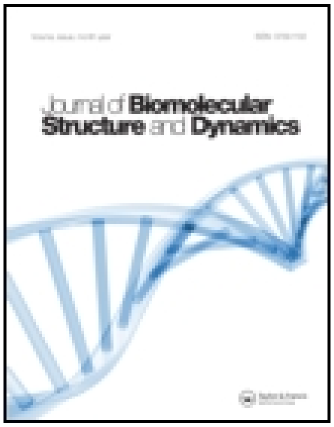


Click for updates

Journal of Biomolecular Structure and Dynamics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tbsd20

iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach

Xuan Xiao[abd], Jian-Liang Min[a], Wei-Zhong Lin[a], Zi Liu[a], Xiang Cheng[a] & Kuo-Chen Chou[cd]

[a] Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333046, China

[b] Information School, ZheJiang Textile & Fashion College, NingBo 315211, China

[c] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, JeddaH 21589, Saudi Arabia

[d] Gordon Life Science Institute, 53 South Cottage Road, Boston 02478, MA, USA

Accepted author version posted online: 16 Dec 2014.Published online: 14 Jan 2015.

To cite this article: Xuan Xiao, Jian-Liang Min, Wei-Zhong Lin, Zi Liu, Xiang Cheng & Kuo-Chen Chou (2015) iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach, Journal of Biomolecular Structure and Dynamics, 33:10, 2221-2233, DOI: 10.1080/07391102.2014.998710

To link to this article: http://dx.doi.org/10.1080/07391102.2014.998710

Journal of Biomolecular Structure and Dynamics, 2015
Vol. 33, No. 10, 2221–2233, http://dx.doi.org/10.1080/07391102.2014.998710

iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via benchmark dataset optimization approach

Xuan Xiao[a,b,d]*, Jian-Liang Min[a], Wei-Zhong Lin[a], Zi Liu[a], Xiang Cheng[a] and Kuo-Chen Chou[c,d]

[a]*Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 333046, China;* [b]*Information School, ZheJiang Textile & Fashion College, NingBo 315211, China;* [c]*Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, JeddaH 21589, Saudi Arabia;* [d]*Gordon Life Science Institute, 53 South Cottage Road, Boston 02478, MA, USA*

Communicated by Ramaswamy H. Sarma

(*Received 9 October 2014; accepted 11 December 2014*)

Information about the interactions of drug compounds with proteins in cellular networking is very important for drug development. Unfortunately, all the existing predictors for identifying drug–protein interactions were trained by a skewed benchmark data-set where the number of non-interactive drug–protein pairs is overwhelmingly larger than that of the interactive ones. Using this kind of highly unbalanced benchmark data-set to train predictors would lead to the outcome that many interactive drug–protein pairs might be mispredicted as non-interactive. Since the minority interactive pairs often contain the most important information for drug design, it is necessary to minimize this kind of misprediction. In this study, we adopted the neighborhood cleaning rule and synthetic minority over-sampling technique to treat the skewed benchmark datasets and balance the positive and negative subsets. The new benchmark datasets thus obtained are called the optimized benchmark datasets, based on which a new predictor called iDrug-Target was developed that contains four sub-predictors: iDrug-GPCR, iDrug-Chl, iDrug-Ezy, and iDrug-NR, specialized for identifying the interactions of drug compounds with GPCRs (G-protein-coupled receptors), ion channels, enzymes, and NR (nuclear receptors), respectively. Rigorous cross-validations on a set of experiment-confirmed datasets have indicated that these new predictors remarkably outperformed the existing ones for the same purpose. To maximize users' convenience, a public accessible Web server for iDrug-Target has been established at http://www.jci-bioinfo.cn/iDrug-Target/, by which users can easily get their desired results. It has not escaped our notice that the aforementioned strategy can be widely used in many other areas as well.

Keywords: iDrug-GPCR; iDrug-Chl; iDrug-Ezy; iDrug-NR; molecular fingerprints; chou's PseAAC; NCR; SMOTE; optimized training data-set; target-jackknife validation

1. Introduction

One of the key steps in rational drug design is to identify the interactions between drugs and targets (Lindsay, 2003; Schenone, Dančík, Wagner, & Clemons, 2013; Sirois, Wei, & Du, 2004). Although molecular docking simulation (Morris et al., 1998) (Ewing, Makino, Skillman, & Kuntz, 2001) is a very useful vehicle for this purpose and has been widely used (see, e.g. (Chou, Wei, Du, Sirois, & Zhong, 2006; Chou, Wei, & Zhong, 2003; Du, Huang, & Wang, 2010, 2009; Li, Wang, Xu, & Wang, 2011; Liao, Gao, & Wei, 2011; Ma, Wang, & Xu, 2012; Wang & Chou, 2011; Wang, Du, Huang, & Zhang, 2009b; Wei, Wang, Du, & Meng, 2009)), to make the molecular docking simulation study feasible, a prerequisite condition is the availability of a reliable 3D (three-dimensional) structure of the target protein. Although X-ray crystallography is a powerful technique for determining the 3D structure of a protein, it is time-consuming and expensive. Besides, to make the technique workable, the protein concerned must be crystallized first. Unfortunately, many proteins, particularly membrane proteins, are hard to crystallize. Although the high-resolution NMR (nuclear magnetic resonance) technique is indeed a very powerful tool in determining the 3D structures of membrane proteins as indicated by a series of recent publications (see, e.g. (Berardi, Shih, Harrison, & Chou, 2011; Call, Wucherpfenning, & Chou, 2010; OuYang et al., 2013; Oxenoid & Chou, 2005; Pielak & Chou, 2010; Pielak, Schnell, & Chou, 2009; Schnell & Chou, 2008; Wang, Piela, & McClintock, 2009a) and a recent review (OuYang & Chou, 2014)), it is also time-consuming and costly. To acquire the 3D structural information in a timely manner, one has to resort to various structural bioinformatics tools (see, e.g. (Chou, 2004b)), particularly the homologous modeling technique as used for many target proteins whose 3D structures were desperately needed during the process of drug

*Corresponding author. Emails: xxiao@gordonlifescience.org, jdzxiaoxuan@163.com development (Chou, 2004a, 2005a, 2005b, 2005c; Chou & Howe, 2002; Chou, Jones, & Heinrikson, 1997; Chou, Tomasselli, & Heinrikson, 2000; Wang & Chou, 2011, 2012; Wang, Du, & Chou, 2007b; Wei, Du, & Sun, 2006; Wang, Wei, Li, & Zheng, 2007a). Unfortunately, the number of dependable templates for developing high quality 3D protein structures by means of homology modeling is very limited (Chou, 2004b).

To overcome the aforementioned barriers, it would be very useful to develop computational methods aimed at identifying the interactions of drug compounds with various protein targets, such as GPCRs (G-protein-coupled receptors), protein channels, enzymes, and NRs (nuclear receptors), in cellular networking based on the sequence information of the latter. The results thus obtained can be used to pre-exclude the compounds identified not interacting with the protein targets, so as to timely stop wasting time and money on those unpromising compounds (Sirois, Hatzakis, & Wei, 2005).

Actually, considerable efforts have been made in this regard. For instance, (He, Zhang, Shi, Hu, & Kong, 2010) developed a powerful computation method for predicting drug–target interaction networks based on the functional groups and biological features. However, no Web server whatsoever was provided for their method, and hence, its practical application value is quite limited for most drug–development scientists. To make up this shortcoming, four Web server predictors, called "iGPCR-Drug" (Xiao, Min, & Wang, 2013b), "iCDI-PseFpt" (Xiao, Min, & Wang, 2013a), "iEzy-Drug" (Min, Xiao, & Chou, 2013), and "iNR-Drug" (Fan, Xiao, & Min, 2014), were recently developed for identifying the interactions of drug compounds with GPCRs, ion channels, enzymes, and NRs in cellular networking, respectively (Figure 1). Their Web sites addresses are listed in Table 1. Although each of the four Web server predictors could yield higher success rate than the original prediction method (He et al., 2010) for the same purpose, the benchmark data-set used to train and test each of the four predictors was taken from (He et al., 2010), and hence has the following problem. For the benchmark data-set in (He et al., 2010), the number of the non-interactive pair samples is much larger than that of the interactive pair samples. Although this might reflect the real world in which the interactive pairs are always the minority compared with the non-interactive ones, using this kind of highly unbalanced benchmark data-set to train a predictor would lead to the outcome that many interactive drug–target pairs might be mispredicted as non-interactive ones (Sun, Wong, & Kamel, 2009). Since the minority interactive drug–target pairs are our focus in drug development, we should take some action to optimize the benchmark data-set so as to minimize this kind of misprediction, and meanwhile avoiding the overprediction as well. This study was initiated in an attempt to address this kind of problem.

