International Journal of Innovative
Computing, Information and Control
Volume **20**, Number **1**, February **2024**

ICIC International ©2024 ISSN 1349-4198
pp. **181–196**

PREDICTION OF CIRCRNA AND DISEASE ASSOCIATION BASED ON TRIPARTITE GRAPH AND SINGULAR VALUE DECOMPOSITION

Bo Wang*, Tingbin Liu, Jingyou Li, Xiaoxin Du and Guangda Zhang

School of Computer and Control Engineering
Qiqihar University
No. 42, Wenhua Street, Qiqihar 161003, P. R. China
{ 2021912336; lijingyou; xiaoxindu; zhangguangda }@qqhru.edu.cn
*Corresponding author: bowangdr@qqhru.edu.cn

Received April 2023; revised September 2023

Abstract. *Studies have proved that circRNA is an important regulatory factor of various pathological processes and can be used as an important biomarker for disease diagnosis, treatment and prognosis. Therefore, determining the circRNA-disease association in humans can understand the functional characteristics of circRNA and the pathogenesis of disease. In this paper, a new model for predicting circRNA-disease association, TPGSVDCDA, is proposed. This model introduces miRNA-disease association information, circRNA sequence information and disease semantic information into the circRNA-disease prediction model. And the nonlinear relationship of circRNA-disease was quantified by using a tripartite graph. By using Jaccard index and singular value decomposition fusion similarity calculation, the data sparsity problem is improved. The experimental results show that TPGSVDCDA achieves quite good performance compared with other methods in LOOCV evaluation index. The robustness of the model was verified in evaluation indicators such as AUC and AUPR, and the accuracy of the circRNA-disease association was further confirmed by the case study.*

Keywords: circRNA and disease association, Similarity improvement, Tripartite graph, Probability propagation mechanism, LOOCV validation

1. **Introduction.** circRNA is a group of endogenous non-coding RNA's with a covalent closed-loop structure, discovered in the 1970s and thought at the time to be a by product [1]. With the rapid development of high-throughput sequencing methods and bioinformatics, many studies have proved that circRNA interacts with other molecules, thus participating in the regulation [2]. For example, circRNA interacts with other types of RNA to regulate the expression of corresponding target genes, including affecting arteriosclerosis, participating in the regulation of mRNA expression and regulating alternative splicing, and binding with RNA-binding proteins to form RNA-protein complexes affecting corresponding biological functions [3]. More importantly, research evidence suggests that circRNA is related to the occurrence and development of diseases, and may be used as a therapeutic target or biomarker of diseases in the course of disease treatment. Zhou and Yu, for example, found that circRNA_010567 inhibits miR-141 by targeting TGF-β1, thereby promoting myocardial fibrosis [4]. Liang et al. found that circ-ABCB10 can promote the proliferation and progression [5]. Many scholars believe that many circRNAs can be used as tumor markers and therapeutic targets [6].

However, due to the long time and high cost of biological experiments, traditional methods have not been able to verify the correlation between circRNA and disease on a large

DOI: 10.24507/ijicic.20.01.181

scale. To address this problem, researchers began working on developing computational prediction methods based on existing data sets, quantifying the probability of association between circRNA and disease, and using the most promising circRNA and disease associations for further biological experimental validation. In this case, the time and cost of biological experiments can be effectively reduced [7]. In recent years, a lot of research has been done on circRNAs, and a series of databases or tools have also been developed, such as circBase [8], CircR2Disease [9], circR2Cancer [10], circRNA disease [11], and Circ2Traits [12]. The establishment of circRNA-disease association dataset can provide data support for the prediction of potential disease-associated circRNAs by computational models. Many computational models have been used to mine potential circRNA-disease association pairs under the assumption that similar circRNAs may have similar associations with the same diseases. These models can be broadly classified into three categories: based on the propagation of information across a network, based on machine learning, and based on deep learning. For example, Fan et al. proposed a computational model of KATZ metrics called KATZHCDA that uses heterogeneous networks for circRNA-disease association prediction [13]. Lei et al. proposed an algorithm named PWCDA that first constructed a heterogeneous network with three similarities, and then calculated the association score between each circRNA pair and disease according to the connection path between each circRNA pair and disease in the heterogeneous network [14]. Li et al. proposed a computational method based on network consistency projection (NCPCDA) to identify the association between circRNA and disease. Firstly, multi-source similarity data is used to construct synthetic circRNA similarity and disease similarity. Then, the circRNA space and disease space were projected onto the interaction network of circRNA and disease respectively, and finally, the predicted circRNA-disease association score matrix was obtained by combining the projection scores of the above two spaces [15]. Lei et al. proposed a computational method based on collaborative filtering recommendation system (ICFCDA), which can deal with the "cold start" problem [16]. Ding et al. proposed a computational model based on random walk and logistic regression (RWLR) to predict the association of circRNA with disease [17]. Xiao et al. proposed a graph-based multi-label learning (GMCDA) method to maintain the local geometric structure of the data by making full use of the different features of the circRNA space and the disease space, and to add graph regularization and mixed gauge constraint terms to the model to help predict [18]. Fan et al. proposed a new approach called MSFCNN, which involves a two-layer convolutional neural network on a feature matrix that incorporates multiple similarity nuclei and interaction features among circRNA, miRNA, and disease [19]. Deepthi and Jereesh proposed a set method called AE-RF based on deep autoencoders and random forest classifiers, which first integrates the similarity of circRNA and disease to construct features. The integrated features are sent to the deep autoencoder to extract hidden biological patterns. Using the extracted depth features, the random forest classifier is trained for association prediction [20]. Xiao et al. proposed an adaptive subspace learning method based on network embeddings (NSL2CD) to predict potential circRNA association with disease [21]. Through the above analysis, we can see that although the current computational model achieves good results, it also has some defects. First, it is not difficult to see that the training data used by the current model is limited, which has an impact on the robustness of the model. At the same time, the lack of training data also brings about the problem of limited coverage. There are about 10,000 potential associations that these models can predict. Second, they are mainly based on a single data description method, and do not combine circRNA with disease behavior information and attribute information in the network to comprehensively define the characteristics of circRNA and disease, resulting in limited predictive performance. Finally, they did not

take account of the heterogeneity of code-non-coding gene-disease association and could not accurately measure circRNA-disease association information.

In order to improve the shortcomings of existing computational models, in this paper, the TPGSVDCDA (association prediction of circRNA and disease based on tripartite graph and singular value) is proposed where decomposition models were used to predict the potential association between circRNA and disease. In summary, the main contributions of this work are as follows.

1) Motivated by the cooperation between non-coding genes (most circRNAs belong to non-coding genes) and coding genes in human diseases, we effectively constructed the association among circRNA, disease and miRNA, and developed a circRNA-disease-miRNA tripartite graph, to better describe the heterogeneity of disease associations encoding non-coding genes.

2) Use Probs mechanism to allocate resources to the tripartite graph and generate recommendations. The Probs mechanism takes account of the contribution of resources moving in both directions, which effectively reduces the unknowable bias in resource allocation and further improves the prediction performance of TPGSVDCDA.

3) By using different biological information, including circRNA sequence information, miRNA-disease association information, circRNA-disease association information, and disease semantic information, potential candidate genes can obtain more information from other diseases and circRNAs.

