Bioinformatics, 31, 2015, i221–i229
doi: 10.1093/bioinformatics/btv256
ISMB/ECCB 2015

Improving compound–protein interaction prediction by building up highly credible negative samples

Hui Liu[1,2,†], Jianjiang Sun[3,†], Jihong Guan[4], Jie Zheng[2] and Shuigeng Zhou[3,*]

[1]Lab of Information Management, Changzhou University, Jiangsu 213164, China, [2]School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore, [3]Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai 200433, China and [4]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Abstract

Motivation: Computational prediction of compound–protein interactions (CPIs) is of great importance for drug design and development, as genome-scale experimental validation of CPIs is not only time-consuming but also prohibitively expensive. With the availability of an increasing number of validated interactions, the performance of computational prediction approaches is severely impended by the lack of reliable negative CPI samples. A systematic method of screening reliable negative sample becomes critical to improving the performance of *in silico* prediction methods.
Results: This article aims at building up a set of highly credible negative samples of CPIs via an *in silico* screening method. As most existing computational models assume that similar compounds are likely to interact with similar target proteins and achieve remarkable performance, it is rational to identify potential negative samples based on the converse negative proposition that the proteins dissimilar to every known/predicted target of a compound are not much likely to be targeted by the compound and vice versa. We integrated various resources, including chemical structures, chemical expression profiles and side effects of compounds, amino acid sequences, protein–protein interaction network and functional annotations of proteins, into a systematic screening framework. We first tested the screened negative samples on six classical classifiers, and all these classifiers achieved remarkably higher performance on our negative samples than on randomly generated negative samples for both *human* and *Caenorhabditis elegans*. We then verified the negative samples on three existing prediction models, including bipartite local model, Gaussian kernel profile and Bayesian matrix factorization, and found that the performances of these models are also significantly improved on the screened negative samples. Moreover, we validated the screened negative samples on a drug bioactivity dataset. Finally, we derived two sets of new interactions by training an support vector machine classifier on the positive interactions annotated in DrugBank and our screened negative interactions. The screened negative samples and the predicted interactions provide the research community with a useful resource for identifying new drug targets and a helpful supplement to the current curated compound–protein databases.
Availability: Supplementary files are available at: http://admis.fudan.edu.cn/negative-cpi/.
Contact: sgzhou@fudan.edu.cn
Supplementary Information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Compound–protein interactions (CPIs) are crucial to the discovery of new drugs by screening candidate compounds and are also helpful for understanding the causes of side effects of existing drugs. Although various biological assays are available, experimental validation of CPIs remains time-consuming and expensive. Therefore, there is a strong incentive to develop computational methods to detect CPIs accurately. Meanwhile, with the rapid growth of public chemical and biological databases, such as the PubChem (Wheeler *et al*., 2006), DrugBank (Wishart *et al*., 2008), SIDER (Kuhn *et al*., 2010), STITCH (Kuhn *et al*., 2014), STRING (Franceschini *et al*., 2013) and Gene Ontology (GO) (Ashburner *et al*., 2013), various kinds of resources, including drug features such as chemical structures, side effects and gene expression profiles under drug treatments and protein features such as amino acid sequences, protein–protein interaction (PPI) networks and functional annotations, become available to the research community and consolidate the basis of computational CPI prediction.

Traditional computational approaches fall roughly into two categories: structure based and ligand based. The structure-based methods depend on the structural information of target proteins that are often unavailable for most protein families. Ligand-based methods get poor performance for those proteins having few or none of the known ligands. Recently, a variety of machine learning-based methods have been proposed and achieved a considerable success by taking the viewpoint of *chemogenomics* (Jaroch and Weinmann, 2006), which integrates the chemical attributes of drug compounds, the genomic attributes of proteins and the known CPIs into a unified mathematical framework. The main rationale underlying the chemogenomics approaches is that *similar* compounds tend to bind *similar* proteins, so that the lack of known ligands for a given protein can be compensated by the availability of known ligands of *similar* proteins and vice versa.

Following the philosophy of chemogenomics, many methods have been proposed by exploiting various types of features and classification algorithms (Tabei and Yamanishi, 2013; Yabuuchi *et al*., 2011; Yamanishi *et al*., 2010). With the chemical structure similarity and protein sequence similarity measures, Bleakley and Yamanishi (2009) proposed the bipartite local model (BLM) to infer CPIs by training local support vector machine (SVM) classifiers based on known interactions. van Laarhoven *et al*. (2011) proposed Gaussian interaction profile (GIP) kernels that exploit the topology of CPI networks. However, these methods still suffer from the lack of known interactions between the drugs and proteins of interest, which often leads to the failure of prediction. Therefore, Mei *et al*. (2013) improved BLM by exploiting the known interactions of neighbors to compensate the lack of interaction information. van Laarhoven and Marchiori (2013) also used the interaction profiles of weighted nearest neighbors to improve the GIP method.

Instead of utilizing the attributes of drugs and proteins separately, more and more researchers combined these attributes into a single feature vector by the concatenation or tensor product operators and then built classifiers based on the integrated features and known CPIs. For example, Jacob and Vert (2008) proposed SVM classifiers with pairwise kernels that were derived, respectively, from similarity measures of drugs and proteins. Yamanishi *et al*. (2008) proposed the bipartite graph inference that maps drugs and proteins into a unified Euclidean feature space in which the distances between drugs and proteins linked by known interactions are minimized and otherwise maximized. The network-based inference originally proposed for personal recommendation (Zhou *et al*., 2010) was also used to identify CPIs (Alaimo *et al*., 2013; Cheng *et al*., 2012). Other methods including kernel-based data fusion (Wang *et al*., 2011), Kernelized Bayesian matrix factorization with twin kernels (KBMF2K) (Gonen, 2012), restricted Boltzmann machine (Wang and Zeng, 2013) and semi-supervised methods (Chen and Zhang, 2013; Xia *et al*., 2010) were successively proposed. In addition to chemical structures and protein sequences, researchers also resorted to other attributes of drugs and proteins to reveal their interacting associations, including the drug expression profiles (Carrella *et al*., 2014; Iorio *et al*., 2010; Wolpaw *et al*., 2011), functional groups (He *et al*., 2010) and side effects (Campillos *et al*., 2008; Mizutani *et al*., 2012) of drugs, signaling pathways and GO annotations (Jaeger *et al*., 2014) of proteins or even the combination of these attributes (Gottlieb *et al*., 2011, 2012; Perlman *et al*., 2011).