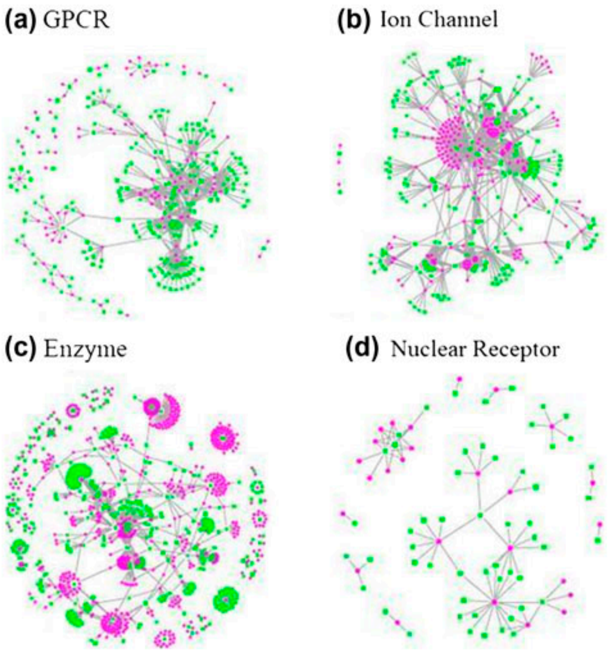


Figure 1. Graphical representation to show the drug–target interactions in cellular networking.
Note: Panel (a) is for drug-GPCR interaction; (b) for drug–channel; (c) for drug–enzyme; and (d) for drug-NR, where the drug is represented by a green square, the target protein by a magenta circle, and the interaction between the two is by a gray edge. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this paper.)

As demonstrated in a series of recent publications (see, e.g. (Chen, Feng, Deng, & Lin, 2014b; Guo, Deng, Xu, Ding, & Lin, 2014; Lin, Deng, Ding, & Chen, 2014; Liu, Xu, Lan, Xu, & Zhou, 2014a; Liu, Xiao, & Qiu, 2015; Liu et al., 2014b; Qiu, Xiao, & Chou, 2014a; Qiu, Xiao, & Lin, 2014b, 2014c; Xu, Wen, Wen, & Wu, 2014c; Xu, Zhou, Liu, He, & Zou, 2014a)) in response to the suggestion from (Chou, 2011), to establish a really useful predictor for a biological system, we need to consider the following steps: (i) select or construct a valid benchmark data-set to train and test the predictor; (ii) formulate the samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target 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; and (v) establish a user-friendly Web server for the predictor that is accessible to the public. Below, let us elaborate how to deal with these steps one by one.

Table 1. A set of four Web server predictors for identifying drug–target interaction networks based on the same benchmark data-set as in He et al. (2010).

Name	Interactive pair	Web site address	Reference
iGPCR-Drug	GPCR and drug	http://www.jci-bioinfo.cn/iGPCR-Drug/	Xiao et al., 2013b
iCDI-PseFpt	Channel and drug	http://www.jci-bioinfo.cn/iCDI-PseFpt/	Xiao et al., 2013a
iEzy-Drug	Enzyme and drug	http://www.jci-bioinfo.cn/iEzy-Drug/	Min et al., 2013
iNR-Drug	Nuclear receptor and drug	http://www.jci-bioinfo.cn/iNR-Drug/	Fan et al., 2014

2. Materials and methods

2.1. Optimization of imbalanced benchmark datasets

The original data used in (He et al., 2010) and (Fan et al., 2014; Min et al., 2013; Xiao et al., 2013a, 2013b) were collected from KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kotera, Hirakawa, Tokimatsu, Goto, & Kanehisa, 2012) at http://www.kegg.jp/kegg/. The original benchmark datasets used for the Web–server predictors iGPCR-Drug (Xiao et al., 2013b), iCDI-PseFpt (Xiao et al., 2013a), iEzy-Drug (Min et al., 2013), and iNR-Drug (Fan et al., 2014) as listed in Table 1 can be summarized as follows:

$$\begin{cases} \mathbb{S}_{\text{GPCR-Drug}}(1860) = \mathbb{S}^{+}_{\text{GPCR-Drug}}(620) \cup \mathbb{S}^{-}_{\text{GPCR-Drug}}(1240) \\ \mathbb{S}_{\text{Chl-PseFpt}}(4116) = \mathbb{S}^{+}_{\text{Chl-Drug}}(1372) \cup \mathbb{S}^{-}_{\text{Chl-Drug}}(2744) \\ \mathbb{S}_{\text{Ezy-Drug}}(8157) = \mathbb{S}^{+}_{\text{Ezy-Drug}}(2719) \cup \mathbb{S}^{-}_{\text{Ezy}-\text{Drug}}(5438) \\ \mathbb{S}_{\text{NR-Drug}}(258) = \mathbb{S}^{+}_{\text{NR-Drug}}(86) \cup \mathbb{S}^{-}_{\text{NR}-\text{Drug}}(172) \end{cases} \tag{1}$$

where $\mathbb{S}_{\text{GPCR-Drug}}(1860)$is the benchmark data-set for the iGPCR-Drug predictor (Xiao et al., 2013b), and it contains 1860 GPCR-drug pairs of which 620 are interactive pairs belonging to the positive subset $\mathbb{S}^{+}_{\text{GPCR-Drug}}(620)$ while 1240 are non-interactive belonging to the negative subset $\mathbb{S}^{-}_{\text{GPCR-Drug}}(1240)$, and $\cup$ represents the union in the set theory; $\mathbb{S}_{\text{Chl-PseFpt}}(4116)$ is the benchmark data-set for the iCDI-PseFpt predictor (Xiao et al., 2013a), and it contains 4116 channel–drug pairs of which 1372 are interactive pairs belonging to the positive subset $\mathbb{S}^{+}_{\text{Chl-Drug}}(1372)$ while 2744 are non-interactive belonging to the negative subset $\mathbb{S}^{-}_{\text{Chl-Drug}}(2744)$; $\mathbb{S}_{\text{Ezy-Drug}}(8157)$ is the benchmark data-set for the iEzy-Drug predictor (Min et al., 2013), and it contains 8157 enzyme–drug pairs of which 2179 are interactive pairs belonging to the positive subset $\mathbb{S}^{+}_{\text{Ezy-Drug}}(2719)$ while 5438 are non-interactive belonging to the negative subset $\mathbb{S}^{-}_{\text{Ezy-Drug}}(5438)$; and $\mathbb{S}_{\text{NR-Drug}}(258)$ is the benchmark data-set for the iNR-Drug predictor (Fan et al., 2014), and it contains 258 NR-drug pairs of which 86 are interactive pairs belonging to the positive subset $\mathbb{S}^{+}_{\text{NR-Drug}}(86)$ while 172 are non-interactive belonging to the negative subset $\mathbb{S}^{-}_{\text{NR-Drug}}(172)$. 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 (Kotera et al., 2012); while the "non-interactive" pair means that its two counterparts are not interacted with each other in the drug–target networks.

All the detailed data for the four benchmark datasets can be found in the Supplementary Materials of (Fan et al., 2014; Min et al., 2013; Xiao et al., 2013a, 2013b) or can be directly downloaded from their corresponding Web server predictors whose Web site addresses are explicitly given in Table 1.

As we can see from Equation (1), for the benchmark data-set used to train and test each of the aforementioned four predictors, the size of the negative subset is two times the size of the positive subset. Although this might reflect the real world in which the non-interactive pairs are always the majority compared with the interactive ones, a predictor trained with such a skewed benchmark data-set would have the consequence that many interactive drug–target pairs might be mispredicted as non-interactive ones (Sun et al., 2009). Actually, what is really most intriguing information for the drug-development scientists is the one about the interactive pairs. Therefore, it is worthwhile to find an effective approach to optimize the unbalanced benchmark data-set and minimize the consequence of this kind of misprediction.

