GPCR-drug Interactions Prediction Using Random Forest with Drug-Association-Matrix-Based Post-Processing Procedure

Jun Hu[1], Yang Li[1], Jing-Yu Yang[1], Hong-Bin Shen[2], and Dong-Jun Yu[1, 3,*]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei 200, Nanjing, China, 210094

[2] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Dongchuan Road 800, Shanghai, China, 200240

[3] Changshu Institute, Nanjing University of Science and Technology, Changshu, China, 215513

[*] Address correspondence to D. J. Yu at njyudj@njust.edu.cn

Tel: +86-025-84316190

Fax: +86-025-84315960

- *J. Hu, , Y. Li, J. Y. Yang, and D. J. Yu are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei Road, Nanjing, 210094, China. phone: 086-025-84316190; fax: 086-025-84315960; E-mail: njyudj@njust.edu.cn*
- *D. J. Yu is with the Changshu Institute, Nanjing University of Science and Technology, Research Road 5, Changshu, 215513, China.*
- *H. B. Shen is with Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai, 200240, China. phone: 086-021-34205320; fax: 086-021-34204022; E-mail: hbshen@sjtu.edu.cn*

ABSTRACT

G-protein-coupled receptors (GPCRs) are important targets of modern medicinal drugs. The accurate identification of interactions between GPCRs and drugs is of significant importance for both protein function annotations and drug discovery. In this paper, a new sequence-based predictor called TargetGDrug is designed and implemented for predicting GPCR-drug interactions. In TargetGDrug, the evolutionary feature of GPCR sequence and the wavelet-based molecular fingerprint feature of drug are integrated to form the combined feature of a GPCR-drug pair; then, the combined feature is fed to a trained random forest (RF) classifier to perform initial prediction; finally, a novel drug-association-matrix-based post-processing procedure is applied to reduce potential false positive or false negative of the initial prediction. Experimental results on benchmark datasets demonstrate the efficacy of the proposed method, and an improvement of 15% in the Matthews correlation coefficient (*MCC*) was observed over independent validation tests when compared with the most recently released sequence-based GPCR-drug interactions predictor. The implemented webserver, together with the datasets used in this study, is freely available for academic use at http://csbio.njust.edu.cn/bioinf/TargetGDrug.

Keywords: GPCR-drug interactions; random forest; drug association matrix; machine learning

1. Introduction

Protein-ligand interactions are indispensable for biological activities and play important roles in virtually all biological processes (Alberts 2008; Kokubo, Tanaka et al. 2011; Gao and Skolnick 2012). The accurate identification of protein-ligand interactions is of significant importance for both protein function analysis and drug design (Schmidtke and Barril 2010; Yang, Roy et al. 2013). Much effort has been made to reveal the intrinsic mechanism of protein-ligand interactions (Roy and Zhang 2012; Nagarajan, Ahmad et al. 2013), and many effective methods (Glaser, Morris et al. 2006; Le Guilloux, Schmidtke et al. 2009; Yu, Hu et al. 2013) have been developed during the past decades.

Membrane proteins are proteins that interact with biological membranes. Membrane proteins are encoded by 20-35% of genes and fulfill many essential functions in the eukaryotic cell (Krogh, Larsson et al. 2001). Previous studies have demonstrated that membrane proteins are implicated in many diseases because they are positioned at the apex of signaling pathways that regulate cellular processes (Hopkins and Groom 2002; Kunji, Chan et al. 2005). G-protein-coupled receptors (GPCRs) are plasma membrane proteins that feature a common topology consisting of seven-transmembrane helices. As the largest family of cell surface receptors (Zhang and Zhang 2010), GPCRs are involved in many diseases, such as cancer, diabetes, neurodegenerative, inflammatory and respiratory disorders, and are pharmacologically important target proteins for new drug development (Roth, Willins et al. 1998; Chou 2005). In fact, GPCRs have been the targets of approximately 40% of all modern medicinal drugs (Chou 2005; Xiao, Min et al. 2013). Hence, accurately identifying GPCR-drug interactions is of significant importance for modern drug design and development (Knowles and Gromo 2003).

Considerable effort has been made to predict protein-drug interactions, and many promising methods have been developed. Roughly speaking, according to the features they used, these methods can be divided into structure-based and sequence-based methods. The structure-based methods, such as docking simulations (Chou, Wei et al. 2003; Cheng, Coleman et al. 2007) and literature text mining (Zhu, Okuno et al. 2005), are structure-based and require protein tertiary structures as inputs to predict protein-drug interactions. The sequence-based methods, such as iGPCR-Drug (Xiao, Min et al. 2013), are computational methods for predicting interactions between GPCRs and drugs only based on the corresponding sequence-based features.

Among the above-mentioned two methods, the structure-based methods directly predict the GPCR-drug interactions with ground-truth 3D structures or predicted structure (Yu, Li et al.), they can provide deeper insight into the underlying mechanism of interactions if compared with non-structural prediction methods (Yamanishi, Araki et al. 2008). However, the intrinsic hydrophobicity of GPCRs makes resolving their tertiary structures a time-consuming and costly work with regular wet-lab experimental methods. During the past decades, much progress has been

made in resolving crystal structures of GPCRs (Granier and Kobilka 2012; Tate 2012). In addition, as there are lots of highly conserved residues in GPCR sequence (Tate 2012), many homologous techniques (Chou 2005; Eswar, Webb et al. 2006; Worth, Kreuchwig et al. 2011) can help acquire the 3D modelling of GPCR by using various bioinformatics tools (Chou 2004), such as BLAST (Schaffer, Aravind et al. 2001). Despite of the big progress made in resolving 3D structure and 3D modelling of GPCR, currently, we still have only modest and patchy coverage of GPCR structural space. Together with the fact that the receptors (GPCRs) are rather flexible, thus resolving and/or modelling the 3D structure of a GPCR is still a very challenging task.

Hence, developing a computational sequence-based method that can predict the interactions between drugs and GPCRs solely from primary GPCR sequences will be especially useful. The sequence-based methods, such as the combination of chemical structure and genomic sequence information (Yamanishi, Araki et al. 2008; He, Zhang et al. 2010; Xiao, Min et al. 2013), are popular for predicting GPCR-drug interaction recently. There are many tools for extracting the useful sequence information, such as *PseAAC* (Shen and Chou 2008) (available at http://www.csbio.sjtu.edu.cn/bioinf/PseAAC/), are developed. Additionally, the chemical fingerprint information of drug is also employed by sequence-based methods for GPCR-drug interactions prediction. The most advantages of the sequence-based methods are that the prediction performance is appreciable and the prediction speed is very fast without protein structure information.

He et al. (He, Zhang et al. 2010) and Yamanishi et al. (Yamanishi, Kotera et al. 2010) carried out a series of studies on the prediction of GPCR-drug interactions. Their results demonstrated the feasibility of performing GPCR-drug interactions prediction from GPCR sequences. Recently, Xiao et al. (Xiao, Min et al. 2013) developed a predictor, called iGPCR-Drug, which is a sequence-based predictor specifically designed for predicting GPCR-drug interactions. The iGPCR-Drug performs GPCR-drug interactions prediction by utilizing the combined features extracted from a GPCR-drug pair and a fuzzy K-Nearest Neighbor (FKNN) classification algorithm (Keller, Gray et al. 1985). More specifically, for a query GPCR-drug pair, iGPCR-Drug first extracts the pseudo amino acid composition feature (*PseAAC*) (Chou 2001) based on the grey model theory and the Fourier-transformed 2D molecular fingerprint (*FT2DMFP*) (O'Boyle, Banck et al. 2011) for the GPCR sequence and the drug contained in the query pair, respectively. The two features are then combined to form the discriminative feature of the query GPCR-drug pair. Finally, a fuzzy K-Nearest Neighbor (FKNN) classification algorithm (Keller, Gray et al. 1985) is utilized to determine whether the query GPCR-drug pair is interactive.

Although iGPCR-Drug achieved great success in predicting GPCR-drug interactions, we were motivated to further analyze the algorithm for three reasons. First, to the best of our knowledge, the

evolutionary information contained in GPCRs, especially, highly conserved residues of GPCRs, has not been well investigated in GPCR-drug interactions prediction; second, the relationship between one drug with each other, which is one of the most important information for predicting the interactions between GPCRs with drugs, has been ignored by iGPCR-Drug; third, although iGPCR-Drug has achieved satisfactory prediction performances, there is still room for further improvement.

In view of these three important issues, we developed a new sequence-based method for prediction of GPCR-drug interactions with high performance. In the proposed method, the 140-D pseudo-position-specific scoring matrix (*PsePSSM*) feature that encodes the evolutionary information of the GPCR sequence (Chou and Shen 2007) and the 128-D wavelet-based molecular fingerprint (*WaveletMFP*) feature of the drug are combined to form the discriminative feature of a query GPCR-drug pair. The combined feature is then fed to a trained random forest (RF) (Breiman 2001) classifier to perform the initial prediction. Finally, a novel drug-association-matrix-based post-processing procedure, which can extract the relationship information between one drug with each other, is applied to reduce potential false positive or false negative resulting from the initial prediction.

Cross-validation tests and independent validation tests on different benchmark datasets demonstrate the efficacy of the proposed method. An improvement of 15% in terms of the Matthews correlation coefficient (*MCC*) was observed over independent validation tests when compared with the most recently released sequence-based GPCR-drug interactions predictor, i.e., iGPCR-Drug (Xiao, Min et al. 2013), demonstrating the better generalization capability of the proposed method.

2. Materials and Methods

2.1 Benchmark Datasets

A. Cross-validation dataset

To objectively evaluate the proposed method and fairly compare it with existing sequence-based GPCR-drug interactions prediction methods, two datasets that have been used in previous studies were chosen as cross-validation benchmark datasets.

Each of the two benchmark datasets can be formulated as

$$V = V^+ \cup V^- \tag{1}$$

where V^+ represents the positive subset that consists only of the interactive GPCR-drug pairs, while V^- is the negative subset that consists of only the non-interactive GPCR-drug pairs, and the symbol $\cup$ is the union in the set theory.

