Bioinformatics, 34(7), 2018, 1164–1173
doi: 10.1093/bioinformatics/btx731
Advance Access Publication Date: 24 November 2017
Original Paper

Systems biology

DDR: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches

Rawan S. Olayan[1], **Haitham Ashoor**[2] **and Vladimir B. Bajic**[1,*]

[1]King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, Thuwal, Saudi Arabia and [2]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut 06032, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on August 3, 2017; revised on October 10, 2017; editorial decision on November 8, 2017; accepted on November 23, 2017

Abstract

Motivation: Finding computationally drug–target interactions (DTIs) is a convenient strategy to identify new DTIs at low cost with reasonable accuracy. However, the current DTI prediction methods suffer the high false positive prediction rate.

Results: We developed DDR, a novel method that improves the DTI prediction accuracy. DDR is based on the use of a heterogeneous graph that contains known DTIs with multiple similarities between drugs and multiple similarities between target proteins. DDR applies non-linear similarity fusion method to combine different similarities. Before fusion, DDR performs a pre-processing step where a subset of similarities is selected in a heuristic process to obtain an optimized combination of similarities. Then, DDR applies a random forest model using different graph-based features extracted from the DTI heterogeneous graph. Using 5-repeats of 10-fold cross-validation, three testing setups, and the weighted average of area under the precision-recall curve (AUPR) scores, we show that DDR significantly reduces the AUPR score error relative to the next best start-of-the-art method for predicting DTIs by 34% when the drugs are new, by 23% when targets are new and by 34% when the drugs and the targets are known but not all DTIs between them are not known. Using independent sources of evidence, we verify as correct 22 out of the top 25 DDR novel predictions. This suggests that DDR can be used as an efficient method to identify correct DTIs.

Availability and implementation: The data and code are provided at https://bitbucket.org/RSO24/ddr/.

Contact: vladimir.bajic@kaust.edu.sa

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

1 Introduction

Drug discovery is the process through which potential beneficial treatment effects or medical uses of a new drug candidate are identified. Distinct phases of drug discovery and development define the initial stage of target identification and validation, compound leads identification, validation and optimization, and different types of preclinical and clinical trials until the final approval by the Food and Drug Administration (FDA) (Paul *et al.*, 2010) is reached. Drugs function through interaction with various molecular targets. We call such interaction drug–target interactions (DTIs). Proteins are one useful group of such targets. Through binding, drugs can either enhance or inhibit functions carried out by proteins (Overington *et al.*, 2006; Santos *et al.*, 2017) and thus affect the disease conditions. Bringing a new drug to the market is a highly challenging and complex process in terms of time and cost. Moreover, the number of newly approved drugs by the FDA is decreasing, illustrating the productivity decline in drug discovery and development (Swinney and Anthony, 2011). However, studies showed that most of the FDA-approved drug molecules exhibit poly-pharmacological properties, i.e. drugs can have interaction with multiple targets, which are not their primary therapeutic targets (i.e. drugs have off-target molecules) (Cichonska *et al.*, 2015), and this is frequently the major cause of undesirable side-effects. One interesting and useful objective is to link the newly identified DTIs of a known drug to the treatment of diseases that are different from diseases for which the drug has been originally developed (Cichonska *et al.*, 2015; Li *et al.*, 2016; Shim and Liu, 2014). The availability of public biomedical databases along with the development of computational approaches has made it possible to provide useful frameworks to partially overcome limitations of the traditional experimental approaches (Vilar and Hripcsak, 2016) and help in finding a new association for the existing drugs with off-target effects. Identifying computationally highly likely DTIs for a known drug can be then employed to identify potential new uses of the drug in question, and this makes a useful strategy in drug repurposing (Chen *et al.*, 2016; Lu *et al.*, 2017; Wu *et al.*, 2017). A part of such solution is the identification of novel DTIs that play an important role in the discovery of additional applications for known drugs, as well as in the understanding of drug's modes of action (Overington *et al.*, 2006; Santos *et al.*, 2017; Schenone *et al.*, 2013). This necessitates development of accurate computational approaches to focus on a smaller number of highly likely targets of a drug for the follow-up experimental validation. However, predicting correct DTI is not sufficient for itself to infer what effect such interaction may have. Additional steps may be needed, such as, for example, to show inhibition of target expression. One approach for computationally inferring such effects may be the utilization of predictive models of activity in appropriate biological assays (Soufan *et al.*, 2016; Soufan *et al.*, 2015) as those in the PubChem resource. As summarized in recent reviews (Chen *et al.*, 2016; Santos *et al.*, 2017), a wide range of databases, web tools and computational methods have emerged with the potential to predict DTIs by learning from interaction data supplemented with information on the similarities among drugs and similarities among proteins (Lu *et al.*, 2017; Wu *et al.*, 2017). However, confirming whether a drug could interact with a target protein requires an additional effort. This is owing to the relatively limited information about interactions between drugs and target proteins (Dobson, 2004; Kanehisa *et al.*, 2006; Menni *et al.*, 2017), as well as the poor characterization of proteins as drug targets (Santos *et al.*, 2017).

Early attempts in computational prediction of DTIs can be categorized into two main groups and include docking simulations and ligand-based approaches (Cheng *et al.*, 2007; Keiser *et al.*, 2007). Docking methods consider the 3D structure of target proteins. However, this approach is extensively time-consuming, and the structural information of targets is not available for all target proteins. Ligand-based methods compare a query ligand with a set of known ligands with target proteins. However, it may not perform well in cases the number of known ligands with target proteins is small.

Public data sources have promoted the development of various strategies for repurposing drugs including genome, phenome, drug chemical structures, biological interactome, biomedical literature text and biological bioassays (Li *et al.*, 2016). Moreover, the accessibility of big data sources, through several databases and biomedical literature of DTI information, provide a useful way to extract different biological interaction profiles and signatures (or descriptors) of drugs and target proteins to discover novel DTIs (Ba-Alawi *et al.*, 2016; Cheng *et al.*, 2012a,b; Ding *et al.*, 2014; Mitchell, 2001; Perlman *et al.*, 2011; Vilar and Hripcsak, 2016; Wu *et al.*, 2017; Yamanishi *et al.*, 2010). On the basis of the guilt-by-association principle, in which chemically similar drugs tend to interact with similar proteins, many methods have been proposed for DTI prediction based on the consideration of similarity measures between drugs or similarities between proteins. Such prediction methods are based on graph inference (Alaimo, 2013; Ba-Alawi *et al.*, 2016; Bleakley and Yamanishi, 2009; Chen *et al.*, 2012a,b; Seal *et al.*, 2015; Wang *et al.*, 2013), machine-learning algorithms (Hao *et al.*, 2017; Lim *et al.*, 2016; Liu *et al.*, 2016; Mei *et al.*, 2013; Perlman *et al.*, 2011; Soufan *et al.*, 2016; van Laarhoven *et al.*, 2011; Yuan *et al.*, 2016), text mining (Zhu *et al.*, 2005) and semantic linked data (Chen *et al.*, 2012a; Fu *et al.*, 2016; Tari and Patel, 2014; Zhu *et al.*, 2014).

