JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 26, Number 0, 2019

Pp. 1–13
DOI: 10.1089/cmb.2019.0063

Research Article

BiRWDDA: A Novel Drug Repositioning Method Based on Multisimilarity Fusion

CHAO-KUN YAN,[1] WEN-XIU WANG,[1] GE ZHANG,[1] JIAN-LIN WANG,[1] and ASHUTOSH PATEL[2]

ABSTRACT

The explosive growth of large-scale biological data enables network-based drug repositioning to be an important way of drug discovery, which can reduce the time and cost of drug discovery efficiently. Many existing approaches always construct drug–disease association network only based on some similarity measuring data for drug or disease, which ignore the impacts of different similarity measuring on predicting performance. In this study, we develop a new computational approach named BiRWDDA, which fused multiple similarity measures and bi-random walk to discover potential associations between drugs and diseases. First, multiple drug–drug similarity and disease–disease similarity are measured. Next, the information entropy of similarities measured based on different data are calculated to select proper similarities of drugs and diseases. Subsequently, improved drug–drug similarity and disease–disease similarity can be obtained by fusing similarities selected. Then, a logistic function is adopted to adjust the improved drug similarity and disease similarity. What is more, a heterogeneous network can be conducted by connecting the drug similarity network and the disease similarity network through known drug–disease associations. Finally, a bi-random walk algorithm is implemented on the heterogeneous network to predict potential drug–disease associations. Experimental results demonstrate that BiRWDDA outperforms the other state-of-the-art methods with average AUC of 0.930. Case studies for five selected drugs further verify the favorable prediction performance.

Keywords: bi-random walk, drug repositioning, heterogeneous network, information entropy, multisimilarity.

1. INTRODUCTION

Traditional drug discovery is a costly, time-consuming, risky, and ineffective process (Pammolli et al., 2011). Relevant surveys show that it always takes about 10–15 years and costs >$800 million to bring a new drug to market (Dudley et al., 2011). In the past decade, although the total worldwide cost of drug R&D has risen up to 141 billion, the number of drug approvals per year remains low (Schuhmacher et al., 2016). For the issue, drug repositioning or drug repurposing that refers to find new uses for existing drugs can

[1]School of Computer and Information Engineering, Henan University, Kaifeng, China.
[2]VU College, Victoria University, Melbourne, Australia.

provide a better risk-versus-reward trade-off and has attracted increasing interests from the research community and pharmaceutical industry (Hurle et al., 2013). So far, some successful repositioned drugs have generated historically high revenues for their patent holders or companies, such as Sildenafil, thalidomide, and raloxifene (Ashburn and Thor, 2004).

Essentially, the objective of drug repositioning is to identify the potential treatment of existing drugs. Currently, most existed repositioned drugs are the consequence of serendipitous observations of unexpected efficacy and side effects of drugs in development or on the market (Lee et al., 2012). To accelerate the drug development process, it needs to develop rational and systematic methods to discover new uses of old drugs on a large scale (Liu et al., 2016).

Generally speaking, existing computational drug repositioning methods can be classified into the following categories: machine learning based, text mining based, and network based (Li et al., 2015). Most machine learning-based methods take randomly generated associations between biomedical entities as negative samples, in which some unreliable negative samples are included and cause biased decision boundary (Cheng et al., 2017). The text mining approaches typically depend on occurrence and co-occurrence statistics of terms to infer associations between biomedical entities (Jelier et al., 2008). For instance, Li and Lu (2012) developed a text mining model to systematically identify PGx relevant relationships between genes, drugs, and diseases from trial records in ClinicalTrials.gov. Owing to name ambiguity between entity types and limited accuracy of text mining techniques, text mining approaches could not obtain desirable performance.

With the increasing accumulation of the topological and structural properties of complex biomedical networks, some network-based approaches have been developed to find new indications for existing drugs (Chen et al., 2015). For example, Chiang and Butte (2009) developed a guilt-by-association method, which measured the relationship between diseases to predict potential drug–disease association. Li and Lu (2013) developed a computational method to discover new uses of existing drugs based on causal inference in a layered drug-target-pathway-gene-disease network. Yu et al. (2015) constructed a tripartite network consisting of drugs, protein complexes and disease, and inferred the weighted relationships between drugs and diseases. Based on the observation that similar drugs are indicated for similar diseases, Gottlieb et al. (2011) utilized multiple drug–drug and disease–disease similarity measures for the prediction task. Napolitano et al. (2013) combined the drug similarities (e.g., target protein similarity and chemical structure similarity) into a single information layer used to train a multiclass SVM classifier. Martínez et al. (2015) have developed DrugNet, a network-based prioritization method, which integrated drugs, disease, and targets to perform disease–drug and drug–disease prioritization. It is conceivable that these methods can help to improve the prediction performance by fusing multiple related sources (Zhang et al., 2017).

However, previous studies seldom utilized the known drug–disease association information of data set to improve similarity measures. Luo et al. (2016) proposed a novel computational method named MBiRW, which utilizes some comprehensive similarity measures and bi-random walk algorithm to predict potential indications for existing drugs.