Most previous approaches used experimentally validated CPIs as positive samples and randomly generated negative samples to learn the prediction models. However, the randomly generated negative samples may include real positive samples not yet known. A classifier trained by using such randomly generated negative samples may yield high cross-validation accuracy but very possibly has poor performance on independent, real test datasets. Screening highly reliable negative samples is therefore critical to improving the accuracy of computational prediction methods. The importance of true-negative interactions was recently highlighted as one of the future developments in predicting drug–target interactions (Ding *et al*., 2014). Motivated by this, we set about to screen *in silico* highly credible negative samples of CPIs. An assumption underlying most computational methods for predicting CPIs is that similar drug compounds are likely to interact with similar target proteins. Our method is based on the converse negative proposition, i.e. the proteins that are dissimilar to every known/predicted target of a given compound are not much likely to be targeted by the compound and vice versa. We integrated various resources of compounds and proteins, including chemical structures, chemical expression profiles and side effects of compounds, amino acid sequences, PPI networks and GO functional annotations of proteins, to a systematic screening framework. We evaluated our method on both human and *Caenorhabditis elegans* data. We first tested our screened negative samples on six classical classifiers, including random forest, L1- and L2-regularized logistic regression, naive Bayes, SVM and *k*-nearest neighbor (kNN). All these classifiers achieved remarkably higher performance on our negative samples than on randomly generated negative samples. We also verified our negative samples on three existing prediction models, including BLM (Bleakley and Yamanishi, 2009), Gaussian kernel profile (van Laarhoven *et al*., 2011) and Bayesian matrix factorization (Gonen, 2012), and found that the performances of these models are also significantly improved on the screened negative samples. Furthermore, we validated our screened negative samples with a drug bioactivity dataset.

Finally, we derived two sets of new CPIs by training an SVM classifier on the positive interactions annotated in DrugBank and our screened negative interactions. These screened negative samples and the predicted interactions can serve the research community as a useful resource for identifying new drug targets and as a helpful supplement to the current curated compound–protein databases.

2 Materials

2.1 Compound–protein interaction

CPIs were retrieved from DrugBank 4.1 (Wishart *et al*., 2008), Matador (Gnther *et al*., 2008) and STITCH 4.0 (Kuhn *et al*., 2014).

DrugBank and Matador are manually curated databases, and STITCH is a comprehensive database that collects CPIs from four different sources: experiments, databases, text mining and predicted interactions. Meanwhile, STITCH provides a score ranging from 0 to 1000 for each interaction, which indicates the confidence of the CPI supported by four types of evidence, i.e. experimental validation, manually curated databases, text mining and predicted interactions. We assigned the interactions from DrugBank and Matador the highest score 1000 because these interactions are supported by biochemical experiments and the literature. Totally, we got 2 290 630 interactions between 367 142 unique compounds and 19 342 proteins of human, and 2 141 740 interactions between 276 294 unique compounds and 11 234 proteins of *C.elegans*. For simplicity, we refer to the created assembly of CPIs as K and denote by a triple $(c_i, p_j, w_{ij}) \in K$ the interaction between drug c_i and protein p_j with confidence score w_{ij} in the rest of the article.

2.2 Chemical data

2.2.1 Chemical structure similarity

Chemical structures (also referred to as *fingerprints*) of drugs were obtained from the PubChem database (Wheeler *et al*., 2006). We calculated the Jaccard score (Jaccard, 1908) of the fingerprints as the chemical structure similarity between compounds. The Jaccard score between compounds c and c' is defined as $|c \cap c'|/|c \cup c'|$, which is the ratio of the number of common substructures between c and c' over the total number of substructures in c and c'. There are totally 881 kinds of substructures used in our analysis for human and *C.elegans*. Applying this operation to all drug pairs, we thus constructed a chemical similarity matrix.

2.2.2 Side effect similarity

Side effects of drugs were downloaded from the SIDER database (Kuhn *et al*., 2010). For the drugs involving in CPIs but not included in the SIDER database, we employed a recently proposed method that predicts side effects based on chemical fragments (Pauwels *et al*., 2011) to predict side effects. Similarly, we computed the Jaccard score of each pair of drugs as side effect similarity based on either their known side effects or top 10 predicted side effects in case they are unknown (Perlman *et al*., 2011).

2.3 Protein data

2.3.1 Sequence similarity

Amino acid sequences of proteins were obtained from the UCSC Table Browser. We computed sequence similarity between proteins using a normalized version of Smith–Waterman score (Smith and Waterman, 2010). The normalized Smith–Waterman score between two proteins g and g' is $\mathrm{sw}(g,g')/\sqrt{\mathrm{sw}(g,g)}\sqrt{\mathrm{sw}(g',g')}$ where $\mathrm{sw}(.,.)$ means the original Smith–Waterman score. Applying this operation to all protein pairs, we got the similarity matrix of protein sequences.

2.3.2 Functional annotation semantic similarity

GO annotations were downloaded from the GO database (Ashburner *et al*., 2013). Semantic similarity score between each pair of proteins was calculated based on the overlap of the GO terms that were associated with the two proteins (Coutoa *et al*., 2007). All three types of ontologies were used in the computation as similar drugs are expected to interact with proteins that act in similar biological processes or have similar molecular functions or reside in similar compartments. We computed the Jaccard score with respect to the GO terms of each pair of proteins as their similarity.

2.3.3 Protein domain similarity

Protein domains were extracted from PFAM database (Punta *et al*., 2012). Each protein was represented by a domain fingerprint (binary vector) whose elements encode the presence or absence of each retained PFAM domain by 1 or 0, respectively. The numbers of PFAM domains for human and *C.elegans* are 1331 and 3837, respectively. We computed the Jaccard score of any two proteins via their domain fingerprints as their similarity.

