Briefings in Bioinformatics, 22(6), 2021, 1–27

https://doi.org/10.1093/bib/bbab286
Review

Circular RNAs and complex diseases: from experimental results to computational models

Chun-Chun Wang, Chen-Di Han, Qi Zhao and Xing Chen

Corresponding authors: Qi Zhao, School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Anshan 114051, China. E-mail: zhaoqi@lnu.edu.cn; Xing Chen, Artificial Intelligence Research Institute, China University of Mining and Technology, Xuzhou 221116, China. E-mail: xingchen@amss.ac.cn

Abstract

Circular RNAs (circRNAs) are a class of single-stranded, covalently closed RNA molecules with a variety of biological functions. Studies have shown that circRNAs are involved in a variety of biological processes and play an important role in the development of various complex diseases, so the identification of circRNA-disease associations would contribute to the diagnosis and treatment of diseases. In this review, we summarize the discovery, classifications and functions of circRNAs and introduce four important diseases associated with circRNAs. Then, we list some significant and publicly accessible databases containing comprehensive annotation resources of circRNAs and experimentally validated circRNA-disease associations. Next, we introduce some state-of-the-art computational models for predicting novel circRNA-disease associations and divide them into two categories, namely network algorithm-based and machine learning-based models. Subsequently, several evaluation methods of prediction performance of these computational models are summarized. Finally, we analyze the advantages and disadvantages of different types of computational models and provide some suggestions to promote the development of circRNA-disease association identification from the perspective of the construction of new computational models and the accumulation of circRNA-related data.

Key words: circRNA; disease; circRNA-disease association prediction; network algorithm; machine learning; computational model

CircRNA

Circular RNAs (circRNAs) are a class of single-stranded, covalently closed RNA molecules, which are produced by backsplicing from pre-mRNAs [1]. During backsplicing, a downstream splice-acceptor site is covalently connected to an upstream splice-donor site [1]. The first circRNA molecules, viroids, were identified more than 40 years ago [2, 3]. Soon after, Hsu *et al*. [4] discovered circRNAs in the cytoplasmic fractions of eukaryotic cell lines through electron microscopy. Furthermore, circRNAs were identified to be produced from self-splicing introns of pre-ribosomal RNA in unicellular eukaryotes [5]. Later, researcher discovered that a small part of circRNAs stem from protein-coding genes in archaea [6]. However, circRNAs were initially treated as 'junk' yielded by splicing errors [7].

As the development of high-throughput RNA sequencing technology and new bioinformatics algorithms, more and more

Chun-Chun Wang is a PhD student of the School of Information and Control Engineering, China University of Mining and Technology. His research interests include bioinformatics, complex network algorithm and machine learning.
Chen-Di Han is a master student of the School of Information and Control Engineering, China University of Mining and Technology. Her research interests include bioinformatics and deep learning.
Qi Zhao, PhD, is a Professor of the School of Computer Science and Software Engineering, University of Science and Technology Liaoning. His research interests include bioinformatics, complex network and machine learning.
Xing Chen, PhD, is a Professor of the China University of Mining and Technology. He is the Associate Dean of Artificial Intelligence Research Institute, China University of Mining and Technology. He is also the Founding Director of the Institute of Bioinformatics, China University of Mining and Technology and Big Data Research Center, China University of Mining and Technology. His research interests include complex disease-related non-coding RNA biomarker prediction, computational models for drug discovery and early detection of human complex disease based on big data and artificial intelligence algorithms.
Submitted: 1 June 2021; **Received (in revised form):** 23 June 2021

circRNAs were discovered in eukaryotes including protists, fungi, plants, insects and mammals [8–13]. CircRNAs are a relatively large family of RNAs and massive circRNAs have been identified, but studies on the classification of circRNAs and the mechanism of loop formation have just begun. CircRNAs mainly include exonic circRNAs (ecircRNAs), exon-intron circRNAs (EIciRNAs) and circular intronic RNAs (ciRNAs) [14]. Among them, ecircRNAs are produced by the exons in the back-splicing process of pre-mRNA, which are abundant in the cytoplasm [15]. The EIciRNAs are widely present in the nucleus, which are formed by the combined action of exons and introns during the back-splicing process [16]. In addition, ciRNAs are formed by introns and are mainly localized in the nucleus [17]. Besides, circRNAs could be generated from more than 10% of expressed gene in the investigated cells and tissues [18, 19]. It can be learned that the expression of circRNAs is broad. Usually, the expression level of circRNA is low [20, 21], but some circRNAs are experimentally verified to be high expressed in specific type of cells or tissues [15, 22]. Moreover, thousands of circRNAs are abundant in the mammalian brain and some of them are upregulated during neurogenesis [23]. These studies demonstrate that circRNAs should not be'junk' and they may have specifically biological functions.

CircRNA function

CircRNAs are usually expressed in only a few cell types, exhibiting significant specificity during tissue and developmental stages. However, some other circRNAs show cross-species conservation [18]. In addition, by comparison with linear exons, the exon sequence of circRNA appears to be more conserved at the third codon position, while the third codon is meaningless at the protein level [21]. These indicate that in addition to encoding proteins, circRNA has other functions.