Recently, several methods are developed to integrate heterogeneous information related to the drug, target protein, and their interaction data, to provide effective and efficient ways to predict new DTIs (Hao *et al.*, 2017; Mei *et al.*, 2013). These methods utilize various types of profiles for drugs and proteins constructed with different biological data. Such DTI prediction methods were developed based on the idea of utilizing heterogeneous networks of known DTIs, similarity between drugs and similarity between target proteins (Hao *et al.*, 2017; Nascimento *et al.*, 2016; Perlman *et al.*, 2011; Zong *et al.*, 2017). These methods demonstrate that utilizing different measures of similarity between drugs and target proteins results in improved performance compared to other methods that are based on using only single similarity for drugs and single similarity for target proteins. Moreover, a prediction method (Hao *et al.*, 2017) that is based on non-linear integration of similarity measures shows better performance than the other methods based on the linearly combined similarity measures (Mei *et al.*, 2013; Nascimento *et al.*, 2016; van Laarhoven *et al.*, 2011). However, these studies indicated that the DTI prediction performance of different methods varies significantly with and depends heavily on the similarity measures used. These require development of computational methods that optimize combination of multiple similarity measures with the aim to improve the DTI prediction accuracy (for more detailed information see Supplementary Material, Related work).

In this study, aiming to further improve the accuracy of DTI prediction, we developed DDR, an efficient DTI prediction method that in a novel way determines through a heuristic method an optimized combination of similarity measures between drugs and between target proteins used in the prediction model. To predict DTIs, DDR integrates information from different types of drug–drug and target–target similarity measures and then, it applies a random forest (RF) model using graph-based features. On different representative datasets and under various test setups, and using different performance measure, we show that DDR significantly outperforms the other state-of-the-arts methods by dramatically reducing the error. Using independent sources of evidence, we verified as correct 22 out of the top 25 DDR novel predictions. This suggests that DDR can be used as an efficient method to identify correct DTIs.

2 Materials

2.1 Datasets

2.1.1 DTI data

Five datasets were used to evaluate the performance of the proposed DDR method in DTI prediction. Each dataset contains three types of information: (i) the known DTIs for humans, (ii) multiple drug similarity measures and (iii) multiple target proteins similarity measures.

A frequently considered gold standard dataset (we name it Yamanishi_08) was originally compiled by Yamanishi *et al.* (2008) and was used as a reference in many studies (Ba-Alawi *et al.*, 2016; Lim *et al.*, 2016; Lu *et al.*, 2017; Mei *et al.*, 2013). This dataset contains known DTIs as retrieved from KEGG BRITE (Kanehisa *et al.*, 2006), BRENDA (Schomburg *et al.*, 2004), SuperTarget (Gunther *et al.*, 2008) and DrugBank databases (Wishart *et al.*, 2008). In Yamanishi_08, the information on DTI is classified according to the target proteins of drugs into the following four groups: (i) enzymes (E), (ii) ion channels (IC), (iii) G-protein-coupled receptors (GPCR) and (iv) nuclear receptors (NR). Thus, Yamanishi_08 dataset is composed of the four datasets corresponding to the classes of target proteins.

The fifth dataset is DrugBank_FDA, which is extracted from 5.0.3 version of DrugBank database (Wishart *et al.*, 2008). We only extracted DTI information of drugs approved by the FDA and single human target proteins; these proteins are not part of protein complexes. Table 1 summarizes the statistics of these datasets. Note that, the ratios of known (positive) versus non-existing (not known, negative) DTIs in all datasets are variable. This reflects practical situations where the number of true DTIs is considered to be much smaller than that of non-interacting drug–targets.

2.1.2 Similarity measures for drugs and for target proteins

We computed multiple similarity measures for drugs and for target proteins, respectively, where all similarity values were normalized to the range [0, 1].

For the first four benchmark datasets from Yamanishi_08, the similarities between drug pairs and between target protein pairs were calculated based on information from different sources and from Nascimento *et al.* (2016). For drugs, distinct chemical structure fingerprints, side-effects profiles and the Gaussian interaction profile (GIP) were considered as drug information sources for calculation of the drug similarities. On the other hand, the similarities of target proteins were calculated based on various amino acid sequence profiles of proteins, as well as different parameterizations of the Mismatch (MIS) and the Spectrum (SPEC) kernels, target proteins functional annotation based on Gene Ontology (GO) terms, proximity within the protein–protein interaction (PPI) network and the GIP for target proteins.

For the fifth benchmark dataset, DrugBank_FDA, we computed different similarity measures between drugs based on: different types of molecular fingerprints, drug interaction profile, drug side-effects profile, drug profile of the anatomical therapeutic class (ATC) coding system, drug-induced gene expression profile, drug disease profiles, drug pathways profiles and GIP. Furthermore, different target protein similarity measures were calculated based on protein amino acid sequence, their GO annotations, proximity in the PPI network, GIP, protein domain profiles and gene expression similarity profiles of protein encoding genes. Chemical structures of drugs were extracted from DrugBank (Wishart *et al.*, 2008), while the target protein sequences were extracted from UniProt (Boutet *et al.*, 2016). Supplementary Table S1 shows the summary of multiple similarity measures calculated for drugs and target proteins in the DrugBank_FDA dataset, as well as describing their importance and tools used to calculate them.

As a summary, all different similarity measures between drugs and between target proteins for the first four datasets are recomputed/available and collected from Nascimento *et al.* (2016). For DrugBank_FDA dataset, all different similarity measures between drugs and between target proteins are calculated in this study, since there is no available similarity measures data obtained for such dataset.

3 Methods

3.1 Problem description

We define a set of DTIs, which consists of a set of drugs D and a set of target proteins T, where $D = \{d_i, i = 1, \ldots, m\}$ and $T = \{t_j, j = 1, \ldots, n\}$, in which m represents the number of drugs and n represents the number of target proteins. The interactions between D and T are represented as a binary matrix Y such that if d_i interacts with t_j, then $y_{ij} = 1$, if not then $y_{ij} = 0$. We also define the set of similarity matrices between drugs in D as D_s, where similarity matrices have dimensions of m x m; we define the set of similarity matrices between target proteins in T as T_s, where similarity matrices have dimensions of n x n. Element values in different similarity matrices represent how much are drugs or target proteins similar to each other based on different measures. All elements in each matrix have values in the range of [0, 1]. A similarity value close to 0 indicates that two elements are not similar to each other while a similarity value close to 1 represents the most similar elements. Given the matrix Y, and matrices in D_s and T_s, our goal is to predict novel (i.e. unknown) interactions in Y.

3.2 Description of the DDR method

The heterogeneous DTI graph is a weighted graph that is constructed with m nodes from the drug set and n nodes from the set of target proteins. The edge between two drug nodes or two target protein nodes represents the similarity between them and is weighted by the similarity value obtained from the similarity fusion step. The edge between a drug and a target protein represents a known DTI and is weighted by 1. A path structure of a path that starts at a D node and ends up at a T node describes a subgraph that sequentially links of drug and target protein nodes. For example, a path Drug$_1$–Drug$_2$–Target$_1$ connects the Drug$_1$ node with the Target$_1$ node through the similarity edge between Drug$_1$ and Drug$_2$ and via the interaction edge between Drug$_2$ and Target$_1$. The path structure of this path is D–D–T. All paths with more than one edge and without loops, starting at a D node and ending at a T node, and having the same path structure define a path-category on the heterogeneous DTI graph.

