Briefings in Bioinformatics, 21(4), 2020, 1356–1367

doi: 10.1093/bib/bbz057
Advance Access Publication Date: 4 June 2019
Problem solving protocol

iCircDA-MF: identification of circRNA-disease associations based on matrix factorization

Hang Wei and Bin Liu

Corresponding author: Bin Liu, Harbin Institute of Technology, HIT Campus Shenzhen University Town, Xili, Shenzhen, 518055, China and School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. Tel: (+86) 0755-26033283; E-mail: bliu@insun.hit.edu.cn

Abstract

Circular RNAs (circRNAs) are a group of novel discovered non-coding RNAs with closed-loop structure, which play critical roles in various biological processes. Identifying associations between circRNAs and diseases is critical for exploring the complex disease mechanism and facilitating disease-targeted therapy. Although several computational predictors have been proposed, their performance is still limited. In this study, a novel computational method called iCircDA-MF is proposed. Because the circRNA-disease associations with experimental validation are very limited, the potential circRNA-disease associations are calculated based on the circRNA similarity and disease similarity extracted from the disease semantic information and the known associations of circRNA-gene, gene-disease and circRNA-disease. The circRNA-disease interaction profiles are then updated by the neighbour interaction profiles so as to correct the false negative associations. Finally, the matrix factorization is performed on the updated circRNA-disease interaction profiles to predict the circRNA-disease associations. The experimental results on a widely used benchmark dataset showed that iCircDA-MF outperforms other state-of-the-art predictors and can identify new circRNA-disease associations effectively.

Key words: circRNA-disease associations; matrix factorization; circRNA similarity; disease similarity

Introduction

As newly discovered non-coding RNAs, circular RNAs (circRNAs) have attracted more and more attention. Comparing with other linear non-coding RNAs such as microRNAs (miRNAs) and long non-coding RNAs (lncRNAs), circRNAs lack 5′ and 3′ polyadenylated tails and form a covalently closed continuous loop structure [1].

Existing experimental results indicate that non-coding RNAs play crucial roles in many biological processes [2, 3]. For example, miRNAs normally regulate gene expression via base-pairing with complementary sequences of mRNAs [4]. CircRNAs exert biological functions by acting as miRNA sponges [5], regulators of RNA binding proteins [6]. Several non-coding RNAs have been confirmed to implicate in the developments of various diseases [7, 8]. For example, CDR1as is one of the earliest well-studied circRNAs, which can regulate miRNAs in tumour cells [9], and it is significantly differentially expressed in many diseases including hepatocellular carcinoma [10], colorectal cancer [11, 12] and neurological disorders [13]. Therefore, identifying potential associations between non-coding RNAs and diseases draws increasing attention to understand the complex disease mechanism and discover therapeutic targets. In the past decade, increasing experimentally supported associations of miRNA-disease and lncRNA-disease have been detected [14–17]; meanwhile, many computational approaches have been proposed

Hang Wei is a PhD candidate at the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. Her expertise is in bioinformatics.
Bin Liu is a professor at the School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, China, and School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China. His expertise is in bioinformatics, nature language processing and machine learning.
Submitted: 31 January 2019; **Received (in revised form):** 13 March 2019

to discover new associations of miRNA-disease [18, 19] and lncRNA-disease [20]. Compared with linear non-coding RNA-disease, the research on circRNA-disease is still in the preliminary stage. The recently constructed experimentally supported circRNA-disease association databases (CircR2Disease [21], Circ2Disease [22], circRNADisease [23]) provide an opportunity to identify the potential circRNA-disease associations via the computational approaches. However, available experimental verified circRNA-disease associations and circRNA-related information are limited. It is still challenging to develop effective predictors to identify circRNA-disease associations from the available circRNA-related information and known circRNA-disease associations.

Recently, three computational predictors have been proposed to identify circRNA-disease associations. PWCDA calculates an association score for each circRNA-disease pair based on paths connecting them in a heterogeneous network [24], and DWNN-RLS predicts circRNA-disease associations based on regularized least squares of Kronecker product kernel [25]. KATZHCDA identifies potential circRNA-disease associations by considering the number of walks between nodes and walk length in a heterogeneous network [26]. All the three models have obtained encouraging results, and play important roles in the development of computational methods for circRNA-disease association identification. However, they are suffering from certain problems or limitations: (i) The existing predictors are based on incompletely related biological information, failing to accurately measure the circRNA similarity and disease similarity. (ii) The experimentally verified circRNA-disease associations are very limited, and there are many false negative associations and negative associations. Performed on such noisy and spare circRNA-disease association network, the predictors tend to detect many false negative associations.

This study is initiated in an attempt to overcome these problems by developing a novel computational predictor for identifying circRNA-disease associations. The proposed predictor is called iCircDA-MF. The iCircDA-MF predictor has the following advantages: (i) Disease semantic information and the known associations of circRNA-gene, gene-disease, circRNA-disease are employed to accurately measure the circRNA similarity and disease similarity. (ii) The false negative associations can be detected and corrected via the neighbour interaction profiles based on the circRNA similarity and disease similarity. (iii) Matrix factorization is employed to extract the latent features from the noisy and spare circRNA-disease association network. The experimental results on a benchmark dataset showed that iCircDA-MF outperformed other competing methods, and can effectively predict novel circRNA-disease associations. Furthermore, we showed that the performance of iCircDA-MF was underestimated because of the limitation of the known associations with experimental validation.

Materials and Methods

Benchmark dataset

A recently established benchmark dataset circR2Disease [21] based on the experimentally verified circRNA–disease associations was employed to evaluate the performance of various methods. After removing duplicate and non-human circRNA-disease associations, 649 experimentally verified circRNA-disease associations were obtained, including 583 circRNAs and 88 diseases. The benchmark dataset can be represented as

$$\mathbb{S} = \mathbb{S}^{+} \bigcup \mathbb{S}^{-}, \tag{1}$$

where $\bigcup$ is the symbol for union in the set theory, $\mathbb{S}^{+}$ denotes the positive subset that contains 649 experimentally verified circRNA-disease associations and $\mathbb{S}^{-}$ is the negative subset that contains 50 655 circRNA-disease associations without experimental validation. The benchmark dataset $\mathbb{S}$ is given in Supporting Information S1.

Method overview

In this study, a novel method called iCircDA-MF is proposed to identify potential circRNA-disease associations. The iCircDA-MF consists of three steps, and its framework is shown in Figure 1. Firstly, the similarity matrix of diseases is constructed based on disease semantic information and known circRNA-disease associations. The similarity matrix of circRNAs is constructed based on known associations of circRNA-gene, gene-disease and circRNA-disease. Secondly, reformulated circRNA-disease association adjacency matrix is generated based on the interaction profiles of similar circRNAs and similar diseases. Finally, the matrix factorization is exploited to compute the circRNA-disease association scores.

Construct disease similarity matrix

Disease similarity based on Gaussian interaction profile kernel

Gaussian interaction profile (GIP) kernel was proposed to calculate the network topologic similarity of biological entities [27], and it has been widely applied to construct disease similarity matrix [28–30]. According to the biological assumption that similar diseases show similar interaction patterns with circRNAs [24–26], the topologic information of circRNA-disease association network can be used to measure disease similarity. Therefore, GIP kernel was used to calculate the similarity between diseases $\mathbf{D}_{\mathbf{Gip}}\left(d_i, d_j\right)$ as follows [27]:

$$\mathbf{D}_{\mathbf{Gip}}\left(d_i, d_j\right) = \exp\left(-\gamma_d \left\|\mathbf{A}\left(d_i\right) - \mathbf{A}\left(d_j\right)\right\|^2\right), \tag{2}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix} \tag{3}$$

$$\gamma_d = \gamma_d{}' / \left(\frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{A}\left(d_i\right)\right\|^2\right), \tag{4}$$

where the binary vector $\mathbf{A}\left(d_i\right)$ is the ith column in matrix $\mathbf{A}$, which is the interaction profile of disease d_i representing whether disease d_i is associated with each circRNA or not. γ_d is a regulation parameter, which controls the kernel bandwidth.$\gamma_d{}'$ is the original bandwidth and is defined as 1 according to the previous study [27]. $\mathbf{A}$ is the adjacency matrix of circRNA-disease association dataset $\mathbb{S}$ (cf. (1)). m is the number of circRNAs, and n is the number of diseases. The element $a_{i,j}$ in row i column j of $\mathbf{A}$ is 1 if circRNA c_i is related to the disease d_j with experimental validation, otherwise 0. The size of $\mathbf{D}_{\mathbf{Gip}}$ is $n \times n$, and $\mathbf{D}_{\mathbf{Gip}}\left(d_i, d_j\right)$ represents the GIP kernel similarity score between d_i and d_j.

