Bioinformatics, 32(7), 2016, 1057–1064
doi: 10.1093/bioinformatics/btv695
Advance Access Publication Date: 26 November 2015
Original Paper

Systems biology

Identification of drug–target interaction from interactome network with 'guilt-by-association' principle and topology features

Zhan-Chao Li[1,*], Meng-Hua Huang[1], Wen-Qian Zhong[1], Zhi-Qing Liu[1], Yun Xie[1], Zong Dai[3] and Xiao-Yong Zou[2,3,*]

[1]School of Chemistry and Chemical Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, People's Republic of China, [2]SYSU-CMU Shunde International Joint Research Institute, Shunde 528300, People's Republic of China and [3]School of Chemistry and Chemical Engineering, Sun Yat-Sen University, Guangzhou 510275, People's Republic of China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 19, 2015; revised on November 16, 2015; accepted on November 19, 2015

Abstract

Motivation: Identifying drug–target protein interaction is a crucial step in the process of drug research and development. Wet-lab experiment are laborious, time-consuming and expensive. Hence, there is a strong demand for the development of a novel theoretical method to identify potential interaction between drug and target protein.

Results: We use all known proteins and drugs to construct a nodes- and edges-weighted biological relevant interactome network. On the basis of the 'guilt-by-association' principle, novel network topology features are proposed to characterize interaction pairs and random forest algorithm is employed to identify potential drug–protein interaction. Accuracy of 92.53% derived from the 10-fold cross-validation is about 10% higher than that of the existing method. We identify 2272 potential drug–target interactions, some of which are associated with diseases, such as Torg-Winchester syndrome and rhabdomyosarcoma. The proposed method can not only accurately predict the interaction between drug molecule and target protein, but also help disease treatment and drug discovery.

Contacts: zhanchao8052@gmail.com or ceszxy@mail.sysu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Most drugs are a class of small molecule compounds which act by activating or inhibiting the biological activity or function of the specific target proteins. Identifying interaction between drug and target can help to not only reduce the time, cost and failure rates for new developing drug (Kim *et al.*, 2013), but also decipher the mechanism of drug action (Bantscheff and Drewes, 2012) and pathomechanism of disease (Wang *et al.*, 2012).

Experimental methods containing affinity chromatography and protein microarray have been widely adopted to identify drug–target interaction (Cheng *et al.*, 2011). However, these methods are ineffective in the face of massive biomolecules data. With the completion of the Human Genome Project and development of combinatorial chemistry, abundant biological and chemical data have been generated and deposited in the various database. As of January 2015, the UniProtKB (Boutet *et al.*, 2007) and ChEMBL (Gaulton *et al.*, 2012) contain 142 140 human protein entities and 1 638 394 compound entities, respectively. Undoubtedly, it is impossible to carry out an exhaustive experiment for detecting all possible interactions between these proteins and compounds, because the pairwise combinations are astronomical. It motivates researchers to develop a theoretical method to not only quickly and accurately

discriminate the potential drug–target interactions, but also guide experimentalists and provide supporting evidence for their experimental results (Gonen, 2012).

Currently, many studies show that relationship between disease and drug is far more complex than the 'one gene, one drug, one disease' paradigm (Pei *et al.*, 2014). Since proteins often interact with each other and with other molecules in cells to form an interaction network, network-based researches have emerged as a promising alternative to accelerate the identification of drug–target interaction (Alaimo *et al.*, 2013; Chen *et al.*, 2012; Cheng *et al.*, 2012a,b; Mei *et al.*, 2013; Xia *et al.*, 2010; Yamanishi *et al.*, 2008; Yildirim *et al.*, 2007). These works generally model the drug–target interactions as a bipartite graph, in which vertex corresponds to either drug molecule or target protein, edge corresponds to drug–target interaction. Then, various strategies are utilized to mine the bipartite graph and predict potential drug–target interaction based on the known interaction.

In spite of the advances in bipartite graph-based methods, it is a challenge to accurately identify drug–target interaction. The first reason is that most existing approaches only consider the interaction of drug–target, and neglect the interaction of protein–protein and relationship of drug–drug. Indeed, target protein usually exerts their function by interacting with other protein rather than working alone at the cellular level (Baudot *et al.*, 2012; Chi and Hou, 2011). Structurally different drugs can bind to the multiple targets and express similar or identical bioactivities. Additionally, more than one drug can be synergistically used in the treatment of diseases. The second one is that bipartite graph-based methods do not consider the false positive of protein–protein interaction. Currently, protein–protein interaction data mainly comes from the yeast two-hybrid and affinity purification with mass spectrometry. But, the two methods usually suffer from high false positive rate due to the technical limitation. The third one is that drug and protein are only considered as nodes in mathematics, and ignore their biological or chemical properties. In fact, the interaction between drug and target is influenced by many factors, such as volume, shape and charge of drug compound as well as hydrophobicity, polarity and tertiary structure of target protein. The fourth one is that only known targets are utilized to construct the bipartite graph. Therefore, they are usually unable to discover the new interaction of drug–target excluded in the constructed network. The last one is that the information of network topology is not adopted to recognize drug–target interaction. Research of protein–protein interaction network have shown that target proteins usually have higher degree and betweenness centrality than non-target proteins (Hase *et al.*, 2009; Kotlyar *et al.*, 2012; Yao and Rzhetsky, 2008), implying that network topology information can be utilized to differentiate target proteins from non-target proteins.

