Computational Drug Repositioning with Random Walk on a Heterogeneous Network

Huimin Luo, Jianxin Wang, Min Li, Junwei Luo, Peng Ni, Kaijie Zhao, Fang-Xiang Wu, and Yi Pan

Abstract—Drug repositioning is an efficient and promising strategy to identify new indications for existing drugs, which can improve the productivity of traditional drug discovery and development. Rapid advances in high-throughput technologies have generated various types of biomedical data over the past decades, which lay the foundations for furthering the development of computational drug repositioning approaches. Although many researches have tried to improve the repositioning accuracy by integrating information from multiple sources and different levels, it is still appealing to further investigate how to efficiently exploit valuable data for drug repositioning. In this study, we propose an efficient approach, Random Walk on a Heterogeneous Network for Drug Repositioning (RWHNDR), to prioritize candidate drugs for diseases. First, an integrated heterogeneous network is constructed by combining multiple sources including drugs, drug targets, diseases and disease genes data. Then, a random walk model is developed to capture the global information of the heterogeneous network. RWHNDR takes advantage of drug targets and disease genes data more comprehensively for drug repositioning. The experiment results show that our approach can achieve better performance, compared with other state-of-the-art approaches which prioritized candidate drugs based on multi-source data.

Index Terms—Drug repositioning, random walk, heterogeneous network

1 Introduction

Despite the increasing investments in pharmaceutical research and development (R&D), the number of new drugs that are approved by the US Food and Drug Administration (FDA) annually remains low [1]. Drug discovery and development is still a risky, time-consuming and tremendously costly process. Indeed, bringing a new drug from discovery to market involves multiple research stages, and it can take about 15 years and US$800 millions [2], [3]. Meanwhile, the probability of success of drug discovery is relatively low due to efficacy and safety in clinical trials, and approximately 30 percent of the failures are linked to clinical toxicology [4], [5], [6]. In light of these challenges, drug repositioning has become an increasingly important part of the drug development landscape. The aim of drug repositioning is to identify and develop new therapeutic indications for existing drugs (referred to as indication discovery) and apply the newly identified drugs to the treatment of diseases other than the drug's originally intended disease [7]. As repositioning candidates have frequently gone through several phases of development for their original indication, then these phases common to de novo drug development can be bypassed to reduce time and risk of drug discovery [5]. In recent years, governments, academic researchers, and the pharmaceutical companies have launched large-scale funding and activities to support drug repositioning-related researches [8].

With the constant growth of drug-related and disease-related data, there have been a number of computational approaches to repurpose drugs, including machine learning, network, text mining and semantic inference based approaches [8]. Among these methods, network-based strategy is increasingly attracting much attention from the pharmaceutical community in recent years and widely used in computational drug repositioning due to the advances of high-throughput technology and a growing number of available data sources (e.g., genetic, pharmacogenomics, clinical, chemical agent, etc.). For example, Chiang and Butte [9] developed a network-based method to predict novel drug-disease associations based on guilt-by-association principle. This method produced novel drug indications based on shared treatment profiles from any disease pairs which shared at least one FDA approved drug in common. Wu et al. [10] applied graph clustering algorithms to detect disease-drug modules in a weighted disease and drug heterogeneous network. Based on the identified modules, all drug-disease combinations of a module were assembled as drug repositioning candidates. Chen et al. [11] applied inference method to predict potential drug-disease associations only based on topology information of the

- *H. Luo is with the School of Information Science and Engineering, Central South University, Changsha 410083, China, and the School of Computer and Information Engineering, Henan University, KaiFeng 475001, China. E-mail: luohuimin@csu.edu.cn.*
- *J. Wang, M. Li, J. Luo, P. Ni, and K. Zhao are with the School of Information Science and Engineering, Central South University, Changsha 410083, China. E-mail: {jxwang, limin}@mail.csu.edu.cn, {luojunwei, nipeng, kay.zkj}@csu.edu.cn.*
- *F.-X. Wu is with the Division of Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, Saskatoon, SKS7N5A9, Canada. E-mail: faw341@mail.usask.ca.*
- *Y. Pan is with the Department of Computer Science, Georgia State University, Atlanta, GA 30302. E-mail: yipan@gsu.edu.*

Manuscript received 11 Apr. 2017; revised 15 Nov. 2017; accepted 18 Apr. 2018. Date of publication 2 May 2018; date of current version 5 Dec. 2019. (Corresponding author: Jianxin Wang.)
Digital Object Identifier no. 10.1109/TCBB.2018.2832078

constructed drug-disease bipartite network model. Luo et al. [12] proposed a computational method to find novel indications for existing drugs by applying comprehensive similarity measures and a bi-random walk algorithm. Moreover, matrix completion algorithm has been proposed to fill drug-disease association matrix and identify potential treatments for diseases [13]. Although these approaches have achieved better prediction performance in identifying novel disease-drug associations, biological target networks have not been integrated in the constructed drug-disease network model of the above researches.

Along with the increase of identified substantial drug-target and disease-gene associations in recent researches, drug targets and disease genes could be further integrated for drug repositioning. For instance, Wang et al. [14] proposed a computational framework, TL_HGBI, based on a heterogeneous network model and applied it on drug repositioning by using existing omics data about diseases, drugs and drug targets. Martínez et al. [15] have proposed a novel network-based prioritization method to identify novel drug indications. By propagating information in the drug-disease-protein network, the proposed method identified potential drug-disease associations. These two approaches have demonstrated that integrating multi-source data could improve prediction performance in drug repositioning. However, for TL_HGBI, existing validated disease-gene information was not used in the prioritizing candidate drugs for diseases. For DrugNet, the prioritization process was performed by propagating information from drug network to disease network across target network or not, while the propagation from disease network to drug network was not utilized to assist the prediction. To address these problems, we propose a novel drug repositioning method based on random walk, RWHNDR, which integrates multi-source data effectively and exploits global network information to predict and prioritize potential drugs for diseases. Random walk models have been widely and successfully used in bioinformatics research [16], [17], [18].

In this study, target information has been fully utilized in drug repositioning. First, we constructed a heterogeneous network containing six sub networks, namely drug network, disease network, target network, drug-disease, drug-target network and target-disease network. Then, our proposed method, RWHNDR, extended the random walk model on the constructed heterogeneous network to predict candidate pharmacological treatments for diseases. Two recent network-based drug repositioning methods, which integrated targets data, were compared with RWHNDR to evaluate their prediction performances. The experiment results demonstrated that RWHNDR could obtain better prediction power not only with respect to diseases with known associated drugs but also for new diseases without any drug information. The effect of target data on predicting candidate drugs for diseases was also validated. In case studies, the top candidate drugs for four different diseases were examined, and many top ranked drugs were strongly supported by recent studies. Our research confirmed the validity of integrating drug targets and disease genes information and the effectiveness of applying the extended random walk model to drug repositioning.

TABLE 1
Statistics of the Dataset Used in This Study

	drugs	diseases	targets
	593	313	1,076
Dataset	drug-disease associations 1,933	drug-target associations 2,706	disease-target associations 483

2 Materials and Methods

In this study, we propose a novel approach, named RWHNDR, to prioritize candidate drugs for diseases. We first give brief descriptions of the used dataset and then construct a heterogeneous network by integrating multi-source data. Finally a computational prediction method based on random walk model is developed to rank potential drugs for diseases on the constructed heterogeneous network.

