BRIEFINGS IN BIOINFORMATICS. VOL 15. NO 5. 734–747
Advance Access published on 11 August 2013
doi:10.1093/bib/bbt056

Similarity-based machine learning methods for predicting drug–target interactions: a brief review

Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka and Shanfeng Zhu

Submitted: 17th February 2013; Received (in revised form): 15th July 2013

Abstract

Computationally predicting drug–target interactions is useful to select possible drug (or target) candidates for further biochemical verification. We focus on machine learning-based approaches, particularly similarity-based methods that use drug and target similarities, which show relationships among drugs and those among targets, respectively. These two similarities represent two emerging concepts, the chemical space and the genomic space. Typically, the methods combine these two types of similarities to generate models for predicting new drug–target interactions. This process is also closely related to a lot of work in pharmacogenomics or chemical biology that attempt to understand the relationships between the chemical and genomic spaces. This background makes the similarity-based approaches attractive and promising. This article reviews the similarity-based machine learning methods for predicting drug–target interactions, which are state-of-the-art and have aroused great interest in bioinformatics. We describe each of these methods briefly, and empirically compare these methods under a uniform experimental setting to explore their advantages and limitations.

Keywords: *drug discovery; drug–target interaction prediction; machine learning; drug similarity; target similarity*

INTRODUCTION

Interactions between drugs and targets (proteins) are of importance in drug research, such as facilitating the process of drug discovery [1], drug side-effect prediction [2, 3] and drug repurposing [4–6]. Drugs have specific impacts on multiple proteins called targets by changing pharmaceutical functions of the targets [7], such as enzymes, ion channels, G protein-coupled receptors (GPCRs) and nuclear receptors. Known drug–target interactions, however, are very limited [8–10]. In fact, PubChem [11] contains around 35 million compounds, although only <7000 compounds have target protein information. This gives us a strong incentive to develop more effective and efficient methods to predict drug–target interactions.

Biochemical experiments or *in vitro* methods for finding drug–target interaction are extremely costly and time-consuming [12–14]. In contrast, computational or *in silico* methods can find potential interactions for *in vitro* validation more efficiently [1, 15]. Two major *in silico* approaches are docking

Corresponding author. Shanfeng Zhu, School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200433, China. Tel: +8621-55664712; Fax: +8621-65654253; E-mail: zhusf@fudan.edu.cn

Hao Ding is a postgraduate student at the Shanghai Key Lab of Intelligent Information Processing and School of Computer Science in Fudan University. His research interests include machine learning, data mining and their applications in bioinformatics.

Ichigaku Takigawa is an assistant professor at the Creative Research Institution and Division of Computer Science, Graduate School of Information Science and Technology in Hokkaido University. His research interests include complex molecular interactions, computational genomics and genetics, chemical genomics and pharmacogenomics.

Hiroshi Mamitsuka is a professor at the Institute for Chemical Research in Kyoto University, jointly appointed as a professor at the School of Pharmaceutical Sciences of the same university. His research interests include machine learning, data mining and their applications in bioinformatics and chemoinformatics.

Shanfeng Zhu is an associate professor at the Shanghai Key Lab of Intelligent Information Processing and School of Computer Science, Fudan University, Shanghai, China. His research interests include machine learning, data mining and their applications in bioinformatics and information retrieval.

simulation and machine learning. Docking simulation is widely used in biology but has two serious problems: (1) we need to know the three-dimensional structure of a target to compute the binding of each drug candidate to the target [16–20], but the three-dimensional structures of many targets, especially GPCRs, are still unavailable [21, 22]; and (2) simulation is time-consuming, in that a large amount of computational resources are needed. On the other hand, machine learning is more efficient, allowing larger-scale predictions than docking simulation, and thus examining a larger number of promising candidates for further experimental screening. In this review, we focus on machine learning-based methods for predicting drug–target interactions.

Machine learning-based methods proposed so far can be classified into three types:

(i) **Feature vector-based approach**: The input of general machine learning is instances, which can be represented by feature vectors. In our setting, instances are drug–target interactions, and the feature vectors can be generated by combining (structural) chemical descriptors of drugs and sequences of targets. Then with these inputs, we can use any standard machine learning method, such as support vector machines (SVMs) [23–25].

(ii) **Similarity-based approach**: We can generate a similarity matrix for drugs, where the (i,j)-element of the matrix is the similarity of drug i and drug j, typically being computed by chemical structures. Likewise, a target similarity matrix can be generated by protein sequence alignment. Recently, these two similarity matrices have been used in many methods, including kernel regression [26], bipartite local method (BLM) [27], pairwise kernel method (PKM) [28], Laplacian regularized least squares (LapRLS) [29], net Laplacian regularized least squares (NetLapRLS) [29], Gaussian interaction profile (GIP) [30] and kernelized Bayesian matrix factorization with twin kernels (KBMF2K) [31]. These similarity-based methods have a number of clear advantages: (1) compared with feature vector-based approaches, similarity-based approaches do not need feature extraction or selection, which is usually a complex and difficult process. (2) Computing similarity measures such as chemical structure similarity for drugs and genomic sequence similarity for targets have been already fully developed and widely used. (3) Similarity-based approaches can be directly related to well-developed kernel methods, which can provide high-performance prediction results. (4) Similarity matrices show the relationships among drugs and genes, being consistent with recent concepts, the *chemical space* and the *genomic space*, respectively. These advantages make similarity-based approaches more promising than other approaches. In addition, for drugs or targets, not only a single matrix but also different similarity matrices can be combined [32].

(iii) **Other approaches**: We can use other information, including pharmacological information of drugs [33] and biomedical documents, from which implicit co-occurrent compound–protein relations can be extracted by text mining techniques [34]. However, one major drawback of the relations from documents is that they might not be real drug–target interactions.

In light of the properties of these three approaches, we focus on similarity-based machine learning methods in this review.

Predicting drug–target interactions has currently attracted much attention in bioinformatics and chemoinformatics. There already exist three recent reviews with different emphasis [35–37] and one special issue [38], to the best of our knowledge. However, these reviews and special issue have not been written from a viewpoint of developing machine learning methods. More detailed comparison and analysis of machine learning methods would be useful for biologists and chemists to choose the most suitable model and for computer scientists to develop higher-performance prediction methods. In particular, the most promising similarity-based methods should be checked more carefully. In addition, latest methods, such as GIP, KBMF2K and network-based inference (NBI), have not been included in any of the reviews and the special issue. In this review, we focus on similarity-based machine learning methods and systematically compare these methods, models and information used. Furthermore, we extensively compare the performance of several representative methods under a uniform experimental setting. We will summarize the performance advantages and drawbacks of the compared methods, and discuss future perspectives.

The rest of this article is organized as follows: GENERAL FRAMEWORK describes the data

and information used in the latest similarity-based machine learning methods. METHODS briefly explains these methods. EXPERIMENTS empirically compares the performance of the methods under a uniform experimental setting. DISCUSSION AND CONCLUSION discusses the advantages, limitations and future perspectives of prediction methods.

GENERAL FRAMEWORK

Data sources

Machine learning methods are data-driven. Data on drugs, targets and drug–target interactions are available in the following databases: KEGG BRITE [9], BRENDA [39], SuperTarget [40], DrugBank [41], DCDB [42] and ChEBI [43]. Table 1 shows two datasets, Dataset 1 and Dataset 2, that have been used in experiments of recent similarity-based machine learning methods.

