GRTR: Drug-Disease Association Prediction Based on Graph Regularized Transductive Regression on Heterogeneous Network

Qiao Zhu, Jiawei Luo(✉), Pingjian Ding, and Qiu Xiao

College of Computer Science and Electronic Engineering,
Collaboration and Innovation Center for Digital Chinese Medicine
in Hunan Province, Hunan University, Changsha 410082, China
luojiawei@hnu.edu.cn

Abstract. Computational drug repositioning helps to decipher the complex relations among drugs, targets, and diseases at a system level. However, most existing computational methods are biased towards known drugs-disease associations already verified by biological experiments. It is difficult to achieve excellent performance with sparse known drug-disease associations. In this article, we present a graph regularized transductive regression method (GRTR) to predict novel drug-disease associations. The proposed method first constructs a heterogeneous graph consisting of three interlinked sub-graphs including drugs, diseases and targets from multiple sources and adopts preliminary estimation of drug-related disease to initial unknown drug-disease associations for unlabeled drugs. Since the known drug-disease associations are sparse, graph regularized transductive regression is used to score and rank drug-disease associations iteratively. In the computational experiments, the proposed method achieves better performance than others in terms of AUC and AUPR. Moreover, the varying of parameters is shown to verify the importance of preliminary estimation in GRTR. Case studies on several selected drugs further confirm the practicality of our method in discovering potential indications for drugs.

Keywords: Transductive regression · Drug repositioning
Drug-disease association · Graph regularization · Heterogeneous network

1 Introduction

Traditional drug development faces difficulties relating to the expensive, time consuming and high risk of failure. Studies have demonstrated that drug repositioning, which aims to discovery new indications for existing drugs, offers a promising alternative to drug development. Some successful repositioned drugs (e.g. Sildenafil, thalidomide, raloxifene) have historically generated high revenues for their patent holders or companies [1]. Compared to in vivo experimental methods for drug repositioning, in silico approaches are efficient at identifying potential drug-disease association, and thus significantly reduce research costs. Therefore, it is necessary to develop a computational method for identifying drug-disease associations.

F. Zhang et al. (Eds.): ISBRA 2018, LNBI 10847, pp. 13–25, 2018.
https://doi.org/10.1007/978-3-319-94968-0_2

To date, much effort has been allocated to developing computational approaches for predicting drug-disease associations. Conventional computational methods mainly depend on two strategies, the network-based method and feature-based method. A key idea behind network-based algorithms is the construction of complex biological networks with large-scale biological data. Wang et al. [2] proposed a drug-disease heterogeneous network model termed Heterogeneous Graph Based Inference (HGBI) and extended the algorithm to a three-layer network (HL_HGBI), adding a new layer of the target information [3]. However, the assumption was that drugs should have diverse indications and diseases should have diverse treatments. Martínez et al. [4] constructed a complex network which included drugs, diseases and proteins. Protein interactions were used as a bridge to perform DrugNet, a general network-based prioritization based on a propagation flow algorithm. Luo et al. [5] exploited known drug-disease associations to devise the drug-drug and disease-disease similarity measures, then building a drug-disease heterogeneous network, on which a bi-random walk algorithm was adopted to predict novel potential associations between drugs and diseases.

Much attention has also been devoted to introducing feature-based methods. Bleakley et al. provided a supervised learning approach [i.e. support vector machine (SVM)] on a bipartite local model (BLM) from chemical and genomic data [6]. Mei et al. [7] proposed BLM-NII, combining BLM with a neighbor-based interaction-profile inferring(NII) procedure. Gottlieb et al. [8] conducted multiple drug-drug and disease-disease similarity measures as classification features, implementing a classification algorithm named PREDICT to infer potential drug indications. Yang et al. [9] calculated relevance scores between drugs and diseases from a drug-target-pathway-gene-disease network and learnt a probabilistic matrix factorization model (PMF) based on known drug-disease associations to classify drug-disease associations. However, most of these approaches rely on the known association information and directly set the weight of unknown disease-drug associations to zero. This is perhaps the major reason that existing methods can't obtain a satisfactory performance based on sparse known associations validated by biological experiments.

