Journal of Molecular Modeling (2020) 26: 60
https://doi.org/10.1007/s00894-020-4315-x

ORIGINAL PAPER

DDAPRED: a computational method for predicting drug repositioning using regularized logistic matrix factorization

Xiaofeng Wang[1] · Renxiang Yan[2,3]

Received: 25 December 2019 / Accepted: 28 January 2020 / Published online: 15 February 2020

Abstract

Due to rising development costs and stagnant product outputs of traditional drug discovery methods, drug repositioning, which discovers new indications for existing drugs, has attracted increasing interest. Computational drug repositioning can integrate prioritization information and accelerate time lines even further. However, most existing methods for predicting drug repositioning have low precisions. The present article proposed a new method named DDAPRED (https://github.com/nongdaxiaofeng/DDAPRED) for drug repositioning prediction. The method integrated multiple sources of drug similarity and disease similarity information, and it used the regularized logistic matrix decomposition method to significantly improve the prediction performance. In 5-fold cross-validation, the areas under the receiver operating characteristic curve (AUROC) and the precision-recall curve (AUPRC) of DDAPRED reached 0.932 and 0.438, respectively, exceeding other methods. The present study also analyzed the parameters influencing the model performance and the effect of different drug similarity information in-depth, and it verified the treatment relationship of the top 50 predictions with unknown relationships in the training set, further demonstrating the practicability of our method.

Keywords Drug-disease association · Logistic matrix factorization · Drug repositioning · Prediction

Introduction

Traditional drug discovery starts with the concept of a single target acting through a specific mechanism or a phenotypic screen in which the model system is screened for efficacious compounds [1]. Over the past few years, rising development costs and stagnant product outputs of traditional methods have become major reasons for the growing interest in other drug development methods, such as drug repositioning [2]. Drug repositioning discovers new indications for existing drugs, and it can expand the indications for drugs with safety data in human patients and reduce the time line by more than 10 years compared with traditional methods [3]. Drug repositioning has achieved many successes, including arsenic for acute promyelocytic leukemia [4], viagra for erectile dysfunction [3], and fish oil for Raynaud's syndrome [5].

Computational drug repositioning designs automated and systematic workflows to generate new indications for drug candidates, which can integrate prioritization information and accelerate time lines even further. Most computational methods used a "guilt by association" approach to discover new indications of existing drugs [6]. Based on the observation that similar drugs are indicated for similar diseases, the PREDICT model utilizes multiple drug-drug and disease-disease similarity measures, and it builds a logistic regression model to predict potential drug indications for either novel or approved drugs [7]. PreDR defines a kernel function of correlate drugs with diseases, and it predicts novel drug-disease interactions using a support vector machine (SVM) [8]. TL_HGBI uses an iterative algorithm on a heterogeneous network model, which incorporates drug target information, to infer drug repositioning [9]. Laplacian regularized sparse subspace learning (LRSSL) integrates drug chemical, drug target

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00894-020-4315-x) contains supplementary material, which is available to authorized users.

✉ Renxiang Yan
yanrenxiang@fzu.edu.cn

[1] College of Mathematics and Computer Science, Shanxi Normal University, Linfen 041004, China

[2] College of Biological Science and Engineering, Fuzhou University, Fuzhou 350106, Fujian, China

[3] Fujian Key Laboratory of Marine Enzyme Engineering, Fuzhou 350116, Fujian, China

domain information, and target annotations for indication prediction of drugs based on LRSSL [10]. SCMFDD employs a similarity constrained matrix factorization method for the drug-disease association prediction based on known drug-disease associations, drug features, and disease semantic information [11].