In this study, we propose a novel computational drug repositioning method named BiRWDDA, which fused multiple similarity measures and utilizes bi-random walk algorithm to find potential drug–disease associations. First, based on different available biological data, BiRWDDA calculates multiple drug–drug similarity and disease–disease similarity, separately. Next, the information entropy is calculated for different drug similarity and disease similarity to select a set of more informative similarity set. Subsequently, to obtain the improved drug–drug similarity and disease–disease similarity we construct similarity fusion based on the aforementioned similarity selection. Then BiRWDDA adopts a logistic function to adjust the improved drug similarity and disease similarity. Finally, bi-random walk algorithm is adopted on the heterogeneous network to predict new drug–disease associations. Furthermore, we utilize 10-fold cross validation to evaluate the performance of BiRWDDA. The experiment results show that the proposed model obtains superior performance in predicting novel indications for existing drugs.

2. MATERIALS AND METHODS

In this section, a novel multisimilarity fusion drug repositioning method using bi-random walk algorithm was used to predict new indications for approved drugs. First, we introduce the data set used in this article. Second, we compute multiple similarity measures and then implement similarity selection to get an optimized combination of similarity measures for drugs and diseases, respectively. Next, a heterogeneous

network is constructed by connecting the drug similarity network and the disease similarity network through known drug–disease associations. Finally, bi-random walk algorithm is implemented on the heterogeneous network to predict potential drug–disease associations. Figure 1 shows the flowchart of BiRWDDA.

2.1. Data set

The gold standard data sets used in this work is obtained from Gottlieb et al. (2011), which is collected from multiple data sources. This data set includes 1933 known interactions between 593 drugs and 313 diseases. Drugs are collected from DrugBank (Wishart et al., 2007), and diseases are registered in the Online Mendelian in Man (OMIM) database (Hamosh et al., 2005). Figure 2 demonstrates the statistics of this data set in detail.

2.2. Similarity measures

In this section, we present the process of improved similarity measure. First, we define and compute four drug–drug similarity measures and three disease–disease similarity measures based on some drug-related properties and disease-related properties, separately; Then, informative similarity measures can be selected based on information entropy and fused into new drug similarity and disease similarity; Finally, we adjust similarity values by applying a logical function to obtain improved similarity for drug and disease, respectively.

2.2.1. Drug similarity measures. In this section, four drug–drug similarity measures are introduced as follows:

1. Sequence based: The protein sequence information of the drugs is obtained from the UniProt database (Apweiler et al., 2004), based on a Smith–Waterman sequence alignment score (Smith et al., 1985)

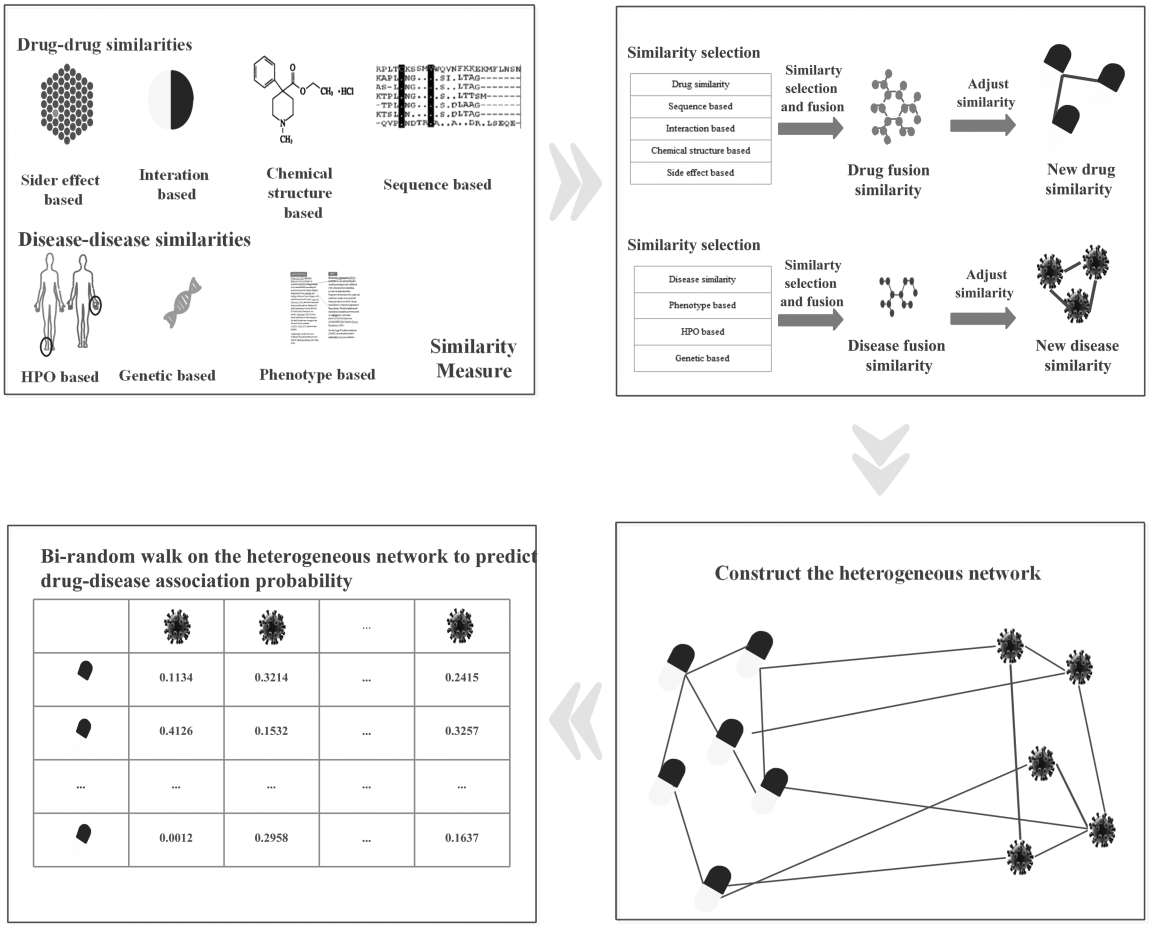

FIG. 1. The flowchart of BiRWDDA.

