*Bioinformatics*, 32(17), 2016, 2664–2671
doi: 10.1093/bioinformatics/btw228
Advance Access Publication Date: 5 May 2016
Original Paper



Systems biology

# Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm

**Huimin Luo[1,2], Jianxin Wang[1,*], Min Li[1], Junwei Luo[1], Xiaoqing Peng[1], Fang-Xiang Wu[3] and Yi Pan[4]**

[1]School of Information Science and Engineering, Central South University, ChangSha, 410083, China, [2]School of Computer and Information Engineering, Henan University, KaiFeng 475001, China, [3]Division of Biomedical Engineering, University of Saskatchewan, Saskatchewan S7N 5A9, Canada and [4]Department of Computer Science, Georgia State University, Atlanta, GA 30302-3994, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on January 5, 2016; revised on March 27, 2016; accepted on April 18, 2016

## Abstract

**Motivation:** Drug repositioning, which aims to identify new indications for existing drugs, offers a promising alternative to reduce the total time and cost of traditional drug development. Many computational strategies for drug repositioning have been proposed, which are based on similarities among drugs and diseases. Current studies typically use either only drug-related properties (e.g. chemical structures) or only disease-related properties (e.g. phenotypes) to calculate drug or disease similarity, respectively, while not taking into account the influence of known drug–disease association information on the similarity measures.

**Results:** In this article, based on the assumption that similar drugs are normally associated with similar diseases and vice versa, we propose a novel computational method named MBiRW, which utilizes some comprehensive similarity measures and Bi-Random walk (BiRW) algorithm to identify potential novel indications for a given drug. By integrating drug or disease features information with known drug–disease associations, the comprehensive similarity measures are firstly developed to calculate similarity for drugs and diseases. Then drug similarity network and disease similarity network are constructed, and they are incorporated into a heterogeneous network with known drug–disease interactions. Based on the drug–disease heterogeneous network, BiRW algorithm is adopted to predict novel potential drug–disease associations. Computational experiment results from various datasets demonstrate that the proposed approach has reliable prediction performance and outperforms several recent computational drug repositioning approaches. Moreover, case studies of five selected drugs further confirm the superior performance of our method to discover potential indications for drugs practically.

**Availability and Implementation:** http://github.com//bioinfomaticsCSU/MBiRW.

**Contact:** jxwang@mail.csu.edu.cn

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

## 1 Introduction

The last decades have witnessed impressive advances in genomics, life sciences and technology. However, drug discovery process is still time consuming, risky and tremendously expensive (Li *et al.*, 2015). Even with the continuously increasing investments of research and development (R&D), the number of new drugs approved by the U.S. Food and Drug Administration (FDA) has been declining since the late 1990s (Grabowski *et al.*, 2004). In light of the challenge in traditional drug discovery, identifying new indications for existing drugs, also known as drug repositioning or drug repurposing, has attracted increasing interests from both the pharmaceutical industry and research community (Hurle *et al.*, 2013). Drug repositioning can speed up the process of drug development and reduce risk, as repositioning candidates have already passed all necessary tests common to de novo drug discovery and development. Therefore, drug repositioning has consequently become a major strategy and played a key role in drug discovery and development. In addition, several successful repositioned drugs (e.g. Sildenafil, thalidomide, raloxifene) have generated historically high revenues for their patent holders or companies (Ashburn *et al.*, 2004).

The goal of drug repositioning is to find potential new uses for existing drugs and apply the newly identified drugs to the treatment of diseases other than the drug's originally intended disease (Shim and Liu, 2014). Recently, some computational methods have been proposed to predict direct drug–disease associations for drug repositioning. For example, Chiang and Butte (2009) developed a network-based, guilt-by-association method, to predict potential drug–disease associations. This methodology recommended the novel use of a drug based on the assumption that if two diseases share some similar treatment profiles, then the drugs used for only one of the two diseases could also be used for the other. However, the method is biased toward both older drugs with diverse uses and diseases with diverse treatments. Gottlieb et al. (2011) conducted multiple drug–drug and disease–disease similarity measures to construct discriminate features, and implemented a classification algorithm named PREDICT to predict novel drug indications. Wu et al. (2013) constructed a weighted disease and drug heterogeneous network based on known disease–gene and drug–target relationships, and applied network clustering to identify drug candidates. Wang et al. (2014) proposed a novel heterogeneous network model which integrated drug repositioning and target prediction into one unified computational framework. Based on the framework, an iterative algorithm was developed to rank candidate drugs for each disease. Martínez *et al.* (2015) developed a network-based prioritization method named DrugNet, which simultaneously integrated information about diseases, drugs and targets to perform drug–disease and disease–drug prioritization. Chen *et al.* (2015) formulated drug–disease association prediction problem as recommending preferable diseases for drugs and adopted two existing recommendation methods, ProbS and HeatS, to infer drug–disease associations directly. Most of these approaches usually perform drug–disease prediction by exploiting drug similarity and disease similarity, while similarity measures are often based on some important drug- or disease-related properties. However, previous studies seldom utilized the known drug–disease interaction information of dataset for defining similarity measures, and yet can be exploited to improve similarity measures.

