Bioinformatics, 36(11), 2020, 3474–3481
doi: 10.1093/bioinformatics/btaa157
Advance Access Publication Date: 7 March 2020
Original Paper

Systems biology

A graph regularized generalized matrix factorization model for predicting links in biomedical bipartite networks

Zi-Chao Zhang[1,2,†], Xiao-Fei Zhang [3,†], Min Wu [4], Le Ou-Yang[1,*], Xing-Ming Zhao [2,5] and Xiao-Li Li[4]

[1]Guangdong Key Laboratory of Intelligent Information Processing, Key Laboratory of Media Security, Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen University, Shenzhen 518060, China, [2]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China, [3]School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, China, [4]Institute for Infocomm Research (I2R), A*STAR, 138632, Singapore and [5]Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence, Ministry of Education, 200433 China

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Lenore Cowen

Received on August 5, 2019; revised on February 5, 2020; editorial decision on February 29, 2020; accepted on March 3, 2020

Abstract

Motivation: Predicting potential links in biomedical bipartite networks can provide useful insights into the diagnosis and treatment of complex diseases and the discovery of novel drug targets. Computational methods have been proposed recently to predict potential links for various biomedical bipartite networks. However, existing methods are usually rely on the coverage of known links, which may encounter difficulties when dealing with new nodes without any known link information.

Results: In this study, we propose a new link prediction method, named graph regularized generalized matrix factorization (GRGMF), to identify potential links in biomedical bipartite networks. First, we formulate a generalized matrix factorization model to exploit the latent patterns behind observed links. In particular, it can take into account the neighborhood information of each node when learning the latent representation for each node, and the neighborhood information of each node can be learned adaptively. Second, we introduce two graph regularization terms to draw support from affinity information of each node derived from external databases to enhance the learning of latent representations. We conduct extensive experiments on six real datasets. Experiment results show that GRGMF can achieve competitive performance on all these datasets, which demonstrate the effectiveness of GRGMF in prediction potential links in biomedical bipartite networks.

Availability and implementation: The package is available at https://github.com/happyalfred2016/GRGMF.

Contact: leouyang@szu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Advances in network science promotes the development of network medicine (Barabási *et al.*, 2011; Ideker and Sharan, 2008). Network-based approaches have been shown effective for multiple clinical and biological applications (Barabási *et al.*, 2011). In particular, the analysis of biomedical networks can provide useful insights into the diagnosis and prognosis of complex diseases. An important subset of biomedical networks is bipartite, which includes two different classes of nodes such that every edge connects a node in one class to a node in the other class. In biomedical bipartite networks, one class of the nodes is usually composed of molecular components, such as genes, microRNAs or proteins, and the other class is usually composed of various indicators of human diseases, such as symptoms and adverse drug effects (Pavlopoulos *et al.*, 2018). Identifying the potential links between two classes of nodes in biomedical bipartite networks may help to identify new disease genes, new drugs, new targets and new biomarkers for complex diseases (Barabási *et al.*, 2011; Pavlopoulos *et al.*, 2018).

Identifying links in biomedical bipartite networks via *in vivo* or biochemical experimental methods can be extremely costly (Luo *et al.*, 2017). Computational prediction of potential links in

biomedical bipartite networks can efficiently guide *in vivo* validation and significantly reduce the cost required for disease diagnosis and drug discovery. Thus, much effort has been devoted to developing computational methods for predicting potential links in biomedical bipartite networks (Ezzat *et al.*, 2017, 2018; Luo *et al.*, 2017; Xiao *et al.*, 2018). For example, computational prediction of interactions between drugs/compounds and targets has been widely used to complement wet-lab experiments for drug discovery and drug repositioning (Ezzat *et al.*, 2017, 2018; Liu *et al.*, 2016b; Zheng *et al.*, 2013). To investigate the genetic complexities of complex diseases and the correlation among discrete disease phenotypes, many network-based computational approaches have been developed for identifying the associations between genes and diseases (Barabási *et al.*, 2011; van Dam *et al.*, 2017). Similarly, as microRNAs (miRNA) have been proven to play important roles in the development and prognosis of human complex diseases, discovering the potential associations between miRNAs and diseases via computational approaches provides a low-cost and efficient way to understand the molecular mechanisms and pathogenesis of complex diseases (Chen *et al.*, 2018; Xiao *et al.*, 2018; You *et al.*, 2017).

Generally, existing computational methods for link prediction can be roughly divided into three categories, that is, feature-based classification methods, network diffusion-based methods and matrix factorization-based methods (Ezzat *et al.*, 2018). The key idea behind feature-based machine learning methods is the 'guilt-by-association' assumption, that is, similar nodes in one class may be connected to same or similar nodes and vice versa (Luo *et al.*, 2017). By representing each pair of nodes as a feature vector, link prediction problem is usually formulated as a binary classification task. For instance, Bleakley and Yamanishi (2009) proposed a bipartite local model to predict drug–target interactions (DTIs) based on support vector machines. Mei *et al.* (2013) extended this model by adding a neighbor-based interaction-profile inferring procedure. Network diffusion-based methods utilize graph-based approaches for influence propagation in biomedical bipartite networks and predict potential links. For example, Xuan *et al.* (2015) proposed a random walk-based method to predict miRNA-disease associations (MDAs). By utilizing some comprehensive similarity measures to calculate the similarity networks for drugs and diseases, Luo *et al.* (2016) introduced a bi-random walk algorithm to predict potential associations between drugs and diseases. Finally, matrix factorization-based methods take the biadjacency matrix of a bipartite network as input and decompose the input matrix into the product of two or more low-rank matrix factors. For example, Zheng *et al.* (2013) proposed a factor model which could project drugs, targets and their similarity matrices into a common feature space for DTI prediction. Liu *et al.* (2016b) proposed a powerful logistic matrix factorization-based model named NRLMF for predicting DTIs. Xiao *et al.* (2018) proposed a graph regularized non-negative matrix factorization model to predict potential associations between miRNAs and diseases.

On the other hand, with the influx of heterogeneous biological data, predicting the potential links in biomedical bipartite networks by integrating multiple related data sources has become one of the most promising trends (Liu *et al.*, 2016a). Matrix factorization-based methods have been extensively explored to integrate multiple data sources (Ding *et al.*, 2010; Gu and Zhou, 2009a, b; Klema and Laub, 1980; Liu *et al.*, 2016b; Wang *et al.*, 2014, 2018; Xue *et al.*, 2017; You *et al.*, 2017). This type of methods becomes popular because they can jointly explore the intrinsic structure of multiple data sources without the need to develop separate models for individual data sources (Fu *et al.*, 2018). Moreover, it is easier for them to incorporate additional information by imposing suitable constraints on the latent factor matrices (Gu and Zhou, 2009a, b). However, most of existing matrix factorization-based methods strongly rely on the known link information to learn the latent factor matrices, which means it is hard for these methods to predict the potential links involving new nodes. Although some methods have been proposed to learn the latent representations of new nodes by drawing support from external information (e.g. neighborhood information) of nodes (Liu *et al.*, 2016b; Xiao *et al.*, 2018), these methods usually achieve this goal by imposing regularization constraints between the latent representations of a new node and its neighbors, which makes them sensitive to the external information and strongly rely on the input data sources.