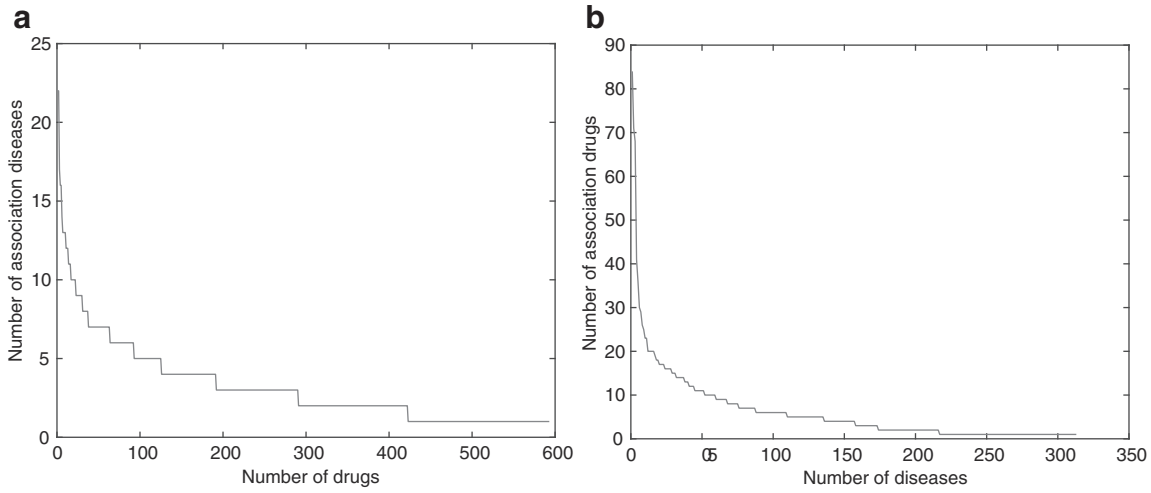

FIG. 2. Statistics of drug–disease incidence matrix. **(a)** Number of indicated diseases per drug. **(b)** Number of drugs per indicated disease.

between the corresponding drug-related genes. After the normalization suggested in Bleakley and Yamanishi (2009), we divide the Smith–Waterman score by the geometric mean of the scores obtained from aligning each sequence against itself. For sequence s and s', the normalized drug sequence similarity calculation is demonstrated as follows:

$$sim_{ss'} = \frac{sw(s, s')}{\sqrt{sw(s, s)}\sqrt{sw(s', s')}}, \tag{1}$$

where $sim_{ss'}$ represents the similarity between drugs s and s', $sw(.,.)$ represent the original Smith–Waterman score.

2. Interaction based: Drug interaction refers to the compound effect of the patient taking two or more drugs at the same time or within a certain period, which can enhance the efficacy or reduce the side effects and can also weaken the efficacy or cause the undue poison side effect. Therefore, the interaction between drugs is of great significance in the process of drug repositioning. Drug interaction information can be extracted from the DrugBank. Then the Jaccard score is used to calculate the similarity of drug interactions, as follows:

$$sim_{ii'} = \frac{|I_i \cap I_{i'}|}{|I_i \cup I_{i'}|}, \tag{2}$$

where I_i and $I_{i'}$ represent the set of drugs rated on drugs i and i'.

3. Chemical structure based: The canonical SMILES (Weininger, 1988) of the drug are obtained from DrugBank. The Chemical Development Kit (Steinbeck et al., 2006) is used to compute the similarity between two drugs as the Tanimoto score of their 2D chemical fingerprints.
4. Side effect based: Drug side effects are obtained from SIDER (Kuhn et al., 2010). We augmented this list by side effect predictions for drugs that are not included in SIDER based on their chemical properties (Atias and Sharan, 2011). We define the similarity of two drugs in terms of the Jaccard score between known side effects of the drug or between the top 10 predicted side effects in an unknown situation.

2.2.2. Disease similarity measures. In this section, three disease–disease similarity measures are introduced as follows:

1. Phenotype based: Disease phenotypes similarity $simd_{ph}$ is computed using MimMiner (Van Driel et al., 2006), which is constructed by calculating the similarity between MeSH terms (Lipscomb, 2000) appearing in the medical description of diseases extracted from the OMIM database.

2. Semantic phenotypic similarity: Semantic phenotypic similarity information of the diseases is obtained from the OMIM database. The semantic similarity score $simd_{hp}$ is calculated based on Resnik (1999) by using the hierarchical structure of the HPO (Robinson and Mundlos, 2010) together with the mapping provided by HPO between ontology nodes and OMIM diseases.
3. Genetic based: The genetic signatures of the diseases are collected from gene expression experiments, the Jaccard score is used between each pair of signatures, considering the direction of the response of each gene.