The first dataset, which consists of 314 interactive GPCR-drug pairs (V^+)

(http://cbio.ensmp.fr/~yyamanishi/pharmaco/), was constructed by Yamanishi et al. (Yamanishi, Araki et al. 2008; Yamanishi, Kotera et al. 2010). We artificially synthesized 628 non-interactive pairs (V^-) with the same procedure described in (He, Zhang et al. 2010). The obtained dataset $V = V^+ \cup V^-$ contains 84 distinct GPCR sequences and 105 distinct drugs and is denoted as D84.

The second benchmark dataset, which was a larger dataset collected by He et al. (He, Zhang et al. 2010), contains 1860 GPCR-drug pairs, among which 620 and 1240 GPCR-drug pairs constitute the positive subset V^+ and negative subset V^-, respectively. The 620 pairs in V^+, which have been experimentally verified as "interactive", were collected from the KEGG database (Kanehisa, Goto et al. 2006) at the time of writing their paper. The 1240 "non-interactive" GPCR-drug pairs in V^- were artificially synthesized as follows: (a) each of the pairs in V^+ was separated into a single GPCR and drug; (b) each of the single GPCRs was re-coupled with each of the single drugs into synthesized pairs in such a way that none of the synthesized pairs appeared in V^+; and (c) the synthesized pairs were randomly selected until the number of selected pairs was twice the number of pairs in V^+. The final dataset $V = V^+ \cup V^-$ contains 92 distinct GPCR sequences and 217 distinct drugs and is denoted as D92 for the convenience of subsequent description. However, as increasing GPCR-drug pairs are discovered and updated by the broad availability of GPCR assays and widespread drug repositioning efforts, we found that 15 GPCR-drug pairs in the positive subset of D92 are falsely labeled as "non-interactive" by checking the recently released KEGG database (Kanehisa, Goto et al. 2006). As we known, less mistaken information contained in the training dataset may lead to that a better predictor will be developed (Quince, Lanzen et al. 2011). Therefore, we refined D92 by moving the 15 falsely labeled GPCR-drug pairs from the negative subset to the positive subset of D92 to enhance the effectiveness of our proposed method. The obtained dataset is denoted as D92M, which consists of 635 "interactive" and 1225 "non-interactive" pairs.

Note that D84 is an early constructed dataset which consists of 314 positive pairs and 628 negative pairs, while D92M is a much larger dataset built after the construction date of D84. In fact, D92M include 313 out of the 314 positive pairs in D84 and 322 newly annotated positive pairs. There are only 42 negative pairs shared by D84 and D92M. We used D84 because we need to compare the proposed method with several existing GPCR-drug interaction predictors that have been evaluated on D84.

B. Independent validation dataset

Evaluating a newly developed predictor by comparing it with existing predictors using the same datasets may potentially lead to optimistically biased results in the sense that the new predictor's characteristics over-fit the used datasets (Boulesteix 2010). It has become a routine procedure to evaluate the generalization capability of a predictor with an independent validation test. We therefore constructed an independent-validation dataset. The positive subset of the independent

validation dataset is constructed as follows:

Step 1: All protein-drug pairs that interact with any one of the 217 distinct drugs in D92M are extracted from the KEGG database (Kanehisa, Goto et al. 2006) (http://www.genome.jp/kegg/). At the time of writing, the number of extracted protein-drug pairs is 904.

Step 2: All pairs, among which the protein components are not GPCRs, are removed from the extracted protein-drug pairs obtained in Step 1; whether a protein is a GPCR can be determined by searching the protein code against the GPCRDB database available at http://www.gpcr.org/7tm/.

Step 3: All pairs that appeared in the positive subset of D92M are removed from the GPCR-drug pairs obtained in Step 2, and the remaining 130 interactive pairs constitute the positive subset.

Furthermore, we artificially synthesized 260 non-interactive pairs to constitute the negative subset of the independent validation dataset with the same procedure used to generate the negative subset of D92M. To guarantee the independence of the validation dataset, none of the 260 non-interactive pairs appears in the negative subsets of D84 and D92M. The final independent validation dataset, denoted as Check390, consists of the 130 interactive pairs and 260 non-interactive pairs obtained above.

Table 1 summarizes the detailed data statistics of the three benchmark datasets. All the data listed in Table 1 can be found in Supporting Information S1.

Table 1. Statistics of the three benchmark datasets used for evaluating the performance of GPCR-drug interactions predictions.

Dataset	No. of Distinct GPCRs	No. of Distinct Drugs	No. of Positive Pairs	No. of Negative Pairs	Total No. of Pairs
D84	84	105	314	628	942
D92M	92	217	635	1225	1860
Check390	38	171	130	260	390

2.2 Feature Representation

One of the critical steps in designing a machine-learning based predictor is to effectively transform the targets into fixed-length feature vectors. In this study, for each GPCR-drug pair, we extracted the pseudo position specific scoring matrix (*PsePSSM*) feature and the wavelet-based molecular fingerprint (*WaveletMFP*) feature for the GPCR sequence and the drug, respectively; the feature vector of a GPCR-drug pair is then obtained by combining its corresponding *PsePSSM* and *WaveletMFP* features.

2.2.1 Pseudo Position Specific Scoring Matrix Feature of GPCR

The pseudo position specific scoring matrix (*PsePSSM*) feature encodes both evolutionary and sequence-order information of a protein sequence (Chou and Shen 2007) and has been widely used

in related protein attribute prediction problems (Yu, Wu et al. 2012; Zia-Ur-Rehman and Khan 2012; Yu, Hu et al. 2013; Nanni, Brahnam et al. 2014). We extracted *PsePSSM* feature of a GPCR sequence as follows:

For a GPCR sequence $\mathbf{P}$ with L amino acid residues, we obtained its $L\times 20$ position specific scoring matrix (*PSSM*), denoted as $\mathbf{P}_{PSSM}$, using PSI-BLAST (Schaffer 2001) with default parameter settings as follows:

$$\mathbf{P}_{PSSM}=\left[x_{i,j}\right]_{i=1,2,3,\cdots L;\,j=1,2,3,\cdots,20} \tag{2}$$

where $x_{i,j}$ represents the score of the amino acid residue i in the GPCR sequence being mutated to amino acid type j during the evolution process. Note that here we use the numerical code 1, 2,$\cdots$, 20 to represent the 20 native amino acid types according to the alphabetical order of their single-character codes. Then, each element x contained in $\mathbf{P}_{PSSM}$ is normalized by the logistic function $f(x)=1/\left(1+\mathrm{e}^{-x}\right)$ (i.e., $x_{i,j}^{Normalized}=1/\left(1+\mathrm{e}^{-x_{i,j}}\right)$) and the normalized *PSSM* is obtained as follows:

$$\mathbf{P}_{PSSM}^{Normalized}=\left[x_{i,j}^{Normalized}\right]_{i=1,2,3,\cdots L;\,j=1,2,3,\cdots,20} \tag{3}$$

Next, we extract fixed-length *PsePSSM* feature of a GPCR sequence from its corresponding $\mathbf{P}_{PSSM}^{Normalized}$ as follows:

First, the 20-D *PSSM* composition feature, denoted as $\mathbf{F}_{\overline{PSSM}}$, is calculated from $\mathbf{P}_{PSSM}^{Normalized}$ as follows:

$$\mathbf{F}_{\overline{PSSM}}=\left(\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_{20}\right)^T \tag{4}$$

where

$$\overline{x}_j=\frac{1}{L}\sum\nolimits_{i=1}^{L}x_{i,j}^{Normalized} \tag{5}$$

In $\mathbf{F}_{\overline{PSSM}}$, $\overline{x}_j$ represents the average score of the amino acid residues in a GPCR sequence being mutated to amino acid type j during the evolution process.

Then, as Fig. 1 shown, we extract the sequence-order information contained in $\mathbf{P}_{PSSM}^{Normalized}$ by calculating the correlation factor of each column of a $\mathbf{P}_{PSSM}^{Normalized}$ as follows:

$$\boldsymbol{\theta}^{\xi}=\left(\theta_1^{\xi},\theta_2^{\xi},\cdots,\theta_{20}^{\xi}\right)^T \tag{6}$$

where $\theta_j^{\xi}=\frac{1}{L-\xi}\sum\nolimits_{t=1}^{L-\xi}\left(x_{t,j}^{Normalized}-x_{t+\xi,j}^{Normalized}\right)^2$, $1\le j\le 20$, $0\le\xi\le G$, and $G<L$. The scalar quantity θ_j^{ξ} is the correlation factor obtained by coupling the ξ-most contiguous scores in $\mathbf{P}_{PSSM}^{Normalized}$ along the GPCR sequence for the amino acid type j.

Finally, the *PsePSSM* feature is obtained by combining $\mathbf{F}_{\overline{PSSM}}$ and $\boldsymbol{\theta}_{\xi}$ as follows:

$$\mathbf{f}_{PsePSSM}^{\xi}=\begin{pmatrix}\mathbf{F}_{\overline{PSSM}}\\ \boldsymbol{\theta}^{\xi}\end{pmatrix}=\left(\overline{x}_1,\overline{x}_2,\cdots,\overline{x}_{20},\theta_1^{\xi},\theta_2^{\xi},\cdots,\theta_{20}^{\xi}\right)^T \tag{7}$$

Note that $\mathbf{f}_{PsePSSM}^{\xi}$ degenerates to $\mathbf{F}_{\overline{PSSM}}$ when $\xi=0$.

