Predicting Drug-Disease Associations Based on Network Consistency Projection

Qiang Zhang[1,2], Zonglan Zuo[1], Rui Yan[2], Chunhou Zheng[1(✉)], and Fa Zhang[2(✉)]

[1] College of Computer Science and Technology, Anhui University, Hefei, China
[2] High Performance Computer Research Center, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
zhangfa@ict.ac.cn

Abstract. With the increasing cost of traditional drug discovery, drug repositioning methods at low cost have attracting increasing attention. The generation of large amounts of biomedical data also provides unprecedented opportunities for drug repositioning research. However, how to effectively integrate different types of data is still a challenge for drug repositioning. In this paper, we propose a computational method using **N**etwork **C**onsistency **P**rojection for **D**rug-**D**isease **A**ssociation (NCPDDA) prediction. First of all, our method proposes a new method for calculating one type of disease similarity. Moreover, since effective integration of data from multiple sources can improve prediction performance, the NCPDDA integrates multiple kinds of similarities. Then, considering that noise may affect the prediction performance of the model, the NCPDDA uses the similarity network fusion method to reduce the impact of noise. Finally, the network consistency projection is used to predict potential drug-disease associations. NCPDDA is compared with several classical drug repositioning methods, and the experimental results show that NCPDDA is superior to these methods. Moreover, the study of several representative drugs proves the practicality of NCPDDA in practical application.

Keywords: Drug repositioning · Drug-disease association · Network consistency projection · Similarity network fusion

1 Introduction

In the traditional drug development, the successful development of a completely new drug is behind the high investment of more than 800 million dollars and more than a decade of continuous efforts of researchers [1]. Even so, usually only one in ten drugs are approved by the FDA (Food and Drug Administration) for actual treatment each year [2]. Given these challenges, drug repositioning, which is the discovery of new indications for existing drugs, has attracted increasing attention. Compared with traditional drug development strategies, the computational drug repositioning methods have advantages in terms of time, cost, and risk reduction because the repositioned drug has already passed some preclinical trials. There are many typical examples of drug repositioning:

D.-S. Huang et al. (Eds.): ICIC 2021, LNAI 12838, pp. 591–602, 2021.
https://doi.org/10.1007/978-3-030-84532-2_53

for example, the sildenafil, developed for individuals with heart diseases, repositioned for erectile dysfunction [3]. Besides, the aspirin, developed for mild or moderate pain and repositioned for inhibit thrombosis. These successful drug repositioning cases have further promoted the development of drug repositioning.

Recently, many methods for computational drug repositioning have been proposed, including machine learning-based methods and network-based methods. By integrating drug similarities and disease similarities, Gottlieb et al. proposed a machine learning model that uses logistic regression classifiers to infer drug repositioning [4]. Liang et al. integrated multiple types of similarities and proposed Laplacian regularized sparse subspace learning (LRSSL) method to predict new indications for drugs [5]. Based on the drug similarity and disease similarity, Yang et al. proposed a bounded nuclear norm regularization (BNNR) method that is robust to the noise in the data and makes the prediction scores of all drug-disease associations within the range of (0, 1) [6]. Adding target information into the BNNR model leads to a significant increase in the calculation cost of BNNR. Yang et al. further proposed a method based on overlap matrix completion [7]. Besides, based on known drug-disease associations, disease similarity and five types of drug similarities, Zhang et al. proposed a similarity constrained matrix factorization method to predict potential drug-disease associations [8]. Moreover, with the advent of the era of big data and the progress of science and technology, a large number of biological and clinical data that can be processed by computer technology have been produced. Because the network has ability to integrate data from multiple sources, network-based approaches are widely used in drug repositioning. Wang et al. proposed a three-layer heterogeneous network model using an iterative algorithm to achieve drug repositioning [9]. Martínez et al. proposed a network-based prioritization approach to infer new relationships between drugs and diseases [10]. Luo et al. proposed a drug repositioning method based on comprehensive similarity measures and Bi-Random walk algorithm [11]. Even though the above methods have accelerated the speed of drug repositioning, there are still some limitations. Some approaches are limited to a single type of data, such as considering only phenotype similarity of diseases and chemical structure similarity of drugs. Since different types of data reflect different aspects of drugs or diseases, the mechanism of action of drugs or diseases can be understood more clearly by integrating multi-source data. In addition, drug or disease data may contain noise, which can reduce the accuracy of prediction.

