Journal of Translational Medicine

RESEARCH

Open Access

A weighted non-negative matrix factorization approach to predict potential associations between drug and disease

Mei-Neng Wang[1], Xue-Jun Xie[1], Zhu-Hong You[2*], De-Wu Ding[1*] and Leon Wong[3,4]

Abstract

Background: Associations of drugs with diseases provide important information for expediting drug development. Due to the number of known drug-disease associations is still insufficient, and considering that inferring associations between them through traditional in vitro experiments is time-consuming and costly. Therefore, more accurate and reliable computational methods urgent need to be developed to predict potential associations of drugs with diseases.

Methods: In this study, we present the model called weighted graph regularized collaborative non-negative matrix factorization for drug-disease association prediction (WNMFDDA). More specifically, we first calculated the drug similarity and disease similarity based on the chemical structures of drugs and medical description information of diseases, respectively. Then, to extend the model to work for new drugs and diseases, weighted *K* nearest neighbor was used as a preprocessing step to reconstruct the interaction score profiles of drugs with diseases. Finally, a graph regularized non-negative matrix factorization model was used to identify potential associations between drug and disease.

Results: During the cross-validation process, WNMFDDA achieved the AUC values of 0.939 and 0.952 on Fdataset and Cdataset under ten-fold cross validation, respectively, which outperforms other competing prediction methods. Moreover, case studies for several drugs and diseases were carried out to further verify the predictive performance of WNMFDDA. As a result, 13(Doxorubicin), 13(Amiodarone), 12(Obesity) and 12(Asthma) of the top 15 corresponding candidate diseases or drugs were confirmed by existing databases.

Conclusions: The experimental results adequately demonstrated that WNMFDDA is a very effective method for drug-disease association prediction. We believe that WNMFDDA is helpful for relevant biomedical researchers in follow-up studies.

Keywords: Drug-disease association, Weighted nearest neighbor, Graph regularization, Non-negative matrix factorization

Background

In the past few decades, people have made remarkable progress in life sciences and genomics. However, the development of a new drug is still a high-risky, tremendously expensive and time-consuming process [1, 2]. On average, it takes about 15 years and costs more than $ 800 million to discover and bring a new drug to the market [3, 4]. Although tremendous investment in new drugs

*Correspondence: zhuhongyou@nwpu.edu.cn; dwding2008@aliyun.com

[1] School of Mathematics and Computer Science, Yichun University, Yichun 336000, Jiangxi, China
[2] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
Full list of author information is available at the end of the article

design and discovery, the number of new drugs authorized by the U.S. Food and Drug Administration (FDA) has remained low since the 1990s [5, 6]. About 90% new drugs designed for specific diseases fail the first phase of clinical trials, which means that new drugs design and discovery are becoming more and more costly [7]. In light of these challenges, repositioning of already commercialized drugs, which aims to identify and discover the new therapeutic uses for these drugs, is attracting strong increasing interests from the biomedical researchers and pharmaceutical companies [8]. Since existing drugs have been proven safe through various clinical trials, drug repositioning can lower risk, shorten the process of drug development, and are more likely to be approved by regulatory authorities [9]. Therefore, drug repositioning plays an important role in drug research and development. Nowadays, some existing drugs (e.g. Minoxidil, Thalidomide, Sildenafil) have been successfully repositioned in clinical trials, which have saved new drug development costs and created great economic value for related pharmaceutical companies [10]. For example, Minoxidil, originally commercialized to prevent high blood pressure, was repositioned to treat the androgenic alopecia; Thalidomide was marketed to use as a sedative, it was later repurposed as a treatment to insomnia and nausea [11, 12]. Compared with the development of a novel drug for specific indications, drug repositioning costs only about $ 300 million and can shorten the drug development cycle by more than half [10, 13]. To this end, more and more existing drugs are being repurposed to treat diseases other than those originally intended [14].

In fact, drug repositioning can be seen as identifying the associations between drug and disease. Although some associations of drugs with diseases have been verified in clinical trials, many of them are still undiscovered. In recent years, some computational approaches have been developed to infer associations between drug and disease for drug repositioning, such as semantic inference [1], network analysis [15], text mining [16] and machine learning [17], etc. For example, Napolitano et al. trained a multi-class Support Vector Machine (SVM) classifier based on drugs similarities to identify potential drug indications [18]. Gottlieb et al. constructed classification features by integrating disease similarities and drug similarities, and scored the new associations of drugs with diseases to predict novel therapeutic indications by implementing a logistic regression classification algorithm [19]. Based on the hypothesis that different diseases with similar treatments can be treated with similar drugs, Chiang et al. developed a "guilt-by-association" principle approach to infer potential relationships between drug and disease [20]. Yang et al. developed a causal network linking drug-target-pathway-gene-disease to calculate association scores of drugs with diseases. Based on known drug-disease associations, a probabilistic matrix factorization model is learned to classify drug–disease associations, and novel associations of drugs with diseases were predicted according to the calculated association scores and association types [21]. However, these methods fail to predict associations of novel drugs without any known related disease.

At present, with the generation of large-scale high-throughput biological data, researchers are increasingly concerned how to establish complex biomolecular interaction networks for predicting their associations. Martínez et al. have developed a novel model, DrugNet, to infer new treatments for diseases and novel therapeutic indications for drugs [22]. This method predicts drug-disease potential associations by prioritizing based on a heterogeneous network which was integrated biological information about drugs, targets and diseases. Wang et al. proposed three-layer heterogeneous network-based computational method named TL-HGBI, which performs drug repositioning by applying known drug-disease associations and drug, disease and target similarities [23]. Luo et al. presented a new prediction model MBiRW, which utilized Bi-Random walk algorithm to infer new drug indications based on the assumption that similar drugs tend to be associated with the different diseases that with similar treatments [24].

In fact, predicting novel indications for existing drugs can be considered as a recommendation system problem. Recently, recommendation system models have been used to predict associations between biomolecules (e.g. drug-target interactions, circRNA-disease associations) [25, 26]. Luo et al. developed a drug repositioning recommendation system (DRRS) to infer new indications for existing drugs, which used fast Singular Value Thresholding (SVT) algorithm to complete the association adjacency matrix of drug with disease [27]. Similar to finding missing interactions in an adjacency matrix, matrix factorization is well applied in collaborative filtering recommendation algorithms [28]. Recent studies have shown that matrix factorization technique has been successfully used in recommender system and link prediction for data representation [29, 30], especially in the field of bioinformatics [31–33]. Inspired by these, we can view the drug-disease association prediction problem as a recommender system task and used matrix factorization to predict.