3 Methods

3.1 Integration of multiple similarities

We have computed multiple similarity measures from different features for both drugs and proteins as mentioned above. For drugs c_i and c_j, we formulate them into a single comprehensive similarity measure as below:

$$\mathrm{CS}_{ij} = 1 - \prod_n (1 - \mathrm{cs}_{ij}^{(n)}) \tag{1}$$

in which $\mathrm{cs}_{ij}^{(n)}$ $(n = 1, 2)$ represents the similarity measure derived from features of chemical structure and side effect, respectively. Note that similar formulation was also adopted by STITCH (Kuhn *et al*., 2014), as it can be easily extended to integrate more similarity measures. Similarly, we computed the comprehensive similarity between proteins p_i and p_j by

$$\mathrm{PS}_{ij} = 1 - \prod_n (1 - \mathrm{ps}_{ij}^{(n)}), \tag{2}$$

where $\mathrm{ps}_{ij}^{(n)}$ $(n = 1, 2, 3)$ represents the similarity measure derived from sequence similarity, functional annotation semantic similarity and protein domain similarity, respectively.

3.2 The screening framework

Most existing prediction methods for CPIs (positive samples) are based on the assumption that similar compounds are likely to interact with the proteins that are similar to the corresponding known target proteins. Our basic idea was inspired by the converse negative proposition of this assumption. Specifically, we assume that a protein dissimilar to every known/predicted target of a compound is not much likely to be targeted by this compound, and on the other hand, a compound not similar to any known/predicted compound targeting a protein is not much likely to target this protein. For simplicity, we refer them as *protein dissimilarity rule* and *drug dissimilarity rule*, respectively. Both rules are simultaneously applied in our screening framework so as to identify the most reliable negative samples of CPIs. Different from existing prediction methods that often depend on known CPIs for making reliable predictions, our negative sample screening framework exploits both validated and predicted CPIs. Figure 1 shows the flowchart of our method. Here, the three green dashed-line boxes show the data resources used in our screening framework, and the protein dissimilarities and drug dissimilarities are, respectively, computed so as to gain a combined score for each candidate negative sample. We summarize the screening steps as follows:

1. Compute the integrated similarity of each pair of compounds/proteins via Equation (1)/Equation (2). Build the assembly K of known/predicted CPIs as mentioned above.
2. Build the set of candidate negative interactions from all possible interactions excluding the created assembly K of known/predicted CPIs. We take the candidate negative interaction

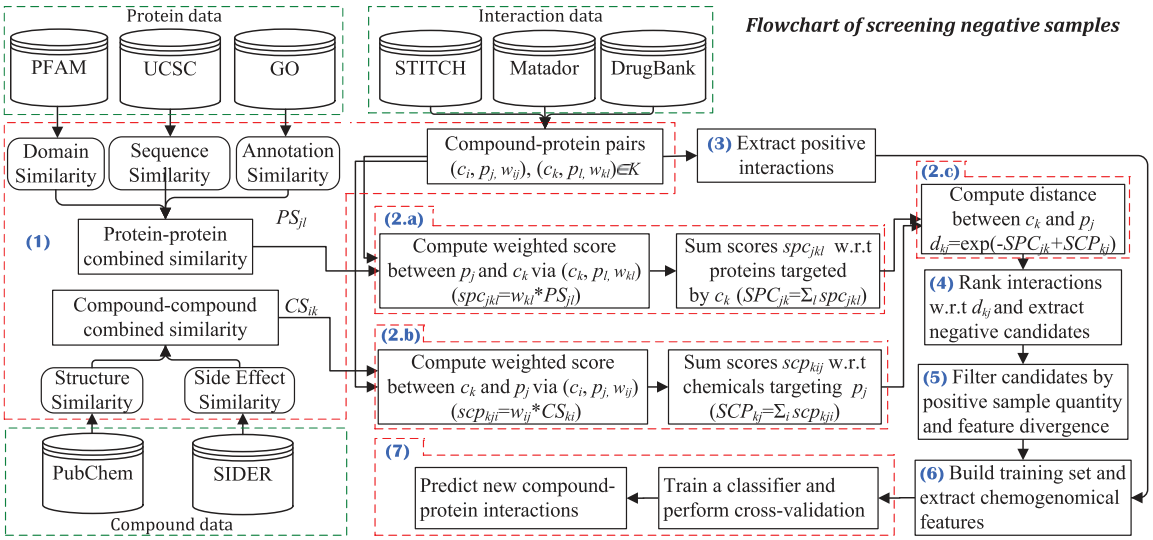

Fig. 1. The flowchart of our negative CPI screening framework. Three green dashed-line boxes show the data resources used in our screening process, and the red dashed-line boxes represent the screening steps that include multiple operations

between compound k and protein j, denoted by (c_k, p_j, d_{kj}) with d_{kj} indicating the distance between compound k and protein j, as an example to demonstrate the screening process. Figure 2 is to illustrate the process of calculating d_{kj}.

a. For any protein p_l targeted by c_k in K, we compute the weighted score $\text{spc}_{jkl} = w_{kl} * \text{PS}_{jl}$ that indicates the possibility of protein p_j being targeted by compound c_k in consideration of the similarity between p_j and p_l. Taking into account the similarity between p_j and each known/predicted protein p_l targeted by compound c_k, i.e. $(c_k, p_l, w_{kl}) \in K$, we calculate the combined score by summing up the weighted scores spc_{jkl} with respect to l, and thus obtain $\text{SPC}_{jk} = \sum_l \text{spc}_{jkl}$.

b. Similarly, we compute the weighted score $\text{scp}_{kji} = w_{ij} * \text{CS}_{ik}$ that represents the possibility of compound c_k targeting p_j in consideration of the similarity between c_k and c_i. Considering the similarity between c_k and each known/predicted compound c_i targeting protein p_j, i.e. $(c_i, p_j, w_{ij}) \in K$, we calculate the combined score by summing up the weighted scores spc_{kji} with respect to i and thus obtain $\text{SCP}_{kj} = \sum_i \text{scp}_{kji}$.

c. For compound c_k and protein p_j, we define the distance between c_k and p_j as below:

$$d_{kj} = e^{-(\text{SPC}_{jk} + \text{SCP}_{kj})}. \quad (3)$$

d_{kj} is the final score representing the possibility that compound c_k does not target protein p_j. The larger d_{kj} is, the higher the possibility of c_k not targeting p_j is.