In this article, we propose a graph regularized transductive regression (GRTR) method to deal with the problem of the sparse known associations for drug-disease association prediction. A three-layer heterogeneous network composed of drugs, diseases and targets is constructed from multiple datasets. Then we approximately calculate drug related diseases from local neighborhood information and adjust the weight of links with diseases based on it. Through a transductive regression model with graph regularization, the relevance score for potential drug-disease associations will be iteratively updated and all drugs ranked by their scores and judged whether they are related to a disease. Compared to the previous nine advanced prediction methods, GRTR performs better in terms of AUC and AUPR. Furthermore, the effect of varying weighted parameters and the effect of preliminarily estimating drug-related disease are analyzed. Case studies on the selected drugs and targets further exhibit the predictive ability of drug-disease association.

2 Methods

The overall process of predicting new drug-disease associations by GRTR is displayed in Fig. 1. GRTR first constructs a three-layer heterogeneous network composed of drugs, diseases and targets. Next, local information is obtained based on preliminary estimates for drug-related disease from the distribution of diseases associated with neighbor nodes in the heterogeneous network. Finally, using the heterogeneous network, the known relationships with diseases and preliminary estimation results as inputs, GRTR adopt graph regularization transductive regression to score and rank drug-disease associations iteratively.

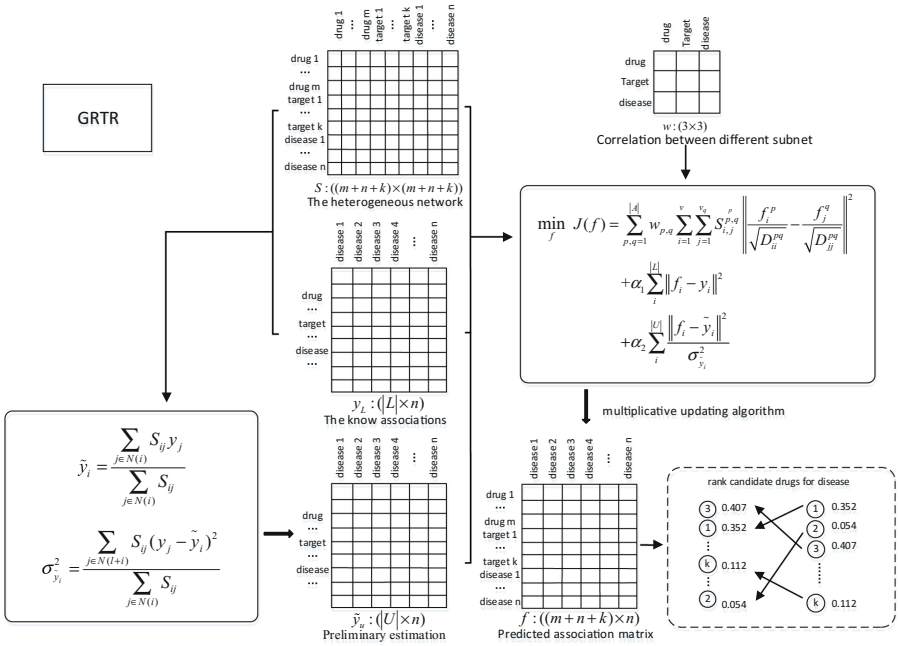

Fig. 1. GRTR workflow. Given the inputs of the heterogeneous network matrix **S** and the matrix of known association $\mathbf{y_L}$, we first obtain preliminary estimates for drug related diseases $\mathbf{y_u}$ using neighbor distribution information. We then score and rank drug-disease associations iteratively based on graph regularization transductive regression. The top rank drugs for each disease in the predicted association matrix **f** are treated as the candidate drugs for those diseases for further experimental investigation.