In this study, we develop a novel link prediction model named graph regularized generalized matrix factorization (GRGMF) to infer potential links in biomedical bipartite networks (Fig. 1). In particular, we first propose a generalized matrix factorization (GMF) framework to formulate the link prediction task. Here, we incorporate the neighborhood information of all nodes into the matrix factorization model, which makes it possible for our model to learn the latent representations of new nodes. Moreover, instead of simply using predefined metrics to calculate the neighborhood information of each node, we introduce a novel algorithm to learn the neighborhood information of each node adaptively. Then, we incorporate two graph regularization terms into the GMF framework to draw support from other existing public databases to promote the learning of latent representation for each node. To evaluate the performance of various link prediction methods comprehensively, we consider three different cross-validation (CV) settings. The experiment results on six real datasets under three CV settings demonstrate the effectiveness of our GRGMF in predicting potential links in biomedical bipartite networks. Furthermore, case studies on three types of cancers also demonstrate the effectiveness of our GRGMF in discovering novel MDAs.

2 Materials and methods

2.1 Notations and problem statement

A biomedical bipartite network can be modeled as a bipartite graph $G = (A, B, E)$, where $A = \{a_i\}_{i=1}^{m}$ and $B = \{b_j\}_{j=1}^{n}$ are two disjoint sets of nodes that contain $|A| = m$ and $|B| = n$ nodes, respectively, and each edge/link in edge set E connects a node in A to a node in B. Here, the nodes in A and B correspond to two different types of

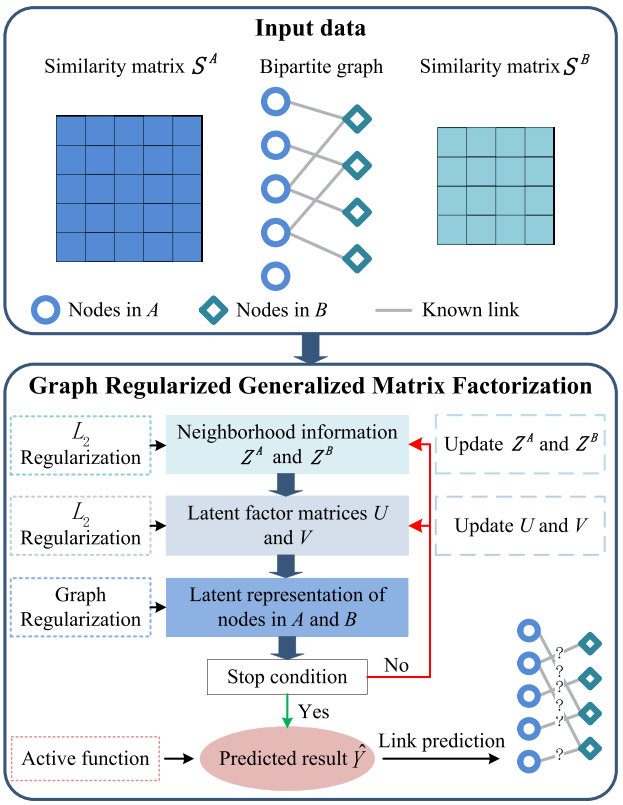

Fig. 1. The overall workflow of GRGMF for detecting potential links in biomedical bipartite networks

bioentities in the biomedical bipartite network, and the edges in E correspond to the associations (or connections) between two types of bioentities. A biadjacency matrix $Y \in \{0,1\}^{m\times n}$ can be used to represent the bipartite graph, where $Y_{ij} = 1$ if there is a link between nodes a_i and b_j, and $Y_{ij} = 0$ if the link does not exist or is still unknown. The problem of link prediction in a biomedical bipartite network is to estimate the scores of unobserved entries in the corresponding biadjacency matrix Y, and rank the candidate links according to the predicted scores such that the top-ranked links could be treated as the potential links.

Based on the assumption that similar bioentities tend to have similar link patterns, taking into account the similarities within each type of bioentities may help to improve the accuracy of link prediction (Liu *et al.*, 2016b; Xiao *et al.*, 2018). With the development of high-throughput experimental techniques, we could collect the feature vectors of bioentities from external databases (Luo *et al.*, 2017). Based on these feature vectors and some predefined metrics, we can calculate the similarities within each type of bioentities. Taking DTI as example, the similarity between a pair of drugs can be computed based on the chemical structures of their compounds and the similarity between a pair of targets can be calculated based on their amino acid sequences (Liu *et al.*, 2016b; Luo *et al.*, 2017). Thus, in this study, we assume that the similarities within each type of bioentities can be prepared in advance. Let $S^A \in \mathbb{R}^{m\times m}$ and $S^B \in \mathbb{R}^{n\times n}$ denote two similarity matrices that describe the similarities within nodes in A and B derived from external databases via predefined metrics. In general, S^A and S^B are symmetric matrices, and the values of the elements in S^A and S^B are ranging from 0 to 1.

Thus, in this study, given a biadjacency matrix Y which indicates the observed links, and two similarity matrices S^A and S^B which describe the similarities among nodes in A and B derived from external databases, the task of various link prediction methods is to predict potential links from unobserved entries in Y based on these three input data.

2.2 Generalized matrix factorization

In this study, we develop the link prediction model based on GMF (He *et al.*, 2017), which is the generalization of matrix factorization under the neural collaborative filtering framework. In particular, under the setting of GMF (He *et al.*, 2017), the input feature vectors $\mathbf{z}_i^A$ and $\mathbf{z}_j^B$ for nodes a_i and b_j are two binarized sparse vectors with one-hot encoding, which indicate the identities for nodes a_i and b_j (i.e. $\mathbf{z}_i^A \in \{0,1\}^{1\times m}$ and $\mathbf{z}_j^B \in \{0,1\}^{1\times n}$ are two zero vectors except for $\mathbf{z}_i^A(1,i) = 1$ and $\mathbf{z}_j^B(1,j) = 1$). Let $U \in \mathbb{R}^{m\times K}$ and $V \in \mathbb{R}^{n\times K}$ denote the latent factor matrices (K denotes the dimension of the latent space) for node sets A and B, respectively. GMF estimates the probability of a link between a_i and b_j as

$$\hat{Y}_{ij} = g(\mathbf{w}((\mathbf{z}_i^A U) \odot (\mathbf{z}_j^B V))^T). \tag{1}$$

where $\odot$ denotes the element-wise product of vectors, $g(\cdot)$ and $\mathbf{w} \in \mathbb{R}^{1\times K}$ denote the active function and weight vector of output layer. According to this definition, if $g(\cdot)$ is an identity function and $\mathbf{w}$ is a uniform vector of 1, the above model degenerates into a standard matrix factorization model.