Major drug similarity matrices can be categorized into three types: (1) chemical structure-based similarity that can be computed from chemical structures of drugs by using databases such as DrugBank [41] or KEGG LIGAND [9]; (2) side-effect-based similarity that can be computed from side-effect information of drugs in databases such as SIDER [44] and (3) gene-expression-based similarity that is computed from the response of gene expression to drugs, which can be retrieved from databases such as the Connectivity Map project [45].

Typical target similarities can be classified into the following three types: (1) sequence-based similarity can be computed using sequence information from databases such as KEGG GENES [9]; (2) protein–protein interaction (PPI) network-based similarity can be derived from the distance between two targets in the PPI network [46–50] and (3) Gene Ontology (GO) semantic similarity can be computed using GO annotations [51] in some databases such as UniProt [52].

It is noteworthy that the similarity between two drug–target interactions can be directly computed [28, 53]. One method will be shown in Pairwise kernel method.

Learning and prediction

The general framework of machine learning for predicting drug–target interactions has two stages: (1) training a model and (2) predicting the interaction of a given drug–target pair by the trained model. A key underlying assumption of similarity-based machine learning methods is that similar drugs tend to share similar targets and vice versa [54–56].

Learning can be of three types: supervised, semi-supervised or unsupervised learning. In supervised learning, training data with known labels are used to train a prediction model. Techniques used in supervised learning are SVM, kernel methods, logistic regression etc. In semi-supervised learning, training data with and without labels are used in the learning process. Typical techniques in semi-supervised learning use a graph-based representation. Unsupervised learning does not need labels, and unlabeled data are the input. Typical techniques are clustering methods.

Prediction can be done for three cases: (1) predicting a new target that can interact with a drug that already has one or more targets, (2) predicting a new drug that can interact with a target that already has one or more drugs or (3) predicting a new interaction for the pair of a drug and a target that already has one or more interactions.

METHODS

We briefly introduce the procedure of recent key similarity-based methods, which are nearest neighbor (NN), BLM, PKM, LapRLS, NetLapRLS, GIP and KBMF2K. Note that we do not raise kernel regression [26], as BLM [27] was empirically proven to outperform the kernel regression method by systematic experiments already. In addition, we briefly explain NBI, which is not a similarity-based machine learning method but proposed recently to allow

Table 1: Two datasets used in publications of similarity-based machine learning methods: Dataset 1 was used in BLM, LapRLS, NetLapRLS, GIP, KBMF2K and NBI, whereas Dataset 2 was used in PKM only

Dataset	Number of drugs (databases)	Number of targets (databases)	Number of interactions (databases)
Dataset 1	932 (KEGG LIGAND)	989 (KEGG GENES)	5127 (KEGG BRITE, BRENDA, SuperTarget, DrugBank)
Dataset 2	1205 (KEGG LIGAND)	889 (KEGG GENES)	2782 (KEGG BRITE)

predictions by using drug–target interactions. We use the following notation throughout this article: let $D = \{d_1, d_2, \ldots, d_m\}$ be a drug set, and $T = \{t_1, t_2, \ldots, t_n\}$ be a target set. Let $\mathbf{S}_d$ be a drug similarity matrix, where the (i,j) element of $\mathbf{S}_d$ is denoted by $s_d(d_i,d_j)$, which is the similarity score between drugs d_i and d_j. Similarly, let $\mathbf{S}_t$ be a target similarity matrix, where the (i,j) element of $\mathbf{S}_t$ is denoted by $s_t(t_i,t_j)$, being the similarity score between targets t_i and t_j. We assume that they are given and can be an input of any method. Let $\mathbf{Y}$ be a binary matrix of true labels of drug–target interactions, where if drug d_i and target t_j interact with each other, $\mathbf{Y}_{ij} = 1$; otherwise $\mathbf{Y}_{ij} = 0$. Let $\mathbf{F}$ be a score function (or matrix), where the (i,j)-element of $\mathbf{F}$, i.e. $\mathbf{F}_{ij}$, shows the score that drug d_i and target t_j interact with each other. The objectives of the methods in this section are to estimate $\mathbf{F}$ so that $\mathbf{F}$ should be consistent with $\mathbf{Y}$.

Nearest neighbor

NN is used as a baseline method in [27]. Let $\mathbf{y}_{d_i}$ be a binary vector, called *interaction profile* of drug d_i, where the j-th element of $\mathbf{y}_{d_i}$ is 1 if drug d_i interacts with target t_j; otherwise 0. Similarly, let $\mathbf{y}_{t_j}$ be a binary vector, called interaction profile of target t_j.

For new drug d_{new}, NN computes interaction profile $\mathbf{y}_{d_{new}}$ of d_{new} as follows:

$$\mathbf{y}_{d_{new}} = s_d(d_{new}, d_{nearest})\mathbf{y}_{d_{nearest}},$$

where $d_{nearest}$ is the known, most similar drug, i.e. $d_{nearest} = \arg\max_d s_d(d_{new}, d)$. Similarly, NN computes interaction profile $\mathbf{y}_{t_{new}}$ of t_{new} as follows:

$$\mathbf{y}_{t_{new}} = s_t(t_{new}, t_{nearest})\mathbf{y}_{t_{nearest}},$$

where $t_{nearest}$ is the known, most similar target, i.e. $t_{nearest} = \arg\max_t s_t(t_{new}, t)$. In this way, we can fill all elements of $\mathbf{y}_{d_{new}}$ and $\mathbf{y}_{t_{new}}$, by which we can predict the score of pairs (d_{new}, t_j) and (d_i, t_{new}), respectively, for all i and j. This means that there are two ways of predicting the score of any pair (d_i, t_j), for which the score is unknown. Thus in this case, we can simply take the average of the two possible scores.

NN is very efficient, using which we can predict many new drug–target pairs quickly.

Bipartite local models

BLM extends the idea, called *local models*, which was proposed in [27, 57, 58]. BLM turns the problem of predicting edges in a bipartite graph into a binary supervised problem.

Drug–target interactions are represented by a bipartite graph. Figure 1 shows an example of the bipartite graph, where circles are drugs $\{d_1, d_2, \ldots, d_m\}$ and squares are targets $\{t_1, t_2, \ldots, t_n\}$. If drug d_i interacts with target t_j, then edge $\mathbf{Y}_{ij}$ connects nodes d_i and t_j. Suppose that we predict whether t_3 interacts with d_2 in Figure 1.

We first focus on d_2. We check whether a known target interacts with d_2 and give a label of $+1$ if so; otherwise a label of -1. We repeat this operation for all known targets. SVM does not need feature vectors of examples but instead similarity (kernel) between instances [59]. Thus we use the similarity matrix of targets with the generated labels of all known targets to train an SVM classifier. We then use the trained classifier to predict the label of t_3, i.e. whether t_3 interacts with d_2. Figure 2 shows this process schematically in the left-hand side.

The reverse way is also possible. That is, we first repeat labeling all known drugs by whether interacting with t_3 or not. We then train an SVM classifier to predict if d_2 interacts with t_3. This part is shown in the right-hand side of Figure 2. The final prediction is obtained by averaging the two obtained scores.

BLM can build a classifier for specific drug–target interactions, which, however, causes a serious computational problem, because BLM has to train unique classifiers for each of all possible pairs, including new drugs or targets.

