Bioinformatics, 32(2), 2016, 226–234
doi: 10.1093/bioinformatics/btv528
Advance Access Publication Date: 28 September 2015
Original Paper

OXFORD

Systems biology

Computational probing protein–protein interactions targeting small molecules

Yong-Cui Wang[1,*], Shi-Long Chen[1], Nai-Yang Deng[2] and Yong Wang[3,*]

[1]Key Laboratory of Adaptation and Evolution of Plateau Biota, Northwest Institute of Plateau Biology, Chinese Academy of Sciences, Xining 810001, China, [2]College of Science, China Agricultural University, Beijing 100083, China and [3]National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, 100190, China

*To whom correspondence should be addressed
Associate Editor: Burkhard Rost

Received on March 2, 2015; revised on August 4, 2015; accepted on August 31, 2015

Abstract

Motivation: With the booming of interactome studies, a lot of interactions can be measured in a high throughput way and large scale datasets are available. It is becoming apparent that many different types of interactions can be potential drug targets. Compared with inhibition of a single protein, inhibition of protein–protein interaction (PPI) is promising to improve the specificity with fewer adverse side-effects. Also it greatly broadens the drug target search space, which makes the drug target discovery difficult. Computational methods are highly desired to efficiently provide candidates for further experiments and hold the promise to greatly accelerate the discovery of novel drug targets.

Results: Here, we propose a machine learning method to predict PPI targets in a genomic-wide scale. Specifically, we develop a computational method, named as PrePPItar, to Predict PPIs as drug targets by uncovering the potential associations between drugs and PPIs. First, we survey the databases and manually construct a gold-standard positive dataset for drug and PPI interactions. This effort leads to a dataset with 227 associations among 63 PPIs and 113 FDA-approved drugs and allows us to build models to learn the association rules from the data. Second, we characterize drugs by profiling in chemical structure, drug ATC-code annotation, and side-effect space and represent PPI similarity by a symmetrical S-kernel based on protein amino acid sequence. Then the drugs and PPIs are correlated by Kronecker product kernel. Finally, a support vector machine (SVM), is trained to predict novel associations between drugs and PPIs. We validate our PrePPItar method on the well-established gold-standard dataset by cross-validation. We find that all chemical structure, drug ATC-code, and side-effect information are predictive for PPI target. Moreover, we can increase the PPI target prediction coverage by integrating multiple data sources. Follow-up database search and pathway analysis indicate that our new predictions are worthy of future experimental validation.

Conclusion: In conclusion, PrePPItar can serve as a useful tool for PPI target discovery and provides a general heterogeneous data integrative framework.

Availability and implementation: PrePPItar is available at http://doc.aporc.org/wiki/PrePPItar.

Contact: ycwang@nwipb.cas.cn or ywang@amss.ac.cn

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

1 Introduction

Systems biology emphasizes to understand the complex biological systems from interactions beyond the single molecules. This paradigm shift pushes the generations of large sets of interactions called interactome. One typical example is to measure protein–protein interaction by yeast-two-hybrid and mass spectrometry. Protein–protein interactions (PPIs) are building blocks for the majority of biological processes in the living cell. Beyond the single protein, they are natural to serves as the basic elements to be perturbed by drugs in the treatment, cure, prevention, or diagnosis of disease. Indeed, targeting PPIs with small molecular inhibitors is of increasing interests both in academia as well as in the pharmaceutical industry (Arkin and Wells, 2004). Many different types of interactions can be inhibited using drug-like small molecules (Archakov *et al.*, 2003; Pawson and Nash, 2003; Wells and McClendon, 2007). Searching potential drug PPI targets can serve both therapeutic purpose and useful chemical tool for basic science (Valkov *et al.*, 2012). Compared with inhibition of a single protein, inhibition of PPIs is promising to broaden the drug target search space. Moreover, drugs targeting such interactions are more specific and likely to act with fewer side-effects than conventional medication influencing whole cell functions (Klussmann and Scott, 2008). However, it's challenging to find PPI target by wet experiment since it suffers from the dynamics of PPIs and the particular problem of more exposed and less defined binding sites (Valkov *et al.*, 2012). Computational methods are then highly desired to provide candidates for further experimental verification and then accelerate the mapping of PPI targets on a large scale.

The initial challenge is the discovery of specific PPIs and in turn identification of those molecules that are 'druggable' (White *et al*, 2008). To address this issue, computational techniques were developed to identify small molecules as PPI inhibitors through a rationalization of the PPI inhibitor chemical space (Arnout *et al.*, 2013; Neugebauer *et al.*, 2007; Reynès *et al.*, 2010; Villoutreix, 2014) and the design of PPI-focused compound libraries (Basse *et al.*, 2013; Hamon *et al.*, 2013; Labbè *et al.*, 2013), which deposited the structures of PPIs as protein complexes and their ligands. Following the identification of a suitable PPI, the next step involves study of the binding interface and ligand design. Within a binding interface, only a small number of highly conserved amino acid residues: 'hot spots', are crucial for the interaction. Thus the key problem in PPI inhibitors discovery becomes looking for the location of 'hot spots' from large surface area of the PPI interface. Previous work mainly focused on fragment-based methods, which attempted to discover the compounds that are druggable and ligandable by matching three-dimensional structure of molecules with the surface area of PPI interface (Valkov *et al.*, 2012). Structure information is required in fragment-based methods, however, only a small fraction of proteins have well-established structures, thus this class of methods is restricted in small-scale.

With the rapid development of high-throughput experimental methodologies, genome-wide PPIs in some model organisms, including *Escherichia coli* (Butland *et al.*, 2005), *Helicobacter pylori* (Rain *et al.*, 2001), *Saccharomyces cerevisiae* (Gavin, 2002; Uetz *et al.*, 2000; Ito *et al.*, 2001), *Caenorhabditis elegans* (Li, 2004), *Drosophila melanogaster* (Giot *et al.*, 2003), and *Homo sapiens* (Stelzl *et al.*, 2005; Rual, 2005), have been established. Meanwhile, some curated databases deposit high quality drug targets data, including Kyoto Encyclopedia of Genes and Genomes (KEGG), Biomolecular Relations in Information Transmission and Expression (BRITE) (Kanehisa *et al.*, 2006), BRENDA (Schomburg *et al.*, 2004), SuperTarget (Günther *et al.*, 2008), and DrugBank (Wishart *et al.*, 2008). These valuable information sources together provide a great opportunity to understand the mechanism of PPI targets from the viewpoint of interactome. That is, drug-PPI associations could be constructed based on genome-wide PPIs and high quality drug targets data. Most importantly, the understandable rules for drug-PPI associations can be learned by a statistical predictor based on these associations.