In this study, we use the NCR (neighborhood cleaning rule) (Laurikkala, 2001) and the SMOTE (synthetic minority over-sampling technique) (Chawla, Bowyer, Hall, & Kegelmeyer, 2011) treatments to optimize the aforementioned skewed benchmark datasets. The former is to remove some redundant negative samples from the negative subset so as to reduce its statistical noise, which can be likened to the sample-screening procedure in computational proteomics (see, e.g. (Chou & Shen, 2006)). The latter is to add some hypothetical positive samples into the positive subset so as to enhance the ability in identifying the interactive pairs, which can be likened to the seed-propagation approach in Zhang and Chou (1995) and the Monte Calo sampling approach in Chou (1993), Zhang and Chou (1992) for expanding the positive subsets.

In this study, we applied the NCR treatment (Laurikkala, 2001) according to the following criteria: (i) for each of the samples in the benchmark data-set, find its three nearest neighbors; (ii) if the sample concerned belongs to a negative subset and at least two of its three nearest neighbors belong to the positive subset, remove the sample from the benchmark data-set; (iii) if, however, it belongs to a positive subset, then remove those of its nearest neighbors from the benchmark data-set that belong to the negative subset; (iv) if the number of samples in a negative subset is less than 200 such as in the case of NR–drug system, no action will be taken to remove the negative samples.

After the aforementioned NCR treatment, the number of samples in each of the four negative subsets was reduced, and hence, Equation (1) would become

$$\begin{cases} \mathbb{S}_{\text{GPCR-Drug}}(1428) = \mathbb{S}^{+}_{\text{GPCR-Drug}}(620) \cup \mathbb{S}^{-}_{\text{GPCR-Drug}}(808) \\ \mathbb{S}_{\text{Chl-Drug}}(3309) = \mathbb{S}^{+}_{\text{Chl-Drug}}(1372) \cup \mathbb{S}^{-}_{\text{Chl-Drug}}(1937) \\ \mathbb{S}_{\text{Ezy-Drug}}(6957) = \mathbb{S}^{+}_{\text{Ezy-Drug}}(2719) \cup \mathbb{S}^{-}_{\text{Ezy-Drug}}(4240) \\ \mathbb{S}_{\text{NR-Drug}}(258) = \mathbb{S}^{+}_{\text{NR-Drug}}(86) \cup \mathbb{S}^{-}_{\text{NR-Drug}}(172) \end{cases} \tag{2}$$

See Supporting Information S1 for the detailed data obtained by the NCR treatment described above.

Subsequently, to further optimize the benchmark datasets of Equation (2), the SMOTE approach (Chawla et al., 2011) was adopted to create some hypothetical samples for the positive subsets by the linear interpolation scheme. Finally, the benchmark datasets thus obtained can be formulated as

$$\begin{cases} \mathbb{S}_{\text{GPCR-Drug}}(1616) = \mathbb{S}^{+}_{\text{GPCR-Drug}}(808) \cup \mathbb{S}^{-}_{\text{GPCR-Drug}}(808) \\ \mathbb{S}_{\text{Chl-Drug}}(3874) = \mathbb{S}^{+}_{\text{Chl-Drug}}(1937) \cup \mathbb{S}^{-}_{\text{Chl-Drug}}(1937) \\ \mathbb{S}_{\text{Ezy-Drug}}(8480) = \mathbb{S}^{+}_{\text{Ezy-Drug}}(4240) \cup \mathbb{S}^{-}_{\text{Ezy-Drug}}(4240) \\ \mathbb{S}_{\text{NR-Drug}}(344) = \mathbb{S}^{+}_{\text{NR-Drug}}(172) \cup \mathbb{S}^{-}_{\text{NR-Drug}}(172) \end{cases} \tag{3}$$

As we can see from Equation (3), the four optimized benchmark datasets via the NCR (Laurikkala, 2001) and SMOTE (Chawla et al., 2011) treatments are well balanced out, each having its positive and negative subset equal to each other in size.

Note that the hypothetical samples generated via the linear interpolation scheme in SMOTE can only be expressed by their feature vectors as defined in the next section, but not real sample codes as given in the Online Supporting Information S1. Nevertheless, it would be perfectly reasonable to do so since the data directly used to train a predictor were actually the samples' feature vectors, but not their codes. This is the key to optimize an imbalanced benchmark data-set in the current study, and the rationale of such an interesting approach will be further elucidated in Section 3.2 later.

To provide an intuitive picture, a flowchart is given in Figure 2 to illustrate the process of how to optimize an imbalance benchmark data-set.

2.2. *Sample representation*

Since each of the samples in the current network system contains a drug compound and a target protein. The latter can be a GPCR, ion-channel, enzyme, or NR.

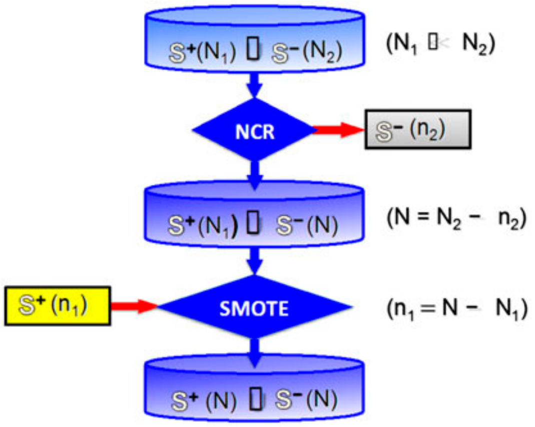


Figure 2. A flowchart to show the process of converting an imbalanced benchmark data-set to a balanced one by NCR (neighborhood cleaning rule) and SMOTE.
Note: In the figure, N_1 and N_2 represent the numbers of samples in the original positive and negative subsets, respectively; n_2, the number of the negative samples removed by the NCR treatment; and n_1, the number of the positive hypothetical samples created by SMOTE and added to the final balanced benchmark data-set. See the relevant text for further explanation.

For the drug compound **D**, we used the two-dimensional molecular fingerprint (Steffen, Kogej, Tyrchan, & Engkvist, 2009; Willett, 2006; Yap, 2011) to represent it that contains 256 components as elaborated in Xiao et al. (2013b); that is,

$$\mathbf{D} = [\, d_1 \quad \cdots \quad d_j \quad \cdots \quad d_{256} \,]^{\mathbf{T}} \tag{4}$$

where $d_j (j = 1, 2, \ldots, 256)$ have been clearly defined by Equations (2)–(5) of Xiao et al. (2013b), and hence, there is no need to repeat here, and **T** is the matrix transpose operator.

For the protein molecule **P**, such as GPCR, ion-channel, enzyme, or NR, we used the pseudo amino acid composition (Chou, 2001c, 2005d) or Chou's PseAAC (Cao, Xu, & Liang, 2013; Du, Gu, & Jiao, 2014; Lin & Lapointe, 2013; Zhong & Zhou, 2014) to represent it, which contains 480 components as elaborated in Min et al. (2013); that is,

$$\mathbf{P} = [\, \psi_1 \quad \psi_2 \quad \cdots \quad \psi_u \quad \cdots \quad \psi_{480} \,]^{\mathbf{T}} \tag{5}$$

where $\psi_u (u = 1, 2, \cdots, 480)$ have been given in Equations 12–23 in (Min et al., 2013), and hence, there is no need to repeat here.

Thus, the drug–protein pair Φ can be formulated by combining Equations (4) and (5) via an orthogonal sum as given below

$$\Phi = \mathbf{D} \oplus \mathbf{P} = [\, \phi_1 \quad \phi_2 \quad \cdots \quad \phi_u \quad \cdots \quad \phi_{736} \,]^{\mathbf{T}} \tag{6}$$

where the symbol $\oplus$ represents the sign of orthogonal sum (Chou & Cai, 2006), and $\phi_u (u = 1, 2, \ldots, 736)$ is the u-th component after combining the 256 components in Equation (4) and 480 components in Equation (5). Note that all the components in Equation (6) are subjected to a standard conversion as described by the following equation:

$$\phi_u \Leftarrow \frac{\phi_u\text{-}\langle\phi\rangle}{\mathrm{SD}(\phi)} \tag{7}$$

where $\langle\phi\rangle$ means the average of the 736 components in Equation (6), and SD means the corresponding standard deviation. The converted values obtained by Equation (7) will have a zero mean value, and will remain unchanged if they go through the same conversion procedure again (Chou & Shen, 2007).

As an illustration, the 736 standard converted components for each of the 172 positive samples (including 86 hypothetical samples created by SMOTE) or each of the 172 negative samples in the NR-drug system (cf. Equation (3)) are given in the Supporting Information S2.