DDR workflow (Fig. 1) depicts several steps including: (i) inferring interaction profile for new drugs and for new target proteins,

Table 1. Summary of the five datasets (Yamanishi_08 and DrugBank_FDA) used in this study

Datasets	Target classes	Number of drugs	Number of target proteins	Number of known DTIs
Yamanishi_08	NR	54	26	90
	GPCR	223	95	635
	IC	210	204	1476
	E	445	664	2926
DrugBank_FDA	Multi-class	1482	1408	9881

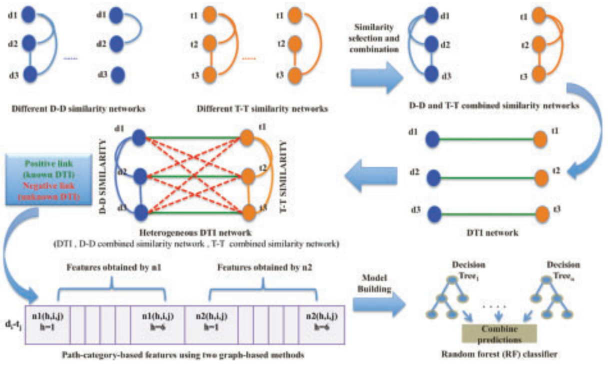
d1
d2
d3
t1
t2
t3
Different D-D similarity networks
Different T-T similarity networks
Similarity selection and combination
D-D and T-T combined similarity networks
Positive link (known DTI)
Negative link (unknown DTI)
D-D SIMILARITY
T-T SIMILARITY
Heterogeneous DTI network
(DTI , D-D combined similarity network , T-T combined similarity network)
DTI network
Features obtained by n1
Features obtained by n2
Model Building
Decision Tree1
Decision Treen
d_i-t_j
n1(h,i,j) h=1
n1(h,i,j) h=6
n2(h,i,j) h=1
n2(h,i,j) h=6
Path-category-based features using two graph-based methods
Combine predictions
Random forest (RF) classifier

Fig. 1. Flowchart of DDR method. DDR consists of several steps including: (i) Similarity selection, where a subset of similarity measures is selected in a heuristic process. (ii) Similarity fusion, with the goal to combine the selected similarity measures into one final composite similarity that combines information from similarities determined in (i). (iii) Path-category-based feature extraction, where the feature vector corresponds to drug and target protein pairs, i.e. for (d_i, t_j) pair, features are determined as the vector composed of the 12 (*i, j*) elements obtained by two graph-based scores, namely, *n*1(*h, i, j*) and *n*2(*h, i, j*) for each specific path-category $C_{h,}$ $h = 1, 2, \ldots, 6$. (iv) Building DTI prediction model using RF, where both positive and negative data are provided; positive data contain known links between drugs and target proteins and represent positive labels, while negative data contain unknown DTI links that are treated as negative labels

(ii) similarity measure selection, (iii) similarity fusion, (iv) path-category-based features extraction, (v) building DTI prediction model using RF.

3.2.1 Inferring interaction profile

Inferring the DTIs profiles for new drugs and target proteins is used only with the GIP similarity calculation. A drug is called new if it does not have any known target protein to interact with, while a target protein is called new if it is not targeted by any known drug. Since the GIP similarity is constructed based on training DTI data only, the GIP similarity cannot be computed for drugs or target proteins that do not have known DTIs in the training data. So, we enhance the GIP similarity calculation by inferring interaction profiles for new drugs and for new target proteins, in cases where DTIs for drugs or for target proteins are missing from the training data. This inference is made based on the interaction profiles of such drugs or target proteins. Drugs (or target proteins) with high similarities to a new drug (or a new target protein) are said to be the neighbors of the drug (the target protein). This interaction profile inferring technique is based on Mei *et al.* (2013). For example, the inferred value of interaction for a new drug with a specific target protein is represented as the ratio of the sum of similarity values for drug neighbors interacting with this target protein relative to the total sum of all neighbors' similarity values. For DDR, we subjectively set the number of neighbors to 5.

3.2.2 Similarity selection: selection of an optimized set of similarities

Combining all similarity types may introduce noise in the data as some similarities have more information than others. In order to select a more robust similarity set, we implement similarity selection procedure (Supplementary Figure S1) that is able to select a set of informative and less-redundant set of similarities for drugs and for target proteins, separately. This is done through a heuristic process, where a subset of similarity measures is selected forming an optimized (possibly the best) combination of similarities for our problem.

To select set of informative similarities, our procedure goes as follows:

(i) Calculate the average entropy for each similarity matrix to determine how much information each similarity carries. For a similarity matrix *M* (target–target similarity or drug–drug similarity) of size $k \times k$, where k represents the number of drugs (or target proteins), with elements $m_{i,j}$, we calculate entropy E_i for each row i as:

$$E_i = -\sum\nolimits_{j=1}^{k} p_{ij} \log\left(p_{ij}\right), \text{ where}$$

$$p_{ij} = \frac{m_{ij}}{\sum\limits_{j=1}^{k} m_{ij}}.$$

After that, we average the entropy values of all matrix rows to get the final average entropy value that describes how informative a similarity matrix is. (ii) Rank the matrices according to their average entropy values in ascending order. The lower the average entropy value is the less random information the similarity matrix carries. Then, remove similarity matrices with high average entropy that contain more random information with average entropy value greater than $c_1 \log(k)$, where c_1 is a constant that controls how much information each similarity carries; thus, c_1 controls level of entropy to be selected; $\log(k)$ represents the maximum entropy value. (iii) Calculate the pairwise similarity measure between similarity matrices from different data sources, based on the Euclidean distance, as follows. To assess the information overlap between any two similarity matrices, we constructed the pairwise similarity matrix between all similarity measures based on Euclidean distance as follows: given two matrices for similarity measures A and B, we reorganize each similarity matrix into vectors (V_A and V_B) and then compute the Euclidean distance d as

$$d = \sqrt{\sum\nolimits_{i=1}^{k^2} \left(V_{A_i} - V_{B_i}\right)^2}.$$

We converted distance values to similarity E_s as

$$E_s = \frac{1}{1 + d}.$$

(iv) After obtaining a set of informative similarities matrices, the redundant similarity matrices are removed as follows: the procedure starts with the similarity measure matrix having the lowest average entropy value and eliminates all other similarity measure matrices with E_s value larger than a threshold c_2. After that, the procedure continues with the next similarity matrix in the ranked list until the whole list of the similarity matrices is exhausted. At the end, the remaining list of similarity measures is reported as the selected set with small redundancy of informative similarity measures for drugs and target proteins. In this study, we subjectively set c_1 to 0.7 and c_2 to 0.6. We applied this procedure to select the set of informative less-redundant similarity measures of drugs and target proteins, separately.