Here, we develop a machine learning framework to **Pr**edict **PPI**s **tar**gets (PrePPItar) by dissecting the drug-PPI associations in a large-scale manner. We observed that the current available data sources describe drug's biological function in living cell from different levels and different aspects. For example, drug's chemical structure provides information by the 'structure determines function' paradigm. ATC-code annotation provides the therapeutic effect at molecular level, and side-effect hints the unwanted effect at phenotype level. One straightforward assumption is that drugs similar in one or more data source metrics will interact with similar PPIs. We demonstrate that drugs with similar chemical structures, ATC-codes, or side-effects indeed associate with similar PPIs. Then we propose the idea to integrate heterogeneous chemical structure, ATC-code, and side-effect information sources. Specifically, drug and PPI are first characterized by their similarity-based profiles, and a kernel function is then defined to correlate them. Finally, the potential drug-PPI associations are inferred by training a machine learning model, i.e. support vector machines (SVM), which is motivated by statistical learning theory (Vapnik, 1995, 1998) and has been successful on many different classification problems in bioinformatics (Schölkopf *et al.*, 2004). PrePPItar overcomes the main difficulty to integrate these data sources from structure, ATC classification, and side-effect level, which are highly heterogenous.

To make the learning feasible and validate PrePPItar by cross validation, we construct a well-established dataset from scratch. In total, 227 associations among 63 PPIs and 113 FDA-approved drugs are collected from HPRD (Prasad *et al.*, 2009) and DrugBank (Wishart *et al.*, 2008). We find that all chemical structure, ATC-code annotation, and side-effect information are predictive in different ways. Moreover, drug-PPI associations can be uncovered by combination of these three properties. In addition, database search and pathway analysis indicate that our new predictions are worthy of future experimental validation. Compared with single protein target prediction, PPI target prediction can discover some dissimilar targets, that is, it is promising to broaden the drug target search space.

2 Materials and methods

2.1 Methods

We propose a novel algorithm, PrePPItar, to predict PPI targets by identifying the potential associations between drugs and PPIs based on kernel fusion of heterogenous data sources. The schematic illustration of PrePPItar is shown in Figure 1A. The functional role of drug is characterized by its molecular structure, molecular function, and phenotype data. PrePPItar aims to optimally integrate those three data sources and connect drug with PPI more accurately.

The PPI target prediction problem is converted into an interaction prediction problem: determining whether a give pair of drug-PPI is associated or not. We introduce SVM-based algorithm to cope with this prediction task. The algorithm works in three phases (Fig. 1): (B) Collecting known drug targeting PPIs as gold-standard positives in a bipartite graph. (C) Modelling drug-drug and PPI-PPI similarity metrics. Drug similarity is derived from chemical structures, ATC-codes, and side-effects. PPI similarity is calculated

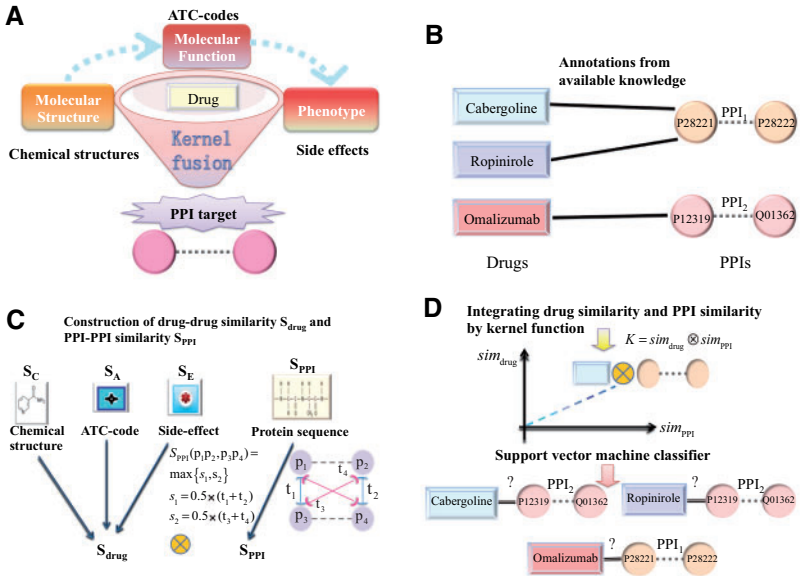

Fig. 1. The flowchart of PrePPITar. (**A**) The schematic plot for our PrePPItar method. PrePPItar applies the kernel fusion method to integrate multiple information about drug, including chemical structure, ATC-code, and drug side-effect to detect the interactions between drugs and PPIs. (**B**) Collecting known associations between drugs and PPIs as gold standard positives in a bipartite graph. (**C**) Calculating drug-drug and PPI-PPI similarity metrics. Where, $t_i, i = 1, 2, 3, 4$, are the sequence similarity among proteins. (**D**) Relating the similarity among drugs and similarity among PPIs by Kronecker product kernel, and applying SVM-based algorithm to predict the unknown associations between drugs and PPIs

by a symmetrical S-kernel based on protein amino acid sequence (Shen *et al.*, 2007). (D) Correlating the similarity among drugs and similarity among PPIs by Kronecker product kernel, and applying SVM algorithm to predict the unknown associations between drugs and PPIs. To implement the SVM-based algorithm, the kernel function and standard training dataset are needed.

Given two drug-PPI pairs, we consider to construct a kernel function which potentially correlates with their similarity. Since the kernel function represents the similarities among the training samples in some sense (Hofmann *et al.*, 2008), we put more efforts on the similarity among drugs and similarity among PPIs rather than on the representing profiles for drugs and PPIs.

2.2 Drug similarity

The chemical structures, drug ATC-codes, and side-effects are used to represent the similarity among drugs, respectively. S_C is constructed to represent drug chemical structure similarity. Each row (or column) of this matrix is the chemical structure similarity profile for a single drug, where the chemical structure similarity between two drugs d and d' is computed by a graph-based method for comparing pairwise chemical structures, SIMilar COMPound (SIMCOMP) (Hattori *et al.*, 2003), which is widely used in single protein target prediction (Yamanishi *et al.*, 2008; Wang *et al.*, 2010, 2011).

S_A is utilized to denote the drug therapeutic similarity matrix. Each row (or column) of this matrix is the therapeutic similarity profile for a single drug, where the ATC-code based similarity between drugs d and d' is calculated as following:

$$S_A(d, d') = \max_{t_i \in T(d), t_j \in T(d')} sim(t_i, t_j), \quad (1)$$

where $T(d)$ and $T(d')$ are the sets of ATC-code annotations of corresponding drugs. The similarity between two ATC-codes (t_i and t_j) is calculated by the equation of $sim(t_i, t_j) = w(t_i)w(t_j)\exp(-\gamma d(t_i, t_j))$, where $d(t_i, t_j)$ is the shortest distance between two ATC-codes t_i and t_j in the hierarchical structure of the ATC classification system. $w(t_i)$ and $w(t_j)$ represent the weights of the corresponding ATC-codes, and are defined as the inverse of ATC-code frequencies, which means that more emphasis was put on the specific codes rather than the general ones (Yamanishi *et al.*, 2010). γ is a predefined parameter (set to be 0.25 in this study).

