived features. PLoS One 6, e23505.

Wang, S., Zaharevitz, D.W., Sharma, R., Marquez, V.E., Lewin, N.E., Du, L., Blumberg, P. M., Milne, G.W., 1994. The discovery of novel, structurally diverse protein kinase C agonists through computer 3D-database pharmacophore search. Molecular modeling studies. Journal of Medicinal Chemistry 37, 4479–4489.

Willett, P., 2011. Similarity searching using 2D structural fingerprints. Methods in Molecular Biology 672, 133–158.

Wootton, J.C., Federhen, S., 1993. Statistics of local complexity in amino acid sequences and sequence databases. Computers and Chemistry 17, 149–163.

Wu, Z.C., Xiao, X., Chou, K.C., 2012. iLoc-Gpos: a multi-layer classifier for predicting the subcellular localization of singleplex and multiplex gram-positive bacterial proteins. Protein and Peptide Letters 19, 4–14.

Xiao, X., Chou, K.C., 2007. Digital coding of amino acids based on hydrophobic index. Protein and Peptide Letters 14, 871–875.

Xiao, X., Lin, W.Z., Chou, K.C., 2008. Using gray dynamic modeling and pseudo amino acid composition to predict protein structural classes. Journal of Computational Chemistry 29, 2018–2024.

Xiao, X., Wu, Z.C., Chou, K.C., 2011a. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. Journal of Theoretical Biology 284, 42–51.

Xiao, X., Wang, P., Chou, K.C., 2011b. Quat-2L: a web-server for predicting protein quaternary structural attributes. Molecular Diversity 15, 149–155.

Xiao, X., Wang, P., Chou, K.C., 2011c. 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.

Xiao, X., Wang, P., Chou, K.C., 2011d. 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.

Xiao, X., Wu, Z.C., Chou, K.C., 2011e. A multi-label classifier for predicting the subcellular localization of gram-negative bacterial proteins with both single and multiple sites. PLoS One 6, e20592.

Xiao, X., Wang, P., Chou, K.C., 2012. iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. PLoS One 7, e30869.

Xu, Y., Ding, J., Wu, L.Y., Chou, K.C., 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.

Xue, L., Bajorath, J., 2000. Molecular descriptors in chemoinformatics, computational combinatorial chemistry, and virtual screening. Combinatorial Chemistry and High Throughput Screening 3, 363–372.

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–i240.

Yuan, Y., Pei, J., Lai, L., 2013. Binding site detection and druggability prediction of protein targets for structure-based drug design. Current Pharmaceutical Design 19, 2326–2333.

Zhang, C.T., Chou, K.C., Maggiora, G.M., 1995. Predicting protein structural classes from amino acid composition: application of fuzzy clustering. Protein Engineering 8, 425–435.

Zhang, S.W., Zhang, Y.L., Yang, H.F., Zhao, C.H., 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.

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–ii251.

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 and Peptide Letters 19, 890–903.

Zou, D., He, Z., He, J., Xia, Y., 2011. Supersecondary structure prediction using Chou's pseudo amino acid composition. Journal of Computational Chemistry 32, 271–278.