2.1 Dataset

The gold standard dataset is obtained from [19], which includes 1,933 validated drug-disease associations involving 593 drugs and 313 diseases. To be specific, these drugs are obtained from DrugBank database [20], which is a comprehensive biomedical database containing detailed drug data and drug targets information. Chemical structure or substructure information has been used to calculate similarity between drugs [21], [22]. In our study, the SMILES [23] data describing drug chemical structures are retrieved from DrugBank. The similarity of two drugs is measured by the Tanimoto score of their 2D chemical fingerprints calculated with the Chemical Development Kit (CDK) [24]. Diseases are collected from human phenotypes defined in the Online Mendelian Inheritance in Man (OMIM) database [25], which is a large database containing genes and disease phenotypes established by domain experts. Disease-disease similarity is calculated by MimMiner [26], which measures the degree of similarity of various diseases in terms of Mesh terms appearing in the medical description of diseases.

The associations between drugs and target proteins are retrieved from DrugBank, and 2,706 drug-target associations only involving 593 drugs in our gold standard dataset are collected. Moreover, 483 disease-gene associations involving 313 phenotype diseases in our gold standard dataset are obtained from OMIM database. Targets and genes included in drug-target and disease-gene associations are mapped into proteins in Uniprot database [27], then 1,076 protein targets with UniprotKB identifiers are collected finally. Moreover, disease-target associations are obtained through the above mapping. The target-target similarity is measured based on the amino acid sequence information retrieved from Uniprot. Rcpi [28], an R package, is used to calculate target-target similarity based on sequence alignment. The numbers of drugs, diseases, targets and all the associations used in this study are shown in Table 1.

2.2 Construction of the Heterogeneous Network

In detail, the heterogeneous network is composed of six subnetworks, namely drug network, disease network, target network, drug-target network, disease-target network and drug-disease network, respectively.

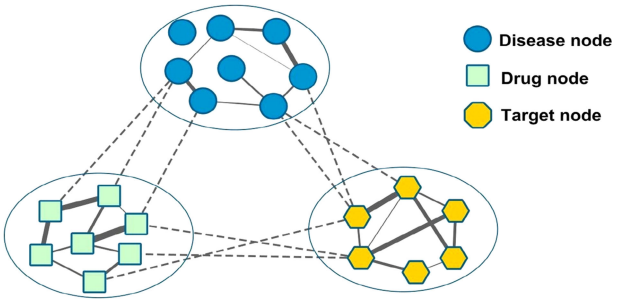


Fig. 1. A heterogeneous network consisting of drug network, disease network, target network, drug-disease network, drug-target network and target-disease network. The solid line denotes the intra-similarities, where the thickness of line represents the degree of similarity. The dashed line denotes the known disease-drug, drug-target or target-disease associations.

For drug network, let $R = \{r_1, r_2, \ldots, r_m\}$ denotes m drugs, and the weight of the edge between drugs r_i and r_j is set as the chemical structure similarity between them. For disease network, let $D = \{d_1, d_2, \ldots, d_n\}$ denotes n diseases, and the weight of the edge between diseases d_i and d_j is set as the phenotype similarity between them. For target network, let $T = \{t_1, t_2, \ldots, t_p\}$ denotes the p targets, and the weight of the edge between targets t_i and t_j is set as the sequence similarity between them.

The drug-disease network contains n diseases and m drugs, if there exists an association between disease d_i and drug r_j, the edge weight of d_i and r_j is initially assigned as 1 and otherwise 0. Similarly, the drug-target network includes m drugs and p targets, if there exists an association between drug r_i and target t_j, the edge weight of r_i and t_j is initially assigned as 1 and otherwise 0. Likewise, the target-disease network consists of p targets and n diseases, if there exists an association between target t_i and disease d_j, the edge weight of t_i and d_j is initially assigned as 1 and otherwise 0. A_{DR}, A_{RT} and A_{DT} are defined as the adjacency matrices of disease-drug network, drug-target network and disease-target network, respectively.

Finally, the heterogeneous network is constructed by connecting drug network, disease network and target network via the corresponding association networks, as shown in Fig. 1. The heterogeneous network can be represented by an adjacency matrix A as follows,

$$A = \begin{bmatrix} A_{RR} & A_{RT} & A_{RD} \\ A_{TR} & A_{TT} & A_{TD} \\ A_{DR} & A_{DT} & A_{DD} \end{bmatrix}, \tag{1}$$

where the diagonal sub-matrices A_{RR}, A_{TT}, and A_{DD}, are the corresponding adjacency matrices of drug network, target network and disease network, respectively. Besides, the off-diagonal sub-matrices A_{RD}, A_{TR} and A_{TD} represent the transpose matrices of A_{DR}, A_{RT} and A_{DT}, respectively.

2.3 Random Walk on the Heterogeneous Network

Based on the constructed heterogeneous network, RWHNDR simulates the process of the random walk on the heterogeneous network to predict candidate drugs for the query disease under consideration. RWHNDR is based on a random walk with restart (RWR) model. Generally speaking, RWR algorithm presented by [29] is described as a random walker's iterative transition from one node to its neighbor when starting from a set of given seed node. Formally, RWR can be defined as follows,

$$P_{t+1} = (1-\gamma)M^T P_t + \gamma P_0, \tag{2}$$

where parameter γ represents the restart probability, the walker can restart from seed nodes with probability γ or move on with probability $(1-\gamma)$ at each step of the random walk process. M is the transition matrix, and the element M_{ij} denotes probability that the walker transits from node i to node j. M^T is the transpose of matrix M. The initial probability vector P_0 is constructed by assigning equal probabilities to nodes in the seed node set. P_t denotes the probability vector at step t, and the ith element of P_t is the probability that the random walker is on node i at step t. After a number of iteration steps, if the difference between P_t and P_{t+1} falls below a small threshold, which is set as 10^{-10} in our application, we consider the walker reaches a steady-state P.

The process of prioritizing candidate drugs for a given disease by RWHNDR can be described as follows.

Step 1: Set the initial probability vector P_0;

RWHNDR allows restarting the walk at seed nodes with a probability γ in each step. When predicting potential drugs for a given disease i, d_i denotes as the seed node in the disease network. If drug j is associated with disease i, r_j is considered as the seed node in the drug network. Additionally, target proteins associated with disease i represent seed nodes in the target network.

According to the seed nodes in the three networks, the initial probability vector P_0 containing pr_0, pt_0 and pd_0, can be determined as follows. pr_0 represents the initial probability of drug network, equal initial scores are assigned for each drug node, and the sum is equal to 1. Similarly, the initial probability of the target network pt_0 is formed. To construct the initial probability pd_0 of the disease network, the probability of the query disease node i is assigned to be 1, and the probabilities of other disease nodes are assigned to be 0. Therefore, the initial probability of the heterogeneous network is defined as

$$P_0 = \begin{bmatrix} \lambda_R \cdot pr_0 \\ \lambda_T \cdot pt_0 \\ (1-\lambda_R-\lambda_T) \cdot pd_0 \end{bmatrix}, \tag{3}$$

where parameters λ_R, λ_T and $(1-\lambda_R-\lambda_T)$ weight the importance of drug network, target network and disease network, respectively. If the value of λ_R is larger than λ_T and $(1-\lambda_R-\lambda_T)$, the random walker has higher tendency to return to the drug seeds, indicating that the drug network is more important in disease-drug prediction.