3.2.3 Similarity fusion

Given the selected subsets of similarity measures obtained previously for drugs and for target proteins, respectively, the goal of the similarity fusion step is to combine multiple similarity measures into one final composite similarity that captures the necessary information from different similarities. Thus, given a set of multiple similarity measures for drugs and for target proteins, respectively, we computed the final fused similarity measure following the similarity network fusion (SNF) method developed in Wang *et al.* (2014). We represent each similarity measure by $k \times k$ similarity matrix $M = (m_{i,\ j})$, where $m_{i,\ j}$ equals to the similarity value between d_i/d_j or t_i/t_j indicating how much they are similar.

The SNF combines multiple similarity measures into a single fused similarity by a nonlinear method based on message-passing theory. It iteratively updates every similarity network with information from the other networks, using *K*-nearest neighbors, making it more similar to the others. The SNF method can capture common as well as complementary information across different measures of similarities. We applied the SNF method to integrate multiple drug–drug similarities and target–target similarities, separately.

3.2.4 Path-category-based features

After obtaining the combined similarity for drugs and for target proteins, respectively, we augmented the combined similarities with the known DTIs to construct a heterogeneous DTI graph. Based on this heterogeneous graph, we extracted 12 path-category-based features that we used to build a DTI prediction model. In this study, we work with path-categories of lengths 2 and 3 (but not longer, because of the computational cost). When we restrict paths to start at drug nodes and end at target protein nodes, there are only two path-categories with paths of length 2, having path structures (D–D–T) and (D–T–T), and four path-categories with paths of length 3, having path structures (D–D–D–T), (D–D–T–T), (D–T–D–T) and (D–T–T–T). Thus, we will consider these six path-categories through which drug nodes could connect to target protein nodes. We define matrices $S1_h$ and $S2_h$ associated with each path-category $C_{h,}$ $h = 1, 2, \ldots, 6$, that we consider. To do this, we start with a given drug d_i to reach a given target protein t_j through a specific path-category C_h. We restrict traversing the graph to retrieve all paths passing only through the *K*-nearest neighbors of drugs to d_i and only through the *K*-nearest neighbors of target proteins to t_j. In this study, we subjectively set the number of nearest neighbors K to 5. The set of such paths we denote as R_{ijh}. Next, for each path p_q from $R_{ijh,}$ we calculate an edge-weight product value s obtained by multiplying all weights w_x of edges of p_q as follows:

$$s(h,i,j,q) = \prod_{\forall w_x \ \in \ p_q,\ \ p_q \ \in \ R_{ijh}} w_x.$$

Using the s values calculated for all paths p_q from R_{ijh}, we calculate scores $s1$ and $s2$ as follows:

$$s1(h,i,j) = \sum_{\forall q:p_q \ \in \ R_{ijh}} s(h,i,j,q).$$

Thus, for each path-category C_h, we obtained a matrix $S1_h$ with elements $s1(h,\ i,\ j)$. Also, for each path-category C_h, we obtained a matrix $S2_h$ with elements $s2(h,\ i,\ j)$ determined as:

$$s2(h,i,j) = \ \max_{\forall q:p_q \ \in \ R_{ijh}}(s(h,i,j,q)).$$

Finally, we normalized matrices $S1_h$ and $S2_h$ to adjust for the overall connectivity of the network, where the elements of the normalized matrices are:

$$nr(h,i,j) = \frac{sr(h,i,j)}{\sum_j sr(h,i,j)},$$

where $r = 1$ or 2. The normalized matrices are now $N1_h$ with elements $n1(h,\ i,\ j)$ and $N2_h$ with elements $n2(h,\ i,\ j)$ calculated as shown above.

In total, DDR defines 12 different path-category-based matrices, namely $N1_h$, $N2_h$, where $h = 1, 2, \ldots, 6$, which contain feature values. These matrices have the same number of rows (corresponding to drugs) and the same number of columns (corresponding to target proteins).

3.2.5 RF classification model for DTI prediction

To predict DTI, DDR utilizes supervised machine learning model based on the RF classifier (Ho, 1995). RF has been shown to be an effective tool in prediction, as it runs efficiently on large datasets and is less prone to over-fitting. We implemented the RF predictive model using scikit-learn (Pedregosa *et al.*, 2011). The inputs to the RF correspond to drug and target protein pairs, i.e. for the (d_i, t_j) pair, the feature vector is determined as the vector composed of the $(i,\ j)$ elements of matrices $N1_h$ and $N2_h$. Since $h = 1, 2, \ldots, 6$, these feature vectors contain 12 elements each. In order to learn from highly imbalanced data, in this study we adjusted the RF class weights to be inversely proportional to the number of class labels for each class in the training data. Two important parameters are set when building the RF model: The number of trees in the forest (n_estimators) was set to be in the range of [100, 600] trees and a function to measure the quality of a split (criterion) where we used Gini index and entropy based functions. To construct the prediction model, both positive and negative data are provided as either known DTIs to represent positive labels or unknown DTIs that are treated as negative labels.

3.3 Experimental setting and performance evaluation

To facilitate the comparison with other methods, we performed cross-validation (CV) and hold-out type tests. First, we evaluated the performance of the DDR method for DTI prediction using CV experiments obtained under three different settings of prediction tasks as in Pahikkala *et al.* (2015). The experiments were performed separately for each dataset used in this study (the four gold standard datasets from Yamanishi_08 and the DrugBank_FDA dataset). The three prediction settings correspond to the cases when: (a) the drugs are new, (b) the target proteins are new and (c) the drugs and their target proteins are known but all interactions between them are not necessarily known. Cases (a) and (b) correspond to the situation when there are no DTIs in the training data for such drugs or target proteins, while case (c) corresponds to the situation when there are DTI in the training data for such drugs or target proteins. We name settings (a), (b) and (c) as S_{D}, S_{T} and S_{P}, respectively.

For each dataset, a prediction model in each setting is built using a dataset of positive and negative labels split into the training and testing sets. This procedure is followed for each fold in 10-fold CV and the whole process is repeated 5 times, each time with a different random seed used for random selection for the split into training and testing sets. In each fold of the CV process, all interactions y_{ij} in Y matrix that belong to the testing set in that fold are set to zero, i.e. they were excluded from consideration. The resulting matrix with removed testing DTIs is Y_{train}. In each fold, the model learns interactions from Y_{train} and then constructs the GIP similarity. Then, we select the best set of similarity measures (according to DDR's heuristic procedure). After that, we use all selected similarities separately for drugs and separately for target proteins, to generate a fused similarity matrix for drugs and a fused similarity matrix for target proteins. Based on Y_{train} and the two generated fused similarity matrices, we construct a heterogeneous DTI graph, where we extract path-category-based features as explained before, and score them using two graph-based scores. Finally, we train the RF model on the training set for that fold until the best area under the precision-recall (AUPR) is obtained. Then, using the trained model, we predict and evaluate predictions on the testing set for that fold.

Moreover, we performed the hold-out tests derived from 9881 DTIs from DrugBank_FDA dataset under the same S_D, S_T and S_P settings of prediction tasks. In the hold-out test, we split the data into 80% for training and 20% for testing.