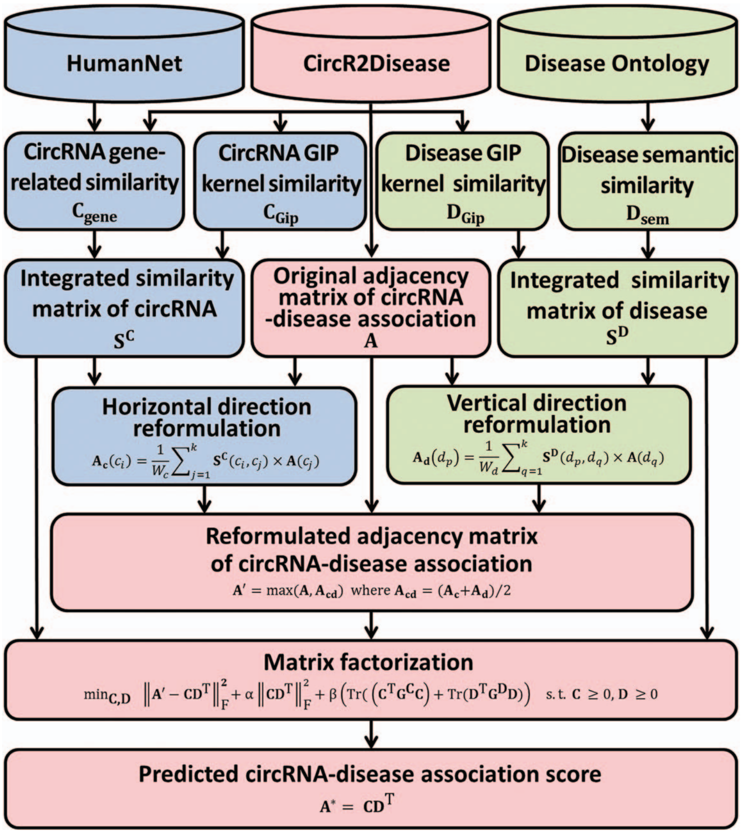
HumanNet
CircR2Disease
Disease Ontology
CircRNA gene-related similarity $\mathbf{C}_{\text{gene}}$
CircRNA GIP kernel similarity $\mathbf{C}_{\text{Gip}}$
Disease GIP kernel similarity $\mathbf{D}_{\text{Gip}}$
Disease semantic similarity $\mathbf{D}_{\text{sem}}$
Integrated similarity matrix of circRNA $\mathbf{S}^{\mathrm{C}}$
Original adjacency matrix of circRNA -disease association $\mathbf{A}$
Integrated similarity matrix of disease $\mathbf{S}^{\mathrm{D}}$
Horizontal direction reformulation
$\mathbf{A}_{\mathbf{c}}(c_i) = \frac{1}{W_c}\sum_{j=1}^{k} \mathbf{S}^{\mathrm{C}}(c_i, c_j) \times \mathbf{A}(c_j)$
Vertical direction reformulation
$\mathbf{A}_{\mathbf{d}}(d_p) = \frac{1}{W_d}\sum_{q=1}^{k} \mathbf{S}^{\mathrm{D}}(d_p, d_q) \times \mathbf{A}(d_q)$
Reformulated adjacency matrix of circRNA-disease association
$\mathbf{A}' = \max(\mathbf{A}, \mathbf{A}_{\mathbf{cd}})$ where $\mathbf{A}_{\mathbf{cd}} = (\mathbf{A}_{\mathbf{c}} + \mathbf{A}_{\mathbf{d}})/2$
Matrix factorization
$\min_{\mathbf{C},\mathbf{D}} \left\|\mathbf{A}' - \mathbf{C}\mathbf{D}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} + \alpha \left\|\mathbf{C}\mathbf{D}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} + \beta \left(\mathrm{Tr}(\left(\mathbf{C}^{\mathrm{T}}\mathbf{G}^{\mathrm{C}}\mathbf{C}\right) + \mathrm{Tr}(\mathbf{D}^{\mathrm{T}}\mathbf{G}^{\mathrm{D}}\mathbf{D})\right)$ s.t. $\mathbf{C} \geq 0, \mathbf{D} \geq 0$
Predicted circRNA-disease association score
$\mathbf{A}^{*} = \mathbf{C}\mathbf{D}^{\mathrm{T}}$

Figure 1. The framework of iCircDA-MF.

Disease similarity based on disease semantic

Disease can be represented as a directed acyclic graph (DAG) based on its semantic associations of terms in ontology, where nodes denote diseases and edges denote associations between diseases. Similar diseases will share more common parts of DAGs. One of the most popular methods for computing the similarity of terms in ontology was proposed by Wang *et al*. [31], and it has been widely utilized for computing the semantic similarity of diseases in recent years [29, 32–34]. In this study, disease semantic related annotation terms were obtained from the Disease Ontology [35]. The disease semantic similarity matrix $\mathbf{D}_{\text{sem}}$ can be calculated by [32–34]:

$$\mathbf{D}_{\text{sem}}\left(d_i, d_j\right) = \frac{\sum_{t \in \mathrm{T}_{d_i} \cap \mathrm{T}_{d_j}} \left(\mathrm{S}_{d_i}(t) + \mathrm{S}_{d_j}(t)\right)}{\sum_{t \in \mathrm{T}_{d_i}} \mathrm{S}_{d_i}(t) + \sum_{t \in \mathrm{T}_{d_j}} \mathrm{S}_{d_j}(t)}, \tag{5}$$

where $\mathbf{D}_{\text{sem}}\left(d_i, d_j\right)$ denotes the semantic similarity score between disease d_i and d_j. T_d denotes the set of diseases in the DAG graph of disease d, and $\mathrm{S}_d(t)$ denotes the semantic contribution of disease $t \in \mathrm{T}_d$ related to disease d calculated by

$$\begin{cases} \mathrm{S}_d(t) = \max\left\{\omega * \mathrm{S}_d\left(t'\right) | t' \in \text{children of } (t)\right\} & \text{if } t \neq d \\ \mathrm{S}_d(t) = 1 & \text{otherwise} \end{cases}, \tag{6}$$

where ω is the semantic contribution factor, which is set as 0.5 following the previous study [33].

Construct integrated disease similarity matrix

Measuring disease similarity from different perspectives can generate a more comprehensive view of similarity between diseases. Based on the aforementioned similarity measures, the two disease similarity matrices $\mathbf{D}_{\text{Gip}}$(cf. (2)) and $\mathbf{D}_{\text{sem}}$(cf. (5)) can be obtained and integrated to provide a more comprehensive disease similarity, where $\mathbf{D}_{\text{Gip}}$ represents the network topologic similarity between diseases and $\mathbf{D}_{\text{sem}}$ describes the semantic similarity between diseases.

Because some diseases have no corresponding disease semantic terms, we merged the two similarity matrices following previous studies [36, 37] so as to complement the missing similarity information of $\mathbf{D}_{\text{sem}}$. The final disease similarity matrix $\mathbf{S}^{\mathrm{D}}$ was computed by [36, 37]:

$$\mathbf{S}^{\mathrm{D}}\left(d_i, d_j\right) = \begin{cases} \frac{\mathbf{D}_{\text{sem}}\left(d_i, d_j\right) + \mathbf{D}_{\text{Gip}}\left(d_i, d_j\right)}{2} & \text{if } \mathbf{D}_{\text{sem}}\left(d_i, d_j\right) \neq 0 \\ \mathbf{D}_{\text{Gip}}\left(d_i, d_j\right) & \text{otherwise} \end{cases}. \tag{7}$$

The size of $\mathbf{S}^{\mathrm{D}}$ is $n \times n$; each element in the matrix $\mathbf{S}^{\mathrm{D}}$ indicates the similarity score between diseases.

Construct circRNA similarity matrix

CircRNA similarity based on GIP kernel

GIP kernel has also been widely utilized to calculate biomolecule similarity [38–40]. Because similar circRNAs tend to show similar associations with diseases [24–26], the circRNA similarity matrix $\mathbf{C}_{\text{Gip}}\left(c_i, c_j\right)$ was calculated by GIP kernel [27] based on the circRNA-disease association network:

$$\mathbf{C}_{\mathbf{Gip}}\left(c_i, c_j\right) = \exp\left(-\gamma_c \left\| \mathbf{A}\left(c_i\right) - \mathbf{A}\left(c_j\right) \right\|^2\right) \tag{8}$$

$$\gamma_c = \gamma_c{}' / \left(\frac{1}{m} \sum_{i=1}^{m} \left\| \mathbf{A}\left(c_i\right) \right\|^2 \right), \tag{9}$$

where the binary vector $\mathbf{A}\left(c_i\right)$ is the ith row in the adjacency matrix **A** (cf. (3)), which is the interaction profile of circRNA c_i representing whether circRNA c_i is associated with each disease or not. γ_c is a regulation parameter, which controls the kernel bandwidth.$\gamma_c{}'$ is the original bandwidth and is defined as 1 based on [27]. The size of $\mathbf{C}_{\mathbf{Gip}}$ is $m \times m$, and $\mathbf{C}_{\mathbf{Gip}}\left(c_i, c_j\right)$ represents the GIP kernel similarity score between circRNA c_i and c_j.

CircRNA similarity based on gene-related similarity

It is known that circRNAs can act as competing endogenous RNAs to regulate genes [41]. Because similar RNAs tend to regulate similar genes, genes have been widely used to infer RNA similarity [42, 43]. In this study, gene-disease associations were collected from HumanNet v2 [44, 45]. CircRNAs and their related gene target information were downloaded from circR2Disease [21]. Considering the close relevance between genes and circRNAs, circRNA gene-related similarity matrix $\mathbf{C}_{\mathbf{gene}}$was calculated by

$$\mathbf{C}_{\mathbf{gene}} = \mathbf{R} \times \mathbf{G} \times \mathbf{R}^{\mathrm{T}}, \tag{10}$$

where T is the transpose operator. **R** is a matrix representing the relevance of circRNAs and genes; the elements in **R** are 1 if circRNAs and genes are relevant, otherwise they are 0. **G** is the GIP kernel similarity matrix for genes, which is computed based on gene-disease associations. The element in $\mathbf{C}_{\mathbf{gene}}$ represents similarity score between each two circRNAs from their gene regulation perspective.