2.1 Heterogeneous Network Construction

The three-layer heterogeneous network consists of three nodes types: drug nodes, disease nodes and target nodes. Suppose that m, n and k are the number of drugs, diseases and targets, respectively. $S^{11} = \left\{S_{i,j}^{11}\right\}_{i=1,j=1}^{m,m}$ is an adjacency matrix of the drugs similarity network, $S^{22} = \left\{S_{i,j}^{22}\right\}_{i=1,j=1}^{k,k}$ is an adjacency matrix of the protein interaction network and $S^{33} = \left\{S_{i,j}^{33}\right\}_{i=1,j=1}^{n,n}$ is an adjacency matrix of the disease similarity network. Drug similarities can be calculated based on their chemical structures. Disease similarities and protein-protein interactions can be obtained from online datasets. We connect the above three subnetworks using experimentally verified drug-disease associations ($S^{13} = \left\{S_{i,j}^{13}\right\}_{i=1,j=1}^{m,n}$), target-disease associations ($S^{23} = \left\{S_{i,j}^{23}\right\}_{i=1,j=1}^{k,n}$) and drug-target associations ($S^{12} = \left\{S_{i,j}^{12}\right\}_{i=1,j=1}^{m,k}$) to form a heterogeneous network. The adjacency matrix of the heterogeneous network can be represented as follows:

$$S = \begin{pmatrix} S^{11} & S^{12} & S^{13} \\ (S^{12})^T & S^{22} & S^{23} \\ (S^{13})^T & (S^{23})^T & S^{33} \end{pmatrix}$$

where $(\cdot)^T$ represents the transpose of a matrix.

2.2 Preliminary Estimation of Drug Related Disease

In our research, the node with no known associations with a disease is unlabeled while other nodes are labeled. Preliminary estimation for the related diseases for an unlabeled drug is a local estimation. According to the assumption that drugs which are 'close together' will have associations with the same disease [10], we will consider neighborhood information based on the equal combination of diseases which have association with neighbor nodes in the heterogeneous network. Firstly, the neighbors of a drug i can be defined by the nearest labeled nodes N in the heterogeneous network.

$$N(i) = \left\{j \,|\, S_{ij} > \sigma, 1 \le i \le m+n+k, 1 \le j \le m+n+k\right\} \tag{1}$$

where σ is a threshold and in this paper $\sigma = 0.5$. Then we use the mean distribution of the neighbor's disease to describe the biological network's local information and obtain preliminary estimations for the related diseases ($\tilde{y}$).

$$\tilde{y}_i = \frac{\sum\limits_{j \in N(i)} S_{i,j} y_j}{\sum\limits_{j \in N(i)} S_{i,j}} \tag{2}$$

where y_j denotes the known associations between nodes j and diseases. Here, diseases can be understood as discrete variables. Hence, the variance of a neighbor's disease distribution ($\sigma^2_{\tilde{y}}$) can be obtained as follows:

$$\sigma^2_{\tilde{y}_i} = \frac{\sum\limits_{j \in N(l+i)} S_{i,j}(y_j - \tilde{y}_i)^2}{\sum\limits_{j \in N(i)} S_{i,j}} \tag{3}$$

2.3 Graph Regularized Transductive Regression

The main idea of our prediction method is based on transductive regression which is one of the most popular methods for imbalanced (sparse) data analysis, because prediction though transductive regression can lead to good knowledge extraction of the hidden network structure [11]. Wan et al. [12] presented a graph regularization-based transductive regression (Grempt) method using a symmetry meta-path to deal with label prediction on heterogeneous information networks, which have performed satisfactorily. In order to address the limitations of the symmetry meta-path, we revise the objective function's first term to directly consider different links classes in the heterogeneous network. The revised objective function is defined as follows:

$$\begin{aligned} J(f) = & \sum_{p,q=1}^{|A|} w_{p,q} \sum_{i=1}^{v_p} \sum_{j=1}^{v_q} S_{i,j}^{p,q} \left\| \frac{f_i^p}{\sqrt{D_{ii}^{pq}}} - \frac{f_j^q}{\sqrt{D_{jj}^{pq}}} \right\|^2 \\ & + \alpha_1 \sum_{i}^{|L|} \|f_i - y_i\|^2 \\ & + \alpha_2 \sum_{i}^{|U|} \frac{\|f_i - \tilde{y}_i\|^2}{\sigma^2_{\tilde{y}_i}} \end{aligned} \tag{4}$$

where α_1 and α_2 are two regularization coefficients which balance the different components of the model. A = {drug, disease, target} is the network nodes category, $w_{p,q}$ is the correlation between categories A_p and A_q $(p, q \in \{1, 2, \ldots, |A|\})$, v_p is the number of nodes which belong to category A_p, $S_{i,j}^{pq}$ is the relevance between object $i \in A_p$ and object $j \in A_q$ in the network, f_i and f_i^p are the prediction results of node i where p denotes $i \in A_p$, D_{ii}^{pq} is the sum of the i-th row in S^{pq}, L is the labeled nodes set

which has an association with disease and U is the unlabeled node set. The model consists of 3 functions and each one corresponds to different meaning:

- The first part of the objective function is the global smoothness item, which formulates that similar nodes are likely to be associated with similar diseases.
- The second term of the objective function minimizes the difference between the predicted results and the known association.
- The last term formulates a regularization item to minimize the difference between the predicted results and the preliminary estimation from local characteristics.

The global minimum is calculated by differentiating (4) with respect to f_L^p and f_U^p respectively, which gives:

$$\begin{aligned}\frac{\partial J(f)}{\partial f_L^p} = &\sum_{p,q,p\neq q}^{|A|} 2w_{p,q}(f_L^p - R_{LL}^{pq}f_L^q - R_{LU}^{pq}f_U^q) \\ &+ 4w_{p,p}(f_L^p - R_{LL}^{pp}f_L^p - R_{LU}^{pp}f_U^p) + 2\alpha_1(f_L^p - y_L^p)\end{aligned} \tag{5}$$

$$\begin{aligned}\frac{\partial J(f)}{\partial f_U^p} = &\sum_{p,q,p\neq q}^{|A|} 2w_{p,q}(f_U^p - R_{UL}^{pq}f_L^q - R_{UU}^{pq}f_U^q) \\ &+ 4w_{p,p}(f_U^p - R_{UL}^{pp}f_L^p - R_{UU}^{pp}f_U^p) + 2\frac{\alpha_1}{\sigma_{\tilde{y}^p}^2}(f_L^p - \tilde{y}^p)\end{aligned} \tag{6}$$

where f_L^p denotes the prediction result of labeled nodes belonging to A_p and f_U^p denotes the prediction result of unlabeled nodes belonging to A_p. $R^{pq} = (D^{pq})^{-\frac{1}{2}}S^{pq}(D^{qp})^{-\frac{1}{2}}$ is the integration of the whole heterogeneous network, which can be rearranged based on labeled and unlabeled objectives.

$$R^{pq} = \begin{pmatrix} R_{LL}^{pq} & R_{LU}^{pq} \\ R_{UL}^{pq} & R_{UU}^{pq} \end{pmatrix} \cdot \; p, q \in \{1, 2, \ldots, |A|\}$$

Suggested that $\frac{\partial J(f)}{\partial f_L^p} = 0$ and $\frac{\partial J(f)}{\partial f_U^p} = 0$, the closed-form solution is obtained. However, the iterative solution is sometimes preferable [13]. The detail steps of GRTR to predict potential associations are described in Algorithm 1.

Algorithm 1. GRTR

Input: $R, D, T, G_{RD}, G_{TD}, G_{RT}, w, y, m, n, k, \sigma$;

Output: f

1. Use R, D, T, G_{RD}, G_{TD} and G_{RT} to build the heterogeneous network S;
2. Get $\tilde{y}$ and $\sigma_{\tilde{y}}^2$ from the preliminary estimation of drug-related disease based on (2) and (3);
3. Initialize $f(0) = (y_L^T, \tilde{y}_L^T)^T$ and $t = 0$;
4. Repeat
5. $p, q \in \{1,2, \ldots, |A|\}$;
6. $$f_L^p(t+1) = \frac{\alpha_1 y_L{}^p + \sum_{p,q,p\neq q}^{|A|} w_{p,q}\left(R_{LL}^{pq} f_L^q(t) + R_{LU}^{pq} f_U^q(t)\right) + 2w_{p,p}\left(R_{LL}^{pp} f_L^p(t) + R_{LU}^{pp} f_U^p(t)\right)}{\sum_{p,q,p\neq q}^{|A|} w_{p,q} + 4w_{p,p} + \alpha_1};$$
7. $$f_U^p(t+1) = \frac{\frac{\alpha_2 \tilde{y}^p}{\sigma_{\tilde{y}^p}^2} + \sum_{p,q,p\neq q}^{|A|} w_{p,q}\left(R_{UL}^{pq} f_L^q(t) + R_{UU}^{pq} f_U^q(t)\right) + 2w_{p,p}\left(R_{UL}^{pp} f_L^p(t) + R_{UU}^{pp} f_U^p(t)\right)}{\sum_{p,q,p\neq q}^{|A|} w_{p,q} + 4w_{p,p} + \frac{\alpha_2}{\sigma_{\tilde{y}^p}^2}};$$
8. $t = t + 1$;
9. Repeat until convergence;
10. $f = \left\{f^{1^T}, f^{2^T}, \ldots, f^{|A|^T}\right\}^T$. Every unlabeled drug is assigned to the disease label which is on top r scores in each row of f.