Step 2: Construct the transition matrix M;

The random walker first starts from some seed nodes based on the initial probability, and transits from the current node randomly to its direct neighbors in the heterogeneous network or restarts from seed nodes at each step. Therefore, for each node, the transition probability to other nodes should be calculated. The transition matrix M of the heterogeneous network is defined as follows.

$$M = \begin{bmatrix} M_{RR} & M_{RT} & M_{RD} \\ M_{TR} & M_{TT} & M_{TD} \\ M_{DR} & M_{DT} & M_{DD} \end{bmatrix}. \quad (4)$$

In Equation (4), there are nine sub-matrices, which involve three intra-transition matrices and six inter-transition matrices. M_{RR} is the intra-transition matrix of drug network, which includes the probabilities from one drug to the other drugs in the random walk. Similarly, M_{TT} and M_{DD} represent the intra-transition matrices of target network and disease network, respectively. The other sub-matrices are inter-transition matrix among networks. M_{RD} is defined as the transition matrix from drug network to disease network, M_{RT} is the transition matrix from drug network to target network, M_{DR} is the transition matrix from disease network to drug network, M_{DT} is the transition matrix from disease network to target network, M_{TR} the is transition matrix from target network to drug network, and M_{TD} is the transition matrix from target network to disease network, respectively.

Along the heterogeneous network, the random walker can move within the current network or jump to the other networks with some jumping probabilities. For example, when the random walker stands at a node in the disease network, he may move on by choosing to stay in this network, or jump to the drug network or the target network with some probabilities. Therefore the jumping probability between any two different networks should be decided. We define parameter λ_{DR} as the jumping probability between disease network (D) and drug network (R), λ_{DT} as the jumping probability between disease network (D) and target network (T), and λ_{RT} as the jumping probability between drug network (R) and target network (T). If the walker stays at a disease node associated with some drug nodes and target nodes, he may jump to the drug network or the target network with probability λ_{DR} or λ_{DT}, respectively, or move to the other nodes in disease network with probability $(1 - \lambda_{DR} - \lambda_{DT})$.

Based on the above analysis, each sub-matrix in Equation (4) can be calculated based on the corresponding adjacency matrix defined in Equation (1). In Equation (4), the three intra-transition probability matrices are built based on the similarity data and the known association information of the corresponding network. For example, M_{DD} is defined as follows.

$$M_{DD}(i,j) = \begin{cases} A_{DD}(i,j)/\sum_j A_{DD}(i,j) \\ \quad if\ \sum_j A_{DR}(i,j) = 0, \sum_j A_{DT}(i,j) = 0; \\ (1-\lambda_{DR}) \cdot A_{DD}(i,j)/\sum_j A_{DD}(i,j) \\ \quad if\ \sum_j A_{DR}(i,j) \neq 0, \sum_j A_{DT}(i,j) = 0; \\ (1-\lambda_{DT}) \cdot A_{DD}(i,j)/\sum_j A_{DD}(i,j) \\ \quad if\ \sum_j A_{DR}(i,j) = 0, \sum_j A_{DT}(i,j) \neq 0; \\ (1-\lambda_{DR}-\lambda_{DT}) \cdot A_{DD}(i,j)/\sum_j A_{DD}(i,j) \\ \quad if\ \sum_j A_{DR}(i,j) \neq 0, \sum_j A_{DT}(i,j) \neq 0; \end{cases} \quad (5)$$

When the random walker stands on a node in disease network D, if this node does not have associations with nodes in drug network R and target network T, the walker can only move on within disease network D; if this node has associations with some nodes in drug network R while having no associations with nodes in target network T, the walker can move on within disease network D with probability $(1 - \lambda_{DR})$; if this node has associations with some nodes in target network T while having no associations with nodes in drug network R, the walker can move on within disease network D with probability $(1 - \lambda_{DT})$; if this node have associations with some nodes in drug network R and target network T simultaneously, the walker can move on within disease network D with probability $(1 - \lambda_{DR} - \lambda_{DT})$. The other two intra-transition probability matrices M_{RR} and M_{TT} are defined similarly.

Moreover, six inter-transition matrices in M are decided only based on the known association information. For instance, transition matrices M_{DR} is defined

$$M_{DR}(i,j) = \begin{cases} \lambda_{DR} \cdot A_{DR}(i,j)/\sum_j A_{DR}(i,j); & if\ \sum_j A_{DR}(i,j) \neq 0 \\ 0 & otherwise \end{cases} \quad (6)$$

When the random walker is staying at a node in disease network D, if this node is associated with some nodes in drug network R, then he can jump to drug network R with probability λ_{DR}. Otherwise, he can not transfer to drug network R. Similarly, the other inter-transition matrices, M_{RD}, M_{DT}, M_{TD}, M_{RT} and M_{TR}, can be defined.

Step 3: Implement the random walk on the heterogeneous network;

Based on the constructed heterogeneous network, the initial probability vector P_0 and transition matrix M are defined as in Step 1 and Step 2, respectively. The random walk model is developed for the heterogeneous network. When Equation (2) converges to the steady state, the probability vector P can be obtained, and the ith element of P is the final probability that the random walker stays at node i.

The probability vector P includes three parts: P_r, P_t and P_d. In which, P_r contains the probability scores of all drugs associated with the query disease, P_t contains the probability scores of all targets associated with the query disease, and P_d contains the probability scores of all diseases associated with the query disease. All candidate drugs in the datasets are ranked in descending order according to their steady probability scores assigned by RWHNDR in P_r. For each candidate drug, the larger the score, the more possible it is associated with the query disease. The algorithm for predicting candidate drugs of a given disease by RWHNDR is shown in Fig. 2.

3 Experiments and Results

In this section, the predictive performance of RWHNDR is comprehensively evaluated using the gold standard dataset. First, the evaluation metrics are introduced. Then, we compare RWHNDR with other network-based algorithms for prioritizing candidate drugs for the query disease. Next, the effect of integrating target information on the prediction performance is discussed. Finally, case studies are conducted to further illustrate the effectiveness of the proposed method.

3.1 Evaluation Metrics

RWHNDR can prioritize candidate drugs for a given disease in each prediction. In the gold standard dataset, each disease has about 6.18 drugs on average. Therefore, it is

Algorithm RWHNDR: Predict candidate drugs for a specific disease d;
Input: Adjacency matrices: A_{RR}, A_{DD}, A_{TT}, A_{DR}, A_{RT}, A_{DT};
Parameters: γ, λ_{DR}, λ_{DT}, λ_{RT}, λ_R, λ_T;
Output: *d_result* stores predicted scores of all candidate drugs associated with disease d;
RWHNDR (A_{RR}, A_{DD}, A_{TT}, A_{DR}, A_{RT}, A_{DT}, γ, λ_{DR}, λ_{DT}, λ_{RT}, λ_R, λ_T)
1. Construct adjacency matrix A of the heterogeneous network;
2. Calculate the number of drugs *Rnum*;
3. Construct the transition matrix M;
4. Set the initial probability vector P_0;
5. $P_t = P_0$;
6. While (P has not converged)
7. $P_{t+1} = (1-\gamma)M^T P_t + \gamma P_0$;
8. End While
9. Predicted scores of all drugs associated with disease d are stored in *M_d_result*;
M_d_result = $P_{t+1}(1: Rnum)$;
10. Predicted scores of all candidate drugs are stored in *d_result*;
11. All candidate drugs are ranked by their scores;

Fig. 2. Algorithm for predicting candidate drugs of a given disease.

appropriate to adopt leave-one-out cross-validation to systematically evaluate the performance of RWHNDR.