2. **Materials and Methods.**

2.1. **circRNA-disease association.** In this article, the circRNA-disease information is downloaded from the circR2Cancer database, and the current version of circR2Cancer contains 1,439 associations between 1,135 circRNAs and 82 types of cancer. After removing circRNAs that did not match the genetic symbols, 647 confirmed circRNA disease associations were obtained, including 514 circRNAs and 62 diseases. Sequence information and gene symbol information for circRNA are available from a number of public databases, such as circBase, deepBase [22], CircNet [23], and circRNADb [24]. In order to build a more complete circRNA sequence dataset, circRNA sequence information was downloaded from the database circBase. In order to integrate multi-source information, miRNA-disease information was downloaded from the miR2Disease database [25]. Disease semantic information was downloaded from the MeSH database [26]. TPGSVDCDA could provide predictive scores for approximately 90,000 unconfirmed circRNA-disease associations. We hope that these improvements will better serve circRNA-disease researchers as a way to move the field forward.

2.2. **Fusing similarities.** In this paper, due to the high level sparsity of the data set and other problems, the calculation process time is long and the algorithm time complexity is high. The calculation by cosine similarity [27] will produce misleading results, compared with traditional similarity measurement methods, Jaccard similarity [28]. SVD's excellent performance in reducing and compressing data makes SVD an important tool for dealing with sparse matrix problems. Therefore, we propose a similarity algorithm based on Jaccard similarity and SVD to calculate the similarity on the tripartite graph. Our basic idea is to combine the Jaccard similarity between vertices in the connection prediction with the SVD of the score matrix to form a new similarity algorithm.

2.2.1. *Jaccard index.* Jaccard index, also known as Jaccard similarity coefficient, is used to compare the similarity and difference between finite sample sets. The higher the value of the Jaccard coefficient, the higher the sample similarity. The basic idea is: for node X

in the network, the number of neighbors is defined as $\Gamma(X)$, then the similarity between two nodes is defined as the number of common neighbors of nodes X and Y divided by the number of elements in the union of nodes X and Y, its formula is as follows:

$$jaccard(X,Y) = \frac{|\Gamma(X) \cap \Gamma(Y)|}{|\Gamma(X) \cup \Gamma(Y)|} \tag{1}$$

In the formula, the value of $|\Gamma(X) \cap \Gamma(Y)|$ is equal to the sum of the product of the corresponding elements of the rows corresponding to vertices X and Y in the network adjacency matrix.

2.2.2. *Singular value decomposition.* Singular value decomposition (SVD) [29] is an important matrix decomposition in linear algebra, which is a generalization of spectral analysis theory to arbitrary matrices. For the circRNA-disease matrix A of order $m \times n$, we decompose the matrix A into three parts U, Σ, V, where U is the orthogonal matrix of order m, V is the orthogonal matrix of order n, Σ is the rectangular diagonal matrix of $m \times n$, the elements on the non-diagonal matrix Σ are all 0, and the elements on the diagonal are the singular values $a_1, a_2, \ldots, a_n$ of the matrix A. We assume that they are arranged in order from largest to smallest, that is, $a_1 \geq a_2 \geq \cdots \geq a_n \geq 0$. Then the singular value decomposition formula of A is expressed as follows: $A = U\Sigma V^T$.

In practical applications such as principal component analysis [30], we tend to focus only on k singular values of A, that is Σ for retaining k largest principal diagonal elements, forming a diagonal matrix of $k * k$, obtaining a new diagonal matrix Σ_k. Matrix U also takes the first k columns accordingly to form a matrix U_k of order $m * k$, and V is also simplified in this way to obtain a matrix V_k of order $n * k$; then there is the matrix $A_1 = U_k * \Sigma_k * V_k^T$ obtained by reconstruction. It is approximately A_1 equal to A, thus achieving the dimensionality reduction of the matrix as shown in Figure 1.

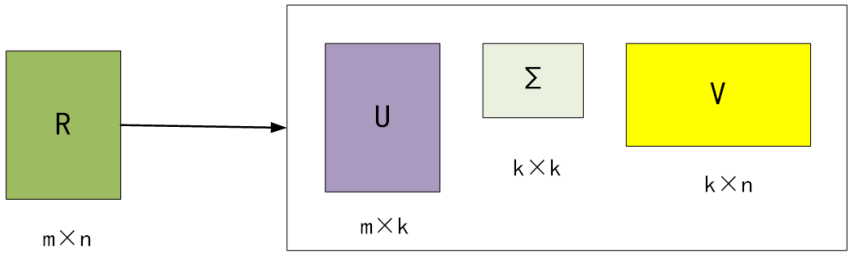

FIGURE 1. Matrix decomposition diagram

What we want to do is based on circRNA or disease recommendations, so we just need to take out the matrix V_k, treat each row of the matrix V_k as a k-dimensional vector, and calculate the similarity between rows to calculate the circRNA and disease similarity matrix. For the sum of the V_k two row vectors $x = (x_1, x_2, \ldots, x_k)$ and $y = (y_1, y_2, \ldots, y_k)$, we use Gaussian kernel similarity to calculate $Sim(x, y)$:

$$Sim(x,y) = GIP(x,y) = e^{-\frac{||x-y||^2}{2\sigma^2}} \tag{2}$$

The basic idea of the Gaussian interaction profile [31]: It is usually defined as a monotone function of the Euclidean distance between any point in space and a central point, and can be denoted as $k(||x-x'||)$, its effect is often local, that is, the value of the function is small when it is far away. It is defined as

$$k(x,x') = e^{-\frac{\|x-x'\|^2}{2\sigma^2}} \tag{3}$$

where x' is the center of the kernel function, $\|x - x'\|$ is the Euclidean distance between the vectors, and the Gaussian kernel function monotonically decreases as the distance between the two vectors increases. σ controls the scope of the Gaussian kernel function, set to 1.

2.2.3. *Calculation of similarity index.* In order to calculate the similarity between circRNA and circRNA, disease and disease, we first use the Jaccard index to find its recommended J_C and J_D similarity matrix. Then, due to the sparsity of circRNA-disease data, many similarity indexes may be 0. To this end, we supplemented the deficiency of the jaccard index similarity matrix based on the similarity obtained by SVD, so as to obtain a complete similarity matrix. Remember that the similarity of X and Y indexes based on Jaccard is $sim_{Jaccard(x,y)}$, and the similarity obtained based on SVD is sim_{SVD}. We define the similarity index after combining the two $sim(x, y)$ as follows:

$$sim(x,y) = \begin{cases} \alpha sim_{Jaccard}(x,y) + (1-\alpha) sim_{SVD}(x,y) & \text{if } sim_{Jaccard}(x,y) \neq 0 \\ sim_{SVD}(x,y) & \text{otherwise} \end{cases} \tag{4}$$

Here, $\alpha \in [0, 1]$, in order to adjust the importance of the two types of similarity parameters, when α is larger, Jaccard similarity is assigned greater importance. The initial value of α is set to 0.4.

