BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 03/06 pp280–291
Volume 3, Number 4, December 2020
DOI: 10.26599/BDMA.2020.9020025

CircRNA-Disease Associations Prediction Based on Metapath2vec++ and Matrix Factorization

Yuchen Zhang, Xiujuan Lei*, Zengqiang Fang, and Yi Pan*

Abstract: Circular RNA (circRNA) is a novel non-coding endogenous RNAs. Evidence has shown that circRNAs are related to many biological processes and play essential roles in different biological functions. Although increasing numbers of circRNAs are discovered using high-throughput sequencing technologies, these techniques are still time-consuming and costly. In this study, we propose a computational method to predict circRNA-disesae associations which is based on metapath2vec++ and matrix factorization with integrated multiple data (called PCD_MVMF). To construct more reliable networks, various aspects are considered. Firstly, circRNA annotation, sequence, and functional similarity networks are established, and disease-related genes and semantics are adopted to construct disease functional and semantic similarity networks. Secondly, metapath2vec++ is applied on an integrated heterogeneous network to learn the embedded features and initial prediction score. Finally, we use matrix factorization, take similarity as a constraint, and optimize it to obtain the final prediction results. Leave-one-out cross-validation, five-fold cross-validation, and f-measure are adopted to evaluate the performance of PCD_MVMF. These evaluation metrics verify that PCD_MVMF has better prediction performance than other methods. To further illustrate the performance of PCD_MVMF, case studies of common diseases are conducted. Therefore, PCD_MVMF can be regarded as a reliable and useful circRNA-disease association prediction tool.

Key words: circular RNAs (circRNAs); circRNA-disease associations; matepath2vec++; matrix factorization

1 Introduction

Recently, circular RNA (circRNA), which is a novel biological molecule circRNA[1], has attracted considerable attention. CircRNA plays essential roles in different biological functions and controls the expressions of genes[2]. In contrast to linear RNAs that have the exposed 3' caps and 5' tails, the structure of circRNA is a closed loop with neither 5' to 3' polarity nor polyadenylated tail[3]. The first circRNA was discovered in the plant viroids[4]. Because of its stable loop structure and low expression level[5,6], circRNAs are always identified as molecular fragments or by-products of transcription. However, with the development of high-throughput sequencing techniques, increasing numbers of circRNAs are discovered gradually. Simultaneously, circRNA-related biological functions illustrate that circRNAs are endogenous, abundant, conserved, and stable in mammalian cells[2,7,8]. The accumulated evidence shows that circRNAs can be divided into four types, namely, exonic circRNAs, which are mainly derived from back-spliced exons[9]; intronic circRNAs, which are predominantly generated by Groups I and II introns, i.e., intron lariats and excised tRNA introns[10]; exon-intron circRNAs[8], which are exons circularized with introns retained between exons; and intergenic circRNAs[11], which consist of two intronic circRNA fragments. Evidence shows that circRNAs

• Yuchen Zhang, Xiujuan Lei, and Zengqiang Fang are with the School of Computer Science, Shaanxi Normal University, Xi'an 710119, China. E-mail: yczhang@snnu.edu.cn; xjlei@snnu.edu.cn; fangzq@snnu.edu.cn.

• Yi Pan is with the Department of Computer Science, Georgia State University, Atlanta, GA 30302, USA. E-mail: yipan@gsu.edu.

∗ To whom correspondence should be addressed.

Manuscript received: 2020-06-02; revised: 2020-10-09; accepted: 2020-10-10

play essential roles or functions in many biological procedures. circRNAs can be regarded as competing endogenous RNAs or miRNAs'sponges, which has been proven by some previous studies of circ-SRY[12], circ-HIPK3[13], and mm9_circ_012559[14]. Meanwhile, some studies show that circRNAs can interact with RNA-Binding Proteins (RBPs)[15]. circRNAs can not only regulate gene transcription processes[16], but can also be translated into proteins[17, 18].

Moreover, circRNAs can affect diverse biological processes, and can be associated with different complex diseases[19]. circRNA has some unique characteristics, such as conservation, abundance, and tissue specificity, which make them potential disease markers, particularly for some tumors[20]. According to the different expression levels of circRNAs in different tissues, we can identify the differential expression between normal people and the patients. Therefore, these differences can help in the prognoses or diagnosis of diseases. Through the use of quantitative Polymerase Chain Reaction (qPCR) techniques, circRNA expression in lung cancer tissues can be compared with that in neighboring normal tissues. CircRNA ciRS-7[21] can be down-regulated in lung cancer tissues or cells, whereas both circRNA circRNA_100876[22] and hsa_circ_0013958[23] can be up-regulated in lung cancer tissues, cells or plasma. Through the microarray chip technique, some circRNAs expression considerably differs between gastric cancer tissues and neighboring normal tissues, which indicates that circRNAs can be regarded as a biomarker for gastric cancer diagnosis and progression[24]. For example, both circRNA circPVT1[25] and hsa_circ_0000096[26] can affect gastric tissues or cells through the down-regulation mechanism. Moreover, circRNAs can function as miRNA sponges or gene regulators. For example, circRNA hsa_circ_001569[27] is a sponge of miRNA miR-145, which can promote the expression of its target genes in colorectal cancer cells.

In addition, some circRNA-related databases are established. circBase[28] is one of the earliest circRNA-related databases, which provides the location on chromosomes, base sequences, and target genes of circRNA. CircRNADb[29] is also a widely used circRNA database, which has collected a larger number of circRNA annotations data extracted from genomic information, exon splicing, and genome sequences. To analyze circRNA expression in different tissues, exoRBase[30] is set up to provide the circRNA, lncRNA, and mRNA information of human blood exosomes. The CircNet[31] database employs the circRNA expression in RNA-seq samples to identify the circRNA regulation pathways and tissue-specific expression profiles systematically. Moreover, some databases provide information on the associations between circRNAs and diseases. Circ2Traits[32] utilizes circRNA-miRNA, miRNA-disease, and disease-Single Nucleotide Polymorphisms (SNPs) associations to produce circRNA-disease associations. Recently, researchers have focused on the associations between circRNA individuals and single diseases. To make the disease-circRNA research more efficient, some databases, such as circR2Disease[33] (http://bioinfo.snnu.edu.cn/CircR2Disease/), circRNA disease[34], and Circ2Disease[35], collect the information of scattered circRNA-disease associations manually by extracting from thousands of literature.

Although high-throughput sequencing techniques have already been applied to identify circRNA-disease associations, these techniques have some limitations. These techniques can extract circRNA-disease associations with high accuracy, but are still time-consuming and costly. Thus, many scholars have investigated the use of computational methods to identify the circRNA-disease association. Xiao et al.[36] proposed a manifold regularization learning framework for predicting circRNA-disease associations. Yan et al.[37] developed a method called DWNN-RLS to predict circRNA-disease associations based on the regularized least squares of the Kronecker product kernel. The matrix factorization model was also used in this area[38]. A Graph Convolutional Network (GCN), which combines multiple features of nodes and networks, has also been developed[39, 40]. Lei et al.[41] used the gradient boosting decision tree to predict circRNA-disease associations. Lei et al.[42] also adopted the collaboration filtering recommendation system to explore potential relationships. Fan et al.[43] proposed a novel method called MSFCNN for predicting circRNA-disease association. In this study, we propose a novel computational method to predict circRNA-disease which is based on metapath2vec++ and matrix factorization association with multiple biological data sources (called PCD_MVMF). First, the initial circRNA-disease association data are downloaded from circR2Diaeae[33] database. A total of 212 circRNA-disease associations based on 42 disease individuals and

200 circRNA individuals is screened out from the initial dataset. On the basis of 42 disease individuals, diseases functional and semantic similarity scores are computed to construct disease similarity networks. The circRNA annotation, functional, and sequence similarity scores are computed to build circRNA similarity networks based on 200 circRNA entries. Then, different disease similarity networks and circRNA similarity networks are integrated into the final circRNA and disease similarity networks. Afterward, the metapath2vec++ model[44] is adopted to learn the embedded features. Metapath2vec++ is a method of representation learning that learns the network topology and determines embedded features of each node. Finally, we use these embedded features and matrix factorization to predict the circRNA-disease associations. To evaluate the performance of our proposed computational method, several metrics, such as Leave-One-Out Cross-Validation (LOOCV), 5-fold Cross-Validation (CV), and f-measure, are applied. To obtain more reliable evaluation results, case studies of some common diseases are conducted. The overall framework of the method is shown in Fig. 1.