2.3. Similarity selection

It may introduce noise into the data if we combine all similarity information. To choose a more robust set of similarities, we utilize information entropy to select similarity; the procedure goes as follows:

1. The average entropy of each similarity measure is calculated, which decides how much information each similarity takes. For a similarity matrix M for drug or disease, we calculate entropy E_i for each row i as:

$$E_i = -\sum_{j=1}^{k} \frac{m_{ij}}{\sum_{j=1}^{k} m_{ij}} \log\left(\frac{m_{ij}}{\sum_{j=1}^{k} m_{ij}}\right), \tag{3}$$

where, m_{ij} represent the similarity value between drug (or disease) i and drug (or disease) j. k denotes the number of drugs (or diseases).

2. To get the final average entropy value, we average the entropy values of all matrix rows, which describe how informative a similarity is.

$$E_{mean} = \frac{\sum_{i=1}^{k} E_i}{k}. \tag{4}$$

The smaller the average entropy means less random information is introduced in the similarity measure.

We calculate the information entropy of four drug–drug similarity measures and three disease–disease similarity measures.

According to the results shown in Table 1, comparing with other similarity measure, drug sequence based and drug interaction based have smaller information entropy. For similarity measure, smaller information entropy means more informative than they can provide. Likewise, for disease similarity measure, we can see from Table 2 that phenotype- and semantic-based measures are more informative. As a result, these similarity measures with smaller information entropy are selected to conduct the following similarity fusion.

2.4. Similarity fusion

Based on similarity selection, given the selected similarity measures obtained previously for drugs and diseases, respectively, the purpose of the similarity fusion is to combine multiple similarity measures into

Table 1. Information Entropy of Different Drug Similarity Measure

Drug similarity matrix	*Information entropy*
Sequence similarity	7.926456540713852
Interaction similarity	8.206457749641498
Chemical similarity	9.047256161965553
Side effect similarity	8.453597635459150

Table 2. Information Entropy of Different Disease Similarity Measure

Disease similarity matrix	*Information entropy*
Phenotype similarity	7.538069246904729
Semantic phenotypic similarity	7.978509314340557
Gene information similarity	8.035719478595299

one similarity measure that captures the information from different similarities. The smaller the average entropy means less random information is introduced in the similarity measure. Thus, we computed the fused similarity measure, which is described as follows:

$$frsim = \begin{cases} 0 & if\ r1\ and\ r2 = 0 \\ r1 & if\ r2 = 0 \\ r2 & if\ r1 = 0 \\ r1 + \frac{r2}{2} & otherwise \end{cases} . \tag{5}$$

In particular, $r1$ is drug sequence similarity or disease phenotype similarity, $r2$ is drug interaction similarity or disease semantic phenotypic similarity; we can generate the fusion similarity through $r1$ and r2.

2.5. Adjusting the similarity matrices

According to the aforementioned fusion similarity for drugs and diseases. We execute the correlation analysis on drug similarity and disease similarity by taking into account known drug–disease association's information of the gold standard data set.

Drug pairs with similarity values within the range of [0, 0.3] have an insignificant probability of treating common diseases, and drug pairs with similarity values within the scope of [0.6, 1] have a significant probability of treating common disease. Then for disease pairs with similarity values within the range of [0, 0.3] have an insignificant probability of sharing drugs, and disease pairs with similarity values within the scope of [0.6, 1] have a significant probability of sharing drugs. The evidence is shown in Supplementary Figure S1. In this study, we adjust similarity by applying a logistic function (Vanunu et al., 2010), as follows:

$$L(s_i, s_j) = \frac{1}{1 + e^{(c \cdot sim(s_i, s_j) + d)}}, \tag{6}$$

where $sim(s_i, s_j)$ is the drug pairs similarity or disease pairs similarity, c and d are parameters that are used to control the adjustment of the sim. Then for similarity values $sim(s_i, s_j) \in [0, 0.3]$, we set $L(s_i, s_j) \approx 0$; and for $sim(s_i, s_j) \in [0.6, 1]$, $L(s_i, s_j) \approx 1$. when $sim(s_i, s_j) = 0$, we set $L(s_i, s_j) = 0.0001$, which set d as $\log(9999)$, and c as -15. In our study, L denotes the final similarity.

2.6. Construction of the heterogeneous network

Based on the drug similarity and disease similarity calculated earlier, both the drug similarity network and the disease similarity network can be constructed. In the drug similarity network, let $R = \{R_1, R_2, \ldots, R_m\}$ denote the node set of m drugs. The edge between two drugs is weighted by the similarity value of these two drugs. In the disease similarity network, let $D = \{D_1, D_2, \ldots, D_n\}$ denote the node set of n diseases. The edge between two diseases is weighted by the similarity value of these two diseases.

A drug–disease association network is a heterogeneous network composed of a drug similarity network, a disease network, and the drug–disease association network modeled by a bipartite graph $G_{R,D}$. $G_{R,D} = \{\{R, D\}, \{E_{rr}, E_{dd}, E_{rd}\}\{W_{rr}, W_{dd}, W_{rd}\}\}$. Let E_{rr}, E_{dd}, and E_{rd} denote drug–drug, disease–disease, and drug–disease edges, respectively, and W_{rr}, W_{dd}, and W_{rd} represent the weights on these three kinds of edges. If there has been a known association between drug R_m and disease D_n, W_{rd} is initially set to 1; otherwise, it is initially set to 0.