In this work, we propose a method using network consistency projection for drug-disease association (NCPDDA) prediction. Firstly, a method for calculating disease similarity based on drug characteristics is proposed. Secondly, the NCPDDA uses the similarity network fusion method to integrate four kinds of drug similarities and three kinds of disease similarities respectively, because the similarity network fusion method can effectively fuse different types of data and reduce noise. Finally, based on drug-disease association network and the integrated drug similarity network and disease similarity network, the network consistency projection is used to obtain the prediction scores of all drug-disease associations. The overall framework of the NCPDDA method is depicted in Fig. 1. The experimental results show that our method has better prediction performance than other three classical methods. In the case studies section, the new indications for

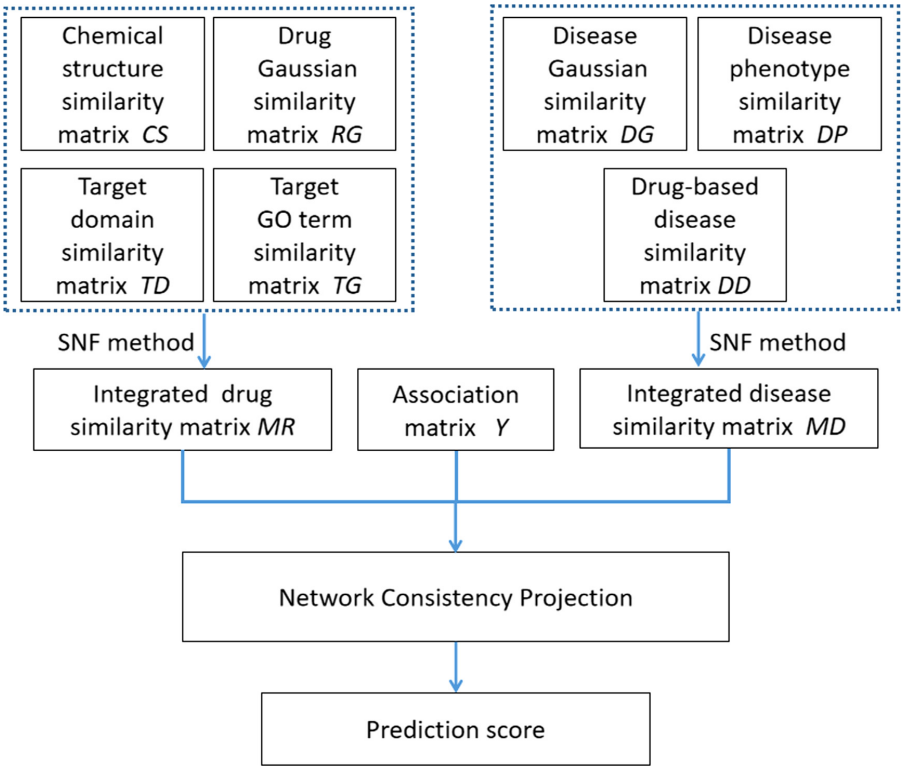


Fig. 1. The overall framework of the NCPDDA method.

four drugs are validated, and the experimental results further confirms the practicality of our method.

2 Materials and Methods

2.1 Dataset

Drug-disease Associations. The LRSSL dataset including 763 drugs, 681 diseases and 3,051 known drug-disease associations is used by our method. A binary matrix $Y \in R^{m \times n}$ is used to represent drug-disease associations, where m and n are the number of drugs and diseases, respectively. If the drug r_i is associated with the disease d_j, then $Y(i, j) = 1$, otherwise it is 0.