2 Material and method

2.1 CircRNA-disease associations

The circRNA-disease associations data used in our paper, which have been verified by experiments, are screened out from the circR2Disease[33] database (http://bioinfo.snnu.edu.cn/CircR2Disease/). The initial dataset has 725 circRNA-disease associations in the initial dataset which includes 661 circRNA entries and 100 disease entries. To integrate other features on circRNAs and diseases, 200 circRNA individuals and 42 disease individuals are selected on the basis of 212 circRNA-disease associations, which are represented by the adjacent matrix $\boldsymbol{A}$. If the circRNA c_i is associated with the disease d_j, $\boldsymbol{A}(i,j)$ is equal to 1. Otherwise, $\boldsymbol{A}(i,j)$ is equal to 0.

2.2 CircRNA similarity network

2.2.1 CircRNA annotation similarity network

The circRNA annotation similarity network is constructed using the circRNA target-gene-related Gene Ontology (GO) terms. On the basis of the 200 circRNA entries, circRNA target-gene-GO terms data are downloaded from the Human Protein Reference Database (HPRD)[45]. All of the matching GO terms data are used to calculate the circRNA annotation similarity scores, which are adopted to construct the circRNA annotation similarity network ($\boldsymbol{C}_{\text{AS}}$). In the study, an information content algorithm[46] is used to calculate the similarity score of two circRNAs. Specifically, the similarity score between the circRNA c_i and c_j is calculated as follows:

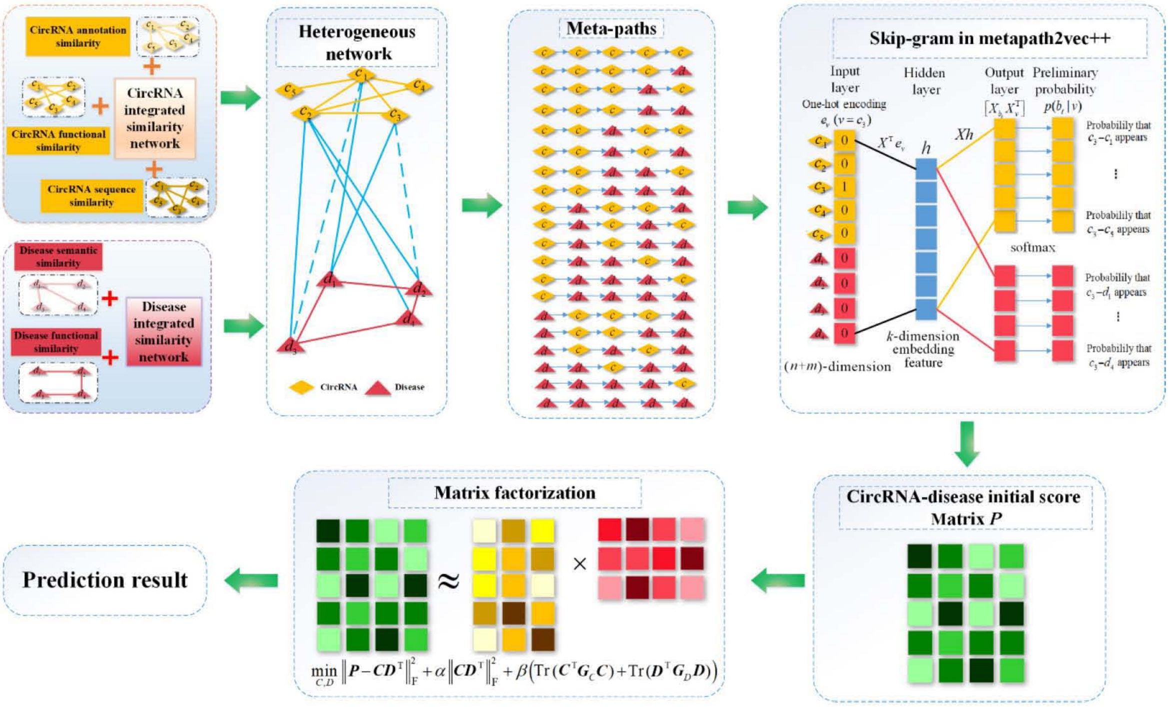
CircRNA annotation similarity
CircRNA functional similarity
CircRNA sequence similarity
CircRNA integrated similarity network
Disease semantic similarity
Disease functional similarity
Disease integrated similarity network
Heterogeneous network
CircRNA
Disease
Meta-paths
Skip-gram in metapath2vec++
Input layer
One-hot encoding
e_v ($v=c_3$)
Hidden layer
Output layer
$[X_{b_t} X_v^{\mathrm{T}}]$
Preliminary probability
$p(b_t|v)$
$X^{\mathrm{T}}e_v$
h
Xh
Probabilily that c_3–c_1 appears
Probabilily that c_3–c_5 appears
softmax
Probabilily that c_3–d_1 appears
Probabilily that c_3–d_4 appears
k-dimension embedding feature
$(n+m)$-dimension
CircRNA-disease initial score Matrix P
Matrix factorization
≈
×
$\min_{C,D}\|P-CD^{\mathrm{T}}\|_F^2+\alpha\|CD^{\mathrm{T}}\|_F^2+\beta(\mathrm{Tr}(C^{\mathrm{T}}G_C C)+\mathrm{Tr}(D^{\mathrm{T}}G_D D))$
Prediction result

Fig. 1 Overall framework of PCD_MVMF.

$$C_{\mathrm{AS}}(i,j)=\frac{2\times\log(\mathrm{Pro}(c_i\cup c_j))}{\log(\mathrm{Pro}(c_i))+\log(Pro(c_j))} \quad (1)$$

where $C_{\mathrm{AS}}(i,\ j)$ denotes the similarity score between the circRNAs c_i and c_j. Pro(c_i) (Pro(c_j)) denotes the proportion between the number of target-genes-related GO terms of circRNA $c_i(c_j)$ and the total number of all the circRNAs target-gene-related GO terms. Pro($c_i\cup c_j$) denotes the proportion of the number of the union of the circRNAs c_i and c_j target-gene-related GO terms and the total number of GO terms.

2.2.2 CircRNA sequence similarity network

To consider the circRNA sequence information, base sequence of circRNA is adopted to calculate the circRNA sequence similarity scores. On the basis of the matching 200 circRNA individuals, circRNA base sequence data are extracted from the circBase[28] database. To calculate the circRNA sequence similarity scores, a sequence alignment algorithm, called Smith-Waterman (SW) pairwise alignment algorithm, is packaged using the python tool, Biopython[47]. In this study, C_{SS} denotes the circRNA sequence similarity network. The weight of each edge in C_{SS} is normalized as follows:

$$C_{\mathrm{SS}}(i,j)=\frac{\mathrm{SW}_s(c_i,c_j)}{\max\left(\mathrm{SW}_s(c_i,c_i),\mathrm{SW}_s(c_j,c_j)\right)} \quad (2)$$

where $\mathrm{SW}_s(c_i,c_j)$ is the Smith-Waterman pairwise alignment score between circRNAs c_i and c_j.