In this study, we present a novel prediction method MBiRW, which adopts effective mechanism to measure similarity for drugs and diseases, respectively, and applies Bi-Random walk (BiRW) algorithm to predict potential indications for existing drugs. In the new similarity measures, similarity calculated based on drug- or disease-related properties is improved in two aspects. First, as previous studies (Van Driel *et al.*, 2006; Vanunu et al., 2010; Wang *et al.*, 2014) have found that weak similarity provides little information for interaction inference, we adjust those weak similarities which are not informative for drug–disease prediction by correlation analysis. Second, two drugs are considered to be more similar if they have common indications, or there exist other drugs which have common indications with them simultaneously. Drugs and diseases are clustered based on their shared commons, and similarity for drugs or diseases which belong to the same cluster is adjusted. What's more, as it has been demonstrated better performance than other random-walk methods on disease gene prioritization, BiRW algorithm (Xie *et al.*, 2012) is adopted to predict potential associations between drugs and diseases in this study.

Furthermore, we compare and evaluate the performance of MBiRW on various datasets by adopting common metrics. Experimental results demonstrate that MBiRW has the superior capability to discover potential disease indications for drugs.

## 2 Materials and methods

In this section, the dataset used in this study is firstly presented. Then, based on the assumption that similar drugs often indicate similar diseases, a novel approach of integrating comprehensive similarity measures with Bi-directional Random Walk(BiRW), referred to MBiRW, is proposed to predict potential associations between drugs and diseases.

Generally speaking, the overall prediction procedure of MBiRW consists of three steps described as follows. Based on the collected dataset, comprehensive similarity measures are firstly implemented to measure similarity for drugs and diseases. Then, a heterogeneous network consisting of drug similarity network, disease similarity network and drug–disease interactions, is constructed. Last, based on the heterogeneous network, BiRW is implemented to rank candidate diseases for drugs.

### 2.1 Dataset

The gold standard dataset used in this study is obtained from the Supplementary Material of paper Gottlieb et al. (2011), which collected comprehensive associations between drugs and diseases from multiple data sources. For this dataset, there are 1933 known drug–disease associations involving 593 drugs registered in DrugBank (Wishart *et al.*, 2008) and 313 diseases listed in the Online Mendelian Inheritance in Man (OMIM) (Hamosh *et al.*, 2002).

### 2.2 Similarity measures

The similarity measures can be described briefly as follows: first, calculate drug similarity $Sim_{r^p}$ and disease similarity $Sim_{d^p}$ based on some drug-related properties and disease-related properties, respectively; second, analyze and evaluate the predictive power of various similarity values calculated above, and adjust these similarity values based on the analysis results to obtain new drug similarity $Sim_{r^l}$ and disease similarity $Sim_{d^l}$; last, cluster drugs and diseases based on the known drug–disease associations, respectively, then according to clustering results, $Sim_{r^l}$ and $Sim_{d^l}$ can be further improved to obtain comprehensive similarity measures $Sim_r$ for drugs and $Sim_d$ for diseases.

#### 2.2.1 Drug similarity measure

Based on the description above, the similarity between drugs can be measured in three steps.

*Step 1: Similarity measured based on drug-related properties.* In this study, drug similarity $Sim_{r^p}$ is calculated based on their chemical structures. SMILES (Simplified Molecular Input Line Entry Specification) is a line notation for describing chemical structures, the Canonical SMILES (Weininger, 1988) of all drugs are downloaded from DrugBank, and the Chemical Development Kit (Steinbeck *et al.*, 2006) is used to calculate the similarity of two drugs as the Tanimoto score (Tanimoto, 1957) of their 2D chemical fingerprints. According to $Sim_{r^p}$, known drug–disease association information is incorporated to improve drugs similarities in the following steps.

*Step 2: Analysis of similarity.* According to the findings that weak similarities provide little information for prediction in previous researches, we analyze the correlation between similarity $Sim_{r^p}$ of two drugs and the existence of their common diseases, and transform similarity $Sim_{r^p}$ which is not informative for prediction into some value close to zero.

The correlation analysis procedure can be described as follows: (i) the range of value of $Sim_{r^p}$, which is [0,1], is divided into ten equal subranges. For each subrange, the percentage of drug pairs sharing common diseases is calculated. (ii) All values of $Sim_{r^p}$ are randomly permutated for drug pairs using Fisher-Yates shuffling (Fisher and Yates, 1938) to form randomized drug similarities, and the percentage is calculated in the same way as step one. Overall, the average of 10 randomized results is used as a control for background signal. (iii) Based on the above results, a similarity threshold LSim is determined. That is to say, values of drug pairs whose similarity value is smaller than LSim has a small probability of sharing common diseases and thus is not informative for prediction. For $Sim_{r^p}$ of the gold standard dataset, the drug similarity correlation analysis result is shown in Figure 3(A). In our work, we adjust $Sim_{r^p}$ by applying the logistic function which has been used by Vanunu *et al.* (2010) to modify the phenotypic similarities between diseases associated with genes.

The logistic function is defined as follows,

$$L(x) = \frac{1}{1 + e^{(cx+d)}} \tag{1}$$

where $x$ denotes the value of $Sim_{r^p}$ value, $c$ and $d$ are the parameters which can be tuned to control the adjustment of $Sim_{r^p}$. Those small similarity values can be transformed into some values close to zero. At the same time, those large similarity values will be enlarged by logistic function. By applying the above procedure, drug similarity $Sim_{r^p}$ is transformed to new similarity $Sim_{r^l}$.

*Step 3: Clustering drugs based on known drug–disease associations.* As mentioned above, drug similarity $Sim_{r^l}$ has been obtained by adjusting $Sim_{r^p}$ according to the similarity analysis results. Next, based on the assumption that two drugs are more similar if they share indications, or there exist other drugs which share indications with them simultaneously. The known sharing information can be further utilized to improve drug similarity $Sim_{r^l}$ as follows.