Similarities of Drugs. The drug data includes drug chemical structure, drug target proteins domain and gene ontology (GO) term of drug targets. Based on the chemical fingerprints extracted from PubChem database [12], the Jaccard coefficient is used to calculate similarity, and the drug chemical structure similarity matrix $CS \in R^{m \times m}$ is obtained. Similarly, based on the drug target domain and drug target gene ontology term extracted from InterPro database [13] and UniProt database [14], the drug target domain similarity matrix *TD* and the drug target GO term similarity matrix *TG* are calculated.

Similarity of Diseases. MimMiner [15] is used to calculate the similarity of diseases, and the disease phenotype similarity matrix $DP \in R^{n\times n}$ is obtained. MimMiner quantify the similarities between diseases according to the Medical Subject Headings (MeSH) Vocabulary terms that appeared in the medical descriptors of diseases in the OMIM (Online Mendelian Inheritance in Man) database.

2.2 Disease Similarity Based on Drug Characteristics

Since similar diseases are often associated with similar drugs, we calculate a type of disease similarity based on three types of drug similarities. According to [16 and 17], we firstly obtain the drug sets r_a and r_b, which are related to the disease d_a and d_b, respectively. Then, the similarity between disease d_a and d_b can be computed as follows:

$$DD(d_a, d_b) = \frac{\sum_{i=1}^{u} \max\limits_{1\le j\le v}\left(RS\left(r_{ai}, r_{bj}\right)\right) + \sum_{j=1}^{v} \max\limits_{1\le i\le u}\left(RS\left(r_{bj}, r_{ai}\right)\right)}{u+v} \tag{1}$$

Where $RS(r_{ai}, r_{bj})$ is the drug similarity of r_{ai} and r_{bj} belonging to r_a and r_b, respectively; u and v are the number of drugs included by r_a and r_b. We use the similarity network fusion (SNF) method [18] to integrate the above three drug similarity matrices into the drug similarity matrix RS to calculate the disease similarity matrix DD, $DD \in R^{n\times n}$. The flow chart for calculating disease similarity based on drug characteristics is shown in Fig. 2. It is important to note that in the cross-validation, some diseases may not have associated drugs, making it impossible to calculate the disease similarity matrix. Therefore, we initialize $Y' = Y$, and update the matrix Y' as follows:

$$Y'(i) = \frac{1}{\sum_{d_j\in N_{d_i}} DP(i,j)} \sum_{d_j\in N_{d_i}} DP(i,j)Y(j) \tag{2}$$

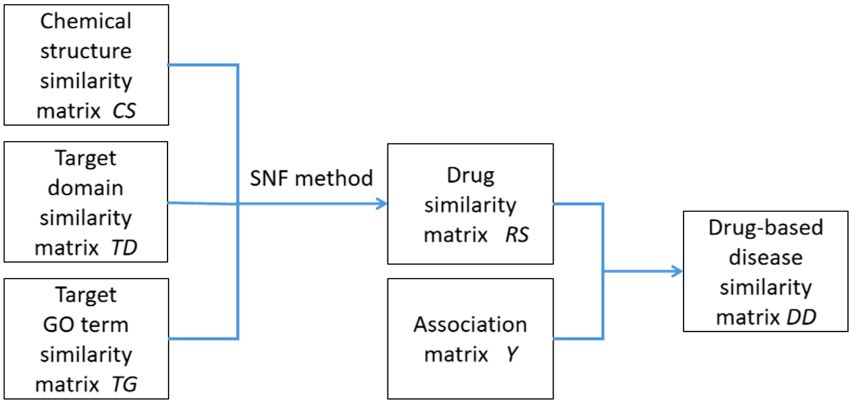


Fig. 2. The flow chart for calculating disease similarity based on drug characteristics.