2.3. **circRNA-disease-miRNA tripartite graph.**

2.3.1. *Introduction to the tripartite diagram.* A tripartite graph is typically constructed from data from three heterogeneous data sources [32]. In this paper, the tripartite graph is constructed by circRNA-disease and miRNA-disease, in which circRNA similarity and disease similarity are added to process isolated nodes. Inspired by the previous recommendation using the tripartite graph of user, commodity and label, in this paper, the TPGSVDCDA model is first constructed, and the overview in identifying potential circRNA disease associations can be simply summarized as the following five steps. First, circBase database is used to obtain circRNA sequence information, circRNA-disease association information is obtained from circR2Cancer database, miRNA-disease information is obtained from miR2Disease database, and disease semantic information is obtained from MeSH. The data were processed into circRNA-disease association matrix (514×62) and miRNA-disease association matrix (461×62). Secondly, the circRNA-disease association information and miRNA-disease association information were used to construct the circRNA-disease-miRNA tripartite graph. Third, the mixed similarity of Jaccard index and SVD is added to TPGSVDCDA to calculate circRNA similarity and disease similarity, and the interaction profile is calculated for isolated nodes and the change vector is integrated into the adjacency matrix for further resource allocation. Fourthly, the Probs (probability spreading) mechanism is used to allocate resources to the tripartite graphs, and the final circRNA-disease prediction matrix is obtained. Fifth, the prediction scores were calculated and ranked for leave-one cross-validation. We ranked the predicted scores of all circRNAs for each disease in descending order, the higher the score, the more likely the two are related. The circRNA-disease-miRNA tripartite graph $T(C, D, MI, E)$, where $C = (c_1, c_2, c_3, \ldots, c_m)$, $D = (d_1, d_2, d_3, \ldots, d_n)$ and $MI = (mi_1, mi_2, mi_3, \ldots, mi_r)$ are nodal sets of m circRNA, n diseases, and r miRNAs, respectively. E represents the set of interactions (edges) between C and D, D and MI nodes. A tripartite graph can also be expressed as the sum of two adjacency matrices $A^{CD} = \left\{a_{ij}^{CD}\right\}_{m\times n}$, and $A^{MID} = \left\{a_{ij}^{MID}\right\}_{r\times n}$, where, if $circRNAC_i$ is associated with a $diseased_j$, $a_{ij}^{CD} = 1$ and otherwise $a_{ij}^{CD} = 0$, where $circRNAC_i$ is associated with a $diseased_j$ pair as unknown. Similarly, if the

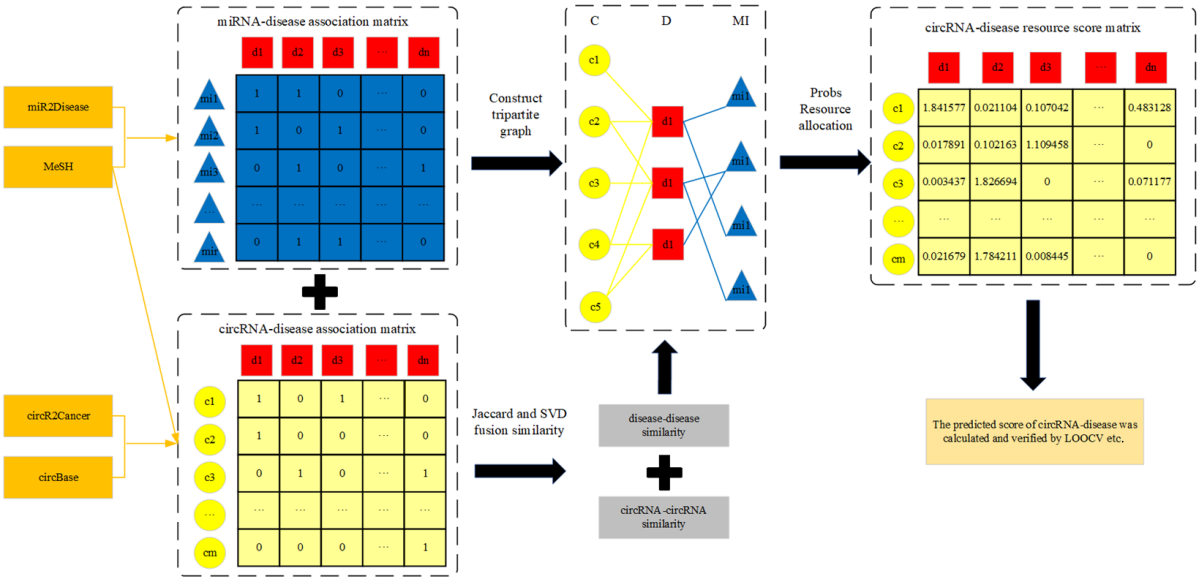

FIGURE 2. Flowchart of circRNA and disease association prediction model based on tripartite graph and singular value decomposition

$diseased_j$ is associated with a $geneMI_i$, $a_{ij}^{MID} = 1$, then 0 if not. The TPGSVDCDA model flow chart is shown in Figure 2.

2.3.2. *Tripartite graph resource recommendation.* The recommendation system based on tripartite graph to be implemented in this paper uses a mechanism called Probs (probability propagation). Here is an introduction.

The resource allocation in the circRNA-disease-miRNA tripartite graph recommendation system is shown in Figure 3. The working principle is as follows. The tripartite graph consists of 5 circRNAs, 4 miRNAs and 3 diseases. The yellow circle, red square and blue triangle represent circRNA, disease and miRNA, respectively. We focus on the potential circRNA-disease association, so take the circRNA-disease association in Figure 3 as an example. In Figure 3(1), initial resources x, y and z are allocated to the above three disease nodes, respectively. In Figure 3(2), from the above node to the below node in the way of equal probability propagation, the so-called equal probability propagation, that is, the resources of each node are equally transmitted to each node with which there is

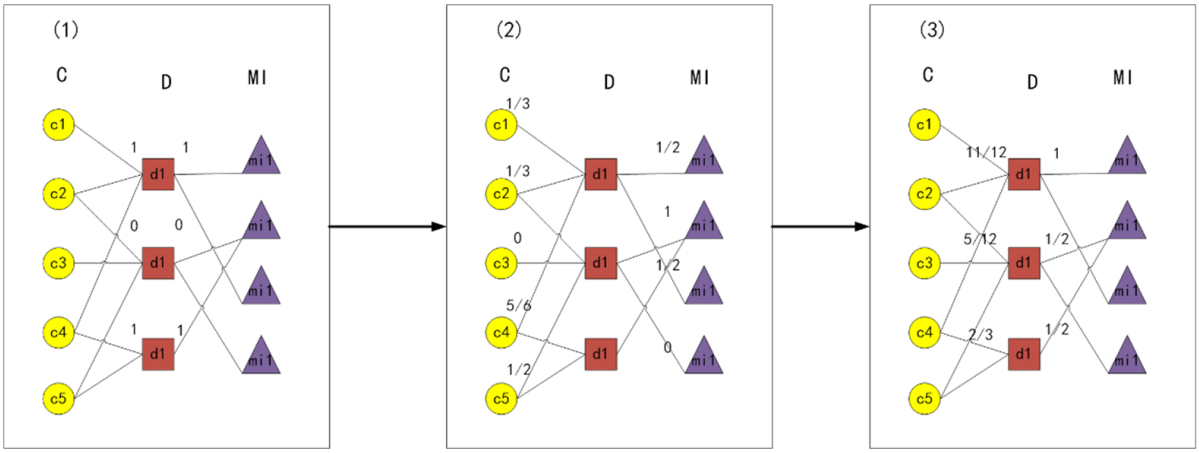

FIGURE 3. Resource allocation graph

a side connection. In Figure 3(3), the resources are transmitted back to the above nodes by equal probability propagation, and it can be seen that the resources of the original three nodes change from x, y, z to $\left(\frac{2}{3}x+\frac{1}{6}y+\frac{1}{4}z,\ \frac{1}{6}x+\frac{2}{3}y+\frac{1}{4}z,\ \frac{1}{6}x+\frac{1}{6}y+\frac{1}{2}z\right)$. In the tripartite graph, there is no edge connection between the nodes of the same part, so the relationship between the nodes of the same part cannot be found directly. After this two-step propagation mechanism, each node is mixed with the resources of other nodes. For example, d1 contains 1/6 of the resources from d2 and 1/4 of the resources from d3. Obviously, d3 is more important to d1 nodes than d2 nodes. Using these coefficients after propagation to represent the weight of the relationship between some nodes, we use x', y', z' to represent the resources after secondary propagation, then

$$\begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = \begin{pmatrix} \frac{2}{3} & \frac{1}{6} & \frac{1}{4} \\ \frac{1}{6} & \frac{2}{3} & \frac{1}{4} \\ \frac{1}{6} & \frac{1}{6} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

The above numerical matrix is the relation weight matrix between nodes, for example, the relation weight between d1 nodes and d3 nodes is 1/4, note that this is an asymmetric; the relation weight between d3 nodes and d1 nodes is 1/6, which is exactly the hidden asymmetry, which is conducive to the circRNA-disease association prediction analysis.