In view of these reasons, we develop a novel theoretical computation method to infer drug–target interaction based on the 'guilt-by-association' principle and network topology features. Instead of using only the bipartite network of drug–target interaction, we construct a drug–target interactome network containing three subnetworks of protein–protein interaction, drug–target interaction and drug–drug relationship to grasp the complex interaction relationship between drug and target. Each node in the interactome network is weighted by using either protein primary sequence descriptors or drug molecular structure features to characterize their attributes. In the subnetwork of protein–protein interaction and drug–drug relationship, edge is weighted based on the protein interaction probability score and the number of common target protein, respectively. With the use of the network topology information, we develop a novel network topology feature vector to characterize drug–target interaction. The underlying assumption is 'guilt-by-association', in which a target protein is likely to interact with a drug if the majority of protein's neighbors (i.e. they can directly interact with the protein in the subnetwork of protein–protein interaction) can interact with the drug, vice versa. Finally, we employ random forest (RF) algorithm (Leo, 2001) to construct model for identifying potential drug–target interaction.

2 Methods

2.1 Construction of drug–target interactome network

In order to construct the subnetwork of protein–protein interaction, we retrieve the human protein–protein interaction data from the HIPPIE database (Schaefer *et al.*, 2012). After deleting self-interactions, repeated interactions and interactions with interaction confidence score 0, a protein–protein interaction subnetwork is acquired (Supplementary file PPIN.xlsx). Protein primary structure is retrieved from the UniprotKB/Swiss-Prot (Boutet *et al.*, 2007) based on the access number (AC). To reflect the protein properties, we calculate primary structure descriptors (i.e. vertex weights) with 1767-dimensions and containing amino acid composition (20-dimensions), dipeptide composition (400-dimensions), normalized Moreau-Broto autocorrelation (400-dimensions), Moran autocorrelation (400-dimensions), Geary autocorrelation (400-dimensions), composition (21-dimensions), transition (21-dimensions) and distribution (105-dimensions).

In order to build the subnetwork of drug–drug relationship (Supplementary file DDRN.xlsx), we download the data of small molecule drug with detailed structural formula and molecular weight less than 1000 Da from the DrugBank database (Law *et al.*, 2014). After removing those compounds that their target proteins do not belong to the human species and include them in the built subnetwork of protein–protein interaction, we retrieve the information of molecular structure from the file of all.sdf.zip based on the identification (ID). Each vertex of the subnetwork is weighted by using molecular fingerprint feature vector with 1024-dimensions by the calculation of the PaDEL-Descriptor software (Yap, 2011). In the binary vector, elements 1 and 0 indicate the presence or absence of a specific chemical substructure. The molecular fingerprint feature has been widely used in the studies of quantitative structure-activity relationship (Dimova *et al.*, 2013; Rabal *et al.*, 2015). Two drugs are interconnected through a weighted edge if they share at least one target protein. The weight of edge is defined $|C|/\min(|A|,|B|)$, where $|A|$ and $|B|$ represent the number of target protein of chemical A and chemical B, $|C|$ is the number of shared target protein, and $\min(|A|, |B|)$ denotes the smallest value between $|A|$ and $|B|$. According to the definition, the edge weight locates in the range of [0, 1]. A larger weight value means that two drugs have more common target proteins, higher similar structure and bioactivity.

Based on the downloaded drug information, we construct the drug–target interaction subnetwork (Supplementary file DTIN.xlsx), in which a drug belonging to the subnetwork of drug–drug relationship and a target belonging to the subnetwork of protein–protein interaction is connected to each other if drug and target interact, and the corresponding weight is set to 1.

Finally, the drug–target interactome network is constructed by mapping AC of target protein in the drug–target interaction subnetwork to AC of protein in the protein–protein interaction network, and ID of drug in the drug–target interaction subnetwork to ID of drug in the drug–drug relationship subnetwork. The interactome is

composed of 240 300 edges and 17 695 nodes. In all these edges, 153 749 edges represent protein–protein interactions, 77 713 drug–drug relationships, 8838 drug–target interactions. In all these nodes, 14 086 are proteins in which 1523 are drug targets, 3609 are small molecule drugs in which 1038 are approved by the FDA, the remaining are experimental compounds, illicit or withdrawn drugs.

2.2 Characterization of drug–target interaction

In a network-based method for predicting drug–target interaction, it is one of the most important things to develop a suitable encoding method for extracting core and essential attributes of drug–target interaction. However, it is by no means easy to characterize the attributes, because it is highly related to the structures of drug molecule and target protein, and deeply hidden in the complicated interaction network. In order to solve the problem, we utilize the 'guilt-by-association' principle to encode the interaction information. Based on the principle, we assume that a drug molecule tends to attack a target protein if the majority of the drug's neighbors in the interactome network interacts with the target, vice versa.

For an interaction between drug d and target protein p, we firstly find the neighbor molecules of drug d in the subnetwork of drug–drug relationship and the corresponding edge weights. By considering the topology information of neighbor molecules, vertex weights and edge weights, we further calculate the network topology feature of the drug d (DNTF_d) based on the following Eq. (1).

$$\mathrm{DNTF}_d = \frac{1}{N}\sum_{j=1}^{N} D_d(i) \times E_{d,j} \times D_j(i) \quad (i = 1, \quad 2,\ldots\ldots, 1024) \quad (1)$$

where, $D_d(i)$ and $D_j(i)$ represent the ith node weight of drug d and neighbor molecule j, respectively. $E_{d,j}$ implies the edge weight between d and j. N is the number of neighbor molecules. According to the definition, DNTF_d is a feature vector with 1024 dimensions. If the neighbor molecules do not exist (i.e. $N = 0$), we define $\mathrm{DNTF}_d = D_d$.