In this paper, we propose a new computational method named WNMFDDA to infer the unknown associations of drugs with diseases, which is based on weighted graph regularized collaborative non-negative matrix factorization. Distinct from previous methods, graph Laplacian regularization is introduced to prevent overfitting, which

can ensure close drugs or diseases are sufficiently close to each other in the corresponding latent feature space; Tikhonov (L_2) is used to guarantee that the solution obtained from matrix factorization is smooth. In addition, in order to extend our model to work for new drugs (or new diseases) and reduce the impact of sparse associations on prediction performance, weighted K-nearest neighbor is utilized to rebuild the association adjacency matrix between drug and disease before performing matrix factorization. We carry out ten-fold cross validation to verify the performance of WNMFDDA and compared it with several classical models. The experimental results of cross validation show that WNMFDDA obtains better performance than other compared models. Case studies on drugs and diseases also demonstrate that our proposed approach is reliable in identifying drug-disease potential associations.

Methods and materials

Method overview

To identify potential associations between drug and disease, we propose a new computational model named WNMFDDA. The proposed method mainly process (See Fig. 1) contains three steps: (i) We measure the drug similarity and disease similarity based on chemical structures of drugs and medical description information of diseases, respectively. (ii) To extend WNMFDDA to predicting novel diseases and drugs, the adjacency matrix of drug with disease is reformulated based on weighted K-nearest neighbor profiles of drug and disease. (iii) Graph regularized collaborative matrix factorization is performed on the updated adjacency matrix to obtain the final score matrix.

Datasets

The dataset (Fdataset) used in this work was obtained from Gottlieb et al. [19], which is comprised multiple data sources, and is considered as the golden standard datasets for predicting potential associations between drug and disease. After deleting the duplicate association pairs, a total of 1933 experimentally verified associations between 593 drugs and 313 diseases are collected for prediction. Diseases and drugs are obtained from the Online Mendelian Inheritance in Man (OMIM) database [34] and DrugBank database [35], respectively. Here, we construct the drug-disease association adjacency matrix $Y^{n\times m}$ based on the known associations, n is the number of drugs and m is the number of diseases. Let $R=\{r_1,r_2,\cdots,r_n\}$ and $D=\{d_1,d_2,\cdots,d_m\}$ represent the set of n drugs and m diseases. In the original adjacency matrix $Y\in R^{n\times m}$, the value of $Y(i,j)$ is set 1 if drug r_i relates with disease d_j, otherwise it is 0. Finally, the original adjacency matrix $Y\in R^{593\times 313}$, the drug similarity matrix and disease similarity matrix are used to identify the associations of drugs with diseases based on WNMFDDA.

Similarity for drugs and diseases

In this work, the drug similarity matrix is denoted by $S^R\in R^{593\times 593}$. we calculate the drug-drug similarity using the Chemical Development Kit (CDK) [36] based on Simplified Molecular Input Line Entry Specification (SMILES) chemical structures [37], and the Tanimoto score of their 2D chemical fingerprints is used as representing the pair of drug similarity [38].

The disease similarity matrix is denoted by $S^D\in R^{313\times 313}$. The similarities between diseases are derived from MimMiner [39], which measures the pairwise disease semantic similarity through text mining based on the medical description information in the OMIM database [34].

Weighted graph regularized collaborative non-negative matrix factorization for predicting drug-disease associations

Reformulate association adjacency matrix of drug with disease

Due to many of non-interactions of drugs or diseases in the original adjacency matrix (i.e. their values are 0 in Y) that could be potential true interactions, which may lead to poor performance in predicting the potential drug-disease associations. In order to solve the above mentioned problem, we perform weighted K-nearest neighbor (WKNN) profiles to construct novel interaction profiles of drug and disease.

For each drug r_p, we sort all other drugs in descending order according to their similarities with r_p. Then, the new interaction profile of drug r_p is obtained according to its K -nearest known drugs (each drug has at least one confirmed association), and their corresponding K interaction profiles are as follows:

$$Y_r(r_p)=\frac{1}{\sum_{1\le i\le K}S^R(r_i,r_p)}\sum_{i=1}^{K}w_iY(r_i) \tag{1}$$

where

$$w_i=a^{i-1}*S^R(r_i,r_p) \tag{2}$$

$a\in[0,1]$ isadecayterm. w_i is a weight coefficient, it means that the more similar r_i to r_p, the larger weight is assigned. $Y(r_i)=(Y_{i1},Y_{i2},\cdots,Y_{im})$ denotes the interaction profile for drug r_i, which is the *ith* row vector of adjacency matrix Y.

Similar to drugs, for each disease d_q, the new interaction profiles of disease d_q can be calculated as follows:

$$Y_d(d_q) = \frac{1}{\sum_{1\le j\le K} S^D(d_j,d_q)} \sum_{j=1}^{K} w_j Y(d_j) \tag{3}$$

$$w_j = a^{j-1} * S^D(d_j,d_q) \tag{4}$$

where, w_i is a weight coefficient. $Y(d_j) = (Y_{1j}, Y_{2j}, \cdots, Y_{nj})$ represents the interaction profile for disease d_j, which is the *jth* column vector of adjacency matrix Y.

Thereafter, we merge the new interaction profiles of drug and disease by $Y_{rd} = (Y_r + Y_d)/2$. Finally, the original adjacency matrix Y is updated by replacing $Y_{ij} = 0$ with related likelihood score as follows:

$$Y = \max(Y, Y_{rd}) \tag{5}$$

The model of WNMFDDA

Non-negative matrix factorization (NMF) is one of the most popular multidimensional data processing tools in research fields such as bioinformatics and pattern recognition [40–42]. The purpose of NMF is to decompose a nonnegative matrix Y into two low-dimensional nonnegative matrices, and makes their product approximation to the original matrix Y. Therefore, for drug-disease adjacency matrix $Y^{n\times m}$, it can be decomposed into two low-rank feature matrices, $A^{k\times n}$ and $B^{k\times m}$, and $Y \cong A^T B (k \le \min(n,m))$. The objective function for predicting drug-disease associations can be mathematically formulated as follows:

$$\min_{A,B} \|Y - A^T B\|_F^2 \qquad s.t. A \ge 0, B \ge 0 \tag{6}$$

where $\|\bullet\|_F$ denotes the Frobenius norm. To enhance generalization capability and solve the problem that the standard NMF in formula (6) fails to discover the intrinsic geometrical of drug space and disease space, we introduce Laplacian regularization to constrain non-negative matrix factorization which can ensure that close drugs or diseases are sufficiently close to each other in corresponding latent feature space. The optimization problem can be written as follows:

$$\min_{A,B} \|Y - A^T B\|_F^2 + \lambda \left(\sum_{i\le j}^{n} \|a_i - a_j\|^2 S_{ij}^R + \sum_{i\le j}^{m} \|b_i - b_j\|^2 S_{ij}^D \right) s.t. A \ge 0, B \ge 0 \tag{7}$$

where $R_1 = \sum_{i\le j}^{n} \|a_i - a_j\|^2 S_{ij}^R$ and $R_2 = \sum_{i\le j}^{m} \|b_i - b_j\|^2 S_{ij}^D$ are the Laplacian regularization terms. a_i and b_i are *ith* column of matrices A and B, respectively. λ is the regularization parameter.