2.3.3. *TPGSVDCDA model implementation.* We modeled the prediction process of circRNA-disease association as resource allocation of circRNA-disease-miRNA tripartite graph. An example of resource allocation in a simple tripartite graph is shown in Figure 3. For a particular circRNAc_i, the initial resource located on the disease d_j is defined as

$$f(c_i) = a_{ij}^{CD}, \quad j = 1, 2, \ldots, n \tag{5}$$

If we select circRNAc_1 in Figure 3(1) as the target circRNA, the initial resource vector is represented as $f(d_1) = (1, 0, 1)$. The resource allocation process in TGPSVDCDA consists of two steps. In the first step of allocation, the initial resource is assigned from a node in D to both a node in C and a node in MI, as if the resource matrix assigned to C is $W_{D\to C}$:

$$W_{D\to C} = \left\{w_{ij}^{CD}\right\}_{n\times n} \tag{6}$$

$$w_{ij}^{CD} = \frac{1}{k_c(c_i)} \sum_{j=1}^{n} \frac{a_{ij}^{CD} a_{xj}^{CD}}{k_c^*(d_j)} \tag{7}$$

In the formula, n represents the number of disease species, and w_{ij}^{CD} represents the contribution from the i to the j node in the disease node in the circRNA-disease resource allocation. a_{ij}^{CD} and a_{xj}^{CD} in the $m*n$ adjacency matrix are composed of circRNA-disease, the entities in column j of row i and column j of row x respectively, $k_c(c_i) = \sum_{j=1}^{n} a_{ij}^{CD}$ represents the number of disease species associated with circRNA i and $k_d^*(d_i) = \sum_{x=1}^{m} a_{ij}^{CD}$ represents the number of circRNA species associated with disease j.

In the second step, resources are transferred from the nodes C and MI back to node D. The resource matrix from C back to D is $W_{C\to D}$:

$$W_{C\to D} = \left\{w_{ij}^{CD}\right\}_{m\times m} \tag{8}$$

$$w_{ij} = \frac{1}{k_c(c_i)} \sum_{j=1}^{n} \frac{a_{ij}^{CD} a_{js}^{CD}}{k_d(d_j)} \tag{9}$$

where m represents the number of circRNA species, w_{ij} is the contribution of the resource moving from the j-th node to the i node in C which can be described as the similarity between circRNAc_i and circRNAc_j. $k_c(c_i) = \sum_{j=1}^{n} a_{ij}^{CD}$ is the number of diseases associated with circRNAc_i, called c_i degrees. Again, $k_d(d_j) = \sum_{s=1}^{m} a_{js}^{CD}$ represents the degree of the d_j node in D.

We further modify the resource allocation algorithm to take account of the level of consistency between the bidirectional moving contributions of the resource, which reflects the effect of co-selection (c_i, c_j) between the contribution of c_i to c_j and c_j to c_i, because the higher the consistency of the two objects, the higher the similarity. Therefore, we defined a consistency-based resource allocation to represent the circRNA-disease association as follows:

$$w'_{ij} = w_{ij} + \frac{w_{ji}}{\sum_{j=1}^{n} w_{j'i}} \tag{10}$$

where w'_{ij} represents the sum of resource allocation contributions between the i-th and j-th nodes in C, and the corresponding weight matrix is rewritten as $W' = \{w'_{ij}\}_{m \times m'}$. Combining the adjacency matrix A^{CD} and the weight matrix W', the final resource f_1 located on D nodes is defined as

$$f_1 = W' \times A^{CD} \tag{11}$$

In the resource allocation between miRNA and disease, the same initial resource on D node is assigned from D node to miRNA node and then transferred back, and the final f_2 resource vector located on D node can be calculated as

$$f_2 = \sum_{s=1}^{r} \frac{a_{js}^{MID}}{k_{mi}(mi_i)} \sum_{j=1}^{n} \frac{a_{ij}^{CD}}{k'_d(d_j)} \tag{12}$$

where $k_{mi}(mi_i) = \sum_{j=1}^{n} a_{ij}^{MID}$ representing the degree of miRNA mi_i in MI, $k'_d(d_j) = \sum_{s=1}^{r} a_{ij}^{MID}$ is the number of disease d_j associated miRNAs. By weighing f_1 and f_2, the final resource score R_{score} used to measure potential circRNA-associated disease, is defined by weighting and performing as follows:

$$R_{score} = \gamma \cdot f_1 + (1 - \gamma) \cdot f_2 \tag{13}$$

where the parameter $\gamma \in [0, 1]$ is adjustable to balance the contribution of circRNA and miRNA to the disease association, with a default value of 0.6.

The inference process of the isolated nodes present when constructing the tripartite graph is as follows. First, we calculate the similarity $SIM(s_{new}, s_i)$ between an isolated node (for example, a new circRNA) and its neighbors, which is calculated by the circRNA expression similarity of an isolated circRNA or the disease semantic similarity of an isolated disease. Second, we compute the interaction profile by the following from S_{new}:

$$S_{new} = \sum_{i=1}^{n_{new}} (SIM(S_{new}, S_i)) \cdot a_i \tag{14}$$

where a_i is the interaction profile vector, and S_{new} reflects the potential relationship between isolated nodes and disease by considering the interaction between neighboring nodes and disease, and then integrates it into the tripartite graph for further resource allocation.

3. Experimental Results and Analysis.

3.1. **Evaluation indicators.** Leave-One-Out Cross Validation (LOOCV) was implemented on our derived circRNA-disease association matrix [33] to assess the performance of TPGSVDCDA in inferring potential associations between circRNA-diseases. AUC values were calculated based on the corresponding area under the ROC curve. Calculate the AUPR value based on the PR curve. The definition is shown as follows:

$$recall = TP_{Rate} = \frac{TP}{TP + FN} \tag{15}$$

$$FP_{Rate} = \frac{FP}{FP + TN} \tag{16}$$

$$precision = \frac{TP}{TP + FP} \tag{17}$$

The horizontal axis of the ROC curve is FP_{Rate}, and the vertical axis is TP_{Rate}.

$$AUC = \frac{\sum_{ins_i \in positiceclass} rank_{insi} - \frac{M \times (M+1)}{2}}{M \times N} \tag{18}$$

where M is the positive sample, N is the negative sample, $insi$ represents the serial number of the Article i sample.

The horizontal axis of the PR curve is TP_{Rate}, and the vertical axis is $precision$.

$$AUPR = \sum_{i=1}^{n-1} (recall_{i+1} - recall_i) \cdot (precision_i) \tag{19}$$

where n is the total number of positive and negative samples.

$$F\text{-}score = \frac{2 \cdot (precision \cdot recall)}{(precision + recall)} \tag{20}$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{21}$$

3.2. **Evaluation of model prediction ability.** For comparison of self-efficacy, AUC, AUPR, F-score and MCC were evaluated using LOOCV and 5-fold CV, 10-fold CV, respectively. For example, the results are shown in Table 1, and it can be seen from the table that this model has good robustness.