Although several methods for predicting drug repositioning have been elegantly developed, there is still room to improve the performance. LRSSL and SCMFDD are reported to perform better than previous methods [11]. However, their algorithms for them to converge are very time consuming. This article presents a new method named DDAPRED for drug repositioning prediction, which greatly improves the prediction performance compared with LRSSL and SCMFDD. In addition, it requires very little time (about 10 min) to train the DDAPRED model in a general personal computer, which is much less than that of LRSSL and SCMFDD. The DDAPRED model integrates multiple drug similarities and disease similarities, and it uses the regularized logistic matrix factorization. In 5-fold cross-validation, the area under the ROC curve (AUROC) and the precision-recall curve (AUPRC) of DDAPRED reached 0.9322 and 0.4377, respectively, which exceeded other methods. The parameters influencing the model performance and the importance of each item in the optimization model were analyzed. Fifty samples with the highest predictive scores were extracted from samples labeled as unknown relationships in the training dataset. The treatment relationships were verified by searching the literature and KEGG database [12], which further confirmed the prediction accuracy and practicability of DDAPRED.

Materials and methods

Dataset

To build and test the model, the dataset used by the LRSSL method was downloaded. The dataset contained 3051 treatment relationships involving 763 FDA-approved drugs and 681 diseases, which were represented by a binary adjacency matrix Y of 763 rows and 681 columns. If a drug and a disease had a treatment relationship, they were treated as associative. In the matrix Y, 1 represented known treatment relationships and 0 denoted unknown relationships. The rows of Y were named association profiles of the drugs, and the columns were named association profiles of the diseases. This dataset also included the disease semantic similarity matrix [13], drug chemical structure information, drug target domain information, and drug target gene ontology (GO) information [14]. Specifically, the three kinds of drug information were represented by three 763-row binary matrices, and each row elements represented whether the drug contains the corresponding chemical substructures, target domains, or target GO items. The row of the matrices was named drug profile. The numbers of the drugs and the diseases were denoted as m and n, respectively, in the present article.

Methods

The flowchart of DDAPRED is shown in Fig. 1. The flowchart consists of the following four steps to build the prediction model: (1) infer association profiles of novel drugs and novel diseases; (2) calculate and integrate similarities between drugs and similarities between diseases; (3) regularize similarities; and (4) build the optimization model.

Inferring association profiles for novel drugs and novel diseases

When cross-validation was performed, all of the association profile elements for some drugs or diseases may be zero, which was named novel drugs or novel diseases. The association profile of the novel drug was inferred using the weighted sum of profiles of five drugs that were most similar to it, which was computed using the following formula:

$$\text{profile}(i) = \sum_{k \in N_i} \frac{S(i,k)}{\sum_{k \in N_i} S(i,k)} \text{profile}(k) \quad (1)$$

where i is the index of the novel drug; N_i is the index set of the five most similar drugs to the novel drug; and $S(i, k)$ is the similarity score between drug_i and drug_k. The process to infer the association profile of the novel disease was similar. When inferring association profiles of novel diseases, the disease semantic similarity was used. When inferring association profiles of novel drugs, the integrated similarity of drug chemical structure profile, target GO profile, and target domain profile was used. The calculation of the similarity scores between drugs and between diseases is introduced in the next section.

Calculating drug or disease similarities

For the association profile, the Gauss kernel [15] was used to compute the similarities between drugs and between diseases with the following formula:

$$G(x,y) = \exp\left(-\frac{\|x-y\|_2^{\,2}}{\lambda}\right) \quad (2)$$

where x and y are two profiles; $\|\cdot\|_2$ is the 2-norm; and λ is the average of 2-norms of association profiles of drugs or diseases.

For other types of profiles, the Jaccard coefficient [16] was used to compute the similarity. The following formula of the Jaccard coefficient between profiles x and y was used:

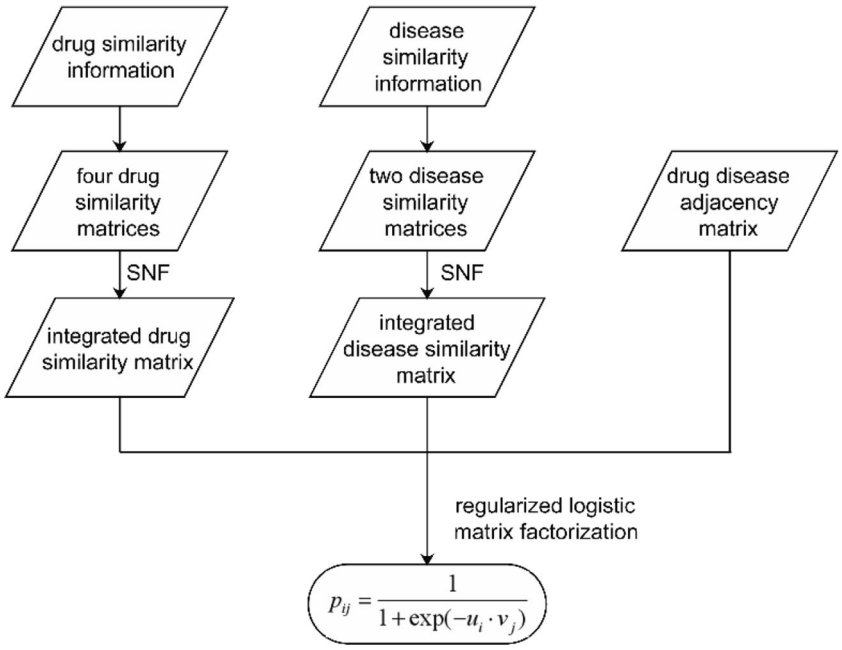

Fig. 1 The DDAPRED flowchart. Based on different drug and disease information, DDAPRED calculated four drug similarity matrices and two disease similarity matrices, and then integrated them into one drug similarity matrix and one disease similarity matrix using SNF method. Based on the drug similarity matrix, the disease similarity matrix, and the drug-disease adjacency matrix, the logistic matrix factorization mapped drug_i and disease_j into vectors u_i and v_j and their probability to be associated is p_{ij}

$$J(x,y) = \frac{x \cdot y}{\|x\|_1 + \|y\|_1 - x \cdot y} \tag{3}$$

where $\cdot$ is the dot product and $\|\cdot\|_1$ is the 1-norm.

Based on the drug chemical structure profile, target GO profile, target domain profile, and association profile, four similarity matrices for the drugs were obtained. Using the similar network fusion (SNF) method [17], these four matrices were further fused into one similarity matrix denoted as S^{drug}. Based on the disease association profile, a similarity matrix was obtained for the diseases. Using SNF method, this matrix was combined with the disease semantic similarity matrix, which was denoted as S^{disease}.

Regularizing the similarity matrix

To improve the prediction accuracy, two regularized similarity matrices, A and B, were created for drugs and diseases. The (i, j) element of A was defined as follows:

$$a_{ij} = \begin{cases} s_{ij}^{\text{drug}} & \text{if } \text{drug}_j \in N(\text{drug}_i) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $N(\text{drug}_i)$ is the set of the five most similar drugs to drug_i and s_{ij}^{drug} is the (i, j) element of S^{drug}. The (i, j) element of B was defined as follows:

$$b_{ij} = \begin{cases} s_{ij}^{\text{disease}} & \text{if } \text{disease}_j \in N(\text{disease}_i) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $N(\text{disease}_i)$ is the set of the five most similar diseases to disease_i and s_{ij}^{disease} is the (i, j) element of S^{disease}.

Building the optimization model

The objective function of the optimization model contained three parts. The first part used the logistic matrix factorization technique [18], which has been shown to be effective in drug-target interaction identification [19] and miRNA-disease association prediction [20]. The drugs and diseases were mapped to a shared latent space. For drug_i and disease_j, two vectors (u_i and v_j) of length r were used to represent them, where $r << \min(m, n)$. The probability for drug_i and disease_j to be associative was as follows:

$$p_{ij} = \frac{1}{1 + \exp(-u_i \cdot v_j)} \tag{6}$$

Supposing that all the relationships were independent in the training set, the occurrence probability for all the training samples was as follows:

$$P(Y) = \prod_{1 \le i \le m, 1 \le j \le n} p_{ij}^{y_{ij}} \left(1 - p_{ij}\right)^{1 - y_{ij}} \tag{7}$$