Recent studies on manifold learning theory and spectral graph theory have shown that the local geometric structure and topological structure of original data points can be leaved unchanged by the p-nearest neighbor graph when these points are mapped from high-dimensional space to low-dimensional space [43, 44]. In addition, drugs and diseases in the same cluster are more possible to have similar characteristics, and the sparse similarity matrix has been effectively applied to graph regularization [45]. As a graph clustering method, p-nearest neighbor is used to construct the graphs (S^{R*} and S^{D*}) for drug space and disease space. Therefore, we can obtain the following weight matrix W^R of drug according to the drug similarity matrix S^R:

$$W_{ij}^R = \begin{cases} 1, & i \in N_p(r_j) \& j \in N_p(r_i) \\ 0, & i \notin N_p(r_j) \& j \notin N_p(r_i) \\ 0.5, & otherwise \end{cases} \tag{8}$$

Here, $N_p(r_i)$ and $N_p(r_j)$ represent the sets of p-nearest neighbors of drug r_i and drug r_j. Then, the graph matrix S^{R*} for drugs is defined as follows:

$$\forall i,j S_{ij}^{R*} = S_{ij}^R W_{ij}^R \tag{9}$$

Similarly, based on the disease similarity matrix S^D, the graph matrix S^{D*} for diseases is determined by:

$$\forall i,j S_{ij}^{D*} = S_{ij}^D W_{ij}^D \tag{10}$$

Then, the optimization problem is formularized as follows:

$$\min_{A,B} \|Y - A^T B\|_F^2 + \lambda \left(\sum_{i\le j}^{n} \|a_i - a_j\|^2 S_{ij}^{R*} + \sum_{i\le j}^{m} \|b_i - b_j\|^2 S_{ij}^{D*} \right) \qquad s.t. A \ge 0, B \ge 0 \tag{11}$$

where $R_1^* = \sum_{i\le j}^{n} \|a_i - a_j\|^2 S_{ij}^{R*}$ and $R_2^* = \sum_{i\le j}^{m} \|b_i - b_j\|^2 S_{ij}^{D*}$ are the graph Laplacian regularization terms. In order to avoid overfitting and guarantee the A and B smoothness, Tikhonov (L_2) regularization terms are incorporated into the Eq. (11) [46]. Finally, the optimization problem of WNMFDDA can be transformed into:

$$\begin{aligned}&\min_{A,B}\|Y-A^TB\|_F^2+\lambda\left(\sum_{i\le j}^{n}\|a_i-a_j\|^2S_{ij}^{R*}\right.\\&\left.+\sum_{i\le j}^{m}\|b_i-b_j\|^2S_{ij}^{D*}\right)+\beta\left(\|A\|_F^2+\|B\|_F^2\right)\\&s.t.A\ge 0,B\ge 0\end{aligned}\tag{12}$$

and

$$\begin{aligned}\mathcal{L}_f=&Tr\left(YY^T\right)-2Tr\left(YB^TA\right)+Tr\left(A^TBB^TA\right)\\&+\lambda Tr\left(AL_rA^T\right)+\lambda Tr\left(BL_dB^T\right)\\&+\beta Tr\left(AA^T\right)+\beta Tr\left(BB^T\right)+Tr\left(\Phi A^T\right)\\&+Tr\left(\Psi B^T\right)\end{aligned}\tag{16}$$

The partial derivatives of $\mathcal{L}_f$ to A and B are as follows:

$$\sum_{i\le j}^{n}\|a_i-a_j\|^2S_{ij}^{R*}=\sum_{j=1}^{n}a_j^Ta_j\sum_{i,j=1}^{n}S_{ij}^{R*}-\sum_{i,j=1}^{n}a_i^Ta_jS_{ij}^{R*}=Tr\left(AD_rA^T\right)-Tr\left(AS^{R*}A^T\right)=Tr\left(AL_rA^T\right)\tag{13}$$

$$\begin{aligned}\sum_{i\le j}^{m}\|b_i-b_j\|^2S_{ij}^{D*}=&Tr\left(BD_dB^T\right)-Tr\left(BS^{D*}B^T\right)\\=&Tr\left(BL_dB^T\right)\end{aligned}\tag{14}$$

where β is the regularization parameter. $Tr(\bullet)$ is the trace of a matrix. $D_r=\sum_{i=1}^{n}S_{ij}^{R*}$ and $D_d=\sum_{i=1}^{m}S_{ij}^{D*}$ are the diagonal matrices; $L_r=D_r-S^{R*}$ and $L_d=D_d-S^{D*}$ denote the graph Laplacian matrices with respect to S^{R*} and S^{D*} [47]. The Eq. (12) can be rewritten as:

$$\begin{aligned}&\min_{A,B}\|Y-A^TB\|_F^2+\lambda\left(\sum_{i\le j}^{n}\|a_i-a_j\|^2S_{ij}^{R*}\right.\\&\left.+\sum_{i\le j}^{m}\|b_i-b_j\|^2S_{ij}^{D*}\right)+\beta\left(\|A\|_F^2+\|B\|_F^2\right)\\&=Tr\left(YY^T\right)-2Tr\left(YB^TA\right)\\&\quad+Tr\left(A^TBB^TA\right)\\&\quad+\lambda Tr\left(AL_rA^T\right)+\lambda Tr\left(BL_dB^T\right)\\&\quad+\beta Tr\left(AA^T\right)+\beta Tr\left(BB^T\right)\end{aligned}\tag{15}$$