2.7. The BiRWDDA method

In this study, we use bi-random walk algorithm on the heterogeneous network to predict drug–disease association.

The drug similarity network and the disease similarity network contain diverse topologies and structures; therefore, the optimal number of random walk steps may be disparate on the two networks. To solve this problem, we restrict the number of random walk steps on the two network by setting two parameters l and r as the numbers of maximal iterations. The iterative process is written as follows:

Random walk on the drug similarity network:

$$R_r = \alpha \cdot RS \cdot RE_{t-1} + (1-\alpha) \cdot A. \tag{7}$$

Random walk on the disease similarity network:

$$R_d = \alpha \cdot RE_{t-1} \cdot DS + (1-\alpha) \cdot A, \tag{8}$$

where α stands for the decay factor; $RS^{m \cdot m}$, $DS^{n \cdot n}$, and $A^{m \cdot n}$ denote the matrix of the drug similarity network, the disease similarity network, and the drug–disease association network, respectively, where m is the number of drugs and n is the number of diseases. The objective is based on the heterogeneous drug–disease association network to predict the missing associations by reconstructing an association matrix $R(i, j)$. R_r and R_d refer to the correlations between drug and disease based on the walk on the two networks, respectively. $R_r(i, j)$ and $R_d(i, j)$ represent the predicted probability between drug R_i and disease D_j. R is the average output from drug similarity network and disease similarity network in each step. The complete BiRWDDA algorithm for inferring potential drug–disease associations is outlined as in Algorithm 1.

Algorithm 1. BiRWDDA

Input: drug set R, disease set D, drug–disease association adjacency matrix A, parameters α l and r.
Output: predicted drug–disease association matrix RE.
bi-random Walk(R, D, A, l, r)

1. Construct drug similarity matrix RS_{seq}, RS_{ddi}, RS_{che}, and RS_{sid}, and disease similarity matrix DS_{phe}, DS_{hp}, and DS_{gene};
2. Select drug similarity matrix RS_{seq}, RS_{ddi} and disease similarity matrix DS_{phe}, DS_{hp};
3. Combine drug similarity $CSimR$, disease similarity $CSimD$;
4. Get new drug similarity $SimR$, new disease similarity $SimD$;
5. //Laplacian normalization $D_{simR}(i, i)$ is the sum of row i of $SimR$
6. $RS = \mathrm{SimR}^{-1/2} \cdot D_{SimR} \cdot SimR^{-1/2}$
7. //Laplacian normalization $D_{simD}(i, i)$ is the sum of row i of $SimD$
8. $DS = SimD^{-1/2} \cdot D_{SimD} \cdot SimD^{-1/2}$
9. //RE_0 is the initial probability
10. $RE_0 = A = A/sum(A)$
11. for $r = 1$ to max(l,r)
12. $rflag = dflag = 1$
13. if $t <= l$
14. //random walk on the drug similarity networks
15. $R_r = \alpha \cdot RS \cdot RE_{t-1} + (1-\alpha) \cdot A$
16. $rflag = 1$
17. **e**ndif
18. if $t <= r$
19. //random walk on the disease similarity networks
20. $R_d = \alpha \cdot RE_{t-1} \cdot DS + (1-\alpha) \cdot A$
21. $dflag = 1$
22. endif
23. //combination of the results
24. $RE_t = (rflag \cdot Rr + dflag \cdot Rd)/(rflag + dflag)$
25. endfor
26. return(RE)

3. EXPERIMENTS AND RESULTS

In this section, we use the golden standard data sets to evaluate the prediction ability of our BiRWDDA model. First, we introduce evaluation metrics used in this article. Then, we compare BiRWDDA with other several state-of-the-art models. Next, we conduct case studies to verify the ability of BiRWDDA in identifying new indications.

3.1. Evaluation metrics

To test the prediction ability of BiRWDDA, we implement 10-fold cross validation on the gold standard data sets to compute the association probabilities of drug–disease pairs. All known drug–disease pairs in the golden data sets are randomly divided into 10 equal subsets. In each round of 10-fold cross validation, each subset is held out in turn as the test set, whereas the remaining data are merged with the training set. After performing BiRWDDA, each drug–disease pairs is assigned a predicted score. In terms of the predicted score, for each drug, the test drug–disease pairs and the candidate associations (all uncertain drug–disease pairs until now) are sorted in descending order. For a given rank threshold, true positive rate (TPR), false positive rate (FPR), Precision and Recall can be obtained. TPR is the proportion of all known drug–disease pairs are correctly predicted, FPR is the proportion of all unconfirmed drug–disease pairs that are predicted, Precision is the proportion of all known drug–disease pairs that appear in the ranked list based on a given threshold, Recall is the same as TPR. We further computed various TPR, PPR, Precision, and Recall with different thresholds to construct receiver operating characteristic (ROC) curve and the Precision–Recall curve. The AUC (the area under of ROC curves) and precision are utilized to evaluate the performance of the prediction models.

To reduce the random set division bias, 10-fold cross validation is repeated 10 times by randomly dividing the sets in each time.

3.2. Effect of parameters

There are three parameters in BiRWDDA, the parameter α is the decay factor, the parameters l and r are the numbers of maximal iterations in the drug and disease similarity network, respectively. To test the impact of these three parameters, we set different values for these three parameters and got the value of AUC by 10-fold cross validation. The experimental result is reported in Supplementary Table S1. It shows that BiRWDDA achieves better performance when parameters l and r are equal. Based on the AUC values, the three parameters used in BiRWDDA are chosen as $\alpha = 0.2$, $l = 2$ and $r = 2$ in our study.