Construct integrated circRNA similarity matrix

Through the above similarity measures, two circRNA similarity matrices can be obtained, including $\mathbf{C}_{\mathbf{Gip}}$ (cf. (8)),$\mathbf{C}_{\mathbf{gene}}$ (cf. (10)). However, $\mathbf{C}_{\mathbf{gene}}$ is incompleteness because some circRNAs in the circRNA-disease association dataset $\mathbb{S}$ (cf. (1)) lack the corresponding experimentally supported circRNA-gene associations. In this regard, the two similarity matrices were merged to complement the missing similarity information by [36, 37]:

$$\mathbf{S}^{\mathrm{C}}\left(c_i, c_j\right) = \begin{cases} \frac{\mathbf{C}_{\mathbf{gene}}\left(c_i, c_j\right) + \mathbf{C}_{\mathbf{Gip}}\left(c_i, c_j\right)}{2} & \text{if } \mathbf{C}_{\mathbf{gene}}\left(c_i, c_j\right) \neq 0 \\ \mathbf{C}_{\mathbf{Gip}}\left(c_i, c_j\right) & \text{otherwise} \end{cases}. \tag{11}$$

The size of $\mathbf{S}^{\mathrm{C}}$ is $m \times m$; each element in the matrix $\mathbf{S}^{\mathrm{C}}$ indicates the similarity score between circRNAs.

Reformulate circRNA-disease association adjacency matrix

The elements in the original circRNA-disease association adjacency matrix **A** (cf. (3)) are binary values, representing whether circRNAs are associated with diseases or not. Because the known circRNA-disease associations are still limited, there are many false negative associations whose values are zero in **A** (cf. (3)). In order to reduce the noise, the circRNA-disease association adjacency matrix **A** (cf. (3)) was reformulated based on [46].

The vertical direction reformulation

For each disease d_p, all other diseases were ranked in descending order according to their similarities with d_p. In the reformulated circRNA-disease association adjacency matrix $\mathbf{A}_{\mathbf{d}}$, the interaction profile for d_p was calculated based on the corresponding interaction profiles of the top k similar diseases with d_p [46]:

$$\mathbf{A}_{\mathbf{d}}\left(d_p\right) = \frac{1}{\mathrm{W}_d} \sum\nolimits_{q=1}^{k} \mathbf{S}^{\mathrm{D}}\left(d_p, d_q\right) \times \mathbf{A}\left(d_q\right), \tag{12}$$

where $\mathrm{W}_d = \sum_{1 \le q \le k} \mathbf{S}^{\mathrm{D}}\left(d_p, d_q\right)$, $\mathbf{S}^{\mathrm{D}}$ (cf. (7)) is the disease similarity matrix. $\mathbf{A}\left(d_q\right)$ represents the corresponding interaction profile of disease d_q, which is the qth column of original association adjacency matrix **A** (cf. (3)).

The horizontal direction reformulation

For each circRNA c_i, all other circRNAs were ranked in descending order according to their similarities with c_i. In the reformulated circRNA-disease association adjacency matrix $\mathbf{A}_{\mathbf{c}}$, the interaction profile for c_i was calculated based on the corresponding interaction profiles of the top k similar circRNAs with c_i [46]:

$$\mathbf{A}_{\mathbf{c}}\left(c_i\right) = \frac{1}{\mathrm{W}_c} \sum\nolimits_{j=1}^{k} \mathbf{S}^{\mathrm{C}}\left(c_i, c_j\right) \times \mathbf{A}\left(c_j\right), \tag{13}$$

where $\mathrm{W}_c = \sum_{1 \le j \le k} \mathbf{S}^{\mathrm{C}}\left(c_i, c_j\right)$, $\mathbf{S}^{\mathrm{C}}$ (cf. (11)) is the circRNA similarity matrix. $\mathbf{A}\left(c_j\right)$ represents the corresponding interaction profile of circRNA $c_{i,}$ which is the ith row of original association adjacency matrix **A** (cf. (3)).

Final reformulation

The two reformulated matrices $\mathbf{A}_{\mathbf{d}}$ and $\mathbf{A}_{\mathbf{c}}$ were merged by $\mathbf{A}_{\mathbf{cd}} = \left(\mathbf{A}_{\mathbf{c}} + \mathbf{A}_{\mathbf{d}}\right)/2$. The final reformulated circRNA-disease association adjacency matrix $\mathbf{A}'$ was calculated by

$$\mathbf{A}' = \max\left(\mathbf{A}, \mathbf{A}_{\mathbf{cd}}\right). \tag{14}$$

Predict circRNA-disease association by matrix factorization

As matrix factorization can detect the intrinsic structure of data effectively; it shows powerful ability to identify meaningful information in recommendation problems [47–49]. In recent years, identifying associations between two biological entities can be considered as recommendation problem. Matrix factorization has been successfully applied to detect potential associations of drug-target [46, 50, 51], miRNA-disease [37, 38, 52–54], lncRNA-disease [55, 56], etc. In this study, iCircDA-MF is based on basic non-negative matrix factorization, which can detect the latent features from sparse matrix and ensure the non-negativity of the predictive association scores. The reformulated circRNA-disease association adjacency matrix $\mathbf{A}'$ (cf. (14)) can be decomposed into two low-dimension matrices via optimizing the following objective function [57, 58]:

$$\min_{\mathbf{C},\mathbf{D}} \quad \left\| \mathbf{A}' - \mathbf{C}\mathbf{D}^{\mathrm{T}} \right\|_{\mathrm{F}}^{2} \quad \text{s.t. } \mathbf{C} \ge 0, \mathbf{D} \ge 0, \tag{15}$$

where **C** and **D** are the latent feature matrices of circRNAs and diseases, and their dimensions are $m \times r$ and $n \times r$, respectively. r is the subspace dimensionality, T is the transpose operator and $\left\| \bullet \right\|_F$ represents Frobenius norm [59].

Because similar circRNAs tend to be associated with same or similar diseases and vice versa, two biological constraint terms are considered. For circRNAs, if two circRNAs are similar, the distance between their latent feature vectors is close. Two data points in the disease latent feature space are close to each other if their corresponding diseases are similar. It has been proved that the graph Laplacian can help reduce noise [60] and enhance predictive performance [54] by constraining the geometrical structure of latent feature space [61]. To further reduce the noise in $\mathbf{A}'$ (cf. (14)) and ensure the biological meaning of the latent feature spaces, two biological constraint terms based on graph regularization are introduced and formularized as follows:

$$\begin{array}{l} \frac{1}{2}\sum_{i,j=1}^{m} \left\|\mathbf{C}\left(c_i\right) - \mathbf{C}\left(c_j\right)\right\|^2 \mathbf{S}_{ij}^{\mathrm{C}} = \mathrm{Tr}\left(\mathbf{C}^{\mathrm{T}}\mathbf{G}^{\mathrm{C}}\mathbf{C}\right) \\ \frac{1}{2}\sum_{p,q=1}^{n} \left\|\mathbf{D}\left(d_p\right) - \mathbf{D}\left(d_q\right)\right\|^2 \mathbf{S}_{pq}^{\mathrm{D}} = \mathrm{Tr}\left(\mathbf{D}^{\mathrm{T}}\mathbf{G}^{\mathrm{D}}\mathbf{D}\right) \end{array}, \tag{16}$$

where the Euclidean norm is used to measure the distance in each latent feature space. $\mathbf{S}^{\mathrm{C}}$ (cf. (11)) and $\mathbf{S}^{\mathrm{D}}$ (cf. (7)) are circRNA similarity matrix and disease similarity matrix. $\mathbf{C}\left(c_i\right)$ and $\mathbf{D}\left(d_p\right)$ are the ith and pth rows of **C** and **D**, representing the feature vectors for circRNA c_i and disease d_p. $\mathrm{Tr}\left(\bullet\right)$ represents the trace of a matrix. $\mathbf{G}^{\mathrm{C}} = \mathbf{I}^{\mathrm{C}} - \mathbf{S}^{\mathrm{C}}$ and $\mathbf{G}^{\mathrm{D}} = \mathbf{I}^{\mathrm{D}} - \mathbf{S}^{\mathrm{D}}$ are the graph Laplacian matrices for circRNA similarity matrix and disease similarity matrix, respectively. $\mathbf{I}^{\mathrm{C}}$ and $\mathbf{I}^{\mathrm{D}}$ are two diagonal matrices; the elements in $\mathbf{I}^{\mathrm{C}}$ and $\mathbf{I}^{\mathrm{D}}$ are row sums of $\mathbf{S}^{\mathrm{C}}$ and $\mathbf{S}^{\mathrm{D}}$, respectively. Finally, Frobenius norm regularization term is added to avoid overfitting and ensure the smoothness of target space. Therefore, integrating basic objective function (cf. (15)), two graph regularization terms (cf. (16)) and Frobenius norm regularization term, the final objective function in iCircDA-MF is formularized as follows:

$$\begin{array}{l} \min_{\mathbf{C},\mathbf{D}} \quad \left\|\mathbf{A}' - \mathbf{C}\mathbf{D}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \alpha\left\|\mathbf{C}\mathbf{D}^{\mathrm{T}}\right\|_{\mathrm{F}}^2 + \beta\left(\mathrm{Tr}\left(\mathbf{C}^{\mathrm{T}}\mathbf{G}^{\mathrm{C}}\mathbf{C}\right) + \mathrm{Tr}\left(\mathbf{D}^{\mathrm{T}}\mathbf{G}^{\mathrm{D}}\mathbf{D}\right)\right) \\ \text{s.t.} \ \ \mathbf{C} \geq 0, \mathbf{D} \geq 0 \end{array}, \tag{17}$$

where α and β are regularization coefficients. We solved the optimization problem by introducing Lagrange multipliers and Karush–Kuhn–Tucker conditions [62]. Through calculating the partial derivative for **C** and **D**, the updating rules for $\mathbf{C}\left(c_{it}\right)$ and $\mathbf{D}\left(d_{pt}\right)$ are as follows:

$$\begin{array}{l} \mathbf{C}\left(c_{it}\right) = \mathbf{C}\left(c_{it}\right) \frac{\left(\mathbf{A}'\mathbf{D} + \beta\mathbf{S}^{\mathrm{C}}\mathbf{C}\right)_{it}}{\left((\alpha+1)\mathbf{C}\mathbf{D}^{\mathrm{T}}\mathbf{D} + \beta\mathbf{I}^{\mathrm{C}}\mathbf{C}\right)_{it}} \\ \mathbf{D}\left(d_{pt}\right) = \mathbf{D}\left(d_{pt}\right) \frac{\left(\left(\mathbf{A}'\right)^{\mathrm{T}}\mathbf{C} + \beta\mathbf{S}^{\mathrm{D}}\mathbf{D}\right)_{pt}}{\left((\alpha+1)\mathbf{D}\mathbf{C}^{\mathrm{T}}\mathbf{C} + \beta\mathbf{I}^{\mathrm{D}}\mathbf{D}\right)_{pt}} \end{array}. \tag{18}$$

The low-dimension latent feature matrices **C** and **D** were updated via (18) until convergence. Finally, the predicted circRNA-disease association adjacency matrix was obtained by $\mathbf{A}^* = \mathbf{C}\mathbf{D}^{\mathrm{T}}$. The larger value of the element in $\mathbf{A}^*$ represents the higher relevance between the corresponding circRNA and disease.