Optimization algorithm

In this work, the optimization problem of objective function Eq. (15) is solved by using Lagrange multipliers method. We introduce Lagrange multipliers $\Phi=\{\phi_{ki}\}$ and $\Psi=\{\psi_{kj}\}$ to constrain $a_{ki}\ge 0$ and $b_{kj}\ge 0$, respectively. The corresponding Lagrange function $\mathcal{L}_f$ of Eq. (15) is represented as follows:

$$\frac{\partial\mathcal{L}_f}{\partial A}=-2BY^T+2BB^TA+2\lambda AL_r+2\beta A+\Phi\tag{17}$$

$$\frac{\partial\mathcal{L}_f}{\partial B}=-2AY+2AA^TB+2\lambda BL_d+2\beta B+\Psi\tag{18}$$

The Karush–Kuhn–Tucker (KKT) constraint conditions $\phi_{ki}a_{ki}=0$ and $\psi_{kj}b_{kj}=0$ are used in the following equations for a_{ki} and b_{kj} [48]:

$$\begin{aligned}&-\left(BY^T\right)_{ki}a_{ki}+\left(BB^TA\right)_{ki}a_{ki}\\&+\left[\lambda A\left(D_r-S^{R*}\right)\right]_{ki}a_{ki}+(\beta A)_{ki}a_{ki}=0\end{aligned}\tag{19}$$

$$\begin{aligned}&-(AY)_{kj}b_{kj}+\left(AA^TB\right)_{kj}b_{kj}\\&+\left[\lambda B\left(D_d-S^{D*}\right)\right]_{kj}b_{kj}+(\beta B)_{kj}b_{kj}=0\end{aligned}\tag{20}$$

Finally, the updating rules for a_{ki} and b_{kj} can be determined as follows:

$$a_{ki}\leftarrow a_{ki}\frac{BY^T+\lambda AS^{R*}}{\beta A+\lambda AD_r+BB^TA}\tag{21}$$

$$b_{kj}\leftarrow b_{kj}\frac{AY+\lambda BS^{D*}}{\beta B+\lambda BD_d+AA^TB}\tag{22}$$

We update the matrices A and B with Eq. (21) and Eq. (22) until convergence. The predicted association score matrix for drug-disease pairs is obtained by $Y_P=A^TB$. Then, we prioritize the disease-associated drugs (or drug-associated diseases) on the basis of correlation scores in matrix Y_P. Generally, the higher the drug-disease pair score, the more likely they are to be related. The whole algorithm of WNMFDDA is exhibited in Table 1.

Table 1 The algorithm for predicting drug-disease associations

Algorithm: *WNMFDDA*

Input: Matrix $Y \in R^{n\times m}$, $S^R \in R^{n\times n}$ and $S^D \in R^{m\times m}$, decay term a, neighborhood sizes K and p, subspace dimensionality k, regularization parameter λ and β.

Output: Predicted association matrix Y_P;

1. calculate weighted K-nearest neighbor (*KNN*) profiles for drugs;

for each drug $r_p \in R = \{r_1, r_2, \cdots, r_n\}$ **do**

for $i \leftarrow 1$ **to** K **do**

$w_i = a^{i-1} * S^R(r_i, r_p)$; // $r_i \in KNN(r_p, S^R, K)$.

end for

$Y_r(r_p) = \frac{1}{\sum_{1\le i\le K} S^R(r_i, r_p)} \sum_{i=1}^{K} w_i Y(r_i)$

end for

2. calculate weighted K-nearest neighbor (*KNN*) profiles for diseases;

for each disease $d_q \in D = \{d_1, d_2, \cdots, d_m\}$ **do**

for $j \leftarrow 1$ **to** K **do**

$w_j = a^{j-1} * S^D(d_j, d_q)$; // $d_j \in KNN(d_q, S^D, K)$.

end for

$Y_d(d_q) = \frac{1}{\sum_{1\le j\le K} S^D(d_j, d_q)} \sum_{j=1}^{K} w_j Y(d_j)$

end for

3. merge the new interaction profiles $Y_{rd} = (Y_r + Y_d)/2$;
4. update association adjacency matrix $Y = \max(Y, Y_{rd})$;
5. construct sparse similarity matrix S^{R*}, S^{D*} based on S^R, S^D;
6. randomly initialize two nonnegative matrices $A^{k\times n}$ and $B^{k\times m}$;
7. repeat

update A and B based on the following rules until convergence

$$a_{ki} \leftarrow a_{ki} \frac{BY^T + \lambda A S^{R*}}{\beta A + \lambda A D_r + BB^T A}$$

$$b_{kj} \leftarrow b_{kj} \frac{AY + \lambda B S^{D*}}{\beta B + \lambda B D_d + AA^T B}$$

8. return $Y_P = A^T B$.

Results and discussion

Experimental settings

To systematically assess the ability of WNMFDDA in predicting potential associations of drugs with diseases, we conduct ten-fold cross validation (10-CV) experiments based on known drug-disease associations. In the golden dataset, 1933 known associations of drugs with diseases are randomly divided into ten roughly equal parts, while the other unconfirmed pairs are regarded as candidate associations. In each cross validation, each part is served as a test set in turn, and the remaining parts are treated as the training set.

AUC is widely applied for assessing the prediction models [49]. Since the known drug-disease associations are much less than unknown associations between them, the sensitivity (Sen., also known as recall) and Precision (Pre.)