3 Experiment and Results

3.1 Dataset

Experimentally confirmed drug-disease associations and drug–target associations are both downloaded from the supplementary material of [8]. Gottlieb et al. have collected 1933 known drug-disease associations involving 593 drugs registered in DrugBank [14] and 313 diseases listed in the Online Mendelian Inheritance in Man (OMIM) database [15]. At last, we get 2814 known drug–target associations between 593 drugs and 777 proteins.

The interactions between diseases and proteins are obtained from DisGeNET [16], for a total of 10010 relationships between 3221 proteins and 313 diseases.

The disease–disease similarity network is downloaded from Online Mendelian Inheritance in Man Mining Tool (MimMiner) [17]. According to the MimMiner database, disease–disease similarities have already been normalized to the range [0, 1].

The protein–protein interaction network is built using 37039 binary interactions among 9465 genes in the Human Protein Reference Database (HPRD) [18].

The drug–drug similarities are calculated based on their chemical structures. First, the chemical structures of all drug compounds in the Canonical Simplified Molecular Input Line-Entry System (SMILES) format [19] are downloaded from DrugBank. Then, the Chemical Development Kit [20] is used to calculate a binary fingerprint for each drug. Finally, Tanimoto score [21] of two drugs was calculated based on their fingerprints, which was in the range of [0, 1].

3.2 Parameters Selection and the Effect of Preliminary Estimation for Drug-Related Disease

There are three parameters w, α_1 and α_2 in our prediction. w controls the importance of different network. α_1 and α_2 control the contribution of known labeled objects and preliminary estimation, respectively. We set $w = 1$ for easy. To determine the optimal configuration of α_1 and α_2, we firstly let both increase from 0 to 1 in increments of 0.05 and record the change in AUC. The results can be seen in Fig. 2(a), in which AUC value increases rapidly as both α_1 and α_2 increase, and then became steadily reaching the maximum AUC value. However, in order to determine the general future trend as α_1 and α_2 become larger, we also vary them from 1 to 200, as demonstrated in Fig. 2(b). AUC value rapidly decreases in the range $0 \leq \alpha_1 \leq 10$ and then remains almost constant in the range $10 < \alpha_1 \leq 200$ which shows the result is not improved for these regions. But, there is an opposite trend for α_2, which first rises rapidly in the range of $0 \leq \alpha_2 \leq 10$ and after keeping a short constant, it decreases in the range $30 \leq \alpha_2 \leq 200$. Finally, we select $\alpha_1 = 1$ and $\alpha_2 = 20$ for getting a better prediction result. Although α_2 is much larger than α_1, it fits with the reality that preliminary estimation for drug-related disease information is more important than it is for predicting new relations.

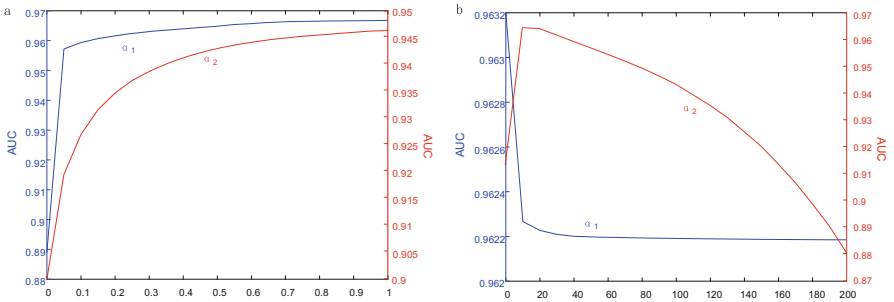

Fig. 2. The influence of different α_1 and α_2 values on AUC. (a): 0–1 in 0.05 increments. (b): 0–200 in 10 increments.