2.2.3 CircRNA functional similarity network

To calculate the functional similarity score of two circRNAs, similar diseases that are associated with them need to be considered. Thereby, the semantic similarity between one disease and a group of disease is adopted to calculate the maximum similarity score between one disease gt and a group of disease GT, which is defined as $S_{\max}$ (gt, GT) illustrated as follows:

$$S_{\max}(\mathrm{gt},\mathrm{GT})=\max_{l\leqslant i\leqslant t}(S(\mathrm{gt},\mathrm{GT}(i))) \quad (3)$$

The circRNA functional similarity between two circRNAs is calculated as follows:

$$C_{\mathrm{FS}}(i,j)=\frac{\sum\limits_{1<l<n} S_{\max}(\mathrm{gt}_{il},\mathrm{GT}_j)+\sum\limits_{1<q<m} S_{\max}(\mathrm{gt}_{jq},\mathrm{GT}_i)}{n+m} \quad (4)$$

where C_{FS} is the circRNA functional similarity network. GT_i and GT_j denote the circRNA c_i and c_j related diseases sets, respectively. gt_{il} and gt_{jq} denote one disease of the disease sets in GT_i and GT_j, respectively. Moreover, n and m denote the number of circRNAs c_i and c_j related diseases, respectively.

2.3 Disease similarity network

2.3.1 Disease semantic similarity network

To calculate the disease semantic similarity score between two disease entries, first, the initial names of the 42 matching diseases are replaced by the corresponding Disease Ontology ID (DOID) manually on the basis of the disease ontology database[48]. Afterward, an R/conductor package named DOSE[49] for disease ontology semantic and enrichment analysis is adopted to calculate the semantic similarity score between each two disease entries. Each disease-entry-related DOID is inputted into DOSE, which is used to set the weight of each edge in the disease semantic similarity network. In this study, D_{SS} denotes the disease semantic similarity network.

2.3.2 Disease functional similarity network

The information on disease functional features should also be considered, because it can provide a more reliable similarity score of each disease pair. In this study, disease-related genes are adopted to describe disease functions. Thus, disease-corresponding genes are downloaded from the DisGeNET[50] database which has collected 3 815 056 gene-disease associations between 16 666 genes and 13 172 diseases. In this study, a statistical algorithm, i.e., the Jaccard index, is used to calculate the disease functional similarity score as follows:

$$D_{\mathrm{FS}}(i,j)=\frac{|\mathrm{DG}_i\cap\mathrm{DG}_j|}{|\mathrm{DG}_i\cup\mathrm{DG}_j|} \quad (5)$$

where D_{FS} represents the disease functional similarity network. $D_{\mathrm{FS}}(i,\ j)$ is the weight of the edge between disease d_i and d_j. DG_i and DG_j illustrate the disease d_i- and d_j-related genes datasets, respectively.

2.4 Integration similarity network for circRNAs and diseases

After all of the aforementioned circRNA and disease similarity networks are constructed, the circRNA annotation, sequence, and functional similarity networks, as well as the disease semantic and functional similarity networks are integrated into the finial circRNA and disease similarity networks for computational modeling. The final combination circRNA similarity network C_{S} and disease similarity network D_{S} are calculated as follows, respectively:

$$C_{\mathrm{S}}(i,j)=\frac{C_{\mathrm{AS}}(i,j)+C_{\mathrm{SS}}(i,j)+C_{\mathrm{FS}}(i,j)}{3} \quad (6)$$

$$D_{\mathrm{S}}(i,j)=\frac{D_{\mathrm{SS}}(i,j)+D_{\mathrm{FS}}(i,j)}{2} \quad (7)$$

where $\boldsymbol{C}_{\mathrm{S}}$ and $\boldsymbol{D}_{\mathrm{S}}$ denote the integrated circRNA and disease similarity networks, respectively. $\boldsymbol{C}_{\mathrm{S}}(i,\ j)$ and $\boldsymbol{D}_{\mathrm{S}}(i,\ j)$ are the final similarity scores between two circRNAs and diseases, respectively.

2.5 Heterogeneous networks and node representation

A heterogeneous network is defined as a graph $G = (V, E, T_V)$ by combining $\boldsymbol{C}_{\mathrm{S}}$, $\boldsymbol{D}_{\mathrm{S}}$, and $\boldsymbol{A}$. V denotes the nodes, E denotes the edges, and T_V denotes the types of the nodes. To construct the metapath later, we select an unauthorized heterogeneous network. In the circRNA and disease similarity network, the nearest n_s $(= 5)$ neighbors of each node are selected and retained, and the remaining edges between the other neighbors are discarded. For each node in the heterogeneous network, the final goal of representation learning is to obtain their embedded vector $\boldsymbol{X} \in \mathbf{R}^{|V| \times k}, k \ll |V|$. $\boldsymbol{X}$ contains the structural relationships among them.

2.6 Metapath2vec++

Metapath2vec is a heterogeneous network representation learning algorithm[44]. The objective of metapath2vec is to maximize the network probability in consideration of multiple types of nodes and edges. Metapath2vec++ further distinguishes the types of nodes in the objective function and optimization process. Similar to deepwalk[51] and node2vec[52], metapath2vec++ is also based on word2vec's skip-gram model[53], which is used to predict the local neighboring nodes (background words) of the target node (given word). In contrast to deepwalk and node2vec, metapath2vec++ uses metapaths when generating node sequences.

2.6.1 Metapath based on random walk

In previous models, such as deepwalk[51] and node2vec[52], random walks and bias random algorithms are mainly used to generate node sequences. However, this study mainly investigates the representational relationship between heterogeneous entities. Thus, metapaths are used as the node sequence to be generated. A metapath scheme ρ is defined as a path that is denoted in the form of $V_1 \xrightarrow{R_1} V_2 \xrightarrow{R_2} \cdots V_t \xrightarrow{R_t} V_{t+1} \xrightarrow{R_{t+1}} \cdots V_l$, R denotes the relationship between two types of nodes.

Metapath2vec++ uses heterogeneous random walk to generate the paths of multiple types of nodes. At step i, the transition probability $\mathrm{tp}(v^{i+1}|v_t^i, \rho)$ represents the probability that the i-th node of the t-th type moves to the next point $i + 1$ on the metapath ρ. The calculation is defined as follows:

$$\mathrm{tp}(v^{i+1} \mid v_t^i, \rho) = \begin{cases} \dfrac{1}{|N_{t+1}(v_t^i)|}, & (v^{t+1}, v_t^i) \in E; \phi(v^{i+1}) = t+1; \\ 0, & (v^{t+1}, v_t^i) \in E; \phi(v^{i+1}) \neq t+1; \\ 0, & (v^{t+1}, v_t^i) \notin E \end{cases} \tag{8}$$

where $v_t^i \in V_t, N_{t+1}(v_t^i)$ denotes the $(t + 1)$-th type neighborhood of node v_t^i, and $\phi(v^{i+1})$ is the type of node v^{i+1}. In the standard metapath2vec++, the metapath is usually symmetrical with the same type of nodes at the beginning and end, and often only one path needs to be defined. However, the circRNA-diseae association investigated in this study is a heterogeneous relationship and the network scale is small, i.e., we can use a different metapath of length 5. According to the combination strategy, there are 32 metapaths in total. However, some of these paths are the inverse order of others, Thus, these paths can be deleted. For example, "circRNA-disease-circRNA-circRNA-circRNA" and "circRNA-circRNA-circRNA-disease-circRNA" can be regarded as the same, thus, "circRNA-disease-circRNA-circRNA-circRNA" can be deleted. In the end we used 20 kinds of metapath2vec++, as shown in Fig. 1.