Pairwise kernel method

SVM in general needs a similarity matrix (kernel) of instances with labels [28]. PKM is a straightforward

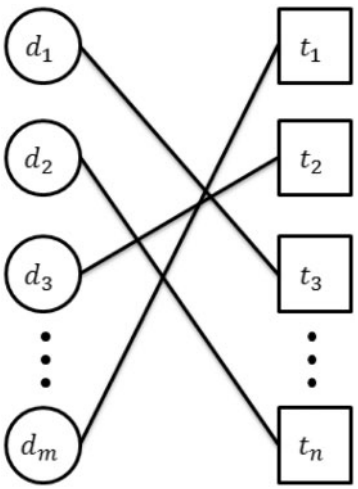

Figure 1: Bipartite graph representing a drug-protein interaction network.

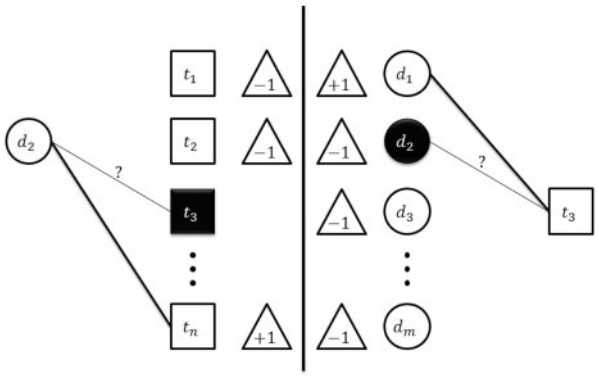

Figure 2: Procedure of predicting (d_2, t_3) by BLM. Circles indicate drugs and squares indicate targets. The label in a triangle indicates that the corresponding pair interacts (+1) or does not (−1).

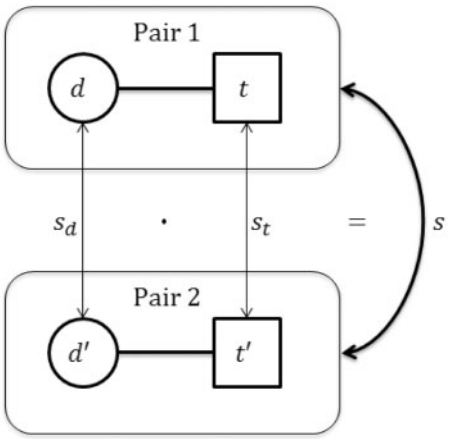

Figure 3: Schematic figure of PKM. Similarity between (d, t) and (d', t') can be computed by the inner product of the drug similarity between d and d' and the target similarity between t and t'.

SVM-based method and uses drug–target interactions as instances, implying that the similarity between drug–target pairs needs to be computed. PKM thus computes the similarity (pairwise kernel) of drug–target pairs from the drug similarity score and the target similarity score as follows:

$$s((d,t),(d',t')) = s_d(d,d') \cdot s_t(t,t') \quad (1)$$

Figure 3 is a schematic figure of this process, where the similarity between drug–target pairs is denoted by **S**.

PKM then uses the similarity matrix (kernel) of drug–target pairs with known labels to train an SVM classifier, which can then predict the scores of arbitrary drug–target pairs. PKM is more efficient than BLM, as only one classifier needs to be trained, by which all new drug–target pairs can be predicted. However, practically, the kernel matrix can be very large. For example, for 600 drugs and 500 target proteins, which are both standard numbers, the size of the kernel matrix is (600 × 500) × (600 × 500). In practice, we cannot use all instances for training, and negative instances must be randomly sampled to meet the size limitation of the main memory.

Laplacian regularized least squares and Net Laplacian regularized least squares

LapRLS attempts to directly estimate interaction score matrix **F** for drugs and targets, separately, which we denote by $\mathbf{F}_d$ and $\mathbf{F}_t$, respectively. For drugs, LapRLS minimizes the squared loss between **Y** and $\mathbf{F}_d$ with a regularized term of $\mathbf{S}_d$ and $\mathbf{F}_d$ [29]. This minimization leads to an analytical solution by which $\mathbf{F}_d$ can be updated by a rule containing $\mathbf{S}_d$ and **Y**. The same procedure can be performed for targets. Finally, **F** is obtained by averaging over $\mathbf{F}_d$ and $\mathbf{F}_t$. LapRLS is efficient because it can predict the scores of all drug–target pairs at one time. NetLapRLS is a modification of LapRLS to consider drug–target interactions more directly [29]. Simply, NetLapRLS is the same regularized least squares method as LapRLS, except that $\mathbf{S}_d$ in LapRLS is replaced with another matrix that considers drug–target interactions.

Gaussian interaction profile

More than one method is proposed in [30]. We select a method called RLS-Kron average in [30] as GIP, as this method achieves the best performance among the methods proposed in [30]. At the first step, GIP generates a Gaussian kernel from the interaction profiles. In this step, GIP considers drugs and targets separately, and so we first explain the drug side. For drugs d_i and d_j, a Gaussian kernel is given as follows:

$$K_{\mathrm{GIP}}(d_i,d_j) = \exp(-\gamma_d|\mathbf{y}_{d_i} - \mathbf{y}_{d_j}|^2),$$

where γ_d is a parameter that controls the width of the Gaussian distribution. GIP then linearly combines the Gaussian kernel with the drug similarity matrix into the kernel to be used further. This means that GIP can consider both the drug similarity score $\mathbf{S}_d$ and the interaction profile similarity. In other words, GIP can consider only the interaction profile similarity, which means that GIP works without $\mathbf{S}_d$. The totally same procedure is done for targets. We then

proceed to the second step, where GIP incorporates the idea of PKM, which computes a pairwise kernel over drug–target pairs that are shown in Equation (1). Finally, the resultant kernel matrix is used in the straight-forward framework of regularized least squares, which minimize the square loss between the score function and the true labels with a regularizer.

GIP keeps the same efficiency level as LapRLS but has a serious limitation that it cannot predict targets of a new drug or drugs of a new target, because it uses the interaction profile that cannot be computed for new drugs (and targets).

Kernelized Bayesian matrix factorization with twin kernels

The idea behind KBMF2K is first to project the drug and target spaces into two low-dimensional spaces through kernels (similarity matrices) and then to estimate drug–target interactions under the low-dimensional spaces [31]. More concretely, for drugs, KBMF2K has parameter matrix $\mathbf{A}_d$ to project drug similarity matrix (kernel) $\mathbf{S}_d$ into low-dimensional space $\mathbf{G}_d$. Similarly, target similarity matrix $\mathbf{S}_t$ is projected by parameter matrix $\mathbf{A}_t$ into low-dimensional space $\mathbf{G}_t$. Then drug–target interaction score matrix $\mathbf{F}$ is estimated to be consistent with both two low-dimensional spaces, $\mathbf{G}_d$ and $\mathbf{G}_t$, i.e. $\mathbf{F} = \mathbf{G}_d^T \mathbf{G}_t$. Figure 4 is a graphical model of KBMF2K. KBMF2K is inefficient, because estimating three matrices, $\mathbf{F}$, $\mathbf{A}_d$ and $\mathbf{A}_t$, is an iterative process, starting with random initial values.

Network-based inference

As shown in BLM, drug–target interactions can be represented by a bipartite graph. NBI considers transition processes over the bipartite graph to compute the score if a drug and a target interact with each other [60]. In the bipartite graph, let w_i be the degree of node i, i.e. the number of edges connecting to node i.