In this study (following previous studies: He *et al.*, 2017; Liu *et al.*, 2016b), we use the sigmoid function $f(x) = 1/(1+e^{-x})$ as $g(\cdot)$ and set $\mathbf{w}$ to be a uniform vector of 1. Note that under this setting, Equation (1) is equivalent to the logistic matrix factorization model proposed by Johnson (2014), which can be formulated as

$$\hat{Y}_{ij} = \frac{1}{(1+e^{-U_{i\cdot}V_{j\cdot}^T})}. \tag{2}$$

where $U_{i\cdot} \in \mathbb{R}^{1\times K}$ and $V_{j\cdot} \in \mathbb{R}^{1\times K}$ denote the i-th row of U and j-th row of V, respectively (where $U_{i\cdot}$ and $V_{j\cdot}$ can be treated as the latent factors that describe nodes a_i and b_j, respectively). For link prediction tasks, elements with a numerical value of 1 in biadjacency matrix Y (i.e. $Y_{ij} = 1$) denote the observed links that have been experimentally verified, whereas elements with zero values in Y are unknown pairs that may contain some unknown links. Thus, when using elements in Y as training samples, in order to train a more accurate model for link prediction (similar to previous studies: Liu *et al.*, 2016b), we could introduce a weighting strategy to assign higher importance to known link pairs than unknown pairs. By assuming that all the training samples in Y are independent, the likelihood function can be defined as follows

$$p(Y|U,V) = \prod_{Y_{ij}=1} \hat{Y}_{ij}^c \prod_{Y_{i'j'}=0} (1-\hat{Y}_{i'j'}). \tag{3}$$

where c is a constant used to control the weight assigned to observed connections (Liu *et al.*, 2016b). Based on this model, the link prediction task is to estimate latent vector matrices U and V by maximizing the above likelihood function. However, for new nodes that do not have any known links, it is hard for the above model to learn their latent factors accurately, as no information can be used to train the model. Based on the assumption that similar nodes tend to have similar link patterns, treating the neighborhood information of each node as its features may help to promote the performance of link prediction. Thus, we would like $\mathbf{z}_i^A$ and $\mathbf{z}_j^B$ to represent the neighborhood information of nodes a_i and b_j. In particular, we use $Z^A = [\mathbf{z}_1^A; \mathbf{z}_2^A; \ldots; \mathbf{z}_m^A] \in [0,1]^{m\times m}$ and $Z^B = [(\mathbf{z}_1^B)^T, (\mathbf{z}_2^B)^T, \ldots, (\mathbf{z}_n^B)^T] \in [0,1]^{n\times n}$ to capture the neighborhood information of nodes in A and B, respectively. Here, $Z_{ii'}^A$ describes the probability of node $a_{i'}$ to be the neighbor of node a_i, while $Z_{jj'}^B$ describes the probability of node $b_{j'}$ to be the neighbor of node b_j. Thus, in our model, Equation (2) becomes

$$\hat{Y}_{ij} = \frac{1}{(1+e^{-Z_{i\cdot}^A U V^T Z_{j\cdot}^B})}. \tag{4}$$

where $Z_{i\cdot}^A$ denotes the i-th row of Z^A and $Z_{\cdot j}^B$ denotes the j-th column of Z^B. By taking Equation (4) into Equation (3), we could learn the full latent factor matrices U and V based on observed links in Y and the neighborhood information of each node. Therefore, for new nodes that do not have any known links, we could still predict their potential links once we get their neighborhood information. Although we could compute the neighbors of each node via predefined metrics based on existing public databases, the neighborhood information provided by these predefined similarity matrices may include some noise data, which will mislead the learning processes. Thus, in this study, instead of directly using the similarity matrices computed from other databases to define Z^A and Z^B, we introduce a novel adaptive learning algorithm to learn Z^A and Z^B by adaptively assigning neighbors for each node.

In particular, to guarantee the probability properties of $Z_{i\cdot}^A$ and $Z_{\cdot j}^B$, we introduce extra constraints $\sum_{i'=1}^m Z_{ii'}^A = 1$ and $\sum_{j'=1}^n Z_{jj'}^B = 1$ on Z^A and Z^B, respectively. To avoid the generation of trivial estimations for $Z_{i\cdot}^A$ and $Z_{\cdot j}^B$, e.g. only the nearest node can be defined as the neighbor of node a_i with probability 1, two L_2 regularization terms are imposed on $Z_{i\cdot}^A$ and $Z_{\cdot j}^B$, respectively. Moreover, to reduce overfitting, we also impose L_2 regularization on U and V. By taking the negative logarithm of the likelihood function (3) and adding the above constraints and regularization terms, the model parameters can then be learned by minimizing the following objective function

$$\begin{aligned} \min_{Z^A,Z^B,U,V} \quad & \sum_{i=1}^m \sum_{j=1}^n [(cY_{ij} + 1 - Y_{ij}) \log(1 + e^{(Z_{i\cdot}^A U V^T Z_{\cdot j}^B)}) \\ & - cY_{ij}(Z_{i\cdot}^A U V^T Z_{\cdot j}^B)] + \lambda(||U||_F^2 + ||V||_F^2) \\ & + \beta(\textstyle\sum_i ||Z_{i\cdot}^A||_2^2 + \sum_j ||Z_{j\cdot}^B||_2^2), \\ s.t. \quad & 0 \le Z_{ii'}^A, Z_{jj'}^B \le 1, \sum_{i'=1}^m Z_{ii'}^A = 1, \sum_{j'=1}^n Z_{jj'}^B = 1. \end{aligned} \tag{5}$$

2.3 Graph regularization

Note that the similarity matrices S^A and S^B for nodes in A and B that are derived from other existing public databases can still provide important insights for the link prediction task (Liu *et al.*, 2016b). In this study, based on the assumption that similar nodes

tend to have similar latent representations and similar link patterns, the nearest neighbors of each node are utilized to improve the accuracy of link prediction. In particular, we use k-nearest-neighbors of each node to define its neighborhood information, e.g. the k-nearest-neighbors of a_i is denoted by $N(a_i)$. That is, for nodes in A, we generate a neighborhood matrix S^{AN} from S^A as follows

$$S^{AN}_{ii'} = \begin{cases} S^A_{ii'} & a_{i'} \in N(a_i), \\ 1 & i = i', \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

Similarly, the neighborhood matrix S^{BN} is generate from S^B as follows

$$S^{BN}_{jj'} = \begin{cases} S^B_{jj'} & b_{j'} \in N(b_j), \\ 1 & j = j', \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