Secondly, we search the neighbor molecules of target protein p in the subnetwork of protein–protein interaction and the corresponding edge weights. We also calculate the network topology feature of target protein p (PNTF_p) according to the Eq. (2).

$$\mathrm{PNTF}_p = \frac{1}{N}\sum_{j=1}^{N} P_p(i) \times E_{p,j} \times P_j(i) \qquad (i = 1, \quad 2,\ldots\ldots, 1767) \quad (2)$$

In this equation, $P_p(i)$ and $P_j(i)$ indicate the ith node weight of target protein p and neighbor j, respectively. $E_{p,j}$ is the interaction confidence score (i.e. edge weight) between target protein p and protein j. N is the number of neighbor proteins. According to the Eq. (2), PNTF_p is a feature vector with 1767 dimensions. If the target protein has no neighbors (i.e. $N = 0$), we define $\mathrm{PNTF}_p = P_p$.

Finally, we formulate the interaction between drug d and target p as a feature vector with 2791 (1024 + 1767) dimensions by concatenating DNTF_d and PNTF_p. This combinational network topology features consider simultaneously the interaction attributes from drug molecules and target proteins.

2.3 Construction of model and assessment of performance

In our study, the identification of drug–target interaction can be viewed as a missing edge prediction problem, meaning that the subnetwork of drug–target interaction is an incomplete graph with missing edges. Our goal is to make use of the observed edges to construct a classifier and predict whether an edge exists between potential drug molecule and target protein. Matlab version of the RF algorithm downloaded at http://code.google.com/p/randomforest-matlab/ is adopted, because the prediction can be formulated as a binary classification problem (i.e. the presence or absence of an edge between drug node and protein node) from the perspective of statistics and machine learning.

For the binary classification problem, false sample of drug–target interaction (i.e. non-interaction pair between drug and target) is essential for constructing classifier. Unfortunately, there is no database dedicated to the collection of experimentally verified negative drug–target interaction. Therefore, we use the following strategy to generate non-interaction sample: (1) randomly couple any protein in the subnetwork of protein–protein interaction and any drug in the subnetwork of drug–drug relationship into a false sample of drug–protein interaction, (2) eliminate the negative sample contained in the collected dataset of true drug–target interaction, (3) repeat step (1) and (2) until the number of negative sample reaches the same number as the positive sample. The generated negative samples are unlikely to contain a large number of positive drug–target interaction samples that have not been proven until now, because the fact that the amount of true drug–target interaction is smaller than that of false drug–target interaction. Finally, a benchmark dataset containing all true drug–target interactions and false drug–target interactions with equal size to positive samples is established. The proportion of '1:1' can overcome the limitation of the larger number of negative samples and result in unbiased prediction.

Taking into account the comparison with other existing methods, we carry out 10-fold cross-validation, in which the benchmark dataset is randomly split into 10 equally sized subsets. Every time, one distinct subset and remaining nine subsets are utilized to train and test classifier, respectively. In addition to the conventional measures of accuracy (*Acc*), sensitivity (*Sen*), specificity (*Spe*), Precision (*Pre*) and Matthews correlation coefficient (*Mcc*), we also employ receiver operating characteristic curve (ROC) and precision recall (also known as *Sen*) curve (PRC) to estimate the predictive ability of the model. The two curves are obtained by plotting *Sen* versus 1-*Sen* and *Pre* versus recall by gradually changing threshold. We adopt the trapezoidal approximation to calculate the areas under ROC and PRC for assessing prediction performance, because the uppermost curve with the largest area indicates that the constructed model is the most outstanding.

Flowchart of the developed method is illustrated in Figure 1, and the steps are follows:

Step 1. Retrieve the information of human protein–protein interaction and drug–target interaction from the HIPPIE and DrugBank database, respectively.

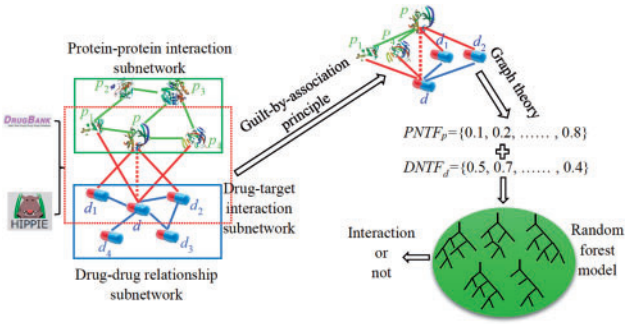

Fig. 1. Flowchart of the current method

Step 2. Construct the interactome network containing three subnetworks of protein–protein interaction, drug–drug relationship and drug–target interaction. Nodes and edges are weighted based on the information of protein sequence, molecular structure and interaction among them.

Step 3. Calculate the network topological feature to characterize drug–target interaction based on the 'guilt-by-association' principle and graph theory.

Step 4. Utilize RF algorithm to construct classifier for identifying potential drug–target interaction pairs. In this step, the number of trees (*ntree*) contained in forest and the number of variables (*mtry*) randomly selected at each node of tree are set to 100 and default value (i.e. square root of the number of all features), respectively. Two-thirds and one-third of the benchmark dataset are considered as in-bag samples and out-of-bag samples, respectively.