Suppose that we predict the interaction between drug d_i and target t_j. NBI first computes *connection score* $v_{t_j \to t_l}$ between target t_j and target t_l, by summing up $\frac{1}{w_k}$ of drug d_k, which connects to both targets t_j and t_l, for all d_k, as follows:

$$v_{t_j \to t_l} = \sum_{k | t_j \to d_k \to t_l} \frac{1}{w_k}$$

This means that if two targets share a larger number of drugs, the connection score increases. In particular, if the intermediate drug d_k has a lower degree, the connection score is larger. Then NBI further extends the connection (from target t_l) to drug d_i. Connection score $v_{t_j \to d_i}$ through target t_l can be computed by summing up $v_{t_j \to t_l}$ weighted by $\frac{1}{w_l}$ for all targets t_l as follows:

$$v_{t_j \to d_i} = \sum_{l | t_l \to d_i} \frac{1}{w_l} v_{t_j \to t_l}$$

This means that again if drug d_i connects to a larger number of targets, the connection score increases, being weighted by the connection score of the targets. In particular, if the target has a lower degree, the added connection score is larger. Finally, the connection score $v_{t_j \to d_i}$ is assigned to the score function $\mathbf{F}$ as follows:

$$\mathbf{F}_{ij} = v_{t_j \to d_i}$$

In a general machine learning sense, NBI is not necessarily a machine learning method and also not a similarity-based method. However, NBI earns the score function from given drug–target interactions, where drug–target interactions can be replaced with the similarity over drug–target pairs. Thus we add NBI to this review. Note that NBI cannot predict targets of a new drug or drugs of a new target, because a new drug (or a new target) has no edges to targets (or drugs), by which we cannot compute the connection scores.

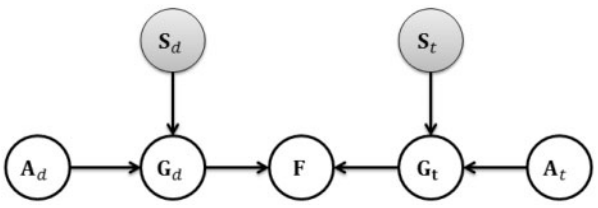

Figure 4: Graphical model of KBMF2K.

EXPERIMENTS

We empirically checked the performance of the recent similarity-based machine learning methods.

Data

We used exactly the same data as those in [27].

Drug–target interaction data

Two datasets have been used in the experiments of recent similarity-based machine learning methods (See Datasets 1 and 2 in Table 1). However, Dataset 2 is not a drug–target dataset but a ligand–protein

dataset, where ligands are mostly chemical compounds but not drugs. In addition, PKM, the only method that used Dataset 2, was not originally intended to predict drug–target interactions, but protein–ligand interactions. Furthermore, most (more than 77%) of the drugs in Dataset 2 also appear in Dataset 1, which means that there are very large overlaps between these two datasets. For these reasons, in our experiments, we just considered Dataset 1, which has four subsets, namely, *Nuclear receptors*, *GPCRs*, *Ion channels* and *Enzymes*, obtained from KEGG BRITE [9], BRENDA [39], SuperTarget [40] and DrugBank [41]. Table 2 shows the statistics of these four subsets.

Drug similarity

Drug similarity was computed from the chemical structures of drugs (obtained from KEGG LIGAND [9]) by using SIMCOMP [61], which computes the drug similarity between two drugs d and d' as follows:

$$s_d(d,d') = |d \cap d'|/|d \cup d'|,$$

where $|d \cap d'|$ is the number of all substructures shared by d and d' and $|d \cup d'|$ is the number of all substructures that either of d and d' has.

Target similarity

Target similarity was computed from target sequences (obtained from KEGG GENES [9]) by using a normalized Smith–Waterman score [62] of targets t and t' as follows:

$$s_t(t,t') = \frac{SW(t,t')}{\sqrt{SW(t,t)}\sqrt{SW(t',t')}},$$

where $SW(\cdot,\cdot)$ is the original Smith–Waterman score.

Settings

Procedure

We compared the performance of the eight similarity-based methods—NN, BLM, PKM, LapRLS, NetLapRLS, GIP, KBMF2K and NBI—under five trials of 10-fold cross-validation (CV). Note that the same folds were used for all methods. Further note that the size of the Nuclear receptor subset is much smaller than the other subsets, by which random division in CV might be less stable and so the resultant performance on this subset might be also more unstable and less significant than the other subsets. The CV was done in three different manners: (1) drug prediction: all drugs were divided into 10-folds, (2) target prediction: all targets were divided into 10-folds and (3) pair prediction: all drug–target interactions were divided into 10-folds. For drug prediction, in each round of 10-fold CV, 90% of rows in Y are used as training data, and the remaining 10% of rows in Y are used as test data. For target prediction, in each round, 90% of columns in Y are used as training data, and the remaining 10% of columns in Y are used as test data. For pair prediction, in each round, 90% of elements in Y are used as training data, and the remaining 10% of elements in Y are used as test data. Note that drug, target and pair predictions correspond to predicting new drugs, new targets and new drug–target interactions, respectively. GIP and NBI cannot be applied to the drug and target predictions, by which only six methods were compared in the drug and target predictions.

A standard measure for evaluating prediction results in supervised learning is AUC (Area Under the Receiver Operating Characteristic curve), which is not affected by the ratio of positives to negatives. Drug–target interactions have much fewer positives than negatives, and false-positives should be weighed more. AUPR (Area Under the Precision-Recall curve) punishes false-positives more than AUC [63]. We thus evaluated the performance of the eight methods by both AUC and AUPR. We denote AUC obtained by drug, target and pair predictions of CV by AUCd, AUCt and AUCp, respectively. Similarly, AUPR obtained by the three

Table 2: The details of Dataset I, which was used in our experiments, downloaded from (http://cbio.ensmp.fr/yyamanishi/bipartitelocal/) [27]

Statistics	Nuclear receptor	GPCR	Ion channel	Enzyme
Number of drugs	54	223	210	445
Number of targets	26	95	204	664
Number of drug–target interactions	90	635	1476	2926

types of CV are denoted by AUPRd, AUPRt and AUPRp, respectively.

Parameters

We implemented NN and NBI, according to [27] and [60], respectively. We used LIBSVM [64] for implementing BLM and PKM. For BLM and PKM, we used nested CV to evaluate their performances. The inner-loop CV is used to tune the regularization coefficient C of each SVM classifier in the range of {0.01,0.1,1,10,100,1000}. For PKM, in each round of 10-fold CV, the training set (9-folds) was divided into three partitions of the same size to select the best-performed C in terms of AUC by 3-fold CV. We then trained the prediction model using all training data (9-folds) with the selected C, and made predictions on the test data (1-fold). For BLM, due to the scarcity of positive examples for drug and target classifiers, we used leave-one-out CV and accuracy as the evaluation metric in the inner loop of nested CV. We randomly chose the same number of negative instances as that of positive instances in the implementation of PKM, due to the limitation of the main memory size. We implemented LapRLS and NetLapRLS, which were run by the same parameter values as those in [29]. For GIP and KBMF2K, we used the software provided by the authors, which were run by the same parameter settings as those in [30] and [31], respectively. The detailed parameter settings are shown in the Supplement. The software availability of all similarity-based methods (except NN) is summarized in Table 3.