Where i represents the index of disease d_i, which has no associated drug in Y'; N_{d_i} denotes the set of k neighbor nodes of d_i; $Y'(i)$ and $Y(j)$ represent the i-th column of Y' and the j-th column of Y, respectively; $DP(i, j)$ denotes the disease phenotype similarity between disease d_i and disease d_j. Then, the similarity matrix DD is recalculated based on the updated matrix Y'.

2.3 Gaussian Similarity of Drugs and Diseases

Since the more the number of common diseases (or drugs), the more similar the drugs (or diseases), the Gaussian similarity of drugs and diseases can be calculated according to the association matrix Y. Firstly, i-th row of the association matrix Y is used to represent the association profile of drug r_i. The association profile is a binary vector, and a value in the association profile indicates whether the drug (or disease) is association with a certain disease (or drug). Similarly, we can obtain association profile of drug r_j. Then, the Gaussian similarity between drug r_i and drug r_j can be calculated as follows:

$$RG\left(r_i, r_j\right) = exp\left(-r_r \|Y(i,:) - Y(j,:)\|^2\right) \tag{3}$$

$$r_r = 1/\left(\frac{1}{m}\sum\nolimits_{i=1}^{m} \|Y(i,:)\|^2\right) \tag{4}$$

Where r_r is responsible for controlling the bandwidth of Gaussian kernel and is the ratio of the number of drugs to the average number of diseases associated with each drug.

Similarly, the disease Gaussian similarity can be calculated as follows:

$$DG\left(d_i, d_j\right) = exp\left(-r_d \|Y(:,i) - Y(:,j)\|^2\right) \tag{5}$$

$$r_d = 1/\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} \|Y(:,i)\|^2\right) \tag{6}$$

Where r_d is computed similarly to r_r. Similar to the similarity matrix DD, the similarity matrix RG and DG need to be recalculated in the cross-validation.

2.4 Integrated Similarity for Drugs and Diseases

In this section, the similarity matrices of drugs and diseases are respectively integrated to implement the network consistency projection algorithm. By using the similarity network fusion (SNF) method [18], the above four drug similarity matrices are integrated into the drug similarity matrix MR, and the above three disease similarity matrices are integrated into the disease similarity matrix MD. SNF, which updates the similarity network corresponding to each similarity matrix in each iteration to make it closer to other networks, is a nonlinear method based on message passing theory. In the above way, SNF can capture common and complementary information between different networks and reduce noise.

Taking the integration of three disease similarity networks as an example, the main process of similarity network fusion is introduced. Firstly, the matrices $W^{(1)}$, $W^{(2)}$ and $W^{(3)}$ are used to represent three different disease similarity networks, and the correspondingly state matrices $P^{(1)}$, $P^{(2)}$ and $P^{(3)}$ are obtained as follows:

$$P(i,j) = \begin{cases} \frac{W(i,j)}{2\sum_{k\neq i} W(i,k)}, j \neq i \\ 1/2, j = i \end{cases} \tag{7}$$

Where $W(i, j)$ denotes the similarity of node i and node j.

Secondly, the correspondingly kernel matrices $S^{(1)}$, $S^{(2)}$ and $S^{(3)}$ are calculated by using the matrices $W^{(1)}$, $W^{(2)}$, and $W^{(3)}$ as follows:

$$S(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k\in N_i} W(i,k)}, j \in N_i \\ 0, otherwise \end{cases} \tag{8}$$

Where N_i represents the set of K (empirically set to 20) neighbors of node i.