First, we construct a new weighted drug sharing network named as SdrugNet based on the known drug–disease associations. In SdrugNet, let $SR = \{r_1, r_2, \ldots, r_m\}$ denote the set of $m$ drugs, and edge weight denotes the number of common diseases shared by corresponding drug pair. After that, we cluster SdrugNet to identify potential drug clusters, and improve $Sim_{r^l}$ between drugs that belong to the same cluster. As a graph clustering method, ClusterONE (Nepusz *et al.*, 2012) can identify cohesive modules in weighted networks and has been used to detect meaningful modules for drug repositioning (Wu *et al.*, 2013; Yu *et al.*, 2015). In view of its good performance in generating overlapping clusters for weighted networks, ClusterONE is adopted to identify clusters in our research.

The cohesiveness of a cluster V is defined by ClusterONE as follows:

$$f(V) = \frac{W_{\text{in}}(V)}{(W_{\text{in}}(V) + W_{\text{bound}}(V) + P(V))} \tag{2}$$

where $W_{\text{in}}(V)$ denotes the total weight of edges within a group of vertices $V$, $W_{\text{bound}}(V)$ represents the total weight of edges connecting this group to the rest of graph, and $P(V)$ is the penalty term. The quality of each identified cluster can be evaluated by cohesiveness of the cluster.

We suppose that drugs belonging to the same cluster tend to behave more similarly. Here, for drugs $r_i$ and $r_j$ locating in the same cluster $C$, $QC$ represents the cohesiveness of cluster $C$, the comprehensive drug similarity $Sim_r$ between $r_i$ and $r_j$ is defined as $w * Sim_{r^l}$, where parameter $w$ is set as $(1 + QC)$. Moreover, for comprehensive similarity between two different drugs which is equal or greater than 1, we replace it with 0.99.

#### 2.2.2 Disease similarity measure

Disease similarity $Sim_{d^p}$ based on disease phenotypes is calculated using MimMiner (Van Driel *et al.*, 2006), which measure disease similarity by computing similarity between MeSH terms (Lipscomb, 2000) appearing in the medical description of diseases from the OMIM database.

Next, diseases similarity $Sim_{d^p}$ is improved based on the adjusted approaches used in drug similarity measure. The correlation between the similarity of two diseases and the existence of their common drugs is firstly analyzed by taking known association information into account, and new disease similarity $Sim_{d^l}$ is obtained by adjusting $Sim_{d^p}$ based on the correlation analysis. Then, disease sharing network named as SdiseaseNet is constructed based on known drug–disease associations. In SdiseaseNet, let $SD = \{d_1, d_2, \ldots, d_n\}$ represents the set with $n$ diseases, and edge weight denotes the number of common drugs shared by corresponding disease pair. Diseases are clustered by applying ClusterONE on SdiseaseNet, and $Sim_{d^l}$ between diseases which locate in the same cluster is enhanced to obtain comprehensive disease similarity $Sim_d$ in the same way as for drugs.

### 2.3 Construction of the heterogeneous network

Based on the aforementioned similarity measures for drugs and diseases, drug similarity network and disease similarity network are constructed. In drug similarity network, let $R = \{r_1, r_2, \ldots, r_m\}$ denote the set of $m$ drugs. The edge between drugs $r_i$ and $r_j$ is weighted by comprehensive similarity $Sim_r$ between the two drugs. In disease similarity network, let $D = \{d_1, d_2, \ldots, d_n\}$ denote the set of $n$ diseases. The edge between diseases $d_i$ and $d_j$ is weighted by comprehensive similarity $Sim_d$ between the two diseases.

Besides, the drug–disease associations can be modeled as a bipartite graph $G(V, E)$, where $V(G) = \{R, D\}$, $E(G) \subseteq R \times D$, $E(G) = \{e_{ij}$, edge between drug $r_i$ and disease $d_j\}$. If there exists a known association between drug $r_i$ and disease $d_j$, the weight of edge between $r_i$ and $d_j$ is initially set to 1, otherwise, it is initially set to 0.

Obviously, the drug–disease network can be considered as a heterogeneous network, which is constructed by connecting the drug similarity network and disease similarity network via the bipartite

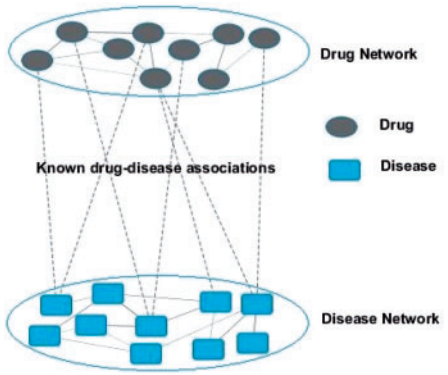


**Fig. 1** A two-layer drug–disease heterogeneous network consisting of drug–drug similarity network, disease–disease similarity network and drug–disease interactions while two networks are connected by the drug–disease associations. In drug–disease heterogeneous network, the weights of solid lines denote the intra-similarities, the dashed lines denote the known drug–disease associations

graph of drug–disease interactions. A prototype example of the heterogeneous network is illustrated in Figure 1.