2.6.2 Heterogeneous skip-gram model

Metapath2vec++ uses a heterogeneous skip-gram model to generate node vectors. The skip-gram model was originally used in word2vec[53] to predict the context (background) of a given center word. After being extended into the network, the local neighbors of a given node can be predicted. For a given node v and its local t-th type neighboring node b_t, the probability that they appear in the sequence window at the same time is $p(b_t|v)$. Our goal is to make the sum of such probabilities as large as possible. Metapath2vec++'s objective function is derived as follows:

$$\arg\max \sum_{v \in V} \sum_{t \in T_V} \sum_{b_t \in N_t(v)} \log p\,(b_t\,|\,v) \tag{9}$$

where N_t (v) denotes v's neighborhood with the t-th type of nodes. $p(b_t|v)$ is commonly defined as a softmax function and is adjusted to the specific type of node, i.e.,

$$p\,(b_t \mid v) = \frac{\mathrm{e}^{\boldsymbol{X}_{b_t} \boldsymbol{X}_v^{\mathrm{T}}}}{\sum\limits_{u_t \in V_t} \mathrm{e}^{\boldsymbol{X}_{u_t} \boldsymbol{X}_v^{\mathrm{T}}}} \tag{10}$$

where V_t denotes the node set of the t-th type nodes, $\boldsymbol{X}$ is the embedded feature matrix, and $\boldsymbol{X}_v$, $\boldsymbol{X}_{b_t}$, and $\boldsymbol{X}_{u_t}$ denote the v-th, b_t-th, and u_t-th row of $\boldsymbol{X}$, respectively.

When optimizing the objective function (i.e., Eq. (8)), because the number of nodes is often very large, the calculation cost of Eq. (9) will become abnormally large. Thus, there are two mainstream methods, one is to extract negative sampling method, another is hierarchical softmax. The core purpose of both methods is to reduce the computational size of V. According to Predictive Text Embedding (PTE)[54], the negative sampling method was applied. For events, node b_t is a local neighbor of node v, which is regarded as a mixture of two independent events. One is that b_t and v appear in the sequence window at the same time, and the other is that M noise nodes do not appear in the sequence window at the same time as node v. Therefore, we have the following objective:

$$O(\boldsymbol{X}) = \log\sigma(\boldsymbol{X}_{b_t}\boldsymbol{X}_v^{\mathrm{T}}) + \sum_{m=1}^{M} E_{u_t^m \sim P_t(u_t)}[\log\sigma(-\boldsymbol{X}_{u_t^m}\boldsymbol{X}_v^{\mathrm{T}})] \tag{11}$$

where $\sigma(x) = 1/(1 + \mathrm{e}^{-x})$ denotes the sigmoid function. $P_t(u_t)$ denotes the negative sampling distribution. $E_{u_t^m \sim P_t(u_t)}$ is an expectation when u_t^m obeys the distribution, where u_t^m is a node of the t-th type obtained by negative sampling. We observe that original calculation scale was V, now it is reduced to M, i.e., $M \ll V$. To maximize Eq. (11), one node can have two vectors, which are the vectors of the center node and neighboring nodes. We usually use the center node vector to represent the embedded features of nodes. Based on this, we can obtain the initial score (preliminary probability) of the circRNA-disease association $\boldsymbol{P} = \boldsymbol{X}_{\text{disease}}\boldsymbol{X}_{\text{circRNA}}^{\mathrm{T}}$. $\boldsymbol{X}_{\text{circRNA}}$ and $\boldsymbol{X}_{\text{disease}}$ correspond to the rows of circRNA and disease in matrix $\boldsymbol{X}$, respectively. Figure 1 shows an eventual representation of metapath2vec++, where each node v is encoded as e_v using one-hot, each embedded feature h from the $\boldsymbol{X}$ matrix is multiplied the vectors of other nodes, and then softmax calculation is performed. Finally, occurrence probability between each node and other nodes is determined. The relationship between circRNA c_3 and other circRNAs and diseases is illustrated in Fig. 1.

2.7 Matrix factorization

Because metapath2vec++ uses an unweighted graph when generating metapaths, the similarities between circRNAs and diseases are not well utilized. Following Wei and Liu[38], after deriving the initial scoring matrix, we continue to use matrix factorization for further optimization. The matrix factorization objective function in PCD_PVMF is formularized as follows:

$$\min_{C,D} \left\|\boldsymbol{P} - \boldsymbol{C}\boldsymbol{D}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} + \alpha\left\|\boldsymbol{C}\boldsymbol{D}^{\mathrm{T}}\right\|_{\mathrm{F}}^{2} + \beta(\mathrm{Tr}(\boldsymbol{C}^{\mathrm{T}}\boldsymbol{G}_C\boldsymbol{C}) + \mathrm{Tr}(\boldsymbol{D}^{\mathrm{T}}\boldsymbol{G}_D\boldsymbol{D})) \tag{12}$$

where matrixes $\boldsymbol{C}$ and $\boldsymbol{D}$ are factorization factors of $\boldsymbol{P}$ matrix, which can be expressed as feature matrices of circRNA and disease. $\boldsymbol{G}_C = \boldsymbol{I}_C - \boldsymbol{C}_S$ and $\boldsymbol{G}_D = \boldsymbol{I}_D - \boldsymbol{D}_S$ denote the graph Laplacian matrices for the circRNA and disease similarity matrices. $\boldsymbol{I}_C$ and $\boldsymbol{I}_D$ are two diagonal matrices, and the elements in $\boldsymbol{I}_C$ and $\boldsymbol{I}_D$ are row sums of $\boldsymbol{C}_S$ and $\boldsymbol{D}_S$, respectively. α and β are regularization coefficients. The PCD_MVMF solved the optimization problem by introducing Lagrange multipliers and Karush-Kuhn-Tucker (KKT) conditions[55]. The updating rules of matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ defined as follows:

$$\boldsymbol{C}_{ij} \leftarrow \boldsymbol{C}_{ij}\frac{(\boldsymbol{P}\boldsymbol{D} + \beta\boldsymbol{C}_S\boldsymbol{C})_{ij}}{((\alpha + 1)\boldsymbol{C}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{D} + \beta\boldsymbol{I}_C\boldsymbol{C})_{ij}},$$
$$\boldsymbol{D}_{ij} \leftarrow \boldsymbol{D}_{ij}\frac{(\boldsymbol{P}^{\mathrm{T}}\boldsymbol{C} + \beta\cdot\boldsymbol{D}_S\boldsymbol{D})_{ij}}{((\alpha + 1)\boldsymbol{D}\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} + \beta\cdot\boldsymbol{I}_D\boldsymbol{D})_{ij}} \tag{13}$$

At the beginning of the iteration, $\boldsymbol{C} = \boldsymbol{X}_{\text{circRNA}}$ and $\boldsymbol{D} = \boldsymbol{X}_{\text{disease}}$. Finally, the predicted circRNA-disease association result matrix is $\boldsymbol{P}^{*} = \boldsymbol{C}\boldsymbol{D}^{\mathrm{T}}$. The larger value of the element in $\boldsymbol{P}^{*}$, the higher the relevance between the corresponding circRNA and disease.

3 Result

3.1 Performance metrics

To evaluate the performance of our proposed computational method, several metrics are applied in this study. Two main metric methods are adopted to appraise the performance of our method. First, the Receiver Operating Characteristic (ROC) curve is drawn on the basis of the True Positive Rate (TPR) and False Positive Rate (FPR). Second, the precision, recall, and f-measure are applied to evaluate the performance of our proposed computational method. Precision is the ratio between the number of true positive samples that are predicted as true samples and the number of positive samples for prediction. Recall is the ratio between the number of true positive samples that are predicted as true samples and the total number of true positive samples. F-measure is the harmonic mean score of precision and recall, which is more reliable and valid.