3. Build the set of positive interactions from two manually curated databases: DrugBank (Wishart *et al.*, 2008) and Matador (Gnther *et al.*, 2008).
4. Rank the potential negative CPIs according to the scores obtained by Equation (3), and those with the highest scores are taken to form the set of negative sample candidates.
5. The negative sample candidates are further filtered by using feature divergence of compound and protein, as described in Section 3.3.
6. Combining the positive interactions and negative interactions, we get a gold standard set of CPIs. On the basis of the chemical substructures and protein PFAM domains, we construct the tensor product for each CPI, so that each interaction is represented by a vector in the chemogenomical space.
7. Train a classifier (e.g. SVM) by using the chemogenomical feature vectors, tune the model parameters via cross-validation and finally predict new CPIs.

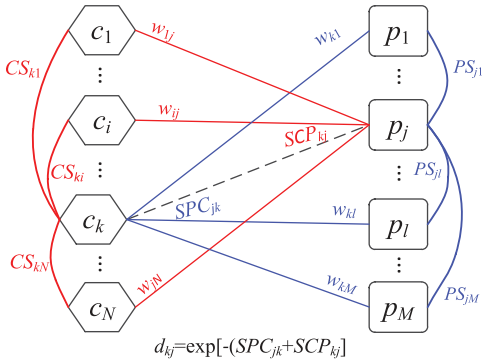

Fig. 2. Schematic diagram of calculating the score d_{kj} for a candidate compound–protein negative sample (c_k, p_j, d_{kj}). Two colors, blue and red, are used to differentiate the weights and similarities for calculating two combined scores SPC_{jk} and SCP_{kj}, respectively

Conceptually, the confidence values of the known/predicted interactions w_{kl} and w_{ij} are propagated to the candidate negative interactions via similar proteins and compounds in Step 2(a) and Step 2(b). The protein similarity linking compound c_k to protein j via protein p_l is formulated by spc_{kjl} in Step 2(a), and chemical similarity linking protein p_j to compound c_k via compound c_i is formulated by scp_{kji} in Step 2(b). In Step 2(c), the two resulting scores are combined according to Equation (3), which embodies the protein dissimilarity rule and the drug dissimilarity rules through a negative exponent function. In particular, the known/predicted interactions function as a bridge to link compounds and proteins that do not form potential interactions of high probability.

3.3 Filtering by feature divergence

It is known that compounds with similar chemical features may have greatly different binding bioactivity (activity cliff) (Sun and Bajorath, 2012). On the other hand, compounds with completely different core structures could potentially target similar proteins (scaffold hopping) (Sun *et al*., 2012). When the number of validated/predicted target proteins of a specific compound is small and thus covers limited proteomic features, it is likely that some proteins screened via proteomic feature dissimilarity based on all known target proteins are actually the targets of the compound. From the perspective of protein, similar situation maybe exists, when the number of validated/predicted compounds targeting a protein is small. Thus, we require that the number of validated/predicted interactions participated by the protein and the compound of each negative sample candidate should be larger than some predefined threshold. By setting the threshold to 15, we got 23 392 compounds and 10 757 proteins of human, 33 353 compounds and 7584 proteins of *C.elegans*, which were used to construct the negative samples.

Moreover, we expect that the features of the proteins targeted by a specific compound differ from each other as largely as possible, so that our dissimilarity rules can exclude more specious candidates that have similar features to known target proteins. Similarly, the more different the chemical features of the compounds targeting a specific protein are, the more false-positive targeting compounds could be excluded by our dissimilarity rules. In other words, the credibility of the screened negative samples is positively correlated to the feature divergence of the proteins (compounds) in validated/predicted interactions associated to a specific compound (protein). Therefore, we exploited the feature divergence to further screen the candidate negative samples. Since variance is a commonly used measure for evaluating data divergence, we carried out statistical test to check whether the similarity variance of the subset of proteins (compounds) in interactions associated to each compound (protein) in the candidate negative samples is significantly larger than the population variance. Take compounds as example, the similarity variance of the compound population is 0.0335, which can be easily computed based on CS_{ij} (see Equation 1). For a subset of n compounds interacting with a protein, the null hypothesis is that the sample variance is less than the compound population variance, and the alternate hypothesis is the opposite of the null hypothesis, then the sample variance follows χ^2 distribution with degree of freedom $n - 1$. With a significance level 0.05, we filtered out more specious candidates and finally obtained 384 916 negative samples between 14 613 unique compounds and 2229 unique proteins of human by setting the threshold of d_{kj} to 0.9 ($\approx e^{-0.1}$). For *C.elegans*, we finally got 88 261 negative samples between 2224 unique proteins and 5278 unique compounds by setting the threshold of d_{kj} to 0.368.

4 Results

4.1 Performance evaluation protocol

To conduct an objective and fair evaluation on the negative CPIs screened by our method, we first built the positive samples from the manually curated databases DrugBank and Matador and then generated two sets of negative samples: one was generated by randomly sampling compound–protein pairs not included in the positive samples, the other was extracted from the list of negative samples screened by our method. We evaluated the screened negative samples by comparing the performances of both six classical classifiers and three existing predictive methods on the same set of positive samples combining with screened and randomly generated negative samples, respectively. We selected the top 384 916 screened negative samples (the dataset is available in the Supplementary Material) from the ranking list as candidates and used some of them in the experiments.

As shown in Supplementary Figure S1, the frequency distribution of interactions is biased to only a small portion of compounds/proteins, indicating that random sampling over the whole interactions might cover only a limited number of compounds and proteins. Therefore, as in Tabei and Yamanishi (2013) and Yamanishi *et al*. (2014), we used two protocols: pairwise cross-validation and blockwise cross-validation, to evaluate our negative samples against randomly generated negative samples. Concretely, pairwise cross-validation assumes that the aim is to detect missing interactions between known ligand compounds and known target proteins with information of interaction partners, while blockwise cross-validation assumes that the goal is to detect new interactions for new ligand compounds and target proteins with no information of interaction partners. Pairwise cross-validation was performed in 3 steps: (i) the CPIs in the gold standard set are randomly split into five subsets of roughly equal size; (ii) each subset is taken in turn as a test set and the remaining four subsets are used to train a predictive model, whose prediction accuracy on the test set is then evaluated and (iii) the average prediction accuracy over the 5-folds is used as the final performance measure. Instead of splitting interactions, blockwise cross-validation randomly splits the compounds and proteins in the gold standard set into five subsets, respectively. Each compound subset and each protein subset are taken in turn and combined as a test set, and then a predictive model is trained on the compound–target pairs included in the remaining four compound subsets and four protein subsets and is further evaluated on the test set. Finally, the average prediction accuracy over the 5-folds is calculated.