The matrix S_E is applied to represent the drug similarity matrix under their side-effect measurement. Each row (or column) is the side-effect based similarity profile for a single drug. The drug similarity under their side-effects measurement is defined as the weighted cosine correlation coefficient: $S_E(d, d') = \frac{\sum_{k=1}^{M} w_k z_k z'_k}{\sqrt{\sum_{k=1}^{M} w_k z_k^2}\sqrt{\sum_{k=1}^{M} w_k z'^2_k}}$, where z and z' are binary vectors for d and d', representing the presence or absence of corresponding side-effect in SIDER database (http://sideeffects.embl.de/). w_k is the weight function for the kth side-effect defined as $w_k = exp(-f_k^2/\sigma^2 h^2)$, where f_k is the frequency of the kth side-effect in the data, and M is the total number of side-effects, σ is the standard derivation of $\{f_k\}_{k=1}^{M}$, and h is a constant (set to 10 in this study).

2.3 PPI similarity

Similar to drug similarity profiles, we also apply similarity profile to represent PPIs. We notice the fact that PPI is symmetrical, i.e. the similarity between PPIs is independent to their proteins' order. Therefore, a symmetrical S-kernel function, which was used in human PPI prediction (Shen *et al.*, 2007), is introduced here to measure PPIs' similarity. Specifically, the similarity between PPI_1 and PPI_2 is calculated by the following:

$$S_{PPI}(PPI_1, PPI_2) = \max\{s_1, s_2\}, \quad (2)$$

where $s_1 = \frac{S_q(p_1,p_2)+S_q(p'_1,p'_2)}{2}$, $s_2 = \frac{S_q(p_1,p'_2)+S_q(p'_1,p_2)}{2}$. The sequence data is used to measure protein similarity due to the rapidly developed sequencing techniques. The sequence similarities S_q among the proteins are defined by a normalized version of Smith–Waterman scores (Smith and Waterman, 1981). They are calculated by 'swalign' function in Matlab Bioinformatics toolbox. The matrix S_{PPI} is then

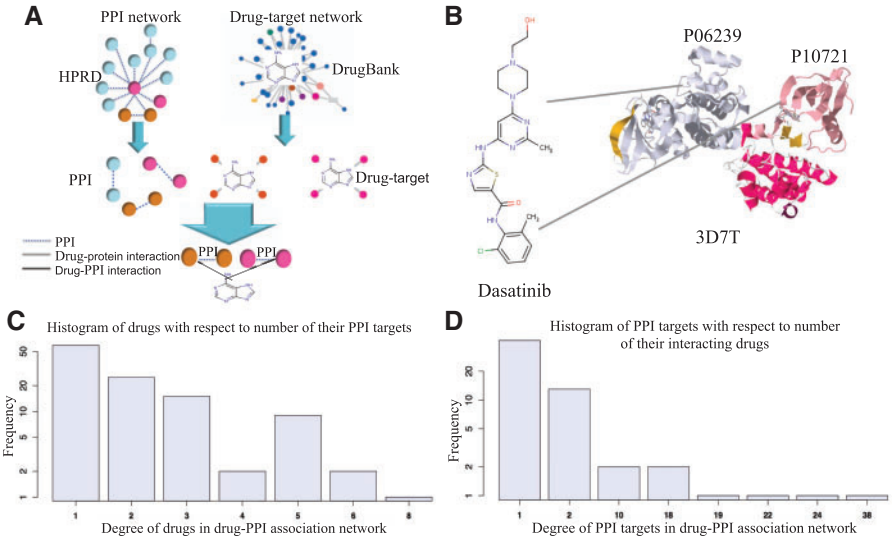

Fig. 2. The properties of gold-standard dataset. (A) The scheme of gold-standard dataset generation. First, collecting protein targets form DrugBank; Second, mapping proteins targeting the same drug into PPI network; Finally, generating PPI targets by selecting the interacting targets in PPI network. (B) an example of drug-PPI target: drug Dasatinib targets PPI: P06239–P10721, and protein P06239 (left part of 3D7T) and P10721 (up-right part of 3D7T) are in the same complex: 3D7T. (C) The histogram of drugs with respect to their PPI targets in drug-PPI target network. Drug Muromonab has 8 PPI targets. (D) The histogram of PPI targets with respect to their interacting drugs in drug-PPI target network. The most-targeted PPIs are P35348–P35368, P28221–P28222, P35348–P25100 and P12314–P08637

utilized to represent the PPI similarity. Each row (or column) is the similarity profile for a single PPI.

2.4 The kernel function and drug-PPI association prediction

With the representation of drugs and PPIs by their similarity profiles, the kernel function with drug-PPI pairs can be calculated as Kronecker product kernel (Basilico and Hofmann, 2004; Ben-Hur and Noble, 2005; Hue and Vert, 2010; Oyama and Manning, 2004):

$$K_{drug-PPI} = S_{drug} \otimes S_{PPI}, \tag{3}$$

where S_{drug} can be any one of S_C, S_A and S_E or their combination. In this paper, 'chem' denotes the case when $S_{drug} = S_C$, 'ATC-code' denotes the case when $S_{drug} = S_A$, 'side-effect' denotes the case when $S_{drug} = S_E$, and 'comb' denotes the case when $S_{drug} = \max(S_C, S_A, S_E)$, which means drug similar in one or more than one metrics will target similar PPIs. Taken together, the rationale behind our kernel function construction scheme for drug-PPI pairs is that two drug-PPI pairs are similar only when the corresponding compounds and PPIs are simultaneously similar supported by different lines of evidences.

With the above kernel function construction scheme, the PPI target prediction task is ready to be formalized as a binary classification problem with the kernel function feeding to SVM. As the kernel function (3) is feeded to the SVM learning scheme, the predictor can be calculated by the equation of $f(\boldsymbol{x}) = \mathrm{sgn}(\sum_{i=1}^{l} \alpha_i^* y_i K_{drug-PPI}(\boldsymbol{x}_i, \boldsymbol{x}) + b^*)$, where α^* is the solution of the following optimization problem

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K_{drug-PPI}(\boldsymbol{x}_i, \boldsymbol{x}_j) - \sum_{j=1}^{l} \alpha_j, \tag{4}$$

$$\text{s.t.} \sum_{i=1}^{l} y_i \alpha_i = 0, 0 \le \alpha_i \le C, \; i = 1, \ldots, l, \tag{5}$$

and b^* can be obtained as follows. If there exists $\alpha_j^* \in (0, C)$, $j = 1, \cdots, l$, then $b^* = y_j - \sum_{i=1}^{l} y_i \alpha_i^* K_{drug-PPI}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Here $\boldsymbol{x}$ represent a given drug-PPI pair.