To make S^{AN} and S^{BN} symmetric, we set $S^{AN} = (S^{AN} + (S^{AN})^T)/2$ and $S^{BN} = (S^{BN} + (S^{BN})^T)/2$. Here, instead of considering all neighbors provided by S^A and S^B which may potentially introduce noisy information, we just leverage the nearest neighbors for each node to enhance the prediction accuracy. We will discuss the effects of the number of nearest neighbors k in the next section. Based on S^{AN} and S^{BN}, we introduce the following Laplacian regularizer to enforce nodes with high similarities to have similar representations in the latent space

$$\begin{aligned} &\frac{r_1}{2}\sum_{i=1}^{m}\sum_{i'=1}^{m} S^{AN}_{ii'} ||Z^A_{i\cdot}U - Z^A_{i'\cdot}U||^2_F \\ &+\frac{r_2}{2}\sum_{j=1}^{n}\sum_{j'=1}^{n} S^{BN}_{jj'} ||V^T Z^B_{\cdot j} - V^T Z^B_{\cdot j'}||^2_F \\ &= r_1 tr((Z^A U)^T L^{AN} Z^A U) + r_2 tr(V^T Z^B L^{BN} (Z^B)^T V). \end{aligned} \tag{8}$$

where $tr(\cdot)$ denotes the trace of a matrix, $L^{AN} = D^{AN} - S^{AN}$ and $L^{BN} = D^{BN} - S^{BN}$ are the Laplacian matrices of S^{AN} and S^{BN}, respectively (D^{AN} and D^{BN} are two diagonal matrices in which their diagonal elements are $D^{AN}_{ii} = \sum_{i'=1}^{m} S^{AN}_{ii'}$ and $D^{BN}_{jj} = \sum_{j'=1}^{n} S^{BN}_{jj'}$, respectively), and r_1 and r_2 are the parameters to control the strength of the graph regularization.

2.4 Graph regularized generalized matrix factorization

The final link prediction model can be formulated by considering the existing connections and the neighbors of nodes. By plugging Equation (8) into Equation (5), our proposed GRGMF is formulated as follows:

$$\begin{aligned} \min_{Z^A,Z^B,U,V} \quad & J = \sum_{i=1}^{m}\sum_{j=1}^{n} [(cY_{ij} + 1 - Y_{ij}) \log(1 + e^{(Z^A_{i\cdot} U V^T Z^B_{\cdot j})}) \\ & - cY_{ij}(Z^A_{i\cdot} U V^T Z^B_{\cdot j})] + \lambda(||U||^2_F + ||V||^2_F) \\ & + \beta(\textstyle\sum_i ||Z^A_{i\cdot}||^2_2 + \sum_j ||Z^B_{\cdot j}||^2_2) \\ & + r_1 tr((Z^A U)^T L^{AN} Z^A U) \\ & + r_2 tr(V^T Z^B L^{BN} (Z^B)^T V), \\ s.t. \quad & 0 \le Z^A_{ii'}, Z^B_{j'j} \le 1, \sum_{i'=1}^{m} Z^A_{ii'} = 1, \sum_{j'=1}^{n} Z^B_{j'j} = 1. \end{aligned} \tag{9}$$

Figure 1 shows the overall flow of our proposed GRGMF to learn Z^A, Z^B, U and V for link prediction.

2.5 Optimization

The optimization problem in Equation (9) can be solved by an alternating optimization scheme. In particular, each time we optimize the objective function with respect to one parameter while fixing others.

2.5.1 Update U and V with fixed Z^A and Z^B

When Z^A and Z^B are fixed, we adopt an alternating gradient descent procedure to update U and V. In particular, we adopt the adaptive moment estimation (Adam) optimizer proposed by Kingma and Ba (2014) with learning rate $lr = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and batch size is set to the size of training set since we use batch gradient descent.

2.5.2 Update Z^A and Z^B with fixed U and V

When U and V are fixed, a relaxed Majorization–Minimization algorithm (Yang and Oja, 2012a) is employed to update Z^A and Z^B. The partial gradients of Equation (9) with respect to Z^A and Z^B are calculated as follows

$$\begin{aligned} D_A \quad &= \frac{\partial J}{\partial Z^A} = \hat{Y}(UV^T Z^B)^T + 2\beta Z^A \\ &+ (c-1)(Y \odot \hat{Y})(UV^T Z^B)^T - cY(UV^T Z^B)^T \\ &+ 2r_1(D^{AN} Z^A UU^T - S^{AN} Z^A UU^T) \end{aligned} \tag{10}$$

$$\begin{aligned} D_B \quad &= \frac{\partial J}{\partial Z^B} = (Z^A UV^T)^T \hat{Y} + 2\beta Z^B \\ &+ (c-1)(Z^A UV^T)^T (Y \odot \hat{Y}) - c(Z^A UV^T)^T Y \\ &+ 2r_2(VV^T Z^B D^{BN} - VV^T Z^B S^{BN}) \end{aligned} \tag{11}$$

Subsequently, we separate the positive parts and negative parts of D_A and D_B, respectively. That is $D_A = D^+_A - D^-_A$ and $D_B = D^+_B - D^-_B$. Referring to (Ding *et al.*, 2010; Yang and Oja, 2012b), we presented the update formulas of Z^A and Z^B as follows:

$$\begin{aligned} Z^A_{ii'} &\leftarrow Z^A_{ii'} \cdot \frac{d^A_i (D^-_A)_{ii'} + 1}{d^A_i (D^+_A)_{ii'} + e^A_i} \\ Z^B_{j'j} &\leftarrow Z^B_{j'j} \cdot \frac{d^B_j (D^-_B)_{j'j} + 1}{d^B_j (D^+_B)_{j'j} + e^B_j} \end{aligned} \tag{12}$$

where

$$\begin{aligned} d^A_i &= \textstyle\sum_{i'} \frac{Z^A_{ii'}}{(D^+_A)_{ii'}}, \quad e^A_i = \sum_{i'} Z^A_{ii'} \frac{(D^-_A)_{ii'}}{(D^+_A)_{ii'}}, \\ d^B_j &= \textstyle\sum_{j'} \frac{Z^B_{j'j}}{(D^+_B)_{j'j}}, \quad e^B_j = \sum_{j'} Z^B_{j'j} \frac{(D^-_B)_{j'j}}{(D^+_B)_{j'j}}. \end{aligned} \tag{13}$$

The whole process of our GRGMF is summarized in Algorithm 1. In this study, we stop the whole algorithm if $|J_t - J_{t-1}| \le \epsilon |J_{t-1}|$, where J_t denotes the value of objective function at the t-th iteration and the tolerance ϵ is set to 10^{-4}.

3 Experiments

In this section, we first introduce our experimental datasets and settings. Then, we show the experimental results and the comparison among various methods. Lastly, we also present case studies for the top predictions from our GRGMF method.

3.1 Datasets

To evaluate the performances of various link prediction methods, we collect six benchmark datasets for two types of biomedical bipartite networks, namely DTI network and MDA network.