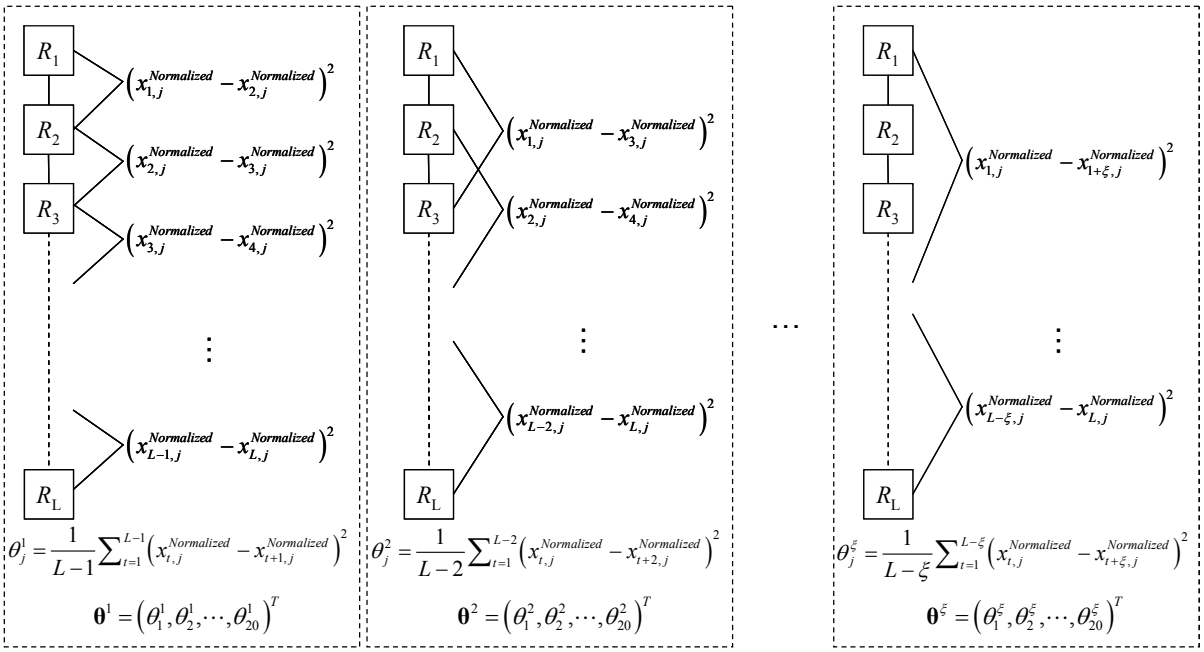


Fig. 1. The extraction schematic of the sequence-order information contained in $\mathbf{P}_{PSSM}^{Normalized}$, where R_1 represents the residue at sequence position 1, R_2 the residue at position 2, and so forth.

According to the original definition of *PsePSSM* given by Chou et al. (Chou and Shen 2007), a protein sequence can be represented by a set of *PsePSSM* feature vectors, denoted as $\{\mathbf{f}_{PsePSSM}^{\xi}\}_{\xi=0}^{G}$, with different values of ξ. However, two disadvantages exist if the feature information represented by $\{\mathbf{f}_{PsePSSM}^{\xi}\}_{\xi=0}^{G}$ is directly utilized (Yu, Hu et al. 2013): first, the combined feature vector will contain redundant information, as each PsePSSM encodes a $\mathbf{F}_{\overline{PSSM}}$; second, the dimensionality of the combined feature vector is very high. To avoid these two disadvantages, a more compact *PsePSSM* feature vector was proposed as follows (Yu, Hu et al. 2013):

$$\mathbf{f}_{PsePSSM}^{G}=\begin{pmatrix}\mathbf{F}_{\overline{PSSM}}\\ \boldsymbol{\theta}^1\\ \vdots\\ \boldsymbol{\theta}^G\end{pmatrix} \tag{8}$$

Accordingly, the dimensionality of the compact *PsePSSM* feature vector represented by Eq. (8) is $20+G\cdot 20$. Because there is no theoretical justification for determining the optimal value of G, we empirically set the value of G to be 6 in this study.

2.2.2 Wavelet-Based Molecular Fingerprint Feature of Drug

Molecular fingerprints are bit-string representations of molecular structures and properties (Eckert and Bajorath 2007). Previous studies have demonstrated the efficacy of molecular fingerprints for the prediction of drug-target interactions. In this study, we also use molecular fingerprints to represent drug molecules, as reported by Xiao et al. (Xiao, Min et al. 2013). The molecular fingerprint of a drug can be extracted as follows:

For a given drug, its MOL file, which contains the detailed chemical structural information of

the drug, is obtained from the KEGG database (http://www.kegg.jp/kegg/) using its drug code; then, the obtained MOL file is converted into a 2D molecular fingerprint file format using OpenBabel (O'Boyle, Banck et al. 2011) software (available at http://openbabel.org/). OpenBabel can generate four types of fingerprints: FP2, FP3, FP4 and MACCS. In this study, the FP2 fingerprint format is used. The FP2 fingerprint encodes each drug into a 256-bit hexadecimal string, among which each component is an integer between 0 and 15.

As we know, the original molecular fingerprint feature may contain some noise and redundant information which may incur negative impacts to the performance of a prediction model. To remove these useless information and dig out the most discriminative feature from a drug's molecular fingerprint (the experimental support for this claim could be found in section 3.2), we perform single-level discrete 1-D wavelet transform (Villasenor, Belzer et al. 1995) on the original 256-bit hexadecimal string. For the convenience of subsequent descriptions, a 256-bit hexadecimal string is considered as a discrete signal, denoted as $s(t)$.

The wavelet transform of a signal $s(t)$ is defined as the summation of the full time frame of the signal $s(t)$ multiplied by a scaled and shifted version of the mother wavelet function $\psi(t)$ (Qiu, Huang et al. 2009). The coefficient $C_s(a,b)$ of the wavelet transform of the signal $s(t)$ is calculated by the formula defined as follows:

$$C_s(a,b)=\frac{1}{\sqrt{a}}\int_0^x s(t)\overline{\psi(\frac{t-b}{a})}dt, \quad a\in\mathbb{R}^+, b\in\mathbb{R} \tag{9}$$

where a and b represent the scaling and shifting values, respectively, x is the length of the signal, and $\overline{\psi(\cdot)}$ is the conjugate function of the wavelet function. The discrete wavelet transform (DWT) has the ability to decompose the signal $s(t)$, i.e., the original drug molecular fingerprint feature in this study, into coefficients at different dilations, and discard the noise contained in the original signal. The DWT can be economically implemented by computing on a dyadic grid of points (Mallat 1989; Qiu, Sun et al. 2010). In this study, we set a=2 and b=1 for DWT implementation. Consequently, there is a binary dilation of 2^{-m} and a dyadic translation of $n2^{-m}$, which can be formulated as follows:

$$\boldsymbol{\psi}_{m,n}(t)=2^{-m/2}\psi(2^{-m}t-n) \tag{10}$$

where m = 1, 2, ⋯, and n = 0, 1, 2, ⋯. The coefficient $C_s(a,b)$ of the signal $s(t)$ can then be obtained by the following formula:

$$C_s(a,b)=\left\langle s(t),\boldsymbol{\psi}_{a,b}(t)\right\rangle=2^{-m/2}\int_0^x s(t)\overline{\psi(2^{-m}t-n)}dt \tag{11}$$

Furthermore, the coefficient $C_s(a,b)$ can be divided into two parts: one is the set of approximation coefficients, denoted as $\{A^{(j)}(n)\}$, representing the high-scale and low-frequency components of $s(t)$ at the j-th decomposition level; the other is the set of detail coefficients, denoted as $\{D^{(j)}(n)\}$, representing the low-scale and high-frequency components of $s(t)$ (Mallat

1989; Qiu, Huang et al. 2009). An efficient algorithm has been proposed by Mallat (Mallat 1989; Mallat 1989) to construct the approximation coefficients and detail coefficients of the DWT as follows:

$$A^{(j)}(n) = \sum_{k \in \mathbb{Z}} \overline{h_{k-2n}} A^{(j-1)}(k) \tag{12}$$

$$D^{(j)}(n) = \sum_{k \in \mathbb{Z}} \overline{g_{k-2n}} D^{(j-1)}(k) \tag{13}$$

where h and g are the high pass and low pass filters, respectively.

In this study, the Haar mother wavelet proposed by Haar (Haar 1910) is utilized as $\psi(t)$ to perform the DWT on the drug molecular fingerprint feature, and the decomposition level j is set to 1 according to our preliminary local tests. Under these configurations, the set of approximation coefficients, denoted as $\{A^{(1)}(n)\}$, is considered as useful information, and the set of detail coefficients, $\{D^{(1)}(n)\}$, is regarded as useless noise. Finally, we obtain the 128-D wavelet-based molecular fingerprint feature, denoted as $\mathbf{f}_{WaveletMFP}$, formed by concatenating the 128 approximation coefficients contained in $\{A^{(1)}(n)\}$.

2.2.3 Combining *PsePSSM* and *WaveletMFP* features

The final feature vector for a GPCR-drug pair is the weighted combination of the *PsePSSM* feature and the *WaveletMFP* feature of GPCR and drug, respectively, contained in the pair:

$$\mathbf{f} = \begin{pmatrix} \mathbf{f}^{G}_{PsePSSM} \\ \lambda \cdot \mathbf{f}_{WaveletMFP} \end{pmatrix} \tag{14}$$

where λ is a weight value. In this study, the value of λ is set to be 0.004 after an optimization procedure performed on the training dataset over cross-validation.

2.3 Classifier Selection

For a given prediction task, the performance will depend not only on the feature used but also on the classifier selection. In this study, four popular classifiers (i.e., optimized evidence theoretic *K* nearest neighbor algorithm (OET-KNN) (Zouhal and Denoeux 1998; Shen and Chou 2007), QuickRBF (Ou 2005; Chen, Ou et al. 2011), support vector machine (SVM) (Vapnik 1998; Fan, Chen et al. 2005; Wong, Hardy et al. 2013), and random forest (RF) (Breiman 2001; Kandaswamy, Pugalenthi et al. 2010)) that have been widely used in bioinformatics fields are chosen for consideration.

According to our comparison study (refer to section 3), random forest (RF) achieves the best prediction performance with the feature developed in the above section using the benchmark datasets. In view of this, we will take RF as the classifier engine in implementation of the web server of the proposed method.

The random forest algorithm (Breiman 2001; Liaw and Wiener 2002) is an ensemble learning

method based on decision trees that adds an additional layer of randomness to bagging. Decision trees in a RF are gradually developed (trained) using a random selection of inputs and random feature selection strategy (Hamby and Hirst 2008). In a trained RF, the decision trees then vote on the class for a given input.