If we treat the drug-PPI pairs with associations publicly available as the positives and the others as the negatives, the training data imbalance problem will arise. Because there are only a relatively small number of known drug-PPI associations. This situation will make the SVM ineffective in determining the class boundary (Wu and Chang, 2003). To maintain a balance between training positives and negatives in SVM training procedure, we randomly select a negative dataset from the unlabeled data (unknown drug-PPI pairs) to make sure that it has the same size with the training positive dataset.

3 Results

3.1 Gold-standard drug-PPI network

Here, we aims to discover novel associations between drugs and PPIs from the data. To this end, we first prepare high quality drug-PPI associations as gold-standard dataset, which can be used to train the model and to validate the performance of the prediction algorithms as a community standard. Specifically, we collect protein targets from DrugBank (Wishart *et al.*, 2008), which contains 5 394 drug-target interactions for 1 417 FDA-approved small molecule compounds. Then we map protein targets which interact with the same drugs into human PPI networks from HPRD (Prasad *et al*, 2009). Finally, by keeping the PPIs associated with the same drug, we got a drug-PPI association network with 227 associations among 63 PPIs and 113 FDA-approved drugs (Supplementary Fig. S2A). The procedure to generate above drug-PPI association network is shown in Figure 2A. An example of PPI target for Dasatinib is drawn in Figure 2B, it shows PPI: P06239–P10721 as target of Dasatinib, due to that protein P06239 and P10721 are in the same complex: 3D7T (Krysta *et al.*, 1998). The histogram of drugs with respect to their PPI targets, and the histogram of PPI targets with respect to their interacting drugs are shown in Figure 2C and D, respectively. Figure 2C shows that, most drugs target only a few PPIs, but some have many PPI targets. For example: Muromonab has 8 PPI targets while most PPIs just interact with a few number of drugs (Fig. 2D). The most-targeted PPIs are P35348–P35368, P28221–P28222, P35348–P25100 and P12314–P08637. Comparing with other targets

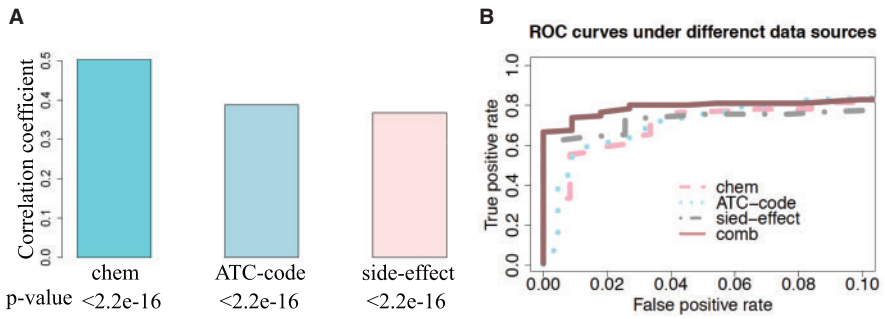

Fig. 3. (A) Barplot of the Pearson's correlation coefficients (PCCs) between chemical structure, ATC-code, and side-effect similarity and PPI similarity. The *P*-values are less than 1e − 16 for all of them. This suggests that all chemical structure, ATC-code annotation, and side-effect similarity correlate with PPI similarity, and PPI similarity is more correlated with drug chemical structure similarity. (B) The ROC curves with FPR less than 0.02. It shows that, 'comb' obtains the highest TPR when FPR is very small

associated with the same drug, PPI targets have higher sequence similarity (Supplementary Fig. S2B). Furthermore 33 PPIs participate in the same IPA canonical pathway (Supplementary Table S5).

We also analyze the existing drug-protein network to see what kind of drugs prefer to target a PPI. Supplementary Fig. S1A and S1B show that, drugs can be divided into two types based on the distance among their interacting proteins in PPI network: Type I drugs tend to target PPI or protein cluster (small average distance of proteins in PPI network) and Type II drugs tend to associate with isolated protein targets. Although a lots of drugs have more than two protein targets, only a small fraction of them which have higher molecular weight and exact mass prefer to bind PPI (Supplementary Table S1). That is, Type I drugs belong to minor population, and PPI targets are only a small fraction of drug protein targets (Supplementary Fig. S1C). This result implies that proteins interacting with the same small molecules tend to have different functions. Just like that the primary target (on-target) results in a therapeutic benefit to the patient, additional target (off-target) may lead to undesirable side effects. Although they interact with the same drug, their functions are different.

3.2 Benchmark datasets and SVM implementation

The lists of the above gold-standard dataset are shown in Supplementary Tables S4–S6. It serves as the benchmark dataset to validate the performance of PrePPItar. It includes 227 drug-PPI associations among 63 PPIs with annotations in HPRD, PDB complex and IPA canonical pathway (Supplementary Table S5) and 113 FDA-approved drugs with their ATC-codes and associated diseases (Supplementary Table S6). Drug targets and target sequences are extracted from DrugBank (Wishart *et al.*, 2008). ATC-codes are extracted from KEGG BRITE (Kanehisa *et al.*, 2006) and DrugBank (Wishart *et al.*, 2008). Drug side-effects are extracted from SIDER (Kuhn *et al.*, 2010).

We train the SVM-based predictor by using *LibSVM* (Chang and Lin, 2011). In our implementation, the penalty parameter C is optimized by a grid search approach with 3-fold cross-validation, and the optimal value of C is 1. To evaluate the performance of our methods, 10-fold cross-validation is introduced. The performance of our proposed method is shown by receiver operating characteristic (ROC) curve (Gribskov and Robinson, 1996), which shows the trade-off between the true positive (correctly predicted associations) rate (TPR) with respect to the false positive (wrongly predicted associations) rate (FPR). Furthermore, considering the complementary role of AUPR (area under precision-recall curve) to AUC in performance evaluation, AUPR is also used to evaluate the performance of PrePPItar. The evaluation criteria used to assess the performance of PrePPItar are shown in Supplementary Table S2.