For each prediction model, at each fold in case of CV, we considered the following evaluation metrics: Based on the methods scores, we define true positive (TP), false negative (FN), false positive (FP) and true negative (TN). We calculate precision, recall and specificity values as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} \;=\; \frac{\text{TN}}{\text{TN} + \text{FP}}$$

We construct precision-recall curve based on different precision and recall values at different cut-offs. Also, we construct the area under the ROC curve (AUC) at various threshold settings, based on different recall values, and false positive rate (FPR) values, calculated as 1 – specificity. Then, we calculate the AUPR and AUC, where the values of AUPR and AUC, separately, over 5-repeats of 10-fold CVs are averaged and reported as the measures of the model performance for each dataset. As the positive and negative data in the datasets used in this study are highly imbalanced, AUPR metric provides a better quality estimate, since it punishes more heavily the existence of FPs than is the case with AUC. Thus, in this study, we mainly used AUPR values to evaluate the performance of the methods, though we also reported the AUC values in Supplementary Material.

As a summary, for the purpose of the fair comparison with the other methods, all methods are subjected to the exactly same conditions of testing and the same datasets [(i) the five trials of 10-fold CV and the same datasets, Yamanishi_08 and DrugBank_FDA dataset and (ii) the same hold-out test based on DrugBank_FDA]. We point out that all methods are evaluated using the same data splits to avoid any type of unwanted bias. Also, we used only training data to develop models.

4 Results

4.1 Comparisons with the state-of-the-art algorithms

First, we compare our proposed DDR method with the following state-of-the-art methods (for more detailed information see Supplementary Material, Related work) namely: COSINE (Lim *et al.*, 2016), DNILMF (Hao *et al.*, 2017), NRLMF (Liu *et al.*, 2016), KRONRLS-MKL (Nascimento *et al.*, 2016) and BLM-NII (Mei *et al.*, 2013) under the same conditions for all methods, i.e. under the three prediction settings (S_P, S_D and S_T) and over five trials of 10-fold CV based on Yamanishi_08 and DrugBank_FDA datasets.

We show that DDR, using 5-repeats of 10-fold CV, achieves higher AUPR values compared with the other methods (Fig. 2). In terms of AUPR, over the five different datasets, DDR, DNILMF, NRLMF, KRONRLS-MKL, COSINE and BLM-NII achieved weighted average of AUPR score under the three different prediction tasks settings as (S_P: 71%, S_D: 53%, S_T: 52%), (S_P: 56%, S_D: 26%, S_T: 37%), (S_P: 50%, S_D: 29%, S_T: 39%), (S_P: 52%, S_D: 20%, S_T: 17%), (S_D: 14%) and (S_P: 35%, S_D: 14%, S_T: 25%), respectively. The weighted average of AUPR is calculated for each of the three settings as

$$\frac{\sum_{i=1}^{5} \text{AUPR}_i \cdot \text{NS}_i}{\text{TS}}$$

where 5 is the number of datasets used in this study, TS is the total number of samples in all datasets and NS_i is the number of samples in i-th dataset.

It should be noted that the COSINE method is specifically tailored for the S_D setting to find target proteins of new drugs with little to no available interaction data; thus, only its results for the S_D setting are shown. Also, we show that DDR, using 5-repeats of 10-fold CV, achieves higher AUC values compared to the other methods under three prediction tasks and over the five different datasets (Supplementary Table S2). Thus, in terms of AUC, over the five different datasets, DDR, DNILMF, NRLMF, KRONRLS-MKL, COSINE and BLM-NII achieved weighted average of AUC score under the three different prediction tasks settings as (S_P: 96%, S_D: 90%, S_T: 89%), (S_P: 95%, S_D: 87%, S_T: 85%), (S_P: 94%, S_D: 85%, S_T: 84%), (S_P: 89%, S_D: 77%, S_T: 83%), (S_D: 79%) and (S_P: 91%, S_D: 73%, S_T: 80%), respectively.

To show more clearly the accuracy improvement by DDR, we define the AUPR score error E as

$$E = 1 - \text{AUPR}$$

while the relative reduction of the AUPR score error of method 1 relative to method 2 we defined as

$$\Delta E = (E_2 - E_1)/E_2,$$

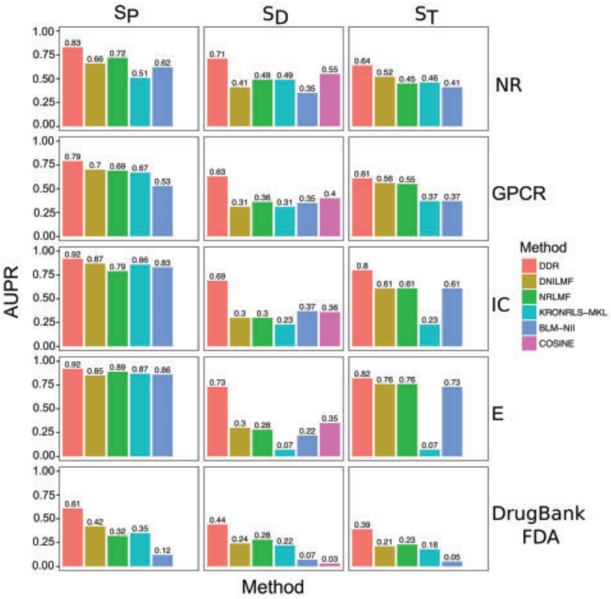

Fig. 2. Comparison results (in terms of AUPR scores) of DDR with the five state of the art methods (DNILMF, NRLMF, KRONRLS-MKL, COSINE and BLM-NII) using 5-repeats of 10-fold CV. Results are obtained under three prediction tasks (S_P, S_D and S_T) over all datasets (NR, GPCR, IC, E and DrugBank_FDA) used in this study. The results for DNILMF, NRLMF, KRONRLS-MKL, COSINE and BLM-NII are obtained using the best parameters reported in the respective publications

where E_1 and E_2 are determined for method 1 and method 2, respectively. Based on individual AUPR values reported from 5-repeats of 10-fold CV experiments, we also calculated the relative reduction of the AUPR error obtained by DDR relative to the next best method across all different testing settings for each dataset. When predicting unknown DTI pairs, as in the S_P setting, DDR significantly reduces AUPR error relative to the next best method by 39%, 30%, 38%, 27% and 33% for NR, GPCR, IC, E and DrugBank_FDA datasets, respectively. For predicting DTIs for new drugs (S_D setting), DDR significantly reduces AUPR error relative to the next best method by 36%, 38%, 51%, 58% and 22%, for NR, GPCR, IC, E and DrugBank_FDA datasets, respectively. Finally, for predicting new target proteins (S_T setting), DDR significantly reduces AUPR error relative to the next best method by 25%, 11%, 49%, 25% and 21%, for NR, GPCR, IC, E and DrugBank_FDA dataset, respectively. As a result, we demonstrate that DDR, reported from 5-repeats of 10-fold CV experiments, achieves significantly more accurate results than the other methods by achieving higher rank (Supplementary Table S3) on different datasets and in all three settings. We also demonstrated that DDR performs significantly better than the other existing methods when known DTIs are missing in the training data. This shows practical assessments of the predictive power of DDR for real scenarios of DTI prediction, as in finding target proteins for new drugs (S_D setting) with no available information about interactions and predicting drugs for new target proteins (S_T setting) (Supplementary Table S3).