The randomness strategy used in RF has been demonstrated to be very effective compared with many other classifiers and is robust against over-fitting (Breiman 2001; Liaw and Wiener 2002). RF can be used to perform both classification and regression. In this study, a random forest classification model is used. We utilized the RF code, which is freely accessible at https://code.google.com/p/randomforest-matlab/, to evaluate and implement the proposed method. Two parameters, i.e., the number of trees to grow (*nTree*) and the number of dimensions randomly sampled as candidates at each split (*mTry*), were set to be 100 and 25, respectively.

For a query GPCR-drug pair, the predicted probability of being "interactive" by a trained RF can be obtained as follows:

Let *nP* be the number of trees in the trained RF that predict a query GPCR-drug pair to be "interactive"; then the predicted probability of being "interactive" for the query GPCR-drug pair is (*nP/ nTree*) × 100%.

2.4 Post-processing Procedure

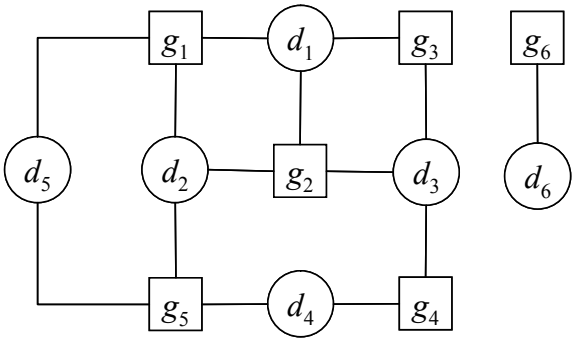
To further reduce predicted false positives and false negatives, we propose a post-processing procedure based on the drug association matrix (DAM). Details of the post-processing procedure are described as follows:

A. Drug Association Matrix

Let V^+ and V^- be the positive and negative subsets consisting of interactive and non-interactive GPCR-drug pairs, respectively. Let $D=\{d_i\}_{i=1}^{N_d}$ and $G=\{g_i\}_{i=1}^{N_p}$ be the set of distinct drugs and GPCR sequences that appear in the positive subset V^+, where N_d and N_p are the number of distinct drugs and GPCR sequences, respectively. The drug association matrix (DAM) is a $N_d \times N_d$ symmetric matrix, among which the value of each element, denoted as $DAM(i,j)$, measures the association degree between drug d_i and d_j. For the convenience of later description, we term $D=\{d_i\}_{i=1}^{N_d}$ as the *association drug set* of a DAM.

For any drug d_i and d_j ($i \neq j$) in D, the knowledge that drug d_i and d_j interact with the same GPCR protein is evidence that increases the likelihood that the two drugs have some type of association. Based on this hypothesis, the DAM for a given training dataset can be calculated as follows: first, the DAM as a $N_d \times N_d$ zero matrix is initialized; then, the value of $DAM(i,j)$ is set to be the number of distinct GPCR sequences in the positive subset V^+ that interact with both drugs d_i and d_j ($i \neq j$); similarly, for each element $DAM(i,i)$ in the diagonal of the DAM, its

value is set to be the number of distinct GPCR proteins that interact with drug d_i found in V^+.



(a)

DAM	1	2	3	4	5	6
1	3	2	2	0	1	0
2	2	3	1	1	2	0
3	2	1	3	1	0	0
4	0	1	1	2	1	0
5	1	2	0	1	2	0
6	0	0	0	0	0	1

(b)

Fig. 2. (a) The synthetic interactive GPCR-drug pairs example. (b) The corresponding DAM result.

Fig. 2 shows the DAM result of a synthetic example. Furthermore, we calculated the DAM on the D92M dataset. Because there are 217 distinct drugs in the D92M dataset, the obtained DAM is a 217×217 symmetric matrix. Fig. 3 plots the DAM as a gray image by transforming the value of each element in DAM to the range of [0, 1].

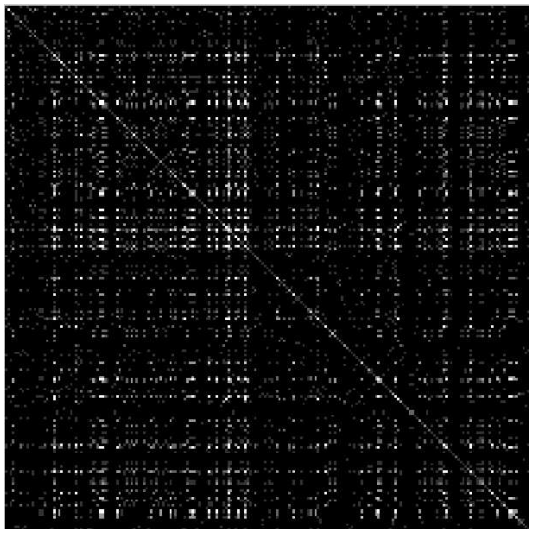


Fig. 3. Image representation of the DAM obtained for the D92M dataset.

We believe that the proposed DAM reflects the knowledge of association degrees between drugs buried in a training dataset, and can be utilized as prior knowledge for post-processing to further improve the performance of GPCR-drug interactions prediction, which will be demonstrated in the subsequent experimental section.

B. Post-processing Procedure: Reducing Possible False Positives and False Negatives

Based on the proposed DAM, we developed a post-processing procedure (PPP) that can

potentially reduce predicted false positives and false negatives.

Let $RFmodel(\mathbf{f}, P_{RFmodel})$ be the trained RF prediction model, where $\mathbf{f}$ is the input feature vector (i.e., the combination of *PsePSSM* and *WaveletMFP* features in this study) for a query GPCR-drug pair, and $P_{RFModel}$ is the set of optimized model parameters. $RFmodel(\mathbf{f}, P_{RFmodel})$ outputs a real scalar value, denoted as p, representing the possibility of being interactive for the query pair. During the initial prediction, a GPCR-drug pair is marked as interactive when its predicted possibility p is larger than the prescribed threshold T.

However, a query pair may potentially be predicted as false positive or false negative if directly applying the above-mentioned prediction model $RFmodel(\mathbf{f}, P_{RFmodel})$. Considering this, we developed the post-processing procedure, which aims to reduce false positive or false negative for a query pair predicted by the initial prediction model $RFmodel(\mathbf{f}, P_{RFmodel})$.

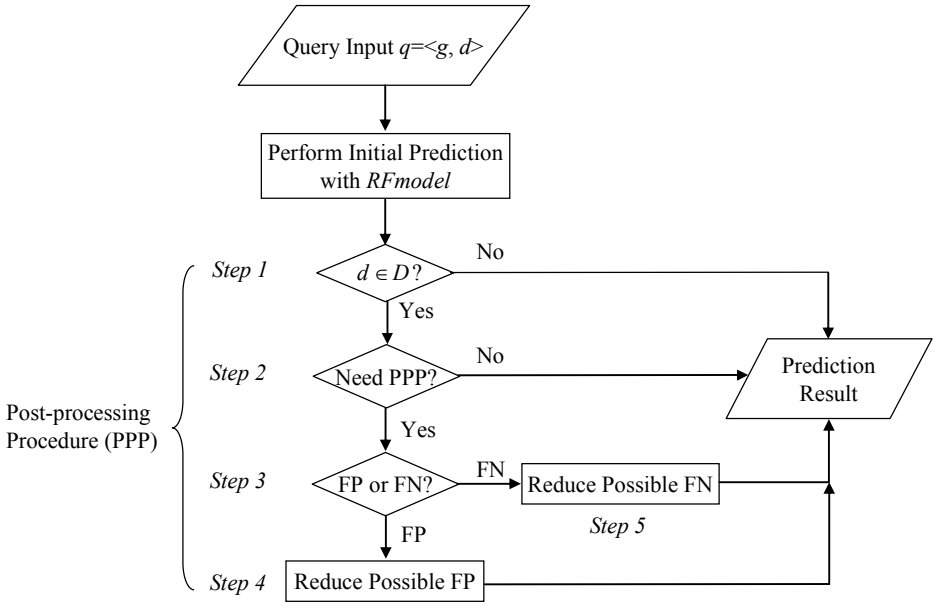


Fig. 4. Flowchart of the proposed post-processing procedure.

Fig. 4 shows the flowchart of the proposed post-processing procedure (PPP). Next, we describe the five steps of the proposed PPP to illustrate how the PPP works.

Let $q =< g, d >$ be a query GPCR-drug pair, $\mathbf{f}_q$ be the feature vector of the query pair, DAM be the drug association matrix, and $D = \{d_i\}_{i=1}^{N_d}$ be the *association drug set* of the DAM.

Step 1: Deciding whether the drug in a query pair exists in the association drug set

The query GPCR-drug pair are separated into a single GPCR sequence and drug, denoted as g and d, respectively.

If the drug d in the query pair exists in the association drug set of the DAM, go to *step 2*; otherwise, the decision can be directly made with the initial prediction model $RFmodel(\mathbf{f}, P_{RFmodel})$. In other words, the decision is made without using the subsequent post-processing procedure when the

drug d in the query pair does not exist in the association drug set. The underlying reason is that there is no corresponding knowledge of association degrees between the drug d and the other drugs in the association drug set that can be used for further post-processing.

Step 2: Deciding whether to perform PPP

First, each of the N_d drugs in the association set of DAM are recoupled with the separated GPCR sequence g in the query pair into pairs: $\{< g, d_i >\}_{i=1}^{N_d}$;

Second, feature vectors are extracted for all re-coupled GPCR-drug pairs:

$$\{\mathbf{f}_i\}_{i=1}^{N_d} \leftarrow \text{FeatureExtraction}(\{< g, d_i >\}_{i=1}^{N_d}) \tag{15}$$

Third, the possibilities of being interactive for all the re-coupled pairs are predicted using the trained prediction model $RFmodel(\mathbf{f}, P_{RFmodel})$:

$$\{P_i\}_{i=1}^{N_d} \leftarrow \text{Predict}(RFmodel(\mathbf{f}, P_{RFmodel}), \{\mathbf{f}_i\}_{i=1}^{N_d}) \tag{16}$$

Let B be the set of those drugs in the association drug set that are predicted as interactive with the GPCR protein g:

$$B = \{d_i \mid P_i \geq T, 1 \leq i \leq N_d\} \tag{17}$$

Then, B is considered as the query-driven knowledge base that can be further used to reduce potential false positive or false negative.