### 2.4 Implementation BiRW algorithm for ranking candidate diseases for each drug

The drug–disease prediction process is modeled as a random walk on the drug similarity network and disease similarity network simultaneously. Considering the diverse topology and structure characteristics of different networks, the optimal number of random walk steps on the two networks may be different. Therefore, two parameters $l$, $r$ are introduced as the numbers of maximal iterations in the left and right random walks on these two networks. The iterative random walk process is described as follows.

Left walk in the drug network:

$$left\_RD_t = \alpha \times MR \times RD_{t-1} + (1-\alpha) \times A \quad (3)$$

Right walk in the disease network:

$$right\_RD_t = \alpha \times RD_{t-1} \times MD + (1-\alpha) \times A \quad (4)$$

where $MR^{m\times m}$, $MD^{n\times n}$ and $A^{m\times n}$ denote adjacency matrices of drug similarity network, disease similarity network and drug–disease association network, respectively, $m$ and $n$ represent the number of drugs and diseases, respectively. $left\_RD_t$ and $right\_RD_t$ represent the predicted drug–disease associations based on walking on drug similarity network and disease similarity network, respectively. The value of element $left\_RD_t(i,j)$ and $right\_RD_t(i,j)$ denotes the probability that drug $r_i$ associates with disease $d_j$. In the iteration process, $RD_t$ is the averaged output from left walks and right walks in each step. The larger the value of $RD_t(i, j)$, the greater the probability that drug $r_i$ associates with disease $d_j$ is. The complete MBiRW algorithm for inferring potential drug–disease associations is outlined as in Figure 2.

**Algorithm MBiRW**
**Input:** Drug set $R$ and disease set $D$, drug-disease association adjacency matrix $A$, parameter $\alpha$, iteration steps $l$ and $r$
**Output:** predicted association matrix $RD$

MBiRW ($R$, $D$, $A$, $\alpha$, $l$, $r$)
1. Calculate similarity of drugs and diseases based on new similarity measures;
2. Construct drug-drug similarity matrix $SimR$ and disease-disease similarity matrix $SimD$;
3. Normalize $SimR$ and $SimD$ to $MR$ and $MD$, respectively;
4. $RD_0 = A = A$ /sum ($A$)
5. for $t = 1$ to max ($l$, $r$)
6. rflag = dflag = 1;
7. if $t <= l$
8. $Rr = \alpha * MR * RD_{t-1} + (1-\alpha) * A$
9. rflag = 1
10. end if
11. if t <= r
12. $Rd = \alpha * RD_{t-1} * MD + (1-\alpha) * A$
13. dflag = 1
14. end if
15. $RD_t$ = (rflag* $Rr$ + dflag* $Rd$) / (rflag + dflag)
16. end for
17. return ($RD$)

**Fig. 2** Description of algorithm MBiRW. It takes $R$, $D$, $A$, the decay factor $\alpha$, left walk steps $l$ and right walk steps $r$ as the inputs, then performs drug–disease association prediction by iteratively updating the values of matrix $RD$ according to matrix $MR$, $MD$, $A$. In algorithm MBiRW, matrix $SimR$ and $SimD$ are normalized by Laplacian normalization. For $SimR$, a diagonal matrix $D_{SimR}$ is defined such that $D_{SimR}(i, i)$ is the sum of row $i$ of $SimR$. We set normalized matrix $MR$ as $SimR^{-1/2} * D_{SimR} * SimR^{-1/2}$, the elements of $MR$ is defined as $MR(i,j) = SimR(i,j) / \sqrt{D_{SimR}(i,i) * D_{SimR}(j,j)}$. The same normalization procedure is applied to matrix $SimD$, and then constructs the normalization matrix $MD$

## 3 Experiments and results

In this section, we firstly introduce several common metrics used to measure the performance of various prediction methods. By applying these evaluation metrics, we performed a comprehensive comparison of MBiRW with the other three methods on the gold standard dataset. Then, case studies are conducted to confirm the ability of MBiRW to find potential indications for drugs. After that, to test the generalization ability of MBiRW, systematic evaluation experiments are further conducted with the other two different datasets.

### 3.1 Evaluation metrics

To systematically evaluate the performance of different methods, we conduct tenfold cross validation, de novo prediction and independent dataset test. In the tenfold cross validation, all known drug–disease associations in gold standard dataset are randomly divided into ten subsets with equal size. In each cross validation trial, one subset is taken in turn as the test set, while the remaining nine subsets constitute the training set. After performing prediction, each association is given a predicted score. According to the final predicted scores, each known association between drug $i$ and disease $j$ in the test set is ranked relative to the candidate associations (all associations which have not been verified to associate with drug $i$ experimentally until now). For a specified rank threshold, TPR (true positive rate) is the fraction of known associations that are correctly predicted, FPR (false positive rate) is the fraction of unknown associations that are predicted, Precision is the fraction of known associations that are ranked within the rank threshold, and Recall is equivalent to TPR. By varying the rank threshold, we can calculate various TPR (true positive rate), FPR (false positive rate), Precision and Recall values. Then receiver operating curve (ROC) and precision-recall curve (PRC) can be drawn based on these measure values to show the performance of the different prediction approaches.

Considering the fact that the predicted top-ranked results are more important in practice, we also measure the performance of all prediction methods in terms of the top-ranked results, i.e. the numbers of correctly retrieved true drug–disease associations based on various top portions. Usually, it is regarded as more effective if the method can rank more true associations in top portions.