2.3. *Operation engine or algorithm*

In this study, the operation engine was SVM (support vector machine), which is based on the structural risk minimization principle from statistical learning theory. SVM has been widely used in the realm of bioinformatics (see, e.g. (Chen et al., 2014b; Chen, Feng, & Lin, 2013, 2014a; Ding, Deng, Yuan, & Liu, 2014; Feng, Chen, & Lin, 2013; Guo et al., 2014; Liu et al., 2014a; Liu et al., 2014b; Liu, Wang, & Chou, 2005; Qiu et al., 2014a; Xu, Wen, Shao, & Deng, 2014b)). The basic idea of SVM is to construct a separating hyper-plane so as to maximize the margin between the positive data-set and negative data-set. The nearest two points to the hyper-plane are called support vectors. SVM first constructs a hyperplane based on the training data-set, and then maps an input vector from the input space into a vector in a higher-dimensional Hilbert space, where the mapping is determined by a kernel function. A trained SVM can output a class label (in our case, interactive pair or non-interactive pair) based on the mapping vector of the input vector. For a brief formulation of SVM and how it works, see the papers (Cai, Zhou, & Chou, 2003; Chou & Cai, 2002); for more details about SVM, see a monograph (Cristianini & Shawe-Taylor, 2000). In this study, the LIBSVM package (Chang & Lin, 2005) was used as an implementation of SVM, which can be downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm/, and the popular radial basis function (RBF) was taken as the kernel function. For the current SVM classifier, there were two uncertain parameters: penalty parameter C and kernel parameter γ. Their values will be given later.

A package of prediction methods thus obtained is called iDrug-Target, which consists of four predictors; that is,

$$\text{iDrug-Target} = \begin{cases} \text{iDrug-GPCR}, & \text{for drug-GPCR interaction} \\ \text{iDrug-Chl}, & \text{for drug-channel interaction} \\ \text{iDrug-Ezy}, & \text{for drug-enzyme interaction} \\ \text{iDrug-NR}, & \text{for drug-NR interaction} \end{cases} \tag{8}$$

where the two parameters for the SVM operation engine are given by

$$\begin{cases} C = 2^{2.80}, \gamma = 2^{-7.0} & \text{for iDrug-GPCR} \\ C = 2^{10.8}, \gamma = 2^{-5.0} & \text{for iDrug-Channel} \\ C = 2^{12.8}, \gamma = 2^{-6.6} & \text{for iDrug-Enzyme} \\ C = 2^{2.40}, \gamma = 2^{-9.0} & \text{for iDrug-NR} \end{cases} \tag{9}$$

which were determined by optimizing the 5-fold cross-validation success rate for each of the four predictors on its corresponding benchmark data-set (cf. Equation (3)) through a two-dimensional grid search as illustrated in Figure 3.

3. Results and discussion

As mentioned in the beginning of this study, one of the important procedures in developing a new predictor is how to properly and objectively evaluate its quality (Chou, 2011), which actually comprises two aspects. One is what metrics should be taken to quantitatively measure the prediction accuracy, and the other is what test method should be used to perform the test. Below, let us address these problems.

3.1. *A set of four metrics for performance measurement*

In order to provide an intuitive and easier-to-understand quantitative scale, here, let us adopt the criteria proposed in Chou (2001a). According to those criteria, the rates of correct predictions for the interactive drug–target pairs in the positive subset and the non-interactive pairs in the negative subset are, respectively, defined by

$$\begin{cases} \Lambda^{+} = \frac{N^{+}-N_{-}^{+}}{N^{+}}, & \text{for interactive drug} - \text{target pairs} \\ \Lambda^{-} = \frac{N^{-}-N_{-}^{-}}{N^{-}}, & \text{for the non} - \text{interactive drug} \ - \text{target pairs} \end{cases} \tag{10}$$

where $N^{+}$is the total number of the interactive drug–target (e.g. drug-GPCR) pairs investigated while N_{-}^{+} the number of the interactive drug–target pairs incorrectly predicted as the non-interactive drug–target pairs; $N^{-}$the total number of the non-interactive drug–target pairs investigated while N_{+}^{-} is the number of the non-interactive drug–target pairs incorrectly predicted as the interactive drug–target pairs. Thus, the overall success prediction rate is given by Chou (2001b)

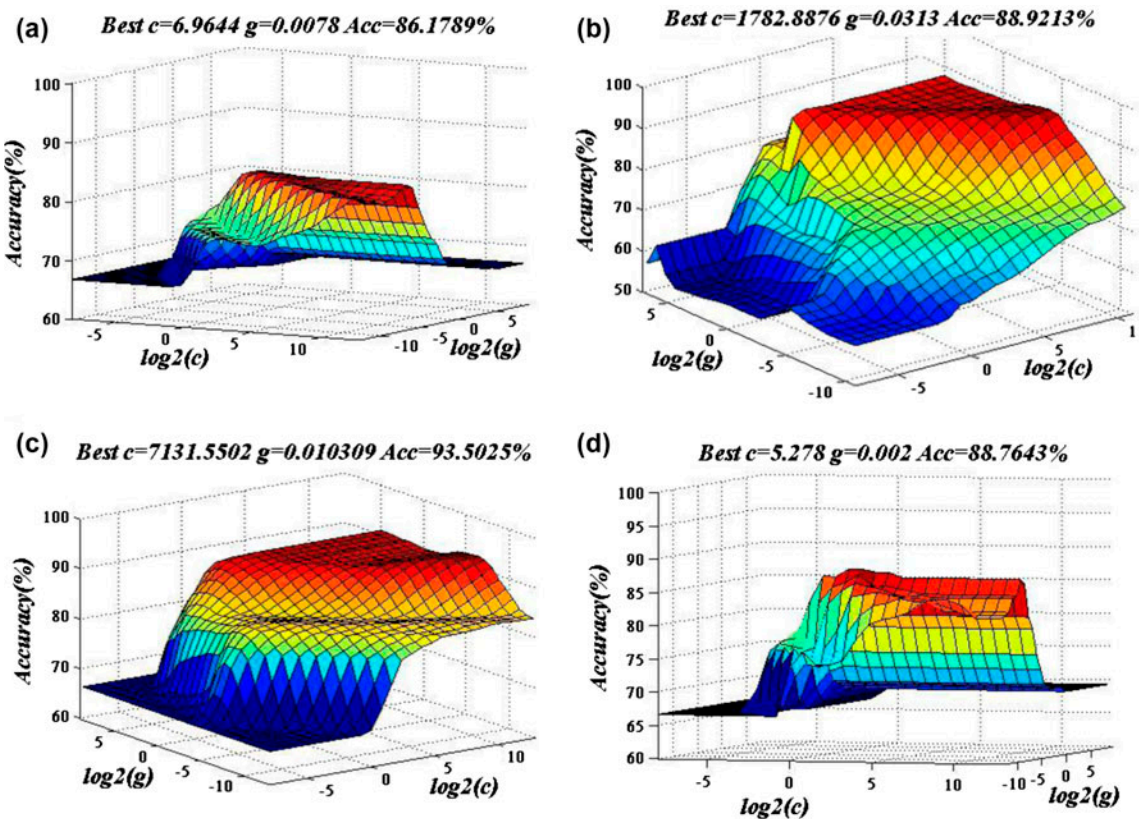


Figure 3. Three-dimensional plot to show how to find the optimal values of C and γ via **a** two-dimensional grid search. Panel (a) for iDrug-GPCR predictor; (b) for iDrug-Chl; (c) for iDrug-Ezy; and (d) for iDrug-NR. See Section 2.3 as well as Equations (8) and (9) for further explanation.

$$\Lambda = \frac{\Lambda^{+}N^{+} + \Lambda^{-}N^{-}}{N^{+} + N^{-}} = 1 - \frac{N_{-}^{+} + N_{+}^{-}}{N^{+} + N^{-}} \tag{11}$$

It is obvious from Equations (10) and (11) that, if and only if none of the interactive drug–target pairs and the non-interactive drug–target pairs are mispredicted, that is,$N_{-}^{+} = N_{+}^{-} = 0$ and $\Lambda^{+} = \Lambda^{-} = 1$, we have the overall success rate $\Lambda = 1$. Otherwise, the overall success rate would be smaller than 1.