3.3. Comparison with other methods

In this section, to evaluate the performance of the proposed model, we compare the prediction performance of BiRWDDA model with some other state-of-the-art methods: MBiRW (Luo et al., 2016), HGBI (Wang et al., 2013). MBiRW uses comprehensive similarity and bi-random walk on the heterogeneous network to predict potential drug–disease associations. HGBI is based on the guilt-by-association principle and an intuitive interpretation of information flow on the heterogeneous graph.

Especially, the parameter setting for each method is set to its default setting or best parameter values, the parameters (λ, l, r) are set to (0.2, 2, 2) for BiRWDDA. MBiRW is set to its default setting, the parameter α is set to 0.3, and l and r are equal to 2. For HGBI, the parameter α is set to its default value 0.4.

The evaluation results of all methods according to ROC curves and PR curves are reported in Figure 3. It can be found that BiRWDDA achieves the highest AUC value of 0.930, whereas MBiRW and HGBI obtain inferior results of 0.919 and 0.814, respectively. In terms of PR curves, we can see that BiRWDDA achieves the best precision, whereas HGBI obtains inferior precision.

3.3. Comparison with the different data source

In our study, multiple data sources include DDI (drug interaction), SEQ (protein sequence), HP (phenotype), and PHE (semantic phenotypic) are measured and integrated to improve the performance of prediction. To evaluate the impact of different integration on the algorithm, we conduct 10-fold cross

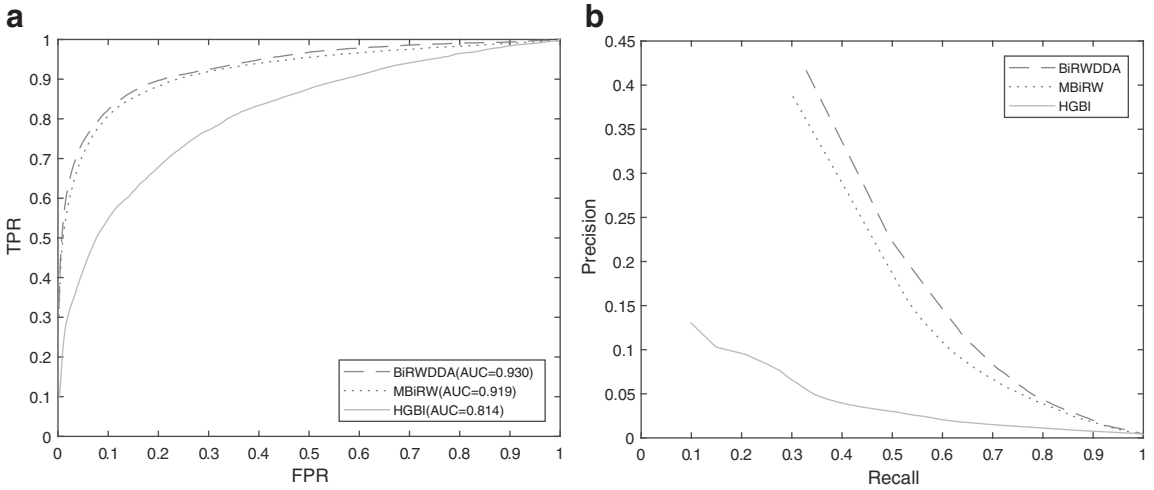

FIG. 3. Prediction results of different methods in identifying potential drug–disease associations. **(a)** ROC curves of prediction results obtained by applying various methods. **(b)** PR curves for the various drug–disease prediction methods. PR, positive rate; ROC, receiver operating characteristic.

validation on these data sets. Figure 4 shows the evaluation result. We can see from Figure 4 that BiRWDDA achieves the best performance (AUC = 0.930), which is based on the principle that it is beneficial to building a robust prediction method with high efficiency by fusing multiple related sources.

4. CASE STUDIES

After confirming the performance of BiRWDDA in terms of 10-fold cross validation, to further validate the capability of our method in predicting new drug–disease indications, all the known drug–disease pairs in the gold standard data set are used as the training set, and the remaining unknown drug–disease pairs are considered to be the candidate associations. By applying BiRWDDA, we can get the prediction scores for all candidate drug–disease associations. For a specific drug, all the candidate diseases are ranked according