Currently, there are many experimental and withdrawn drugs which have no explicit indications in drug-related database, while

these drugs have the potential to indicate some diseases. In our study, de novo prediction test is performed to evaluate the effectiveness of the prediction methods when predicting potential diseases for new drugs (which have no any known associated indications).

## 3.2 Comparison with other methods

MBiRW is compared with the other three network-based prediction methods: NBI (Cheng *et al.*, 2012), HGBI (Wang *et al.*, 2013) and DrugNet (Martínez *et al.*, 2015). NBI conducts prediction based on a two-step diffusion model on a bipartite graph. HGBI is introduced based on the guilt-by-association principle and an intuitive interpretation of information flow on the heterogeneous graph. Although NBI and HGBI were originally developed for drug–target association prediction, while the authors have mentioned that these prediction methods can also be applied in prediction of drug–disease network. In addition, for drug repositioning applications, they have been used to perform drug–disease association prediction in Wang *et al.* (2014). DrugNet is a network-based drug repositioning method, which is able to perform both drug–disease and disease–drug prioritization. In this paper, these methods are evaluated and compared using the same datasets, by cross-validation, independent test set and de novo drug–disease prediction analysis.

### 3.2.1 Drug and disease similarity analysis

Based on the procedure of analysis on similarity described in previous section, we firstly perform the correlation analysis on drug similarity $Sim_{rp}$ by considering known drug–disease associations information of the gold standard dataset, the result is depicted in Figure 3(A). Then, we conduct the similar analysis on disease similarity $Sim_{dp}$ in the same way, and the result is depicted in Figure 3(B).

As we can see from Figure 3, drug pairs with similarity values smaller than 0.4 have an insignificant probability to indicate common diseases, and drug pairs which have similarity values larger than 0.7 have a a significant probability to indicate common diseases. Then for drug similarity values $x \in [0, 0.4]$, we set $L(x) \approx 0$, $L(0) = 0.0001$ and $L(0.4) < 0.01$, which determines d as log(9999), and $c$ as -11 in logistic function (1) for drug similarity. Regarding Figure 3(B), disease pairs which have similarity values lower than 0.3 have an insignificant probability to share drugs, and those which have similarity values higher than 0.6 have a significant probability to share drugs. Then for disease similarity values $x \in [0, 0.3]$, we set $L(x) \approx 0$, $L(0) = 0.0001$ and $L(0.3) < 0.01$, which determines d as log(9999), and $c$ as $-15$ in logistic function (1) for disease similarity. Finally, we performed similarity transformation by using logistic functions determined above to obtain adjusted drug similarity and disease similarity, respectively.

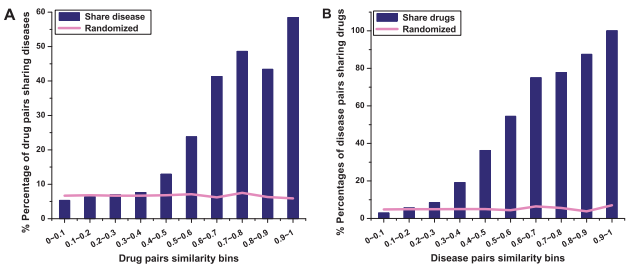


**Fig. 3** (**A**) Drug similarity versus disease sharing relations. (**B**) Disease similarity versus drug sharing relations. The average of 10 randomized similarity datasets was used as a control for background signal (Color version of this figure is available at *Bioinformatics* online.)

### 3.2.2 Tenfold cross validation

The impact of parameters used in MBiRW on the prediction performance is first investigated by conducting cross validation, and the detailed performance measurement results with various parameter settings are reported in Supplementary Table S1. It shows that parameter $\alpha$ ranging from 0.1 to 0.7 has little effect on the prediction performance, and MBiRW achieves better performance when setting parameter $l$ equal to $r$. In this study, we set the parameter $\alpha$ to 0.3, and $l$, $r$ to 2 for MBiRW. The parameter ($\alpha = 0.4$) of HGBI is set to be that used in (Wang *et al.*, 2013). The evaluation results of all methods in terms of ROC curves and top-ranked results of all the drug–disease associations, are reported in Figure 4. It should be pointed out that, in each cross validation trial, we analyze drug similarity and disease similarity, and cluster drugs and diseases again, without using the information about the test drug–disease associations. In terms of AUC (the area under of ROC curve), MBiRW (AUC: 0.917) performs better than other methods.

Moreover, the number of correctly retrieved drug–disease associations is shown in Figure 4(B). For a specified top-ranked threshold, a true drug–disease association is considered as correctly retrieved if the predicted ranking of this association is higher than the specified top-rank threshold. Obviously, when focusing on the top-one results, MBiRW significantly outperforms the other three methods. For example, among the 1933 true drug–disease associations, 586 of them are predicted at the top one based on MBiRW. The top-ranked predictions are particularly important in practice, so MBiRW can be more useful than other methods. We also report the experiment results in terms of the precision-recall curves depicted in Supplementary Figure S1, and MBiRW also shows better performance than other methods in precision and recall analysis.

In general, the prediction results of MBiRW illustrates that the adjusted drug and disease similarities by integrating some prior interaction information can significantly enhance performance of prediction.