3 Results

3.1 Construction and evaluation of model

For the current study, RF model is constructed to identify the potential drug–target interaction. Moreover, various combinations between *ntree* changing from 100 to 1000 with a step size 100 and *mtry* changing from 2^0 to 2^7 with a step size 2^1 are also utilized to construct classifiers. However, the significant improvement for prediction performance is not observed.

In order to investigate the robustness of the current method for random sampling (i.e. generate negative samples), the process of benchmark dataset constructing, classifier building and assessing is repeated 10 times. The statistical results of *Acc*, *Sen*, *Spe*, *Pre* and *Mcc* as well as ROC and PRC derived from the model with the highest *Acc* are shown in Figure 2. From Figure 2A, we can see that our method obtains the statistical average of 92.53%, 94.05%, 91.01%, 91.28% and 0.8510 for *Acc*, *Sen*, *Spe*, *Pre* and *Mcc*, and the corresponding relative standard deviations are only 0.19%, 0.21%, 0.26%, 0.24% and 0.42%. The results indicates that the current method has strong robustness for the sampling of negative samples. Figure 2B exhibits that the proposed algorithm achieves the areas under ROC and PRC up to 0.9799 and 0.9609, revealing the excellent effectiveness of the current method. From the two curves, additionally, we can also observe that the proposed method can effectively capture sufficient information to identify the drug–target interaction with high true-positive rates against low false-positive rates, and with high positive predictive values against low true-positive rates at any threshold. In one word, these results indicate that the proposed method has an excellent capability for predicting the interaction between drug and target.

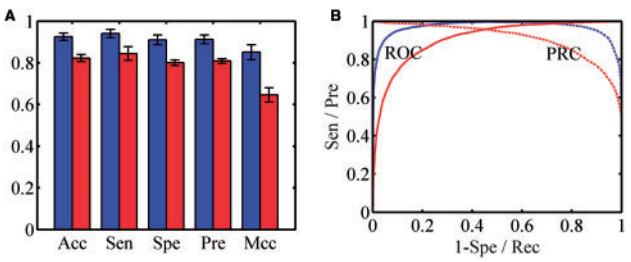

Fig. 2. The results of 10-fold cross-validation based on the various datasets. (**A**) The statistical average results of 10 experiments (The panels indicate the mean values of *Acc, Spe, Sen, Pre* and *Mcc* derived from the current method (blue) and existing method (red), respectively. The vertical bars indicate the standard deviations). (**B**) The ROC and PRC curves based on the developed method (blue) and existing method (red)

3.2 Effectiveness of the 'guilt-by-association' principle

To demonstrate the excellent performance of the current method with the use of the principle of 'guilt-by-association', we perform a comparison with work of Cao *et al.* (2013) based on the constructed 10 benchmark datasets in the Section 3.1. In their study, compound structure and protein sequence are firstly converted into numerical descriptors based on the various methods and feature vectors. Then, the chemical–protein interaction is simply represented by concatenating these chemical and protein descriptor. Finally, the concatenated vector is input into a machine learning model to predict whether an interaction occurs between a given chemical and a specific protein. In the comparison, a feature vector by concatenating our protein descriptors and fingerprint features is used to characterize drug–target interaction following their work. Then, RF algorithm is adopted to construe model and predict interaction between drug and target protein.

The statistical results as well as ROC and PRC derived from the classifier with the highest *Acc* are shown in Figure 2. From Figure2A, we can see that the average *Acc*, *Sen*, *Spe* and *Pre* are 82.27%, 84.47%, 80.07% and 80.91%, which are about 10% lower than those of the current method. *Mcc* is about 0.5 lower than that of the proposed method. Figure 2B shows that the values of *Sen* and *Pre* at the ROC and PRC are always consistently lower than those from the current method when the false-positive rate and recall are changed from 0 to 1. The areas under the two curves are 0.9079 and 0.8983, about 0.07 lower than those of the current method. These results indicate that our approach is superior to the methods from the literature, and the encoding strategy based on the 'guilt-by-association' principle is able to efficiently capture the drug–target interaction information embedded in the interaction network.

3.3 Predictive ability of the developed method for novel drug–target interaction

In order to validate the prediction ability of drug–target interaction, researchers (Ding *et al.*, 2014; Gonen, 2012; Pahikkala *et al.*, 2015; van Laarhoven *et al.*, 2011; Yamanishi *et al.*, 2008) usually utilize four different prediction scenarios to confirm the prediction performance. In these scenarios, targets and drugs contained in the training set are called 'known' whereas those involved in the test set are called 'new'. The four scenarios are: (i) predicting whether a known target and a known drug can interact with each other (i.e. target and drug are enclosed in the training set), (ii) inferring a new target for a known drug (i.e. target and drug are contained in the test set and training set, respectively), (iii) finding a new drug for a known target (i.e. drug and target are involved in the test set and training set, respectively), (iv) estimating whether a new target and a new drug can interact with each other (i.e. target and drug are included in the test set). The first scenario is the most widely used in experimental design for theoretical prediction, in which the objective is to predict the missing interactions without going outside the training space (Pahikkala *et al.*, 2015). The second and third are the real situations with practical value, in which only drug information or target information is used during the model construction phase, and can be employed to study drug reposition, drug off-target, drug combination and drug promiscuity, etc. The fourth is the most rigorous setting, in which neither drug information nor target information is adopted during the model training phase, and can be utilized to design novel drugs. In the current study, we use the fourth scenario to demonstrate the prediction performance by a series of non-redundant datasets.