If B is empty (i.e., $B = \Phi$), then directly make the decision according to the relationship between P_q and T as described in *Step 1*, because there is no query-driven knowledge available for further post-processing; otherwise go to *Step 3*.

Step 3: Determining to reduce false positive or false negative

According to the non-empty query-driven knowledge base B, we determine to reduce possible false positive or false negative as follows:

If the drug d in the query pair $q =< g, d >$ belongs to B (i.e., $d \in B$), then the query pair has the possibility of being mistakenly predicted as false positive by the initial $RFmodel$. Hence go to step 4 to reduce possible false positive.

If the drug d does not belong to B, then the query pair has the possibility of being mistakenly predicted as false negative by the initial $RFmodel$. Hence go to step 5 to reduce possible false negative.

Note that before entering step 4 or 5, the score for each element, denoted as s_{d_i}, contained in B is computed by using following sub-routine:

FOR each $d_i \in B$
  $s_{d_i} = 0$
  FOR each $d_j \in B$
    $s_{d_i} = s_{d_i} + DAM(i, j)$
  END FOR
END FOR

The obtained score set is denoted as $S = \{s_{d_i} \mid d_i \in B\}$.

Note that the larger the score s_{d_i}, the higher the association degree between the drug d_i and the other drugs contained in the query-driven knowledge base B.

Step 4: Reducing possible false positive

Let s_d be the score of the drug d contained in the query pair $q =< g, d >$ and s^{max} be the maximum in the score set $S = \{s_{d_i} \mid d_i \in B\}$ obtained in *Step 3*.

If $s_d < \alpha \cdot s^{max}$, the query pair may possibly be mistakenly predicted as false positive by the initial $RFmodel$ and should be reassigned as negative ("non-interactive") as follows:

$$\text{label}(q) = \begin{cases} \text{non-interactive}, & \text{if } s_d < \alpha \cdot s^{max} \\ \text{interactive}, & \text{else} \end{cases} \tag{18}$$

where $\alpha \in (0,1)$ is a prescribed parameter for reducing possible false positive.

Step 5: Reducing possible false negative

Let $\tilde{B}$ be the set of drugs whose scores are the first m-maximal in S, where m is a prescribed parameter for reducing possible false negative.

Then, we reduce possible false negative as follows: for each $d_j \in \tilde{B}$, if the drug d in the query pair $q =< g, d >$ has a non-zero association degree with d_j, then the query pair may have the possibility of being mistakenly predicted as false negative by the initial $RFmodel$ and should be reassigned as positive ("interactive").

$$\text{label}(q) = \begin{cases} \text{interactive}, & \text{if } d \text{ has non-zero association degree with every drug in } \tilde{B} \\ \text{non-interactive}, & \text{else} \end{cases} \tag{19}$$

It has not escaped our notice that the proposed PPP may also potentially reassign a true positive as a false negative, and a true negative as a false positive. Nevertheless, experimental results in section 3 statistically demonstrate that the overall prediction performance can be further improved by incorporating the proposed PPP.

2.5 Architecture of the TargetGDrug

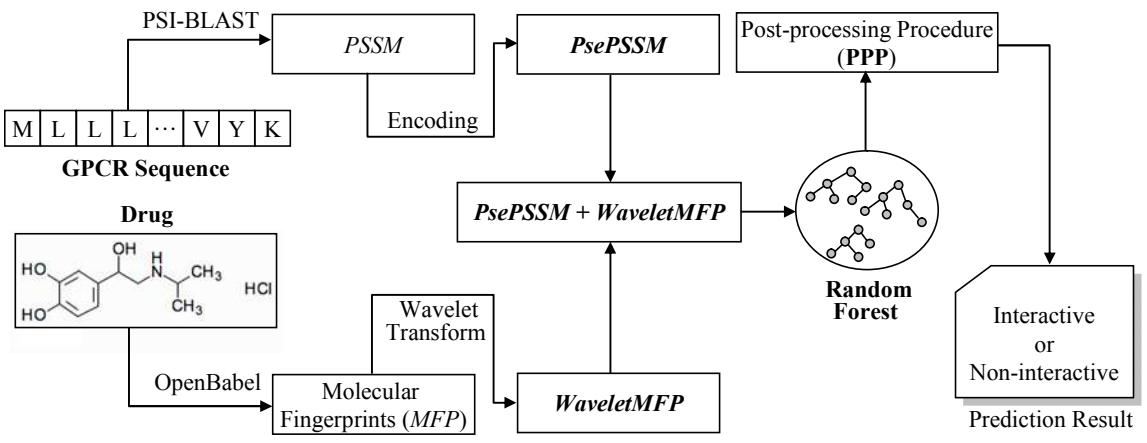


Fig. 5. Architecture of the proposed TargetGDrug.

Fig. 5 illustrates the architecture of the proposed method, called TargetGDrug, for prediction of GPCR-drug interactions. For a query GPCR-drug pair, TargetGDrug first extracts *PsePSSM* and *WaveletMFP* features of the GPCR and the drug contained in the query pair, respectively; then, the two extracted features are combined and fed to the trained random forest (RF) classifier to perform the initial prediction; subsequently, the output of RF will be further processed with the proposed post-processing procedure (PPP) to obtain the final prediction result.

2.6 Evaluation Indexes

In this study, the confusion matrix, also known as a contingency table or an error matrix (Stehman 1997), will be used to represent the prediction performance of a predictor. Fig. 6 illustrates an exemplary confusion matrix, where TP, FP, TN, and FN are the abbreviations for true positives, false positives, true negatives, and false negatives, respectively. For the convenience of subsequent descriptions, we will use [TP FN ; FP TN] to represent a confusion matrix:

		Predicted class: Positive	Predicted class: Negative
True class	Positive	TP (True Positives)	FN (False Negatives)
True class	Negative	FP (False Positives)	TN (True Negatives)

Fig. 6. Confusion matrix for performance evaluation.

From a confusion matrix, five routinely used evaluation indices in this field, i.e., *Specificity* (Sp), *Sensitivity* (Sn), *Accuracy* (Acc), *Strength* (Str), and the Matthews correlation coefficient (MCC), can be computed as follows:

$$Sp = \frac{TN}{TN + FP} \tag{20}$$

$$Sn = \frac{TP}{TP + FN} \tag{21}$$

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{22}$$

$$Str = \frac{Sp + Sn}{2} \tag{23}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{24}$$

For a soft-type classifier, i.e., a classifier that outputs a continuous numeric value to represent the confidence of a sample belonging to the predicted class, its confusion matrix will vary within the classification threshold chosen. Gradually adjusting the classification threshold will produce a series of confusion matrices (He and Garcia 2009). In other words, the five indices listed above (i.e., Sp, Sn, Acc, Str, and MCC) are threshold dependent. Considering that the *MCC* provides the

overall measurement of the quality of the binary predictions (Yu, Hu et al. 2013), we thus reported these threshold-dependent evaluation indices by choosing the threshold, denoted as T_{MaxMCC}, which maximizes the *MCC* value of the predictions. More specifically, we first identify the threshold that maximizes the value of *MCC* of the predictions on the training subset (e.g., D92M) over cross-validation, and the identified threshold was then used to evaluate the performance of the proposed method on the corresponding independent validation subset, i.e., Check390.

3. Experimental Results and Analysis

3.1 The Necessity of Refining Training Dataset

As described in section 2.1, the number of GPCR-drug pairs is continuously increasing under the efforts made by researchers and many GPCR-drug pairs that were previously labeled as non-interactive are recently identified as interactive. Thus, refining training dataset with newly discovered GPCR-drug pairs should be necessary for improving the performance of a prediction model, which will be demonstrated as follows:

We benchmarked the performances of four widely used classifiers, i.e., OET-KNN, QuickRBF, SVM, and RF on D92 and D92M, where D92M is a refinement of D92. More specifically, for each of the considered classifiers, we trained two corresponding classification models on D92 and D92M, respectively; then we tested the two models with the sequences in the independent validation dataset Check390. The comparison results are listed in Table 2.

By observing Table 2, we can find that the performance of each of the considered classifiers obtained on D92M is consistently better than that obtained on D92 concerning *Str* and *MCC*, which are two overall evaluation indexes. Averaged improvements of 3.3% and 6.0% on *Str* and *MCC*, respectively, were observed after refining D92 to D92M. Results in Table 2 empirically demonstrated that refining training dataset with newly discovered GPCR-drug pairs is necessary because it will help to improve the generalization capability of a trained model.

Table 2. Performance comparisons of different classifiers on D92 and D92M with Check390 as independent validation dataset.

Classifier	Training Dataset	*Sn* (%)	*Sp* (%)	*Acc* (%)	*Str* (%)	*MCC*	Confusion Matrix
OET-KNN	D92	60.8	85.4	77.2	73.1	0.47	[79 51; 38 222]
	D92M	67.7	84.2	78.7	75.9	0.52	[88 42; 41 219]
QuickRBF	D92	72.3	76.2	74.9	74.3	0.47	[94 36; 62 198]
	D92M	76.2	77.7	77.2	77.0	0.52	[99 31; 58 202]
SVM	D92	69.2	76.9	74.4	73.1	0.45	[90 40; 60 120]
	D92M	76.2	78.9	78.0	77.6	0.53	[99 31; 55 205]
RF	D92	72.3	77.7	75.1	75.0	0.48	[94 36; 58 202]
	D92M	78.5	78.1	78.2	78.3	0.54	[102 28; 57 203]

3.2 Choosing a Better Classifier for GPCR-drug Interactions Prediction under the Developed Feature Representation

It has been widely acknowledged that the classification/prediction performance depends not only on the feature's discriminative ability but also on the classifier being used. In this section, we will illustrate which classifier is relatively better under the developed feature representation. Table 3 lists performance comparisons between the four considered classifiers on D84 and D92M over leave-one-out cross-validation.