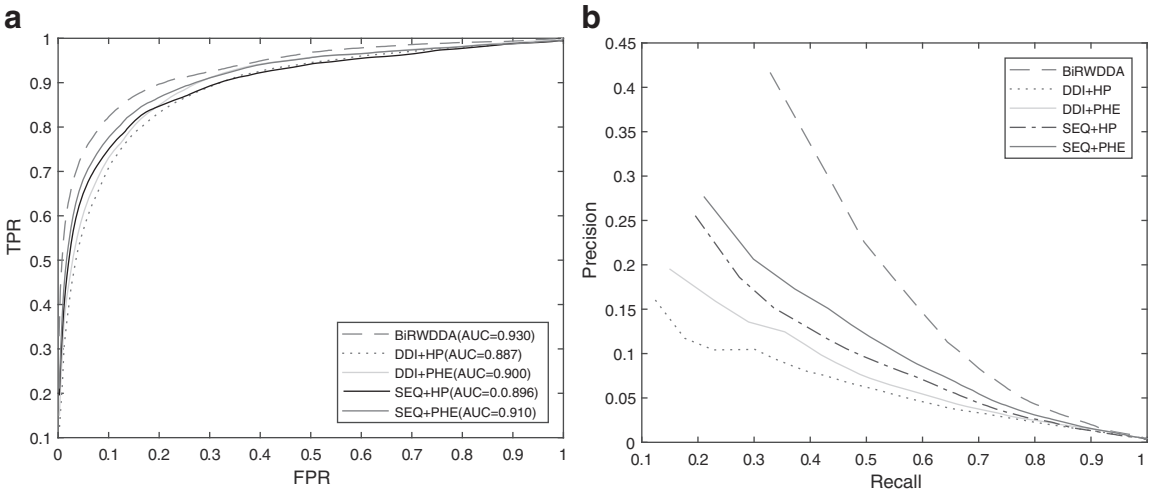

FIG. 4. Prediction results of different data source in identifying potential drug–disease associations. **(a)** ROC curves predict drug–disease associations by using various data sources. **(b)** The comparison between various data sources in terms of PR curves on the golden standard data set.

Table 3. Case Studies About Five Chosen Drugs: Paclitaxel, Prednisone, Risperidone, Doxorubicin, and Dexamethasone

Drug	*Disease*	*Rank*	*Evidence*
DB01229	D236000	1	
Paclitaxel	D276300	2	
	D176807	3	ClinicalTrials.gov
	D273300	4	ClinicalTrials.gov
	D246470	5	
DB00635	D600807	1	ClinicalTrials.gov/KEGG
Prednisone	D603165	2	KEGG
	D601608	3	
	D266600	4	ClinicalTrials.gov/KEGG
	D246470	5	
DB00734	D147530	1	
Risperidone	D161900	2	
	D164230	3	ClinicalTrials.gov
	D608622	4	
	D143465	5	ClinicalTrials.gov
DB00997	D223350	1	
Doxorubicin	D182280	2	ClinicalTrials.gov
	D114500	3	ClinicalTrials.gov
	D267730	4	
	D144700	5	ClinicalTrials.gov
DB01234	D600807	1	ClinicalTrials.gov/KEGG
Dexamethasone	D603165	2	KEGG
	D304790	3	
	D147540	4	
	D151590	5	

For each drug, the top five ranked predictions are listed in the table.

to their prediction scores, and we collect the top 10 predicted diseases as prediction results. For all drugs, the prediction results are listed in Supplementary Table S2.

We conduct case studies for the predicted top-ranked diseases in terms of public biological databases KEGG (Kanehisa et al., 2013) and current clinical trials to verify the prediction results are true or not. In KEGG database, some newly verified drug–disease pairs provide a foundation for our validation. As an example, we choose several drugs and corresponding top five candidate diseases, as shown in Table 3. We find that some novel drug–disease pairs have been confirmed in KEGG database or clinical trials on web ClinicalTrials.gov. For example, Paclitaxel has been predicted potential therapy for Prostate cancer and TGCT (testicular germ cell tumor). Risperidone has been predicted potential therapy for OCD (obsessive-compulsive disorder) and ADHD (attention-deficit/hyperactivity disorder). Doxorubicin has been predicted potential therapy for small cell cancer of the lung, CRC (colorectal cancer), RCC (renal cell carcinoma, nonpapillary). Those have been confirmed in clinical trials. Prednisone has been predicted potential therapy for asthma bronchial, dermatitis atopic, and Crohn's disease. Dexamethasone has been predicted as a potential therapy for prostate cancer and TGCT. They have been verified in clinical trials or KEGG database. These successful case studies show that our proposed method has a strong ability in predicting new drug–disease associations.

5. CONCLUSION

Drug repositioning is a high-efficiency method to discover new indications for a given drug. In this study, we propose a novel drug repositioning approach named BiRWDDA to predict potential drug–disease associations. The main idea of our model is that we have designed novel similarity measures for drugs and

diseases by fusing multisource data, and BiRW algorithm is adopted to perform drug repositioning. First, we compute multiple drug–drug similarity and disease–disease similarity and then combine similarity for drug and disease, respectively. Before applying the combined similarity method, BiRWDDA utilizes information entropy to select similarity because of combining all similarity types may introduce noise in the data. Subsequently, BiRWDDA adjusts similarity by a logistic function to a comprehensive similarity. Next, drug similarity network and disease similarity network are constructed and they are incorporated into a heterogeneous network through known drug–disease associations. Finally, BiRWDDA adopts BiRW algorithm to identify new indications for existing drugs. In 10-fold cross validation, the result shows that BiRWDDA is feasible and effective. Compared with other several state-of-the-art models, the proposed method can effectively improve prediction performance. Case studies further demonstrate that it is reliable in identifying novel indications for existing drugs.

Despite we have confirmed its power according to cross validation and case studies, there are still some limitations. First, the gold standard data set is incomplete, so predictive performance will be limited. It can be solved by increasing drug–disease association discovered. In addition, we plan to integrate more reliable data (such as target data related to drugs and diseases) to construct a comprehensive heterogeneous network in future studies.

ACKNOWLEDGMENT

This study was supported in part by the National Natural Science Foundation of China (nos. 61802113, 61802114, and 61602156).