If we don't use the preliminary estimation for drug-related disease ($\alpha_2 = 0$, $\alpha_1 \neq 0$), the largest AUC is 0.9139. As α_2 gets larger, AUC turns to be larger until reaches the best value. To a certain extent, preliminary estimation for drug-related disease is significant, and can improve predictive ability.

3.3 Compared with Existing Methods

Systematic experiments are performed to evaluate the performance of the presented approach with nine other methods: Weighted Nearest Neighbor-Gaussian Interaction Profile(WNN-GIP) [22], Collaborative Matrix Factorization(CMF) [23], Kernelized Bayesian matrix factorization(KBMF) [24], Neighborhood Regularized Logistic Matrix Factorization(NRLMF) [25], a bipartite local model (BLM) [12], BLM with neighbor-based interaction profile inferring (BLM-NII) [7], comprehensive similarity measures and Bi-Random Walk algorithm (MBiRW) [5], standard LapRLS improved by incorporating a new kernel (NetLapRLS) [26]. We use 10-fold validation to compare GRTR performance. The area under the receiver operating characteristic (ROC) curve (AUC) [27] and the area under the precision recall (PR) curve (AUPR) are used to measure the quality of the predicted drugs for diseases. Figure 3 shows the ROC and PR curves of the 10-fold validation experiments. Table 1 gives the AUC and AUPR values. As expected, the GRTR's AUC value is 0.9668, which outperforms all other competitive methods significantly. GRTR is 2.10% better than the second-best method, NRLMF, which also achieved an impressive result of 0.9465. For AUPR, we observe that the values are lower than those in the original papers. The main reason for this is that the data we used is larger and comparatively sparser. But GRTR also performs well, obtaining the second best in the dataset with the AUPR value of 0.5925. Though GRTR is slightly lower than NRLMF, it is still very competitive among the methods.

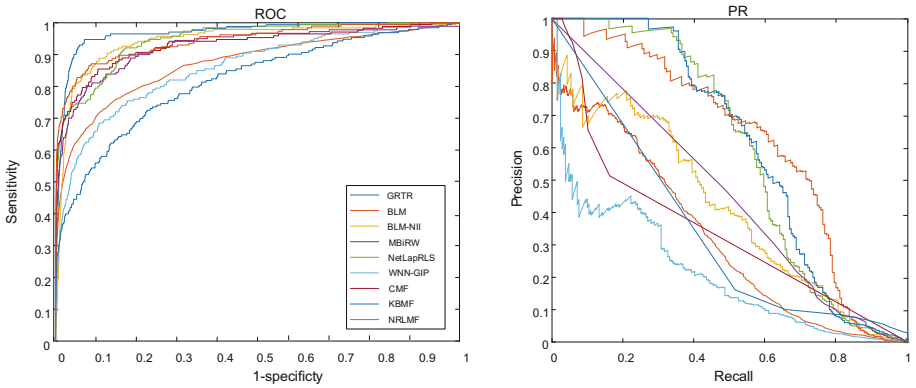

Fig. 3. The ROC and PR curves of GRTR and nine existing methods.

Table 1. AUC and AUPR values of GRTR and nine existing method.

Metric	GRTR	BLM	BLM-NII	MBiRW	NetLapRLS	WNN-GIP	CMF	KBMF	NRLMF
AUC	**0. 9668**	0. 8719	0.9442	0. 9179	0.9444	0. 8584	0.9309	0.8713	0. 9465
AUPR	0.5925	0.3256	0.4075	0.0469	0.5750	0.205	0.3455	0.3463	**0.6790**