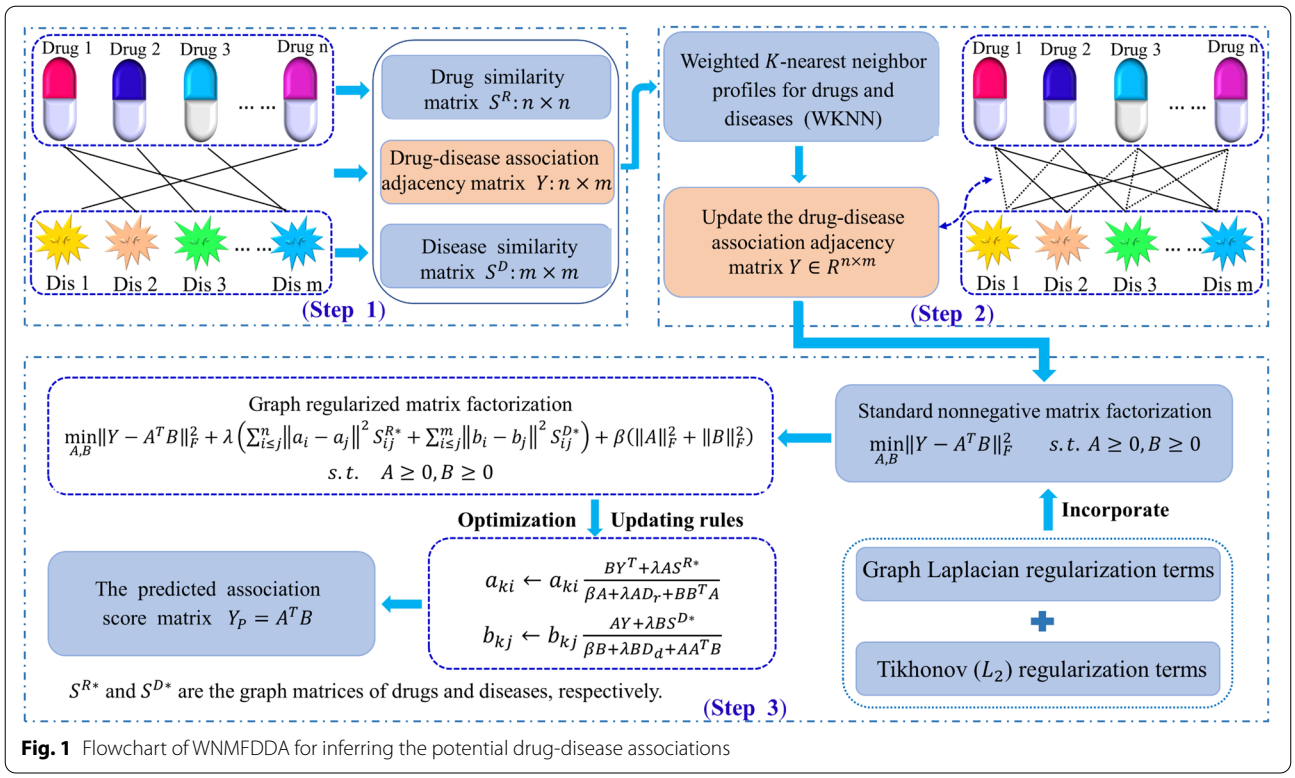

Fig. 1 Flowchart of WNMFDDA for inferring the potential drug-disease associations

are computed as the evaluation metric. In addition, other classification metrics, accuracy (Acc.) and F1-Score, are also used widely [50].

$$Sen. = \frac{TP}{TP + FN} \tag{23}$$

$$Pre. = \frac{TP}{TP + Fp} \tag{24}$$

$$Acc. = \frac{TN + TP}{TN + TP + FN + Fp} \tag{25}$$

$$F1 - Score = \frac{2 \times Pre. \times Sen.}{Pre. + Sen.} \tag{26}$$

In this work, the influence of parameters on WNMFDDA has been analyzed by applying Fdataset. We used grid search to determine the parameter combinations. WNMFDDA has six parameters and their values are considered from the following ranges: decay term $a \in \{0.1, 0.2, \cdots, 1\}$, neighborhood size K is chosen from $\{1, 2, \cdots, 10\}$, subspace dimensionality $k \in \{60, 80, 100, \cdots, 200\}$, regularization coefficients $\lambda \in \{0.02, 0.2, 1, 2\}$ and $\beta \in \{0.002, 0.02, 0.2, 1\}$. At the same time, we set $p = 5$ to construct the graphs for drug

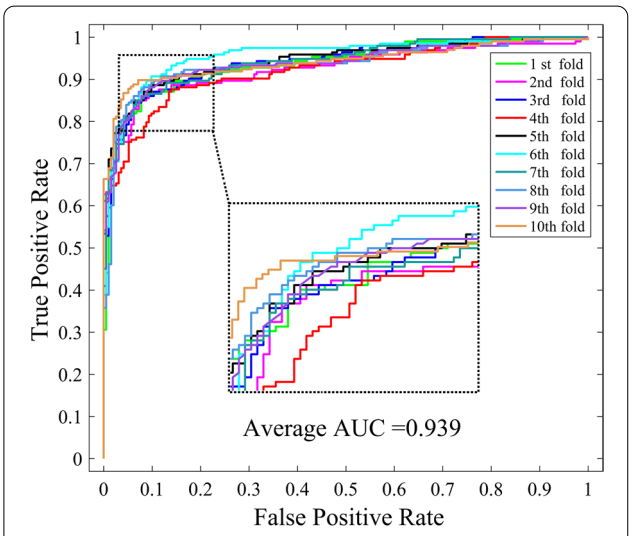

Fig. 2 The ROC curves of WNMFDDA on Fdataset under ten-fold cross validation

Table 2 The average AUC values of WNMFDDA and related methods on Fdataset

Methods	DDRS	MBiRW	HGBI	DrugNet	WNMFDDA
AUC	0.930	0.917	0.829	0.778	0.939

space and disease space according to [43] and [51]. The final optimal parameter combinations are $K = 5, a = 0.5$, $k = 160, \lambda = 1$ and $\beta = 0.02$, which are determined based on AUC values under 10-CV experiments. Meanwhile, we used the best parameter values that recommended by the corresponding authors in compared methods.

Performance evaluation

In this study, ten-fold cross validation was introduced to assess the performance of WNMFDDA. we conduct 10-CV on the Fdataset to compare it with four classical models, including DDRS [27], MBiRW [24], HGBI [23] and DrugNet [22]. As shown in Fig. 2, the AUC value achieved by WNMFDDA is 0.939. The AUC values of WNMFDDA and the other four competing approaches on Fdataset are displayed in Table 2. Specifically, the AUC values of WNMFDDA, DDRS, MBiRW, HGBI and DrugNet are 0.939, 0.930, 0.917, 0.829 and 0.778, respectively. The performance of WNMFDDA method outperforms the compared computational approaches, DDRS, MBiRW, HGBI and DrugNet.

In practice, the predicted top-ranked results are more important than other parts. In this study, the numbers of correctly retrieved true associations between drug and disease from different top portions were counted when all known associations are regarded as the training set. In generally, the method is considered as more reliable if more true associations are discovered on the top portions. At different thresholds, the number of true associations correctly predicted by WNMFDDA are shown in Fig. 3. For example, at the top 20 and 40 of predicted candidate drugs, WNMFDDA correctly identified 1651 (85.41%) and 1819 (94.10%) true associations from all the 1933 known associations, respectively. The experimental