AUTHORS' CONTRIBUTIONS

C.-K.Y. and W.-X.W. conceived and designed the approach. W.-X.W. performed the experiments. C.-K.Y. and W.-X.W. analyzed the data. C.-K.Y., W.-X.W., J.-L.W., and G.Z. wrote the article. All authors read and approved the final article.

AUTHOR DISCLOSURE STATEMENT

The authors declare there are no competing financial interests.

SUPPLEMENTARY MATERIAL

Supplementary Figure S1
Supplementary Table S1
Supplementary Table S2

REFERENCES

Apweiler, R., Bairoch, A., Wu, C.H., et al. 2004. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* 32(suppl_1), D115–D119.

Ashburn, T.T., and Thor, K.B. 2004. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* 3, 673.

Atias, N., and Sharan, R. 2011. An algorithmic framework for predicting side effects of drugs. *J. Comput. Biol.* 18, 207–218.

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

Chen, Y., Li, L., Zhang, G.Q., et al. 2015. Phenome-driven disease genetics prediction toward drug discovery. *Bioinformatics* 31, i276–i283.

Cheng, Z., Huang, K., Wang, Y., et al. 2017. Selecting high-quality negative samples for effectively predicting protein-RNA interactions. *BMC Syst. Biol.* 11, 9.

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.

Dudley, J.T., Deshpande, T., and Butte, A.J. 2011. Exploiting drug–disease relationships for computational drug repositioning. *Brief. Bioinform.* 12, 303–311.

Gottlieb, A., Stein, G.Y., Ruppin, E., et al. 2011. PREDICT: A method for inferring novel drug indications with application to personalized medicine. *Mol. Syst. Biol.* 7, 496.

Hamosh, A., Scott, A.F., Amberger, J.S., et al. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33(suppl_1), D514–D517.

Hurle, M.R., Yang, L., Xie, Q., et al. 2013. Computational drug repositioning: From data to therapeutics. *Clin. Pharmacol. Therap.* 93, 335–341.

Jelier, R., Schuemie, M.J., Veldhoven, A., et al. 2008. Anni 2.0: A multipurpose text-mining tool for the life sciences. *Genome Biol.* 9, R96.

Kanehisa, M., Goto, S., Sato, Y., et al. 2013. Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.* 42(D1), D199–D205.

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

Lee, H.S., Bae, T., Lee, J.H., et al. 2012. Rational drug repositioning guided by an integrated pharmacological network of protein, disease and drug. *BMC Syst. Biol.* 6, 80.

Li, J., and Lu, Z. 2012. Systematic identification of pharmacogenomics information from clinical trials. *J. Biomed. Inf.* 45, 870–878.

Li, J., and Lu, Z. 2013. Pathway-based drug repositioning using causal inference. *BMC Bioinformatics* 14, S3.

Li, J., Zheng, S., Chen, B., et al. 2015. A survey of current trends in computational drug repositioning. *Brief. Bioinform.* 17, 2–12.

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

Liu, H., Song, Y., Guan, J., et al. 2016. Inferring new indications for approved drugs via random walk on drug-disease heterogenous networks. *BMC Bioinformatics* 17, 539.

Luo, H., Wang, J., Li, M., et al. 2016. Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. *Bioinformatics* 32, 2664–2671.

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

Napolitano, F., Zhao, Y., Moreira, V.M., et al. 2013. Drug repositioning: A machine-learning approach through data integration. *J. Cheminform.* 5, 30.

Pammolli, F., Magazzini, L., and Riccaboni, M. 2011. The productivity crisis in pharmaceutical R&D. *Nat. Rev. Drug Discov.* 10, 428.

Resnik, P. 1999. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.* 11, 95–130.

Robinson, P.N., and Mundlos, S. 2010. The human phenotype ontology. *Clin. Genet.* 77, 525–534.

Schuhmacher, A., Gassmann, O., and Hinder, M. 2016. Changing R&D models in research-based pharmaceutical companies. *J. Transl. Med.* 14, 105.

Smith, T.F., Waterman, M.S., and Burks, C. 1985. The statistical distribution of nucleic acid similarities. *Nucleic Acids Res.* 13, 645–656.

Steinbeck, C., Hoppe, C., Kuhn, S., 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.

Van Driel, M.A., Bruggeman, J., Vriend, G., et al. 2006. A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* 14, 535.

Vanunu, O., Magger, O., Ruppin, E., et al. 2010. Associating genes and protein complexes with disease via network propagation. *PLoS Computat. Biol.* 6, e1000641.

Wang, W., Yang, S., and Li, J.I.N.G. 2013. Drug target predictions based on heterogeneous graph inference. *Pac. Symp. Biocomput.* 2013, 53–64.

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

Wishart, D.S., Knox, C., Guo, A.C., et al. 2007. DrugBank: A knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.* 36(suppl_1), D901–D906.

Yu, L., Huang, J., Ma, Z., et al. 2015. Inferring drug-disease associations based on known protein complexes. *BMC Med. Genomics* 8, S2.

Zhang, J., Li, C., Lin, Y., et al. 2017. Computational drug repositioning using collaborative filtering via multi-source fusion. *Expert Syst. Appl.* 84, 281–289.

Address correspondence to:
Ge Zhang, PhD
Room 209, School of Computer and Information Engineering
JinMing Campus, Henan University
Kaifeng City, Henan Province
China

E-mail: zhangge@henu.edu.cn