3.3 Correlation analysis shows chemical structure, drug-target interactions and side-effects are all predictive

Our prediction algorithm is based on the assumption that drugs similar in some metrics will have the same PPI target. First, we validate whether this assumption works or not. Here, chemical structure, ATC-code annotation, and side-effect are utilized to measure drugs' similarity. Thus we check if each similar measurement correlates with PPI similarity, that is drugs with similar structures (ATC-code annotations or side-effects) will target similar PPIs. To this end, we correlate chemical structure, ATC-code annotation, and drug side-effect similarity with PPI similarity, respectively. Pearson's correlation coefficients (PCCs) between drugs similarity in chemical structures, ATC-code annotations, and side-effects metrics and PPI similarity are shown in Figure 3A. It shows that all three PCCs are larger than 0.3, and PCC between PPI similarity and chemical structure similarity goes beyond 0.5. Moreover, the *P*-values are less than 1e − 16 for all three PCCs. All these results suggest that PPI similarity relates with chemical structures, ATC-code annotations, and side-effects. Furthermore, it correlates more with chemical structure similarity. That may come from the fact that chemical structure describes the basic feature of compound. PPI target depends on the structure of binding interface between small molecular and ligand. Thus chemical structure provides direct and more information in identification of PPI target. Scatter plot relating drug structure, ATC-code annotation and side-effect similarity with PPI similarity (Supplementary Fig. S3) also indicates that drugs similar in either structure, ATC-code, or side-effect measurement have similar PPI targets. Moreover, PPI similarity is more correlated with drug chemical structure similarity.

3.4 PPI target prediction by PrePPItar

We firstly validate the performance of each single data source in PPI target prediction. The effects of chemical structure, ATC-code, and side-effect similarity in uncovering the observed drug-PPI associations are shown by replacing the drug similarity matrix S_{drug} in kernel function (3) with S_C, S_A, and S_E, respectively. The ROCs on each data source's effect are displayed in Supplementary Figure S4. The figure shows that, for all the three data sources, ROC curves are beyond the diagonal (random classification) and close to the 0–1 baseline. Since we are more interested in the performance of these

methods when FPR is rather small, we also draw ROC curves with FPR less than 0.02 in Figure 3B. It shows that, 'side-effect' obtains the highest TPR when FPR is very small, and with the number of known associations increasing, 'chem' reveals more observed drug-PPI associations.

The evaluation criteria obtained by PrePPItar are listed in Table 1 when the corresponding F-measure reaches its maximum. From Table 1, we can see that, 'chem' and 'ATC-code' achieve comparable prediction performance, 'side-effect' performs the worst. It means that chemical structure and ATC-code annotation are important to identify observed drug-PPI associations. What's more, 'chem' obtains over 0.90 AUC and over 0.85 sensitivity, 'ATC-code' obtains the best sensitivity of 0.835, and 'side-effect' obtains the AUC of 0.87, and make accuracy, sensitivity, precision, and F-measure over 0.85. All these results suggest that, each such data source will do one's bit in inferring the potential rules from the existing drug-PPI associations. Therefore, combination of these three data sources should produce a much more sophisticated picture of the associations among drugs and PPIs.

Figure 3B and Supplementary Figure S4 show that, 'comb' not only obtains the best ROC curves, but also achieves the high TPR when FPR is less than 0.02. That is, 'comb' can achieve better performance when predicting a small fraction of known drug-PPI associations. In addition, Table 1 shows that 'comb' performs better than using single data source. For example, 'chem' and 'side-effect' make the AUPR 0.917 and 0.906, respectively, while 'comb' obtains an AUPR 0.936. 'chem' and 'ATC-code' obtain the accuracy 0.85 and 0.86, respectively, while 'comb' obtains an accuracy 0.87, which has one percent improvement. These facts demonstrate that all three data sources are very useful in prediction. Combination of them significantly improves the accuracy of drug-PPI association identification.

Table 1. The performance comparison of different data sources for predicting drug-PPI associations

Data source	AUPR	AUC	Acc	Sn	Sp	Pre	F-measure
chem	0.917	0.900	0.857	0.815	0.899	0.881	0.855
ATC-code	0.924	0.916	0.867	**0.835**	0.899	0.892	0.866
side-effect	0.906	0.876	0.833	0.769	0.899	0.882	0.828
comb	**0.936**	**0.921**	**0.873**	0.811	**0.936**	**0.927**	**0.869**

The best predictions obtained are highlighted in bold.

3.5 Comparison with PPI inhibitors and single protein target prediction

Previous work discovered PPI inhibitors by constructing a universal classifier to distinguish drug-like PPI inhibitors from Chemical Libraries. Only chemical properties were used. For example, many studies applied computational method to discover the candidate small molecules binding protein complex based on drug's chemical properties (Arnout *et al.*, 2013; Neugebauer *et al.*, 2007; Reynès *et al.*, 2010; Villoutreix, 2014). While here, we integrated drug chemical structure, phenotype information and protein genomic sequence information to uncover the associations between drugs and physically interacted PPIs. We put more focus on associations between drugs and PPIs, not only included compound information, but also included the protein information. While previous work identified the specific small molecules used as PPI inhibitors based only on drug's chemical properties. This makes the number of studied PPIs is very limited. Instead, we address the issue of drug-PPI associations by extracting information both from drug and PPI sides. Furthermore, we extend the protein complex data to physical interaction data by high throughput experiment. This greatly increased the prediction coverage.

Compared with single target prediction, PrePPItar takes PPI as an basic unit of drug target. By using S-kernel to define the similarity between PPIs, PrePPItar can uncover those protein targets, which have less similarity to a known protein target. This is feasible because S-kernel measures similarity between PPIs as the maximum similarities between proteins involving corresponding PPIs. Take drug NADH (DrugBank: DB00157) as an example (Fig. 4A), the candidate interaction DB00157–O43837 will be missed by single target methods, due to that the similarity between candidate target O43837 and known target P50213 is only 0.392. However, it can be uncovered by direct PPI target method, because of the fact that candidate PPI: O43837–P51553 is similar to known PPI target: P50213–P51553 (with 0.835 similarity). In addition, DrugBank indicates the target role of O43837 for NADH. Besides that, O43837 and P50213 share many GO annotations, including tricarboxylic acid cycle in biological process phase, mitochondrial matrix in cellular component phase, NAD binding, isocitrate dehydrogenase (NAD+) activity and magnesium ion binding in molecular function phase. That is, O43837 and P50213 may work together to perform the same molecular function, such as inhibition by the same drug.