Performance evaluation

In this study, to evaluate the performance of iCircDA-MF on identifying circRNA-disease associations, two cross-validation approaches were implemented, including 5-fold cross-validation and disease-specific cross-validation. For K-fold cross-validation ($K = 5$ in this study) [63], benchmark dataset $\mathbb{S}$ (cf. (1)) is randomly divided into K subsets as follows:

$$\left\{ \begin{array}{l} \mathbb{S} = \mathbb{S}_1 \cup \mathbb{S}_2 \cup \cdots \cup \mathbb{S}_i \cdots \cup \mathbb{S}_K = \bigcup_{i=1}^{K} \mathbb{S}_i \\ \varnothing = \mathbb{S}_1 \cap \mathbb{S}_2 \cap \cdots \cap \mathbb{S}_i \cdots \cap \mathbb{S}_K = \bigcap_{i=1}^{K} \mathbb{S}_i \end{array} \right., \tag{19}$$

where ∪, ∩ and ∅ represent the symbols for union, intersection and empty set in the set theory, respectively, and

$$\left\{ \begin{array}{l} \mathbb{S}_i = \mathbb{S}_i^+ \cup \mathbb{S}_i^- \quad (i = 1, 2, \cdots, K) \\ \mathbb{S}^+ = \mathbb{S}_1^+ \cup \mathbb{S}_2^+ \cup \cdots \cup \mathbb{S}_i^+ \cdots \cup \mathbb{S}_K^+ \\ \mathbb{S}^- = \mathbb{S}_1^- \cup \mathbb{S}_2^- \cup \cdots \cup \mathbb{S}_i^- \cdots \cup \mathbb{S}_K^- \end{array} \right. \tag{20}$$

with

$$\left\{ \begin{array}{l} \left|\mathbb{S}_1^+\right| \approx \left|\mathbb{S}_2^+\right| \approx \cdots \approx \left|\mathbb{S}_i^+\right| \approx \cdots \approx \left|\mathbb{S}_K^+\right| \\ \left|\mathbb{S}_1^-\right| \approx \left|\mathbb{S}_2^-\right| \approx \cdots \approx \left|\mathbb{S}_i^-\right| \approx \cdots \approx \left|\mathbb{S}_K^-\right| \end{array} \right., \tag{21}$$

where $\mathbb{S}_i^+$ is the ith positive sub subset containing the circRNA-disease associations experimentally verified and $\left|\mathbb{S}_i^+\right|$ is the number of samples in $\mathbb{S}_i^+$; $\mathbb{S}_i^-$ is the ith negative sub subset containing the circRNA-disease associations without experimental validation and $\left|\mathbb{S}_i^-\right|$ is the number of samples in $\mathbb{S}_i^-$. Then, one of the subsets is taken as a test set in turn and the remaining K-1 subsets are used as training sets.

For disease-specific cross-validation [24], the positive subset $\mathbb{S}^+$ (cf. (1)) and the negative subset $\mathbb{S}^-$ (cf. (1)) can be formulated as

$$\left\{ \begin{array}{l} \mathbb{S}^+ = \cup_{i=1}^{n} \mathbb{S}_i^{\mathrm{d}+} \\ \mathbb{S}^- = \cup_{i=1}^{n} \mathbb{S}_i^{\mathrm{d}-} \end{array} \right., \tag{22}$$

where n is the number of diseases in $\mathbb{S}$ (cf. (1)); $\mathbb{S}_i^{\mathrm{d}+}$ is the specific positive subset that contains the experimentally verified associations related to the ith disease. $\mathbb{S}_i^{\mathrm{d}-}$ is the specific negative subset that contains the pairs without experimental validation related to the ith disease. For the ith disease, the training samples and test samples are constituted as follows:

$$\left\{ \begin{array}{l} \mathbb{S}^{\mathrm{test}} = a_j \cup \mathbb{S}_i^{\mathrm{d}-} \\ \mathbb{S}^{\mathrm{train}} = \complement_{\mathbb{S}^+} a_j \cup \complement_{\mathbb{S}^-} \mathbb{S}_i^{\mathrm{d}-} \end{array} \right. \quad \left(j \in \left\{1, 2, \cdots, \left|\mathbb{S}_i^{\mathrm{d}+}\right|\right\} \right), \tag{23}$$

where $\complement$ is the symbol for complementary operation; $\left|\mathbb{S}_i^{\mathrm{d}+}\right|$ is the number of samples in $\mathbb{S}_i^{\mathrm{d}+}$. a_j represents the jth association in $\mathbb{S}_i^{\mathrm{d}+}$, which is selected as positive test sample in turn for each run. This process is repeated $\left|\mathbb{S}_i^{\mathrm{d}+}\right|$ times until each association in $\mathbb{S}_i^{\mathrm{d}+}$ is tested.

The AUC (area under the receiver operating characteristics curve) [64–66] scores were used to evaluate the performance of various methods. As a comprehensive evaluation metric, AUC describes the sensitivity and specificity of computational method. A larger value of AUC indicates the better predictive performance of the predictor.

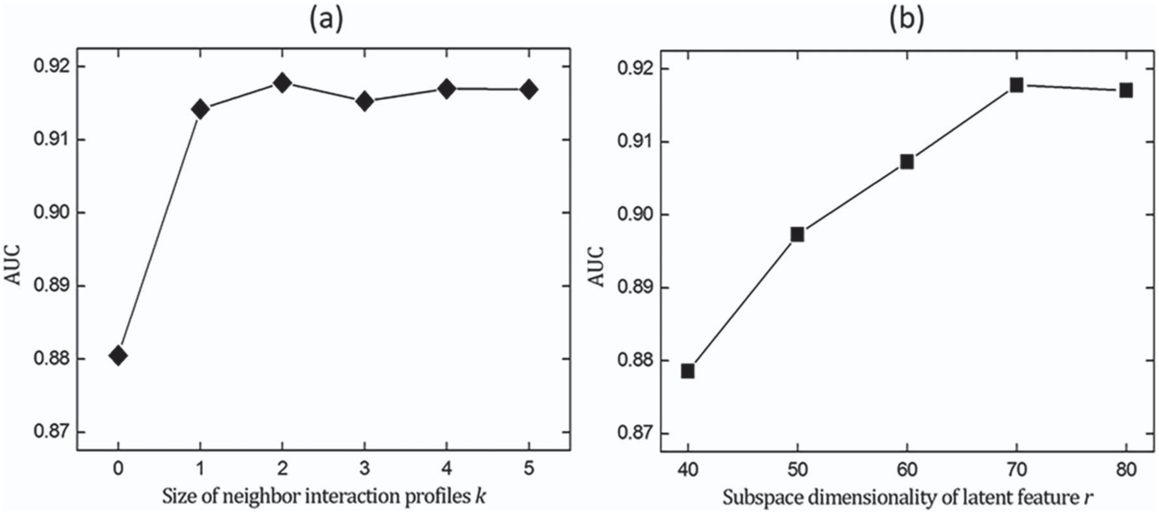

Figure 2. The influence of the parameters k and r on the performance of iCircDA-MF. (**a**) The AUC scores achieved by iCircDA-MF based on differernt k values, and fixed parameters $r = 70, \alpha = 2 \times 10^{-3}$, $\beta = 1 \times 10^{-3}$. (**b**) The AUC scores achieved by iCircDA-MF with different r values and fixed parameters $k = 2, \alpha = 2 \times 10^{-3}$, $\beta = 1 \times 10^{-3}$.

Results and discussion

Parameter optimization

It can be seen that the process of reformulating circRNA-disease association adjacency matrix contains a parameter k representing the size of neighbour interaction profiles. Three parameters: r, α, β in the matrix factorization (see (17)), where r represents the subspace dimensionality of latent feature matrix and α and β are two regularization parameters. In this study, parameter combinations were considered from the following ranges:

$$\begin{cases} 0 \le k \le 5, & \text{with step } \triangle k = 1 \\ 40 \le r \le 80, & \text{with step } \triangle r = 10 \\ \alpha = 2 \times 10^{-i}, & \text{with } 1 \le i \le 3 \\ \beta = 1 \times 10^{-i}, & \text{with } 1 \le i \le 3 \end{cases} . \tag{24}$$

Because the neighbour interaction profiles size k and the subspace dimensionality r are closely related to the number of known circRNA-disease associations, the impacts of k and r on the predictive performance of iCircDA-MF were analysed, and shown in Figure 2a and b, respectively, from which we can see the following: (i) The lowest AUC score is obtained when $k = 0$ representing no neighbour interaction profile information is used to reformulate the sparse circRNA-disease association adjacency matrix $\mathbf{A}$ (cf. (3)). (ii) Compared the performance of iCircDA-MF when $k = 0$, its performance improves rapidly when $k = 1$, and then turns stable, indicating that reformulation of sparse circRNA-disease association adjacency matrix based on the neighbour interaction profiles can improve the performance of iCircDA-MF. (iii) The predictive performance of iCircDA-MF improves as the value of r increases. iCircDA-MF performs best when $r = 70$, and turns stable at $r = 80$.