Then, the key step of the similarity network fusion approach is to iteratively update the state matrices, and the updating process is represented as follows:

$$P_{t+1}^{(1)} = S^{(1)} \times \left(\frac{P_t^{(2)} + P_t^{(3)}}{2} \right) \times \left(S^{(1)} \right)^T \tag{9}$$

$$P_{t+1}^{(2)} = S^{(2)} \times \left(\frac{P_t^{(1)} + P_t^{(3)}}{2} \right) \times \left(S^{(2)} \right)^T \tag{10}$$

$$P_{t+1}^{(3)} = S^{(3)} \times \left(\frac{P_t^{(1)} + P_t^{(2)}}{2} \right) \times \left(S^{(3)} \right)^T \tag{11}$$

Where $P_{t=0}^{(1)} = P^{(1)}, P_{t=0}^{(2)} = P^{(2)}$ and $P_{t=0}^{(3)} = P^{(3)}$. After each iteration, we normalize the state matrices. After t (empirically set to 20) iterations and updates, the integrated similarity matrix P is finally obtained by taking the mean of the state matrices as follows:

$$P = \frac{P_t^{(1)} + P_t^{(2)} + P_t^{(3)}}{3} \tag{12}$$

2.5 Network Consistency Projection Method

Through the above section, the integrated drug similarity matrix *MR* and disease similarity matrix *MD* are obtained. In this section, *MR* and *MD* are also used to represent the integrated drug similarity network and disease similarity network, respectively. Moreover, the association matrix Y is also regarded as drug-disease association network. We perform the network consistency projection method on the integrated drug similarity network and the disease similarity network respectively, and obtain two projection scores, namely the drug space projection score and the disease space projection score.

The drug space projection score and the disease space projection score are combined and normalized to obtain the final prediction score.

The projection of the drug similarity network on the drug-disease association network is the drug space projection, which can be calculated as follows:

$$RSP(i,j) = \frac{MR(i,:) \times Y(:,j)}{|Y(:,j)|} \tag{13}$$

Where *MR(i, :)* is the *i*-th row of the matrix *MR,* representing the similarities between drug r_i and all drugs; *Y(:, j)* is the the association profile of disease d_j; $|Y(:,j)|$ represents the length of the association profile *Y(:, j).*

Similarly, the disease space projection can be computed as follows:

$$DSP(i,j) = \frac{Y(i,:) \times MD(:,j)}{|Y(i,:)|} \tag{14}$$

Where *MD(:, j)* represents the similarities between disease d_j and all diseases.

The final prediction score of drug r_i and disease d_j can be calculated as follows:

$$NCP(i,j) = \frac{RSP(i,j) + DSP(i,j)}{|MR(i,:)| + |MD(:,j)|} \tag{15}$$

3 Experiments

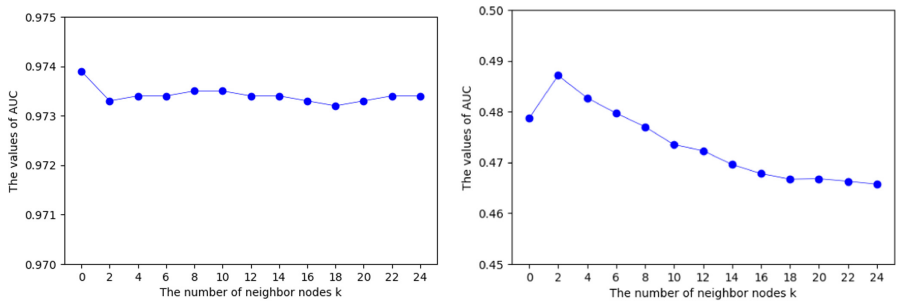
3.1 Evaluation Metrics

To evaluate the ability of our method to predict potential drug-disease associations, a fivefold cross-validation is performed. All known drug-disease associations in the LRSSL data set are randomly divided into 5 roughly equal subsets, each of which is used as the test set in turn, and the remaining subsets are used as the training set. After the performing prediction, the prediction scores of all drug-disease associations are ranked in descending order. If the ranking of the drug-related disease is higher than a specific threshold, it is considered a True Positive (TP) sample; otherwise, it is considered a False Negative (FN) sample. Moreover, if the ranking of a disease not associated with the drug is higher than a specific threshold, it is considered a False Positive (FP) sample; otherwise, it is considered a True Negative (TN) sample. According to different ranking thresholds, the number of samples in each of the above four categories can be calculated to construct receiver operating characteristic (ROC) curve and precision-recall (PR) curve. The area under ROC curve (AUC) and the area under PR curve (AUPR) are used to evaluate the overall performance of the prediction methods.