Table 3. Performance comparisons between OET-KNN, QuickRBF, SVM, and RF on D84 and D92M over leave-one-out cross-validation.

Dataset	Method	*Sn* (%)	*Sp* (%)	*Acc* (%)	*Str* (%)	*MCC*	Confusion Matrix
D84	OET-KNN	74.5	86.3	82.4	80.4	0.61	[234 80; 86 542]
	QuickRBF	72.3	89.3	83.7	80.8	0.62	[227 87; 67 561]
	SVM	71.7	88.5	82.9	80.1	0.61	[225 89; 72 556]
	RF	70.4	91.1	84.2	80.8	0.64	[221 93; 56 572]
D92M	OET-KNN	77.8	88.7	85.0	83.3	0.67	[494 141; 139 1086]
	QuickRBF	74.8	92.4	86.4	83.6	0.69	[475 160; 93 1132]
	SVM	74.2	92.7	86.3	83.5	0.69	[471 164; 90 1135]
	RF	76.5	92.9	87.3	84.7	0.71	[486 149; 87 1138]

From Table 3, we find that all the four considered classifiers achieve satisfactory classification performances with $MCC > 0.60$ and $Str > 0.80$ on both the D84 and D92M datasets, suggesting that it is feasible to perform GPCR-drug interactions prediction using machine-learning based methods. RF consistently performs better, although not significantly, than the other three considered classifiers, (i.e., OET-KNN, QuickRBF, and SVM). Thus, RF is a better choice for GPCR-drug interactions prediction under the proposed feature representation on the two benchmark datasets. In view of this, we will take RF as the classifier in all subsequent experiments, including the implementation of the web server.

Table 4. Performance comparisons between different classifiers under with and without 1-D wavelet transform on D92M over leave-one-out cross-validation.

Classifier	With 1-D Wavelet Transform	*Sn* (%)	*Sp* (%)	*Acc* (%)	*Str* (%)	*MCC*	Confusion Matrix
OET-KNN	Yes	77.8	88.7	85.0	83.3	0.67	[494 141; 139 1086]
	No	59.7	87.2	77.8	73.4	0.49	[379 256;157 1068]
QuickRBF	Yes	74.8	92.4	86.4	83.6	0.69	[475 160; 93 1132]
	No	62.2	88.0	79.2	75.1	0.52	[395 240; 147 1078]
SVM	Yes	74.2	92.7	86.3	83.5	0.69	[471 164; 90 1135]
	No	63.0	89.2	80.3	76.1	0.55	[400 235; 132 1093]
RF	Yes	76.5	92.9	87.3	84.7	0.71	[486 149; 87 1138]
	No	75.9	92.4	86.8	84.2	0.70	[482 153; 93 1132]

Here, we also demonstrated the necessity of performing 1-D wavelet transform for improving the discriminative capability of the original chemical fingerprint feature by comparing the performances between different classifiers under with- and without- wavelet transform on D92M over leave-one-out cross-validation. Experimental results are listed in Table 4. From Table 4, we can find that the prediction performances are consistently improved for all of the four considered

classifiers after performing wavelet transform on the chemical fingerprint feature. Averaged improvements of 6.5% and 12.5% were observed concerning *Str* and *MCC*, which are two overall evaluation indexes. We believe the underlying reason is that the wavelet transform can help to reduce some noise and redundant information contained in original chemical fingerprint so that the following classifier (i.e., RF classifier) can require higher performance.

3.3 Improving Prediction Performance Using the Post-processing Procedure

In this section, we will demonstrate the efficacy of the proposed post-processing procedure for further improving the performance of GPCR-drug interactions predictions. We perform performance comparisons between without and with the post-processing procedure on each of the three benchmark datasets, i.e., D84 and D92M, as follows:

Experiment I: Leave-one-out cross-validation is performed without the post-processing procedure and the results with the threshold (T_{MaxMCC}) that maximizes the *MCC* of the predictions are reported;

Table 5. Performance comparisons between without and with post-processing procedure on D84 and D92M over leave-one-out cross-validation.

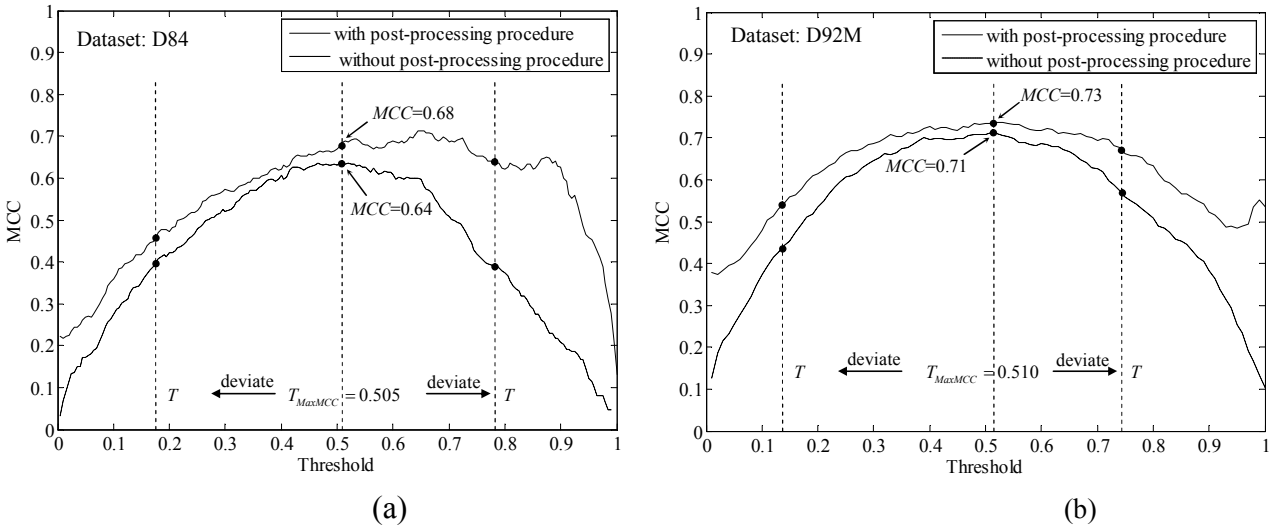
Dataset	With PPP	Threshold (T_{MaxMCC})	*Sn* (%)	*Sp* (%)	*Acc* (%)	*Str* (%)	*MCC*	Confusion Matrix
D84	No	0.505	70.4	91.1	84.2	80.8	0.64	[221 93; 56 572]
	Yes		71.3	93.5	86.1	82.4	0.68	[224 90; 41 587]
D92M	No	0.510	76.5	92.9	87.3	84.7	0.71	[486 149; 87 1138]
	Yes		79.7	92.8	88.3	86.3	0.73	[506 129; 88 1137]

Experiment II: We re-perform leave-one-out cross-validation with the post-processing procedure and the results with the same threshold (T_{MaxMCC}) identified in Experiment I are reported.

Table 5 illustrates the performance comparisons between without and with the post-processing procedure on D84 and D92M over leave-one-out cross-validation.

From Table 5, we can find that the prediction performances on both two benchmark datasets are improved after incorporating the proposed post-processing procedure. An average improvement of 3% in *MCC*, which is the overall measurement of the quality of the binary predictions, is observed on the two considered benchmark datasets. The *Str* of the method with PPP is about 1.6% better than that of the method without PPP on average. Taking the results on D92M as an example, the confusion matrix without the post-processing procedure is [486 149; 87 1138]; while that with the post-processing procedure is improved to [506 129; 88 1137]. Although the number of false positives is increased 1, the number of false negatives is reduced from 149 to 129, respectively. Similar improvements can also be observed on the other two benchmark datasets. Results listed in Table 5 demonstrate the effectiveness of the proposed post-processing procedure for enhancing the performance of GPCR-drug interactions prediction.

As mentioned above, the results listed in Table 5 are obtained with a specific threshold (T_{MaxMCC}) that maximizes the *MCC* of the predictions without the post-processing procedure. To systematically investigate the efficacy of the proposed post-processing procedure, we gradually adjust the threshold value from 0 to 1 with a step size of 0.01, and plot the curve of *MCC* versus threshold both with and without the post-processing procedure. Fig. 7 (a) and (b) illustrate the curves on D84, and D92M, respectively.



(a) (b)

Fig. 7. Comparisons of *MCC* between with and without the post-processing procedure by varying the threshold from 0 to 1 with a step size of 0.01. (a) Comparisons of *MCC* on the D84 dataset, (b) Comparisons of *MCC* on the D92M dataset.

From Fig. 7, two observations can be made.

First, on the two benchmark datasets, the values of *MCC* after incorporating the post-processing procedure are consistently higher than those without the post-processing procedure, with thresholds varying from 0 to 1. This demonstrates the robustness of the proposed post-processing procedure for GPCR-drug interactions prediction.

Second, on the two benchmark datasets, the amplitude of improvement of *MCC* tends to increase when the threshold T deviates away from T_{MaxMCC}, denoting that the proposed post-processing procedure can potentially reduce more false positives and/or false negatives under a threshold that is far away from T_{MaxMCC}.

3.4 Comparisons with Existing Predictors

In this section, we compare the proposed method with several recently published sequence-based GPCR-drug interactions predictors, including iGPCR-Drug (Xiao, Min et al. 2013) and the method by He et al. (He, Zhang et al. 2010). Note that in the subsequent descriptions, unless otherwise stated, the proposed method means that the feature representation developed in section 2.2 is used as the model input, and RF with the post-processing procedure, denoted as RF+PPP, is used as the

prediction engine.

A. Comparisons over Cross-Validation Test

Table 6 summarizes performance comparisons between the proposed method and other sequence-based GPCR-drug interactions predictors on the D84 and D92M datasets over leave-one-out cross-validation. Because iGPCR-Drug (Xiao, Min et al. 2013) does not report its results on the D84 and D92M datasets, we re-implemented this method and then performed leave-one-out cross-validation on D84 and D92M.

Table 6. Performance comparisons of different predictors on D84 and D92M over leave-one-out cross-validation.