The final optimum parameter combinations were optimized based on the AUC scores by 5-fold cross-validation, and their optimized values are $k = 2, r = 70, \alpha = 2 \times 10^{-3}$, $\beta = 1 \times 10^{-3}$.

Incorporating more association information via neighbour interaction profiles

The known circRNA-disease associations in the original circRNA-disease association adjacency matrix $\mathbf{A}$ (cf. (3)) are limited, and most unknown associations are incorrectly labelled as uncorrelated circRNA-disease associations. Based on the biological assumption that similar diseases show similar interaction patterns with circRNAs, and vice versa [24–26], the association information in the original circRNA-disease association adjacency matrix $\mathbf{A}$ (cf. (3)) can be supplemented by the interaction profiles of similar diseases and similar circRNAs. This reformulation approach has been successfully used in drug-target interaction prediction [46]. To illustrate the effect of reformulation of the original sparse association adjacency matrix $\mathbf{A}$ (cf. (3)), a graphical illustration for generating reformulated association adjacency matrix $\mathbf{A}'$ (cf. (14)) is shown in Figure 3, from which we can see the following: (i) The reformulated circRNA-disease association adjacency matrix from horizontal direction $\mathbf{A_c}$ (cf. (13)) based on the interaction profiles of similar circRNAs and that from vertical direction $\mathbf{A_d}$ (cf. (12)) based on the interaction profiles of similar diseases can introduce additional association information. (ii) Compared with the original sparse circRNA-disease association matrix $\mathbf{A}$ (cf. (3)), the final reformulated circRNA-disease association matrix $\mathbf{A}'$ (cf. (14)) is able to correct some false negative associations by assigning their corresponding association scores.

Comparison with highly related methods

To evaluate the performance of iCircDA-MF, 5-fold cross-validation (cf. (19)) was implemented. The performance of iCircDA-MF was compared with five state-of-the-art predictors, including PWCDA [24], RWRHCD [24], HGICD [24], DWNN-RLS [25] and KATZHCDA [26]. Table 1 lists the performance of various methods, from which we can see that the iCircDA-MF obviously outperforms other methods. PWCDA, RWRHCD, HGICD and KATZHCDA are four network-based methods, which construct a bipartite graph by connecting the circRNA network and

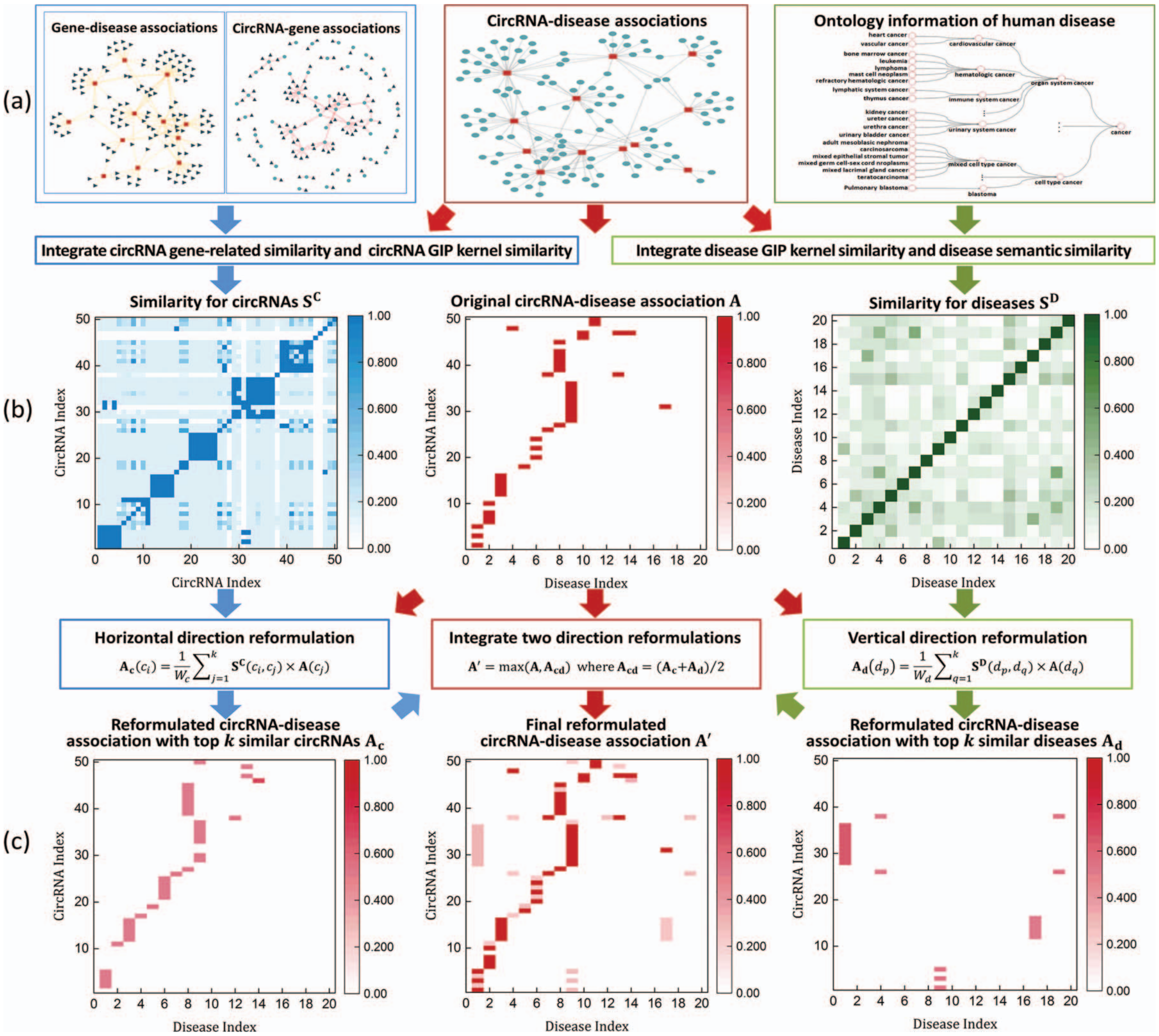
(a)
Gene-disease associations
CircRNA-gene associations
CircRNA-disease associations
Ontology information of human disease
heart cancer
vascular cancer
bone marrow cancer
leukemia
lymphoma
mast cell neoplasm
refractory hematologic cancer
lymphatic system cancer
thymus cancer
kidney cancer
ureter cancer
urethra cancer
urinary bladder cancer
adult mesoblastic nephroma
carcinosarcoma
mixed epithelial stromal tumor
mixed germ cell-sex cord nroplasms
mixed lacrimal gland cancer
teratocarcinoma
Pulmonary blastoma
cardiovascular cancer
hematologic cancer
organ system cancer
immune system cancer
urinary system cancer
cancer
mixed cell type cancer
cell type cancer
blastoma
Integrate circRNA gene-related similarity and circRNA GIP kernel similarity
Integrate disease GIP kernel similarity and disease semantic similarity
(b)
Similarity for circRNAs S^C
Original circRNA-disease association A
Similarity for diseases S^D
CircRNA Index
Disease Index
0.00 0.200 0.400 0.600 0.800 1.00
Horizontal direction reformulation
$A_c(c_i) = \frac{1}{W_c}\sum_{j=1}^{k} S^C(c_i, c_j) \times A(c_j)$
Integrate two direction reformulations
$A' = \max(A, A_{cd})$ where $A_{cd} = (A_c + A_d)/2$
Vertical direction reformulation
$A_d(d_p) = \frac{1}{W_d}\sum_{q=1}^{k} S^D(d_p, d_q) \times A(d_q)$
(c)
Reformulated circRNA-disease association with top k similar circRNAs A_c
Final reformulated circRNA-disease association A′
Reformulated circRNA-disease association with top k similar diseases A_d

Figure 3. A graphical illustration to show generating reformulated association adjacency matrix. We take 50 circRNAs and 20 diseases as an example to illustrate the process of reformulating the original association adjacency matrix **A** (cf. (3)). (**a**) Original biological information. (**b**) The heat maps for circRNA similarity matrix $\mathbf{S}^{\mathbf{C}}$ (cf. (11)), original circRNA-disease association adjacency matrix **A** (cf. (3)) and disease similarity matrix $\mathbf{S}^{\mathbf{D}}$ (cf. (7)). (**c**) The heat maps for reformulated association adjacency matrices when $k = 2$. Left: the reformulated circRNA-disease association matrix from horizontal direction $\mathbf{A_c}$ (cf. (13)) based on the interaction profiles of similar circRNAs. Right: the reformulated circRNA-disease association matrix from vertical direction $\mathbf{A_d}$ (cf. (12)) based on the interaction profiles of similar diseases. Middle: the final reformulated circRNA-disease association matrix $\mathbf{A}'$ (cf. (14)).

disease network using the known circRNA-disease. RWRHCD and HGIMDA infer potential circRNA-disease associations via different extended random walk iterative processing. With iteration increasing, more paths of long length will be introduced in the sparse circRNA-disease network, which may produce much noisy and meaningless information and lead to poor predictive performance. PWCDA and KATZHCDA perform better because they calculate an association score for each circRNA-disease pair by summarizing all paths with a fixed small path length considering the limited experimentally validated circRNA-disease associations. To overcome the problem of sparse known circRNA-disease associations, iCircDA-MF and DWNN-RLS incorporate additional association information by fusing different information sources. DWNN-RLS infers circRNA-disease associations based on regularized least squares of Kronecker product kernel. Compared with DWNN-RLS, iCircDA-MF achieves higher predictive performance, because it can effectively reduce noisy information by considering the geometrical structure and biological meaning of the decomposed low-dimension feature space of circRNAs and diseases in the process of matrix factorization.