Finally, the objective function was as follows:

$$\min O_1 + c_1 O_2 + c_2 O_3 \tag{8}$$

where

$$O_1 = -\ln P(Y) \tag{9}$$

$$O_2 = \frac{1}{2}\left(\sum_{i=1}^{m} \|u_i\|_2^2 + \sum_{j=1}^{n} \|v_j\|_2^2\right) \tag{10}$$

$$O_3 = \frac{1}{2}\left(\sum_{i=1}^{m}\sum_{j=1}^{m} a_{ij}\left\|u_i - u_j\right\|_2^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}\left\|v_i - v_j\right\|_2^2\right) \quad (11)$$

In the objective function, O_1 minimized the negative log-likelihood function of the training samples, and O_2 avoided the occurrence of large elements in the latent vectors. In addition, O_3 made similar drugs or similar diseases have similar latent vectors. The c_1 and c_2 parameters balanced O_1, O_2, and O_3.

An alternating gradient descent method was used to solve the optimization model, the detailing algorithm of which has been clearly described in Liu's paper [19]. After solving the optimization model, the predicted probability for drug$_i$ and disease$_j$ to be associative was obtained by Eq. (5).

Performance evaluation

To estimate the performance of the method, k-fold cross-validation [21] was used. When performing k-fold cross-validation, the training dataset was randomly split into k equal subsets. For each subset, the ratio of positive samples (known associative drug-disease pairs) to negative samples (drug-disease pairs without known relationships) was the same as that of the whole dataset. Of the k subsets, a single subset was retained as the test data, and the remaining k-1 subsets were retained as training data. This process was repeated k times with each of the subsets used exactly once as the test data.

The ROC curve [22] and the precision-recall curve [23] were used to measure the prediction performance. The ROC curve plotted the true positive rate (TPR) against the false positive rate (FPR) at different cutoffs. The precision-recall curve plotted the precision against the recall, which is better than the ROC curve when evaluating binary classifiers on imbalanced datasets [24]. The area under the ROC curve (AUROC) and the area under the precision-recall curve (AUPRC) were also used to measure the prediction performance. The following formulas for TPR (recall), FPR, and precision were used:

$$\mathrm{TPR} = \mathrm{recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \quad (12)$$

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \quad (13)$$

$$\mathrm{precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (14)$$

where TP, FP, TN, and FN are the numbers of true positives, false positives, true negatives, and false negatives, respectively.

Results

The dataset used by the LRSSL method was employed to build and benchmark the DDAPRED method. The workflow of DDAPRED is clearly illustrated in Fig. 1. The aim of DDAPRED was to discover new treatment relationships between the drugs and diseases in the training set. The basic idea of DDAPRED was to embed all drugs and diseases into a latent vector space, in which similar drugs or diseases had similar latent vectors, and to minimize the negative log likelihood of training samples. The performance of DDAPRED was demonstrated with a 5-fold cross-validation, and the efficacy of different drug similarities and the impact of model parameters on model performance were analyzed. Literature and KEGG database searches verified the treatment relationship of the 50 top predictions whose relationships in the training set were marked as unknown.

Performance on 5-fold cross-validation

To obtain a better prediction performance of the present method, the following parameters were set: $c_1 = 0.1$, $c_2 = 30$, and $r = 25$. Fivefold cross-validation was used to evaluate the prediction model. To validate the prediction performance of DDAPRED, DDAPRED was compared with four recently developed methods for predicting drug-target interactions and drug-disease associations, including CMF [25], DTINET [26], LRSSL, and SCMFDD.

CMF and DTINET are methods for predicting drug-target interactions. The source codes of the two methods were downloaded, and the models for predicting drug-disease associations were established. Figure 2 shows the ROC curves and precision-recall curves of CMF, DTINET, and DDAPRED. When the false positive rate was low, the ROC curve of DDAPRED was above the curves of other two methods, and the precision-recall curve of DDAPRED was also above the other two curves. These findings indicated that DDAPRED had fewer false positive errors when the same number of correct treatment relationships was predicted, which is important for experimental scientists because they usually only validate predicted treatment relationships with high scores. DDAPRED also obtained higher AUROC and AUPRC than CMF and DTINET (Table 1), further demonstrating the superiority of DDAPRED.