Table 1. LOOCV, 5-CV, 10-CV for comparison

Test	AUC%	AUPR%	F-score%	MCC%
LOOCV	95.33	79.55	97.16	88.90
5-CV	95.34	79.57	97.11	88.93
10-CV	95.34	79.56	97.15	88.92
Mean	95.34	79.56	97.14	88.92

3.3. **Comparison with other methods.** In order to verify the performance of TPGSVDCDA, this paper compares TPGSVDCDA with five other models, including KATZHCDA, NCPCDA, SIMCCDA, iCDA-CMG [34] and DRGCNCDA [35]. Corresponding ROC curves of different methods are shown in Figure 4. The AUC values of KATZHCDA, NCPCDA, SIMCCDA, iCDA-CMG and DRGCNCDA were 86.69%, 89.56%, 85.21%, 86.24% and 93.99%, respectively, and the AUC value of TPGSVDCDA was 95.33%, which was superior to the other 5 methods.

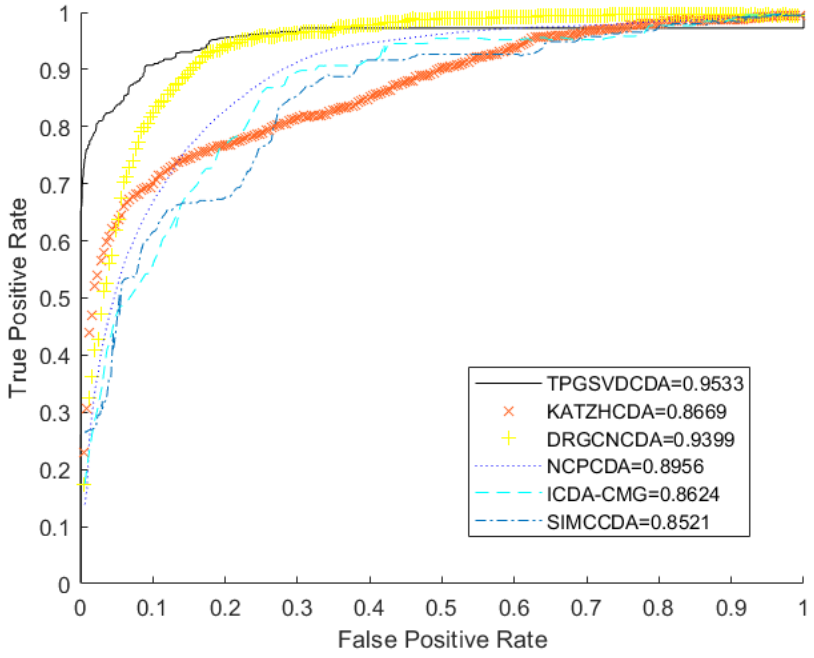

FIGURE 4. Comparison of ROC of TPGSVDCDA and other models

3.4. **Comparative experiment of different similarities in TPGSVDCDA.** In order to verify the effect of different similarities in TPGSVDCDA, this paper compares fusion similarity with Jaccard similarity, cosine similarity, Gauss kernel similarity and Pearson similarity. ROC curves of different similarities are shown in Figure 5. The AUC values of Jaccard similarity, cosine similarity, Gauss kernel similarity and Pearson similarity are 94.46%, 91.47%, 92.21% and 87.09%, respectively, and the AUC value of fusion similarity is 95.33%, which is superior to other similarity calculations.

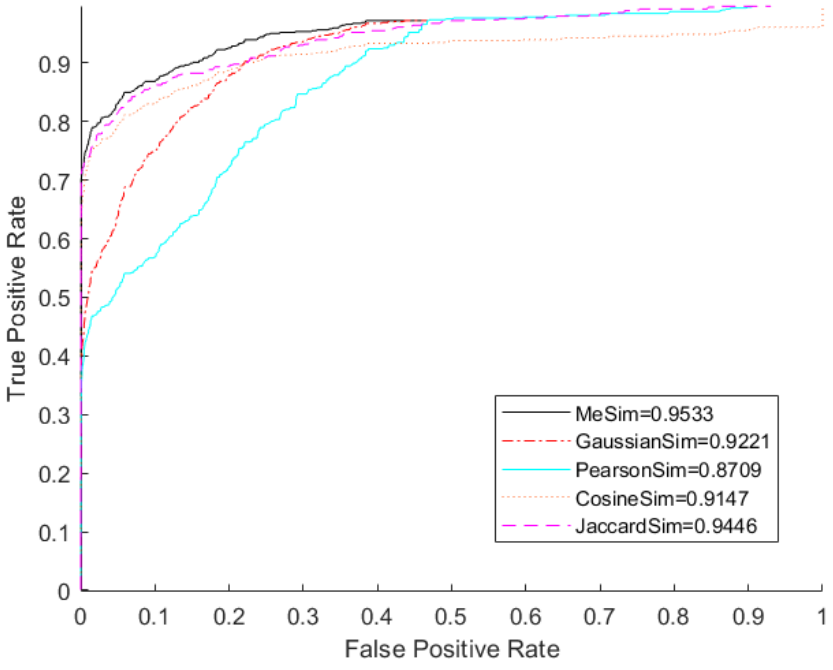

FIGURE 5. Comparison of ROC between fusion similarity and other basic similarities

3.5. **Case study.** To assess the practical value of TPGSVDCDA, we chose gastric cancer [36] to conduct a case study; gastric cancer had sufficient data in the circR2Disease dataset and circR2Cancer dataset to avoid bias due to model defects. By calculating the prediction probability matrix through the candidate set containing unverified circRNAs, we select all known circRNA-disease relationships as the training sample, prioritize all unknown circRNA-disease relationships, and then according to the corresponding prediction scores, we select the top 10 scores by arranging the scores of the probability matrix in descending order, and verified by the already verified databases and literature such as Circr2disease, circR2Cancer and PubMed [37].

The gastric cancer is one of the most common malignant tumors recognized in the world in our country, ranking the third place in the incidence and death of malignant tumor. The study of gastric cancer screening model is of great significance for the early detection of gastric cancer and the early warning of high-risk groups. Screening at the stage of precancerous lesions and early diagnosis and treatment can reduce the incidence and mortality of gastric cancer. Further research on the correlation between gastric cancer and circRNA is helpful to improve the diagnosis and treatment of gastric cancer. This paper aims to provide reference for the establishment and optimization of risk prediction model of high risk groups of gastric cancer in our country. We selected the top 10 circRNAs with prediction scores for validation, and all of the top 10 were validated, as detailed in Table 2. For example, circ_SPECC1 [38] (rank 1) could inhibit the growth and invasion of gastric cancer cells by enhancing the inhibitory effect of miR-526b on the downstream KDM4A/YAP1 pathway through adsorption. circ-NOTCH1 [39] (rank 9) inhibited the transcriptional activity of miR-637 in gastric cancer cells, thereby upregulating the expression of its target gene Apelin and regulating cell proliferation, apoptosis and invasiveness. circ-CEP85L [40] (Rank 3) promotes NFKBIA expression by acting as a sponge for miR-942-5p; thus, inhibiting gastric cancer cell proliferation and invasion, circ-CEP85L is a potential target for the treatment of gastric cancer. Next, we took NFKBIA gene as an example for further analysis to verify whether it is associated with gastric cancer. As shown in Figure 6, in our study, all samples of gastric cancer patients were divided into high expression group and low expression group. Survival analysis showed that the survival time of gastric cancer patients in the low expression group of NFKBIA gene was relatively short. As shown in Figure 7, further results showed that the expression of these genes was significantly lower in the cancer samples than in the normal samples. Based on the above results, we finally concluded that the expression of these genes was significantly positively correlated with the survival time and clinicopathological features of patients with gastric cancer. In addition, STAD enrichment analysis also showed that