Performance for predicting disease specific associations

In order to further illustrate the predictive results of iCircDA-MF for identifying related circRNAs with specific diseases, the diseases in benchmark dataset $\mathbb{S}$ (cf. (1)) were sorted in descending order by their number of experimentally verified associated circRNAs, and the top 10 diseases with most associated circRNAs were selected, and predicted by iCircDA-MF via the disease-specific cross-validation (cf. (23)). The results were shown in Table 2, from which we can see the following: (i) The proposed iCircDA-MF predictor is able to accurately identify the associations between circRNAs and the 10 diseases with AUC scores higher than 0.93. (ii) Among the 10 diseases, 2 diseases (‘Colorectal cancer’ and ‘Esophageal squamous cell carcinoma’) were also predicted by three related methods, including PWCDA

Table 1. AUC scores obtained by various methods via 5-fold cross-validation on the same benchmark dataset $\mathbb{S}$ (cf. (1))

Methods	iCircDA-MF[a]	PWCDA[b]	DWNN-RLS[c]	KATZHCDA[d]	GHICD[b]	RWRHCD[b]
AUC scores	0.9178	0.8900	0.8854	0.7936	0.7290	0.6660

[a] Results obtained by the proposed predictor iCircDA-MF with parameter $k = 2, r = 70, \alpha = 2 \times 10^{-3}, \beta = 1 \times 10^{-3}$.
[b] Results obtained by the predictor reported in [24].
[c] Results obtained by the predictor reported in [25].
[d] Results obtained by the predictor reported in [26].

Table 2. Number of known associations and the corresponding AUC scores achieved by iCircDA-MF via disease-specific cross-validation for the 10 diseases with most associated circRNAs on benchmark dataset $\mathbb{S}$ (cf. (1))

Disease	Number of experimentally validated associated circRNAs	AUC scores[a]
Gastric cancer	55	0.9693
Breast cancer	48	0.9545
Esophageal squamous cell carcinoma	34	0.9520
Glioma	31	0.9578
Colorectal cancer	31	0.9480
Hepatocellular carcinoma	19	0.9503
Osteosarcoma	19	0.9454
Bladder cancer	17	0.9472
Hepatoblastoma	17	0.9537
Papillary thyroid carcinoma	16	0.9359

[a] The results achieved by the proposed predictor iCircDA-MF with parameter $k = 2, r = 70, \alpha = 2 \times 10^{-3}, \beta = 1 \times 10^{-3}$.

[24], RWRHCD [24] and GHICD [24]. Compared with the results reported in [24], iCircDA-MF outperformed these three predictors by 12.3–38.3% and 3.2–41.8% in terms of AUC scores.

Case studies

To illustrate the capability of iCircDA-MF for predicting novel circRNA-disease associations based on the known associations, the performance of iCircDA-MF was further evaluated for predicting novel associated circRNAs for three important diseases, including lung cancer, liver cancer and pancreatic cancer. For each cancer, all the known associations in $\mathbb{S}^{+}$ (cf. (1)) were used to train iCircDA-MF, and the remaining unknown cancer-circRNA pairs were considered as candidates for testing. Then, all the candidates were sorted by the predicted association scores. circ2Disease [22], circRNADisease [23] and literatures in PubMed were used to confirm the predicted associations.

Table 3 and Figure 4 show the association networks of the top 10 circRNAs with highest association scores for each cancer predicted by iCircDA-MF, from which we can see the following: (i) Among the top ranked circRNAs, 'hsa_circ_0000284' and 'has_circ_0001946' are observed to be related to multiple cancers, indicating that the two circRNAs may play crucial roles in the development of cancers. (ii) Most of the predicted circRNAs are really associated with the corresponding cancers (blue circles and lines in Figure 4) as their associations are supported by circ2Disease [22], circRNADisease [23] or experimental literatures in PubMed. For example, 'hsa_circ_0000284' promotes human lung cancer cell proliferation via circHIPK3/miR-379 pathway [67] and regulates pancreatic carcinoma proliferation through the IL6-STAT3 pathway [68]. 'hsa_circ_0023404' is overexpressed in the lung cancer cell [69]. 'hsa_circRNA_103096', 'hsa_circRNA_102032', 'hsa_circRNA_400031', 'hsa_circRNA_100571' and 'hsa_circRNA_102347' are top 5 ranked predicted circRNAs associated with liver cancer, which are significantly differentially expressed between patients of liver cancer following liver transplantation and healthy individuals [70]. 'hsa_circ_0041150' is proved to be dysregulated in pancreatic ductal adenocarcinoma patients, and is proposed to serve as potential roles in pancreatic cancer [71]. The case studies further illustrate that iCircDA-MF can predict reliable novel circRNA-disease associations, and provide valuable circRNA candidates for biological experiments to study the mechanism of diseases.

The performance of iCircDA-MF is underestimated

The benchmark dataset $\mathbb{S}$ (cf. (1)) treats the circRNA-disease associations with experimental validation as the positive samples, and the circRNA-disease associations without experimental validation as the negative samples. However, the experimentally verified circRNA-disease associations are very limited. As shown in the above case studies, many circRNA-disease associations without experimental validation in fact are really circRNA-disease associations. In other words, the predicted circRNA-disease associations by iCircDA-MF labelled as false positives would be true positives, and therefore, its performance would be underestimated. In order to answer this question, we did the following experiment. The label of each association without experimental validation a in $\mathbb{S}^{-}$ (cf. (1)) was inferred by

$$\mathrm{L}(a) = \begin{cases} 1 & \text{if } \mathbf{A}'(a) \geq \lambda \\ 0 & \text{otherwise} \end{cases}, \tag{25}$$

where $\mathrm{L}(a) = 1$ means inferring a as positive, and infer a as negative when $\mathrm{L}(a) = 0$. $\mathbf{A}'(a)$ represents the predicted association score of a in the reformulated circRNA-disease association adjacency matrix $\mathbf{A}'$ (cf. (14)). λ is a threshold of inferring association label.

Figure 5a shows the number of associations without experimental validation inferred as positives based on different values

Table 3. The top 10 novel associations identified by iCircDA-MF for lung cancer, liver cancer and pancreatic cancer

Cancer	Top 10 ranked associations for three cancers					
	Rank	circRNAs	Evidences[a]	Rank	circRNAs	Evidences[a]
Lung cancer	1	hsa_circ_0000284	circRNADisease circ2Disease	6	hsa_circ_0007385	29372377
	2	hsa_circ_0023404	circRNADisease circ2Disease	7	hsa_circ_0014130	29440731
	3	circ-Foxo3	29620202	8	hsa_circ_0008887	unconfirmed
	4	hsa_circ_0043256	circRNADisease	9	hsa_circ_0004214	circRNADisease
	5	hsa_circ_0016760	29440731	10	hsa_circ_0061893	unconfirmed
Liver cancer	1	hsa_circRNA_103096	29609527	6	hsa_circ_0041731	29414822
	2	hsa_circRNA_102032	29609527	7	hsa_circ_0072359	29414822
	3	hsa_circRNA_400031	29609527	8	hsa_circ_0000284	circRNADisease circ2Disease
	4	hsa_circRNA_100571	29609527	9	hsa_circ_0001946	circ2Disease
	5	hsa_circRNA_102347	29609527	10	circDLGAP4	unconfirmed
Pancreatic cancer	1	hsa_circ_0041150	27997903	6	hsa_circ_0005785	27997903
	2	hsa_circ_0000257	27997903	7	hsa_circ_0008719	27997903
	3	hsa_circ_0000284	29255366	8	hsa_circ_0001946	27997903
	4	hsa_circ_0000677	unconfirmed	9	hsa_circ_0002078	unconfirmed
	5	hsa_circ_0005397	27997903	10	hsa_circ_0006913	27997903

[a] The predicted associations are confirmed by circ2Disease [22], circRNADisease [23] or the literatures in PubMed. The PMID of these literatures are provided.