3.2 Cross-validation

Each known circRNA-disease association will be regarded as a test data during the LOOCV process. Each disease related circRNA is remained as a test data in each turn. For example, there are 212 known circRNA-disease associations in this study. First, these 212 known circRNA-disease associations are regarded as the test data in turn. Therefore, 212 iterations are applied to each known circRNA-disease association. Afterward, we can determine the prediction scores between these known circRNAs and diseases. In addition, an extra iteration is adopted to predict the probability score of the remaining circRNA-disease associations. Then, all of the circRNA-disease associations are ranked in descending order. The higher the ranking of the known disease-related circRNAs, the better the performance of our proposed computational method. Based on the changing threshold, we can calculate different FPRs and TPRs, which are used to draw the ROC curve and calculate the corresponding Area Under the Curve (AUC) value. In order to validate the performance of PCD_MVMF, other prediction methods, KATZ[56], BiRW_avg[57], SIMCCDA[58], MRLDC[36], and NCPCDA[59] are compared with PCD_MVMF, as shown in Fig. 2. We set the window size (neighborhood size) to 5 in metapath2vec++. The dimension k of the embedded feature is 64. The number of walks of per node is 5. Batch size was set as 128. In matrix factorization, $\alpha = 0.002$ and $\beta = 0.001$. The parameters of other comparison methods are set according to the default parameters in their literature. As shown in Fig. 2, our method has better performance in LOOCV than other methods.

Afterward, we performed five-fold CV to test our method. We randomly divide all circRNA-disease associations into five equal parts, using four parts as the training set and one part as the test set. Finally, one-fifth of the associations of each test is spliced into a total prediction score matrix. Because each data segmentation is random, each method runs five-fold CV 10 times, and the final evaluation value is averaged. The results of the five-fold CV are shown in Fig. 3. The results show that our method still has good performance (AUC = 0.8539). Notably, because of the sparseness of the network, the AUC value decreases as a whole during the five-fold CV. Particularly for the network-dependent algorithms, such as BiRW_avg, the prediction results are nearly random.

This study also evaluated PCD_MVMF and other methods using precision, recall, and f-measure. For the circRNAs predicted with each disease, the precisions, recalls, and f-measures of the Top-k positions were calculated. Each value of each position is the average of 42 diseases as shown in Figs. 4 – 6. In predicting the circRNA-diseases from Top-5 to Top-50, the precision, recall, and f-measure curves of the PCD_MVMF are always above the curves of the other methods. Thus, PCD_MVMF is superior to other methods.

We also analyzed every single disease, i.e., each column in the predicted matrix. We calculated the AUC values for each disease prediction, and displayed distribution of 42 diseases in the form of a box diagram in Fig. 7. The mean and median of our method are the highest. At the same time, we detected 4 anomalies outside the 1.5 interquartile range (IQR). This finding indicates that only four diseases are not well predicted. The AUC value distributions of other algorithms are scattered, and not all diseases can be effectively predicted.

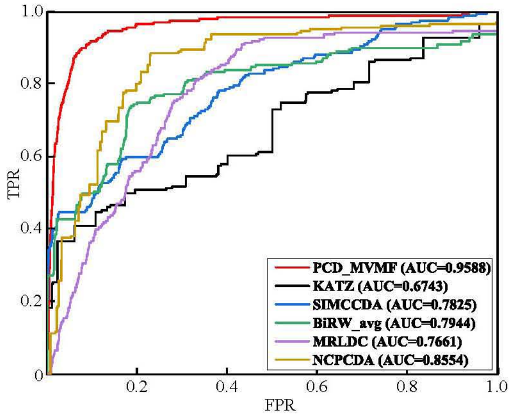

Fig. 2 ROC curves of PCD_MVFM and other computational methods based on LOOCV.

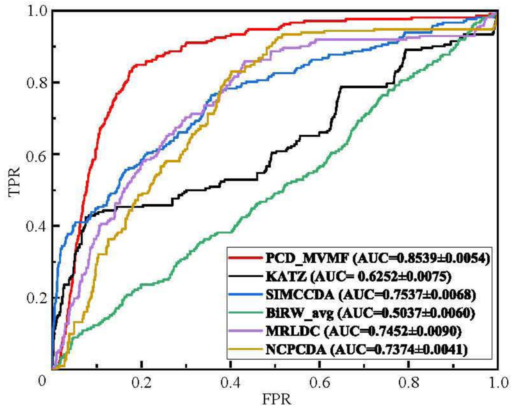

Fig. 3 ROC curves of PCD_MVFM and other computational methods based on five-fold CV.

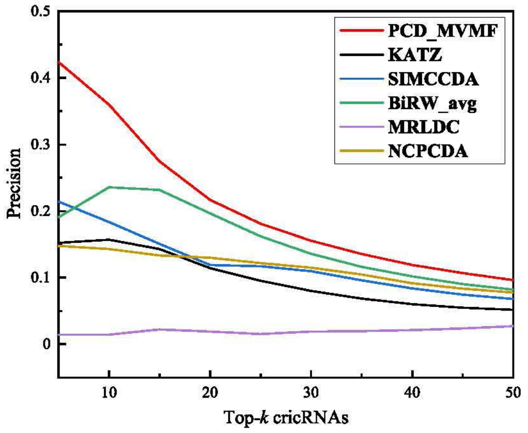

Fig. 4 Average precision of the test set at each Top-*k* position on querying diseases.

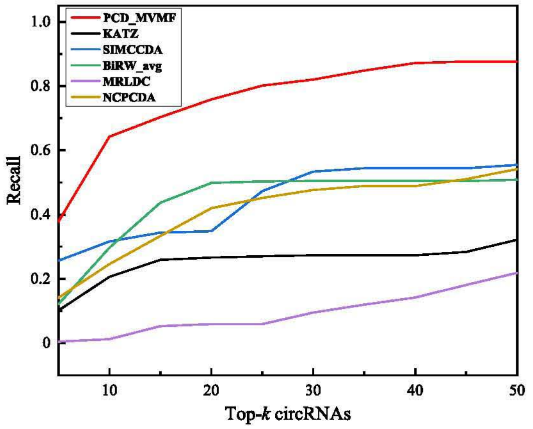

Fig. 5 Average recall of the test set at each Top-*k* position on querying diseases.

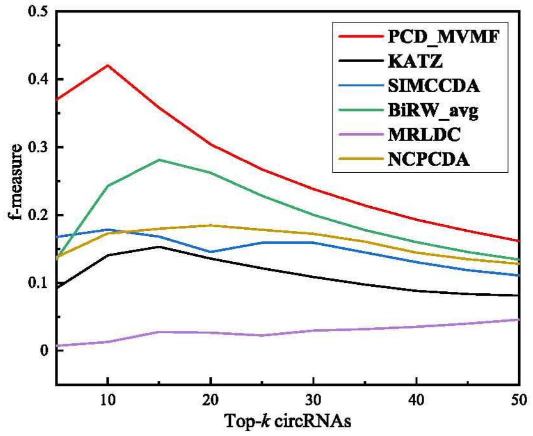

Fig. 6 Average f-measure of the test set at each Top-*k* position on querying diseases.