Dataset	Predictor	Threshold	*Sn* (%)	*Sp* (%)	*Acc* (%)	*Str* (%)	*MCC*	Confusion Matrix
D84	iGPCR-Drug (Xiao, Min et al. 2013) [a]	N/A	77.4	80.1	79.2	78.8	0.56	[243 71; 125 503]
	Proposed Method (RF+PPP)	0.505	71.3	93.5	86.1	82.4	0.68	[224 90; 41 587]
D92M	Method by He et al. (He, Zhang et al. 2010)[b] *	N/A	N/A	N/A	78.5	N/A	N/A	N/A
	iGPCR-Drug (Xiao, Min et al. 2013) *	N/A	80.0	88.3	85.5	84.2	0.68	[496 124; 145 1095]
	iGPCR-Drug (Xiao, Min et al. 2013)[a]	N/A	78.3	91.4	86.9	84.9	0.71	[497 138; 105 1120]
	Proposed Method (RF+PPP)	0.510	79.7	92.8	88.3	86.3	0.73	[506 129; 88 1137]

[a] The re-implementation of iGPCR-Drug (Xiao, Min et al. 2013).
[b] Data excerpted from (He, Zhang et al. 2010).
* The result based on D92

From the results listed in Table 6, we find that the proposed method achieves the highest values of *MCC* and *Acc* on the three benchmark datasets and acts as the best performer. Taking *MCC* as an example, improvements of 12% and 2% are achieved by the proposed method compared with the second-best performer, iGPCR-Drug (Xiao, Min et al. 2013), which is the most recently released sequence-based predictor specifically designed for GPCR-drug interactions prediction.

B. Comparisons over Independent Validation Test

To demonstrate that the proposed method has not been over-optimized on the training datasets, an independent validation dataset as constructed in section 2.1 is used to compare the generalization capability between the proposed method and the other predictors. Note that for the purpose of fair comparisons, the same threshold that maximizes the value of *MCC* for predictions on the training dataset over leave-one-out cross-validation is used for the proposed method to perform an independent validation test. Table 7 lists the performance comparisons between different methods on the independent validation dataset.

By carefully analyzing the results listed in Table 7, several observations can be made as follows:

First, the proposed method again achieves the best performance on the independent validation

dataset with the highest values of *Str* (0.813) and *MCC* (0.60), which are 7.4% and 15% higher than that of iGPCR-Drug (Xiao, Min et al. 2013), respectively. These results demonstrate better generalization capability over the considered iGPCR-Drug predictor. The OET-KNN method achieves the best performance on *Sp* (84.2%) while having the lowest performance on *Sn* (67.7%), denoting that too many false positives are introduced during the predictions.

Second, we find that the prediction performance on the independent validation dataset can be further improved by incorporating the post-processing procedure (PPP) into the RF model. *Str* and *MCC* improvements of 3% and 6% are achieved by RF+PPP, respectively. This observation further demonstrates the effectiveness of the proposed post-processing procedure for GPCR-drug interactions prediction.

Table 7. Performance comparisons between different methods on the independent validation dataset.

Method	Threshold	*Sn* (%)	*Sp* (%)	*Acc* (%)	*Str* (%)	*MCC*	Confusion Matrix
iGPCR-Drug (Xiao, Min et al. 2013)[a]	N/A	80.8	66.9	71.6	73.9	0.45	[105 25; 86 174]
OET-KNN[b]	0.500	67.7	84.2	78.7	76.9	0.52	[88 42; 41 219]
QuickRBF [b]	0.450	76.2	77.7	77.2	77.6	0.52	[99 31; 58 202]
SVM [b]	0.420	76.2	78.9	78.0	77.6	0.53	[99 31; 55 205]
RF [b]	0.510	78.5	78.1	78.2	78.3	0.54	[102 28; 57 203]
Proposed Method (RF+PPP) [b]		83.1	79.6	80.8	81.3	0.60	[108 22; 53 207]

[a] Results obtained by feeding the sequences in the independent validation dataset to the iGPCR-Drug web server.

[b] The results based on D92M with feature representation developed in section 2.2.

3.5 Why Our Proposed Method is Better Than iGPCR-Drug

According to the experimental results in previous sections, it is easy to find that our method, called TargetGDrug, is better than iGPCR-Drug, at least on the considered benchmark datasets. In this section, we will seek to explain why the proposed TargetGDrug perform much better than iGPCR-Drug. The iGPCR-Drug employs *PseAAC* and *FT2DMFP* features, FKNN classifier, and without-PPP method. Our proposed TargetGDrug uses *PsePSSM* and *WaveletMFP* features, RF classifier, and with-PPP method.

Table 8 summarizes the performance comparisons between different methods on the independent validation dataset Check390.

Table 8. Performance comparisons between different methods on the independent validation dataset Check390.

Predictor	Feature representation	Classifier	With-PPP	*Sn* (%)	*Sp* (%)	*Acc* (%)	*Str* (%)	*MCC*
iGPCR-Drug (Xiao, Min et al. 2013)[a]	*PseAAC* and *FT2DMFP*	FKNN	No	80.0	68.1	72.1	74.0	0.45
-	*PseAAC* and *FT2DMFP*	RF	No	62.3	83.8	76.7	73.1	0.47
-	*PsePSSM* and *WaveletMFP*	RF	No	78.5	78.1	78.2	78.3	0.54
TargetGDrug	*PsePSSM* and *WaveletMFP*	RF	Yes	83.1	79.6	80.8	81.3	0.60

[a] The re-implementation of iGPCR-Drug (Xiao, Min et al. 2013).

From Table 8, several observations can be made as follows:

First, *PsePSSM* encodes both the evolutionary and sequence-order (positional) information of a GPCR sequence. Comparing with *PseAAC* of iGPCR-Drug, the dimensionality of *PsePSSM* is much higher thus possibly captures more useful information for classification;

Second, the wavelet transform is the better method than the flourier transform for processing the molecular fingerprint feature of a drug; and the learning ability of RF classifier is much better than that of FKNN;

Third, the drug-association-matrix-based post-processing procedure helps to reduce potential false positive or false negative of the initial predictions.

We believe these three points account for the superiority of the proposed TargetGDrug over iGPCR-Drug.

3.6 Case Study

In order to demonstrate the effectiveness of PPP method, two interactive GPCR-drug pairs (i.e., "hsa:3358 - D01358" and "hsa:1132 - D00525") and one non-interactive GPCR-drug pair (i.e., "hsa:5732 - D02356"), which have not been utilized for training the prediction model, were used for case study.

Table 9. Prediction results of three GPCR-drug pairs by using RF and RF+PPP.

Target	RF	RF+PPP
hsa:3358 - D01358	×	√
hsa:1132 - D00525	×	×
hsa:5732 - D02356	×	√

× denotes error prediction;
√ denotes correct prediction.

We tested the three GPCR-drug pairs using a RF model trained on D92M without PPP procedure. Unfortunately, the three pairs are all mistakenly predicted by RF model, as listed in Table 9. After applying PPP to the initial prediction of RF, we find two pairs, i.e., "hsa:3358 - D01358" and "hsa:5732 - D02356" were correctly predicted. However, "hsa:1132 - D00525 was still mistakenly predicted even after applying PPP. We speculate that the underlying reasons for this phenomenon are as follows:

We find drugs D01358 and D02356 show very promiscuous profiles, i.e., they interact with a number of GPCRs. For example, drug D01358 and D02356 interact with four and seven GPCRs, respectively, in D92M. On the contrary, D00525 interact with only one GPCR in D92M. In other words, there exist much historical information about drug D01358 and D02356, which can be utilized by PPP to effectively rectify those predictions related to D01358 and D02356. Since the historical information about drug D00525 is little, the PPP can not rectify the prediction of "hsa:1132 - D00525".

4. Discussions and Conclusions

In this study, we developed a novel sequence-based GPCR-drug interactions predictor with reasonable accuracy. Experimental results on benchmark datasets demonstrate the superiority of the proposed method over existing sequence-based predictors. The good performance of the proposed method comes from the use of the combined discriminative features of the GPCR-drug pair, the powerful RF classification algorithm, and particularly the post-processing procedure that can potentially reduce predicted false positives and false negatives. A user-friendly web server for the proposed method, called TargetGDrug, has been established and is available at http://csbio.njust.edu.cn/bioinf/TargetGDrug.

It should be pointed out that the modality (or mechanism) of the GPCR-drug interaction action, e.g., agonist/antagonist/inverse agonist, allosteric/orthosteric (Garland 2013; Wang and Lewis 2013), has not been extensively explored in this paper. We will try to further investigate the modality of the GPCR-drug interaction in our future work. In addition, we will also continuously update our model with new identified data from publicly available datasets such as ChemBl (Gaulton, Bellis et al. 2012) (https://www.ebi.ac.uk/chembl/).

ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 61373062, 61222306, 61233011, 91130033, and 61175024), the China Postdoctoral Science Foundation (No. 2013M530260, 2014T70526), "The Six Top Talents" of Jiangsu Province (No. 2013-XXRJ-022), the Graduate Research and Innovation Project of Jiangsu Province (No. KYZZ_0123), the Natural Science Foundation of Jiangsu (No. BK20141403), and the Fundamental Research Funds for the Central Universities (No. 30920130111010). We would also thanks Dr. Liang Xiao for providing essential instructions on the wavelet transform.

REFERENCES

Alberts, B. (2008). Molecular biology of the cell Garland Science, New York, 5th Ed.

Boulesteix, A. L. (2010). "Over-optimism in bioinformatics research." Bioinformatics **26**(3): 437-439.

Breiman, L. (2001). "Random forests." Machine Learning **45**(1): 5-32.

Breiman, L. (2001). "RandomForests." Mach.Learn. **45**(1): 5-32.

Chen, S.-A., Y.-Y. Ou, et al. (2011). "Prediction of transporter targets using efficient RBF networks with PSSM profiles and biochemical properties." Bioinformatics **27**(15): 2062-2067.

Cheng, A. C., R. G. Coleman, et al. (2007). "Structure-based maximal affinity model predicts small-molecule druggability." Nature Biotechnology **25**(1): 71-75.

Chou, K.-C. (2004). "Structural bioinformatics and its impact to biomedical science." Current medicinal chemistry **11**(16): 2105-2134.