For all the 1,933 known drug-disease associations in the gold standard dataset, each association is taken in turn as the test set, while the remaining associations are served as the training set. When we apply the algorithm RWHNDR to the training set, the specific disease and drug involved in the test set are regarded as the test disease and the test drug, respectively. The test disease is considered as a seed node in disease network, drugs except the test drug, and targets associated with the test disease are considered as seeds in drug network and target network, respectively. The test drug and other drugs without known associations with the test disease are considered as candidate drugs. All the candidate drugs are ranked in descending order according to their predicted probability scores associating with the test disease. For a given ranking threshold, the test drug is considered as a true positive (TP) if its ranking is above the threshold; otherwise, it is considered as a false negative (FN). On the other hand, if the ranking of a drug without known association with the test disease is above the threshold, it is considered as a false positive (FP); otherwise, it is considered as a true negative (TN). By varying the ranking threshold, True Positive Rate (TPR) and False Positive Rate (FPR) can be calculated to construct an ROC curve. The area under the ROC curve (AUC) is used to measure the performance of the prediction methods [14].

Actually, the highly ranked candidate drug-disease associations in prediction are more important in computational drug repositioning. Therefore, based on various top portions, we examine the correctly identified test drug-disease associations.

3.2 Comparison with Other Methods

We compare RWHNDR with other two state-of-the-art methods: TL_HGBI [14] and DrugNet [15], to evaluate the performance of RWHNDR in identifying candidate drugs for diseases. TL_HGBI is a three-layer heterogeneous graph model which can capture the relationships among diseases, drugs and targets, based on the guilt-by-association principle and information flow-based methods. DrugNet is a network-based drug repositioning method, which implements propagation flow algorithms and can prioritize candidate drugs for diseases by integrating disease, drug and target information.

In this study, we consider two classes of the drug-disease prediction problems: one is to identify candidate drugs for known diseases, while the other is to predict candidate drugs for new diseases. A known disease is defined as a phenotype with at least one known drug associating with it, while a new disease means one without any known associated drugs. It is obvious that the prediction for known diseases has more information about the concerned disease-drug pair. In view of the above two prediction situations, our proposed approach, TL_HGBI and DrugNet are analyzed and compared.

3.2.1 Prediction for Known Diseases

In the gold standard dataset, there are 216 diseases with more than one associated drugs. These diseases involve 1,836 known disease-drug associations totally. When conducting the leave-one-out cross validation experiment, one of the known associated drugs for the test disease is removed, and there are still other drugs associated with it. Therefore, in this situation, the test disease and its known associated drugs and targets are set as seed nodes. The main purpose of this experiment is to evaluate the ability of these prioritization approaches in predicting novel drugs for known diseases with associated drugs.

For all the methods, we tune the contained parameters to achieve optimal prediction results in the cross-validations. RWHNDR achieves optimal performance when the parameter settings are $\gamma = 0.7$, $\lambda_{DR} = 0.6$, $\lambda_{DT} = 0.3$, $\lambda_{RT} = 0.3$, $\lambda_R = 0.8$, $\lambda_T = 0.1$. In which, the restart probability parameter γ is set according to that in previous studies [16]. TL_HGBI achieves optimal performance when the parameter settings are: $\alpha = 0.1$, *similarity threshold=0.5*. DrugNet achieves optimal performance when parameter α is set to 0.1. Then, these optimal parameter settings are used for methods in the following experiments. The leave-one-out cross-validation experimental results of all methods are depicted in Fig. 3. As seen from Fig. 3, RWHNDR outperforms the other methods greatly in terms of AUC values and the top-ranked results. With regard to the AUC value, RWHNDR can achieve 0.926 while TL_HGBI and DrugNet merely achieve 0.881 and 0.771, respectively. In addition, more true disease-associated drugs are identified in different top portions by RWHNDR. For a specified top-rank threshold, one true drug-disease association is considered to be correctly retrieved if the predicted ranking of this association is higher than the specified top-rank threshold. As a result, among the 1,836 true drug-disease associations, 1,079 associations are predicted as ranking in top 1 percent by RWHNDR, and only 588 and 233 associations are predicted in top 1 percent by the two competing methods, respectively. Therefore, RWHNDR is an efficient computational method for prioritizing candidate drugs for diseases.

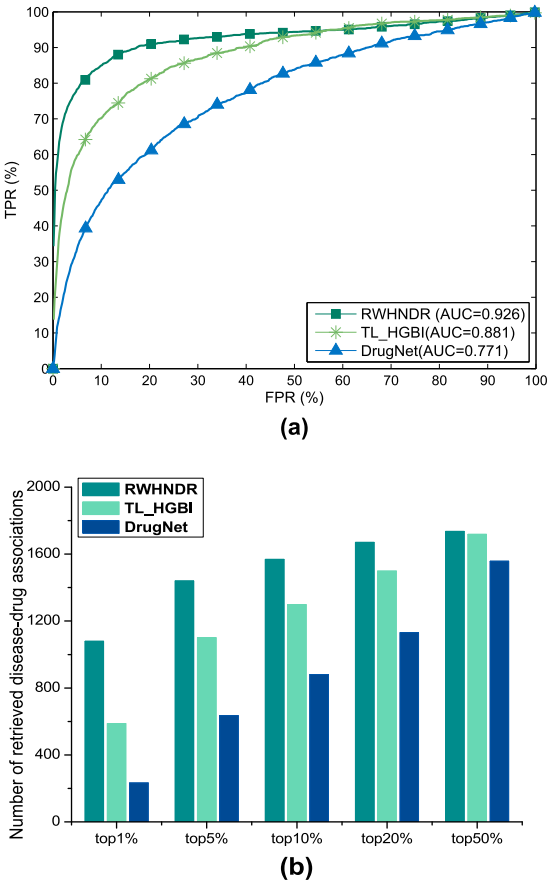


Fig. 3. The performance of different methods in identifying potential drugs for known diseases. (a) ROC curves of prediction results obtained by applying the proposed method and existing approaches. (b) The number of true disease-drug associations are retrieved correctly at different percentiles.

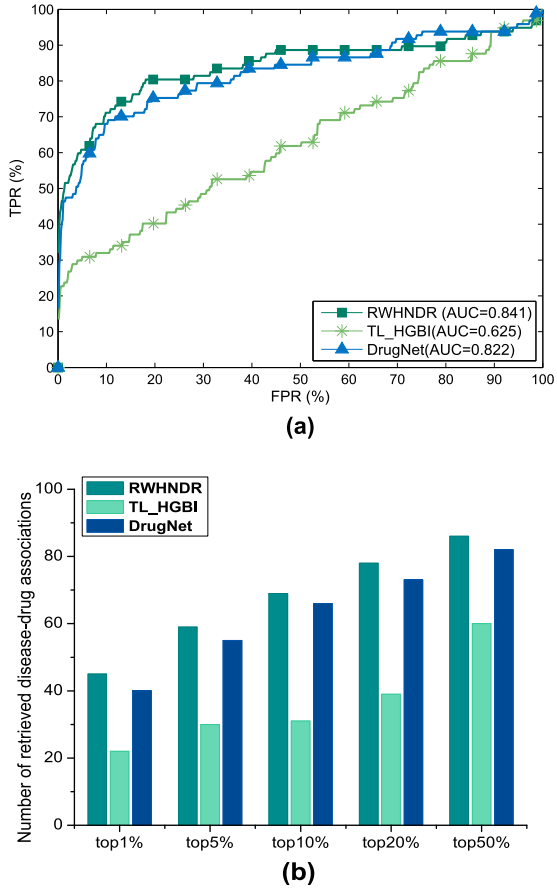


Fig. 4. The performance of different methods in identifying potential drugs for new diseases. (a) ROC curves of prediction results obtained by applying the proposed method and existing approaches. (b) The number of true disease-drug associations are retrieved correctly at different percentiles.