It is instructive, however, to point out that the following equation is often used in literatures for examining the performance quality of a predictor (see, e.g. (Chen, Liu, Yang, & Chou, 2007))

$$\begin{cases} \mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP+FN}} \\ \mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN+FP}} \\ \mathrm{Acc} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \\ \mathrm{MCC} = \frac{(\mathrm{TP}\times\mathrm{TN})-(\mathrm{FP}\times\mathrm{FN})}{\sqrt{(\mathrm{TP+FP})(\mathrm{TP+FN})(\mathrm{TN+FP})(\mathrm{TN+FN})}} \end{cases} \tag{12}$$

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 Mathews correlation coefficient.

The relations between the symbols in Equation (11) and those in Equation (12) are given by

$$\begin{cases} \mathrm{TP} = N^{+} - N_{-}^{+} \\ \mathrm{TN} = N^{-} - N_{+}^{-} \\ \mathrm{FP} = N_{+}^{-} \\ \mathrm{FN} = N_{-}^{+} \end{cases} \tag{13}$$

Substituting Equation (13) into Equation (12) and also considering Equation (11), we obtain

$$\begin{cases} \mathrm{Sn} = 1 - \frac{N_{-}^{+}}{N^{+}} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \frac{N_{+}^{-}}{N^{-}} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \Lambda = 1 - \frac{N_{-}^{+}+N_{+}^{-}}{N^{+}+N^{-}} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \frac{1-\left(\frac{N_{-}^{+}}{N^{+}}+\frac{N_{+}^{-}}{N^{-}}\right)}{\sqrt{\left(1+\frac{N_{+}^{-}-N_{-}^{+}}{N^{+}}\right)\left(1+\frac{N_{-}^{+}-N_{+}^{-}}{N^{-}}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \tag{14}$$

From the above equation, we can see: when $N_{-}^{+} = 0$ meaning that none of the interactive drug-target pairs was mispredicted to be a non-interactive drug–target pairs, we have the sensitivity $\mathrm{Sn} = 1$; while $N_{-}^{+} = N^{+}$ meaning that all the interactive drug–target pairs were mispredicted

Table 2. A comparison of iDrug-Target[1] with the existing predictors on the same experiment-confirmed data.

System	Predictor	Acc (%)	MCC	Sn (%)	Sp (%)
Drug-GPCR	FunD[b]	78.49	N/A	N/A	N/A
	iGPCR-Drug[c]	85.50	.6775	80.00	88.30
	iDrug-Target	**90.32**	**.8066**	97.58	86.69
Drug–channel	FunD[b]	80.78	N/A	N/A	N/A
	iCDI-Drug[d]	87.27	.7233	86.30	87.76
	iDrug-Target[a]	**88.78**	**.7643**	91.98	87.17
Drug–enzyme	FunD[b]	85.48	N/A	N/A	N/A
	iEzy-Drug[e]	91.03	.8039	90.81	91.14
	iDrug-Target[a]	**92.56**	**.8429**	95.99	90.84
Drug-NR	FunD[b]	85.66	N/A	N/A	N/A
	iNR-Drug[f]	89.15	.7519	79.07	94.19
	iDrug-Target[a]	**93.02**	**.8453**	91.86	93.60

[a] Trained by the optimized benchmark datasets as defined in Eq. 3, the iDrug-Target package contains four predictors for identifying the networking interactions of drugs with GPCR, channels, enzymes, and NR, respectively (cf. Eq. 4). The rates reported in this table were derived by the target-jackknife cross-validations on the original experimental benchmark datasets used by FunD (He et al., 2010) and iGPCR-Drug (Xiao et al., 2013b), iCDI-Drug (Xiao et al., 2013a), iEzy-Drug (Min et al., 2013), and iNT-Drug (Fan et al., 2014), respectively. See Section 2.5 for further explanation.
[b] See Ref. He et al., 2010) for the Fund prediction method and its reported success rates.
[c] See Ref. Xiao et al., 2013b for the iGPCR-Drug predictor and its reported success rates.
[d] See Ref. Xiao et al., 2013a for the iCDI-Drug predictor and its reported success rates.
[e] See Ref. Min et al., 2013 for the iEzy-Drug predictor and its reported success rates.
[f] See Ref. Fan et al., 2014 for the iNR-Drug predictor and its reported success rates.

to be the non-interactive drug–target pairs, we have the sensitivity Sn = 0. Likewise, when $N_{+}^{-} = 0$ meaning that none of the non-interactive drug-target pairs was mispredicted, we have the specificity Sp = 1; while $N_{+}^{-} = N^{-}$ meaning that all the non-interactive drug–target pairs were incorrectly predicted as the interactive drug–target pairs, we have the specificity Sp = 0. When $N_{-}^{+} = N_{+}^{-} = 0$ meaning that none of interactive drug–target pairs in the positive subset and none of the non-interactive drug–target pairs in $\mathbb{S}^{-}$ was incorrectly predicted, we have the overall accuracy Acc = Λ = 1; while $N_{-}^{+} = N^{+}$and $N_{+}^{-} = N^{-}$ meaning that all the interactive drug–target pairs in the positive subset and all the non-interactive drug–target pairs in the negative subset were mispredicted, we have the overall accuracy Acc = Λ = 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 drug–target pairs in the positive subset and none of non-interactive drug–target pairs in the negative subset was mispredicted, we have MCC = 1; when $N_{-}^{+} = N^{+}/2$, and $N_{+}^{-} = N^{-}/2$ we have Mcc = 0 meaning that 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 Equation (14) to examine a predictor for its sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient, particularly for its Mathew's correlation coefficient.

It should be pointed out, however, the set of metrics as defined in Equation (14) or Equation (12) is valid only for the single-label systems. For the multi-label systems whose emergence has become more frequent in systems biology (Chou, Wu, & Xiao, 2011, 2012; Lin, Fang, & Xiao, 2013) and systems medicine (Chen, Zeng, Cai, & Feng, 2012; Xiao, Wang, Lin, & Jia, 2013c), a completely different set of metrics as defined in (Chou, 2013) is needed.

3.2. *Jackknife and target-jackknife 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 data-set test, subsampling (or *K*-fold cross-validation) test, and jackknife test (Chou & Zhang, 1995). However, of the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark data-set as elaborated in Chou & Shen (2010) and demonstrated by Equations 28–30 in (Chou, 2011). Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g. (Du, Jiang, & He, 2006; Hajisharifi, Piryaiee, Mohammad Beigi, Behbahani, & Mohabatkar, 2014; Mohabatkar, Beigi, Abdolahi, & Mohsenzadeh, 2013; Mondal & Pai, 2014; Nanni, Brahnam, & Lumini, 2014; Shen & Chou, 2010; Shen, Yang, & Chou, 2007; Xiao, Wu, & Chou, 2011; Xu, Shao, & Wu, 2013)). During the process of jackknife