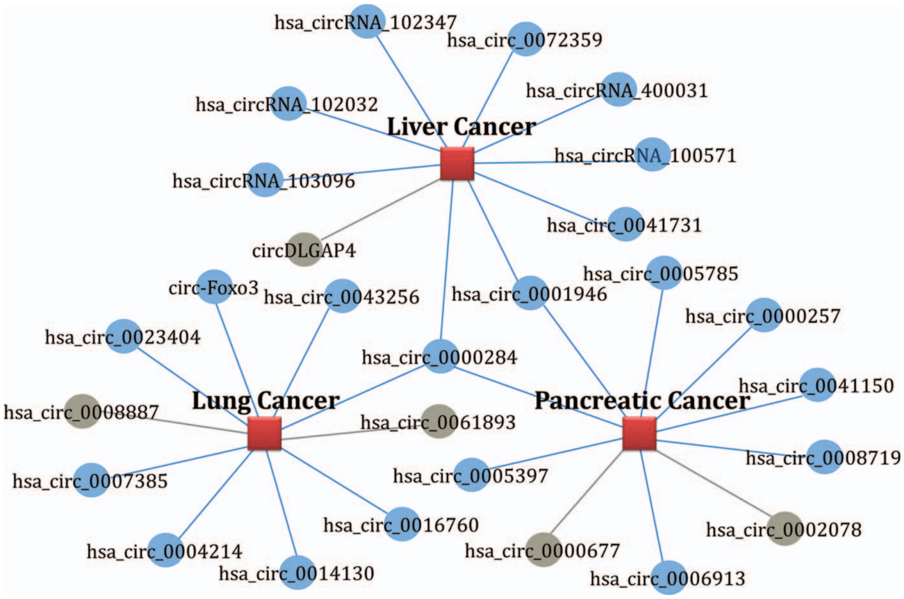

Figure 4. The association networks of the top 10 predicted associations for lung cancer, liver cancer and pancreatic cancer. The three cancers are shown in red squares. The predicted associations and cancer-related circRNAs with experimental validation are shown in blue lines and circles, while the predicted associations and circRNA candidates without experimental validation are shown in grey lines and circles.

of λ, and Figure 5b lists their corresponding AUC scores. From Figure 5, we can draw the following conclusions: (i) The performance of iCircDA-MF can be improved by correcting some negatives as positives based on (25). (ii) Higher λ values result in less added positives, and therefore limited performance improvement is observed. The receiver operating characteristic (ROC) curves of iCircDA-MF based on original labels and corrected labels are shown in Figure 6, from which we can see that the performance of iCircDA-MF is really underestimated. In other words, with the increment of the circRNA-disease associations with experimental validation, the performance of iCircDA-MF will increase as well. This problem also exists in other fields, such as protein disordered region prediction [72], etc. A possible solution is to construct the computational models in a semi-supervised manner trained with the high quality positive samples constructed by the circRNA-disease associations with experimental validation and unlabelled data.

Conclusion

In recent years, more and more experiments have verified that circRNAs are closely related to the development of various diseases. Therefore, identifying associations between circRNAs and diseases can help understand the complex

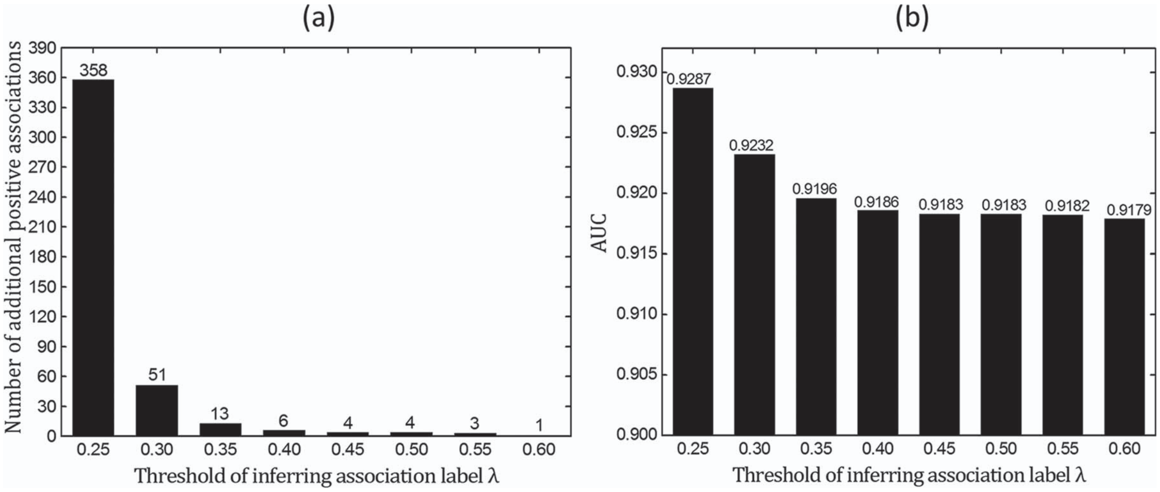

Figure 5. The influence of threshold λ on the predictive performance of iCircDA-MF. (**a**) The relationship between the threshold λ and the number of the corrected positives. (**b**) The influence of threshold λ on the AUC scores obtained by the iCircDA-MF with parameters $k = 2, r = 70, \alpha = 2 \times 10^{-3}, \beta = 1 \times 10^{-3}$ via 5-fold cross-validation on benchmark dataset $\mathbb{S}$ (cf. (1)).

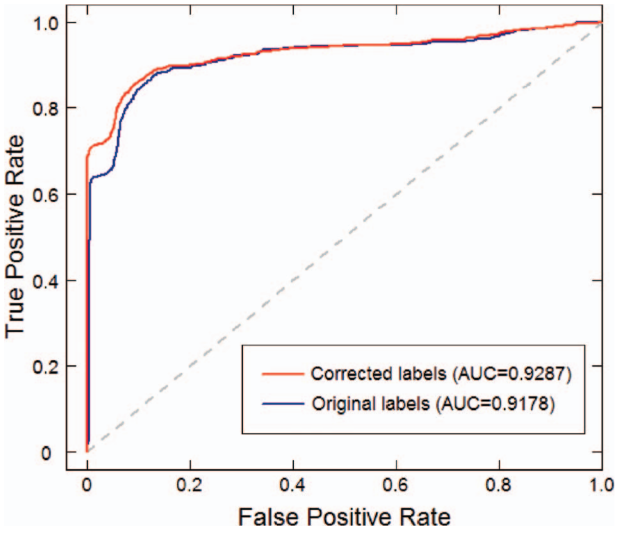

Figure 6. The ROC curves of iCircDA-MF based on original labels and corrected labels. The parameters of iCircDA-MF were set as $k = 2, r = 70, \alpha = 2 \times 10^{-3}$, $\beta = 1 \times 10^{-3}$, and the threshold λ was set as 0.25 with 358 corrected positives.

disease mechanisms and facilitate disease-targeted therapy. It is necessary to develop computational methods to identify circRNA-disease associations.

In this study, we have proposed a novel computational method called iCircDA-MF to identify circRNA-diseases associations. Experimental results showed that it outperformed other state-of-the-art predictors, and can identify new circRNA-diseases associations effectively. Furthermore, we also showed that the performance of iCircDA-MF was underestimated with the current known associations. We anticipate that with the growth of the experimentally validated associations, the predictive performance of iCircDA-MF will be improved and objectively estimated as well.

Three main factors attribute to the performance of iCircDA-MF: (i) iCircDA-MF incorporates multiple biological information to measure circRNA similarity and disease similarity. (ii) A preprocessing step of reformulating the sparse circRNA-disease association adjacency matrix based on their neighbour interaction profiles is employed so as to correct the false negatives in the original association adjacency matrix, and makes iCircDA-MF applicable for predicting the associations of new diseases and circRNAs. (iii) iCircDA-MF computes the association scores based on matrix factorization, which is able to detect meaningful latent features from sparse matrix.

The proposed iCircDA-MF achieves state-of-the-art performance in identifying circRNA-disease associations, but it would be further improved by addressing the following problems. The similarity measurement for circRNAs are only based on the available networks with undetected interactions, leading to noise information and bias to the circRNAs with more known interactions. Because circRNAs sharing similar sequences tend to be associated with similar disease, the sequence-based similarity measure will be useful for calculating circRNA similarity. The disease similarity can be more accurately measured by other disease similarity methods, such as disease functional similarity [73], disease module similarity [74], etc. Finally, more advanced matrix factorization techniques can be applied to remove the noise and discover potential circRNA-disease associations, such as DNRLMF-MDA [38], LWSG_NMF [75], etc.

Key Points

- Due to the circular RNAs (circRNAs) having a close relationship with the progression of various human diseases, it is critical for developing effective computational predictors for circRNA-disease association prediction.
- A predictor called iCircDA-MF was proposed, which incorporated more association information based on the circRNA similarity and disease similarity, and employed matrix factorization to extract the latent features from a new circRNA-disease association network.
- Performance on a widely used benchmark dataset showed that iCircDA-MF outperforms other state-of-the-art predictors. Furthermore, some case studies showed that iCircDA-MF is also able to discover new circRNA-disease associations accurately.

Funding

National Natural Science Foundation of China (No. 61672184, 61732012, 61822306); Fok Ying-Tung Education Foundation for Young Teachers in the Higher Education Institutions of China (161063); Guangdong Special Support Program of Technology Young talents (2016TQ03X618); Shenzhen Overseas High Level Talents Innovation Foundation (Grant No. KQJSCX20170327161949608); Scientific Research Foundation in Shenzhen (JCYJ20180306172207178).