In order to construct non-redundant datasets, we define the similarity between two drug–target interaction pairs. For the two

interaction pairs of p_1–d_1 and p_2–d_2, we first calculate the protein sequence similarity between p_1 and p_2 by using the Needleman–Wunsch algorithm and scoring matrix of BLOSUM50. Second, calculate the drug structure similarity between d_1 and d_2 based on the absolute value of Pearson's correlation coefficient of molecular fingerprint descriptors by using Eq. (3):

$$R_{d_1,d_2} = abs\left(\frac{\sum_{i=1}^{N}\left(D_{d_1}(i)-\overline{D_{d_1}}\right)\left(D_{d_2}(i)-\overline{D_{d_2}}\right)}{\sqrt{\sum_{i=1}^{N}\left(D_{d_1}(i)-\overline{D_{d_1}}\right)^2\times\sum_{i=1}^{N}\left(D_{d_2}(i)-\overline{D_{d_2}}\right)^2}}\right) \quad (3)$$

Here $D_{d_1}(i)$ and $D_{d_2}(i)$ means the ith fingerprint feature value of d_1 and d_2, $\overline{D_{d_1}}$ and $\overline{D_{d_2}}$ are the mean values of all fingerprint feature values of d_1 and d_2. $N = 1024$, and *abs* represents the operation of absolute value. According to the definition, $R_{d1,d2}$ is located in the range of [0, 1]. '1' means that the two drug molecules d_1 and d_2 are completely identical, '0' means that they are completely different. Finally, the similarity between the two interaction pairs is obtained by calculating the mean value of sequence similarity and structure similarity. According to the definition, the similarity of two drug–target interaction pairs is always located between 0 and 1, the number of '1' represents the two interaction pairs with complete identical, '0' shows complete difference. By setting the similarity threshold of 0.2, 0.3, 0.4, . . . , 0.9, we can construct 8 non-redundant drug–target interaction databases, in which the similarity between any two interaction pairs is always lower than a given threshold. We do not set threshold to 0.1, because the corresponding non-redundant dataset only contains 6 true drug–target interaction pairs, the number is too small to have statistical significant.

The results of 10-fold cross-validation test for these non-redundant datasets are listed in Table 1. We can observe that with the decrease of the threshold from 0.9 to 0.2, *Acc*, *Sen*, *Spe*, *Pre* and *Mcc* are also gradually reduced. Our method achieves *Acc* up to 91.97% when the threshold is set to 0.9. *Acc* is always higher than 80% when the threshold is decreased from 0.8 to 0.3. The current method still obtains *Acc* of 77.52% even if the threshold is reduced to 0.2. Therefore, we can conclude that the proposed method has the excellent performance for predicting novel drug–target interaction. Based on these non-redundant datasets, we further perform a comparison with work of Cao *et al.* (2013), and results also demonstrate the superiority of our method (see Supplementary Materials for details).

3.4 Robustness of the proposed model for false positive of protein–protein interactions

We construct the subnetwork of human protein–protein interaction in order to consider the target network topology information in the process of encoding drug–target interaction samples. The constructed subnetwork may contain some false positive interactions due to drawbacks in experimental techniques for detecting protein–protein interaction. In order to overcome the problem of false positive, an interaction confidence score in the range of [0, 1] is assigned to each protein–protein interaction for explicating the possibility of interaction. In this section, the robustness for false positive rate is investigated by constructing 8 subnetworks of protein–protein interaction with the interaction confidence score higher than 0.1, 0.2, . . . , 0.8, respectively. Please note that we do not build and discuss the subnetwork with interaction confidence higher than 0.9, because it is composed of 1781 proteins and 3030 protein–protein interactions, but only one protein (Q13085) is the drug target.

For the various human protein–protein interaction subnetwork, we construct corresponding eight dataset and then carry out 10-fold cross-validation test. From the Table 2, it is illustrated that *Acc*, *Sen*, *Spe*, *Pre* and *Mcc* always fluctuate slightly in the range of [92.55%, 92.75%], [93.99%, 94.44%], [90.90%, 91.40%], [91.21%, 91.63%] and [0.8513, 0.8554], respectively. More stringently, our method still achieves more than 92% of *Acc*, *Sen*, *Spe* and *Pre*, 0.84 of *Mcc* when the interaction confidence score is further increased to 0.7 and 0.8. The results indicates that the current method has a good robustness for false positive rate of protein–protein interaction.

3.5 Identification of potential drug–protein interaction

After verifying the performance of the current method, we conduct a comprehensive prediction for unknown drug–protein interaction pairs by using the model derived from the constructed redundant dataset and with the highest *Acc*. In these unknown interaction pairs, drug and protein comes from the subnetwork of drug–drug relationship and the subnetwork of protein–protein interaction, respectively. And, these unknown interaction pairs are produced by randomly combining drug and protein. Finally, the constructed model identifies a total of 41 749 potential drug–protein interactions between 3406 drugs and 5972 proteins. We rank all the potential interaction pairs according to the votes in descending order, and the results show that 2272 interaction pairs (Supplementary file PredDTIP.xlsx) are the most likely to be putative interactions because their votes are higher than 90 (i.e. in all 100 trees in forest, interaction is predicted by more than 90 trees). Then we check (see Supplementary Materials for details) whether these predicted interaction pairs appear in BindingDB database (Liu *et al.*, 2007), which is a public and web-accessible database of measured binding affinities between proteins and small, drug-like molecules. Results show that there are 6 predicted drug–protein interaction pairs recorded in the BindingDB (Supplementary file ComDTIP.xlsx). By comparing

Table 1. Results of the 10-fold cross-validation test based on the various non-redundant datasets