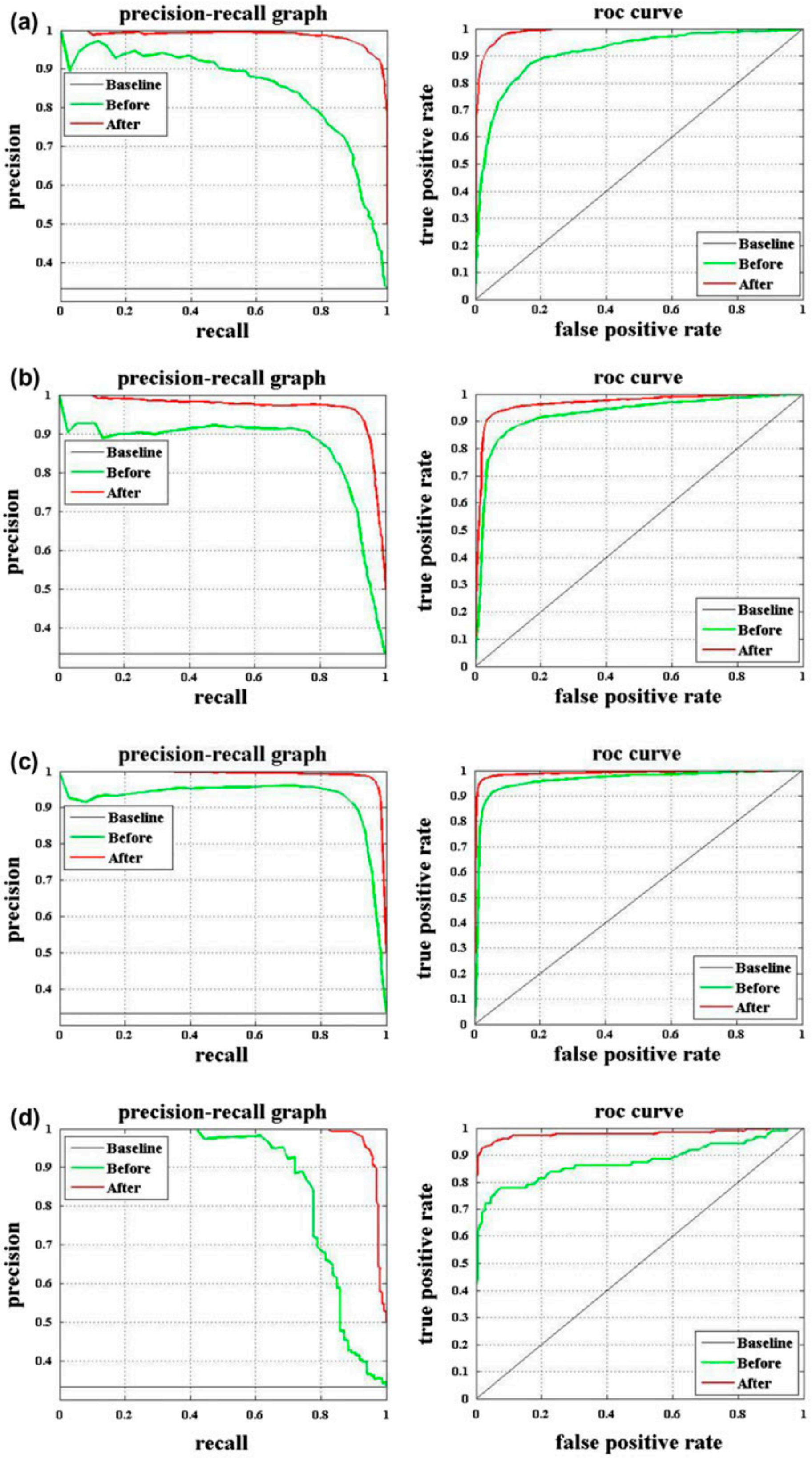
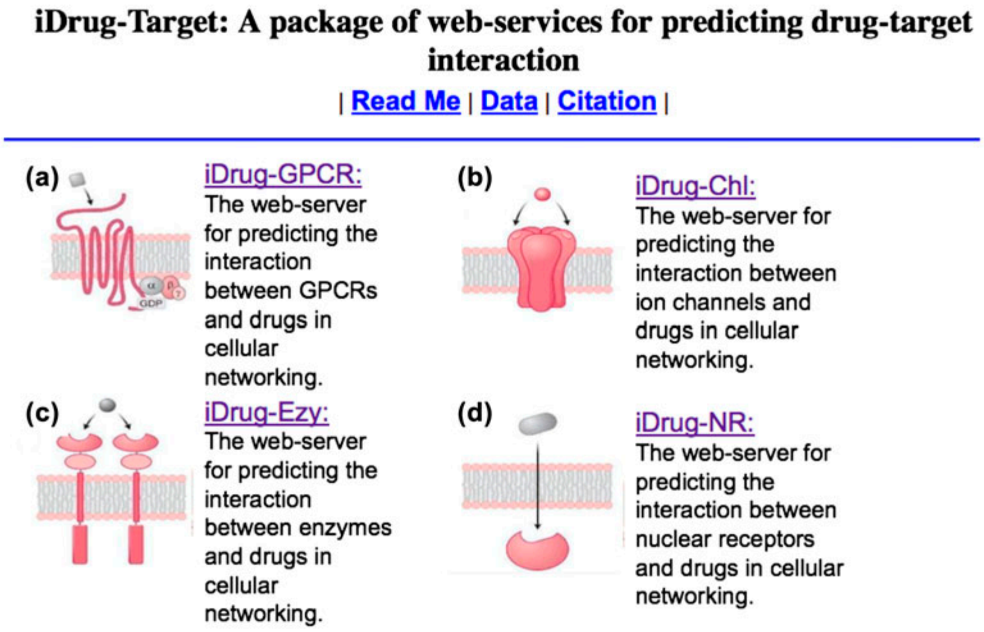


Figure 4. The ROC and PR curves to show the predictor's quality.
Note: The green line is for the existing predictors (a) iGPCR-Drug (Xiao et al., 2013b), (b) iCDL-Drug (Xiao et al., 2013a), (c) iEzy-Drug (Min et al., 2013), and (b) iNR-Drug (Fan et al., 2014) while the red line for the corresponding predictors in the iDrug-Target package. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this paper.)



iDrug-Target: A package of web-services for predicting drug-target interaction

| Read Me | Data | Citation |

(a) iDrug-GPCR: The web-server for predicting the interaction between GPCRs and drugs in cellular networking.

(b) iDrug-Chl: The web-server for predicting the interaction between ion channels and drugs in cellular networking.

(c) iDrug-Ezy: The web-server for predicting the interaction between enzymes and drugs in cellular networking.

(d) iDrug-NR: The web-server for predicting the interaction between nuclear receptors and drugs in cellular networking.

Figure 5. A semi-screenshot to show the top page of the iDrug-Target Web server at http://www.jci-bioinfo.cn/iDrug-Target/.

test, all the samples in the benchmark data-set will be singled out one by one and tested by the predictor trained by the remaining samples.

When conducting the jackknife test on the optimized benchmark data-set of Equation (3), however, some special consideration is needed. Take the optimized benchmark data-set for the GPCR-drug system as an example: both its positive subset and negative subset contain 808 samples. But, of the 808 positive samples, only $(808 - 188) = 620$ are from experimental observations (cf. Equations (2) and (3)) and the rest from the SMOTE treatment (Chawla et al., 2011). Also, in the negative subset, $(1240 - 808) = 432$ experimental samples have been removed by NCR (Laurikkala, 2001) (cf. Equations (1) and (2)). Since the validation should be carried out strictly based on the experimental data only, a special jackknife test, the so-called "target-jackknife test", was introduced. During the process of target-jackknife test, only the experiment-confirmed samples are in turn singled out as a target (or test sample) for cross-validation. Accordingly, although the predictor is trained by the optimized benchmark data-set that includes both experimental and hypothetical samples, only or all, the experiment-confirmed samples are the targets used to count its success rates regardless of whether they are part of a subset or removed from the benchmark data-set during the optimization process. For instance, for the aforementioned GPCR-drug system (cf. Equation (3)), only the 620 experimental positive samples need to be singled out for cross-validation; however, even the 432 experimental negative samples need to be validated as well despite they have been removed from the negative subset during the optimization process.

3.3. *Comparison with the existing predictors*

The scores for the four metrics as defined in Equations (12) or (14) achieved by the current iDrug-Target predictor via the target-jackknife tests are given in Table 2, where for facilitating comparison the corresponding scores by the existing predictors are also listed. From the table, we can see the following. (i) The scores of the overall accuracy (Acc) achieved by the four predictors in the iDrug-Target package are remarkably higher than those of the existing predictors for the same purposes. (ii) The scores of the Mathew's correlation coefficient (MCC) by the iDrug-Target package are also remarkably higher than those of the existing predictors. These facts indicate that the current predictors in the iDrug-Target package not only can yield higher prediction accuracy but also are more stable and consistent.

Shown in Figure 4 is a graphic comparison of the four predictors in the iDrug-Target with their counterparts via the ROC (receiver operating characteristic) curves and PR (precision–recall) curves. As we can see from the figure, the areas under both the ROC and PR curves for the four predictors in the iDrug-Target package are obviously larger than those of their counterparts, indicating a clear improvement of the new predictors in comparison with the old ones.

It is instructive to point out that, although the four predictors in the iDrug-Target package were trained by the four optimized benchmark datasets in which some experimental negative samples were removed from the original benchmark datasets to balance out the sizes of subsets, they were still counted in the target-jackknife cross-validation. On the other hand, although some hypothetical positive samples were added to form the optimized benchmark datasets, only the experimental samples were counted in calculating the metrics scores. In other words, the objects counted during the cross-validation, regardless of whether they are positive or negative samples, are exactly the same as those counted by the other methods in Table 2.

3.4. *Web server and user guide*

For those who are interested in using the iDrug-Target package, but not its mathematical details, a Web server was established. Below, let us give a step-by-step guide on how to use the Web server to get the desired results.