3.2 Parameter Analysis

Different values of the hyperparameter can produce different prediction performance, so it is necessary to determine the optimal value of the hyperparameter to achieve the best performance. In our method, the optimal value of hyperparameter *k* is determined within

the range [0, 24]. As shown in Fig. 3, as the value of k changes, the AUC value does not change significantly, while the AUPR value changes more significantly. In addition, when $k = 2$, the AUPR value reaches the maximum and decreases gradually with the increase of k. Therefore, we set the value of k to 2 as the optimal parameter value, and then the AUC value and AUPR value of NCPDDA are 0.9733 and 0.4871, respectively.



(a) (b)

Fig. 3. The predicted results under different number neighbor nodes settings. (a) AUC values for various settings. (b) AUPR values for various settings.

3.3 Comparison with Other Methods

To evaluate the performance of NCPDDA, we compare it with other three classical approaches: OMC2 [7], BNNR [6], and MBiRW [11]. In order to make a fair comparison, the best hyperparameter values of the other methods are selected according to their publications. As depicted Fig. 4, the AUC value and AUPR value of NCPDDA are superior to those of other methods. Specifically, NCPDDA obtain the best AUC value of 0.9733, while OMC2, BNNR and MBiRW are 0.9342, 0.9101 and 0.9122, respectively. Moreover, NCPDDA also achieve the best AUPR value of 0.4871, which are 8.83%, 9.19% and 17.21% higher than OMC2, BNNR and MBiRW, respectively. The other three methods only use one type of drug similarity and disease similarity, while NCPDDA integrates four types of drug similarities and three types of similarities. The experimental results indicate that the integration of multiple types of similarities can improve the performance of the prediction methods.

3.4 Comparison of Different Similarity Network Fusion Methods

In the process of similarity information fusion, many methods adopt linear fusion methods (for example, mean fusion), while our proposed method uses similarity network fusion (SNF) method. Compared with mean fusion method, SNF can integrate common information and complementary information of various types of data. Besides, the similarities of drugs and diseases may be incomplete or noisy, and SNF can effectively reduce noise.

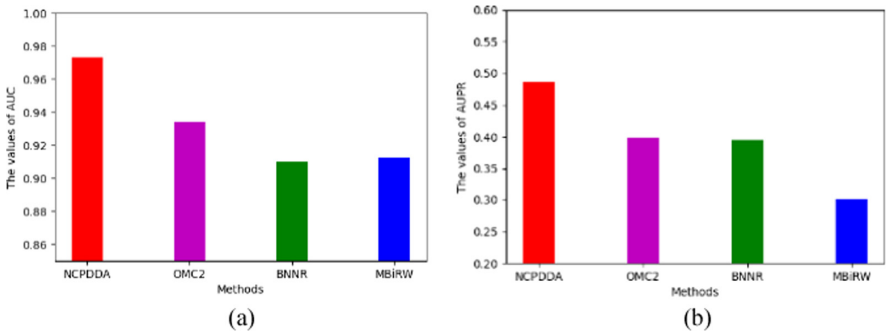


Fig. 4. The predicted results of all methods. (a) AUC values for the various methods. (b) AUPR values for various methods.

We set up two groups of comparative experiments to illustrate the effect of similarity network fusion method. NCPDDA_DrugMean represents that the mean fusion method is used to integrate three types of drug similarities to calculate one type of disease similarity, and NCPDDA_Mean represents that four kinds of drug similarities and three kinds of disease similarities are integrated into one kind of drug similarities and disease similarities respectively by mean fusion method. As shown in Fig. 5, the AUC values of NCPDDA and NCPDDA_DrugMean are approximately equal, while the AUPR value of NCPDDA is 3% higher than NCPDDA_DrugMean. Moreover, the AUC value of NCPDDA is 2.48% higher than NCPDDA_Mean and the AUPR value of NCPDDA is 4.25% higher than NCPDDA_Mean.