Threshold	*Acc* (%)	*Sen* (%)	*Spe* (%)	*Pre* (%)	*Mcc*
0.9	91.97	94.16	89.77	90.20	0.8401
0.8	89.93	93.66	86.19	87.15	0.8008
0.7	88.60	93.28	83.91	85.29	0.7753
0.6	87.80	93.91	81.69	83.69	0.7618
0.5	85.20	91.41	78.99	81.31	0.7094
0.4	83.07	89.59	76.54	79.25	0.6671
0.3	81.34	85.19	77.49	79.10	0.6286
0.2	77.52	81.65	75.42	75.42	0.5523

Table 2. Results of the 10-fold cross-validation test based on the various human protein–protein networks with interaction confidence score higher than 0.1, 0.2, . . . , 0.8

Score	*Acc* (%)	*Sen* (%)	*Spe* (%)	*Pre* (%)	*Mcc*
0.8	92.68	92.37	92.99	92.95	0.8537
0.7	92.13	92.25	92.01	92.03	0.8427
0.6	92.67	94.05	91.28	91.52	0.8537
0.5	92.55	93.99	91.11	91.36	0.8513
0.4	92.60	94.08	91.12	91.37	0.8524
0.3	92.75	94.11	91.40	91.63	0.8554
0.2	92.63	94.35	90.90	91.21	0.8531
0.1	92.74	94.44	91.04	91.33	0.8553

randomly selected 2272 interaction pairs from all predicted interactions with the interactions from the BindingDB database, the expected value is calculated to be 3.5447 (see Supplementary Materials for details), indicating that the number of common drug–target interaction pairs is 3.5447 when randomly comparing our predicted interactions with collected interactions. Therefore, we can conclude that our developed method has a certain capability and reliability in practical usage.

Here, we take the 2 predicted and toped drug–protein interaction pairs as examples to illustrate the application of current method. The binding between experimental drug *N*-{2-[(4′-cyano-1,1′-biphenyl-4-yl)oxy]ethyl}-*N*′-hydroxy-*N*-methylurea (DB06971) and protein 72 kDa type IV collagenase (P08253) identified by our method (98 votes) is annotated in the BindingDB and supported by evidence derived from the reference (Campestre *et al.*, 2006). The DB06971 is an organic compound and belongs to biphenylcarbonitriles. The P08253 is an ubiquitinous metalloproteinase with various functions such as remodeling of the vasculature, angiogenesis, tissue repair tumor invasion, inflammation and atherosclerotic plaque rupture. To further study and confirm the interaction, we perform docking simulations based on the AutoDock program (Santos-Martins *et al.*, 2014) and DS Visualizer software. The three-dimensional structural information of protein and drug are downloaded from the http://www.rcsb.org/pdb (ID: 1CK7) and http://www.drugbank.ca/drugs/DB06971, respectively. The grid center coordinates of box are set to 82.37A, 101.67A and 159.155A. The Lamarckian genetic algorithm is employed to search the docking conformation. Finally, we obtain the optimal docking model with binding energy −8.44 kcal/mol and inhibition constant (K_i) 655.08 nM. Complex model of ligand and receptor as well as their interaction are shown in Figure 3. We can see that there are Van der Waals interactions between drug molecule and amino acid residues Ile478, Gln480, Val523, Tyr524, Pro527, Ala571, Phe572, Ser575, Val620, Val621 and Gln624. The small molecular drug is connected to the target protein through hydrogen bonds between Asn573 and carbonyl group as well as Trp574, Lys578 and hydroxyl. Meanwhile, there are carbon hydrogen bond between the drug and Asp622, π-donor hydrogen bond between phenyl ring and Asn573. Additionally, it is also shown that electrostatic interaction (π-anion) and hydrophobic interaction (π–σ) are formed between phenyl ring and Glu525 as well as phenyl ring and Ala479. These results show that the probability of interaction between the drug and the protein is very high. Some studies (Martignetti *et al.*, 2001; Rouzier *et al.*, 2006; Zankl *et al.*, 2005) have shown that the variation of protein is associated with the multicentric osteolysis, nodulosis and arthropathy (MONA, also known as Torg-Winchester syndrome). Therefore, we can speculate that the drug DB06971 can be used to treat the syndrome through targeting the protein P08253.

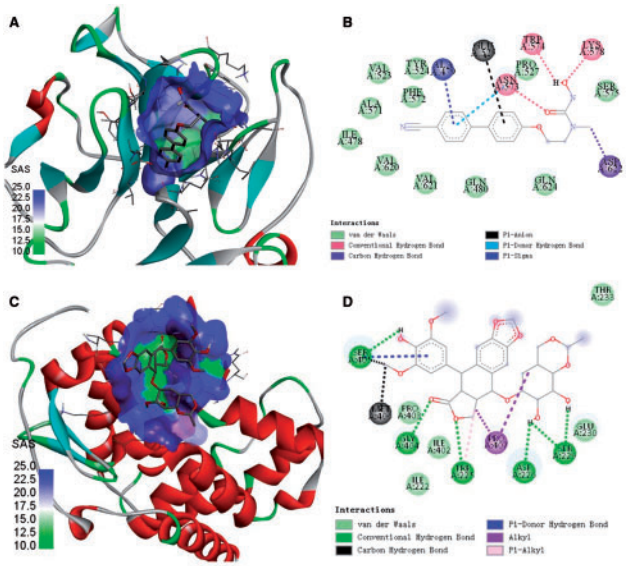

Fig. 3. The complex model (**A**) of ligand DB06971 and receptor P08253 as well as their interactions (**B**). The complex model (**C**) of ligand DB00773 and receptor Q15788 as well as their interactions (**D**)