3.4 Case Study

Here, the capability of our method in predicting novel drug-disease associations is further examined here. One well-known biological database CTD [30] and some references are used to verify the predicted novel drug-disease associations. For each disease, the candidate drugs are ranked based on the prediction scores and the top-10 predicted drugs as prediction results are collected. For instance, 8 of the top 10 potentially related drugs have been directly shown to be linked with Diabetes Mellitus type II (see Table 2), a endocrine system disease and metabolic disease. Lovastatin (DrugBank: DB00227) is predicted to treat it and has been recorded in CTD. Figure 4 presents lovastatin's neighbor drugs and the diseases they can treat. Vitamin d-dependent rickets, osteoporosis and hyperlipoproteinemia are metabolic disease. And barakat syndrome is an endocrine system disease. In addition, we also find many associated genes between those that lovastatin can interact with to treat diabetes mellitus and the neighbors can act on to treat corresponding disease, e.g. there are 1307 genes shared with the pravastatin treating hyperlipoproteinemia, 1460 genes shared with the calcitriol treating vitamin d-dependent Rickets and 762 genes shared with the Ergocalciferol treating barakat syndrome, etc.

Table 2. The top 10 predicted results for diabetes mellitus associated drugs.

Rank	Drug	Evidence
1	Guanfacine	Literature [28]
2	Nalbuphine	
3	Lovastatin	CTD
4	Tamoxifen	CTD
5	Bicalutamide	
6	Promethazine	CTD
7	Risperidone	CTD
8	Dinoprostone	CTD
9	Spironolactone	CTD
10	Carvedilol	Literature [29]

Fig. 4. Lovastatin (DB00227)'s neighbors and diseases can be treated. The yellow circle is the predicted drug, the red circles are the neighbor drugs of the predicted drug and the green circles are the diseases its neighbor can treated. (Color figure online)

4 Conclusions

Identifying drug-disease associations is helpful in reducing the difficulty of drug development and contributing to improved understanding of the underlying complex relations among drugs, targets and diseases. In this work, we systematically studied the problem of predicting drug-disease associations. Conventional methods for drug-disease association prediction mainly achieved unsatisfactory performance for the sparse known associations. However, the number of drug-disease associations verified by biological experiments is far less than that of the potential drug-disease associations. Therefore, GRTR based on graph regularized transductive regression was developed to predict potential drug-disease associations. At first a three-layer heterogeneous network consisting of drugs, diseases and targets was constructed. Afterwards, preliminary estimation for drug-related diseases was conceived from neighbor information. Ultimately, transductive regression strategy was adopted a to predict drug-disease associations on the heterogeneous network. The superior performance of GRTR was validated by cross validation and the top-ranked predictions. Experiment results indicate that our method can predict better than nine other approaches. Furthermore, case studies on several drugs indicated that potential drug-disease association predicted by GRTR could assist in the biomedical research.

Despite the efficiency of GRTR, there are still some limitations which require further optimization. Firstly, our method involved multiple parameters and the establishment of the optimal parameter values is still a challenging problem. Secondly, more biological information can be used to improve predictions. Finally, although higher reliability has been achieved, the current capability of GRTR remains unsatisfactory and necessitates further improvement.

Acknowledgment. This work has been supported by the National Natural Science Foundation of China (Grant No. 61572180).

References

1. Ashburn, T.T., Thor, K.B.: Drug repositioning: identifying and developing new uses for existing drugs. Nat. Rev. Drug Discov. **3**, 673–683 (2004)
2. Wang, W., Yang, S., Li, J.: Drug target predictions based on heterogeneous graph inference. Biocomputing 2013, pp. 53–64. World scientific, Kohala Coast, Hawaii, USA (2012)
3. 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)
4. 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)
5. Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F.-X., Pan, Y.: Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics **32**(17), 2664–2671 (2016)
6. Bleakley, K., Yamanishi, Y.: Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics **25**(18), 2397–2403 (2009)