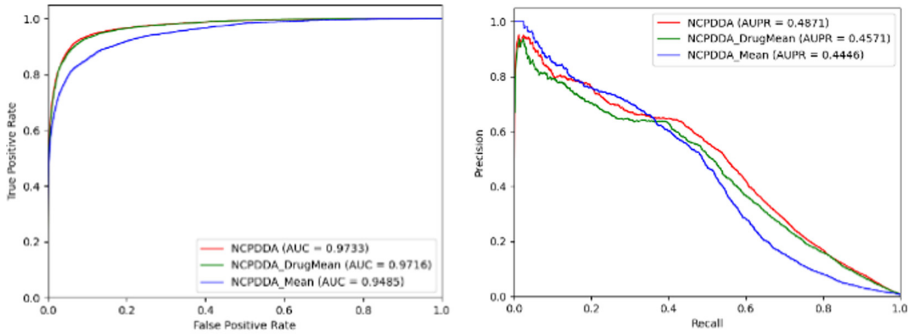


Fig. 5. The performance of different similarity fusion methods. (a) ROC curves of different methods. (b) precision-recall curves of different methods.

3.5 Case Studies

Through the previous experiments, the excellent predictive performance of NCPDDA has been confirmed. In this section, we further validate the ability of the NCPDDA to predict potential associations between drugs and diseases. All known associations in

the LRSSL dataset are treated as the training set to train the NCPDDA model, and the trained NCPDDA is used to get the prediction score of all unknown associations. The candidate diseases for each drug are then ranked according to the predicted scores.

The top 5 candidate diseases for four representative drugs, Levodopa, Capecitabine, Flecainide and Amantadine, are validated by searching authoritative public databases, such as DrugBank [19] and CTD [20]. As shown in Table 1, more than 3 new indications are validated for each representative drug. It further suggests that NCPDDA can be used to predict new indications for drugs in practical applications.

Table 1. The top 5 candidate diseases for the four representative drugs.

Drugs	Top 5 candidate diseases	Evidences
Levodopa	Hyperprolactinemia	CTD
	Psychotic disorders	
	Dyskinesia, drug-induced	CTD
	Schizophrenia	CTD
	Tourette syndrome	
Capecitabine	Stomach neoplasms	CTD
	Carcinoma, basal cell	CTD
	Rectal neoplasms	CTD
	Folic acid deficiency	
	Anemia, megaloblastic	
Flecainide	Ventricular fibrillation	CTD
	Tachycardia, supraventricular	CTD
	Ventricular premature complexes	CTD
	Atrial fibrillation	CTD/DrugBank
	Atrial flutter	CTD
Amantadine	Psychotic disorders	CTD
	Tourette syndrome	CTD
	Hyperprolactinemia	
	Schizophrenia	CTD
	Huntington disease	CTD/DrugBank

Besides, our approach has also identified some novel drug-disease associations, including: Levodopa for psychotic disorders and tourette syndrome; Capecitabine for folic acid deficiency and anemia, megaloblastic; Amantadine for hyperprolactinemia. Although these associations are not recorded in the database, it does not necessarily mean that they do not exist.

4 Conclusion

In this work, we develop a method based on the network consistency projection to achieve drug repositioning. In order to accurately predict potential drug-disease associations, our proposed approach effectively integrates information from multiple sources. In addition, the similarity network fusion is used to integrate the data to reduce the influence of noise. Compared with three classical prediction methods, our method is proved to have excellent performance. In the case studies section, four representative drugs are studied to further prove the effectiveness of our method.