3.2.2 *Prediction for New Diseases*

To evaluate the prediction performance of these methods for new diseases, those diseases which have only one known association are selected from the standard dataset, and 97 diseases are obtained. In the leave-one-out cross validation, when the known disease-drug association with the specific disease is removed, the given disease thus becomes a new disease without any known association information. Therefore, the seed nodes include the test disease and its known associated targets. The experiment results of all methods in terms of ROC curves and top-ranked results of all the drug-disease associations are reported in Fig. 4.

It is shown that RWHNDR (AUC: 0.841) outperforms the other methods, TL_HGBI (AUC: 0.625) and DrugNet (AUC: 0.822) in predicting candidate drugs for new diseases. What's more, there are improvements in prediction performance, especially for the top-ranked results. For example, among the 97 true disease-drug associations, 45 of them are predicted in top 1 percent based on RWHNDR while only 22 and 40 true associations are predicted in top 1 percent by TL_HGBI and DrugNet, respectively.

3.3 Assessing the Impact of Integrating Target Information

In this study, to further evaluate the effect of integrating target information on the prediction performance for known diseases and new diseases, we propose a method named DR_RWRH, to perform prediction without using target information. DR_RWRH implements RWRH [16] on a heterogeneous network involving drug network, disease network and drug-disease network. The difference between RWHNDR and DR_RWRH is the utilization of target information in RWHNDR, which is not used in DR_RWRH. For DR_RWRH, a query disease is considered as seed node in disease network, and drugs associating with the given disease are considered as seed nodes in drug network. Then the random walker starts from these seed nodes, and walks on the heterogeneous network until reaching a steady state. The candidate drug with the largest probability is the most possible drug associated with the query disease.

RWHNDR performs random walks on a heterogeneous network with target information, therefore, it is compared with DR_RWRH to demonstrate the effectiveness of integrating targets on predictions. The parameters used in DR_RWRH are also investigated by conducting cross validations. DR_RWRH achieves the optimal performance when the restart probability is set to 0.7, the jumping probability between drug network and disease network is set to 0.7, and the parameter weighting the importance of drug network is set to 0.9. The parameter settings are used for DR_RWRH in the following predictions. We conduct a comparison of these two approaches for identifying candidate drugs for known diseases and new diseases by applying the leave-one-out cross validation experiments. The results

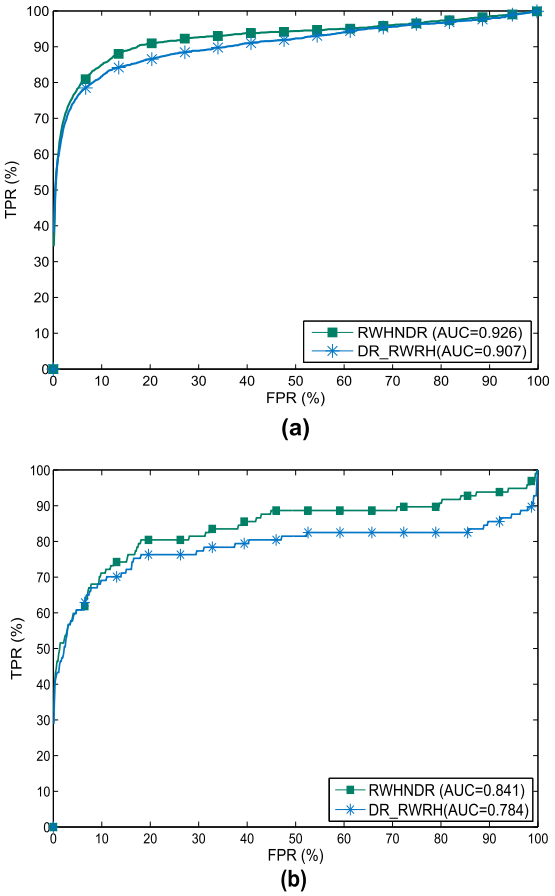


Fig. 5. Comparison results for integrating different information. (a) ROC curves for DR_RWRH and RWHNDR in predicting candidate drugs for known diseases. (b) ROC curves for DR_RWRH and RWHNDR in predicting candidate drugs for new diseases.

of all the known drug-disease associations in terms of AUC values are depicted in Fig. 5.

As seen from the Fig. 5, when prioritizing candidate drugs for known diseases with known associated drugs, RWHNDR achieves an AUC value of 0.926, which is slightly better than that obtained by DR_RWRH (0.907). While predicting candidate drugs for new diseases without any associated drug information, RWHNDR achieves an AUC value of 0.841, which is much higher than that of DR_RWRH (0.784). For new diseases without known drugs, the drug target associations, disease gene associations and similarities of targets can be comprehensively utilized to recommend potential drugs for these new diseases. Meanwhile, the comparison results also indicate that target information is more useful in predicting potential drugs for new diseases.

Based on the above results, we can find the effectiveness of RWHNDR by integrating more useful biological information. Moreover, recent researches have collected and identified more biological data, which can further provide supports to the drug repositioning.

3.4 Case Studies

After confirming the effectiveness of RWHNDR in prioritizing potential drugs for diseases by cross validations, we predict novel drug-disease associations based on all known drug and disease information. All known drug-disease associations in the gold standard dataset are used as training data while those unidentified associations are considered as candidate pairs. We focused on the most possible drugs of each disease predicted by RWHNDR. The validity of the predicted drug-disease pairs is investigated based on the literatures. We conduct case studies on several neurological disorders and common cancers, including Huntington disease (HD), Parkinson disease (PD), breast cancer and lung cancer. When performing prediction for each disease, all the candidate drugs are ranked based on the prediction scores, then the top 5 candidate drugs are selected to examine their relationship with the specific disease.

TABLE 2
The Top 5 Candidate Drugs Indicated for Huntington Disease (HD), Parkinson Disease (PD), Breast Cancer, and Lung Cancer

Disease (OMIM IDs)	Top 5 candidate drugs (DrugBank IDs)	References
Huntington disease (143,100)	Carbamazepine (DB00564)	[30], [31]
	Dantrolene (DB01219)	[32]
	Vigabatrin (DB01080)	
	LORAZEPAM (DB00186)	
	Tizanidine (DB00697)	
Parkinson disease (168,600)	Paclitaxel (DB01229)	
	Docetaxel (DB01248)	
	Biperiden (DB00810)	
	Rivastigmine (DB00989)	[33], [34]
	Levodopa (DB01235)	[35]
Breast cancer (114,480)	Caffeine (DB00201)	
	Ethinyl Estradiol (DB00977)	[36]
	Aspirin (DB00945)	[37], [38]
	Arsenic trioxide (DB01169)	[39], [40]
	Estramustine (DB01196)	
Lung cancer (211,980)	Vincristine (DB00541)	
	Methotrexate (DB00563)	
	Sorafenib (DB00398)	[41]
	Cisplatin (DB00515)	[42]
	Daunorubicin (DB00694)	

We search literatures and find that some top candidate drugs are supported by current researches, which are not found in the gold standard dataset used in this study. For example, HD is an inherited disease of the central nervous system [30]. By performing prediction for HD, two drugs out of the top 5 ranked drugs are well documented as being associated with HD as shown in Table 2. The top-ranked drug is Carbamazepine, which originally is indicated for the treatment of epilepsy and pain associated with true trigeminal neuralgia [20]. Carbamazepine has been demonstrated to be effective in precipitate micturitions in HD with or without nocturnal incontinence [31]. Moreover, in symptomatic treatment of HD, Carbamazepine is found to be effective for depression, paranoia and psychosis in HD [30]. The second-ranked drug is Dantrolene, and it has been demonstrated that dantrolene could be considered as candidates for the treatment of HD and other polyQ-expansion disorders [32].