Several performance measures are used in the following experiments. Denote by TP and FP the numbers of correctly and falsely predicted positive CPIs, TN and FN the numbers of correctly and falsely predicted negative CPIs, the measures are precision [or positive predictive values (PPV)] = TP/(TP + FP), recall (or sensitivity) = TP/(TP+FN), specificity = TN/(FP+TN) and AUC (area under the ROC curve). Especially, the PPV measure reflects the discriminatory power of a classifier to distinguish true positives when the number of negative samples is far larger than that of positive samples. In addition, we report the *precision–recall curve* because it is rather informative when the number of positive examples is small.

4.2 Evaluation on classical classifiers

4.2.1 Pairwise cross-validation

The human dataset that we used includes 3369 positive interactions between 1052 unique compounds and 852 unique proteins, and the *C.elegans* dataset includes 4000 positive interactions between 1434 unique compounds and 2504 unique proteins (the datasets are available in the Supplementary Material). Similar to Tabei and Yamanishi (2013), we evaluated the performance of each classifier when the ratio of negative samples to positive samples increases from 1 to 5. The randomly generated negative samples were produced by randomly sampling pairs of compound and protein not included in the positive samples. For screened negative samples, we got the required number of interactions from the top 384 916 candidates in the ranking list produced by our method. We produced the

Table 1. AUC/recall/precision values of six classical classifiers on screened and randomly generated negative samples of *human* (pairwise cross-validation)

Measure	Negative sample ratio	Naive Bayes		kNN		Random Forest		L1 logistic		L2 logistic		SVM	
		Screened	Random	Screened	Random	Screened	Random	Screened	Random	Screened	Random	Screened	Random
AUC	1	**0.672**	**0.622**	0.860	0.563	0.940	0.647	0.908	0.874	0.911	0.868	0.910	**0.752**
	3	0.672	0.622	0.904	**0.593**	0.954	0.694	**0.917**	**0.879**	**0.920**	**0.873**	0.942	0.705
	5	0.671	0.622	**0.913**	0.589	**0.967**	**0.709**	0.916	0.877	0.920	0.872	**0.951**	0.713
Precision	1	**0.624**	**0.591**	**0.798**	**0.570**	**0.861**	**0.613**	**0.881**	**0.858**	**0.891**	**0.862**	0.966	**0.733**
	3	0.361	0.338	0.716	0.458	0.847	0.529	0.823	0.786	0.837	0.787	0.969	0.700
	5	0.252	0.237	0.684	0.500	0.830	0.514	0.793	0.732	0.804	0.739	**0.969**	0.732
Recall	1	**0.575**	**0.413**	**0.927**	**0.564**	**0.897**	**0.599**	**0.893**	**0.836**	**0.913**	**0.850**	**0.950**	**0.745**
	3	0.560	0.376	0.882	0.306	0.824	0.306	0.749	0.622	0.773	0.631	0.883	0.261
	5	0.555	0.364	0.844	0.205	0.825	0.199	0.649	0.524	0.666	0.522	0.861	0.112

Bold numbers represent the highest performance measures achieved by each method.

chemogenomical features of the positive and negative samples by performing tensor product of chemical substructures and protein domains.

We conducted performance evaluation on six classical classifiers by comparing our screened negative samples against randomly generated negative samples. The six classical classifiers are naive Bayes, random forest, L1-logistic regression, L2-logistic regression, SVM and kNN. The naive Bayes, random forest and kNN were run by using Weka 3.7 (Hall *et al.*, 2009), L1- and L2-logistic regression were run by liblinear 1.94 (Fan *et al.*, 2008) and SVM was run by libsvm 3.17 (Chang and Lin, 2011). All these methods were run on default setting except for kNN where k is set to 1, 3 and 5, respectively. As similar results were obtained for different k values, we reported only the results of $k = 1$. Table 1 shows the AUC, recall and precision measures of the six classifiers on *human* data. We found that the performances of all six classifiers were significantly improved on our screened negative samples in comparison to on randomly generated negative samples. For example, for the six classifiers from naive Bayes to SVM, the average AUC improvement achieved on our screened negatives over the randomly generated negatives is 8.0%, 53.4%, 39.7%, 4.2%, 5.3% and 29.6%, respectively. When the ratio of negative samples increases, the AUC values of most classifiers keep steady or increase slightly. However, we also noticed that the recall and precision measures of most classifiers decrease with the increase of the ratio of negative samples, this is mainly due to the increasingly imbalanced ratio of the negative samples to the positive samples, which leads to the increasing bias of the classification decision boundary against the positive ones. In addition, we obtained similar results on *C.elegans*, as shown in Supplementary Table S1. These empirical results demonstrate the high reliability of our screened negative samples.

4.2.2 Blockwise cross-validation

Here the positive samples are the same as those used in pairwise cross-validation. An equal number of random negative samples to positive samples were selected by the random sampling procedure mentioned above. Also, an equal number of screened negative samples were randomly extracted from the top 384 916 candidates in our ranking list. The six classical classifiers were run in the same way as mentioned above, and the precision–recall curves and AUC histograms are shown in Figure 3 and Supplementary Figure S2. Compared with pairwise validation, the six classifiers perform worse in blockwise cross-validation on both screened and randomly generated negative samples, but their performances are still substantially improved on the screened negative samples. In particular, on randomly generated negative samples, the performance of each classifier deteriorates dramatically under blockwise cross-validation, but most classifiers except for naive Bayes and kNN still achieved relatively high AUC values on the screened negative samples: L1- and L2-logistic regression and SVM achieved AUC values larger than 0.8. As also shown in Supplementary Figure S3, all the six classifiers obtain larger AUC values on screened negative samples than on random negative samples of *C.elegans*, whereas the overall performances of these classifiers decrease slightly in comparison to that on human dataset.