However, although our method has achieved some results, we must acknowledge some limitations of our method. First, our method of calculating the similarities of drugs and diseases may not be optimal, and there may be better methods of calculating the similarities. Second, we should integrate more types of information to further improve prediction performance. In the future research, we will conduct further research on the two points mentioned above.

Acknowledgement. This paper is supported by National Key Research and Development Program of China (No. 2017YFE0103900 and 2017YFA0504702), the NSFC projects Grant (No. 61932018, 62072441 and 62072280), Beijing Municipal Natural Science Foundation Grant (No. L182053).

References

1. Emmert-Streib, F., Tripathi, S., Simoes, R.D.M., Hawwa, A.F., Dehmer, M.: The human disease network: opportunities for classification, diagnosis, and prediction of disorders and disease genes. Syst. Biomed. **1**(1), 20–28 (2013)
2. Weng, L., Zhang, L., Peng, Y., Huang, R.S.: Pharmacogenetics and pharmacogenomics: a bridge to individualized cancer therapy. Pharmacogenomics **14**(3), 315–324 (2013)
3. Ghofrani, H.A., Osterloh, I.H., Grimminger, F.: Sildenafil: from angina to erectile dysfunction to pulmonary hypertension and beyond. Nat. Rev. Drug Discov. **5**(8), 689–702 (2006)
4. Gottlieb, A., Stein, G.Y., Ruppin, E., Sharan, R.: PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol. Syst. Biol. **7**(1), 496 (2011)
5. Liang, X., et al.: LRSSL: predict and interpret drug-disease associations based on data integration using sparse subspace learning. Bioinformatics **33**(8), 1187–1196 (2017)
6. Yang, M., Luo, H., Li, Y., Wang, J.: Drug repositioning based on bounded nuclear norm regularization. Bioinformatics **35**(14), i455–i463 (2019)
7. Yang, M., Luo, H., Li, Y., Wu, F.-X., Wang, J.: Overlap matrix completion for predicting drug-associated indications. PLoS Computat. Biol. **15**(12), e1007541 (2019)
8. Zhang, W., et al.: Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinformatics **19**(1), 1–12 (2018)
9. Wang, W., Yang, S., Zhang, X., Li, J.: Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics **30**(20), 2923–2930 (2014)
10. Martínez, V., Navarro, C., Cano, C., Fajardo, W., Blanco, A.: DrugNet: network-based drug-disease prioritization by integrating heterogeneous data. Artif. Intell. Med. **63**(1), 41–49 (2015)
11. Luo, H., et al.: Drug repositioning based on comprehensive similarity measures and Bi-Random walk algorithm. Bioinformatics **32**(17), 2664–2671 (2016)

12. Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Bryant, S.H.: PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. **37**(suppl_2), W623–W633 (2009)
13. Mitchell, A., et al.: The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. **43**(D1), D213–D221 (2015)
14. UniProt Consortium, U.: The universal protein resource (UniProt) in 2010. Nucleic Acids Res. **38**(suppl_1), D142–D148 (2010)
15. Van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G., Leunissen, J.A.: A text-mining analysis of the human phenome. Eur. J. Hum. Genet. **14**(5), 535–542 (2006)
16. Wang, D., Wang, J., Lu, M., Song, F., Cui, Q.: Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. Bioinformatics **26**(13), 1644–1650 (2010)
17. Xuan, P., Cao, Y., Zhang, T., Wang, X., Pan, S., Shen, T.: Drug repositioning through integration of prior knowledge and projections of drugs and diseases. Bioinformatics **35**(20), 4108–4119 (2019)
18. Wang, B., et al.: Similarity network fusion for aggregating data types on a genomic scale. Nat. Methods **11**(3), 333 (2014)
19. Wishart, D.S., et al.: DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. **36**(suppl_1), D901–D906 (2008)
20. Davis, A.P., et al.: The comparative toxicogenomics database: update 2019. Nucleic Acids Res. **47**(D1), D948–D954 (2019)