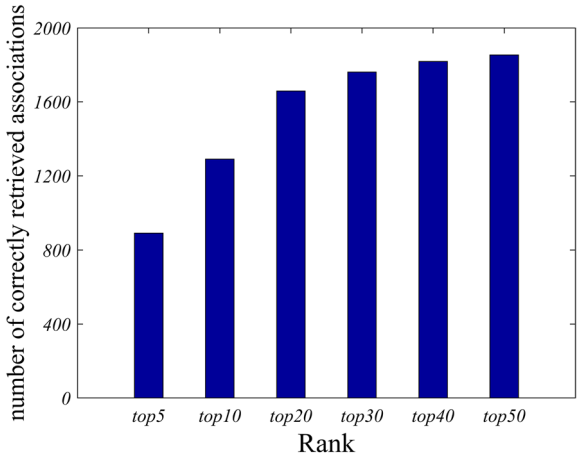

Fig. 3 The number of correctly retrieved drug-disease associations for various ranking thresholds

Table 3 The ten-fold cross validation results achieved by WNMFDDA on Fdataset

Test set	Sen.(%)	Pre.(%)	Acc.(%)	F1-Score(%)
1	86.53	89.78	88.34	88.13
2	87.05	89.84	88.60	88.42
3	86.01	89.73	88.08	87.83
4	81.35	89.20	85.75	85.09
5	87.05	89.84	88.60	88.42
6	89.64	90.10	89.90	89.87
7	86.53	89.78	88.34	88.13
8	88.60	90.00	89.38	89.30
9	86.53	89.78	88.34	88.13
10	89.80	89.80	89.80	89.80
Average	86.91 ± 2.38	89.79 ± 0.24	88.51 ± 1.17	88.31 ± 1.35

results suggest that our model has higher accuracy and lower false positive rate in identifying potential drug-disease associations.

In addition, considering the fact that the known and unknown associations between them are serious imbalance, several classification metrics (i.e. Sen., Pre., Acc. and F1-Score) are calculated at different specificity (Spe.), and are used as evaluation indicators. As shown in Table 3, the average Sen, Pre, Acc and F1-Score are 86.91%, 89.79%, 88.51% and 88.31%, respectively, when Spe is 90%. This result further illustrates that our method is reliable.

Case studies

In this section, to further test the predictive performance of WNMFDDA, we conduct two types of case studies on two drugs and two diseases, respectively. The first type of case study was performed on Doxorubicin drug and Obesity. During the experiment, all known associations on the Fdataset are utilized to train prediction model of WNMFDDA. For Doxorubicin, the top-15 candidate diseases related with Doxorubicin are obtained according to their predicted association scores. Then, we validate these candidate diseases based on the other public biological database: Comparative Toxicogenomics Database (CTD) [52], which provides newly experimentally verified associations between drugs and diseases. Table 4 lists the top-15 predicted candidate diseases for Doxorubicin, 12 out of the top-15 are confirmed by CTD to be associated with Doxorubicin. For example, Doxorubicin, originally indicated for Acute Leukemia, is predicted to treat stomach cancer and confirmed by CTD. As shown in Table 5, 13 out of the top-15 predicted drugs are confirmed by CTD to be associated with Obesity.

In order to illustrate the prediction capability of WNMFDDA on novel diseases /drugs without known associated drugs/diseases, we selected Amiodarone

Table 4 The top-15 candidate diseases associated with Doxorubicin are predicted by GWMFDDA based on known associations in Fdataset

Drug	Rank	Diseases	Evidences	Rank	Diseases	Evidences
Doxorubicin	1	Turcot syndrome	CTD	9	Urinary Bladder Neoplasms	CTD
	2	Lymphoblastic Leukemia, Acute, with Lymphomatous Features	unconfirmed	10	Neuroblastoma	CTD
	3	Breast Neoplasms	CTD	11	Testicular Germ Cell Tumor	CTD
	4	Hodgkin Disease	CTD	12	Multiple Myeloma	CTD
	5	Leukemia, Myeloid, Acute	CTD	13	Carcinoma, Small Cell	CTD
	6	Dohle Bodies And Leukemia	unconfirmed	14	Stomach Neoplasms	CTD
	7	Rhabdomyosarcoma 2	CTD	15	Reticulum Cell Sarcoma	unconfirmed
	8	Osteosarcoma	CTD			

Table 5 The top-15 candidate drugs associated with Obesity are predicted by GWMFDDA based on known associations in Fdataset

Disease	Rank	Drugs	Evidences	Rank	Drugs	Evidences
	1	Benzphetamine	CTD	9	Bupropion	CTD
	2	Phentermine	CTD	10	Amphetamine	CTD
	3	Phenylpropanolamine	CTD	11	Pseudoephedrine	unconfirmed
Obesity	4	Sibutramine	CTD	12	Dextroamphetamine	CTD
	5	Metamfetamine	unconfirmed	13	Ephedrine	CTD
	6	Orlistat	CTD	14	Cimetidine	CTD
	7	Phendimetrazine	CTD	15	Topiramate	CTD
	8	Diethylpropion	CTD			

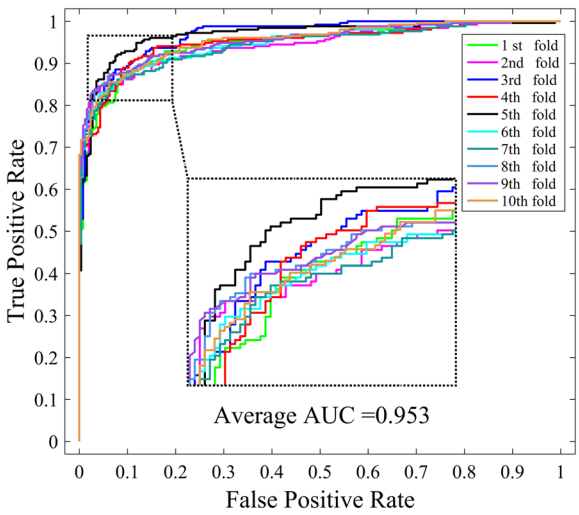

Fig. 4 ROC curves of WNMFDDA on Cdataset under ten-fold cross validation

drug and Asthma disease to perform the second type of case study. For drug Amiodarone, before training the model, all known associations with Amiodarone are removed from the original dataset. Then, we sort all the 313 diseases in descending order according to the correlation scores, and verify the top-15 diseases in the CTD. As shown in Table 6, 12 out of the top-15 drug-disease associations predicted by WNMFDDA are confirmed in the CTD. Similarly, all known associations with Asthma are hidden from the original dataset when we carry out case study to Asthma. The top-15 inferred candidate drugs are displayed in Table 7, 13 out of 15 are verified to be related with the Asthma by CTD. These results further suggest that WNMFDDA is a useful predictor to infer potential associations of diseases with drugs.