For DTI network, we collect five datasets, i.e. nuclear receptors (NR), G-protein coupled receptors (GPCR), ion channels (IC), enzymes (E) and DTINet which are originally provided by Yamanishi *et al.* (2008) and Luo *et al.* (2017) and could be downloaded from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/

and https://github.com/luoyunan/DTINet. Each dataset includes three types of information, i.e., the observed DTIs, the similarities among drugs and the similarities among targets. In particular, for all these five datasets, the similarities among drugs are calculated based on their chemical structures, and the similarities among targets are computed based on their primary amino acid sequences. Please refer to Yamanishi *et al.* (2008) and Luo *et al.* (2017) for more details about these datasets.

For MDA network, we collect one dataset. The experimentally verified associations between miRNAs and diseases are downloaded from *HMDD v2.0* (Li *et al.*, 2014b). The similarities among diseases are derived from MeSH (https://www.nlm.nih.gov/mesh/), based on their semantic similarities (You *et al.*, 2017). The similarities among miRNAs are inferred by integrating the experimentally verified miRNA–gene interactions and the gene functional interaction network, following the method developed in Xiao *et al.* (2018). More details about the similarities among diseases and miRNAs can be found in previous studies (Wang *et al.*, 2010; Xiao *et al.*, 2018; You *et al.*, 2017). In addition, the detailed statistics of the above six datasets are shown in Table 1.

3.2 Evaluation metrics and competing methods

In this study, the performance of various link prediction methods are evaluated in terms of the area under receiver operating characteristic curve (AUC), which has been widely used in previous studies (Gönen, 2012; Liu *et al.*, 2016b; Mei *et al.*, 2013; Shen *et al.*, 2017; Xiao *et al.*, 2018; You *et al.*, 2017). To demonstrate the effectiveness of our method in predicting links in biomedical bipartite networks, we compare our proposed GRGMF method with the following seven state-of-the-art link prediction methods, namely, BLM-NII (Mei *et al.*, 2013), CMF (Zheng *et al.*, 2013), NRLMF (Liu *et al.*, 2016b), PBMDA (You *et al.*, 2017), CMFMDA (Shen *et al.*, 2017), DRCC (Gu and Zhou, 2009a) and GRNMF (Xiao *et al.*, 2018). These methods were originally designed for DTI prediction or MDA prediction and all of them can make use of the similarity matrices (i.e. S^A and S^B) derived from external databases. As deep learning models have gained progress in collaborative filtering, we further compare our proposed method with a deep matrix factorization model named DMF (Xue *et al.*, 2017). Note that DMF only takes the biadjacency matrix as input. To make use of the similarity matrices S^A and S^B, we also extend DMF to incorporate these two extra data and denote this extended variant as DMF-SIM.

Algorithm 1 GRGMF Algorithm

Input: The matrix of known connections $Y \in \{0,1\}^{m\times n}$; The similarity matrices of nodes in A and B, $S^A \in \mathbb{R}^{m\times m}$ and $S^B \in \mathbb{R}^{n\times n}$; Hyper parameters c, k, K, λ, β, r_1 and r_2.
Output: Predicted link matrix $\hat{Y}$.
1: Initialize Z^A and Z^B with S^{AN} and S^{BN} and normalize them by rows and columns, respectively;
2: Initialize U and V based on the initialization method used in Ezzat *et al.* (2017);
3: **While** not converged do
4: **Repeat**
5: Update U and V by using Adam optimizer;
6: **Until Convergence**
7: **Repeat**
8: Update Z^A and Z^B according to Equation (12);
9: **Until Convergence**
10: $P = \frac{1}{1+e^{(-Z^A U V^T Z^B)}}$;
11: Check the convergence conditions.
12: **End while**
13: $\hat{Y} = P$
14: **return** $\hat{Y}$

Table 1. The statistics of the six datasets

Dataset	Number of association	Dimension	Type	Sparsity (%)
NR	90	54 × 26	DTI	93.59
GPCR	635	223 × 95	DTI	97.00
IC	1476	210 × 204	DTI	96.55
E	2926	445 × 664	DTI	99.01
HMDD	5430	495 × 383	MDA	97.14
DTINet	1923	708 × 1512	DTI	99.82

3.3 Experimental settings

Following previous studies (Ezzat *et al.*, 2018; Liu *et al.*, 2016b; Xiao *et al.*, 2018), we evaluate the performance of various methods by performing 5-fold CV. In particular, for each 5-fold CV repetition, we calculate an AUC score for each method, and the final AUC score for each method is obtained by calculating the average AUC scores over 5 repetitions.

In this study, to evaluate the performance of various methods comprehensively, we consider the following three scenarios for CV experiments (Liu *et al.*, 2016b; Xiao *et al.*, 2018):

- CVS1: entries in Y are selected randomly for testing.
- CVS2: row vectors in Y are selected randomly for testing.
- CVS3: column vectors in Y are selected randomly for testing.

Under CVS1, in each round, 80% of the elements in Y are used for training and the remaining 20% are used for testing. Under CVS2 (or CVS3), in each round, 80% of the row vectors (or column vectors) in Y are used for training and the remaining 20% row vectors (or column vectors) are used for testing. The settings CVS2 and CVS3 can test the performance of various methods in predicting links for new nodes (e.g. new drugs or new targets) which do not have any known links. Since we use different random seeds, the information we used for training and testing is different for each repetition.

For all the compared methods, their hyper-parameters are selected from the range provided in the corresponding studies (Liu *et al.*, 2016b; Shen *et al.*, 2017; You *et al.*, 2017; Xue *et al.*, 2017) and their performances are obtained with the best-tuned parameters.

3.4 Results

Figure 2a–c shows the performance of various methods obtained under three different CV settings on six real datasets. Moreover, the detailed AUC scores of various methods are also provided in Supplementary Tables S1–S3. In Supplementary Tables S1–S3, highest score in each column is shown in bold face and the second highest score is underlined. As shown in Figure 2 and Supplementary Tables S1–S3, our proposed GRGMF method can achieve competitive performance on all these six datasets under all CV settings. For DTI datasets, GRGMF achieves comparative performance with NRLMF which is a popular DTI prediction method. While for miRNA-disease dataset HMDD, GRGMF outperforms all the other methods under all CV settings. Moreover, we observe that GRGMF could usually achieve good performance when the dataset is more challenging. For example, the size and sparsity level of E, HMDD and DTINet are larger than other three datasets, which makes it more difficult to predict links accurately. We can find from Figure 2 and Supplementary Tables S1–S3 that our GRGMF could achieve competitive performance on these three datasets. Overall, we can find from these experiment results that it is hard for a method to perform the best in all cases, but our GRGMF could always achieve the Top 2 performance on all datasets under all CV settings. These experiment results indicate the effectiveness of GRGMF in predicting potential links in biomedical bipartite networks.