Moreover, we demonstrated that on weighted average over five datasets, reported from 5-repeats of 10-fold CV experiments, DDR reduces the AUPR score error relative to the next best method by 34% for predicting DTIs as in setting (S_P), by 34% for predicting DTIs as in setting (S_D) and by 23% for predicting DTIs as in setting (S_T). This demonstrates that DDR significantly reduces the AUPR score error compared to the other start-of-the-art methods.

In general, based on our prediction results (Fig. 2), we observe that the results with the prediction model built for each specific class of target proteins (i.e. NR, GPCR, IC, E) are better than the results obtained by building a general model for multiple different target protein classes as in the case of DrugBank_FDA data. This is because each class of target proteins (NR, IC, GPCR, E) has its common characteristics that make them different from other classes. Thus, it is reasonable to expect that the well-designed and trained DTI predictor will capture some of these characteristics. In this way, the DTI prediction models will also be more specific and tuned to the target protein class for which they were developed and less tuned for the other target protein classes. Our results obtained for predicting DTIs using DrugBank_FDA data confirmed that even in this case DDR significantly outperformed all other state-of-the-art methods used in the comparison.

We also performed test on hold-out data using DrugBank_FDA dataset. These tests show that DDR achieves higher AUPR and AUC values compared with the other methods under the three prediction settings (Supplementary Table S4). We demonstrate that based on AUPR values, reported from hold-out tests, the reduction of the AUPR error for DDR relative to the next best method across all different testing settings for DrugBank_FDA dataset are 44% in the S_P setting, 51% in the S_D setting and 29% in the S_T setting.

4.2 Effect of similarity measures on the DDR performance

Similarity between drugs or target proteins plays the most crucial role when trying to predict DTIs for new drugs or new target proteins. Different similarity measures describe data instances differently. Several studies have highlighted the importance of selecting the proper similarity and integrating several similarity types to capture complementary information from several sources (Hao *et al.*, 2017; Nascimento *et al.*, 2016). The proof is the improved accuracy of DTI predictions over single adopted similarity (one for target proteins, one for drugs), and this is why the combining multiple types of similarities is important. We demonstrated that a suitable combination of few similarity measures results in higher accuracy of DTI predictions than when many or all similarity measures are used. Thus, the improvements DDR provide compared to the current combination strategies are that: (i) it applies non-linear similarity fusion method to combine different similarities, (ii) it can handle any number of provided similarities and (iii) it provides a systematic framework to select the most relevant non-redundant similarities. In addition, combining multiple similarity measures into one combined similarity reduces the time complexity and data dimensionality needed by the DDR method compared to the case of building a classification model with multiple features, where each feature is based on scoring a path from a drug to a target protein through each single similarity measure between drugs and each single similarity measure between target proteins.

Thus, our aim is to combine multiple similarity measures into one final composite similarity that captures the necessary information from different similarities between drugs as well as from different similarities between target proteins. Regarding this, we show that DDR achieves higher AUPR values compared to the other methods (Fig. 2). We also compared the DDR performance when combining all similarity measures we used in this study, with the case of combining only the similarity measures we selected in a heuristic process. We observed that the performance of DDR when combining only selected similarities is better than when combining all similarities (Supplementary Table S5).

When we examined the selected similarities over the four datasets in Yamanishi_08, we observed that DDR consistently selects a similar set of similarity measures for drugs and for target proteins (Supplementary Table S6). For the selected similarity of drugs, we observe that the selected similarities are related to network interaction profiles and drug side-effects. It has been highlighted before that the side-effect-based similarity improves the prediction of DTIs, where the assumption is that drugs with similar target protein binding profiles tend to cause similar side-effects, implying a direct correlation between target protein binding and side-effect similarity (Campillos *et al.*, 2008; Vilar and Hripcsak, 2016). It has also been shown that the interaction profiling is an effective tool that can be used for accurate prediction of DTIs (van Laarhoven *et al.*, 2011); the assumption is that two drugs that interact in a similar way with the target proteins in a known DTI network, will also interact in a similar way with new target proteins. For selected similarities of target proteins, we observe that these similarities are constructed based on a specific characteristic of amino acids sequence and closeness in PPI network that have been highlighted before in different benchmarking studies of target protein descriptors to result in a good performance for DTI prediction (Cao, 2015; Deng *et al.*, 2002; Nascimento *et al.*, 2016).

For the DrugBank_FDA dataset (Supplementary Table S6), DDR selected a set of similarity measures for drugs and for target proteins, separately. We note that the information included in different data sources used to calculate the similarity measures between drugs and between target proteins have highly influenced the prediction performance for drugs interacting with multi-class target proteins (i.e. NR, GPCR, etc.). For similarity measures of drugs and target

Table 2. Top ranked 25 novel DTIs predicted by DDR

Drug ID	Drug name	Taregt protein ID	Target protein name	Validation source	Evidence
Dataset: NR					
D00348	Isotretinoin	hsa6256	RXRA	CTD	CTD: D015474, CTD: 6256
D00585	Mifepristone	hsa2099	ESR1	C and PMID	C: 1166117, C: 206, C: 1276308, PMID: 20046055
D00962	Clomiphene citrate	hsa5241	PGR	CTD	CTD: D002996, CTD: 5241
D00182	Norethindrone	hsa2099	ESR1	T3DB and PMID	T3DB: T3D4745, PMID: 23611293
D00951	Medroxyprogesterone acetate	hsa2099	ESR1	DB	DB: DB00603
Dataset: GPCR					
D00049	Niacin	hsa8843	HCAR3	DB	DB: DB00627
D02910	Amiodarone	hsa154	ADRB2	CTD	CTD: D000638, CTD: 154
D02340	Loxapine	hsa1812	DRD1	DB	DB: DB00408
D00726	Metoclopramide	hsa1129	CHRM2	M	M: PC4168
D00674	Naratriptan hydrochloride	hsa3351	HTR1B	DB	DB: DB00952
Dataset: IC					
D02356	Verapamil	hsa6833	ABCC8	PMID	PMID: 21098040
D03365	Nicotine	hsa1137	CHRNA4	DB	DB: DB00184
D00538	Zonisamide	hsa6331	SCN5A	DB	DB: DB00909
D02098	Proparacaine hydrochloride	hsa8645	KCNK5	None	None
D00775	Riluzole	hsa2898	GRIK2	None	None
Dataset: E					
D00139	Methoxsalen	hsa1543	CYP1A1	DB and PMID	DB: DB00553 PMID: 15670584
D00437	Nifedipine	hsa1559	CYP2C9	DB	DB: DB01115
D00410	Metyrapone	hsa1583	CYP11A1	CTD	CTD: D008797, CTD: 1583
D00574	Aminoglutethimide	hsa1589	CYP21A2	M	M: PC2145
D00542	Halothane	hsa1571	CYP2E1	M	M: PC3562
Dataset: DrugBank_FDA					
DB01589	Quazepam	P47870	GABRB2	K	K: D00457
DB00825	Menthol	P35372	OPRM1	None	None
DB00147	Pyridoxal	P04798	CYP1A1	PMID	PMID: 19637937
DB01544	Flunitrazepam	P14867	GABRA1	CTD and K	CTD: D005445, K: D01230
DB02546	Vorinostat	P56524	HDAC4	CTD and C	CTD: C111237, CTD: 9759 C: 98, C: 3524

Note: Most of the top novel interactions (highest prediction score) are confirmed as supported by other existing evidences (public databases or literature) where the following annotation is used to demarcate the source of confirmatory information.