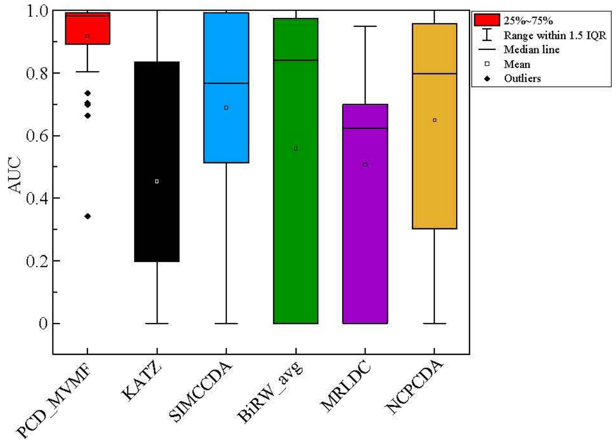

Fig. 7 AUC value distribution of PCD_MVMF and other algorithms in predicting circRNAs for each single disease.

3.3 Case studies

After measuring the performance of the method, we also analyzed the predicted disease-circRNAs. We mainly analyzed the predicted circRNAs of colorectal cancer and lung cancer. By screening 10 newly predicted colorectal cancer and lung cancer circRNAs seperately, we can identify potential research objects. These new relationship does not appear in circR2Disease. We verified their functions through a search of biological literature and reports.

As shown in the Table 1, 9 of the 10 predicted circRNAs are related to colorectal cancer. In the research of Guo et al.[60], hsa_circ_0000069 up-regulates and promotes cell proliferation, migration, and invasion in colorectal cancer. The circRNA hsa_circ_0000567 can be used as a promising diagnostic biomarker for human colorectal cancer[61]. CircRNA_001569 regulates colorectal cancer by targeting mir-145[27]. Xiong et al.[62] investigated the expression matrices in colorectal cancer and determined that multiple circRNAs are differentially expressed, such as hsa_circ_0001824, hsa_circ_0006174, hsa_circ_0008509, and hsa_circ_0007031.

Table 1 Top 10 new colorectal cancer-related candidate circRNAs.

Rank	circRNA name/ID	Evidence
1	hsa_circ_0000069	PMID: 28003761
2	hsa_circ_0000567	PMID: 29333615
3	hsa_circ_0000677/ hsa_circ_001569/circABCC	PMID: 27058418
4	hsa_circ_0001824	PMID: 28656150
5	circRNA_101419/hsa_circ_0032832	–
6	hsa_circ_0006174	PMID: 28656150
7	hsa_circ_0008509	PMID: 28656150
8	hsa_circ_001988/hsa_circ_0001451	PMID: 25624062
9	hsa_circ_0007031	PMID: 28656150
10	hsa_circ_0000504	PMID: 28656150

Note: PMID means PubMed unique identifier.

For lung cancer, we performed the same analysis, and the results are shown in Table 2. Five circRNAs of the 10 predicted lung cancer circRNAs are corroborated by the literature or database. Luo et al.[63] demonstrated new roles of hsa_circ_0000064 in lung cancer. And hsa_circ_0084927 sponges miR-93-5p to inhibit TGF-β signaling so that it affects lung cancer[64]. In addition, circRNA hsa_circ_100395 regulates the miR-1228/TCF21 pathway to inhibit lung cancer progression[65]. In the circFunBase, the differential expressions of hsa_circ_0000284 and hsa_circ_0001946 in lung cancer are also collected[66].

4 Conclusion

In this study, we developed a computational method, called PCD_MVMF, which is based on metapath2vec++ and matrix factorization algorithm, and applied it in a heterogeneous network.

First, various circRNA-related biological data including circRNA sequence data, circRNA target-gene-related GO terms, and functional data are extracted to compute the circRNA sequence, annotation, and functional similarity subnetworks. Disease semantic and related genes are adopted to construct disease semantic similarity and disease functional similarity subnetworks. After that, we construct a heterogeneous network and use metapath2vec++ to conduct representational learning of the network nodes. Finally, the initial scoring matrix is optimized by matrix factorization and the predicted results are obtained. To evaluate the performance of our proposed method, LOOCV, five-fold CV, precision, recall, and f-measure are adopted to test PCD_MVMF. By analyzing the predicted circRNA from colorectal cancer and lung cancer, we determine that circRNAs do have a role in these diseases. Notably, the method has a good predictive effect and practical value.

Table 2 Top 10 new lung cancer-related candidate circRNAs.

Rank	circRNA name/ID	Evidence
1	hsa_circ_0000064	PMID: 29223555
2	hsa_circRNA_104953/hsa_circ_0089310	–
3	hsa_circ_0002908	–
4	circRNA_0084927/hsa_circ_0084927	PMID: 31728016
5	hsa_circRNA_101720/hsa_circ_0002078	–
6	hsa_circRNA_100914/hsa_circ_0023903	–
7	hsa_circRNA_100782/circHIPK3/ hsa_circ_0000284	circFunBase
8	CDR1as/ciRS-7/hsa_circ_0001946	circFunBase
9	hsa_circRNA_100395/hsa_circ_0015278	PMID: 30176158
10	hsa_circ_0091017	–

Although, PCD_MVMF obtains better results than other state-of-the-art computational methods, some limitations could not be ignored. On one hand, metapath2vec++ needs to learn a large number of node sequences (metapaths). When the number of nodes is large, even if the algorithm has corresponding optimization measures, it takes a long time. On the other hand, metapath2vec++ is better at learning and representing network nodes, but not good at predicting, thus, it usually needs to cooperate with other prediction methods in our study. Moreover, the dataset used in this study is small and its predictive capability is limited. In the future, we will further optimize the algorithm while fusing and expanding the dataset.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Nos. 61972451, 61672334, and 61902230) and the Fundamental Research Funds for the Central Universities, Shaanxi Normal University (Nos.GK201901010 and 2018TS079).

References

[1] S. Qu, X. Yang, X. Li, J. Wang, Y. Gao, R. Shang, W. Sun, K. Dou, and H. Li, Circular RNA: A new star of noncoding RNAs, *Cancer Letters*, vol. 365, no. 2, pp. 141–148, 2015.

[2] J. Salzman, C. Gawad, P. Wang, N. Lacayo, and P. O. Brown, Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types, *PLoS One*, vol. 7, no. 2, p. e30733, 2012.

[3] L. Chen, and L. Yang, Regulation of circRNA biogenesis, *RNA Biology*, vol. 12, no. 4, pp. 381–388, 2015.

[4] H. L. Sanger, G. Klotz, D. Riesner, H. J. Gross, and A. K. Kleinschmidt, Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures, *Proceedings of the National Academy of Sciences of the United States of America*, vol. 73, no. 11, pp. 3852–3856, 1976.

[5] C. Cocquerelle, P. Daubersies, M. A. Majerus, J. P. Kerckaert, and B. Bailleul, Splicing with inverted order of exons occurs proximal to large introns, *EMBO Journal*, vol. 11, no. 3, pp. 1095–1098, 1992.

[6] F. A. Saad, L. Vitiello, L. Merlini, M. L. Mostacciuolo, S. Oliviero, and G. A. Danieli, A 3' consensus splice mutation in the human dystrophin gene detected by a screening for intra-exonic deletions, *Human Molecular Genetics*, vol. 1, no. 5, pp. 345–346, 1992.

[7] Y. Zhang, X. Zhang, T. Chen, J. F. Xiang, Q. F. 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.

[8] Z. Li, C. Huang, C. Bao, L. Chen, M. Lin, X. Wang, G. Zhong, B. Yu, W. Hu, L. Dai, et al., Exon-intron circular RNAs regulate transcription in the nucleus, *Nature Structure and Molecular Biology*, vol. 22, no. 3, pp. 256–264, 2015.

[9] J. E. Wilusz and P. A. Sharp, Molecular biology. A circuitous route to noncoding RNA, *Science*, vol. 340, no. 6131, pp. 440–441, 2013.

[10] E. Lasda and R. Parker, Circular RNAs: Diversity of form and function, *RNA*, vol. 20, no. 12, pp. 1829–1842, 2014.