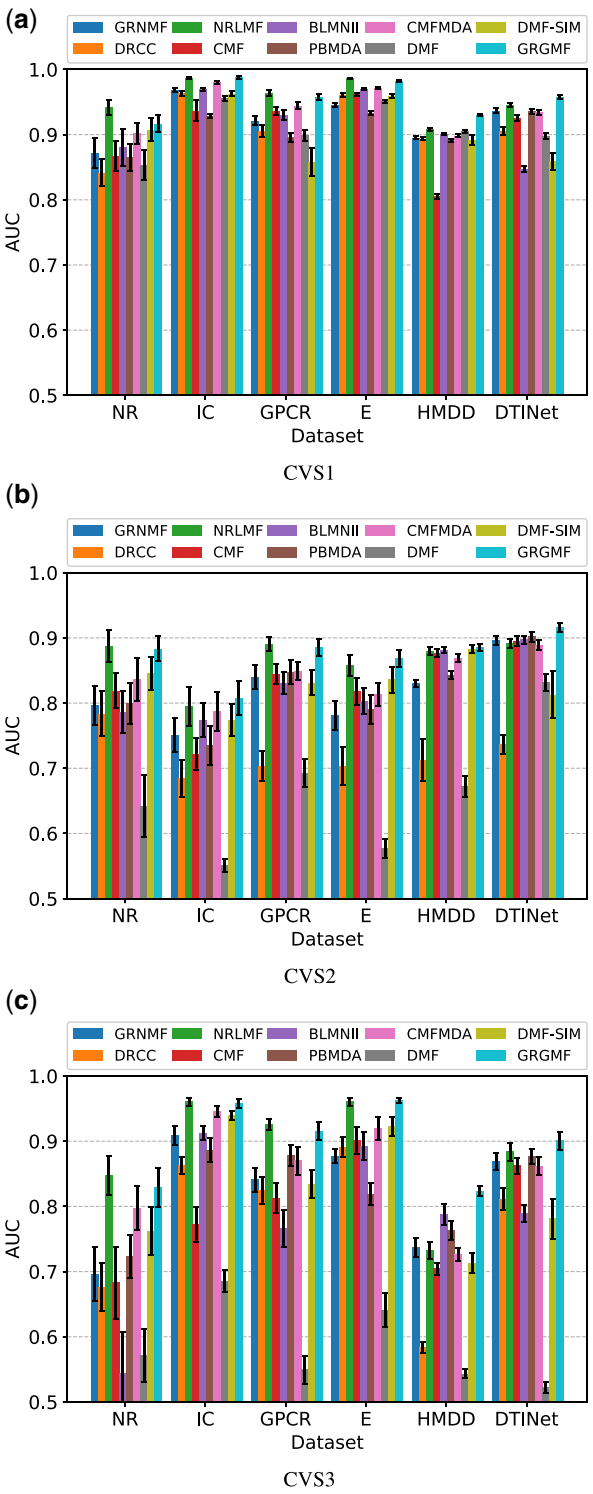

Fig. 2. Performance of various methods on six datasets under three different cross-validation settings. Error bars denote 95% confidence intervals. (**a**) CVS1. (**b**) CVS2. (**c**) CVS3

Note that our GRGMF learns the latent feature matrices U and V and the neighborhood information Z^A and Z^B simultaneously. We can derive a simplified variant of GRGMF, which learns U and V only and fixes Z^A and Z^B as S^A and S^B. Here, we denote this simplified variant as GRGMF−.

By comparing GRGMF and GRGMF−, we can further verify the benefit of learning Z^A and Z^B adaptively. Figure 3 and Supplementary Figures S1 and S2 show the performance of GRGMF and GRGMF− on six datasets. We can observe from these figures that GRGMF outperforms GRGMF− under different CV settings on all datasets, indicating that learning neighbor information adaptively could improve the performance for link prediction.

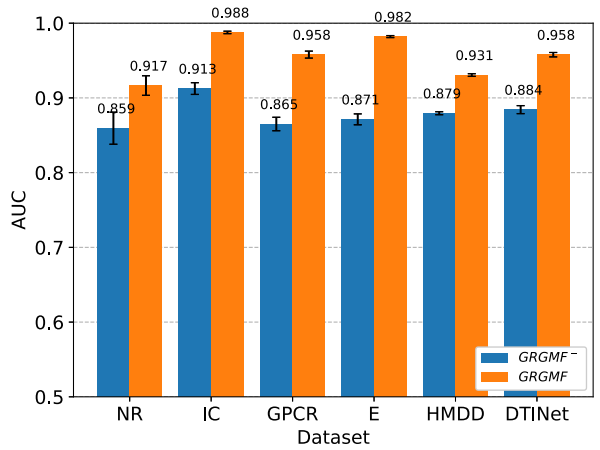

Fig. 3. This figure shows the AUC scores of GRGMF and GRGMF− on six benchmark datasets under CVS1. Error bars denote 95% confidence intervals

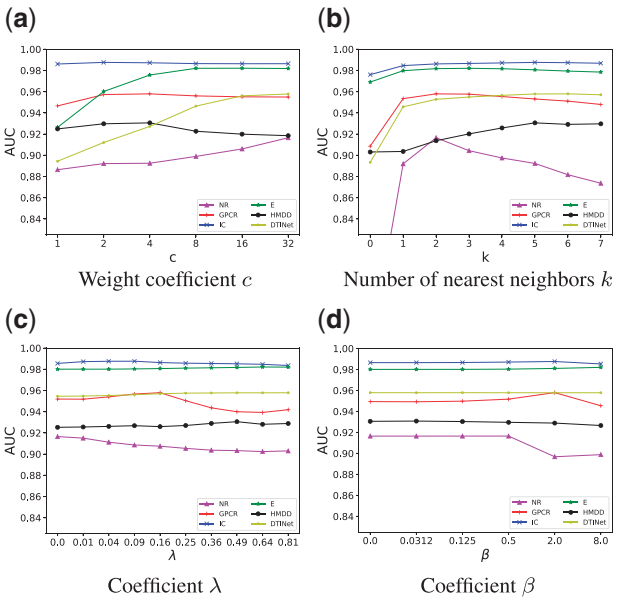

Fig. 4. Performance of GRGMF on six benchmark datasets with different values of c, k, λ and β under CVS1. (**a**) Weight coefficient c. (**b**) Number of nearest neighbors k. (**c**) Coefficient λ. (**d**) Coefficient β

3.5 Parameter analysis

There are several hyper-parameters in GRGMF that need to be tuned, i.e. c, k, K, λ, β, $r1$ and $r2$. We use random search strategy to select hyper-parameters from fixed ranges (Bergstra and Bengio, 2012). c is the weight assigned to known links. In this study, the value of c is select from $\{2^0, 2^1, 2^2, 2^3, 2^4, 2^5\}$. k is used to determine the number of nearest neighbors in S^{AN} and S^{BN}. Here, k is selected from $\{1, 2, \ldots, 7\}$. K denotes the dimensionality of the latent space. The value of K is selected from $\{25, 50, \ldots, 175, 200\}$. λ, β, r_1 and r_2 are four parameters that control the effect of regularization terms in Equation (9), we select λ from $\{0.1^2, 0.2^2, \ldots, 0.8^2, 0.9^2\}$, β from $\{0, 2^{-5}, 2^{-3}, 2^{-1}, 2^1, 2^3\}$, r_1 and r_2 from $\{2^{-3}, 2^{-2}, \ldots, 2^4, 2^5\}$. Next, we show the influence of these parameters under the setting CVS1.