C, ChEMBL; CTD, Comparative Toxicogenomics Database; DB, DrugBank; M, MATADOR; K, KEGG; PMID, PubMed; PC, PubChem Compound.

proteins that have been selected in the sequential heuristic process, we observe that these similarities are related to network interaction profiles (van Laarhoven *et al.*, 2011) and other genome-wide global characteristics of drugs and of target proteins such as drug-diseases profiles and drug-pathways profiles between drugs, drug-induced gene expression profiles of drugs, profiles of drug ATC-codes associations, profiles of GO terms of target proteins and profiles of pathways of target proteins. Using such types of similarities in DTI prediction in numerous studies proved to be effective in describing each drug and target protein in different datasets (Chen *et al.*, 2012a,b; Dudley *et al.*, 2011; Dunkel *et al.*, 2008; Ehsani and Drablos, 2016; Iwata *et al.*, 2017; Pan *et al.*, 2014; Rodriguez-Esteban, 2016; Smith *et al.*, 2012; van Laarhoven *et al.*, 2011; Vilar and Hripcsak, 2016).

4.3 Prediction and validation of novel (unknown) DTIs

To evaluate the utility of DDR, we used it to predict novel DTIs (i.e. those that are not known to be true DTIs) in each of the five datasets, separately. For prediction of novel interactions, we build the predictive model for each dataset used in this study, in which the model is trained using all known interactions (positive labels) in all data folds of CV, and the negative labels are split into train and test sets as in a CV setup. As a result, all unknown DTI (negative labels) are predicted and the top 5 ranked interactions for each dataset are validated. To verify these novel predictions, we considered several reference databases that contain information obtained from curated/experimental/published results on small molecule–protein interactions. Thus, we searched DrugBank (Wishart *et al.*, 2008), KEGG (Kanehisa *et al.*, 2006), ChEMBL (Gaulton *et al.*, 2012), Matador (Gunther *et al.*, 2008), CTD (Davis *et al.*, 2017), T3DB (Wishart *et al.*, 2015) and the biomedical literature to find supporting evidences.

In summary, we evaluated the accuracy of 25 novel DTIs predicted by our method using four datasets of Yamanashi_08 and DrugBank_FDA dataset and confirmed 22 of these novel DTIs as supported by other existing evidence (Table 2).

Furthermore, to demonstrate that the predictions by DDR are not random, we additionally performed the label permutation tests to ensure that the top 5 DTI predictions by DDR in each dataset are not predicted by chance. To do so, we performed the following: we randomly shuffled the network labels (known and unknown) 100 times to produce different 100 random networks. Then, we performed S_P DTI prediction setup on each network. For each dataset and for each novel DTI in the top 5 DTIs based on that dataset, we calculated *P*-value as the percentage of a given novel DTI being ranked in the top 5 DTIs in the 100 random networks. We demonstrated that all predicted novel DTIs have significant *P*-values <0.05 (Supplementary Table S7). Thus, in addition to having DDR novel DTI prediction validated based on other sources, results from the label permutation tests also confirm the reliability of DDR novel DTI predictions.

5 Discussion

This study introduces a novel DTI prediction method, DDR, which utilizes a heterogeneous drug–target graph that contains information about various similarities between drugs and similarities between proteins as drug targets. On different representative datasets, under various test setups, and using AUPR and AUC as the performance measures, we show that DDR clearly outperforms the other state-of-the-art methods we used in the comparison. For these we used CV and hold-out tests. DDR achieves notably higher AUPR values compared to other methods, thus significantly reducing the AUPR score error relative to the next best method.

Moreover, on different datasets and in all three task settings we demonstrate that DDR produces significantly more accurate results than the other methods by achieving higher rank, based on AUPR values. We also demonstrated that DDR performs significantly better than the other existing methods when known DTIs are missing in the training data. This shows practical assessments of the predictive power of DDR for real scenarios of DTI prediction, as in finding target proteins for new drugs (S_D setting) with no available information about interactions and predicting drugs for target proteins that are new (S_T setting).

When we compared DDR performance in case of combining all similarity measures we used in this study with the case of combining only the similarity measures we selected through our heuristic method, we observed that the performance of DDR with selected similarities is better than when combining all similarities.

We observed that the best second method in predicting DTI as in S_P setting, based on the weighted average of AUPR results over the five different datasets is the DNILMF method. This is due to the method followed by DNILMF in employing the nonlinear combination technique of multiple similarity measures for drugs and for target proteins, as well as smoothing the predictions of new drugs and new target proteins by incorporating neighbor information based on the assumption that similar drugs (or target proteins) may contribute to the accuracy of the predictions for their neighbors. On the other hand, in predicting DTIs in both settings of S_D and S_T, we observed that the second best method, based on the weighted average of AUPR results over the five different datasets, is the NRLMF method. This is due to the methodology followed by the NRLMF method in incorporating neighborhood information from most similar drugs and target proteins.

As the current implementation of DDR handles only binary DTI data with the goal of classifying a given DTI as binding (label = 1) or non-binding (label = 0), in future, we plan to extend the functionality of DDR to handle continuous DTI data (i.e. continuous values of binding affinities of drugs and target proteins, He *et al.*, 2017).

6 Conclusion

We presented our method (DDR) that is based on the use of a heterogeneous graph containing information about known DTIs, as well as similarities between drugs and similarities between target proteins obtained from different data sources. DDR utilizes graph mining and machine learning techniques. It is capable of utilizing different similarity measures between drugs, as well as between target proteins. DDR applied non-linear similarity fusion method to combine different similarities for drugs and target proteins. Before applying the combined similarity method, DDR performed a pre-processing step where a subset of similarity types is selected in a heuristic process. This is done to select the best combination of similarities for the tasks in question since using all similarity types introduces noise.

We demonstrated that DDR achieves significantly more accurate results than the other state-of-the-art methods under different prediction tasks settings and using different datasets and different methods of performance evaluation. Finally, we evaluated the accuracy of 25 novel DTIs predicted by our method and confirmed 22 of these novel DTIs as supported by other existing evidences. Thus, DDR proved its practical utility by validating predictions of novel DTIs over different datasets, suggesting that DDR can be used as an efficient method to identify correct DTIs.

Funding

Research reported in this publication was supported by the King Abdullah University of Science and Technology (KAUST) Base Research Funds to VBB (BAS/1/1606-01-01) and by KAUST Office of Sponsored Research Grant No. URF/1/1976-02.

Conflict of interest: none declared.

References

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

Ba-Alawi,W. *et al.* (2016) DASPfind: new efficient method to predict drug-target interactions. *J. Cheminform.*, **8**, 15.

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

Boutet,E. *et al.* (2016) UniProtKB/Swiss-Prot, the manually annotated section of the UniProt knowledgebase: how to use the entry view. *Methods Mol. Biol.*, **1374**, 23–54.