Although side-by-side comparison of drug-PPI prediction and drug-target is difficult, we still design some experiments to systematically compare the proposed approach with three state-of-the-art methods for predicting drug targets (Yamanishi *et al.*, 2010, 2012,

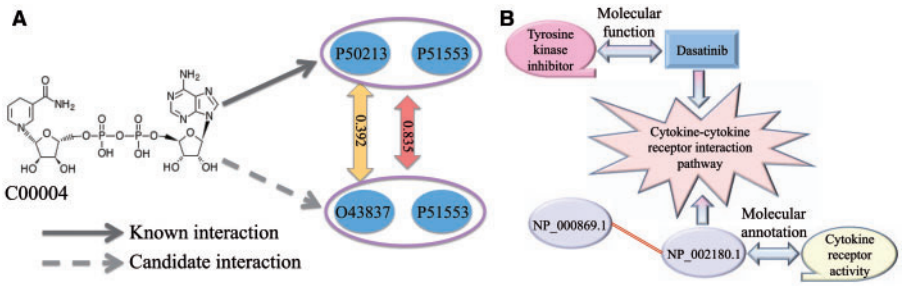

Fig. 4 (**A**) The example shows the promising advantage of inhibition of PPI target. Due to candidate target O43837 has low similarity to known target P50213, the relationship between drug C00004 (NAHD) and protein target O43837 will be missed by prediction of single protein target, but can be discovered by prediction of direct PPI target method with high similarity to known PPI target: P50213–P51553. (**B**) The relationship between drug D03658 (Dasatinib) and PPI (NP_002180.1(Q13261)-NP_000869.1(P14784)). Co-pathway relationship between drug Dasatinib and protein NP_002180.1, and functional analysis indicate that drug D03658 (Dasatinib) will associate with PPI (NP_002180.1-NP_000869.1) with high probability

2014). Firstly, we compare three state-of-the-art methods in web service DINIES (**D**rug-target **I**nteraction **N**etwork **I**nference **E**ngine based on **S**upervised Analysis) with our method. DINIES is a recently constructed web server for predicting unknown drug-target interaction networks from various types of biological data (e.g. chemical structures, drug side effects, amino acid sequences and protein domains) in the framework of supervised network inference (Yamanishi *et al.*, 2014). We first submitted our manually curated drug-PPI dataset to DINIES by taking PPI as the drug target. The results showed that our method outperformed the three methods by higher sensitivity (Supplementary Fig. S5). Then we decomposed our gold-standard drug-PPI network into the drug-protein network by separating PPIs into individual proteins. We ran DINIES on this drug-protein network by integrating drug chemical structure, ATC-code, side-effect data and protein sequence data. We found that our method can greatly increase the coverage of PPI target compared with the three methods. Finally we compiled the example dataset provided by DINIES as a new drug-PPI dataset. The co-pathway information provided by KEGG database is used to approximate PPI. This leads to a drug-PPI network with 1 985 interactions among 53 drugs and 292 co-pathway PPIs (Supplementary Table S7). Our method achieved an AUC of 0.937 by 10-fold cross-validation and outperformed the three methods in DINIES on this dataset. We checked one example that DINIES failed to predict while our method can predict. The drug-protein pair D00312–has:2099 was missed by DINIES because has:2099 is dissimilar with known target of D00312: has:2100 (with 0.37 sequence similarity, and share three Pfam domain in all 335 Pfam domains). While it could be uncovered by our method PrePPItar in prediction D00312-has:2099 has:2100. The evidence used by PrePPItar is as follows. The co-pathway protein pair has:2099 has:2100 is similar to known co-pathway protein pair has:1812 hsa:1813 (with Pfam domain similarity 0.74). has:1812 hsa:1813 is associated with drug D00059. The drug D00059 is very similar to D00312 (side-effect based similarity between D00312 and D00059 is 0.85). Therefore D00312-has:2099 has:2100 can be predicted by our PrePPItar in high confidence. The detailed experiments and the corresponding results are provided in the Supplementary Material. Taken together, our method could not only uncover more drug targets comparing with three state-of-the-art methods for predicting drug targets, but also discover some novel targets failed by the single protein based drug target prediction.

3.6 Novel predictions

By cross-validation, PrePPItar displays its promising performance in predicting observed drug-PPI associations. To test whether it can produce biologically useful predictions, we focus on the unknown (non-interacting) drug-PPI pairs. We trained 'comb' on the gold-standard drug-PPI associations and randomly selected non-interacting drug-PPI pairs from gold standard drug-PPI network, and tested it on 4 050 drug-PPI pairs. This comprises the randomly selected 50 PPIs form HPRD (excluding PPIs appearing in gold-standard dataset) and 81 drugs with known KEGG ID in gold-standard dataset. Our expectation is that 'comb' can discover novel PPI targets for known drugs.

The top five novel predictions are listed in Table S3. For each novel prediction, we look for the drug-target form DrugBank (Wishart *et al.*, 2008), drug treated disease from KEGG BRITE (Kanehisa *et al.*, 2006), the corresponding disease genes from OMIM (Hamosh *et al.*, 2002), the pathway information, disease gene from KEGG BRITE (Kanehisa *et al.*, 2006), and protein cellular component annotation from Uniprot (http://www.uniprot.org/). As an example shown in Figure 4B, drug D03658 (Dasatinib) is an oral multi-BCR/ABL and Src family tyrosine kinase inhibitor approved for usage in patients with chronic myelogenous leukemia (CML). This drug participates in the pathway of hsa04060 (Cytokine-cytokine receptor interaction). Protein NP_002180.1 (Interleukin-15 receptor subunit alpha) has the molecular annotation of cytokine receptor activity and participates in the pathway of hsa04060 (Fig. 4B). These annotations suggest that NP_002180.1 will be the target of D03658 with high probability, and PPI (NP_002180.1-NP_000869.1) will be served as the target of D03658 with high probability. That is, further experiment should focus on those PPIs with one of protein sharing the same pathway.

The analysis of the remaining four predictions is presented in the Supplementary Material. Database search and functional annotation analysis support these novel predictions. All these results suggest that PrePPItar can uncover potential PPI targets of drugs and can provide candidates for further experiments.

4 Discussion and conclusion

In this work, we integrate multiple chemical properties of drugs, such as chemical information, ATC-code annotation, and the drug side-effect. We use the maximum among them to obtain good predictions. However, there are alternative strategies to address the same issue, such as the multiple kernel learning (MKL), which optimizes the weight to integrate kernels (Francis *et al.*, 2004; Sonnenburg *et al.*, 2006; Mehmet and Ethem, 2011). MKL is a unified framework and has elegant model to integrate different data sources. To compare with MKL, we implemented the alternative optimization method. We iteratively obtained optimal weights to integrate kernels and got the optimal decision function. It turns out that MKL achieved an AUC of 0.795, which is lower than both single kernel and maximum kernel. One possible reason is that using maximum weight can effectively remove redundancy. These comparison results suggest that our simplified strategy is the better option for integrating data sources. In addition, MKL will add extra computational complexity. So in practice, it is better to choose the maximum strategy to simplify the model and make it available to large-scale problems.