To see how c and k affect the performance of our GRGMF, we change the values of c and k with the other hyper-parameters being

fixed and show the corresponding AUC scores of GRGMF in Figure 4a and b, respectively. We can find from these two figures that with the increase of the values of c and k, the performance of GRGMF increases initially and decreases or maintain stability after reaching the maximum. We can also find that when c is large enough, the performance of GRGMF tends to become saturated, which is consistent with previous studies (Liu *et al.*, 2016b). These results demonstrate that both c and k contribute to improving the performance of GRGMF. We also analyze the impact of K on the performance of GRGMF and show the results in Supplementary Figure S3. As shown in Supplementary Figure S3, GRGMF is not very sensitive to the value of K and larger K generally achieves better results.

Figure 4c and d shows the change of AUC scores along with the variation of a single hyper-parameter (λ or β), while the other hyper-parameters being fixed. We can observe from Figure 4c that when fixing the values of other parameters and increasing the value of λ, the performance of GRGMF increases initially and decreases after reaching the maximum. As shown in Figure 4d, for some datasets, as the value of β increases, the performance of GRGMF increases initially and decreases after reaching the maximum. We can also find from Figure 4d that for some datasets, such as DTINet, $\beta = 0$ could result in good performance, which means the L_2 regularization on Z^A and Z^B does not work. The L_2 regularization terms imposed on Z^A and Z^B are used to avoid the generation of trivial estimations for Z^A and Z^B (i.e. only the nearest node can be defined as the neighbor of a given node with probability 1). In some cases, even if there is no such constraint, our method will not generate such trivial estimations. That is why in some cases β has no effect of the performance of GRGMF. However, to enhance the performance of our model under different situations, we still keep this term in our model. λ controls the regularization on U and V, which could reduce the risk of overfitting. Overall, the results shown in Figure 4c and d demonstrates that both these two parameters contribute to the improvement of the performance of our model.

r_1 and r_2 control the influence of the graph regularization terms. Figure 5 and Supplementary Figures S4–S8 show the performances of GRGMF on NR, IC, GPCR, E, HMDD and DTINet in terms of AUC score with respect to different combinations of r_1 and r_2. As shown in these figures, for a fixed value of r_1, as the value of r_2 increases, the AUC scores increase initially and decrease after reaching the maximum. Similar properties can be observed with r_1. The results shown in these figures illustrate the necessity of introducing these regularization terms.

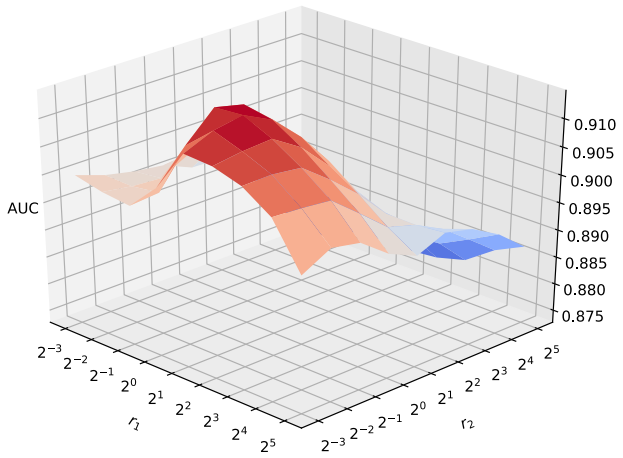

Fig. 5. Performance of GRGMF on NR dataset with different values of r_1 and r_2 under CVS1

4 Case studies

In this section, we conduct case studies to demonstrate the effectiveness of our GRGMF in predicting novel associations.

In our experiments, we focus on predicting novel MDAs using our GRGMF method. In particular, we adopt the optimal hyper-parameters which are obtained under CVS1 and use all the known MDAs to make prediction. For each disease, we rank the candidate miRNAs based on the predicted association scores. We collect three types of evidence to verify our predictions, which are three independent external databases, i.e. dbDEMC (Yang *et al.*, 2017), HMDD3.0 (Huang *et al.*, 2019) and miRCancer (Xie *et al.*, 2013). Note that HMDD3.0 refers to the HMDD v3.0, while we used HMDD v2.0 for training the GRGMF model. The full list of the candidate miRNAs of all diseases predicted by our GRGMF are provided in Supplementary Table S1. We select three diseases, namely, breast neoplasms, lung neoplasms and brain neoplasms, and list the Top 10 predicted candidate miRNAs in Table 2.

As shown in Table 2, the top 10 predicted miRNAs associated with breast neoplasms, lung neoplasms and brain neoplasms are all confirmed by independent external databases, except hsa-mir-92a. Aberrant expression of hsa-mir-92a has been observed in various

Table 2. The Top 10 associated miRNAs for breast neoplasms, lung neoplasms and brain neoplasms predicted by GRGMF

Cancer	Number of confirmed	Top 10 prediction					
		Rank	miRNA	Evidence	Rank	miRNA	Evidence
Breast neoplasms	10	1	hsa-mir-30e	dbDEMC HMDD3.0 miRCancer	6	hsa-mir-15b	dbDEMC HMDD3.0
		2	hsa-mir-130a	dbDEMC HMDD3.0 miRCancer	7	hsa-mir-130b	dbDEMC HMDD3.0
		3	hsa-mir-138	dbDEMC HMDD3.0 miRCancer	8	hsa-mir-302e	dbDEMC
		4	hsa-mir-106a	dbDEMC HMDD3.0	9	hsa-mir-372	dbDEMC HMDD3.0
		5	hsa-mir-98	dbDEMC HMDD3.0 miRCancer	10	hsa-mir-181d	dbDEMC
Lung neoplasms	10	1	hsa-mir-16	dbDEMC HMDD3.0 miRCancer	6	hsa-mir-92b	dbDEMC
		2	hsa-mir-106b	dbDEMC	7	hsa-mir-15a	dbDEMC HMDD3.0 miRCancer
		3	hsa-mir-149	dbDEMC HMDD3.0	8	hsa-mir-195	dbDEMC HMDD3.0 miRCancer
		4	hsa-mir-20b	dbDEMC	9	hsa-mir-23b	dbDEMC
		5	hsa-mir-129	dbDEMC HMDD3.0 miRCancer	10	hsa-mir-193b	dbDEMC
Brain neoplasms	9	1	hsa-mir-221	dbDEMC HMDD3.0	6	hsa-mir-125b	dbDEMC
		2	hsa-mir-155	dbDEMC	7	hsa-mir-92a	Unconfirmed
		3	hsa-mir-16	dbDEMC	8	hsa-mir-31	dbDEMC
		4	hsa-mir-146a	dbDEMC	9	hsa-mir-15a	dbDEMC
		5	hsa-mir-1	dbDEMC HMDD3.0	10	hsa-mir-145	dbDEMC

malignant tumors (Li *et al.*, 2014a). Hsa-mir-92a family has been found to play important roles in tumorigenesis and tumor progression (Li *et al.*, 2014a). Thus, hsa-mir-92a may be associated with the development and progression of brain neoplasms. These promising results demonstrate the effectiveness of GRGMF in predicting novel MDAs.