Etoposide (DB00773) is a semisynthetic derivative of podophyllotoxin and utilized to treat various cancers, such as leukemia, small cell lung cancer, malignant lymphoma, ovarian cancer, choriocarcinoma, testicular cancer, bladder cancer and prostate cancer, by inhibiting DNA synthesis. In DrugBank database, the etoposide attacks two known target proteins (Azarova *et al.*, 2007; Uesaka *et al.*, 2007), namely DNA topoisomerase 2-α (P11388) and DNA topoisomerase 2-β (Q02880). Interestingly, a new target protein, nuclear receptor coactivator 1 (Q15788) is predicted for the drug molecule with 98 votes. And, the new predicted interaction is also supported by evidences derived from the BindingDB. We also perform docking simulations based on the three-dimensional structural information of protein 3KMR in the PDB database and drug DB00773 in the DrugBank database. The grid center coordinates x, y and z of box are set to −7.342A, −12.289A and −14.4A. The optimal complex model with binding energy −8.38 KJ/mol and K_i 655.08 nM is obtained and shown in the Figure 3. We can observe that the drug is connected to the protein by hydrogen bonds between Asp226, Ser229, Ser405 and hydroxyl, Gly404 and carbonyl, Trp225 and oxygen, Met406 and hydrogen (carbon hydrogen bond), Ser405 and hydrogen (carbon hydrogen bond) as well as Ser405 and phenyl ring (π-donor hydrogen bond). Additionally, there are hydrophobic interactions between Pro407 and drug (alkyl hydrophobic) as well as Trp225 and drug (π-alkyl hydrophobic). And, Van der Waals interaction between the drug and the target protein happens through Ile222, Thr233, Glu230, Ile402 and Pro403. A chromosomal aberration involving Q15788 causes rhabdomyosarcoma, which is the most common soft tissue carcinoma in childhood, representing 5–8% of all malignancies in children (Wachtel *et al.*, 2004). Therefore, we can reasonably speculate that the drug may also be repositioned in the treatment of rhabdomyosarcoma.

In one word, all above outcomes enhance the strength of the current method for realistic drug–target interaction prediction application, these putative drug–target interactions provide invaluable information for experimentalist, especially in solving problems associated with drug repositioning, drug combination and drug promiscuity, etc.

4 Discussion

Effectiveness and performance of the current method is evaluated against various datasets through the 10-fold cross-validation and confirmed by comparing with the existing methods, and the results indicate that our method can accurately predict the interaction between drug molecule and target protein. Robustness on the false positive protein–protein interaction is also confirmed based on a series of constructed protein interaction network and the corresponding drug–target interaction datasets. Success of the developed method can be attributed to three aspects. Firstly, using interactome network supplies a comprehensive and system viewpoint to identify

potential drug–target interaction and understand the interaction mechanisms. Secondly, integrating drug molecular structure information with target protein sequence information into a unified framework provides further insight for the nature of the interaction between drug and protein. Thirdly, resorting the 'guild-by-association' principle opens up a new avenue to develop topology features for characterizing drug–target interaction pairs in the context of network.

The main advantages of the current approach are summarized as follows: (i) compared with the structure-based theoretical methods, our method is not constrained by the 3D structure data of targets; (ii) in contrast to experimental methods, the developed approach only takes a few seconds to identify whether a drug targets a protein at the proteome scale; (iii) the proposed approach is able to identify those drug–target interaction pairs with low protein sequence similarity and drug structure similarity; (iv) the developed method can aid in the research of drug repositioning and drug promiscuity by recognizing an interaction between a new target protein and a known drug.

Of course, our approach also has some limitations. For example, it can predict only the binding between proteins and all known drugs because the constructed drug–drug relationship subnetwork only contains the drug molecules from the DrugBank. Currently, we are trying to expand the subnetwork by adding a variety of small molecule compounds with bioactivity based on the database of BindingDB, ChEBI and ChEMBLdb. This strategy can help to discovery novel lead compounds and drug molecules. Besides, the negative drug–target interaction pair with the same size for positive pairs is used to construct dataset in order to overcome prediction biases for samples with larger size. We carry out 10 test, and the statistical results indicate that the current method is robustness for random sampling of negative samples. In fact, the amount of negative samples is much higher than that of positive ones in nature. It is significance to develop novel machine learning algorithm for overcoming the problem, but it is out of the scope of the current study. In addition, we do not take into account the mode of action of the drug molecule due to the little database now devoted to the collection of data. In future research, we will solve this problem with the accumulation of relevant data.

Despite possible limitations, we believe that our approach provides a novel avenue to determine drug–target interaction, helps to fill the existing gap between chemical proteomics and network pharmacology, and has an important impact on the mechanism of action of drugs and targets.

Funding

This work was supported by National Natural Science Foundation of China [21205019, 81171666]; Natural Science Foundation of Guangdong Province [S2013010012135]; Scientific Technology Project of Guangdong Province [2014A040401022, 2015A030401033]; Ph.D. Programs Foundation of the Ministry of Education of China [20110171110014].

Conflict of Interest: none declared.

References

Alaimo,S. *et al.* (2013) Drug–target interaction prediction through domain-tuned network-based inference. *Bioinformatics*, **29**, 2004–2008.