Table 3: Software availability

Method	Public	URL
BLM	Yes	Available on request
PKM	Yes	http://bioinformatics.oxfordjournals.org/content/24/19/2149/suppl/DC1/
LapRLS	No	–
NetLapRLS	No	–
GIP	Yes	http://cs.ru.nl/tvanlaarhoven/drugtarget2011/
KBMF2K	Yes	http://users.ics.tkk.fi/gonen/kbmf2k/
NBI	No	–

Results

Tables 4–9 show AUC and AUPR values of the eight compared methods, with *P*-values (of paired *t*-test between each method and the best method in the same column) shown within brackets to show the statistical significance of the improvement obtained by the best methods, where for each column, the highest value is in boldface. From Tables 4 and 6, we can see that in terms of AUCt and AUCd, PKM performed the best, implying the superiority of the pairwise kernel shown in Equation (1). A comparable performance was obtained by KBMF2K, and only a little lower performance was obtained by LapRLS and NetLapRLS, being further followed by BLM and NN.

Tables 5 and 7 show that in terms of AUPRt and AUPRd, LapRLS performed the best, slightly outperforming NetLapRLS. The performance was followed by KBMF2K and PKM, and further by BLM and NN. Note that both PKM and KBMF2K achieved high AUC performance, whereas this was not the case for AUPR. On the other hand, LapRLS and NetLapRLS were very strong in terms of AUPR. We may say that SVM achieves the best performance in general evaluation, such as AUC, whereas this might not be the case with predicting drug–target interactions, a special situation in machine learning. Another point is that NetLapRLS was worse than LapRLS. The difference of LapRLS and NetLapRLS was only that $\mathbf{S}_d$ ($\mathbf{S}_t$) has drug–target interaction information in NetLapRLS but not in LapRLS. This means that the drug–target information in $\mathbf{S}_d$ ($\mathbf{S}_t$) of NetLapRLS does not work to improve the performance of predicting new drugs or new targets.

Tables 8 and 9 show that for both AUCp and AUPRp, GIP achieved the best performance, implying that the interaction profiles of drugs and targets worked well. For AUCp, however, the performance of GIP, NetLapRLS, KBMF2K, LapRLS and PKM was relatively comparable, being followed by BLM, NBI and NN. For example, for GPCR, AUCp was 0.94–0.95 by the top five methods and was 0.84–0.88 by the other three methods. Similarly for Enzyme, AUCp was 0.96–0.97 by the top five methods and was 0.89–0.93 by the other three methods. On the other hand, for AUPRp, GIP clearly outperformed the other methods, except NetLapRLS, which achieved a slightly lower performance than GIP. This means that the performance difference in AUPRp among the top five methods by AUCp was much clearer. For example, for GPCR, AUPRp by GIP was 0.73, being followed by NetLapRLS of 0.71, KBMF2K of 0.69 and LapRLS of 0.64. The next best AUPRp was NBI of 0.62, being followed by BLM, PKM and

Table 4: AUCt by 5 × 10-fold cross-validation, with *P*-values (of paired *t*-test) within brackets

AUCt	Nuclear receptor	GPCR	Ion channel	Enzyme
NN	0.707 (3.46×10^{-2})	0.646 (9.68×10^{-22})	0.652 (7.02×10^{-34})	0.555 (2.97×10^{-48})
BLM	0.458 (1.21×10^{-16})	0.627 (1.02×10^{-21})	0.881 (9.91×10^{-15})	0.843 (2.70×10^{-16})
PKM	0.688 (1.75×10^{-5})	**0.880**	**0.943**	**0.946**
LapRLS	0.563 (1.03×10^{-11})	0.788 (3.02×10^{-15})	0.920 (1.11×10^{-5})	0.914 (1.34×10^{-12})
NetLapRLS	0.561 (1.06×10^{-11})	0.787 (1.15×10^{-15})	0.916 (9.13×10^{-8})	0.909 (7.80×10^{-13})
KBMF2K	**0.756**	0.837 (8.43×10^{-7})	0.924 (6.98×10^{-5})	0.889 (9.38×10^{-18})

Table 5: AUPRt by 5 × 10-fold cross-validation, with *P*-values (of paired *t*-test) within brackets

AUPRt	Nuclear receptor	GPCR	Ion channel	Enzyme
NN	**0.438**	0.224 (2.31×10^{-14})	0.243 (6.05×10^{-33})	0.088 (2.49×10^{-45})
BLM	0.325 (3.10×10^{-3})	0.367 (2.90×10^{-15})	0.641 (2.26×10^{-21})	0.611 (7.92×10^{-25})
PKM	0.431 (4.30×10^{-1})	0.427 (3.44×10^{-9})	0.684 (3.35×10^{-20})	0.605 (3.29×10^{-30})
LapRLS	0.432 (4.47×10^{-1})	**0.508**	**0.778**	**0.792**
NetLapRLS	0.433 (4.49×10^{-1})	0.503 (3.65×10^{-2})	0.762 (1.35×10^{-9})	0.787 (4.45×10^{-4})
KBMF2K	0.404 (2.19×10^{-1})	0.412 (1.01×10^{-6})	0.725 (6.90×10^{-8})	0.607 (4.48×10^{-28})

Table 6: AUCd by 5 × 10-fold cross-validation, with *P*-values (of paired *t*-test) within brackets

AUCd	Nuclear receptor	GPCR	Ion channel	Enzyme
NN	0.599 (7.96×10^{-19})	0.533 (2.99×10^{-40})	0.518 (1.60×10^{-29})	0.521 (5.19×10^{-42})
BLM	0.693 (4.17×10^{-12})	0.829 (4.66×10^{-13})	0.770 (5.21×10^{-4})	0.781 (5.56×10^{-24})
PKM	**0.847**	**0.872**	0.798 (1.57×10^{-1})	**0.870**
LapRLS	0.820 (3.22×10^{-7})	0.845 (2.69×10^{-9})	0.796 (8.92×10^{-2})	0.801 (3.26×10^{-12})
NetLapRLS	0.819 (2.39×10^{-7})	0.834 (5.74×10^{-11})	0.783 (9.40×10^{-3})	0.791 (4.64×10^{-17})
KBMF2K	0.831 (3.95×10^{-2})	0.844 (1.55×10^{-5})	**0.808**	0.783 (7.01×10^{-12})

Table 7: AUPRd by 5 × 10-fold cross-validation, with *P*-values (of paired *t*-test) within brackets

AUPRd	Nuclear receptor	GPCR	Ion channel	Enzyme
NN	0.230 (3.88×10^{-11})	0.068 (1.25×10^{-31})	0.062 (1.53×10^{-19})	0.032 (2.73×10^{-25})
BLM	0.194 (3.49×10^{-20})	0.210 (6.50×10^{-29})	0.167 (1.51×10^{-20})	0.092 (1.29×10^{-26})
PKM	**0.504**	0.337 (3.11×10^{-10})	0.328 (3.03×10^{-5})	0.267 (5.95×10^{-15})
LapRLS	0.482 (1.79×10^{-2})	0.397 (5.01×10^{-1})	**0.366**	**0.368**
NetLapRLS	0.481 (1.96×10^{-2})	**0.397**	0.343 (1.50×10^{-3})	0.298 (1.36×10^{-12})
KBMF2K	0.450 (3.46×10^{-4})	0.357 (5.14×10^{-6})	0.296 (2.15×10^{-7})	0.253 (3.09×10^{-17})

NN with AUPRp of 0.46–0.50, these values being far lower than that of GIP. Another point of note is that NetLapRLS outperformed LapRLS, implying that incorporating drug–target interactions worked well to boost AUCp and AUPRp.