[11] Y. Gao, J. Wang, and F. Zhao, CIRI: An efficient and unbiased algorithm for de novo circular RNA identification, *Genome Biology*, vol. 16, p. 4, 2015.

[12] T. B. Hansen, T. I. Jensen, B. H. Clausen, J. B. Bramsen, B. Finsen, C. K. Damgaard, and J. Kjems, Natural RNA circles function as efficient microRNA sponges, *Nature*, vol. 495, no. 7441, pp. 384–388, 2013.

[13] Q. Zheng, C. Bao, W. Guo, S. Li, J. Chen, B. Chen, Y. Luo, D. Lyu, Y. Li, G. Shi, et al., Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs, *Nature Communications*, vol. 7, p. 11215, 2016.

[14] K. Wang, B. Long, F. Liu, J. X. Wang, C. Y. Liu, B. Zhao, L. Y. Zhou, T. Sun, M. Wang, T. Yu, et al., A circular RNA protects the heart from pathological hypertrophy and heart failure by targeting miR-223, *European Heart Joutnal*, vol. 37, no. 33, pp. 2602–2611, 2016.

[15] M. Armakola, M. J. Higgins, M. D. Figley, S. J. Barmada, E. A. Scarborough, Z. Diaz, X. Fang, J. Shorter, N. J. Krogan, S. Finkbeiner, et al., Inhibition of RNA lariat debranching enzyme suppresses TDP-43 toxicity in ALS disease models, *Nature Genetics*, vol. 44, no. 12, pp. 1302–1309, 2012.

[16] C. Ragan, G. J. Goodall, and N. E. Shirokikh, Insights into the biogenesis and potential functions of exonic circular RNA, *Scientific reports*, vol. 9, no. 1, p. 2048, 2019.

[17] I. Legnini, G. Di Timoteo, F. Rossi, M. Morlando, F. Briganti, O. Sthandier, A. Fatica, T. Santini, A. Andronache, M. Wade, et al., Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis, *Molecular Cell*, vol. 66, no. 1, pp. 22–37, 2017.

[18] N. R. Pamudurti, O. Bartok, M. Jens, R. Ashwal-Fluss, C. Stottmeister, L. Ruhe, M. Hanan, E. Wyler, D. Perez-Hernandez, E. Ramberger, et al., Translation of CircRNAs, *Molecular Cell*, vol. 66, no. 1, pp. 9–21, 2017.

[19] J. Greene, A. M. Baird, L. Brady, M. Lim, S. G. Gray, R. McDermott, and S. P. Finn, Circular RNAs: Biogenesis, function and role in human diseases, *Frontiers in Molecular Biosciences*, vol. 4, p. 38, 2017.

[20] A. Rybak-Wolf, C. Stottmeister, P. Glazar, M. Jens, N. Pino, S. Giusti, M. Hanan, M. Behm, O. Bartok, R. Ashwal-Fluss, et al., Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed, *Molecular Cell*, vol. 58, no. 5, pp. 870–885, 2015.

[21] D. Barbagallo, A. Condorelli, M. Ragusa, L. Salito, M. Sammito, B. Banelli, R. Caltabiano, G. Barbagallo, A. Zappala, R. Battaglia, et al., Dysregulated miR-671-5p/CDR1-AS/CDR1/VSNL1 axis is involved in glioblastoma multiforme, *Oncotarget*, vol. 7, no. 4, pp. 4746–4759, 2016.

[22] J. Yao, S. Zhao, Q. Liu, M. Lv, D. Zhou, Z. Liao, and K. Nan, Over-expression of CircRNA_100876 in non-small cell lung cancer and its prognostic value, *Pathology Research and Practice*, vol. 213, no. 5, pp. 453–456, 2017.

[23] X. Zhu, X. Wang, S. Wei, Y. Chen, Y. Chen, X. Fan, S. Han, and G. Wu, hsa_circ_0013958: A circular RNA and potential novel biomarker for lung adenocarcinoma, *FEBS Journal*, vol. 284, no. 14, pp. 2170–2182, 2017.

[24] W. Sui, Z. Shi, W. Xue, M. Ou, Y. Zhu, J. Chen, H. Lin, F. Liu, and Y. Dai, Circular RNA and gene expression profiles in gastric cancer based on microarray chip technology, *Oncology Reports*, vol. 37, no. 3, pp. 1804–1814, 2017.

[25] J. Chen, Y. Li, Q. Zheng, C. Bao, J. He, B. Chen, D. Lyu, B. Zheng, Y. Xu, Z. Long, et al., Circular RNA profile identifies circPVT1 as a proliferative factor and prognostic marker in gastric cancer, *Cancer Letters*, vol. 388, pp. 208–219, 2017.

[26] P. Li, H. Chen, S. Chen, X. Mo, T. Li, B. Xiao, R. Yu, and J. Guo, Circular RNA 0000096 affects cell growth and migration in gastric cancer, *British Journal of Cancer*, vol. 116, no. 5, pp. 626–633, 2017.

[27] H. Xie, X. Ren, S. Xin, X. Lan, G. Lu, Y. Lin, S. Yang, Z. Zeng, W. Liao, Y. Q. Ding, et al., Emerging roles of circRNA_001569 targeting miR-145 in the proliferation and invasion of colorectal cancer, *Oncotarget*, vol. 7, no. 18, pp. 26 680–26 691, 2016.

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

[29] 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, p. 34985, 2016.

[30] S. Li, Y. Li, B. Chen, J. Zhao, S. Yu, Y. Tang, Q. Zheng, Y. Li, P. Wang, X. He, et al., exoRBase: A database of circRNA, lncRNA and mRNA in human blood exosomes, *Nucleic Acids Research*, vol. 46, no. D1, pp. D106–D112, 2018.

[31] Y. Liu, J. Li, C. Sun, E. Andrews, R. Chao, F. Lin, S. Weng, S. D. Hsu, C. Huang, C. Cheng, et al., circNet: A database of circular RNAs derived from transcriptome sequencing data, *Nucleic Acids Research*, vol. 44, no. D1, pp. D209–D215, 2016.

[32] 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, p. 283, 2013.

[33] C. Fan, X. Lei, Z. Fang, Q. Jiang, and F. X. Wu, circR2Disease: A manually curated database for experimentally supported circular RNAs associated with various diseases, doi: 10.1093/database/bay044.

[34] 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 and Disease*, vol. 9, no. 5, p. 475, 2018.

[35] D. Yao, L. Zhang, M. Zheng, X. Sun, and Y. Lu, circ2Disease: A manually curated database of experimentally validated circRNAs in human disease, *Scientific reports*, vol. 8, no. 1, p. 11018, 2018.

[36] Q. Xiao, J. Luo, and J. Dai, Computational prediction of human disease-associated circRNAs based on manifold regularization learning framework, *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2661–

2669, 2019.

[37] C. Yan, J. Wang, and F.-X. Wu, DWNN-RLS: Regularized least squares method for predicting circRNA-disease associations, *BMC Bioinformatics*, vol. 19, no. S19, p. 520, 2018.

[38] H. Wei and B. Liu, iCircDA-MF: Identification of circRNA-disease associations based on matrix factorization, *Briefings in Bioinformatics*, vol. 21, no. 4, pp. 1356–1367, 2020.

[39] L. Wang, Z.-H. You, Y.-M. Li, K. Zheng, and Y.-A. Huang, GCNCDA: A new method for predicting circRNA-disease associations based on graph convolutional network algorithm, *PLoS Computational Biology*, vol. 16, no. 5, p. e1007568, 2020.