7. Mei, J.P., Kwoh, C.K., Yang, P., Li, X.L., Zheng, J.: Drug–target interaction prediction by learning from local information and neighbors. Bioinformatics **29**(2), 238–245 (2013)
8. 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)
9. Yang, J., Li, Z., Fan, X., Cheng, Y.: Drug-disease association and drug-repositioning predictions in complex diseases using causal inference-probabilistic matrix factorization. J. Chem. Inf. Model. **54**(9), 2562–2569 (2014)
10. Dudani, S.A.: The distance-weighted K-nearest-neighbor rule. IEEE Trans. Syst. Man. Cybern. SMC **6**(4), 325–327 (1976)
11. Luo, J., Ding, P., Liang, C., Cao, B., Chen, X.: Collective prediction of disease-associated miRNAs based on transduction learning. IEEE/ACM Trans. Comput. Biol. Bioinform. **14** (6), 1468–1475 (2017)
12. Wan, M., Ouyang, Y., Kaplan, L., Han, J.: Graph regularized meta-path based transductive regression in heterogeneous information network. In: Proceedings of the 2015 SIAM International Conference on Data Mining 2015, pp. 918–926 (2015)
13. Xiao, Q., Luo, J., Liang, C., Cai, J., Ding, P.: A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. Bioinformatics **34**(2), 239–248 (2018)
14. Knox, C., Law, V., Jewison, T., Liu, P., Ly, S., Frolkis, A., Pon, A., Banco, K., Mak, C., Neveu, V., Djoumbou, Y., Eisner, R., Guo, A.C., Wishart, D.S.: DrugBank 3.0: a comprehensive resource for Omics research on drugs. Nucleic Acids Res. **39**(suppl_1), D1035–D1041 (2011)
15. Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., McKusick, V.A.: Online mendelian inheritance in man (OMIM) a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. **33**, D514–D517 (2005)
16. Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., Furlong, L.I.: DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. **45**(D1), D833–D839 (2017)
17. Van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G., Leunissen, J.A.M.: A text-mining analysis of the human phenome. Eur. J. Hum. Genet. **14**, 535 (2006)
18. Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D.S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C.J., Kanth, S., Ahmed, M., Kashyap, M.K., Mohmood, R., Ramachandra, Y.L., Krishna, V., Rahiman, B.A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R., Pandey, A.: Human protein reference database—2009 update. Nucleic Acids Res. **37**(suppl_1), D767–D772 (2009)
19. Weininger, D.: SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. J. Chem. Inf. Comput. Sci. **28**(1), 31–36 (1988)
20. Steinbeck, C., Hoppe, C., Kuhn, S., Floris, M., Guha, R., Willighagen, E.L.: Recent developments of the chemistry development kit (CDK) - an open-source java library for chemo- and bioinformatics. Curr. Pharm. Des. **12**(17), 2111–2120 (2006)
21. Tanimoto, T.T.: Elementary mathematical theory of classification and prediction. IBM Internal report, pp. 1–10 (1958)
22. Van Laarhoven, T., Marchiori, E.: Predicting drug-target interactions for new drug compounds using a weighted nearest neighbor profile. PLoS One **8**(6), e66952 (2013)

23. Zheng, X., Ding, H., Mamitsuka, H., Zhu, S.: Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1025–1033. ACM, Chicago, Illinois, USA (2013)
24. Gönen, M.: Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. Bioinformatics **28**(18), 2304–2310 (2012)
25. Liu, Y., Wu, M., Miao, C., Zhao, P., Li, X.-L.: Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comput. Biol. **12**(2), e1004760 (2016)
26. Xia, Z., Wu, L.-Y., Zhou, X., Wong, S.T.: Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst. Biol. **4**(2), S6 (2010)
27. Li, G., Luo, J., Xiao, Q., Liang, C., Ding, P., Cao, B.: Predicting MicroRNA-disease associations using network topological similarity based on deepwalk. IEEE Access **5**, 24032–24039 (2017)
28. Coves, M.J., Gomis, R., Goday, A., Casamitjana, R., Rivera, F., Vilardell, E.: Antihypertensive treatment with guanfacine in patients with type II diabetes mellitus. Med Clin (Barc) **88**(8), 315–317 (1987)
29. Ahmad, A.: Carvedilol can replace insulin in the treatment of type 2 diabetes mellitus. J. Diab. Metab. **8**(2), (2017)
30. Davis, A.P., Grondin, C.J., Johnson, R.J., Sciaky, D., King, B.L., McMorran, R., Wiegers, J., Wiegers, T.C., Mattingly, C.J.: The comparative toxicogenomics database: update 2017. Nucleic Acids Res. **45**(D1), D972–D978 (2017)