Table 2. Top 10 circRNAs associated with gastric cancer

Diseases	Rankings	circRNA	Evidence (PMID)
Stomach Cancer	1	circ-specc1	31349968
Stomach Cancer	2	circatxn7	31997941
Stomach Cancer	3	circ-cep85l	32026471
Stomach Cancer	4	circ-dcaf6	31226266
Stomach Cancer	5	circ-eif4g3	31257089
Stomach Cancer	6	circ-erbb2	30853181
Stomach Cancer	7	circhectd1	34001137
Stomach Cancer	8	circmllt10	31754397
Stomach Cancer	9	circ-notch1	31276627
Stomach Cancer	10	circpdss1	30417526

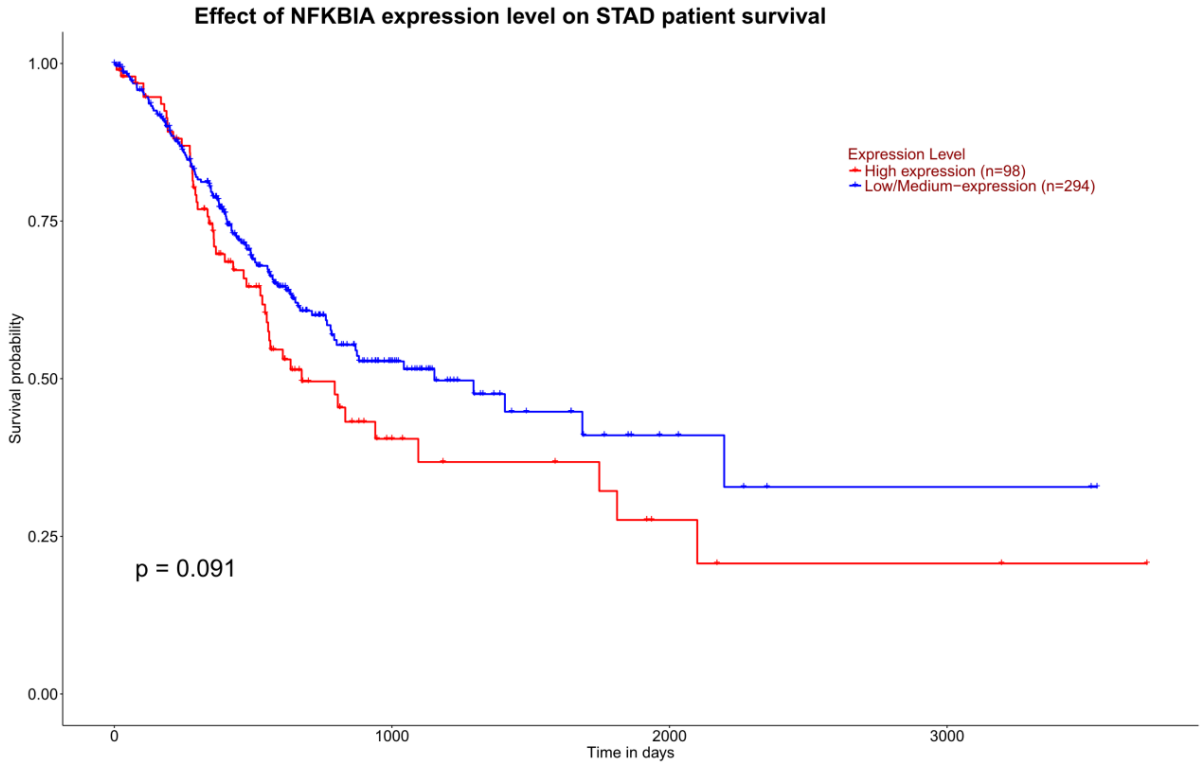

FIGURE 6. Survival analysis of NFKBIA gene in patients with gastric cancer

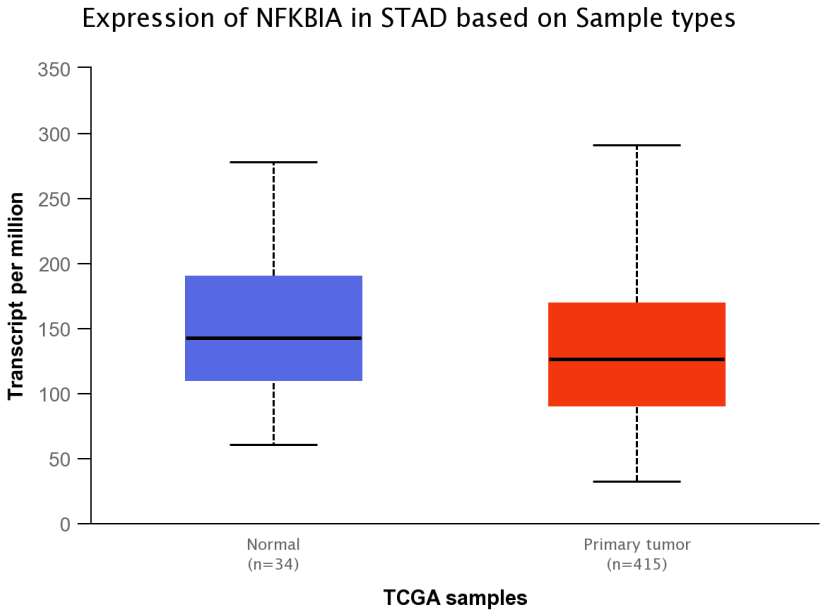

FIGURE 7. Differential expression of NFKBIA gene in normal and tumor samples

for humans, the group with low expression of NFKBIA gene was mainly enriched during the occurrence of immune defects, as shown in Figure 8.

4. **Conclusion.** In this paper, we propose a new computational model, TPGSVDCDA, to identify potential circRNA disease associations by integrating experimentally validated circRNA sequence information, circRNA-disease association information, miRNA-disease