Validation on the other dataset

To further validate the robustness of WNMFDDA, we implement 10-CV to verify the prediction accuracy on the Cdataset. This dataset has been used in previous studies [24, 27], including 663 drugs, 409 diseases and 2532 verified drug-disease associations. These drugs and diseases are obtained from DrugBank database and OMIM database, respectively. The ROC curves of WNMFDDA on Cdataset are drawn in Fig. 4. The average AUC values of WNMFDDA and the compared methods are shown in Table 8. We can see that the average AUC value

Table 6 The top-15 candidate diseases associated with Amiodarone are predicted by GWMFDDA after removing all known associations with Amiodarone based on the Fdataset

Drug	Rank	Diseases	Evidences	Rank	Diseases	Evidences
Amiodarone	1	Breast Neoplasms	CTD	9	Hodgkin Disease	CTD
	2	Lymphoblastic Leukemia, Acute, with Lymphomatous Features	CTD	10	Osteosarcoma	CTD
	3	Leukemia, Myeloid, Acute	CTD	11	Inclusion Body Myopathy With Early-Onset Paget Disease And Frontotemporal Dementia	CTD
	4	Turcot Syndrome	Unconfirmed	12	Urinary Bladder Neoplasms	CTD
	5	Dohle Bodies and Leukemia	Unconfirmed	13	Lung Neoplasms	CTD
	6	Hajdu-Cheney Syndrome	Unconfirmed	14	Carcinoma, Small Cell	CTD
	7	Multiple Myeloma	CTD	15	Fibrous Dysplasia, Polyostotic	CTD
	8	Osteoporosis	CTD			

Table 7 The top-15 candidate drugs associated with Asthma are predicted by GWMFDDA after removing all known associations with Asthma based on the Fdataset

Disease	Rank	Drugs	Evidences	Rank	Drugs	Evidences
Asthma	1	Cromoglicic acid	Unconfirmed	9	Triamcinolone	CTD
	2	Ciprofloxacin	CTD	10	Montelukast	CTD
	3	Budesonide	CTD	11	Beclomethasone	CTD
	4	Pirbuterol	CTD	12	Moxifloxacin	CTD
	5	Salbutamol	CTD	13	Nedocromil	CTD
	6	Zileuton	CTD	14	Formoterol	CTD
	7	Prednisone	CTD	15	Orciprenaline	Unconfirmed
	8	Terbutaline	CTD			

Table 8 AUC values of WNMFDDA and related methods on Cdataset

Methods	DDRS	MBiRW	HGBI	DrugNet	WNMFDDA
AUC	0.947	0.933	0.858	0.804	0.953

of WNMFDDA is 0.953, while DDRS, MBiRW, HGBI and DrugNet are 0.947, 0.933, 0.858 and 0.804, respectively. WNMFDDA achieves the best prediction performance. The superior experiment results on Cdataset also demonstrate that our proposed model is robust and effective in revealing potential associations between drug and disease.

Conclusions

Identifying new indications for existing drugs is a promising alternative to drug development, which not only saves time and costs, but also reduces risks and expedites drug approval. In this work, a model based on weight non-negative matrix factorization, WNMFDDA, was proposed to predict potential drug-disease associations. Different from other traditional computational methods, WNMFDDA reformulate the adjacency association matrix based on weighted K nearest neighbor profiles as a preprocessing step, which enables it to infer potential associations for novel diseases/drugs without any known associated with drugs/diseases. Meanwhile, graph regularized matrix factorization was used to calculate the association scores.

We conducted 10-CV on two datasets and case studies on Fdataset to verify the performance of our developed model. Comprehensive experimental results demonstrate that WNMFDDA outperforms other state-of-the-art approaches, and can effectively infer potential associations between drug and disease. We believe that WNMFDDA is helpful for relevant biomedical researchers in follow-up studies. However, WNMFDDA still has some limitations. Firstly, the number of experimental verified drug-disease associations used in this work is relatively sparse. Secondly, determining the optimal parameter combinations for different biological datasets is still a daunting task. Finally, how to reasonably incorporate more effective drug and disease features to enhance the performance of WNMFDDA deserves further research.

Acknowledgements
We are grateful to all group members in the research group leaded by professor Zhu-Hong You for their valuable suggestions. The authors would like to thank the editors and anonymous reviewers for their reviews.

Author contributions
MNW conceived the algorithm, analyzed it, conducted the experiment, and wrote the manuscript. ZHY and DWD prepared the data set and designed experiment. XXJ and LW analyzed the experiment. All authors read and approved the final manuscript.

Funding
This work is supported in part by the NSFC Excellent Young Scholars Program, under Grants 61722212, in part by the National Science Foundation of China, under Grants 62161050, in part by the Science and Technology Project of Jiangxi Provincial Department of Education, under Grants GJJ190834, GJJ211603.

Availability of data and materials
The datasets that we collected in this work is freely available on https://github.com/meinengwang/WNMFDDA.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Mathematics and Computer Science, Yichun University, Yichun 336000, Jiangxi, China. [2]School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China. [3]Xinjiang Technical Institutes of Physics and Chemistry, Chinese Academy of Sciences, Urumqi 830011, China. [4]University of Chinese Academy of Sciences, Beijing 100049, China.

Received: 18 January 2022 Accepted: 6 November 2022
Published online: 03 December 2022