LRSSL and SCMFDD are two drug-disease association prediction methods, which have been reported to perform better than previous methods. Because running LRSSL and SCMFDD programs was too time consuming, only the AUROC and AUPRC values of the two methods are listed in Table 1. These values were previously computed by Zhang et al. based on 5-fold cross-validation on the LRSSL dataset [11]. The AUC value of DDAPRED was slightly greater than that of SCMFDD but much greater than that of LRSSL.

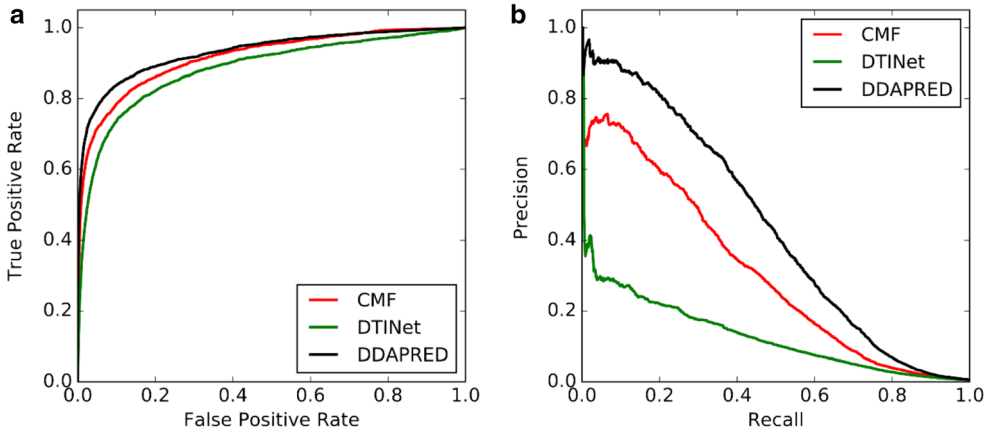

Fig. 2 ROC curves and precision-recall curves of different methods with 5-fold cross-validation. **a** ROC curves. **b** Precision-recall curves

Moreover, the AUPRC of DDAPRED was much greater than that of LRSSL and SCMFDD, demonstrating the better prediction performance of DDAPRED. It should be mentioned that LRSSL performed a 5-fold cross-validation, obtaining an AUROC of 0.908 after removing drugs with only one indication in the training set, but this value was still lower than that of DDAPRED.

Efficacy of drug similarities for prediction

To establish the prediction model, four types of drug similarities were integrated, and they were calculated using four different types of information. To compare the efficacy of the four similarities, each type of similarity was used to build the model, and the model parameters were adjusted to make their performance as good as possible. Table 2 lists the AUROC and AUPRC values of the four models after 5-fold cross-validation based on the four similarities. The AUROC values of the four models were all above 0.89, and the AUPRC values were all above 0.37, indicating the efficacy of the four similarities in modeling. The models based on the target information were the best, which was consistent with the results of LRSSL and SCMFDD. It should be noted that the performance of the model was also related to the quality of the information used to calculate the similarity.

Analysis of model parameters

When compared with other models, the three parameters in the model were set as follows: $c_1 = 0.1$, $c_2 = 30$, and $r = 25$. By changing one parameter and fixing the other two parameters, the influence of the parameters on the model performance was studied. Figure 3 shows the curves of AUROC and AUPRC as the parameters changed with 5-fold cross-validation.

The c_1 parameter is the coefficient of O_2, and O_2 is designed to prevent excessive elements occurring in the latent vectors for drugs and diseases. Figure 3a and b show that when c_1 was small, the performance of the model was better. With the increase of c_1, the performance of the model gradually decreased. When c_1 was 0, that is, the objective function did not contain O_2, the performance of the model changed very little compared with that when $c_1 = 0.1$, which may indicate that O_2 is not important in the model.

The c_2 parameter is the coefficient of O_3. In the objective function, O_3 made similar drugs or diseases have similar latent vectors. Figure 3c and d show that when c_2 increases from 0 to 40, AUROC kept increasing. In contrast, AUPRC increased until c_2 reached 25, and then, it slowly decreased. When c_2 was 0, that is, when O_3 was not included in the objective function, the performance of the model worsened, and the AUROC and AUPRC decreased to approximately 0.80 and 0.22 respectively. These findings showed that the O_3 item was important for predicting the performance of the model. It was difficult to achieve good prediction performance when relying solely on logistic matrix decomposition.