### 3.2.3 De novo drug–disease prediction

To evaluate the capability of MBiRW in predicting potential indications for new drugs, we conduct the de novo drug–disease prediction test. In de novo prediction test, for each queried drug $i$, all known drug–disease associations with drug $i$ are removed. The performance of a prediction method is evaluated by the rank of the removed drug–disease associations relative to candidate associations with drug $i$. In the gold standard dataset, each drug has at least one known associated disease, then we perform de novo prediction test for all drugs. The other three methods compared with MBiRW can also be applied to predict potential diseases for new drugs. The experiment results in terms of ROC curves and top-ranked results are reported in Figure 5.

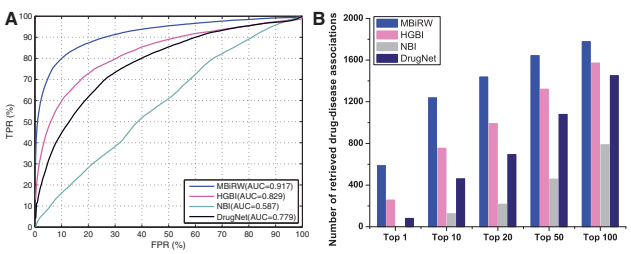


**Fig. 4** (**A**) ROC curves for predicting drug–disease associations using various methods. (**B**) Number of correctly retrieved known drug–disease associations for various rank thresholds (Color version of this figure is available at *Bioinformatics* online.)

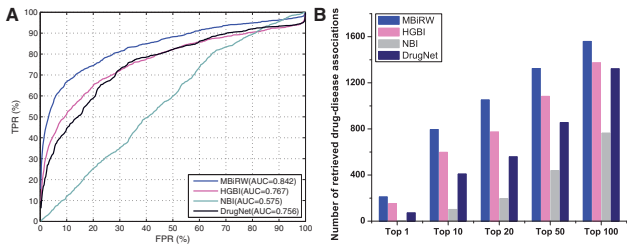


**Fig. 5** De novo test: (**A**) ROC curves of various methods in predicting drug–disease associations for new drugs. (**B**) Number of correctly retrieved known drug–disease associations for various rank thresholds (Color version of this figure is available at *Bioinformatics* online.)

As we can see from Figure 5, MBiRW achieves an AUC value of 0.842, which is better than that of other methods in the same experimental scenario too. Meanwhile, when focusing on the top-ranked prediction results, among the 1933 known drug–disease associations, MBiRW predict 209 of them at the top one. For PR curves analysis results in Supplementary Figure S2, it shows that MBiRW also exhibits the optimal predictions. Overall, all de novo prediction results show that our method can achieve the superior performance.

#### 3.2.4 Independent test set

We also investigate the performance of the various methods for drug–disease association prediction on the independent test set. In order to create the independent test set, the published drug–disease associations are derived from KEGG (Kanehisa *et al.*, 2014) database and literature (Martínez *et al.*, 2015) firstly. Then an independent test set with 144 new drug–disease pairs for 115 drugs can be obtained after removing those associations existing in the gold standard dataset. The test of independent dataset is conducted to measure the performance of all prediction methods, by predicting new drug–disease associations in the independent dataset. The test results are reported in Supplementary Figure S3 in terms of ROC curves, top-ranked analysis and PR curves. From these results, we can clearly see that MBiRW achieves the best AUC value of 0.831, and 15 new associations are ranked at top one. In addition, MBiRW also outperforms other methods in terms of PR curves, and achieves the highest precision. All these results shows that MBiRW achieves the best performance on independent test set, which is consistent with the evaluation on the gold standard dataset.

### 3.3 Measuring the effects of the comprehensive similarity measures

The computation of comprehensive similarity measures includes three steps as described in Section 2.2. We compare four methods (BiRW1, BiRW12, BiRW13 and MBiRW) by cross-validation expereiments. In these methods, various similarity calculation steps are considered in similarity measures. For BiRW1, step one is involved in similarity measures, which means only the drug- or disease-related property information is used to calculate similarity. For BiRW12, both step one and step two were implemented in similarity measures, which means similarity measured by step one is adjusted by corresponding similarity analysis. For BiRW13, both step one and step three are taken in measuring similarity, which means similarity measured by step one is adjusted by taking known drug–disease associations into account. For MBiRW, all three steps are involved in similarity measures.

The comparison results in terms of ROC curve and top-ranked results by these methods are depicted in Supplementary Figure S4. It shows that BiRW1 is inferior to that of all other methods. The AUC value of BiRW1 is 0.87, and only 166 associations can be correctly retrieved at the top one. While the AUC values of BiRW12 and BiRW13 are 0.896 and 0.897, respectively, which means that prediction performance can be improved by integrating known association information into similarity measures. MBiRW outperforms all other methods in terms of AUC values, top-ranked results and PR curves, which further demonstrate the effectiveness of the comprehensive similarity measures.

### 3.4 Case studies

After confirming the excellent performance of our method based on cross validation, the capability of MBiRW in predicting novel drug–disease associations is further examined here. To predict novel indications for drugs, all the known drug–disease interactions in the gold standard are used as the training set, and the remaining drug–disease pairs which are not known to be associated in our gold standard dataset formed the set of candidate drug–disease associations. By applying the method of MBiRW, we can obtain the prediction scores for all candidate drug–disease pairs. MBiRW can predict the potential associations for all the drugs simultaneously. For each drug, the candidate diseases are ranked based on the prediction scores, and the top-10 predicted diseases as prediction results are collected. The predictions for all drugs are listed in Supplementary Table S2.