Table 10 summarizes the features, time and space complexities and experimental results of the eight compared methods.

Dataset 1, the dataset we used in this article, was generated in 2008, implying that currently there must be new drug–target interactions. To evaluate the practical predicting ability of similarity-based machine learning methods further, we checked the names of the top five non-interacting drug–target pairs, which were predicted by the three most high-performance methods, i.e. NetLapRLS, GIP

Table 8: AUCp by 5 × 10-fold cross-validation, with *P*-values (of paired *t*-test) within brackets

AUCp	Nuclear receptor	GPCR	Ion channel	Enzyme
NN	0.820 (4.10×10^{-8})	0.852 (8.31×10^{-33})	0.889 (3.82×10^{-39})	0.898 (1.01×10^{-44})
BLM	0.694 (1.30×10^{-15})	0.884 (7.13×10^{-28})	0.918 (9.09×10^{-29})	0.928 (2.78×10^{-39})
PKM	0.856 (6.33×10^{-6})	0.937 (4.96×10^{-18})	0.967 (8.46×10^{-27})	0.966 (1.69×10^{-14})
LapRLS	0.855 (5.01×10^{-5})	0.941 (1.41×10^{-9})	0.969 (5.41×10^{-18})	0.962 (2.16×10^{-24})
NetLapRLS	0.859 (2.57×10^{-4})	0.946 (5.00×10^{-5})	0.977 (1.26×10^{-6})	0.968 (2.94×10^{-18})
GIP	0.869 (1.70×10^{-2})	**0.951**	0.980 (2.70×10^{-3})	**0.973**
KBMF2K	**0.881**	0.943 (5.68×10^{-7})	**0.982**	0.966 (6.11×10^{-15})
NBI	0.680 (4.53×10^{-21})	0.835 (2.29×10^{-31})	0.928 (1.50×10^{-32})	0.894 (5.27×10^{-42})

Table 9: AUPRp by 5 × 10-fold cross-validation, with *P*-values (of paired *t*-test) within brackets

AUPRp	Nuclear receptor	GPCR	Ion channel	Enzyme
NN	0.530 (6.24×10^{-5})	0.474 (5.95×10^{-39})	0.574 (4.21×10^{-51})	0.659 (2.45×10^{-50})
BLM	0.204 (2.09×10^{-24})	0.464 (4.03×10^{-35})	0.592 (1.45×10^{-39})	0.496 (2.56×10^{-53})
PKM	0.515 (2.35×10^{-6})	0.503 (1.60×10^{-33})	0.681 (1.70×10^{-46})	0.633 (1.89×10^{-49})
LapRLS	0.539 (3.29×10^{-5})	0.640 (3.43×10^{-29})	0.804 (1.08×10^{-38})	0.826 (1.81×10^{-38})
NetLapRLS	0.563 (7.04×10^{-4})	0.708 (4.64×10^{-13})	**0.900**	0.874 (2.17×10^{-22})
GIP	**0.604**	**0.727**	0.898 (5.70×10^{-3})	**0.884**
KBMF2K	0.508 (1.37×10^{-8})	0.686 (1.03×10^{-15})	0.876 (2.72×10^{-15})	0.796 (4.12×10^{-41})
NBI	0.425 (1.47×10^{-13})	0.619 (1.18×10^{-26})	0.832 (6.67×10^{-30})	0.783 (8.39×10^{-38})

Table 10: Methods summary

Method	DT	Techniques	Type	Train	Time test	Space	AUC			AUPR		
							d	t	p	d	t	p
NN	Yes	NN	S	O(I)	$O(n_d + n_t)$	$O(n_d n_t)$	6	6	6	6	6	7
BLM	Yes	SVM & BG	S	$O(n_d n_t (n_d^2 + n_t^2))$	O(I)	$O(n_d n_t)$	5	5	7	5	5	8
PKM	Yes	SVM & PK	S	$O(n_d^3 n_t^3)$	O(I)	$O(n_d^2 n_t^2)$	1	1	5	3	4	6
LapRLS	Yes	RLS	SS	$O(n_d^3 + n_t^3)$	O(I)	$O(n_d n_t)$	3	3	4	1	1	4
NetLapRLS	Yes	RLS	SS	$O(n_d^3 + n_t^3)$	O(I)	$O(n_d n_t)$	4	4	3	2	2	2
GIP	No	RLS & GK	SS	$O(n_d^3 + n_t^3)$	O(I)	$O(n_d n_t)$	–	–	1	–	–	1
KBMF2K	Yes	MF	S	$O(R(n_d^3 + n_t^3)t)$	O(I)	$O(n_d n_t)$	2	2	2	4	3	3
NBI	No	BG	–	O(I)	$O(n_d n_t)$	$O(n_d n_t)$	–	–	8	–	–	5

DT indicates whether drug (and target) prediction is possible. BG, PK, RLS, GK and MF indicate bipartite graphs, pairwise kernel, regularized least squares, Gaussian kernel and matrix factorization, respectively. S and SS indicate supervised learning and semi-supervised learning, respectively. Time and Space indicate time and space complexities, respectively, where n_d and n_t are the number of drugs and targets, respectively, whereas in KBMF2K, R and t are the feature space dimension and the number of iterations, respectively. The d, t and p in the column named AUC indicate AUCd, AUCt and AUCp, respectively, where all methods are ranked by their averaged AUC performance over four subsets. The same is followed in the column named AUPR.

and KBMF2K. We manually checked whether these top five interactions can be found in the latest online version of STITCH [65]. Table 11 lists the top five predictions for Nuclear receptor, where interactions found in STITCH were in boldface. From the table, we can see that all predicted interactions, except three cases, were already in the database (whereas they were not in the training data), demonstrating the high practical predicting ability of the three similarity-based methods.

DISCUSSION AND CONCLUSION

We have reviewed state-of-the-art similarity-based machine learning methods for predicting drug–target interactions. Our empirical results showed

Table 11: Top five predicted interactions by NetLapRLS, GIP and KBMF2K

Rank	Drug	Target
NetLapRLS		
1	**D00182 (Norethisterone)**	**hsa:2099 (ESR1)**
2	**D00348 (Isotretinoin)**	**hsa:5915 (RARB)**
3	**D00348 (Isotretinoin)**	**hsa:5916 (RARG)**
4	**D00075 (Testosterone)**	**hsa:5241 (PGR)**
5	**D00075 (Testosterone)**	**hsa:2099 (ESR1)**
GIP		
1	**D00182 (Norethisterone)**	**hsa:2099 (ESR1)**
2	D00316 (Etretinate)	hsa:6096 (RORB)
3	**D00075 (Testosterone)**	**hsa:5241 (PGR)**
4	D00327 (Testosterone)	hsa:5241 (PGR)
5	**D00348 (Isotretinoin)**	**hsa:5915 (RARB)**
KBMF2K		
1	**D00348 (Isotretinoin)**	**hsa:5915 (RARB)**
2	**D00348 (Isotretinoin)**	**hsa:5916 (RARG)**
3	**D00585 (Mifepristone)**	**hsa:2099 (ESR1)**
4	D01132 (Tazarotene)	hsa:2099 (ESR1)
5	**D00348 (Isotretinoin)**	**hsa:6256 (RXRA)**

that the method with the highest performance varies under different experimental settings and evaluation measures. For example, GIP outperformed other methods in both AUCp and AUPRp, whereas it cannot be applied to other settings. Thus for AUCt and AUCd, PKM and KBMF2K performed the best, whereas LapRLS was the best for AUPRt and AUPRd. Another finding was that drug prediction and target prediction were more difficult than pair prediction, because interactions by new drugs or new targets are unknown. Furthermore, we could see that target prediction was easier than drug prediction, which indicates that target similarity was more useful than drug similarity.