Here, only PPIs in HPRD are included in our gold-standard dataset. It results in a relative small benchmark dataset and limits the usage of this dataset. Therefore, we enlarged the PPI set by introducing a recently constructed structure-based PPI database, called PrePPI, which deposits human interactions in databases HPRD, DIP, IntAct, BioGRID and MINT prior to Aug. 2010. It includes 58 772 interactions among proteins (Zhang *et al.*, 2013; Zhang QC et al., 2013). By selecting the interactions, which correlated with the same drugs in DrugBank database, we got 576 interactions among 229 FDA-approved drugs and 241 PPIs, which includes 85 of 113 drugs and 25 of 63 PPIs in our gold-standard dataset. We validated our PrePPItar on this larger dataset and got the AUC of 0.842. This result further demonstrates the efficiency of PrePPItar by using different gold-standard positive datasets. Another concern about our golden-standard dataset may lie in the fact that only physically interacted PPIs were included here. That is, we only consider the bioactive partners. While protein targets for the same drug may be not just physically interacted. They may have some functional linkage and participate in the same cellular pathway. We will consider extending the dataset by including other types of protein linkages in future, including the whole protein-family, co-pathway, and genetic interactions. In addition, PPIs are dynamic in biological processes. It may vary in different tissues and conditions. Thus extended work

would consider some tissue specific and condition specific data as features. For example, gene expression data can be easily obtained in many different situations. And some studies have already demonstrated that gene expression is important to identify novel drug targets (Kosaka *et al.*, 2013; Saito, 2005; Shibata *et al.*, 2010). Therefore, one could use tissue-specific gene expression under drug different dosage to predict whether a given drug can be used as PPI target in this specific tissue or not.

The PPI inhibition by drugs database, such as iPPI-DB (http://www.ippidb.cdithem.fr) and 2P2Idb (http://2p2idb.cnrs-mrs.fr/), deposit the structures of PPIs as protein complexes and their ligands. The small molecules targeting PPI complex are usually called PPI inhibitors. In our gold-standard dataset, the FDA approved drugs in DrugBank database and their interacting PPIs are included. Specifically, the PPIs in our gold-standard datasets are those physically interacted proteins, which interact with the same drug in DrugBank database. While PPIs in PPI inhibitors database are the protein complexes, which directly bind a specific small molecule. Therefore, PPIs in our gold-standard dataset are not limited to those protein complexes directly binding to a small molecule, but have some other types of relationships with small molecules, such as sensitivity to drug therapy. Here, we aimed to broaden the PPI targets searching space, not just focusing on those structurally interacting PPIs for drugs. This will provide novel candidates for drug discovery.

We utilized sequence information to characterize the similarity among PPIs in this work. The experimental results show that sequence information is predictive in PPI target prediction. One concern is that protein sequence similarity for a protein pair is too strict for protein complex. Because two proteins may form the complex due to complicated reason, such as they are co-expressed and have some functional linkage (Whisstock and Lesk, 2003; Dobson *et al.*, 2004). Therefore, we tried to include protein similarities based on other protein information, such as the Gene Ontology (GO) annotations. We applied the GO annotation linkage as the measurement for protein similarity. Specifically, we downloaded the human gene GO annotation from Gene Ontology Consortium, applied R package named GOSemSim (Yu *et al.*, 2010), which can be used to compute semantic similarity among GO terms, to measure the similarity among proteins. We then used S-kernel to measure the similarity among PPIs and finally got the AUC of 0.927 by using maximum kernel. While sequence-based similarity achieved AUC 0.936 by 'comb'. This indicates that sequence similarity is widely used not only because it may be the easiest data source to access, but also because it could get the competitive results with the other data sources. However, we did find PPIs with similar function but with dissimilar amino acid sequence. For example, PPI: P50213-P51553 and PPI: Q9Y478-Q13131 got similarity 0.85 based on GO annotations, but got similarity 0.04 based on amino acid sequence similarity. Although none of P50213 and P51553 is similar to Q9Y478 or Q13131 in terms of amino acid sequence, P51553 and Q13131 share GO term of GO:0005524, and got similarity 0.601 by running GOSemSim. This result suggested that, different data sources for protein might play complementary roles. Therefore the better way to represent PPI is applying maximum strategy to integrate the multiple protein data sources. In future, we will extend our work to integrate different data sources for protein in PPI target prediction, including protein sequence, GO annotation and so on. Another possible improvement is to use the defined interacting domain in protein sequence and to make the sequence similarity score more accurate.

In summary, we propose a new computational method, PrePPItar, to predict PPI targets by inferring novel associations between drugs and PPIs from a collection of molecular structure, molecular function and phenotype data. Our main contributions here are both in proposing the machine learning framework and integrating the drug and PPI similarity profiles by kernel function to construct the predictive model. Specifically, we characterize the drug similarity profiles form chemical structures, ATC-code annotations, and side-effects. We quantify the PPI similarity by S-kernel. And by casting PPI target prediction into a binary classification problem, a SVM-based predictor is utilized to uncover unknown relationships between drugs and PPIs. The improvement in the evaluation criteria is obtained by combining chemical structures, ATC-code annotations, and side-effects. This is shown by the maximum F-measure on a well-established drug-PPI networks with 227 associations between 113 drugs and 63 PPIs. Taken together, PrePPItar can accurately uncover drug-PPI associations with database and literature evidences. Indeed, the database search and pathway analysis reveal that our novel predictions are worth future experimental validation. Furthermore, prediction of PPI targets will broaden the drug target search space. In conclusion, PrePPItar will promote the further research in drug discovery.

Funding

This work is supported by the National Natural Science Foundation of China (No. 11201470, No. 31270270, No. 61171007, No. 11422108, 11131009 and No. 11371365). YW was also supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (No.XDB13040700).

Conflict of Interest: none declared.

References

Archakov,A.I. *et al.* (2003) Protein–protein interactions as a target for drugs in proteomics. *Proteomics*, **3**, 380–391.

Arnout,V. *et al.* (2013) Protein interface pharmacophore mapping tools for small molecule protein: protein interaction inhibitor discovery. *Curr. Top. Med. Chem.*, **13**, 989–1001.

Arkin,M.R. and Wells,J.A. (2004) Small-molecule inhibitors of protein–protein interactions: progressing towards the dream. *Nat. Rev. Drug Dis.*, **3**, 301–317.

Basilico,J. and Hofmann,T. (2004) A joint framework for collaborative and content filtering. In: 27th Annual International ACM SIGIR Conference.

Basse,M.J. *et al.* (2013) 2P2Idb: a structural database dedicated to orthosteric modulation of protein–protein interactions. *Nucleic. Acids Res.*, **41**, D824–D827.

Ben-Hur,A. and Noble,W.S. (2005) Kernel methods for predicting protein–protein interactions. *Bioinformatics*, **21**, i38–i46.

Butland,G. *et al.* (2005) Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature*, **433**, 531–537.

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

Deng,N.Y. *et al.* (2012) *Support vector machines: optimization based theory, algorithms, and extensions*. CRC Press, United States.

Dobson,P.D. *et al.* (2004) Prediction of protein function in the absence of significant sequence similarity. *Curr. Med. Chem.*, **11**, 2135–2142.