PD is a chronic and progressive central nervous system that mainly affects the motor function. RWHNDR is performed to rank candidate drugs for PD, and two of the top 5 ranked candidate drugs have been retrieved in related researches. In which, the fourth-ranked drug is Rivastigmine, which is a cholinesterase inhibitor that inhibits both butyrylcholinesterase and acetylcholinesterase. It has been

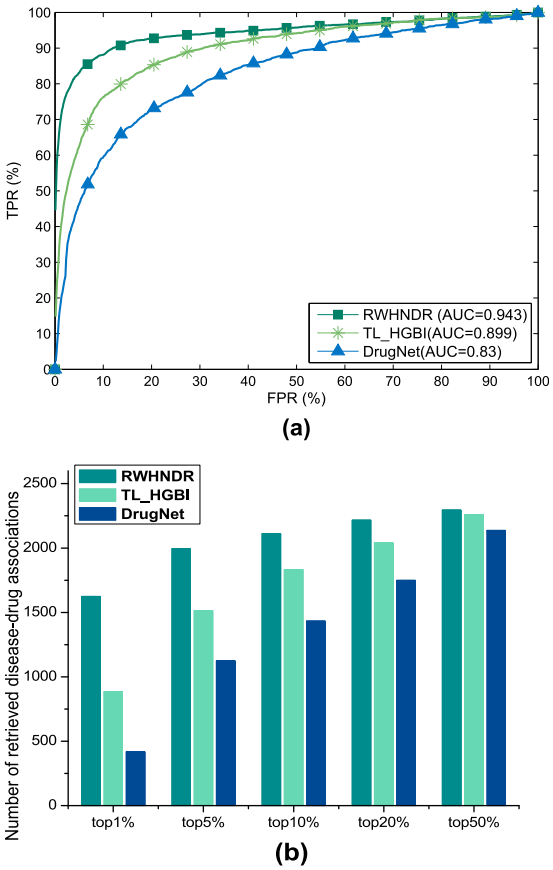


Fig. 6. The performance of different methods in identifying potential drugs for known diseases in new dataset. (a) ROC curves of prediction results obtained by applying the proposed method and existing approaches. (b) The number of true disease-drug associations are retrieved correctly at different percentiles.

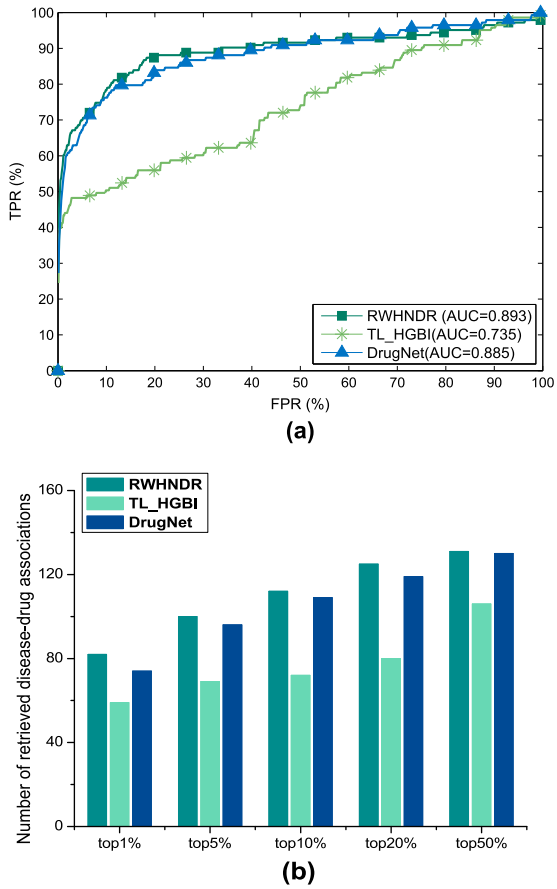


Fig. 7. The performance of different methods in identifying potential drugs for new diseases in new dataset. (a) ROC curves of prediction results obtained by applying the proposed method and existing approaches. (b) The number of true disease-drug associations are retrieved correctly at different percentiles.

reported to produce significant improvements in global ratings of dementia, cognition, and behavioral symptoms among patients with dementia associated with PD [33]. Moreover, Rivastigmine has been shown to improve gait stability and reduce falls in patients with PD [34]. The fifth-ranked drug is Levodopa, which is used to replace dopamine lost in Parkinson's disease [20]. Levodopa is the most effective symptomatic therapy [35].

Breast cancer is one common cancer that develops from breast tissue. By performing RWHNDR for breast cancer, the list of ranked candidate drugs are supported by literatures as reported in Table 2. The second-ranked drug is Ethinyl Estradiol (EE2). Iwase et al. [36] have studied the efficacy of EE2, and concluded that EE2 is beneficial for postmenopausal patients with heavily pre-treated metastatic breast cancer with endocrine therapies. The third-ranked drug is aspirin, studies have suggested that aspirin may inhibit breast cancer metastasis, and found that the use of aspirin was associated with a decrease risk of distant recurrence and breast cancer death [37]. In addition, recent studies suggest that aspirin use may reduce both all-cause and breast cancer-specific mortality [38]. Arsenic trioxide has been investigated in present studies, and results indicated that it has potential as candidate for treatment of breast cancer [39], [40].

Lung cancer is a malignant lung tumor characterized by uncontrolled growth of cells in tissues of the lung. In the top ranked predicted drugs indicated for lung cancer, the third-ranked drug is Sorafenib. In current research, Sorafenib is approved for the treatment of renal cell carcinoma and hepatocellular carcinoma, and has been studied for non-small cell lung cancer [41]. Moreover, Cisplatin-based chemotherapy can improve survival for patients with completely resected non-small-cell lung cancer [42].

As discussed above, some top-ranked candidate drugs are verified by current researches, and the other drugs without validation deserves further investigation to detect their effectiveness. Therefore, our proposed approach, RWHNDR, should have the ability to predict novel drugs for diseases in practice.

3.5 Experimental Evaluation on Other Dataset

To further assess the efficiency of the proposed method on the predicting potential drugs for diseases, the evaluation experiments are conducted on a different dataset. For this new dataset, diseases, drugs and disease-drug associations are obtained from our previous study [12]; drug-target and disease-gene associations are derived from DrugBank and OMIM database, respectively. To sum up, there are 409 diseases, 663 drugs, 1,177 targets, 2,532 disease-drug associations, 567 disease-gene associations and 2,944 drug-target associations. Moreover, there are 266 known diseases which have at least one associated drugs, and 143 new diseases without known associated drugs.

The evaluation results of identifying potential treatments for known diseases are shown in Fig. 6. RWHNDR achieves a much better AUC value of 0.943 in comparison with the other methods. For the 2,389 identified disease-drug associations involved by these 266 known diseases, RWHNDR successfully predicts 1,624 associations within top 1 percent.

In addition, the results in Fig. 7 show that our method is consistently effective in predicting potential drugs for new diseases, and outperforms the other methods. Specifically, our method achieves an AUC value of 0.893. In terms of the top-ranked predicted results, RWHNDR ranks 82 of the 143 identified disease-drug associations within top 1 percent, which is higher than results obtained by the other methods. In conclusion, the above experiment results based on the new dataset further demonstrate the robustness and predictive ability of the proposed method.