References

1. Meng S, Zhou H, Feng Z, *et al*. CircRNA: functions and properties of a novel potential biomarker for cancer. *Mol Cancer* 2017;**16**:94.
2. Wang CY, Wei LY, Guo MZ, *et al*. Computational approaches in detecting non-coding RNA. *Curr Genomics* 2013;**14**:371–7.
3. Wang QC, Wei LY, Guan XJ, *et al*. Briefing in family characteristics of microRNAs and their applications in cancer research. *Biochim Biophys Acta Proteins and Proteom* 2014;**1844**:191–7.
4. Bartel DP. MicroRNAs: target recognition and regulatory functions. *Cell* 2009;**136**:215–33.
5. Hansen TB, Jensen TI, Clausen BH, *et al*. Natural RNA circles function as efficient microRNA sponges. *Nature* 2013;**495**:384–8.
6. Li ZY, Huang C, Bao C, *et al*. Exon-intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol* 2015;**22**:256–64.
7. Wang Y, Mo Y, Gong Z, *et al*. Circular RNAs in human cancer. *Mol Cancer* 2017;**16**:25.
8. Esteller M. Non-coding RNAs in human disease. *Nat Rev Genet* 2011;**12**:861–74.
9. Peng L, Yuan XQ, Li GC. The emerging landscape of circular RNA ciRS-7 in cancer (review). *Oncol Rep* 2015;**33**:2669–74.
10. Yu L, Gong XJ, Sun L, *et al*. The circular RNA Cdr1as act as an oncogene in hepatocellular carcinoma through targeting miR-7 expression. *Plos One* 2016;**11**:e0158347.
11. Weng WH, Wei Q, Toden S, *et al*. Circular RNA ciRS-7—a promising prognostic biomarker and a potential therapeutic target in colorectal cancer. *Clin Cancer Res* 2017;**23**: 3918–28.
12. Tang W, Ji M, He G, *et al*. Silencing CDR1as inhibits colorectal cancer progression through regulating microRNA-7. *Onco Targets Ther* 2017;**10**:2045–56.
13. Floris G, Zhang LB, Follesa P, *et al*. Regulatory role of circular RNAs and neurological disorders. *Mol Neurobiol* 2017;**54**:5156–65.
14. Chen G, Wang Z, Wang D, *et al*. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res* 2013;**41**:D983–6.
15. Li Y, Qiu CX, Tu J, *et al*. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* 2014;**42**:D1070–4.
16. Bao Z, Yang Z, Huang Z, *et al*. LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res* 2019;**47**:D1034–7.
17. Huang Z, Shi J, Gao Y, *et al*. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. *Nucleic Acids Res* 2019;**47**:D1013–7.
18. Zou Q, Li J, Song L, *et al*. Similarity computation strategies in the microRNA-disease network: a survey. *Brief Funct Genomics* 2016;**15**:55–64.
19. Chen X, Xie D, Zhao Q, *et al*. MicroRNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2019;**20**:515–39.
20. Chen X, Yan CC, Zhang X, *et al*. Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2017;**18**:558–76.
21. Fan CY, Lei XJ, Fang ZQ, *et al*. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. *Database* 2018;**2018**:bay044.
22. Yao DX, Zhang L, Zheng MY, *et al*. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Sci Rep* 2018;**8**:11018.
23. Zhao Z, Wang KY, Wu F, *et al*. circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. *Cell Death Dis* 2018;**9**:475.
24. Lei X, Fang Z, Chen L, *et al*. PWCDA: path weighted method for predicting circRNA-disease associations. *Int J Mol Sci* 2018;**19**:3410.
25. Yan C, Wang J, Wu FX. DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. *BMC Bioinformatics* 2018;**19**:520.
26. Fan C, Lei X, Wu F. Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks. *Int J Biol Sci* 2018;**14**:1950–9.
27. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics* 2011;**27**:3036–43.
28. You ZH, Huang ZA, Zhu Z, *et al*. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput Biol* 2017;**13**:e1005455.
29. Luo J, Xiao Q, Liang C, *et al*. Predicting MicroRNA-disease associations using Kronecker regularized least squares based on heterogeneous omics data. *IEEE Access* 2017;**5**:2503–13.
30. Chen X, Yan GY. Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* 2013;**29**:2617–24.
31. Wang JZ, Du ZD, Payattakool R, *et al*. A new method to measure the semantic similarity of GO terms. *Bioinformatics* 2007;**23**:1274–81.
32. Lan W, Li M, Zhao K, *et al*. LDAP: a web server for lncRNA-disease association prediction. *Bioinformatics* 2017;**33**: 458–60.
33. Wang D, Wang JA, Lu M, *et al*. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 2010;**26**:1644–50.
34. Chen X, Wu QF, Yan GY. RKNNMDA: ranking-based KNN for MiRNA-disease association prediction. *RNA Biol* 2017;**14**:952–62.
35. Kibbe WA, Arze C, Felix V, *et al*. Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* 2015;**43**:D1071–8.
36. Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Sci Rep* 2015;**5**:16840.
37. Chen X, Yin J, Qu J, *et al*. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLoS Comput Biol* 2018;**14**:e1006418.
38. Yan C, Wang J, Ni P, *et al*. DNRLMF-MDA:predicting microRNA-disease associations based on similarities of microRNAs and diseases. *IEEE/ACM Trans Comput Biol Bioinform* **2017**:1–12.

39. Chen X, Gong Y, Zhang DH, *et al*. DRMDA: deep representations-based miRNA-disease association prediction. *J Cell Mol Med* 2017;**22**:472–85.
40. Zhao H, Kuang L, Feng X, *et al*. A novel approach based on a weighted interactive network to predict associations of MiRNAs and diseases. *Int J Mol Sci* 2018;**20**:110.
41. Zhong YX, Du YJ, Yang X, *et al*. Circular RNAs function as ceRNAs to regulate and control human cancer progression. *Mol Cancer* 2018;**17**:79.
42. Huang YA, Chan KCC, You ZH. Constructing prediction models from expression profiles for large scale lncRNA-miRNA interaction profiling. *Bioinformatics* 2018;**34**:812–9.
43. Liu Y, Zeng X, He Z, *et al*. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:905–15.
44. Hwang S, Kim CY, Yang S, *et al*. HumanNet v2: human gene networks for disease research. *Nucleic Acids Res* 2018;**47**:D573–80.
45. Lee I, Blom UM, Wang PI, *et al*. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome Res* 2011;**21**:1109–21.
46. Ezzat A, Zhao PL, Wu M, *et al*. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans Comput Biol Bioinform* 2017;**14**:646–56.
47. Lian D, Zhao C, Xie X, *et al*. GeoMF: joint geographical modeling and matrix factorization for point-of-interest recommendation. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY, USA: ACM, 2014, 831–40.
48. He X, Zhang H, Kan M-Y, *et al*. Fast matrix factorization for online recommendation with implicit feedback. In: *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. Pisa, Italy: ACM, 2016, 549–58.
49. Luo X, Zhou MC, Xia YN, *et al*. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Trans Industr Inform* 2014;**10**:1273–84.
50. Wang L, Li XZ, Zhang LX, *et al*. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. *BMC Cancer* 2017;**17**:513.
51. Liu Y, Wu M, Miao CY, *et al*. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. *PLoS Comput Biol* 2016;**12**:e1004760.
52. Li JQ, Rong ZH, Chen X, *et al*. MCMDA: matrix completion for MiRNA-disease association prediction. *Oncotarget* 2017;**8**:21187–99.
53. Zhong Y, Xuan P, Wang X, *et al*. A non-negative matrix factorization based method for predicting disease-associated miRNAs in miRNA-disease bilayer network. *Bioinformatics* 2018;**34**:267–77.
54. Xiao Q, Luo J, Liang C, *et al*. A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations. *Bioinformatics* 2018;**34**:239–48.
55. Lu C, Yang M, Luo F, *et al*. Prediction of lncRNA-disease associations based on inductive matrix completion. *Bioinformatics* 2018;**34**:3357–64.
56. Fu G, Wang J, Domeniconi C, *et al*. Matrix factorization-based data fusion for the prediction of lncRNA-disease associations. *Bioinformatics* 2018;**34**:1529–37.
57. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* 1999;**401**:788–91.
58. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: *Advances in Neural Information Processing Systems Proceedings*, Denver, CO, USA: NIPS, 2000, 556–62.
59. Golub GH, Van Loan CF. *Matrix Computations*, 3rd edn. Baltimore, USA: Johns Hopkins University Press, 1996, xxvii+694 pp.
60. Liu XM, Zhai DM, Zhao DB, *et al*. Progressive image denoising through hybrid graph Laplacian regularization: a unified framework. *IEEE Trans Image Process* 2014;**23**:1491–503.
61. Cai D, He XF, Han JW, *et al*. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans Pattern Anal Mach Intell* 2011;**33**:1548–60.
62. Facchinei F, Kanzow C, Sagratella S. Solving quasi-variational inequalities via their KKT conditions. *Math Program* 2014;**144**:369–412.
63. Liu B, Yang F, Huang DS, *et al*. iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics* 2018;**34**: 33–40.
64. Hanley JA, Mcneil BJ. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology* 1982;**143**:29–36.
65. Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;**27**:861–74.
66. Liu B, Jiang S, Zou Q. HITS-PR-HHblits: protein remote homology detection by combining PageRank and Hyperlink-Induced Topic Search. *Brief Bioinform* 2018. DOI: 10.1093/bib/bby104.
67. Tian F, Wang Y, Xiao Z, *et al*. Circular RNA circHIPK3 promotes NCI-H1299 and NCI-H2170 cell proliferation through miR-379 and its target IGF1. *Zhongguo Fei Ai Za Zhi* 2017;**20**:459–67.
68. Chen G, Shi Y, Zhang Y, *et al*. CircRNA_100782 regulates pancreatic carcinoma proliferation through the IL6-STAT3 pathway. *Onco Targets Ther* 2017;**10**:5783–94.
69. Yao JT, Zhao SH, Liu QP, *et al*. Over-expression of CircRNA_100876 in non-small cell lung cancer and its prognostic value. *Pathol Res Pract* 2017;**213**:453–6.
70. Sui W, Gan Q, Liu F, *et al*. The differentially expressed circular ribonucleic acids of primary hepatic carcinoma following liver transplantation as new diagnostic biomarkers for primary hepatic carcinoma. *Tumour Biol* 2018;**40**:1–8.
71. Li H, Hao X, Wang H, *et al*. Circular RNA expression profile of pancreatic ductal adenocarcinoma revealed by microarray. *Cell Physiol Biochem* 2016;**40**:1334–44.
72. Liu YM, Wang XL, Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief Bioinform* 2019;**20**:330–46.
73. Cheng L, Li J, Ju P, *et al*. SemFunSim: a new method for measuring disease similarity by integrating semantic and gene functional association. *PLoS One* 2014; **9**:e99415.
74. Ni P, Wang J, Zhong P, *et al*. Constructing disease similarity networks based on disease module theory. *IEEE/ACM Trans Comput Biol Bioinform* 2018. DOI: 10.1109/TCBB.2018.2817624.
75. Feng YF, Xiao J, Zhou K, *et al*. A locally weighted sparse graph regularized Non-Negative Matrix Factorization method. *Neurocomputing* 2015;**169**:68–76.