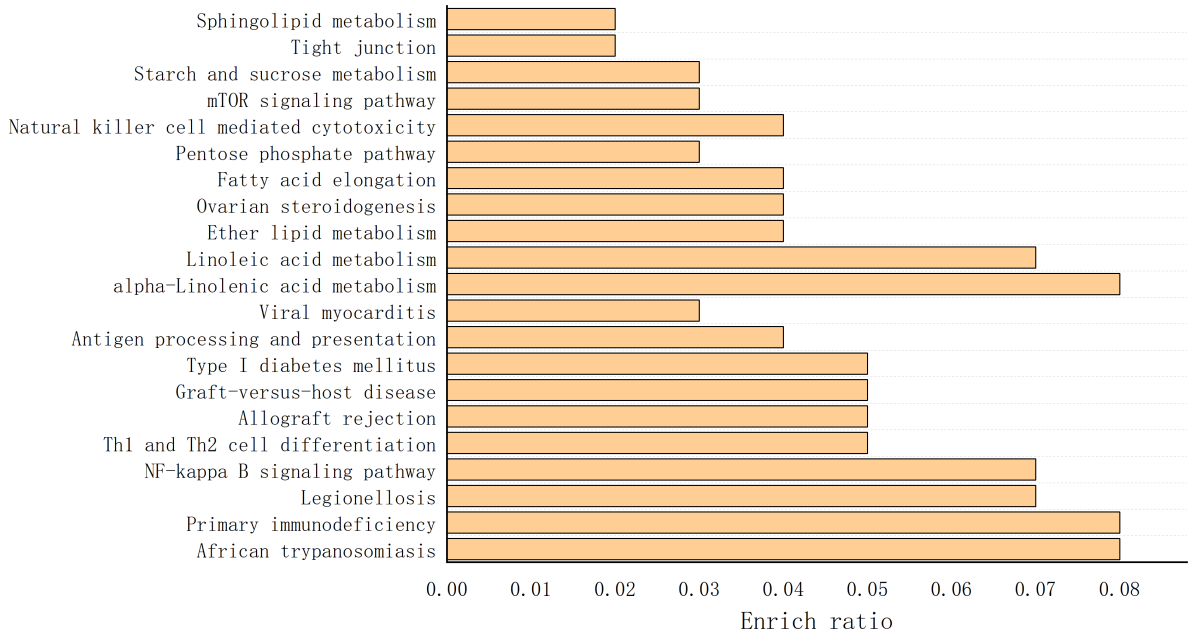

FIGURE 8. Gene set enriched in immune deficiency, based on the KEGG gene set

association information, and disease semantic information. More information about potential candidate genes can be learned from other diseases and circRNAs. Compared with previous approaches, we propose a resource allocation model based on the circRNA-disease-miRNA tripartite graph to better describe the heterogeneity of disease associations encoding non-coding genes. In order to obtain better performance, we then used the Probs propagation mechanism on the constructed circRNA-disease-miRNA tripartite graph to allocate resources to the tripartite graph. The Probs mechanism considered the contribution of resources moving in two directions, and the algorithm effectively reduced the unknowable bias in the resource allocation process. The prediction performance of TPGSVDCDA is further improved. The candidate genes were sequenced. Then by integrating circRNA similarity and disease similarity, TPGSVDCDA can be applied to isolated nodes. Then, Jaccard index and singular value decomposition hybrid recommendation algorithm were used to calculate the similarity of circRNA-disease association matrix to solve the high sparsity of data, so as to better predict the score. TPGSVDCDA has been proved to have good performance and robustness by LOOCV and AUPR experiments. The analysis of the case study provides further evidence that TGPSVDCDA is helpful in identifying potential circRNA disease associations in practice. TPGSVDCDA has potential value in biomedical research for understanding the pathogenesis of diseases and can further improve the quality of disease diagnosis, treatment, prognosis and prevention. Although TPGSVDCDA has achieved relatively good results, there are still some limitations that need further study. First, TPGSVDCDA relies on a tripartite graph topology, so data incompleteness can limit its performance. Therefore, we intend to integrate circRNA-miRNA association information or adopt other biological information in the future. Second, our approach focuses on unweighted tripartite graphs, and finally, the number of experimentally available circRNA disease associations remains relatively small. As biotechnology continues to evolve, the performance of TPGSVDCDA is expected to improve further as more experimental validation associations become available.

Acknowledgment. This work is partially supported by the research project with project number 145209125 funded by the General Research Fund for Higher Education Institutions in Heilongjiang Province. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Y. Zhang, X. Zhang, T. Chen, J. F. Xiang, Q. Yin, Y. H. Xing, S. Zhu, L. Yang and L. L. Chen, Circular intronic long noncoding RNAs, *Molecular Cell*, vol.51, no.6, pp.792-806, 2013.

[2] W. Du, L. Fang, W. Yang, N. Wu, F. Awan, Z. Yang and B. Yang, Induction of tumor apoptosis through a circular RNA enhancing Foxo3 activity, *Cell Death & Differentiation*, vol.24, no.2, pp.357-370, 2017.

[3] P. Li, S. Chen, H. Chen, X. Mo, T. Li, Y. Shao, B. Xiao and J. Guo, Using circular RNA as a novel type of biomarker in the screening of gastric cancer, *Clinica Chimica Acta*, vol.444, pp.132-136, 2015.

[4] B. Zhou and J. Yu, A novel identified circular RNA, circRNA_010567, promotes myocardial fibrosis via suppressing miR-141 by targeting TGF-β1, *Biochemical and Biophysical Research Communications*, vol.487, no.4, pp.769-775, 2017.

[5] H. F. Liang, X. Zhang, B. Liu, G. Jia and W. Li, Circular RNA circ-ABCB10 promotes breast cancer proliferation and progression through sponging miR-1271, *American Journal of Cancer Research*, vol.7, no.7, pp.1566-1576, 2017.

[6] L. Wang, Z. You, Y. Li, K. Zheng and Y. Huang, GCNCDA: A new method for predicting circRNA-disease associations based on Graph Convolutional Network Algorithm, *PLoS Computational Biology*, vol.16, no.5, e1007568, 2020.

[7] L. Wang, X. Yan, M. Liu, K. Song, X. Sun and W. Pan, Prediction of RNA-protein interactions by combining deep convolutional neural network with feature selection ensemble method, *Journal of Theoretical Biology*, vol.461, pp.230-238, 2019.

[8] P. Glažar, P. Papavasileiou and N. Rajewsky, circBase: A database for circular RNAs, *RNA*, vol.20, no.11, pp.1666-1670, 2014.

[9] C. Fan, X. Lei, Z. Fang, Q. Jiang and F. Wu, CircR2Disease: A manually curated database for experimentally supported circular RNAs associated with various diseases, *Database (Oxford)*, DOI: 10.1093/database/bay044, 2018.

[10] W. Lan, M. Zhu, Q. Chen, B. Chen, J. Liu, M. Li and Y. Chen, CircR2Cancer: A manually curated database of associations between circRNAs and cancers, *Database (Oxford)*, DOI: 10.1093/database/baaa085, 2020.

[11] Z. Zhao, K. Wang, F. Wu, W. Wang, K. Zhang, H. Hu, Y. Liu and T. Jiang, circRNA disease: A manually curated database of experimentally supported circRNA-disease associations, *Cell Death & Disease*, vol.9, no.5, DOI: 10.1038/s41419-018-0503-3, 2018.

[12] S. Ghosal, S. Das, R. Sen, P. Basak and J. Chakrabarti, Circ2Traits: A comprehensive database for circular RNA potentially associated with disease and traits, *Frontiers in Genetics*, vol.4, DOI: 10.3389/fgene.2013.00283, 2013.

[13] C. Fan, X. Lei and F. Wu, Prediction of circRNA-disease associations using KATZ model based on heterogeneous networks, *International Journal of Biological Sciences*, vol.14, no.14, pp.1950-1959, 2018.

[14] X. Lei, Z. Fang, L. Chen and F. Wu, PWCDA: Path weighted method for predicting circRNA-disease associations, *International Journal of Molecular Sciences*, vol.19, no.11, DOI: 10.3390/ijms19113410, 2018.

[15] G. Li, Y. Yue, C. Liang, Q. Xiao, P. Ding and J. Luo, NCPCDA: Network consistency projection for circRNA-disease association prediction, *RSC Advances*, vol.9, no.57, pp.33222-33228, 2019.

[16] X. Lei, Z. Fang and L. Guo, Predicting circRNA-disease associations based on improved collaboration filtering recommendation system with multiple data, *Frontiers in Genetics*, vol.10, DOI: 10.3389/fgene.2019.00897, 2019.

[17] Y. Ding, B. Chen, X. Lei, B. Liao and F. Wu, Predicting novel CircRNA-disease associations based on random walk and logistic regression model, *Computational Biology and Chemistry*, vol.87, DOI: 10.1016/j.compbiolchem.2020.107287, 2020.

[18] Q. Xiao, H. Yu, J. Zhong, C. Liang, G. Li, P. Ding and J. Luo, An in-silico method with graph-based multi-label learning for large-scale prediction of circRNA-disease associations, *Genomics*, vol.112, no.5, pp.3407-3415, 2020.

[19] C. Fan, X. Lei and Y. Pan, Prioritizing CircRNA-disease associations with convolutional neural network based on multiple similarity feature fusion, *Frontiers in Genetics*, vol.11, DOI: 10.3389/fgene.2020.540751, 2020.