Azarova,A.M. *et al.* (2007) Roles of DNA topoisomerase II isozymes in chemotherapy and secondary malignancies. *Proc. Natl. Acad. Sci. USA*, **104**, 11014–11019.

Bantscheff,M. and Drewes,G. (2012) Chemoproteomic approaches to drug target identification and drug profiling. *Bioorg. Med. Chem.*, **20**, 1973–1978.

Baudot,A. *et al.* (2012) Network analysis and protein function prediction with the PRODISTIN Web site. *Methods Mol. Biol.*, **804**, 313–326.

Boutet,E. *et al.* (2007) UniProtKB/Swiss-Prot. *Methods Mol. Biol.*, **406**, 89–112.

Campestre,C. *et al.* (2006) N-hydroxyurea as zinc binding group in matrix metalloproteinase inhibition: mode of binding in a complex with MMP-8. *Bioorg. Med. Chem. Lett.*, **16**, 20–24.

CaoD.S. *et al.* (2013) Large-scale prediction of human kinase-inhibitor interactions using protein sequences and molecular topological structures. *Anal. Chem. Acta.*, **792**, 10–18.

Chen,X. *et al.* (2012) Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.*, **8**, 1970–1978.

Cheng,F. *et al.* (2012a) Prediction of drug–target interactions and drug repositioning via network-based inference. *PLos Comput. Biol.*, **8**, e1002503.

Cheng,F. *et al.* (2012b) Prediction of chemical–protein interactions network with weighted network-based inference method. *PLoS One*, **7**, e41064.

Cheng,T. *et al.* (2011) Identifying compound-target associations by combining bioactivity profile similarity search and public databases mining. *J. Chem. Inf. Model.*, **51**, 2440–2448.

Chi,X. and Hou,J. (2011) An iterative approach of protein function prediction. *BMC Bioinformatics*, **12**, 437.

Dimova,D. *et al.* (2013) Quantifying the fingerprint descriptor dependence of structure-activity relationship information on a large scale. *J. Chem. Inf. Model.*, **53**, 2275–2281.

Ding,H. *et al.* (2014) Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Brief. Bioinform.*, **15**, 734–747.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100–D1107.

Gonen,M. (2012) Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinfomratics*, **28**, 2304–2310.

Hase,T. *et al.* (2009) Structure of protein interaction networks and their implications on drug design. *PLoS Comput. Biol.*, **5**, e1000550.

Kim,S. *et al.* (2013) Predicting drug–target interactions using drug–drug interactions. *PLoS One*, **8**, e80129.

Kotlyar,K. *et al.* (2012) Network-based characterization of drug-regulated genes, drug targets, and toxicity. *Methods*, **57**, 499–507.

Law,V. *et al.* (2014) DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res.*, **42**, D1091–D1087.

Liu,T. *et al.* (2007) BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.*, **35**, D198–D201.

LeoB. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Martignetti,J.A. *et al.* (2001) Mutation of the matrix metalloproteinase 2 gene (MMP2) causes a multicentric osteolysis and arthritis syndrome. *Nat. Genet.*, **28**, 261–265.

Mei,J.P. *et al.* (2012) Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, **29**, 238–245.

Pahikkala,T. *et al.* (2015) Toward more realistic drug–target interaction predictions. *Brief. Bioinform.*, **16**, 325–337.

Pei,J. *et al.* (2014) Systems biology brings new dimensions for structure-based drug design. *J. Am. Chem. Soc.*, **136**, 11556–11565.

Rabal,O. *et al.* (2015) Novel scaffold fingerprint (SFP): application in scaffold hopping and scaffold-based selection of diverse compounds. *J. Chem. Inf. Model.*, **55**, 1–18.

Rouzier,C. *et al.* (2006) A novel homozygous MMP2 mutation in a family with winchester syndrome. *Clin. Genet.*, **39**, 271–276.

Santos-Martins,D. *et al.* (2014) AutoDock4(Zn): an improved AutoDock force field for small-molecule docking to zinc metalloproteins. *J. Chem. Inf. Model.*, **54**, 2371–2379.

Schaefer,M.H. *et al.* (2012) HIPPIE: integrating protein interaction networks with experiment based quality scores. *PLoS One*, **7**, e31826.

Uesaka,T. *et al.* (2007) Enhanced expression of DNA topoisomerase II genes in human medulloblastoma and its possible association with etoposide sensitivity. *J. Neurooncol.*, **84**, 119–129.

van Laarhoven,T. *et al*. (2011) Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, **27**, 3036–3043.

Wachtel,M. *et al*. (2004) Gene expression signatures identify rhabdomyosarcoma subtypes and detect a novel t(2;2)(q35;p23) translocation fusing PAX3 to NCOA1. *Cancer Res*., **64**, 5539–5545.

Wang,Y.Y. *et al*. (2012) Predicting drug targets based on protein domains. *Mol. Biosyst*., **8**, 1528–1534.

Xia,Z. *et al*. (2010) Semi-supervised drug–protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol*., **4**, S6.

Yamanishi,Y. *et al*. (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.

Yao,L. and Rzhetsky,A. (2008) Quantitative systems-level determinants of human genes targeted by successful drugs. *Genome Res*., **18**, 206–213.

Yap,C.W. (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J. Comput. Chem*., **32**, 1466–1474.

Yildirim,M.A. *et al*. (2007) Drug–target network. *Nat. Biotechnol*., **25**, 1119–1126.

Zankl,A. *et al*. (2005) Winchester syndrome caused by a homozygous mutation affecting the active site of matrix metalloproteinase 2. *Clin. Genet*., **67**, 261–266.