Francis,R. *et al.* (2004) Multiple kernel learning, conic duality, and the SMO algorithm. In: *Proceedings of the twenty-first international conference on Machine learning (ICML '04)*. ACM, New York, NY, USA

Gavin,A.C. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

Giot,L. *et al.* (2003) A protein interaction map of *Drosophila melanogaster*. *Science*, **302**, 1727–1736.

Gribskov,M. and Robinson,N.L. (1996). Use of receiver operating characteristic (roc) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Günther,S. *et al.* (2008) Supertarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Res.*, **36**, D919–D922.
Hamosh,A. *et al.* (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
Hamon,V. *et al.* (2013) 2P2IHUNTER: a tool for filtering orthosteric protein–protein interaction modulators via a dedicated support vector machine. *J. R. Soc. Interface.*, **11**, 20130860.
Hattori,M. *et al.* (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem. Soc.*, **125**, 11853–11865.
Hofmann, T. *et al.* (2008) Kernel methods in machine learning. *Ann. Stat.*, **36**, 1171–1220.
Hue,M. and Vert,J.-P. (2010) *On learning with kernels for unordered pairs*. ICML, Haifa, Israel. pp. 463–470.
Ito,T. *et al.* (2001) A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, **98**, 4569–4574.
Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, **34**, D354–D357.
Klussmann,E. and Scott,J. (2008) *Protein–Protein Interactions as New Drug Targets*. Springer-Verlag, Berlin.
Kosaka,T. *et al.* (2013) Identification of drug candidate against prostate cancer from the aspect of somatic cell reprogramming. *Cancer Sci.*, **104**, 1017–1026.
Krystal,G.W. *et al.* (1998) Lck associates with and is activated by Kit in a small cell lung cancer cell line: inhibition of SCF-mediated growth by the Src family kinase inhibitor PP1. *Cancer Res.*, **58**, 4660–4666.
Kuhn,M. *et al.* (2010) A side effect resource to capture phenotypic effects of drugs. *Mol. Syst. Biol.*, **6**, 343.
Labbè,C.M. *et al.* (2013) iPPI-DB: a manually curated and interactive database of small non-peptide inhibitors of protein–protein interactions. *Drug Dis. Today*, **18**, 958–968.
Li,S. (2004) A Map of the interactome network of the metazoan *C. elegans*. *Science*, **303**, 540–543.
Mehmet,G. and Ethem,A. (2011) Multiple kernel learning algorithms jour. *Mach. Learn. Res.*, **12**, 2211–2268.
Neugebauer, A. *et al.* (2007) Prediction of protein–protein interaction inhibitors by chemoinformatics and machine learning methods. *J. Med. Chem.*, **50**, 4665–4668.
Oyama,S. and Manning,C.D. (2004) Using feature conjunctions across examples for learning pairwise classifiers. In *European Conference on Machine Learning*. **2004**, pp. 322–333.
Pawson,T. and Nash,P. (2003) Assembly of cell regulatory systems through protein interaction domains. *Science*, **300**, 445–452.
Prasad,T.S.K. *et al.* (2009) Human protein reference database-2009 update. *Nucleic Acids Res.*, **37**, D767–D772.
Rain,J.C. *et al.* (2001) The protein–protein interaction map of Helicobacter pylori. *Nature*, **409**, 211–215.
Reynès,C. *et al.* (2010) Designing focused chemical libraries enriched in protein–protein interaction inhibitors using machine-learning methods. *PLoS Comput. Biol.*, **6**, e1000695.
Rual,J.F. (2005) Towards a proteomescale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.
Saito H. (2005) Mast cell-specific genes–new drug targets/pathogenesis. *Chem. Immunol. Allergy*, **87**, 198–212.
Schomburg,I. *et al.* (2004) Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.
Schölkopf,B. *et al.* (2004) *Support vector machine applications in computational biology*. MIT Press, Cambridge, MA.
Shen,J.W. *et al.* (2007) Predicting protein–protein interactions based only on sequences information. *Proc. Natl Acad. Sci. USA*, **104**, 4337–4341.
Shibata,T. *et al.* (2010) Global downstream pathway analysis reveals a dependence of oncogenic NF-E2-related factor 2 mutation on the mTOR growth signaling pathway. *Cancer Res.*, **70**, 9095–9105.
Sonnenburg,S. *et al.* (2006) Large scale multiple Kernel learning. *J. Mach. Learn. Res.*, **7**, 1531–1565.
Stelzl,U. *et al.* (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.
Smith,T.F. and Waterman,M. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
Uetz,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
Valkov,E. *et al.* (2012) Targeting protein–protein interactions and fragment-based drug discovery. *Top. Curr. Chem.*, **317**, 145–179.
Vapnik,V. (1995) *The nature of statistical learning theory*. Springer, New York.
Vapnik,V. (1998) *Statistical learning theory*. Wiley, New York, U.S.A.
Villoutreix,B.O. *et al.* (2014) Drug-like protein–protein interaction modulators: challenges and opportunities for drug discovery and chemical biology. *Mol. Inf/ Special Issue Strasbourg Summer School Chemoinf.*, **33**, 414–437.
Wang,Y.C. *et al.* (2010) Computationally probing drug-protein interactions via support vector machine. *Lett. Drug Des. Dis.*, **7**, 370–378.
Wang,Y.C. *et al.* (2011) Kernel-based data fusion improves the drug-protein interaction prediction. *Comput. Biol. Chem.*, **35**, 353–362.
Wells,J.A. and McClendon,C.L. (2007) Reaching for high-hanging fruit in drug discovery at protein–protein interfaces. *Nature*, **450**, 1001–1009.
White,A.W. *et al.* (2008) Protein–protein interactions as targets for small-molecule therapeutics in cancer *Expert Rev. Mol. Med.*, **19**, 10:e8.
Whisstock,J.C. and Lesk,A.M. (2003) Prediction of protein function from protein sequence and structure. *Q. Rev. Biophys.*, **36**, 307–340.
Wishart,D.S. *et al.* (2008) Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.
Wu,G. and Chang,E.Y. (2003) Class-boundary alignment for imbalanced dataset learning. In: *ICML 2003 Workshop on Learning from Imbalanced Data Sets*.
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.* (2012) Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, **28**, i611–i618.
Yamanishi Y. *et al.* (2014) DINIES: drug?target interaction network inference engine based on supervised analysis. *Nucleic Acids Res.*, **42**, W39–W45.
Yu,G.C. *et al.* (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.
Zhang,Q.C. *et al.* (2013) Structure-based prediction of protein–protein interactions on a genome-wide scale. *Nature*, **490**, 556–560.
Zhang,Q.C. *et al.* (2013) PrePPI: a structure-informed database of protein–protein interactions. *Nucleic Acids Res.*, **41**, D828–D833.