The r parameter is the length of the latent vector, which was used to represent the drugs and diseases in the latent space. As r increased, the performance of the model gradually improved, but when r increased to 20 and 25, the AUROC and AUPRC stopped increasing.

Analysis of the influence of parameters on the performance of the model identified the importance of each item in the model, and it showed that by adjusting the parameters of the model, a better prediction performance was achieved.

Table 1 Performance of different methods with 5-fold cross-validation

Method	CMF	DTINET	LRSSL	SCMFDD	DDAPRED
AUROC	0.916	0.883	0.825	0.922	0.932
AUPRC	0.307	0.129	0.178	0.267	0.437

Validation of top predictions

Experimental scientists are interested in drug-disease pairs that have high prediction probability to be associative. To demonstrate the usefulness of the present method, the 50 top

Table 2 Prediction performance using single drug similarities

Information	Chemical structure	Target GO	Target domain	Associated diseases
AUROC	0.906	0.924	0.923	0.891
AUPRC	0.401	0.417	0.411	0.378

predictions without known relationships in the training dataset were extracted. Searching the literature and KEGG database validated the potential treatment relationship of the 50 top predictions, and the results are listed in Table S1. The first three predicted treatment relationships were as follows: enalapril and hypertension, ceftriaxone and haemophilus infections, and ampicillin and streptococcal infections. Enalapril protects the function of the kidneys in hypertension, and it may be used in the absence of hypertension [27]. As both ceftriaxone and ampicillin are antibiotics, they can be used to prevent and treat many bacterial infections, including haemophilus and streptococcal infections [28, 29].

Discussion

Compared with other methods, the advantage of DDAPRED is that the AUPRC is much higher than those of other methods; this advantage is useful in practice. According to the precision-recall curves, when the recall value was 0.4, the precision of DDAPRED was 0.57, and the precision of CMF was 0.34. To experimentally identify 40 real treatment relationships, $40/0.34 \approx 118$ CMF top predictions need to be validated, while $40/0.57 \approx 70$ DDAPRED top predictions need to be validated. Thus, the number of DDAPRED predictions requiring verification was much less than that of CMF predictions, which will save time and money.

In addition to the drug information used in DDAPRED, many other drug- and disease-related information can be integrated to infer new drug indications, such as drug-miRNA associations, drug response, and miRNA-disease associations, which may increase the prediction accuracy.

There are several similar problems to drug repositioning prediction, such as drug-target interaction prediction [30] and miRNA-disease association prediction [31]. Methods for these problems can be applied to the drug repositioning