Chou, K.-C. (2005). "Modeling the tertiary structure of human cathepsin-E." Biochemical and Biophysical Research Communications **331**(1): 56-60.

Chou, K.-C., D.-Q. Wei, et al. (2003). "Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS." Biochemical and Biophysical Research Communications **308**(1): 148-151.
Chou, K. C. (2001). "Prediction of protein cellular attributes using pseudo - amino acid composition." Proteins: Structure, Function, and Bioinformatics **43**(3): 246-255.
Chou, K. C. (2005). "Prediction of G-protein-coupled receptor classes." Journal of Proteome Research **4**(4): 1413-1418.
Chou, K. C. and H. B. Shen (2007). "MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM." Biochem Biophys Res Comm **360**: 339-345.
Eckert, H. and J. Bajorath (2007). "Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches." Drug discovery today **12**(5): 225-233.
Eswar, N., B. Webb, et al. (2006). "Comparative protein structure modeling using Modeller." Current protocols in bioinformatics: 5.6. 1-5.6. 30.
Fan, R. E., P. H. Chen, et al. (2005). "Working set selection using second order information for training SVM." Journal of Machine Learning Research **6**: 1889-1918.
Gao, M. and J. Skolnick (2012). "The distribution of ligand-binding pockets around protein-protein interfaces suggests a general mechanism for pocket formation." Proc Natl Acad Sci U S A **109**(10): 3784-3789.
Garland, S. L. (2013). "Are GPCRs still a source of new targets?" Journal of biomolecular screening: 1087057113498418.
Gaulton, A., L. J. Bellis, et al. (2012). "ChEMBL: a large-scale bioactivity database for drug discovery." Nucleic Acids Research **40**(D1): D1100-D1107.
Glaser, F., R. J. Morris, et al. (2006). "A method for localizing ligand binding pockets in protein structures." Proteins: Structure, Function, and Bioinformatics **62**(2): 479-488.
Granier, S. and B. Kobilka (2012). "A new era of GPCR structural and chemical biology." Nature chemical biology **8**(8): 670-673.
Haar, A. (1910). "Zur theorie der orthogonalen funktionensysteme." Mathematische Annalen **69**(3): 331-371.
Hamby, S. E. and J. D. Hirst (2008). "Prediction of glycosylation sites using random forests." BMC Bioinformatics **9**(1): 500.
He, H. and E. A. Garcia (2009). "Learning from Imbalanced Data." IEEE Transactions on Knowledge and Data Engineering **21**(9): 1263-1284.
He, Z., J. Zhang, et al. (2010). "Predicting drug-target interaction networks based on functional groups and biological features." PLoS One **5**(3): e9603.
Hopkins, A. L. and C. R. Groom (2002). "The druggable genome." Nature reviews Drug discovery **1**(9): 727-730.
Kandaswamy, K. K., G. Pugalenthi, et al. (2010). "SVMCRYS: An SVM Approach for the Prediction of Protein Crystallization Propensity from Protein Sequence." Protein and Peptide Letters **17**(4): 423-430.
Kanehisa, M., S. Goto, et al. (2006). "From genomics to chemical genomics: new developments in KEGG." Nucleic Acids Research **34**(suppl 1): D354-D357.
Keller, J. M., M. R. Gray, et al. (1985). "A fuzzy k-nearest neighbor algorithm." IEEE Transactions on Systems, Man and Cybernetics **15**(4): 580-585.
Knowles, J. and G. Gromo (2003). "A guide to drug discovery: Target selection in drug discovery." Nat Rev Drug Discov **2**(1): 63-69.
Kokubo, H., T. Tanaka, et al. (2011). "Ab initio prediction of protein-ligand binding structures by replica-exchange umbrella sampling simulations." J Comput Chem **32**(13): 2810-2821.
Krogh, A., B. Larsson, et al. (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." Journal of Molecular Biology **305**(3): 567-580.
Kunji, E. R., K. W. Chan, et al. (2005). "Eukaryotic membrane protein overproduction in Lactococcus lactis." Current opinion in biotechnology **16**(5): 546-551.
Le Guilloux, V., P. Schmidtke, et al. (2009). "Fpocket: an open source platform for ligand pocket detection." BMC Bioinformatics **10**(1): 168.
Liaw, A. and M. Wiener (2002). "Classification and Regression by random Forest." R news **2**(3): 18-22.
Mallat, S. G. (1989). "Multifrequency channel decompositions of images and wavelet models." IEEE Transactions on Acoustics, Speech and Signal Processing **37**(12): 2091-2110.
Mallat, S. G. (1989). "A theory for multiresolution signal decomposition: the wavelet representation." IEEE Transactions on Pattern Analysis and Machine Intelligence **11**(7): 674-693.
Nagarajan, R., S. Ahmad, et al. (2013). "Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins." Nucleic Acids Res **41**(16): 7606-7014.
Nanni, L., S. Brahnam, et al. (2014). "Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition." J Theor Biol **360C**: 109-116.
O'Boyle, N. M., M. Banck, et al. (2011). "Open Babel: An open chemical toolbox." Journal of cheminformatics **3**(1): 1-14.
Ou, Y. (2005). "QuickRBF: a package for efficient radial basis function networks." Software available at http://csie.org/~yien/quickrbf.
Qiu, J.-D., J.-H. Huang, et al. (2009). "Prediction of G-protein-coupled receptor classes based on the concept of Chou's pseudo amino acid composition: an approach from discrete wavelet transform." Analytical biochemistry **390**(1): 68-73.

Qiu, J.-D., X.-Y. Sun, et al. (2010). "Prediction of the types of membrane proteins based on discrete wavelet transform and support vector machines." The protein journal **29**(2): 114-119.
Quince, C., A. Lanzen, et al. (2011). "Removing noise from pyrosequenced amplicons." BMC Bioinformatics **12**(1): 38.
Roth, B. L., D. L. Willins, et al. (1998). "G protein-coupled receptor (GPCR) trafficking in the central nervous system: relevance for drugs of abuse." Drug and alcohol dependence **51**(1): 73-85.
Roy, A. and Y. Zhang (2012). "Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement." Structure **20**(6): 987-997.
Schaffer, A. A. (2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements." Nucleic Acids Research **29**: 2994-3005.
Schaffer, A. A., L. Aravind, et al. (2001). "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements." Nucleic Acids Research **29**(14): 2994-3005.
Schmidtke, P. and X. Barril (2010). "Understanding and predicting druggability. A high-throughput method for detection of drug binding sites." J Med Chem **53**(15): 5858-5867.
Shen, H.-B. and K.-C. Chou (2007). "EzyPred: a top–down approach for predicting enzyme functional classes and subclasses." Biochemical and Biophysical Research Communications **364**(1): 53-59.
Shen, H.-B. and K.-C. Chou (2008). "PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition." Analytical biochemistry **373**(2): 386-388.
Stehman, S. V. (1997). "Selecting and interpreting measures of thematic classification accuracy." Remote sensing of Environment **62**(1): 77-89.
Tate, C. G. (2012). "A crystal clear solution for determining G-protein-coupled receptor structures." Trends in biochemical sciences **37**(9): 343-352.
Vapnik, V. N., Ed. (1998). Statistical Learning Theory Wiley-Interscience, New York.
Villasenor, J. D., B. Belzer, et al. (1995). "Wavelet filter evaluation for image compression." IEEE Transactions on Image Processing **4**(8): 1053-1060.
Wang, C.-I. A. and R. J. Lewis (2013). "Emerging opportunities for allosteric modulation of G-protein coupled receptors." Biochemical pharmacology **85**(2): 153-162.
Wong, E. S., M. C. Hardy, et al. (2013). "SVM-based prediction of propeptide cleavage sites in spider toxins identifies toxin innovation in an Australian tarantula." PLoS One **8**(7): e66279.
Worth, C. L., A. Kreuchwig, et al. (2011). "GPCR-SSFE: a comprehensive database of G-protein-coupled receptor template predictions and homology models." BMC Bioinformatics **12**(1): 185.
Xiao, X., J. L. Min, et al. (2013). "iGPCR-Drug: A Web Server for Predicting Interaction between GPCRs and Drugs in Cellular Networking." PLoS One **8**(8).
Yamanishi, Y., M. Araki, et al. (2008). "Prediction of drug-target interaction networks from the integration of chemical and genomic spaces." Bioinformatics **24**(13): I232-I240.
Yamanishi, Y., M. Kotera, et al. (2010). "Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework." Bioinformatics **26**(12): i246-i254.
Yang, J., A. Roy, et al. (2013). "Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment." Bioinformatics **29**(20): 2588-2595.
Yu, D.-J., J. Hu, et al. (2013). "Learning protein multi-view features in complex space." Amino Acids **44**(5): 1365-1379.
Yu, D.-J., Y. Li, et al. "Disulfide Connectivity Prediction Based on Modelled Protein 3D Structural Information and Random Forest Regression." IEEE/ACM Transactions on Computational Biology and Bioinformatics(1): 1-1.
Yu, D., J. Hu, et al. (2013). "Designing template-free predictor for targeting protein-ligand binding sites with classifier ensemble and spatial clustering." IEEE/ACM Transactions on Computational Biology and Bioinformatics **10**(4): 994-1008.
Yu, D., X. Wu, et al. (2012). "Enhancing membrane protein subcellular localization prediction by parallel fusion of multi-view features." IEEE Trans Nanobioscience **11**(4): 375-385.
Zhang, J. and Y. Zhang (2010). "GPCRRD: G protein-coupled receptor spatial restraint database for 3D structure modeling and function annotation." Bioinformatics **26**(23): 3004-3005.
Zhu, S., Y. Okuno, et al. (2005). "A probabilistic model for mining implicit 'chemical compound–gene'relations from literature." Bioinformatics **21**(suppl 2): ii245-ii251.
Zia-Ur-Rehman and A. Khan (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 Pept Lett **19**(8): 890-903.
Zouhal, L. M. and T. Denoeux (1998). "An evidence-theoretic K-NN rule with parameter optimization." IEEE Transactions on Systems, Man and Cybernetics **28**: 263-271.