Here, we randomly choose several drugs and present their predictions to conduct case studies. The prediction results are confirmed based on some public databases, current clinical trails and literatures. The public databases include KEGG (Kanehisa *et al.*, 2014), DrugBank and CTD (Davis *et al.*, 2015). In these databases, some new added drug–disease associations provide a foundation for our validation. For each of the selected drugs, the top-5 potential diseases and evidences for the associations with the specified drug are listed in Supplementary Table S3. We find that some top-ranked predictions have been confirmed by existing researches. For instance, Levodopa is predicted to treat Alzheimer's disease, which has already been tested in clinical trail (ClinicalTrials\.gov, 2006a). Doxorubicin, indicated for non-small cell lung cancer (NSCLC), is predicted to treat small cell lung cancer (SCLC). This prediction is confirmed by CTD. In addition, for Doxorubicin, the prediction to treat prostate cancer has been tested in clinical trail (ClinicalTrials\.gov, 2006b). Cabergoline, indicated for Hyperprolactinemia, is predicted to treat Migrane, and this prediction has been studied in literature (Erkulwater and Pillai, 1989). Amantadine, indicated for Parkinson's disease, is predicted to treat Alzheimer's disease, and this treatment has been studied in literature (Cavestro *et al.*, 2006). In addition, Memantine, an amantadine derivative, is an *N*-methyl-D-aspartate (NMDA) receptor antagonist used in the treatment of Alzheimer's disease. These successful prediction instances further confirm that MBiRW has the potential to predict novel disease indications for drugs.

### 3.5 Experiments on other datasets

Most existing studies always conduct performance evaluation experiments based on one specified dataset, and the adaptability of algorithms for different datasets is ignored. To make the prediction results more convincing, experiments mentioned above are further conducted on another two datasets.

First, one new dataset, named as DNdatasets, is obtained from paper (Martínez *et al.*, 2015), which contains 4516 diseases annotated by Disease Ontology (DO) terms, 1490 drugs registered in

DrugBank and 1008 known drug–disease associations which were derived from DrugBank. Besides, another dataset named as Cdatasets, is produced by combining DNdatasets and the gold standard dataset used in this paper. Cdatasets include 663 drugs registered in DrugBank, 409 diseases listed in OMIM database and 2352 known drug–disease associations.

In DNdatasets, drug similarity $Sim_{r^p}$ is measured based on anatomical therapeutic chemical (ATC) codes, and disease similarity $Sim_{d^p}$ is measured based on Disease Ontology (DO) terms. In Cdatasets, drug similarity $Sim_{r^p}$ is measured based on chemical structures and disease similarity $Sim_{d^p}$ is measured based on phenotypes using MimMiner. The performance of MBiRW on DNdatasets and Cdatasets is evaluated in terms of tenfold cross validation and de novo prediction, respectively.

#### 3.5.1 Validation on DNdatasets

Firstly, we perform tenfold cross validation to validate the performance of our method on DNdatasets, the experiment results are depicted in Supplementary Figure S5. The parameters used in MBiRW are the same as those in experiments for evluating the gold standard dataset mentioned above. Here, the parameter settings may not be optimal for DNdatasets, and can be tuned to achieve the best results. According to the cross validation results, the performance of MBiRW and DrugNet is better than other methods, and the AUC value of MBiRW and DrugNet is 0.958 and 0.948, respectively. Although MBiRW only has slight better performance compared with DrugNet in terms of ROC curves and AUC values, while there has significant superiority of MBiRW over other methods in terms of top-ranked analysis and precision results.

Moreover, de novo prediction test is also conducted on DNdatasets. In DNdatasets, there are 550 drugs which have at least one known association, so we perform de novo prediction for these 550 drugs. The prediction results are shown in Supplementary Figure S6. Compared with tenfold cross validation results above, the AUC values of all methods in de novo prediction results improve slightly. The reason may be due to the fact that DNdatasets includes many drugs which have only one known association. In tenfold cross validation, all known associations are divided into ten parts, each part is taken in turn as the test set, and the remaining nine parts are served as the training set. So there exists more than one drug, which has only one related disease and is considered as drug without known associations. However, in each de novo drug–disease prediction, only the known drug–disease associations for the queried drug are removed. That is to say, the number of drugs which have no known disease indications in de novo prediction may be less than that in tenfold cross validation.

#### 3.5.2 Validation on Cdatasets

For Cdatasets, the tenfold cross validation experiments for drug–disease association prediction are implemented, and evaluation results are shown in Supplementary Figure S7. It can be seen from the results, MBiRW obtains AUC value of 0.934 and higher precision, outperforming all other methods significantly. As seen from top-ranked results, among the 2352 known drug–disease associations, 887 associations can be predicted at the top one by MBiRW. In contrast, only 427, 148 and 84 known associations are predicted at top one by HGBI, NBI and DrugNet, respectively.

Each drug has at least one known indication in Cdatasets, so we perform de novo prediction for all drugs, and the results are depicted in Supplementary Figure S8. Compared with tenfold cross validation, we can find both the AUC values and top-ranked results become worse in de novo test. For instance, the AUC value obtained in tenfold cross validation achieved 0.934, however, the AUC value only achieved 0.842 in de novo prediction. The number of top-one predicted associations become 241, which is significantly less than that in tenfold cross validation.

The superior performance evaluation results on DNdatasets and Cdatasets, further demonstrate MBiRW is reliable in predicting potential drug–disease associations. According to the above experiments, we conclude that the prediction performance can be improved by integrating more useful information in similarity measures and adopting BiRW algorithm.