4 Conclusion

This study has proposed a powerful network-based computational method called RWHNDR for predicting novel disease-drug associations. Based on the heterogeneous network constructed from multiple data sources, prioritizing candidate drugs is carried out by implementing the random walk with restart algorithm on the heterogeneous network. We evaluate the predictive performance of the proposed method by conducting the leave-one-out cross validation test. The experiment results have shown that simultaneous integration of information about drugs, drug targets, diseases and disease genes can improve the prediction performance in drug repositioning. Moreover, RWHNDR can effectively predict candidate drugs for new diseases without any associated drugs. Case studies demonstrate the reliability and effectiveness of this method in revealing novel disease-drug associations. RWHNDR achieves better prediction performance when compared with previous methods, which can be attributed to the joint power of several aspects. Multi-source data are used to construct the heterogeneous network, and a random walk method is developed to capture the global multi-source information in the constructed heterogeneous network.

The proposed method RWHNDR performs identifying candidate drugs by walking on the constructed heterogeneous network, and thus the efficacy of RWHNDR is affected by the quality of the heterogeneous network. For future studies, more useful bioinformatics data should be integrated to improve the quality of the heterogeneous network. For example, drug side effects and drug ATC codes data can be utilized to improve the drug similarity. Disease-related gene and protein-protein interaction data can be integrated to further enhance the disease similarity [43]. Additionally, the prediction performance of RWHNDR can be further improved by collecting and incorporating more validated association data. With more disease and drug related studies, we should be able to obtain comprehensive biological data to improve the prediction ability of RWHNDR in further work.

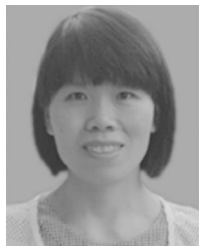
Acknowledgments

This work is supported in part by the National Natural Science Foundation of China (61622213, 61420106009, 61772552, 61732009) and the 111 project (B18059).

References

[1] B. Munos, "Lessons from 60 years of pharmaceutical innovation," *Nature Rev. Drug Discovery*, vol. 8, no. 12, pp. 959–968, 2009.
[2] C. R. Chong and D. J. Sullivan, "New uses for old drugs," *Nature*, vol. 448, no. 7154, pp. 645–646, 2007.
[3] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski, "The price of innovation: New estimates of drug development costs," *J. Health Economics*, vol. 22, no. 2, pp. 151–185, 2003.
[4] A. L. Hopkins, "Network pharmacology: The next paradigm in drug discovery," *Nature Chemical Biol.*, vol. 4, no. 11, pp. 682–690, 2008.
[5] T. T. Ashburn and K. B. Thor, "Drug repositioning: Identifying and developing new uses for existing drugs," *Nature Rev. Drug Discovery*, vol. 3, no. 8, pp. 673–683, 2004.
[6] J. Gilbert, P. Henske, and A. Singh, "Rebuilding big pharma's business model," *In Vivo*, vol. 21, no. 10, pp. 73–80, 2003.
[7] J. S. Shim and J. O. Liu, "Recent advances in drug repositioning for the discovery of new anticancer drugs," *Int. J. Biol. Sci.*, vol. 10, no. 7, pp. 654–463, 2014.
[8] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning," *Briefings Bioinf.*, vol. 17, no. 1, pp. 2–12, 2016.
[9] A. P. Chiang and A. J. Butte, "Systematic evaluation of drug-disease relationships to identify leads for novel drug uses," *Clinical Pharmacology Therapeutics*, vol. 86, no. 5, 2009, Art. no. 507.
[10] C. Wu, R. C. Gudivada, B. J. Aronow, and A. G. Jegga, "Computational drug repositioning through heterogeneous network clustering," *BMC Syst. Biol.*, vol. 7, no. 5, 2013, Art. no. S6.
[11] H. Chen, H. Zhang, Z. Zhang, Y. Cao, and W. Tang, "Network-based inference methods for drug repositioning," *Comput. Math. Methods Med.*, vol. 2015, pp. 1–7, 2015.
[12] H. Luo, J. Wang, M. Li, J. Luo, X. Peng, F. X. Wu, and Y. Pan, "Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm," *Bioinf.*, vol. 32, no. 17, pp. 2664–2671, 2016.
[13] H. Luo, M. Li, S. Wang, Q. Liu, Y. Li, and J. Wang, "Computational drug repositioning using low-rank matrix approximation and randomized algorithms," *Bioinf.*, to be published, doi: 10.1093/bioinformatics/bty013.
[14] W. Wang, S. Yang, X. Zhang, and J. Li, "Drug repositioning by integrating target information through a heterogeneous network model," *Bioinf.*, vol. 30, no. 20, pp. 2923–2930, 2014.
[15] V. Martínez, C. Navarro, C. Cano, W. Fajardo, and A. Blanco, "DrugNet: Network-based drug-disease prioritization by integrating heterogeneous data," *Artif. Intell. Med.*, vol. 63, no. 1, pp. 41–49, 2015.
[16] Y. Li and J. C. Patra, "Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network," *Bioinf.*, vol. 26, no. 9, pp. 1219–1224, 2010.
[17] X. Chen, M. X. Liu, and G. Y. Yan, "Drug-target interaction prediction by random walk on the heterogeneous network," *Mol. BioSystems*, vol. 8, no. 7, pp. 1970–1978, 2012.
[18] W. Peng, W. Lan, J. Zhong, J. Wang, and Y. Pan, "A novel method of predicting microRNA-disease associations based on microRNA, disease, gene and environment factor networks," *Methods*, vol. 124, pp. 69–77, 2017.
[19] A. Gottlieb, G. Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: A method for inferring novel drug indications with application to personalized medicine," *Mol. Syst. Biol.*, vol. 7, no. 1, 2011, Art. no. 496.
[20] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "DrugBank: A knowledgebase for drugs, drug actions and drug targets," *Nucleic Acids Res.*, vol. 36, no. 1, pp. D901–D906, 2008.
[21] W. Lan, J. Wang, M. Li, J. Liu, Y. Li, F. X. Wu, and Y. Pan, "Predicting drug-target interaction using positive-unlabeled learning," *Neurocomputing*, vol. 206, pp. 50–57, 2016.
[22] C. Yan, J. Wang, W. Lan, FX. Wu, and Y. Pan, "SDTRLS: Predicting drug-target interactions for complex diseases based on chemical substructures," *Complexity*, vol. 2017, pp. 1–10, 2017.
[23] D. Weininger, "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules," *J. Chemical Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988.
[24] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen, "The Chemistry Development Kit (CDK): An open-source Java library for chemo-and bioinformatics," *J. Chemical Inf. Comput. Sci.*, vol. 43, no. 2, pp. 493–500, 2003.