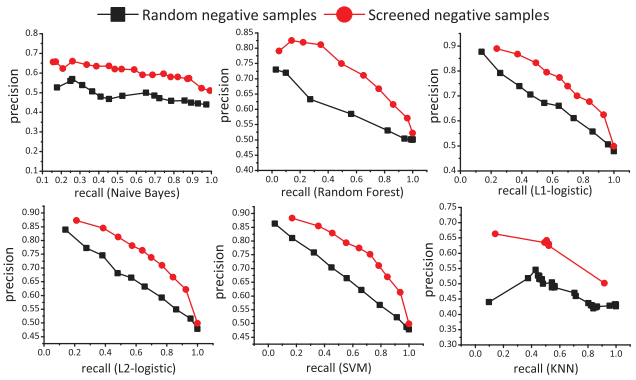

Fig. 3. Precision–recall curves of six classical classifiers on screened and randomly generated negatives of *human* (blockwise cross-validation)

4.3 Evaluation on existing predictive methods

We further checked whether existing predictive methods can achieve higher performance on screened negative samples than on randomly generated negative samples. The evaluated existing methods include BLM (Bleakley and Yamanishi, 2009), RLS-avg and RLS-Kron classifiers with GIP kernels (van Laarhoven *et al.*, 2011), KBMF2K-classification and KBMF2K-regression (Gonen, 2012). RLS-avg and RLS-Kron were run by setting two different groups of parameters, (0.5, 0.5) and (1,1), respectively, and the others were run by default settings. All these methods were originally evaluated on four widely used *human* datasets involving Enzyme, Ion Channel, GPCR and Nuclear Receptor proposed in Yamanishi *et al.* (2008). But these four datasets are small scale and cover only a small number of negative samples screened by our method, so we built another relatively larger dataset of human to evaluate these methods. We got the positive samples from DrugBank and then extracted the negative

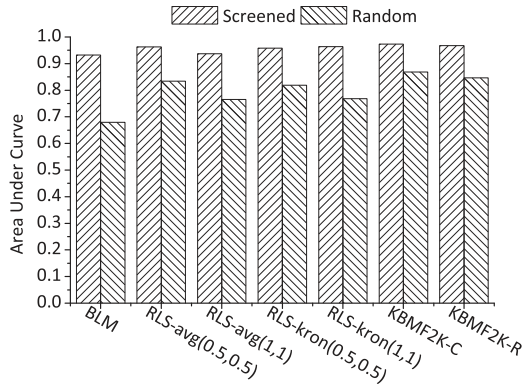

Fig. 4. Histogram of the AUC values achieved by three existing predictive methods on screened and randomly generated negative samples of *human*

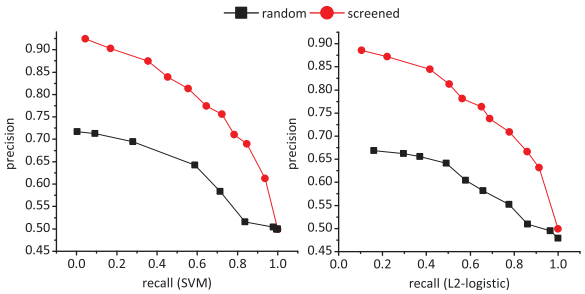

Fig. 5. Precision–recall curves of the SVM and L2-logistic classifiers trained on screened and randomly generated negative samples, evaluated on the kinase bioactivity assay data (Davis *et al.*, 2011)

samples whose compounds and proteins are both involved in the positive samples from our ranking list. The resulting human dataset includes 2315 positive interactions and 2576 negative interactions between 821 unique compounds and 846 unique proteins, and the *C.elegans* dataset includes 463 positive interactions and 1561 negative samples between 543 compounds and 504 proteins (the datasets are available in the Supplementary Material). As these compared methods take as input the chemical structure similarity matrix, protein sequence similarity matrix and CPI matrix, we built the similarity matrices and the interaction matrix as mentioned in Section 2.

The AUC values achieved by these predictive methods on screened and randomly generated negative samples of human are shown in Figure 4. Clearly, all the methods achieved significantly higher performance on the screened negative samples than on the randomly generated negative samples. In particular, BLM had the least AUC (0.679) on the randomly generated negative samples but achieved a comparable AUC (0.932) to other methods on the screened negative samples. KBMF2K-classification and KBMF2K-regression had considerably high AUCs (0.868 and 0.846) on the randomly generated negative samples, but their performances were also significantly improved on the screened negative samples. On *C.elegans*, the performance improvement is more notable than on human for all methods except for KBMF2K-classification and KBMF2K-regression, as shown in Supplementary Figure S4. Although BLM and the four RLS algorithms performed only moderately on the randomly generated negative samples, their performances were substantially boosted on the screened negative samples. This result shows again that our screened negative samples are helpful for improving the performances of existing predictive methods.

4.4 Evaluation on drug bioactivity dataset

The quantitative drug–target bioactivity assays for kinase inhibitors provide experimental observations of the bindings of drug molecules to targets, which enable us to derive both positive and negative interactions. As suggested by Pahikkala *et al.* (2015), recent kinase bioactivity assay data from Davis *et al.* (2011) can be used as an independent benchmark test set for performance evaluation of drug–target prediction methods. This assay reported the quantitative interaction affinity as the dissociation constant (K_d), which reflects how tightly a drug molecular binds to a target protein. The smaller K_d is, the higher the interaction affinity between the chemical compound and the target protein is.

The bioactivity assay included the interactions between 68 unique drugs and 442 unique proteins, from which 20 931 interactions with $K_d \geq 10\,000$ were extracted as negative samples. We got 3564 overlapping interactions between our screened negative samples and the experimentally supported ones. We calculated the frequency distribution of overlapping negative samples with respect to an increasing cutoff threshold of the confidence scores of our screened negative samples, which is shown in Supplementary Figure S5. It can be seen that the confidence scores of about 80% overlapping negative samples are more than 0.5, i.e. the screened negative samples with larger confidence scores are more likely supported by drug bioactivity experiments, which indicates the high credibility of our screened negative samples. Furthermore, we used the threshold 30.00nM of K_d suggested by Davis *et al.* (2011) to extract positive samples and obtained 1867 positive interactions. Together with the same number of negative interactions, we got an independent test set, which was used to evaluate the SVM and L2-logistic classifiers trained on screened and randomly generated negative samples, respectively. We chose SVM and L2-logistic regression for performance evaluation because they are, respectively, binary classification and realistic regression representatives. Figure 5 presents their precision–recall curves, which show that the classifiers trained on our screened negative samples greatly outperform those trained on the random negative samples.