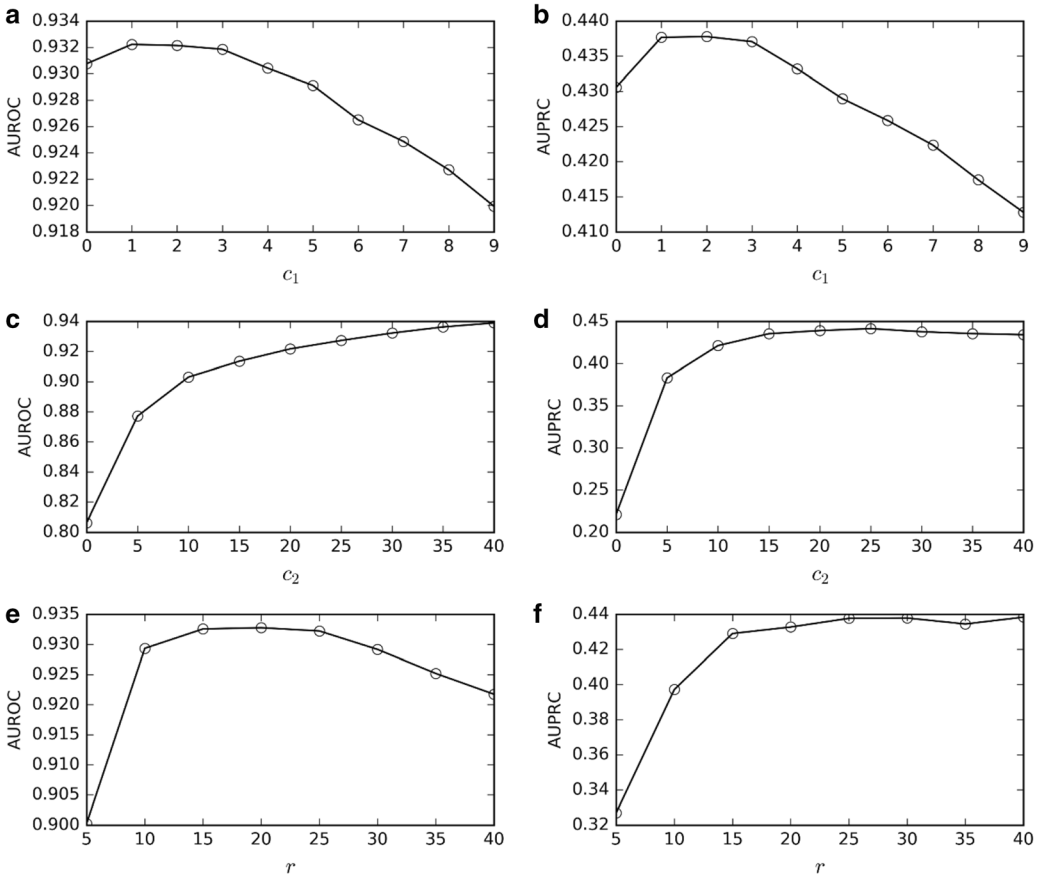

Fig. 3 AUROC and AUPRC as parameters changed. **a**, **c**, and **e** AUROC versus c_1, c_2, and r, respectively. **b**, **d**, and **f** AUPRC versus c_1, c_2, and r, respectively

prediction, which may improve the prediction performance. For example, CMF was developed for drug-target interaction prediction. However, CMF is competitive compared with LRSSL and SCMFDD at drug repositioning prediction.

In recent years, many excellent learning methods have emerged in the field of network representation learning, which embed the network nodes into latent vectors, such as DeepWalk [32], node2vec [33], and graph convolutional networks (GCN) [34]. Introducing these methods into drug repositioning prediction may also promote the development of this field.

Conclusions

In conclusion, a novel computational method named DDAPRED was developed for drug repositioning prediction. This method integrated different types of drug similarities and disease similarities, and it used logistic matrix factorization with regularization. DDAPRED was demonstrated to be excellent in prediction performance compared with other methods. By using each type of drug similarity to build the model, the efficacy of the four similarities for prediction was proved. The model parameter analysis illustrated that the neighborhood regularization item was very important as it greatly improved the AUROC and AUPRC. The 50 top DDAPRED predictions without known relationships in the training dataset were validated to have potential treatment relationships through searching the literature and KEGG database, which further demonstrates the usefulness of the present method.

Funding information This work was supported by the National Natural Science Foundation of China (31500673 and 31571300).