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

Cao,D.S. *et al.* (2015) Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics*, **31**, 279–281.

Chen, B. *et al.* (2012a) Assessing drug target association using semantic linked data. *PLoS Comput. Biol.*, **8**, e1002574.

Chen, L. *et al.* (2012b) Predicting anatomical therapeutic chemical (ATC) classification of drugs by integrating chemical-chemical interactions and similarities. *PLoS One*, **7**, e35254.

Chen,X. *et al.* (2016) Drug-target interaction prediction: databases, web servers and computational models. *Brief. Bioinform.*, **17**, 696–712.

Cheng,A.C. *et al.* (2007) Structure-based maximal affinity model predicts small-molecule druggability. *Nat. Biotechnol.*, **25**, 71–75.

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

Cichonska,A. *et al.* (2015) Identification of drug candidates and repurposing opportunities through compound-target interaction networks. *Expert Opin. Drug Discov.*, **10**, 1333–1345.

Davis,A.P. *et al.* (2017) The comparative toxicogenomics database: update 2017. *Nucleic Acids Res.*, **45**, D972–D978.

Deng,M. *et al.* (2002) Prediction of protein function using protein–protein interaction data. *Proc. IEEE Comput. Soc. Bioinform. Conf.*, **1**, 197–206.

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

Dobson,C.M. (2004) Chemical space and biology. *Nature*, **432**, 824–828.

Dudley,J.T. *et al.* (2011) Exploiting drug-disease relationships for computational drug repositioning. *Brief. Bioinform.*, **12**, 303–311.

Dunkel,M. *et al.* (2008) SuperPred: drug classification and target prediction. *Nucleic Acids Res.*, **36**, W55–W59.

Ehsani,R., and Drabløs,F. (2016) TopoICSim: a new semantic similarity measure based on gene ontology. *BMC Bioinformatics*, **17**, 296.

Fu,G. *et al.* (2016) Predicting drug target interactions using meta-path-based semantic network analysis. *BMC Bioinformatics*, **17**, 160.

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

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

Hao,M. *et al.* (2017) Predicting drug-target interactions by dual-network integrated logistic matrix factorization. *Sci. Rep.*, **7**, 40376.

He,T. *et al.* (2017) SimBoost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *J. Cheminform.*, **9**, 24.

Ho,T.K. (1995) *Random decision forests*. In: *Document Analysis and Recognition, 1995, Proceedings of the Third International Conference on* IEEE, p. 278–282.

Iwata,M. *et al.* (2017) Elucidating the modes of action for bioactive compounds in a cell-specific manner by large-scale chemically-induced transcriptomics. *Sci. Rep.*, **7**, 40164.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res.*, **34**, D354–D357.

Keiser,M.J. *et al.* (2007) Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.*, **25**, 197–206.

Li,J. *et al.* (2016) A survey of current trends in computational drug repositioning. *Brief. Bioinform.*, **17**, 2–12.

Lim,H. *et al.* (2016) Improved genome-scale multi-target virtual screening via a novel collaborative filtering approach to cold-start problem. *Sci. Rep.*, **6**, 38860.

Liu,Y. *et al.* (2016) Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol.*, **12**, e1004760.

Lu,Y. *et al.* (2017) Link prediction in drug-target interactions network using similarity indices. *BMC Bioinformatics*, **18**, 39.

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

Menni,C. *et al.* (2017) Mixing omics: combining genetics and metabolomics to study rheumatic diseases. *Nat. Rev. Rheumatol.*, **13**, 174–181.

Mitchell,J.B. (2001) 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.*, **41**, 1617–1622.

Nascimento,A.C. *et al.* (2016) A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinformatics*, **17**, 46.

Overington,J.P. *et al.* (2006) How many drug targets are there? *Nat. Rev. Drug Discov.*, **5**, 993–996.

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

Pan,Y. *et al.* (2014) Pathway analysis for drug repositioning based on public database mining. *J. Chem. Inf.*, **54**, 407–418.

Paul,S.M. *et al.* (2010) How to improve R&D productivity: the pharmaceutical industry's grand challenge. *Nat. Rev. Drug Discov.*, **9**, 203–214.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

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

Rodriguez-Esteban,R. (2016) A drug-centric view of drug development: how drugs spread from disease to disease. *PLoS Comput. Biol.*, **12**, e1004852.

Santos,R. *et al.* (2017) A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.*, **16**, 19–34.

Schenone,M. *et al.* (2013) Target identification and mechanism of action in chemical biology and drug discovery. *Nat. Chem. Biol.*, **9**, 232–240.

Schomburg,I. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.

Seal, A. *et al.* (2015) Optimizing drug-target interaction prediction based on random walk on heterogeneous networks. *J. Cheminform.*, **7**, 40.

Shim,J.S., and Liu,J.O. (2014) Recent advances in drug repositioning for the discovery of new anticancer drugs. *Int. J. Biol. Sci.*, **10**, 654–663.

Smith,S.B. *et al.* (2012) Identification of common biological pathways and drug targets across multiple respiratory viruses based on human host gene expression analysis. *PLoS One*, **7**, e33174.

Soufan,O. *et al.* (2015) Mining chemical activity status from high-throughput screening assays. *PLoS One*, **10**, e0144426.

Soufan,O. *et al.* (2016) DRABAL: novel method to mine large high-throughput screening assays using Bayesian active learning. *J. Cheminform.*, **8**, 64.

Swinney,D.C., and Anthony,J. (2011) How were new medicines discovered? *Nat. Rev. Drug Discov.*, **10**, 507–519.

Tari,L.B., and Patel,J.H. (2014) Systematic drug repurposing through text mining. *Methods Mol. Biol.*, **1159**, 253–267.

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

Vilar,S., and Hripcsak,G. (2016) The role of drug profiles as similarity metrics: applications to repurposing, adverse effects detection and drug-drug interactions. *Brief. Bioinform.*, **18**, 670–681.

Wang,W. *et al.* (2013) Drug target predictions based on heterogeneous graph inference. *Pac. Symp. Biocomput.*, 53–64.

Wang,B. *et al.* (2014) Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods*, **11**, 333–337.

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

Wishart,D. *et al.* (2015) T3DB: the toxic exposome database. *Nucleic Acids Res.*, **43**, D928–D934.

Wu,Z. *et al.* (2017) SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug-target interactions and drug repositioning. *Brief. Bioinform.*, **18**, 333–347.

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.

Yuan,Q. *et al.* (2016) DrugE-Rank: improving drug-target interaction prediction of new candidate drugs or targets by ensemble learning to rank. *Bioinformatics*, **32**, i18–i27.

Zhu,S. *et al.* (2005) A probabilistic model for mining implicit 'chemical compound-gene' relations from literature. *Bioinformatics*, **21**, ii245–ii251.

Zhu,Q. *et al.* (2014) Exploring the pharmacogenomics knowledge base (PharmGKB) for repositioning breast cancer drugs by leveraging Web ontology language (OWL) and cheminformatics approaches. *Pac. Symp. Biocomput.*, 172–182.

Zong,N. *et al.* (2017) Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations. *Bioinformatics*, **33**, 2337–2344.