Step 1. Open the Web server at http://www.jci-bioinfo.cn/iDrug-Target/, and you will see the top page of the iDrug-Target on your computer screen, as shown in Figure 5. Click on the Read Me button to see a brief introduction about the iDrug-Target package and the caveat when using it.

Step 2. Click one of the four predictors according to your need. For instance, if you wish to predict the drug-GPCR interaction, click the button iDrug-GPCR, and follow the instructions on the screen to get your desired results.

Step 3. If you wish to predict the interaction of drugs with other targets, click the Close button to bring you back to the top page. Then, repeat Step 2 but click a different predictor such as iDrug-Chl, iDrug-Ezy, or iDrug-NR as you desire.

4. Conclusion

The strategy of optimizing the training data-set via the NCR (Laurikkala, 2001) and SMOTE (Chawla et al., 2011) approaches can remarkably improve the prediction quality of a predictor, as indicated by the rigorous target-jackknife tests in which only the experiment-confirmed data were examined. This is particularly true for the case when the predictor was originally trained by a highly unbalanced or skewed benchmark data-set in which the negative subset data-set is overwhelmingly larger than the positive one.

It is anticipated that the new package called **iDrug-Target** developed in this paper with the optimized training datasets will become a very useful high throughput toll for both basic research and drug development.

It is anticipated that the current strategy and novel technique can also be used to improve all those existing statistical predictors that were trained by highly unbalanced training datasets.

Acknowledgments

This work was partially supported by the National Nature Science Foundation of China (No. 31260273), the Jiangxi Provincial Foreign Scientific and Technological Cooperation Project (No.20120BDH80023), Natural Science Foundation of Jiangxi Province, China (No. 20122BAB211033, 20122BAB201044, 20122BAB2010), the LuoDi plan of the Department of Education of JiangXi Province (KJLD12083), 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.

Funding

National Nature Science Foundation of China [Grant Number 31260273].

References

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.

Chang, C., Lin, C. 2005. LIBSVM: A library for support vector machines (2001) [Software].

Cai, Y. D., Zhou, G. P., & Chou, K. C. (2003). Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical Journal, 84*, 3257–3263.

Call, M. E., Wucherpfennig, K. W., & Chou, J. J. (2010). The structural basis for intramembrane assembly of an activating immunoreceptor complex. *Nature Immunology, 11*, 1023–1029.

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.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2011). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357.

Chen, W., Feng, P. M., Deng, E. Z., & Lin, H. (2014b). iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical Biochemistry, 462*, 76–83.

Chen, W., Feng, P. M., & Lin, H. (2013). iRSpot-PseDNC: Identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Research, 41*, e68.

Chen, W., Feng, P. M., & Lin, H. (2014). iSS-PseDNC: Identifying splicing sites using pseudo dinucleotide composition. *Biomed Research International, 2014*, 623149.

Chen, J., Liu, H., Yang, J., & Chou, K. C. (2007). Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. *Amino Acids, 33*, 423–428.

Chen, L., Zeng, W. M., Cai, Y. D., & Feng, K. Y. (2012). Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical–chemical interactions and similarities. *PLoS ONE, 7*, e35254.

Chou, K. C. (1993). A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. *Journal of Biological Chemistry, 268*, 16938–16948.

Chou, K. C. (2001a). Prediction of protein signal sequences and their cleavage sites. *Proteins: Struct, Funct, Genet, 42*, 136–139.

Chou, K. C. (2001b). Prediction of signal peptides using scaled window. *Peptides, 22*, 1973–1979.

Chou, K. C. (2001c). Prediction of protein cellular attributes using pseudo amino acid composition. *PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol. 44, 60), 43*, 246–255.

Chou, K. C. (2004a). Insights from modelling three-dimensional structures of the human potassium and sodium channels. *Journal of Proteome Research, 3*, 856–861.

Chou, K. C. (2004b). Review: Structural bioinformatics and its impact to biomedical science. *Current Medicinal Chemistry, 11*, 2105–2134.

Chou, K. C. (2005a). Insights from modeling the 3D structure of DNA−CBF3b complex. *Journal of Proteome Research, 4*, 1657–1660.

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. (2005c). Modeling the tertiary structure of human cathepsin-E. *Biochemical and Biophysical Research Communications, 331*, 56–60.

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

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. (2013). Some remarks on predicting multi-label attributes in molecular biosystems. *Molecular BioSystems, 9*, 1092–1100.

Chou, K. C., & Cai, Y. D. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry, 277*, 45765–45769.

Chou, K. C., & Cai, Y. D. (2006). Predicting protein–protein interactions from sequences in a hybridization space. *Journal of Proteome Research, 5*, 316–322.

Chou, K. C., & Howe, W. J. (2002). Prediction of the tertiary structure of the beta-secretase zymogen. *Biochemistry and Biophysics Research Communication, 292*, 702–708.

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., & Shen, H. B. (2006). Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic *K*-nearest neighbor classifiers. *Journal of Proteome Research, 5*, 1888–1897.

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. (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.

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., Du, Q. S., Sirois, S., & Zhong, W. Z. (2006). Review: Progress in computational approach to drug development against SARS. *Current Medicinal Chemistry, 13*, 3263–3270.

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 (Erratum: ibid., 2003, Vol. 310, 675). *Biochemistry and Biophysics Research Communication, 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.

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

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction of support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Ding, H., Deng, E. Z., Yuan, L. F., & Liu (2014). iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. *BioMed Research International, 2014*, 286419.

Du, P., Gu, S., & Jiao, Y. (2014). PseAAC-general: Fast building various modes of general form of chou's pseudo-amino acid composition for large-scale protein datasets. *International Journal of Molecular Sciences, 15*, 3495–3506.

Du, Q. S., Huang, R. B., & Wang, C. H. (2009). Energetic analysis of the two controversial drug binding sites of the M2 proton channel in influenza A virus. *Journal of Theoretical Biology, 259*, 159–164.

Du, Q. S., Huang, R. B., & Wang, S. Q. (2010). Designing inhibitors of M2 proton channel against H1N1 swine influenza virus. *PLoS ONE, 5*, e9388.

Du, Q. S., Jiang, Z. Q., & He, W. Z. (2006). Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction. *Journal of Biomolecular Structure and Dynamics, 23*, 635–640.

Ewing, T. J., Makino, S., Skillman, A. G., & Kuntz, I. D. (2001). DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer-Aided Molecular Design, 15*, 411–428.

Fan, Y. N., Xiao, X., & Min, J. L. (2014). iNR-Drug: Predicting the interaction of drugs with nuclear receptors in cellular networking. *International Journal of Molecular Sciences, 15*, 4915–4937.

Feng, P. M., Chen, W., & Lin, H. (2013). iHSP-PseRAAAC: Identifying the heat shock protein families using pseudo reduced amino acid alphabet composition. *Analytical Biochemistry, 442*, 118–125.

Guo, S. H., Deng, E. Z., Xu, L. Q., Ding, H., & Lin, H. (2014). iNuc-PseKNC: A sequence based predictor for predicting nucleosome positioning in genomes with pseudo k–tuple nucleotide composition. *Bioinformatics, 30*, 1522–1529.

Hajisharifi, Z., Piryaiee, M., Mohammad Beigi, M., Behbahani, M., & Mohabatkar, H. (2014). Predicting anticancer peptides with Chou's pseudo amino acid composition and investigating their mutagenicity via Ames test. *Journal of Theoretical Biology, 341*, 34–40.

He, Z., Zhang, J., Shi, X. H., Hu, L. L., & Kong, X. (2010). Predicting drug–target interaction networks based on functional groups and biological features. *PLoS ONE, 5*, e9603.

Kotera, M., Hirakawa, M., Tokimatsu, T., Goto, S., & Kanehisa, M. (2012). The KEGG databases and tools facilitating omics analysis: Latest developments involving human diseases and pharmaceuticals. *Methods in Molecular Biology, 802*, 19–39.

Laurikkala, J. (2001). *Improving identification of difficult small classes by balancing class distribution* (pp. 63–66). Berlin Heidelberg: Springer.

Li, X. B., Wang, S. Q., Xu, W. R., & Wang, R. L. (2011). Novel inhibitor design for hemagglutinin against H1N1 influenza virus by core hopping method. *PLoS ONE, 6*, e28111.

Liao, Q. H., Gao, Q. Z., & Wei, J. (2011). Docking and molecular dynamics study on the inhibitory activity of novel inhibitors on epidermal growth factor receptor (EGFR). *Medicinal Chemistry, 7*, 24–31.