## 4 Conclusion

Drug repositioning is a promising alternative for the drug development. In this work, we have presented a novel method named MBiRW uncovering the potential associations between drugs and diseases. The main contribution of our study is that we have devised novel similarity measures for drugs and diseases by integrating known drug–disease associations information effectively, and performed drug repositioning based on BiRW algorithm. The procedure of the similarity measures involve calculating drug and disease similarities based on some prior biological features; conducting correlation analysis on similarity values, and adjusting similarity by a logistic function based on the analysis results; clustering drugs and diseases based on the known drug–disease associations, and improving drug and disease similarity based on the identified clusters. Based on the calculated comprehensive similarities, drug network and disease network are constructed. Furthermore drug network is connected with disease network via known drug–disease associations to construct a drug–disease heterogeneous network. Finally, BiRW algorithm is adopted to prioritize candidate diseases for each drug and the effectiveness is validated on the collected datasets. In cross validation, all experimental results show that our method can effectively improve prediction performance compared with other approaches. Case studies about several drugs indicate that potential drug–disease association predicted by MBiRW could be more useful for biomedical research.

MBiRW is one powerful computational method that can predict candidate diseases for different drugs simultaneously, and also can predict novel disease indications for drugs without any known associated diseases information effectively. We have validated its prediction power in terms of cross validation and case studies. While the results of our approach are promising, its limitations should be acknowledged. First, MBiRW only uses the known drug–disease association information to improve the similarity of drug pairs and disease pairs. It should be noted that there exists more prior biological relevant information which can be reasonably utilized to improve similarity measures. Second, considering that there exists substantial targets data related with drugs and diseases, so we plan to construct comprehensive heterogenous network for drug repositioning by taking target information into account.

*Funding*: This work is supported in part by the National Natural Science Foundation of China under Grant No. 61232001, No.61370024, No. 61428209, and the Program for New Century Excellent Talents in University under Grant NCET-12-0547.

*Conflict of Interest*: none declared.

## References

Ashburn,T.T. *et al*. (2004) Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov*., **3**, 673–683.

Cavestro,C. *et al*. (2006) High prolactin levels as a worsening factor for migraine. *J. Headache Pain*, **7**, 83–89.

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

Chen,H. *et al*. (2015) Network-based inference methods for drug repositioning. *Comput. Math. Methods Med*., **2015**, 130620.

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

Chiang,A.P. and Butte,A.J. (2009) Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Therap*., **86**, 507–510.

ClinicalTrials.gov (2006a) Docetaxel, Doxorubicin, and Prednisone in Treating Patients With Advanced Prostate Cancer That Has Not Responded to Hormone Therapy.

ClinicalTrials.gov (2006b) Dopaminergic enhancement of learning and memory in healthy adults and patients with dementia/mild cognitive impairment.

Davis,A.P. *et al*. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res*., **43**, D914–D920.

Erkulwater,S. and Pillai,R. (1989) Amantadine and the end-stage dementia of Alzheimerś type. *Southern Med. J*., **82**, 550–554.

Fisher,R.A. and Yates,F. (1938) Statistical tables for biological, agricultural and medical research. Edinburgh: Oliver and Boyd.

Hamosh,A. *et al*. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **30**, 52–55.

Hurle,M.R. *et al*. (2013) Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Therap*., **93**, 335–341.

Gottlieb,A. *et al*. (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol*., **77**, 496.

Grabowski,H. (2004) Are the economics of pharmaceutical research and development changing? *Pharmacoeconomics*, **22**, 15–24.

Kanehisa,M. *et al*. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res*., **42**, D199–D205.

Li,J. *et al*. (2015) A survey of current trends in computational drug repositioning. *Brief. Bioinf*., **1**, 11.

Lipscomb,C.E. (2000) Medical subject headings (MeSH). *Bull. Med. Libr. Assoc*., **88**, 265–266.

Martínez,V. *et al*. (2015) DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. *Artif. Intell. Med*., **63**, 41–49.

Nepusz,T. *et al*. (2012) Detecting overlapping protein complexes in protein–protein interaction networks. *Nat. Methods*, **9**, 471–472.

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.

Steinbeck,C. *et al*. (2006) Recent developments of the chemistry development kit(CDK)-an open-source java library for chemo- and bioinformatics. *Curr. Pharm. Des*., **12**, 2111–2120.

Tanimoto,T. (1957) An Elementary Mathematical theory of Classification and Prediction. Internal IBM Technical Report.

Wang,W. *et al*. (2013) Drug target predictions based on heterogeneous graph inference. *Pac. Symp. Biocomput*, **18**, 53–64.

Wang,W. *et al*. (2014) Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics*, **30**, 2923–2930.

Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model*., **28**, 31–36.

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

Wu,C. *et al*. (2013) Computational drug repositioning through heterogeneous network clustering. *BMC Syst. Biol*., **7**, S6.

Van Driel,M.A. *et al*. (2006) A text-mining analysis of the human phenome. *Eur. J. Human Genet*., **14**, 535–542.

Vanunu,O. *et al*. (2010) Associating genes and protein complexes with disease via network propagation. *PLoS Comput. Biol*., **6**, e1000641.

Xie,M. *et al*. (2012) Prioritizing disease genes by bi-random walk. *Adv. Knowl. Discov. Data Mining*, **2**, LNCS, 7302, 292–303.

Yu,L. *et al*. (2015) Inferring drug–disease associations based on known protein complexes. *BMC Med. Genomics*, **8**, S2.