[20] K. Deepthi and A. Jereesh, Inferring potential circRNA-disease associations via deep autoencoder-based classification, *Molecular Diagnosis & Therapy*, vol.25, no.1, pp.87-97, 2021.

[21] Q. Xiao, Y. Fu, Y. Yang, J. Dai and J. Luo, NSL2CD: Identifying potential circRNA-disease associations based on network embedding and subspace learning, *Briefings in Bioinformatics*, vol.22, no.6, DOI: 10.1093/bib/bbab177, 2021.

[22] J. Yang, P. Shao, H. Zhou, Y. Chen and L. Qu, deepBase: A database for deeply annotating and mining deep sequencing data, *Nucleic Acids Research*, vol.38, pp.D123-D130, 2010.

[23] Y. Liu, J. Li, C. Sun, E. Andrews, R. Chao, F. Lin, S. Weng, S. Hsu, C. Huang, C. Cheng, C. Liu and H. Huang, CircNet: A database of circular RNAs derived from transcriptome sequencing data, *Nucleic Acids Research*, vol.44, pp.D209-D215, 2016.

[24] X. Chen, P. Han, T. Zhou, X. Guo, X. Song and Y. Li, circRNADb: A comprehensive database for human circular RNAs with protein-coding annotations, *Scientific Reports*, vol.6, DOI: 10.1038/srep34985, 2016.

[25] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang and Y. Liu, miR2Disease: A manually curated database for microRNA deregulation in human disease, *Nucleic Acids Research*, vol.37, pp.D98-D104, 2009.

[26] J. Wang, Z. Du, R. Payattakool, P. Yu and C. Chen, A new method to measure the semantic similarity of GO terms, *Bioinformatics*, vol.23, no.10, pp.1274-1281, 2007.

[27] B. Wang, C. Zhang, X. Du and J. Zhang, lncRNA-disease association prediction based on latent factor model and projection, *Scientific Reports*, vol.11, no.1, DOI: 10.1038/s41598-021-99493-5, 2021.

[28] L. Wang, Z. You, J. Li and Y. Huang, IMS-CDA: Prediction of circRNA-disease associations from the integration of multisource similarity information with deep stacked autoencoder model, *IEEE Transactions on Cybernetics*, vol.51, no.11, pp.5522-5531, 2021.

[29] J. Li, J. Li, M. Kong, D. Wang, K. Fu and J. Shi, SVDNVLDA: Predicting lncRNA-disease associations by singular value decomposition and node2vec, *BMC Bioinformatics*, vol.22, no.1, DOI: 10.1186/s12859-021-04457-1, 2021.

[30] R. Zhu, Y. Wang, J. Liu and L. Dai, IPCARF: Improving lncRNA-disease association prediction using incremental principal component analysis feature selection and a random forest classifier, *BMC Bioinformatics*, vol.22, no.1, DOI: 10.1186/s12859-021-04104-9, 2021.

[31] M. Niu, Q. Zou and C. Wang, GMNN2CD: Identification of circRNA-disease associations based on variational inference and graph Markov neural networks, *Bioinformatics*, vol.38, no.8, pp.2246-2253, DOI: 10.1093/bioinformatics/btac079, 2022.

[32] S. Alaimo, R. Giugno and A. Pulvirenti, ncPred: ncRNA-disease association prediction through tripartite network-based inference, *Frontiers in Bioengineering and Biotechnology*, vol.2, DOI: 10.3389/fbioe.2014.00071, 2014.

[33] W. Sun, Constrained role-engineering optimization using Boolean matrix decomposition and integer linear programming techniques, *International Journal of Innovative Computing, Information and Control*, vol.18, no.4, pp.1037-1053, 2022.

[34] Q. Xiao, J. Zhong, X. Tang and J. Luo, iCDA-CMG: Identifying circRNA-disease associations by federating multi-similarity fusion and collective matrix completion, *Molecular Genetics and Genomics*, vol.296, no.1, pp.223-233, 2021.

[35] W. Lan, H. Zhang, Y. Dong, Q. Chen, J. Cao, W. Peng, J. Liu and M. Li, DRGCNCDA: Predicting circRNA-disease interactions based on knowledge graph and disentangled relational graph convolutional network, *Methods*, vol.208, pp.35-41, 2022.

[36] M. Zeng, C. Lu, F. Zhang, Y. Li, F. Wu, Y. Li and M. Li, SDLDA: lncRNA-disease association prediction based on singular value decomposition and deep learning, *Methods*, vol.179, pp.73-80, 2020.

[37] Y. Chen, J. Wang, C. Wang, M. Liu and Q. Zou, Deep learning models for disease-associated circRNA prediction: A review, *Briefings in Bioinformatics*, vol.23, no.6, DOI: 10.1093/bib/bbac364, 2022.

[38] L. Chen, L. Wang and X. Ma, Circ_SPECC1 enhances the inhibition of miR-526b on downstream KDM4A/YAP1 pathway to regulate the growth and invasion of gastric cancer cells, *Biochemical and Biophysical Research Communications*, vol.517, no.2, pp.253-259, 2019.

[39] E. Guan, X. Xu and F. Xue, circ-NOTCH1 acts as a sponge of miR-637 and affects the expression of its target gene Apelin to regulate gastric cancer cell growth, *Biochemistry and Cell Biology*, vol.98, no.2, pp.164-170, 2020.

[40] J. Lu, Y. Wang, X. Huang, J. Xie, J. Wang, J. Lin, Q. Chen, L. Cao, C. Huang, C. Zheng and P. Li, circ-CEP85L suppresses the proliferation and invasion of gastric cancer by regulating NFKBIA expression via miR-942-5p, *Journal of Cellular Physiology*, vol.235, no.9, pp.6287-6299, 2020.

Author Biography

Bo Wang, professor and master tutor, received his Ph.D. in Computer Science from Harbin Engineering University, Harbin, China, in 2021. Currently, he is the discipline leader of computer application technology, the person in charge of electronic information field, and the dean of the Department of Computer Science and Technology at Qiqihar University. His research interests include big data analysis and mining, lncRNA and disease association prediction, deep learning.

Tingbin Liu, who received his bachelor's degree in Computer Science and Technology from Taishan University in Shandong Province, China, in 2017, is currently studying for a master's degree in Qiqihar University in Heilongjiang Province. His main research interest is circRNA and disease association prediction.

Jingyou Li received the Ph.D. certificate of completion in Information and Communication Engineering, Harbin Engineering University, Harbin, Heilongjiang, China, in 2019. His research interests include intelligent information processing, information security, data mining, and cloud computing. He received the M.E. degree in Communication and Information Systems, Harbin Engineering University, in 2007. He was also a visiting scholar in the School of Computer Science and Technology, Beijing University of Posts and Telecommunications (2011-2012). He is a professor and Master's supervisor at School of Computer and Control Engineering, Qiqihar University.

Xiaoxin Du, master and master tutor, graduated from China University of Mining and Technology, China, in 2007, majoring in Computer Science and Technology. She is an associate professor of Software Engineering, School of Computer and Control Engineering, Qiqihar University. Her research interests include swarm intelligent optimization algorithm and application, big data analysis and mining, multimodal medical image enhancement and registration.

Guangda Zhang received the M.E. degree in Communication and Information Systems, Harbin Engineering University, Harbin, Heilongjiang, China, in 2007. Her research interests include computer networks, information security, and big data analysis. She was also a visiting scholar in the School of Information Science and Technology, Beijing Normal University (2015-2016). She is an associate professor at School of Computer and Control Engineering, Qiqihar University.