The assumption of similarity-based methods is that similar targets share similar drugs and vice versa. Low prediction performance on Nuclear receptor can be explained by this assumption: in Nuclear receptor, the average number of interacting targets per drug is the minimum among the four subsets. This is the case with the number of interacting drugs per target. For example, 26 targets and 90 drug–target interactions in Nuclear receptor mean the average number of interacting drugs per target is 3.46, whereas this number is 6.68, 7.23 and 4.40 for GPCR, Ion channel and Enzyme, respectively. Similarly, the average number of interacting targets per drug for Nuclear receptor is 1.66, whereas it is 2.84, 7.02 and 6.57 for GPCR, Ion channel and Enzyme, respectively.

We then focused on some particular interactions, which were in the training data but could not be predicted well, and examined why these interactions could not be predicted. Particularly, our focus was on the interactions, to which very low prediction scores were assigned by all three high-performance methods, NetLapRLS, GIP and KBMF2K. In Nuclear receptor, we found that, for example, D05341 (palmitate) and hsa:3174 (HNF4G), D00506 (phenobarbital) and hsa:9970 (NR1I3) and D00163 (chenodeoxycholic acid) and hsa:9971 (NR1H4), the three interactions had very low scores by all the three methods. Our finding from these three interactions was that each interaction of all these three pairs was the only interaction of the corresponding drug and the corresponding target. It is reasonable that this type of isolated interactions cannot be predicted well by similarity-based methods, which points out the weakness of similarity-based methods.

Here we raise general issues regarding machine learning methods for predicting drug–target interactions.

(i) **Negative sample selection**: Negative instances are currently all or randomly selected non-interacting drug–target pairs. Such negative instances might include potential drug–target interactions, which may be a possible reason why some concepts, such as polypharmacology or drug promiscuity, have been highlighted recently [1, 66]. In fact, experimentally measured negatives are not reported and unavailable, but they might improve the performance of prediction methods.

(ii) **Prediction tools or Web site**: There are already a great number of methods proposed for predicting drug–target interactions. Source programs or software of some methods are distributed, but more easy-to-use libraries or Web servers are unavailable so far. They would be definitely helpful for practitioners in drug discovery to use the current most powerful prediction methods.

(iii) **Interpretability in prediction results**: High AUC or AUPR would not be a sufficient condition for prediction methods. The prediction results should be comprehensible and hopefully provide some biological evidence, which can help biologists to decide which drugs or targets should be selected for the next biochemical verification.

Regarding similarity-based methods, although the methods have showed high AUC and AUPR in our experiments, the underlying assumption, i.e. that similar drugs share similar targets, can be a clear limitation. In recent times, most drug companies often select less costly and less risky ways to create new products. That is, companies start with compounds that are chemically similar to already known drugs and optimize the efficacy or reduce the side-effect of these drugs by modifying their structures within a limited range. As a result, chemical compounds of drug–target interactions are, in some cases, highly similar to each other. In reality, so-called 'me-too drugs' [67] have occupied a substantial portion among the drugs approved by the U.S. Food and Drug Administration. This situation makes the prediction evaluation of similarity-based methods too optimistic, which further makes it now difficult to develop totally innovative new drugs. Thus developing computational methods beyond the structure (or sequence) similarity would be a future direction. For example, using different types of data, such as side-effect similarity [35, 68], might be one promising line.

SUPPLEMENTARY DATA

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Machine learning-based approaches for predicting drug–target interactions, particularly similarity-based methods that use drug and target similarities, have attracted intensive interest.
- Different strategies have been adopted, and their predicting performances vary with different metrics and prediction tasks.
- Overall, for predicting interactions of new drugs or targets, in terms of AUC, PKM performed the best, and in terms of AUPR, LapRLS performed the best; for predicting interactions of known drugs or targets, GIP was the best method.
- *In silico* prediction of drug–target interaction can be further improved by enhancing the procedure used to select negative samples, developing user-friendly prediction tools or Web sites and exploring the interpretability in prediction results.

FUNDING

The National Nature Science Foundation of China (No. 61170097), Scientific Research Starting Foundation for Returned Overseas Chinese Scholars, Ministry of Education, China, and Japan Society for the Promotion of Science (JSPS) Invitation Fellowship. KAKENHI (Nos.23710233 and 24300054) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. Shanfeng Zhu would like to thank the China Scholarship Council for the financial support on his visit at University of Illinois at Urbana-Champaign.

References

1. Hopkins AL. Drug discovery: predicting promiscuity. *Nature* 2009;**462**:167–8.
2. Lounkine E, Keiser MJ, Whitebread S, *et al*. Large-scale prediction and testing of drug activity on side-effect targets. *Nature* 2012;**486**:361–7.
3. Pauwels E, Stoven V, Yamanishi Y. Predicting drug side-effect profiles: a chemical fragment-based approach. *BMC Bioinformatics* 2011;**12**:169.
4. Dudley JT, Deshpande T, Butte AJ. Exploiting drug-disease relationships for computational drug repositioning. *Brief Bioinform* 2011;**12**(4):303–11.
5. Swamidass SJ. Mining small-molecule screens to repurpose drugs. *Brief Bioinform* 2011;**12**(4):327–35.
6. Moriaud F, Richard SB, Adcock SA, *et al*. Identify drug repurposing candidates by mining the Protein Data Bank. *Brief Bioinform* 2011;**12**(4):336–40.
7. Hopkins AL, Groom CR. The druggable genome. *Nat Rev Drug Discov* 2002;**1**(9):727–30.
8. Dobson CM. Chemical space and biology. *Nature* 2004;**432**:824–8.
9. Kanehisa M, Goto S, Hattori M, *et al*. From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006;**34**:D354–7.
10. Stockwell BR. Chemical genetics: ligand-based discovery of gene function. *Nat Rev Genet* 2000;**1**(2):116–25.
11. Sayers EW, Barrett T, Benson DA, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2012;**40**:D13–25.
12. Whitebread S, Hamon J, Bojanic D, *et al*. Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development. *Drug Discov Today* 2005;**10**(21):1421–33.
13. Haggarty SJ, Koeller KM, Wong JC, *et al*. Multidimensional chemical genetic analysis of diversity-oriented synthesis-derived deacetylase inhibitors using cell-based assays. *Chem Biol* 2003;**10**(5):383–96.
14. Kuruvilla FG, Shamji AF, Sternson SM, *et al*. Dissecting glucose signaling with diversity-oriented synthesis and small-molecule microarrays. *Nature* 2002;**416**:653–7.
15. Manly CJ, Louise-May S, Hammer JD. The impact of informatics and computational chemistry on synthesis and screening. *Drug Discov Today* 2001;**6**(21):1101–10.
16. Shoichet BK, Bodian DL, Kuntz ID. Molecular docking using shape descriptors. *J Comput Chem* 1992;**13**(3):380–97.
17. Rarey M, Kramer B, Lengauer T, *et al*. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* 1996;**261**(3):470–89.
18. Halperin I, Ma B, Wolfson H, *et al*. Principles of docking: an overview of search algorithms and a guide to scoring functions. *Proteins* 2002;**47**(4):409–43.