CircRNAs as microRNA sponges

In 2013, Hansen *et al.* [24] found that hsa_circRNA_105055 has more than 70 miR-7 binding sites. Further functional studies have showed that ciRS-7 strongly restrains the activity of miR-7, which in turn leads to an increase in the target level of miR-7. They also demonstrated that hsa_circRNA_105055 and miR-7 have overlapping co-expression in mouse brain tissue [24]. In addition, the sex-determining region Y (Sry)^9 of hsa_circRNA_105055 has 16 microRNA (miRNA)-138 binding sites [24]. Moreover, researchers have demonstrated that circ-HIPK3, circ-ITCH and mm9-circ-012559 can act as miRNA sponges [25–27]. The above findings indicate that circRNA is very common as miRNA sponge.

CircRNAs regulate the expression of parental genes

Different types of circRNAs have different ways of regulating their parental genes. Specifically, ciRNAs promote transcription of genes by binding to Pol II. Zhang *et al.* [17] found that knocking out ciRNA can suppress the expression of its parental gene. For the specific ciRNA ci-ankrd52, it aggregates into the transcriptional site and acts as a positive regulator of Pol II transcription. For EIciRNA, it binds to U1 snRNP to form EIciRNA-U1 snRNP complexes, which further binds to Pol II, thereby promoting transcription of the parental gene [17]. Besides, Li *et al.* [16] found that EIciRNAs can regulate gene expression in the nucleus, which mainly enhances the expression of the parental gene in *cis* and affects transcriptional regulation through the interaction between U1 snRNA and EIciRNA. In addition, ecircRNA, containing miRNA response elements, can bind to miRNA and indirectly regulate the expression of its parent mRNA. Li *et al.* [28] found that hsa_circRNA_001141 binds to miR-7 and miR-214 in lung cancer cells and enhances the expression of ITCH, thereby inhibiting the activity of Wnt/β-catenin.

Competition with pre-mRNA splicing

The pre-mRNA can undergo typical linear splicing to produce mRNA during processing, while nonlinear splicing generates circRNA. Recent studies have found that increasing the efficiency of linear splicing can significantly reduce the abundance of circRNA [29]. When the length of the introns flanking the circRNA is longer, the efficiency of typical linear splicing is reduced, while the efficiency of cyclization is increased [30]. The above findings indicate that circRNA can compete with the pre-mRNA during transcription.

CircRNA-disease associations

Previous functional analysis of circRNAs has demonstrated that a circRNA, hsa_circRNA_105055, contains more than 70 miRNA target sites and can act as a miRNA sponge [24]. Besides, some studies have indicated that circRNAs can regulate protein functions [16, 31]. As biological functions of circRNAs were discovered, circRNAs are receiving the attention of researchers. In the field of human health, more and more studies have shown that circRNAs have close associations with human complex diseases [32–34]. In the following, we will introduce several common cancers and their associated circRNAs.

Gastric cancer

Gastric cancer, one of the top five cancers in the world [35]. In 2019, 27 510 patients were newly diagnosed with gastric cancer and 11 140 patients died because of gastric cancer in the USA [36]. Therefore, it is necessary to discover and explore pathogenesis for the early diagnosis, prevention and treatment of gastric cancer. So far, increasing experiments have shown that circRNAs play an irreplaceable role in the development of gastric cancer [37]. Li *et al.* [38] found that there were 343 differentially expressed (DE) circRNAs by comparing the gastric cancer patients' plasma and plasma of healthy control, and then, the two techniques of reverse-transcription real-time polymerase chain reaction (RT-PCR) [39] and RT-droplet digital PCR (RT-ddPCR) [40] were used to determine the expression level of circRNAs. More concretely, patients with low expression levels of hsa_circ_0001017 or hsa_circ_0061276 in plasma have a shorter overall survival than patients with higher expression levels [38]. In addition, circRNA-0026 regulates RNA transcription, RNA metabolism and gene expression in gastric cancer [41]. Moreover, biological studies have found that knocking out hsa_circ_0047905, hsa_circ_0138960 and has-circRNA7690–15 in gastric cancer cells down-regulates the expression of the parental gene [42]. Inhibition of the expression of these three circRNAs can inhibit the proliferation and invasion of gastric cancer cells [42].

Breast cancer

Breast cancer is one of the major cancer types among women worldwide, and 12% of women are diagnosed with breast cancer during their lifetime in the USA [43]. Common symptoms of breast cancer include: a lump in the breast, a change in breast shape and red or scaly skin. CircRNAs are closely related to the formation and development of breast cancer, and recent studies have found that the expression of some circRNAs can be used to

prevent breast cancer [44–46]. For example, hsa_circ_0001982 in breast cancer tissues inhibits breast cancer cell proliferation and induces apoptosis by targeting miR-143 [44]. In addition, knocking out hsa_circRNA_005239 can inhibit the proliferation and promote the apoptosis in triple negative breast cancer [46]. There are also some circRNAs that can be used as potential biomarkers for breast cancer detection. For example, Yin *et al.* [45] found that the expression level of hsa_circ_0001785 in plasma of breast cancer patients is significantly different from that in preoperative, postoperative and healthy individuals, which demonstrates that hsa_circ_0001785 can act as a diagnostic biomarker for breast cancer.