References

1. Li J, Zheng S, Chen B, Butte AJ, Swamidass SJ, Lu Z. A survey of current trends in computational drug repositioning. Brief Bioinform. 2016;17(1):2–12.
2. Paul SM, Mytelka DS, Dunwiddie CT, Persinger CC, Munos BH, Lindborg SR, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nat Rev Drug Discov. 2010;9(3):203–14.
3. Adams CP, Brantner VV. Estimating the cost of new drug development: is it really $802 million? Health Aff. 2006;25(2):420–8.
4. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ. 2003;22(2):151–85.
5. Grabowski H. Are the economics of pharmaceutical research and development changing? Pharmacoeconomics. 2004;22(2):15–24.
6. Kinch MS, Griesenauer RH. 2017 in review: FDA approvals of new molecular entities. Drug Discov Today. 2018;23(8):1469–73.
7. Krantz A. Diversification of the drug discovery process. Nat Biotechnol. 1998;16(13):1294.
8. Hurle M, Yang L, Xie Q, Rajpal D, Sanseau P, Agarwal P. Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther. 2013;93(4):335–41.
9. Yella JK, Yaddanapudi S, Wang Y, Jegga AG. Changing trends in computational drug repositioning. Pharmaceuticals. 2018;11(2):57.
10. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004;3(8):673–83.
11. Graul AI, Sorbera L, Pina P, Tell M, Cruces E, Rosa E, et al. The year's new drugs & biologics-2009. Drug News Perspect. 2010;23(1):7–36.
12. Sardana D, Zhu C, Zhang M, Gudivada RC, Yang L, Jegga AG. Drug repositioning for orphan diseases. Brief Bioinform. 2011;12(4):346–56.
13. Nosengo N. Can you teach old drugs new tricks? Nature. 2016;534(7607):314–6.
14. Shim JS, Liu JO. Recent advances in drug repositioning for the discovery of new anticancer drugs. Int J Biol Sci. 2014;10(7):654.
15. Oh M, Ahn J, Yoon Y. A network-based classification model for deriving novel drug-disease associations and assessing their molecular actions. PLoS ONE. 2014;9(10):e111668.
16. Yang H, Spasic I, Keane JA, Nenadic G. A text mining approach to the prediction of disease status from clinical discharge summaries. J Am Med Inform Assoc. 2009;16(4):596–600.
17. Chen X, Yan G-Y. Semi-supervised learning for potential human microRNA-disease associations inference. Sci Rep. 2014;4:5501.
18. Napolitano F, Zhao Y, Moreira VM, Tagliaferri R, Kere J, D'Amato M, et al. Drug repositioning: a machine-learning approach through data integration. J Cheminformatics. 2013;5(1):30.
19. Gottlieb A, Stein GY, Ruppin E, Sharan R. PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol. 2011;7(1):496.
20. Chiang AP, Butte AJ. Systematic evaluation of drug–disease relationships to identify leads for novel drug uses. Clin Pharmacol Ther. 2009;86(5):507–10.
21. 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. 2014;54(9):2562–9.
22. Martinez V, Navarro C, Cano C, Fajardo W, Blanco A. DrugNet: Network-based drug–disease prioritization by integrating heterogeneous data. Artif Intell Med. 2015;63(1):41–9.
23. Wang W, Yang S, Zhang X, Li J. Drug repositioning by integrating target information through a heterogeneous network model. Bioinformatics. 2014;30(20):2923–30.
24. Luo H, Wang J, Li M, Luo J, Peng X, Wu F-X, et al. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. Bioinformatics. 2016;32(17):2664–71.
25. Alaimo S, Giugno R, Pulvirenti A. Recommendation techniques for drug–target interaction prediction and drug repositioning. Data mining techniques for the life sciences. Springer; 2016. p. 441–62.
26. Wang M, Xie X, You Z, Wong L, Li L, Chen Z, editors. Weighted nonnegative matrix factorization based on multi-source fusion information for predicting CircRNA-disease associations. In: International conference on intelligent computing. Springer; 2021.
27. Luo H, Li M, Wang S, Liu Q, Li Y, Wang J. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. Bioinformatics. 2018;34(11):1904–12.
28. Huang Y-A, You Z-H, Chen X, Huang Z-A, Zhang S, Yan G-Y. Prediction of microbe–disease association from the integration of neighbor and graph with collaborative recommendation model. J Transl Med. 2017;15(1):209.
29. Luo X, Zhou M, Li S, You Z, Xia Y, Zhu Q. A nonnegative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method. IEEE Trans Neural Netw Learn Syst. 2015;27(3):579–92.
30. Luo X, Zhou M, Xia Y, Zhu Q. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. IEEE Trans Industr Inf. 2014;10(2):1273–84.
31. Jiang X, Hu X, Xu W. Microbiome data representation by joint nonnegative matrix factorization with laplacian regularization. IEEE/ACM Trans Comput Biol Bioinf. 2015;14(2):353–9.
32. Zhang W, Yue X, Lin W, Wu W, Liu R, Huang F, et al. Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinformatics. 2018;19(1):1–12.
33. Fu G, Wang J, Domeniconi C, Yu G. Matrix factorization-based data fusion for the prediction of lncRNA–disease associations. Bioinformatics. 2018;34(9):1529–37.
34. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(suppl_1):D514–7.
35. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34(suppl_1):D668–72.

36. Steinbeck C, Han Y, Kuhn S, Horlacher O, Luttmann E, Willighagen E. The Chemistry Development Kit (CDK): an open-source Java library for chemo- and bioinformatics. J Chem Inf Comput Sci. 2003;43(2):493–500.
37. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. J Chem Inf Comput Sci. 1988;28(1):31–6.
38. Tanimoto TT. Elementary mathematical theory of classification and prediction. 1958.
39. Van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. Eur J Hum Genet. 2006;14(5):535–42.
40. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999;401(6755):788–91.
41. Liu Y, Wang S-L, Zhang J-F. Prediction of microbe–disease associations by graph regularized non-negative matrix factorization. J Comput Biol. 2018;25(12):1385–94.
42. Wang M-N, You Z-H, Wang L, Li L-P, Zheng K. LDGRNMF: LncRNA-disease associations prediction based on graph regularized non-negative matrix factorization. Neurocomputing. 2021;424:236–45.
43. Cai D, He X, Han J, Huang TS. Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell. 2010;33(8):1548–60.
44. You Z-H, Lei Y-K, Gui J, Huang D-S, Zhou X. Using manifold embedding for assessing and predicting protein interactions from high-throughput experimental data. Bioinformatics. 2010;26(21):2744–51.
45. Ezzat A, Zhao P, Wu M, Li X-L, Kwoh C-K. Drug-target interaction prediction with graph regularized matrix factorization. IEEE/ACM Trans Comput Biol Bioinf. 2016;14(3):646–56.
46. Guan N, Tao D, Luo Z, Yuan B. Manifold regularized discriminative non-negative matrix factorization with fast gradient descent. IEEE Trans Image Process. 2011;20(7):2030–48.
47. Liu X, Zhai D, Zhao D, Zhai G, Gao W. Progressive image denoising through hybrid graph Laplacian regularization: a unified framework. IEEE Trans Image Process. 2014;23(4):1491–503.
48. Facchinei F, Kanzow C, Sagratella S. Solving quasi-variational inequalities via their KKT conditions. Math Program. 2014;144(1–2):369–412.
49. Hajian-Tilaki K. Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. Caspian J Intern Med. 2013;4(2):627.
50. 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 Bioinf. 2016;14(6):1468–75.
51. Li X, Cui G, Dong Y. Graph regularized non-negative low-rank matrix factorization for image clustering. IEEE Trans Cybern. 2016;47(11):3840–53.
52. Davis AP, Grondin CJ, Johnson RJ, Sciaky D, McMorran R, Wiegers J, et al. The comparative toxicogenomics database: update 2019. Nucleic Acids Res. 2019;47(D1):D948–54.

Publisher's Note