Lin, H., Deng, E. Z., Ding, H., & Chen, W. (2014). iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Research, 42*, 12961–12972.

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

Lin, S. X., & Lapointe, J. (2013). Theoretical and experimental biology in one. *Journal Biomedical Science and Engineering (JBiSE), 6*, 435–442.

Lindsay, M. A. (2003). Target discovery. *Nature Reviews Drug Discovery, 2*, 831–838.

Liu, H., Wang, M., & Chou, K. C. (2005). Low-frequency Fourier spectrum for predicting membrane protein types. *Biochem Biophys Research Communication (BBRC), 336*, 737–739.

Liu, Z., Xiao, X., Qiu, W. R. (2015). iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. *Analytical Biochemistry*, in press.

Liu, B., Xu, J., Lan, X., Xu, R., & Zhou, J. (2014a). iDNA-Prot|dis: Identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS ONE, 9*, e106691.

Liu, B., Zhang, D., Xu, R., Xu, J., Wang, X., & Chen, Q. (2014b). Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics, 30*, 472–479.

Ma, Y., Wang, S. Q., & Xu, W. R. (2012). Design novel dual agonists for treating type-2 diabetes by targeting peroxisome proliferator-activated receptors with core hopping approach. *PLoS ONE, 7*, e38546.

Min, J. L., Xiao, X., & Chou, K. C. (2013). iEzy-Drug: A web server for identifying the interaction between enzymes and drugs in cellular networking. *Biomedical Research International, 2013*, 701317.

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.

Mondal, S., & Pai, P. P. (2014). Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. *Journal of Theoretical Biology, 356*, 30–35.

Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R., Hart, W. E., Belew, R. K., & Olson, A. J. (1998). Automated docking using a lamarckian genetic algorithm and empirical binding free energy function. *Journal of Computational Chemistry, 19*, 1639–1662.

Nanni, L., Brahnam, S., & Lumini, A. (2014). Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. *Journal of Theoretical Biology, 360C*, 109–116.

OuYang, B., & Chou, J. J. (2014). The minimalist architectures of viroporins and their therapeutic implications. *Biochimica et Biophysica Acta, 1838*, 1058–1067.

OuYang, B., Xie, S., Berardi, M. J., Zhao, X. M., Dev, J., Yu, W., … Chou, J. J. (2013). Unusual architecture of the p7 channel from hepatitis C virus. *Nature, 498*, 521–525.

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, 102*, 10870–10875.

Pielak, R. M., & Chou, J. J. (2010). Solution NMR structure of the V27A drug resistant mutant of influenza A M2 channel. *Biochemical and Biophysical Research Communications, 401*, 58–63.

Pielak, R. M., Schnell, Jason R., & Chou, J. J. (2009). Mechanism of drug inhibition and drug resistance of influenza A M2 channel. *Proceedings of National Academy of Science, USA, 106*, 7379–7384.

Qiu, W. R., Xiao, X., & Chou, K. C. (2014a). iRSpot-TNCPseAAC: Identify recombination spots with trinucleotide composition and pseudo amino acid components. *International Journal of Molecular Sciences, 15*, 1746–1766.

Qiu, W. R., Xiao, X., & Lin, W. Z. (2014b). iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model. *Journal of Biomolecular Structure and Dynamics (JBSD)*, doi:10.1080/07391102.2014.968875.

Qiu, W. R., Xiao, X., & Lin, W. Z. (2014c). iMethyl-PseAAC: Identification of protein methylation sites via a pseudo amino acid composition approach. *Biomedical Research International, 2014*, 947416.

Schenone, M., Dančík, V., Wagner, B. K., & Clemons, P. A. (2013). Target identification and mechanism of action in chemical biology and drug discovery. *Nature Chemical Biology, 9*, 232–240.

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. (2010). Virus-mPLoc: A fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. *Journal of Bimoleculor Structural Dynamics, 28*, 175–186.

Shen, H. B., Yang, J., & Chou, K. C. (2007). Euk-PLoc: An ensemble classifier for large-scale eukaryotic protein subcellular location prediction. *Amino Acids, 33*, 57–67.

Sirois, S., Hatzakis, G. E., & Wei, D. Q. (2005). Assessment of chemical libraries for their druggability. *Computational Biology and Chemistry, 29*, 55–67.

Sirois, S., Wei, D. Q., & Du, Q. S. (2004). Virtual screening for SARS-CoV protease based on KZ7088 pharmacophore points. *Journal of Chemical Information and Computer Sciences, 44*, 1111–1122.

Steffen, A., Kogej, T., Tyrchan, C., & Engkvist, O. (2009). comparison of molecular fingerprint methods on the basis of biological profile data. *Journal of Chemical Information and Modeling, 49*, 338–347.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence, 23*, 687–719.

Wang, J. F., & Chou, K. C. (2011). Insights from modeling the 3D structure of New Delhi metallo-beta-lactamase and its binding interactions with antibiotic drugs. *PLoS ONE, 6*, e18414.

Wang, J. F., & Chou, K. C. (2012). Insights into the mutation–induced HHH syndrome from modeling human mitochondrial ornithine transporter–1. *PLoS ONE, 7*, e31048.

Wang, S. Q., Du, Q. S., & Chou, K. C. (2007b). Study of drug resistance of chicken influenza A virus (H5N1) from homology-modeled 3D structures of neuraminidases. *Biochemical and Biophysical Research Communications, 354*, 634–640.

Wang, S. Q., Du, Q. S., Huang, R. B., & Zhang, D. W. (2009b). Insights from investigating the interaction of oseltamivir (Tamiflu) with neuraminidase of the 2009 H1N1 swine flu virus. *Biochemical and Biophysical Research Communications, 386*, 432–436.

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

Wang, J. F., Wei, D. Q., Li, L., Zheng, S. Y. (2007a). 3D structure modeling of cytochrome P450 2C19 and its implication for personalized drug design. *Biochemical and Biophysical Research Communications (Corrigendum: ibid, 2007, Vol. 357, 330),* 355, 513–519.

Wei, D. Q., Du, Q. S., & Sun, H. (2006). Insights from modeling the 3D structure of H5N1 influenza virus neuraminidase and its binding interactions with ligands. *Biochemical and Biophysical Research Communications, 344*, 1048–1055.

Wei, H., Wang, C. H., Du, Q. S., & Meng, J. (2009). Investigation into adamantane-based M2 inhibitors with FB-QSAR. *Medicinal Chemistry, 5*, 305–317.

Willett, P. (2006). Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today, 11*, 1046–1053.

Xiao, X., Min, J. L., & Wang, P. (2013a). iCDI-PseFpt: Identify the channel-drug interaction in cellular networking with PseAAC and molecular fingerprints. *Journal of Theoretical Biology, 337C*, 71–79.

Xiao, X., Min, J. L., & Wang, P. (2013b). iGPCR–Drug: A web server for predicting interaction between GPCRs and drugs in cellular networking. *PLoS ONE, 8*, e72234.

Xiao, X., Wang, P., Lin, W. Z., & Jia, J. H. (2013c). iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Analytical Biochemistry, 436*, 168–177.

Xiao, X., Wu, Z. C., & Chou, K. C. (2011). 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.

Xu, Y., Shao, X. J., & Wu, L. Y. (2013). iSNO-AAPair: Incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. *PeerJ, 1*, e171.

Xu, Y., Wen, X., Shao, X. J., & Deng, N. Y. (2014b). iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *International Journal of Molecular Sciences, 15*, 7594–7610.

Xu, Y., Wen, X., Wen, L. S., & Wu, L. Y. (2014c). iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. *PLoS ONE, 9*, e105018.

Xu, R., Zhou, J., Liu, B., He, Y. A., & Zou, Q. (2014a). Identification of DNA-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-*n*-gram approach. *Journal of Biomolecular Structure & Dynamics*. doi:10.1080/07391102.2014.968624

Yap, C. W. (2011). PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry, 32*, 1466–1474.

Zhang, C. T., & Chou, K. C. (1992). Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition. *Biophysical Journal, 63*, 1523–1529.

Zhang, C. T., & Chou, K. C. (1995). An analysis of protein folding type prediction by seed-propagated sampling and jackknife test. *Journal of Protein Chemistry, 14*, 583–593.

Zhong, W. Z., & Zhou, S. F. (2014). Molecular science for drug development and biomedicine. *International Journal of Molecular Sciences, 15*, 20072–20078.