19. Shoichet BK, McGovern SL, Wei B, *et al*. Lead discovery using molecular docking. *Curr Opin Chem Biol* 2002;**6**(4): 439–46.
20. Cheng AC, Coleman RG, Smyth KT, *et al*. Structure-based maximal affinity model predicts small-molecule druggability. *Nat Biotechnol* 2007;**25**(1):71–5.
21. Ballesteros J, Palczewski K. G protein-coupled receptor drug discovery: implications from the crystal structure of rhodopsin. *Curr Opin Drug Discov Devel* 2001;**4**(5):561–74.
22. Klabunde T, Hessler G. Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem* 2002; **3**(10):928–44.
23. Nagamine N, Sakakibara Y. Statistical prediction of protein chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics* 2007;**23**(15):2004–12.
24. Nagamine N, Shirakawa T, Minato Y, *et al*. Integrating statistical predictions and experimental verifications for enhancing protein-chemical interaction predictions in virtual screening. *PLoS Comput Biol* 2009;**5**(6):e1000397.
25. Yabuuchi H, Niijima S, Takematsu H, *et al*. Analysis of multiple compound-protein interactions reveals novel bioactive molecules. *Mol Syst Biol* 2011;**7**:472.
26. Yamanishi Y, Araki M, Gutteridge A, *et al*. Prediction of drug-target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics* 2008;**24**(13): i232–40.
27. Bleakley K, Yamanishi Y. Supervised prediction of drug-target interactions using bipartite local models. *Bioinformatics* 2009;**25**(18):2397–403.
28. Jacob L, Vert JP. Protein-ligand interaction prediction: an improved chemogenomics approach. *Bioinformatics* 2008; **24**(19):2149–56.
29. Xia Z, Wu LY, Zhou X, *et al*. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. *BMC Syst Biol* 2010;**4(Suppl 2)**:S6.
30. Van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 2011;**27**(21):3036–43.
31. Gonen M. Predicting drug-target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* 2012;**28**(18):2304–10.
32. Perlman L, Gottlieb A, Atias N, *et al*. Combining drug and gene similarity measures for drug-target elucidation. *J Comput Biol* 2011;**18**(2):133–45.
33. Yamanishi Y, Kotera M, Kanehisa M, *et al*. Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. *Bioinformatics* 2010;**26**(12):i246–54.
34. Zhu S, Okuno Y, Tsujimoto G, *et al*. A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics* 2005;**21(Suppl 2)**: ii245–51.
35. Kuhn M, Campillos M, Gonzalez P, *et al*. Large-scale prediction of drug-target relationships. *FEBS Lett* 2008;**582**(8): 1283–90.
36. Iskar M, Zeller G, Zhao XM, *et al*. Drug discovery in the age of systems biology: the rise of computational approaches for data integration. *Curr Opin Biotechnol* 2012; **23**(4):609–16.
37. Koutsoukas A, Simms B, Kirchmair J, *et al*. From in silico target prediction to multi-target drug design: current databases, methods and applications. *J Proteomics* 2011; **74**(12):2554–74.
38. Sanseau P, Koehler J. Editorial: computational methods for drug repurposing. *Brief Bioinform* 2011;**12**(4):301–2.
39. Schomburg I, Chang A, Ebeling C, *et al*. BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res* 2004;**32**:D431–3.
40. Gunther S, Kuhn M, Dunkel M, *et al*. SuperTarget and matador: resources for exploring drug-target relationships. *Nucleic Acids Res* 2008;**36**:D919–22.
41. Wishart DS, Knox C, Guo AC, *et al*. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res* 2008;**36**:D901–6.
42. Liu Y, Hu B, Fu C, *et al*. DCDB: drug combination database. *Bioinformatics* 2010;**26**(4):587–8.
43. Brooksbank C, Cameron G, Thornton J. The European Bioinformatics Institute's data resources. *Nucleic Acids Res* 2010;**38**:D17–25.
44. Kuhn M, Campillos M, Letunic I, *et al*. A side effect resource to capture phenotypic effects of drugs. *Mol Syst Biol* 2010;**6**:343.
45. Lamb J, Crawford ED, Peck D, *et al*. The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* 2006;**313**: 1929–35.
46. Breitkreutz BJ, Stark C, Reguly T, *et al*. The BioGRID interaction database: 2008 update. *Nucleic Acids Res* 2008; **36**:D637–40.
47. Ewing RM, Chu P, Elisma F, *et al*. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol* 2007;**3**:89.
48. Rual JF, Venkatesan K, Hao T, *et al*. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;**437**:1173–8.
49. Stelzl U, Worm U, Lalowski M, *et al*. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 2005;**122**(6):957–68.
50. Xenarios I, Salwinski L, Duan XJ, *et al*. DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res* 2002; **30**(1):303–5.
51. Ashburner M, Ball CA, Blake JA, *et al*. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000;**25**(1):25–9.
52. Jain E, Bairoch A, Duvaud S, *et al*. Infrastructure for the life sciences: design and implementation of the UniProt website. *BMC Bioinformatics* 2009;**10**:136.
53. Takigawa I, Tsuda K, Mamitsuka H. Mining significant substructure pairs for interpreting polypharmacology in drug-target network. *PLoS One* 2011;**6**(2):e16999.
54. Mitchell JB. The relationship between the sequence identities of alpha helical proteins in the PDB and the molecular similarities of their ligands. *J Chem Inf Comput Sci* 2001;**41**(6): 1617–22.
55. Schuffenhauer A, Floersheim P, Acklin P, *et al*. Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comput Sci* 2003;**43**(2): 391–405.
56. Klabunde T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br J Pharmacol* 2007; **152**(1):5–7.

57. Bleakley K, Biau G, Vert JP. Supervised reconstruction of biological networks with local models. *Bioinformatics* 2007; **23**(13):i57–65.

58. Mordelet F, Vert JP. SIRENE: supervised inference of regulatory networks. *Bioinformatics* 2008;**24**(16):i76–82.

59. Vapnik VN. *Statistical Learning Theory*. New York: Wiley, 1998.

60. Cheng F, Liu C, Jiang J, *et al*. Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput Biol* 2012;**8**(5):e1002503.

61. Hattori M, Okuno Y, Goto S, *et al*. Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J Am Chem Soc* 2003;**125**(39):11853–65.

62. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**(1):195–7.

63. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. *ICML'06 Proceedings of the 23rd international conference on machine learning Pittsburgh, USA, 2006*. pp. 233–40. ACM, New York, 2006.

64. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**(3): Article 27.

65. Kuhn M, Szklarczyk D, Franceschini A, *et al*. STITCH 3: zooming in on protein-chemical interactions. *Nucleic Acids Res* 2012;**40**:D876–80.

66. Keiser MJ, Setola V, Irwin JJ, *et al*. Predicting new molecular targets for known drugs. *Nature* 2009;**462**:175–81.

67. Garattini S. Are me-too drugs justified? *J Nephrol* 1997; **10**(6):283–94.

68. Campillos M, Kuhn M, Gavin AC. Drug target identification using side-effect similarity. *Science* 2008;**321**:263–6.