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Hurle MR, Yang L, Xie Q, Rajpal DK, Sanseau P, Agarwal P (2013) Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther 93(4):335–341
2. Sleigh SH, Barton CL (2010) Repurposing strategies for therapeutics. Pharmaceut Med 24(3):151–159
3. Ashburn TT, Thor KB (2004) Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov 3: 673
4. Soignet SL, Maslak P, Wang Z-G, Jhanwar S, Calleja E, Dardashti LJ, Corso D, DeBlasio A, Gabrilove J, Scheinberg DA et al (1998) Complete remission after treatment of acute promyelocytic leukemia with arsenic trioxide. New Engl J Med 339(19):1341–1348
5. Swanson DR (1990) Medical literature as a potential source of new knowledge. Bull Med Libr Assoc 78(1):29–37
6. Chang A, Butte A (2009) Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther 86(5):507–510
7. Gottlieb A, Stein GY, Ruppin E, Sharan R (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 7(1):496
8. Wang Y, Chen S, Deng N, Wang Y (2013) Drug repositioning by kernel-based integration of molecular structure, molecular activity, and phenotype data. PLoS One 8(11):e78518
9. Wang W, Yang S, Zhang X, Li J (2014) Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics 30(20):2923–2930
10. Liang X, Zhang P, Yan L, Fu Y, Peng F, Qu L, Shao M, Chen Y, Chen Z (2017) LRSSL: predict and interpret drug–disease associations based on data integration using sparse subspace learning. Bioinformatics 33(8):1187–1196
11. Zhang W, Yue X, Lin W, Wu W, Liu R, Huang F, Liu F (2018) Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinformatics 19(1):233
12. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45(D1):D353–D361
13. Mathur S, Dinakarpandian D (2012) Finding disease similarity based on implicit semantic similarity. J Biomed Inform 45(2): 363–371
14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25
15. van Laarhoven T, Nabuurs SB, Marchiori E (2011) Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 27(21):3036–3043
16. Jaccard P (1912) The distribution of the flora in the alpine zone.1. New Phytol 11(2):37–50
17. Wang B, Mezlini AM, Demir F, Fiume M, Tu Z, Brudno M, Haibe-Kains B, Goldenberg A (2014) Similarity network fusion for aggregating data types on a genomic scale. Nat Methods 11:333
18. Ban T, Ohue M, Akiyama Y (2019) NRLMFβ: Beta-distribution-rescored neighborhood regularized logistic matrix factorization for improving the performance of drug–target interaction prediction. Biochem Biophys Rep 18:100615
19. Liu Y, Wu M, Miao C, Zhao P, Li X-L (2016) Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS Comp Biol 12(2):e1004760
20. He B-S, Qu J, Zhao Q (2018) Identifying and exploiting potential miRNA-disease associations with neighborhood regularized logistic matrix factorization. Front Genet 9:303
21. Wang X, Yan R, Li J, Song J (2016) SOHPRED: a new bioinformatics tool for the characterization and prediction of human S-sulfenylation sites. Mol BioSyst 12(9):2849–2858
22. Fan J, Upadhye S, Worster A (2015) Understanding receiver operating characteristic (ROC) curves. CJEM 8(1):19–20
23. Boyd K, Eng KH, Page CD (2013) Area under the precision-recall curve: point estimates and confidence intervals. In: Blockeel H. et al (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2013. Lecture Notes in Computer Science, pp 451–466
24. Saito T, Rehmsmeier M (2015) The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10(3):e0118432
25. Zheng X, Ding H, Mamitsuka H, Zhu S (2013) 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. Chicago, Illinois, USA, p 1025–1033. 2487670: ACM

26. Luo Y, Zhao X, Zhou J, Yang J, Zhang Y, Kuang W, Peng J, Chen L, Zeng J (2017) A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. Nat Commun 8(1):573
27. McMurray JJ (2010) Systolic heart failure. N Engl J Med 362(3): 228–238
28. Richards DM, Heel RC, Brogden RN, Speight TM, Avery GS (1984) Ceftriaxone. Drugs 27(6):469–527
29. De Cueto M, Sanchez M-J, Sampedro A, Miranda J-A, Herruzo A-J, Rosa-Fraile M (1998) Timing of intrapartum ampicillin and prevention of vertical transmission of group B streptococcus. Obstet Gynecol 91(1):112–114
30. Chen X, Yan CC, Zhang X, Zhang X, Dai F, Yin J, Zhang Y (2016) Drug-target interaction prediction: databases, web servers and computational models. Brief Bioinform 17(4):696–712
31. Liu Y, Zeng X, He Z, Zou Q (2017) Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans Comput Biol Bioinform 14(4):905–915
32. Perozzi B, Al-Rfou R, Skiena S (2014) DeepWalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York, New York, USA, p 701–710. 2623732: ACM
33. Grover A, Leskovec J (2016) node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA, p 855–864. 2939754: ACM
34. Schlichtkrull M, Kipf TN, Bloem P, Berg R, Titov I, Welling M (2018) Modeling relational data with graph convolutional networks. In: Gangemi A. et al. (eds) The Semantic Web. ESWC 2018. Lecture Notes in Computer Science, pp 593–607