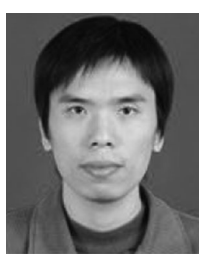
[25] A. Hamosh, A. F. Scott, J. Amberger, C. Bocchini, D. Valle, and V. A. McKusick, "Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 52–55, 2002.
[26] M. A. Van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, and J. A. Leunissen, "A text-mining analysis of the human phenome," *Eur. J. Human Genetics*, vol. 14, no. 5, pp. 535–542, 2006.
[27] UniProt Consortium, "UniProt: A hub for protein information," *Nucleic Acids Res.*, vol. 43, no. D1, pp. D204–D212, 2015.
[28] D. S. Cao, N. Xiao, Q. S. Xu, and A. F. Chen, "Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds, and their interactions," *Bioinf.*, vol. 31, no. 2, pp. 279–281, 2015.
[29] S. Köhler, S. Bauer, D. Horn, and P. N. Robinson, "Walking the interactome for prioritization of candidate disease genes," *Amer. J. Human Genetics*, vol. 82, no. 4, pp. 949–958, 2008.
[30] S. N. Tyagi, L. K. Tyagi, R. Shekhar, M. Singh, and M. Kori, "Symptomatic treatment and management of Huntington's disease: An overview," *Global J. Pharmacology*, vol. 4, no. 1, pp. 06–12, 2010.
[31] V. Cochen, J. D. Degos, and A. C. Bachoud-Lvi, "Efficiency of carbamazepine in the treatment of micturitional disturbances in Huntington disease," *Neurology*, vol. 55, no. 12, pp. 1934–1934, 2000.
[32] X. Chen, J. Wu, S. Lvovskaya, E. Herndon, C. Supnet, and I. Bezprozvanny, "Dantrolene is neuroprotective in Huntington's disease transgenic mouse model," *Mol. Neurodegeneration*, vol. 6, no. 1, 2011, Art. no. 81.
[33] M. Emre, D. Aarsland, A. Albanese, E. J. Byrne, G. Deuschl, P. P. De Deyn, F. Durif, J. Kulisevsky, T. Laar, A. Lees, W. Poewe, A. Robillard, M. M. Rosa, E. Wolters, P. Quarg, S. Tekin, and R. Lane, "Rivastigmine for dementia associated with Parkinson's disease," *New England J. Med.*, vol. 351, no. 24, pp. 2509–2518, 2004.
[34] E. J. Henderson, S. R. Lord, M. A. Brodie, D. M. Gaunt, A. D. Lawrence, J. C. T. Close, A. L. Whone, and Y. Ben-Shlomo, "Rivastigmine for gait stability in patients with parkinson's disease (respond): A randomised, double-blind, placebo-controlled, phase 2 trial," *Lancet Neurology*, vol. 15, no. 3, pp. 249–258,2016.
[35] R. Katzenschlager and A. J. Lees, "Treatment of Parkinson's disease: Levodopa as the first choice," *J. Neurology*, vol. 249, pp. II19–II24, 2002.
[36] H. Iwase, Y. Yamamoto, M. Yamamoto-Ibusuki, K. I. Murakami, Y. Okumura, S. Tomita, T. Inao, Y. Honda, Y. Omoto, and K. I. Iyama, "Ethinylestradiol is beneficial for postmenopausal patients with heavily pre-treated metastatic breast cancer after prior aromatase inhibitor treatment: A prospective study," *Brit. J. Cancer*, vol. 109, no. 6, pp. 1537–1542, 2013.
[37] M. D. Holmes, W. Y. Chen, L. Li, E. Hertzmark, D. Spiegelman, and S. E. Hankinson, "Aspirin intake and survival after breast cancer," *J. Clinical Oncology*, vol. 28, no. 9, pp. 1467–1472, 2010.
[38] D. M. Fraser, F. M. Sullivan, A. M. Thompson, and C. McCowan, "Aspirin use and survival after the diagnosis of breast cancer: A population-based cohort study," *Brit. J. Cancer*, vol. 111, no. 3, pp. 623–627, 2014.
[39] S. K. Chow, J. Y. Chan, and K. P. Fung, "Inhibition of cell proliferation and the action mechanisms of arsenic trioxide (As2O3) on human breast cancer cells," *J. Cellular Biochemistry*, vol. 93, no. 1, pp. 173–187, 2004.
[40] X. Li, X. Ding, and T. E. Adrian, "Arsenic trioxide causes redistribution of cell cycle, caspase activation, and GADD expression in human colonic, breast, and pancreatic cancer cells," *Cancer Investigation*, vol. 22, no. 3, pp. 389–400, 2004.
[41] J. Zhang, K. A. Gold, and E. Kim, "Sorafenib in non-small cell lung cancer," *Expert Opinion Investigational Drugs*, vol. 21, no. 9, pp. 1417–1426, 2012.
[42] J. P. Pignon, "Cisplatin-based adjuvant chemotherapy in patients with completely resected non-small-cell lung cancer," *New England J. Med.*, vol. 350, no. 4, pp. 351–360, 2004.
[43] P. Ni, J. Wang, P. Zhong, Y. Li, F. X. Wu, and Y. Pan, "Constructing disease similarity networks based on disease module theory," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/TCBB.2018.2817624.



Huimin Luo is working toward the PhD degree in the School of Information Science and Engineering, Central South University, Changsha, Hunan, P.R. China. Her current research interests include bioinformatics and systems biology.

Jianxin Wang received the BEng and MEng degrees in computer engineering from Central South University, China, in 1992 and 1996, respectively, and the PhD degree in computer science from Central South University, China, in 2001. He is the vice dean and a professor with the School of Information Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics, and computer network. He is a senior member of the IEEE.

Min Li received the PhD degree in computer science from Central South University, China, in 2008. She is currently a professor with the School of Information Science and Engineering, Central South University, Changsha, Hunan, P.R. China. Her main research interests include bioinformatics and systems biology.



Junwei Luo is working toward the PhD degree in the School of Information Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His main research interests include genome assembly and sequence analysis.

Peng Ni is working toward the master's degree in the School of Information Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His main research interests include bioinformatics and data mining.

Kaijie Zhao is working toward the master's degree in the School of Information Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His main research interests include bioinformatics and data mining.

Fang-Xiang Wu (M'06-SM'11) received the BSc and MSc degrees in applied mathematics from the Dalian University of Technology, Dalian, China, in 1990 and 1993, respectively. He received the first PhD degree in control theory and its applications from Northwestern Polytechnical University, Xian, China, in 1998, and the second PhD degree in biomedical engineering from the University of Saskatchewan (U of S), Saskatoon, Canada, in 2004. During 2004-2005, he worked as a postdoctoral fellow with the Laval University Medical Research Center, Quebec City, Canada. He is currently a professor with the Division of Biomedical Engineering and the Department of Mechanical Engineering, University of Saskatchewan. His current research interests include computational and systems biology, genomic and proteomic data analysis, biological system identification and parameter estimation, and applications of control theory to biological systems. He has published more than 260 technical papers in refereed journals and conference proceedings. He is serving as an editorial board member of three international journals, the guest editor of several international journals, and as the program committee chair or member of several international conferences. He has also reviewed papers for many international journals. He is a senior member of the IEEE.

Yi Pan received the BEng and MEng degrees in computer engineering from Tsinghua University, China, in 1982 and 1984, respectively, and the PhD degree in computer science from the University of Pittsburgh, Pittsburgh, Pennsylvania, in 1991. He is a regents professor of computer science and an interim associate dean and chair of biology with Georgia State University, Atlanta, Georgia. He joined Georgia State University, in 2000 and was promoted to full professor, in 2004, named a Distinguished University Professor, in 2013, and designated a regents' professor (the highest recognition given to a faculty member by the University System of Georgia), in 2015. He served as the chair of the Computer Science Department from 2005-2013. He is also a visiting Changjiang Chair Professor with Central South University, China. His profile has been featured as a distinguished alumnus in both the *Tsinghua Alumni Newsletter* and the *University of Pittsburgh CS Alumni Newsletter*. His research interests include parallel and cloud computing, wireless networks, and bioinformatics. He has published more than 330 papers including more than 180 SCI journal papers and 60 IEEE/ACM Transactions papers. In addition, he has edited/authored 40 books. His work has been cited more than 6,500 times. He has served as an editor-in-chief or editorial board member of 15 journals including seven IEEE Transactions. He received many awards including an IEEE Transactions Best Paper Award, four other international conference or journal Best Paper Awards, four IBM Faculty Awards, two JSPS senior invitation fellowships, IEEE BIBE Outstanding Achievement Award,NSF Research Opportunity Award, and AFOSR summer faculty research fellowship. He has organized many international conferences and delivered keynote speeches at more than 50 international conferences around the world.