4.5 Prediction of new interactions

After confirming the quality of our screened negative samples, we built two sets of predictions of potential CPIs on *human*. The first is a relatively small-scale prediction set built based on a subset of compounds and proteins included in DrugBank. Specifically, we extracted 2675 interactions from DrugBank as positive samples and select an equal number of negative samples from our screened ranking list and then train an SVM classifier based on the chemogenomical features to predict potential interactions. The trained SVM classifier predicted about 390 838 new CPIs from all possible 896 304 interactions whose compounds and proteins are included in DrugBank. We extracted the top 50 interactions for each compound to get a set of 35 425 interactions, in which 1093 predictions were annotated in DrugBank and 3224 predictions (3224/35 425≈9.2%) were annotated in STITCH. Note that only 18 580 interactions are recorded in STITCH for all possible 896 304 (18 580/896 304≈2.1%), thus our predictions rank these curated interactions high and give priority to highly credible interactions.

As a confirmative example, we examined the predicted interactions regarding *Donepezil*, a centrally acting reversible acetylcholinesterase inhibitor compound that is therapeutically used in the palliative treatment of Alzheimer's disease (Birks and Harvey, 2006). Our method predicted 253 target proteins that include

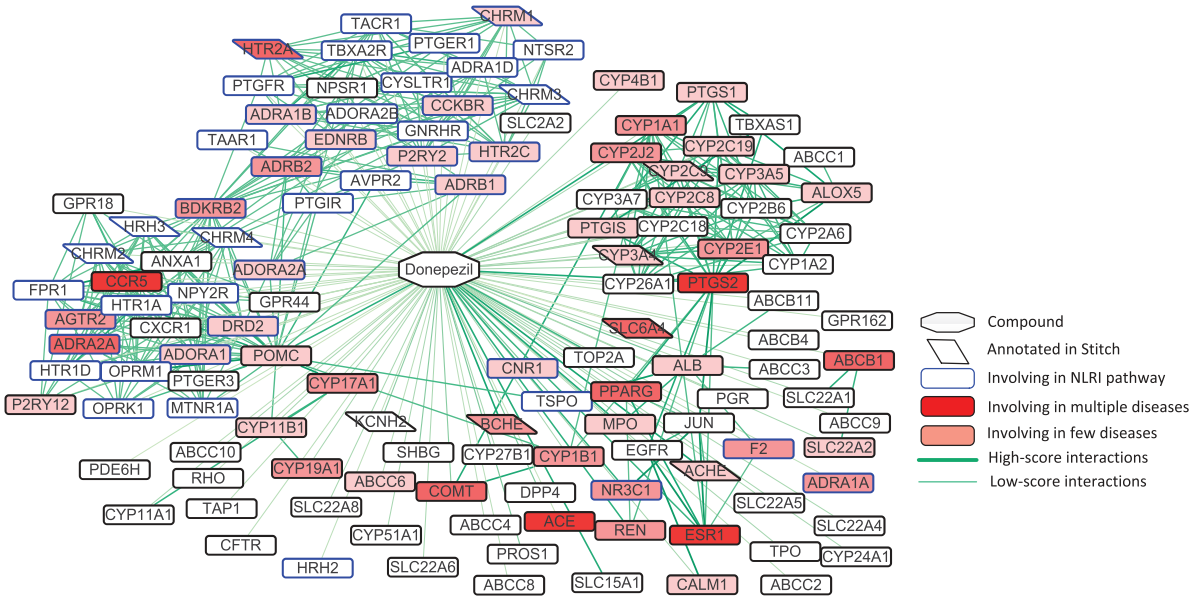

Fig. 6. Predicted target proteins of *Donepezil* and related functional annotations, including neuroactive ligand-receptor interaction pathways, diseases recorded in DrugBank and STITCH

the cholinesterase coding genes ACHE and BCHE, which are two main targets of Donepezil annotated in DrugBank and STITCH. In fact, the set of target proteins covers all 17 interactions annotated in STITCH whose associated proteins are included in the test set, as shown in Figure 6. To confirm other new predictions, we inspected the functional annotations of the top 125 target proteins by using DAVID (Huang *et al.*, 2009) and found that 45 proteins are highly enriched in the neuroactive ligand-receptor interaction pathway (P value = 4.7E-32). These proteins are closely related to many diseases including multiple kinds of psychotic disorders. Furthermore, these proteins are significantly associated in various mental and nervous diseases, such as hypertension (P value = 1.5E-11), Alzheimer's disease (P value = 3.1E-4), Parkinson's disease (P value = 3.0E-4) and arteriosclerotic vascular disease (P value = 1.3E-3). Figure 6 gives an illustration of Donepezil's target proteins and related functional annotations (for the detailed list of proteins involved in the pathways and diseases, please see the Supplementary Table S2).

In addition, to facilitate the research community, we built the second set (a large-scale one) of predictions by constructing a large training set that consists of all 6354 interactions included in DrugBank and the equal number of screened negative samples. The trained SVM classifier predicts more than 6 340 000 CPIs (please refer to the Supplementary Material for detail). These new predictions would be helpful for identifying truly druggable targets in new drug design.

5 Discussion and conclusion

The identification of interactions between compounds and proteins plays an important role in the genomic drug discovery. However, experimental validation of CPIs is still laborious and expensive, although various high-throughput biochemical assays are available. *In silico* prediction methods are appealing to guide experimental design and to provide supporting evidence for the experimental results. Methods based on machine learning have been proposed and demonstrated encouraging performance. However, their performance and robustness depend on the training set in which negative samples have equal importance to positive samples. Unfortunately, our knowledge of negative samples of CPIs is extreme limited which restricts severely the performance of computational methods. This problem motivated us to propose a systematic screening workflow to identify reliable negative CPIs. To the best of our knowledge, this is the first work devoted to screen reliable negative samples of CPIs.

Our screening framework is based on the assumption that the proteins dissimilar to any known/predicted target of a given compound are not much likely to be targeted by the compound and vice versa. In the view of chemogenomical space, we managed to find those compound–protein pairs that locate far from all positive samples in the chemogenomical space as negative samples, which really contributed to the performance improvement of both classical classifiers and existing computational methods. Furthermore, the compounds and proteins associated with a small number of known interactions were excluded to reduce the possibility of taking real interactions as negative interactions due to activity cliff and scaffold hopping. The feature divergence filtering further consolidated the strength of our dissimilarity rules. Extensive experiments demonstrated that our screened negative samples are highly credible and helpful for identifying CPIs.