[40] T. B. Mudiyanselage, X. Lei, N. Senanayake, Y. Zhang, and Y. Pan, Graph convolution networks using message passing and multi-dource dimilarity features for predicting circRNA-disease association, arXiv preprint arXiv: 2009.07173, 2020.

[41] X. Lei and Z. Fang, GBDTCDA: Predicting circRNA-disease associations based on gradient boosting decision tree with multiple biological data fusion, *International Journal of Biological Sciences*, vol. 15, no. 13, pp. 2911–2924, 2019.

[42] 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, p. 897, 2019.

[43] 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, p. 1042, 2020.

[44] Y. Dong, N. V. Chawla, and A. Swami, metapath2vec: Scalable representation learning for heterogeneous networks, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 135–144.

[45] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, et al., Human protein reference database—2009 update, *Nucleic Acids Research*, vol. 37, no. Database issue, pp. D767–D772, 2009.

[46] D. Lin, An Information-theoretic definition of similarity, in *Proceedings of the Fifteenth International Conference on Machine Learning*, Madison, WI, USA, 1998, pp. 296–304.

[47] P. J. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, et al., Biopython: Freely available python tools for computational molecular biology and bioinformatics, *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[48] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, et al., Disease ontology 2015 update: An expanded and updated database of human diseases for linking biomedical knowledge through disease data, *Nucleic Acids Research*, vol. 43, no. Database issue, pp. D1071–D1078, 2015.

[49] G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He, DOSE: An R/Bioconductor package for disease ontology semantic and enrichment analysis, *Bioinformatics*, vol. 31, no. 4, pp. 608–609, 2015.

[50] J. Piñero, À. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, J. García-García, F. Sanz, and L. I. Furlong, DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants, *Nucleic Acids Research*, vol. 45, no. D1, pp. D833–D839, 2017.

[51] B. Perozzi, R. Al-Rfou, and S. Skiena, Deepwalk: Online learning of social representations, in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2014, pp. 701–710.

[52] A. Grover and J. Leskovec, node2vec: Scalable feature learning for networks, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 855–864.

[53] T. Mikolov, K. Chen, G. Corrado, and J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.

[54] J. Tang, M. Qu, and Q. Mei, PTE: Predictive text embedding through large-scale heterogeneous text networks, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, Australia, 2015, pp. 1165–1174.

[55] F. Facchinei, C. Kanzow, and S. Sagratella, Solving quasi-variational inequalities via their KKT conditions, *Mathematical Programming*, vol. 144, nos. 1&2, pp. 369–412, 2014.

[56] C. Fan, X. Lei, and F.-X. 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.

[57] M. Xie, T. Hwang, and R. Kuang, Prioritizing disease genes by bi-random walk, in *Proceedings of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Kuala Lumpur, Malaysia, 2012, pp. 292–303.

[58] M. Li, M. Liu, Y. Bin, and J. Xia, Prediction of circRNA-disease associations based on inductive matrix completion, *BMC Medical Genomics*, vol. 13, no. Suppl 5, p. 42, 2020.

[59] 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. 33 222–33 228, 2019.

[60] J. Guo, J. Li, C. Zhu, W. Feng, J. Shao, L. Wan, M. Huang, and J. He, Comprehensive profile of differentially expressed circular RNAs reveals that hsa_circ_0000069 is upregulated and promotes cell proliferation, migration, and invasion in colorectal cancer, *OncoTargets and Therapy*, vol. 9, pp. 7451–7458, 2016.

[61] J. Wang, X. Li, L. Lu, L. He, H. Hu, and Z. Xu, Circular RNA hsa_circ_0000567 can be used as a promising diagnostic biomarker for human colorectal cancer, *Journal of Clinical Laboratory Analysis*, vol. 32, no. 5, p. e22379, 2018.

[62] W. Xiong, Y. Ai, Y. Li, Q. Ye, Z. Chen, J. Qin, Q. Liu, H. Wang, Y. Ju, W. Li, ,et al., Microarray analysis of circular RNA expression profile associated with 5-fluorouracil-based chemoradiation resistance in colorectal cancer cells,

Biomed Research International, vol. 2017, p. 8421614, 2017.

[63] Y. Luo, X. Zhu, K. Huang, Q. Zhang, Y. Fan, P. Yan, and J. Wen, Emerging roles of circular RNA hsa_circ_0000064 in the proliferation and metastasis of lung cancer, *Biomed Pharmacother*, vol. 96, pp. 892–898, 2017.

[64] W. Huang, Y. Yang, J. Wu, Y. Niu, Y. Yao, J. Zhang, X. Huang, S. Liang, R. Chen, S. Chen, et al., Circular RNA cESRP1 sensitises small cell lung cancer cells to chemotherapy by sponging miR-93-5p to inhibit TGF-β signalling, *Cell Death Differ*, vol. 27, no. 5, pp. 1709–1727, 2020.

[65] D. Chen, W. Ma, Z. Ke, and F. Xie, circRNA hsa_circ_100395 regulates miR-1228/TCF21 pathway to inhibit lung cancer progression, *Cell Cycle*, vol. 17, no. 16, pp. 2080–2090, 2018.

[66] X. Meng, D. Hu, P. Zhang, Q. Chen, and M. Chen, circFunBase: A database for functional circular RNAs, doi: 10.1093/database/baz003.

Yuchen Zhang received the BS degree from Xi'an Technological University, Xi'an, China in 2015. He is currently taking a successive postgraduate and doctoral program in the School of Computer Science, Shaanxi Normal University, Xi'an, China. His current research interests include bioinformatics and intelligent computing, as well as deep learning.

Xiujuan Lei received the MS and PhD degrees from Northwestern Polytechnical University, Xi'an, China in 2001 and 2005, respectively. She is currently a professor with the School of Computer Science, Shaanxi Normal University, Xi'an, China. Her research interests include bioinformatics, swarm intelligent optimization, data mining, and deep learning.

Zengqiang Fang received the BS degree from Xi'an University of Technology, Xi'an, China in 2017. He is currently pursuing the MS degree in Shaanxi Normal University, Xi'an, China. His research interests include machine learning, bioinformatics, and data mining.

Yi Pan received the BEng and MEng degrees in computer engineering from Tsinghua University, China in 1982 and 1984, respectively, and the PhD degree in computer science from the University of Pittsburgh, USA in 1991. His profile has been featured as a distinguished alumnus in both Tsinghua alumni newsletter and University of Pittsburgh CS alumni newsletter.

He is currently a Regents' professor and has served as the chair of Computer Science Department at Georgia State University since January 2006. He has also served as an interim associate dean and the chair of Biology Department during 2013–2017. He joined Georgia State University in 2000 and was promoted to full professor in 2004, named a distinguished university professor in 2013, and designated a Regents' professor (the highest recognition given to a faculty member by the University System of Georgia) in 2015. He has published more than 450 papers, including over 250 journal papers with more than 100 papers published in IEEE/ACM transactions/journals. In addition, he has edited/authored 43 books. His work has been cited more than 14 100 times based on Google Scholar and his current h-index is 72. He has served as an editor-in-chief or editorial board member for 20 journals including 7 IEEE transactions. Currently, he is serving as an associate editor-in-chief of *IEEE/ACM Transactions on Computational Biology and Bioinformatics*. He is the recipient of many awards, including one IEEE Transactions Best Paper Award, five Best Paper Awards of IEEE and other international conference or journal, 4 IBM Faculty Awards, 2 JSPS Senior Invitation Fellowships, IEEE BIBE Outstanding Achievement Award, IEEE Outstanding Leadership Award, NSF Research Opportunity Award, and AFOSR Summer Faculty Research Fellowship. He has organized numerous international conferences and delivered keynote speeches at over 60 international conferences around the world.

His current research interests mainly include bioinformatics and health informatics using big data analytics, cloud computing, and machine learning technologies.