5 Discussion

In this study, we propose a new matrix factorization-based method, named GRGMF, for predicting potential links in biomedical bipartite networks. Our GRGMF could not only effectively utilize the observed links to predict potential links but also draw support from existing external similarity information to enhance the prediction performance. The experiment results on six real datasets demonstrate the effectiveness of our GRGMF in predicting potential links in biomedical bipartite networks. Moreover, case studies on predicted MDAs also demonstrate the effectiveness of our GRGMF in discovering novel MDAs.

Finally, we summarize our contributions as follows. First, we formulated a GMF model which learns the latent factor of each node based on its neighborhood information. Second, instead of utilizing the similarity matrices deriving from external-related databases with predefined metrics, our model could learn the neighbor information for each node adaptively and further promote the prediction of potential links. Third, we conduct extensive experiments, which demonstrate the effectiveness of the proposed GRGMF method.

Funding

This work was supported by the National Natural Science Foundation of China [61602309, 11871026, 61932008, 61772368], Shenzhen Fundamental Research Program [JCYJ20170817095210760], Guangdong Basic and Applied Basic Research Foundation [2019A1515011384], Natural Science Foundation of Hubei province [ZRMS2018001337], Natural Science Foundation of Shanghai [17ZR1445600], Shanghai Municipal Science and Technology Major Project [2018SHZDZX01] and ZJLab, and Chinese National-level Undergraduate Training Programs for Innovation and Entrepreneurship [201710590016].

Conflict of Interest: none declared.

References

Barabási,A.-L. *et al.* (2011) Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.*, **12**, 56–68.

Bergstra,J. and Bengio,Y. (2012) Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, **13**, 281–305.

Bleakley,K. and Yamanishi,Y. (2009) Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, **25**, 2397–2403.

Chen,X. *et al.* (2018) MDHGI: matrix decomposition and heterogeneous graph inference for mirna-disease association prediction. *PLoS Comput. Biol.*, **14**, e1006418.

Ding,C.H. *et al.* (2010) Convex and semi-nonnegative matrix factorizations. *IEEE Trans. Pattern Anal. Mach. Intell.*, **32**, 45–55.

Ezzat,A. *et al.* (2017) Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **14**, 646–656.

Ezzat,A. *et al.* (2018) Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief. Bioinform.*, **20**, 1337–1357.

Fu,G. *et al.* (2018) Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics*, **34**, 1529–1537.

Gönen,M. (2012) Predicting drug–target interactions from chemical and genomic kernels using bayesian matrix factorization. *Bioinformatics*, **28**, 2304–2310.

Gu,Q. and Zhou,J. (2009a) Co-clustering on manifold. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 359–368. ACM.

Gu,Q. and Zhou,J. (2009b) Transductive classification via dual regularization. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 439–454. Springer.

He,X. *et al.* (2017) Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 173–182. International World Wide Web Conferences Steering Committe.

Huang,Z. *et al.* (2019) Hmdd v3. 0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res.*, **47**, D1013–D1017.

Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res.*, **18**, 644–652.

Johnson,C.C. (2014) Logistic matrix factorization for implicit feedback data. *Adv. Neural Inf. Process. Syst.*, **27**, **1–9**.

Kingma,D.P. and Ba,J. (2014) Adam: A method for stochastic optimization. *In: Bengio,Y. and LeCun,Y. (eds.). 3rd International Conference on Learning Representations, {ICLR} 2015, San Diego, CA*.

Klema,V. and Laub,A. (1980) The singular value decomposition: its computation and some applications. *IEEE Trans. Automat. Contr.*, **25**, 164–176.

Li,M. *et al.* (2014a) miR-92a family and their target genes in tumorigenesis and metastasis. *Exp. Cell Res.*, **323**, 1–6.

Li,Y. *et al.* (2014b) HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.*, **42**, D1070–D1074.

Liu,Y. *et al.* (2016a) Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **14**, 905–915.

Liu,Y. *et al.* (2016b) Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput. Biol.*, **12**, e1004760.

Luo,H. *et al.* (2016) Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics*, **32**, 2664–2671.

Luo,Y. *et al.* (2017) A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nat. Commun.*, **8**, 573.

Mei,J.-P. *et al.* (2013) Drug–target interaction prediction by learning from local information and neighbors. *Bioinformatics*, **29**, 238–245.

Pavlopoulos,G.A. *et al.* (2018) Bipartite graphs in systems biology and medicine: a survey of methods and applications. *Gigascience*, **7**, 1–31.

Shen,Z. *et al.* (2017) miRNA-disease association prediction with collaborative matrix factorization. *Complexity*, **2017**, 1–9.

van Dam,S. *et al.* (2017) Gene co-expression analysis for functional classification and gene–disease predictions. *Brief. Bioinform.*, **19**, 575–592.

Wang,D. *et al.* (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, **26**, 1644–1650.

Wang,Q. *et al.* (2018) Imputing structured missing values in spatial data with clustered adversarial matrix factorization. In *2018 IEEE International Conference on Data Mining (ICDM)*, pp. 1284–1289. IEEE.

Wang,Z. *et al.* (2014) Rank-one matrix pursuit for matrix completion. In *International Conference on Machine Learning*, pp. 91–99.

Xiao,Q. *et al.* (2018) A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics*, **34**, 239–248.

Xie,B. *et al.* (2013) miRCancer: a microRNA–cancer association database constructed by text mining on literature. *Bioinformatics*, **29**, 638–644.

Xuan,P. *et al.* (2015) Prediction of potential disease-associated microRNAs based on random walk. *Bioinformatics*, **31**, 1805–1815.

Xue,H.-J. *et al.* (2017) Deep matrix factorization models for recommender systems. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pp. 3203–3209.

Yamanishi,Y. *et al.* (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.

Yang,Z. and Oja,E. (2012a) Clustering by low-rank doubly stochastic matrix decomposition. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pp. 707–714. Omnipress.

Yang,Z. and Oja,E. (2012b) Clustering by low-rank doubly stochastic matrix decomposition. In: *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, Edinburgh, Scotland, pp. 831–838. Omnipress, New York, NY.

Yang,Z. *et al.* (2017) dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers. *Nucleic Acids Res.*, **45**, D812–D818.

You,Z.-H. *et al.* (2017) PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput. Biol.*, **13**, e1005455.

Zheng,X. *et al.* (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*, pp. 1025–1033. ACM.