On the basis of the screened negative samples and positive samples obtained from DrugBank, we carried out prediction of potential CPIs on *human* and *C.elegans* by training SVM classifiers on the chemogenomical features. Also, we gave a confirmative example that the newly predicted target proteins of Donepezil are highly enriched in mental and nervous pathways and diseases. In summary, our screened negative samples and predictions provide the research community with a useful resource for identifying drug targets and constitute a helpful supplement to the current curated compound–protein databases.

Funding

This work was supported by the National Natural Science Foundation of China under grant no. 31300707, no. 61272380 and no. 61173118, China, and MOE AcRF Tier 2 grant ARC 39/13 (MOE2013-T2-1-079), Ministry of Education, Singapore.

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.

Ashburner,M. *et al.* (2013) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Birks,J. and Harvey, R.J. (2006) Donepezil for dementia due to Alzheimer's disease. *Cochrane Database Syst. Rev.*, **1**, CD001190.

Bleakley,K. and Yamanishi,Y. (2009) Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics*, **25**, 2397–2403.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Carrella,D. *et al.* (2014) Mantra 2.0: an online collaborative resource for drug mode of action and repurposing by network analysis. *Bioinformatics*, **30**, 1787–1788.

Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 1–27.

Chen,H. and Zhang,Z. (2013) A semi-supervised method for drug-target interaction prediction with consistency in networks. *PLoS One*, **8**, e62975.

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

Coutoa,F.M. *et al.* (2007) Measuring semantic similarity between Gene Ontology terms. *Data Knowl. Eng.*, **61**, 137–152.

Davis,M.I. *et al.* (2011) Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.*, **29**, 1046–1051.

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

Fan,R.E. *et al.* (2008) LIBLINEAR: a library for large linear classification. *J. Machine Learning Res.*, **9**, 1871–1874.

Franceschini,A. *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.*, **41**(Database issue), D808–D815.

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

Gottlieb,A. *et al.* (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.*, **7**, 496.

Gottlieb,A. *et al.* (2012) INDI: a computational framework for inferring drug interactions and their associated recommendations. *Mol. Syst. biol.*, **8**, 592.

Gnther,S. *et al.* (2008) SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**, D919–D922.

Hall,M. *et al.* (2009) The WEKA data mining software: an update. *SIGKDD Explor. Newsl.*, **11**, 10–18.

He,Z. *et al.* (2010) Predicting drug-target interaction networks based on functional groups and biological features. *PloS One*, **5**, e9603.

Hu,Y. and Bajorath,J. (2012) Extending the activity cliff concept: structural categorization of activity cliffs and systematic identification of different types of cliffs in the ChEMBL database. *J. Chem. Inf. Model.*, **52**, 1806–1811.

Huang,D.W. *et al.* (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Iorio,F. *et al.* (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proc. Natl. Acad. Sci. USA*, **107**, 14621–14626.

Jaccard,P. (1908) Nouvelles recherches sur la distribution florale. *Bul. Soc. Vaudoise Sci. Nat.*, **44**, 223–270.

Jacob,L. and Vert,J.P. (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics*, **24**, 2149–2156.

Jaeger,S. *et al.* (2014) Causal network models for predicting compound targets and driving pathways in cancer. *J. Biomol. Screen.*, **19**, 791–802.

Jaroch,S.E. and Weinmann,H. (eds) (2006) *Chemical genomics: small molecule probes to study cellular function*. Ernst Schering Research Foundation Workshop. Springer, Berlin, pp. 1–20.

Kuhn,M. *et al.* (2010) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol*, **6**, 343.

Kuhn,M. *et al.* (2014) STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.*, **42**(D1), D401–D407.

Lamb,J. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Mei,J.P. *et al.* (2013) Drug-target interaction prediction by learning from local information and neighbors. *Bioinformatics*, **29**, 238–245.

Metz,J.T. *et al.* (2011) Navigating the kinome. *Nat. Chem. Biol.*, **7**, 200–202.

Mizutani,S. *et al.* (2012) Relating drug-protein interaction network with drug side effects. *Bioinformatics*, **28**, i522–i528.

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

Pauwels,E. *et al.* (2011) Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics*, **12**, 169.

Perlman,L. *et al.* (2011) Combining drug and gene similarity measures for drug-target elucidation. *J. Comput. Biol.*, **18**, 133–145.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**(Database issue), D290–D301.

Smith,T.F. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.

Sun,H. *et al.* (2012) Classification of scaffold-hopping approaches. *Drug. Discov. Today*, **17**, 44–57.

Tabei,Y. and Yamanishi,Y. (2013) Scalable prediction of compound-protein interactions using minwise hashing. *BMC Syst. Biol.*, **7**, S3.

van Laarhoven,T. and Marchiori,E. (2013) Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. *PLoS One*, **8**, e66952.

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

Wang,Y. and Zeng,J. (2013) Predicting drug-target interactions using restricted Boltzmann machines. *Bioinformatics*, **29**, i126–i134.

Wang,Y.C. *et al.* (2011) Kernel-based data fusion improves the drug-protein interaction prediction. *Comput. Biol Chem.*, **35**, 353–362.

Wheeler,D.L. *et al.* (2006) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **34**, D173–D180.

Wishart,D.S. *et al.* (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.

Wolpaw,A.J. *et al.* (2011) Modulatory profiling identifies mechanisms of small molecule-induced cell death. *Proc. Natl. Acad. Sci. USA*, **108**, E771–E780.

Xia,Z. *et al.* (2010) Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.*, **4**(Suppl 2), S6.

Yabuuchi,H. *et al.* (2011) Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol. Syst. Biol.*, **7**, 472.

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

Yamanishi,Y. *et al.* (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics*, **26**, i246–i254.

Yamanishi,Y. *et al.* (2014) DINIES: drug-target interaction network inference engine based on supervised analysis. *Nucleic Acids Res.*, **42**, W39–W45.

Zhou,T. *et al.* (2010) Bipartite network projection and personal recommendation. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **76**, 046115.