Lung cancer

Lung cancer is characterized by uncontrolled growth of cells in the lung tissue. It is reported that 85% of lung cancer is caused by long-term smoking [47]. Other factors that cause lung cancer include genetic factors, secondhand smoke or air pollution [48, 49]. The circRNA of hsa_circRNA_001141 in lung cancer tissues has been shown to suppress the development of lung cancer by enhancing the expression of its parental gene ITCH [28], while hsa_circ_0013958 in lung cancer cells can promote the proliferation of lung cancer cells and inhibit apoptosis [50]. Besides, Yao *et al.* [51] found that circRNA_100876 is abnormally expressed in non-small cell lung cancer. In addition, the higher the expression level of circRNA_100876, the lower the survival rate [51]. Therefore, circRNA_100876 can be used as biomarker for early detection and screening of lung cancer.

Pancreatic cancer

Pancreatic cancer is usually caused by uncontrolled growth, division and spread of cells in the pancreas [52]. Symptoms usually manifest as digestive problems including: weight loss, indigestion, back pain, nausea and so on [53]. Studies have found that smoking or lack of exercise and long-term heavy drinking may lead to chronic pancreatitis [54]. Guo *et al.* [55] demonstrated the dysregulation of circRNA expression in pancreatic cancer tissues using qRT-PCR. In addition, they predicted that multiple circRNAs have complementary sequences to miR-15a / miR-506 and different miRNA binding sites in the seed region [55]. Furthermore, Chen *et al.* [56] found that circRNA_100782 regulates the proliferation of BxPC3 pancreatic cancer cells by interacting with miR-124.

There is increasing evidence that circRNAs are related with the development and invasion of complex diseases, although most of the action mechanisms are still unknown [57]. Besides, circRNAs could be novel biomarkers for human cancers [58]. Therefore, identifying associations between circRNAs and diseases would facilitate the diagnosis, prevention and prognosis of human complex diseases.

Databases

Data collection about circRNAs, diseases and circRNA-disease associations is an important premise when researchers identify novel circRNA-disease associations by bioinformatics methods. In addition, the systematic collection and management of the information about circRNAs and circRNA-disease relationships is important for further inspection of the underlying molecular mechanism of circRNAs. In this section, we introduce some important databases, from which researchers could obtain circRNA related data more conveniently. These databases can be divided into two categories. Specifically, the first type of databases record circRNA-disease associations (see Table 1). The second type of databases provide comprehensive annotation resources for circRNAs (see Table 2). More detailed introduction of these databases can be seen from Supplementary Materials available online at https://academic.oup.com/bib.

Computational models

As the development of high-throughput sequencing technology and bioinformatics analysis methods, more and more circRNAs are identified. However, the function and mechanism of circRNAs are unclear in most cases. In addition, researchers discover that the occurrence and development of various diseases including cancer are associated with circRNAs. Identifying and studying circRNA-disease associations is important for understanding the function and molecular mechanism of circRNAs. In addition, circRNA-disease association identification is meaningful for the early detection, early diagnosis and effective treatment of diseases. However, it is time-consuming and laborious to discover novel circRNA-disease relationships directly by biological experiments. Computational models could effectively predict potential circRNA-disease associations for further experimental verification, which would save many resources.

During recent years, scientists have successively proposed some computational models for predicting potential circRNA-disease associations based on distinct algorithms. These computational models can be roughly divided into two categories, namely network algorithm-based models and machine learning-based models (see Table 3). In this section, we mainly introduce the general steps of construction of different models and the main advantages or limitations of these models. The main symbols utilized throughout this sections are listed in Table 4.

Network algorithm-based models

In network algorithm-based models, circRNA similarity network, disease similarity network and circRNA-disease association network are usually utilized to construct a heterogeneous network. Then, the corresponding algorithm is used to predict potential relationships based on the heterogeneous network.

PWCDA

Lei *et al.* [77] developed the model of Path Weighed method for predicting CircRNA-Disease Associations (PWCDA) (see Figure 1). The same model has been used for potential miRNA-disease association prediction before [78]. They first construct a heterogeneous network, which is composed of circRNA similarity network, disease similarity network and circRNA-disease association network. Then, PWCDA searches all the paths between circRNA c_i and disease d_j with the length less than η by depth-first search (DFS) algorithm. The path set can be described as $\{p_1, p_2, \ldots, p_k, \ldots, p_{m_{i,j}}\}$, where the variable $m_{i,j}$ denotes the number of searched paths between circRNA c_i and disease d_j. Finally, the predicted score between c_i and d_j can be calculated by accumulating all contributing scores (CS) of paths in $\{p_1, p_2, \ldots, p_k, \ldots, p_{m_{i,j}}\}$. The $CS(p_k)$ of the path $p_k = \{e_{k_1}, e_{k_2}, \ldots, e_{k_n}\}$ is defined as follows:

$$CS\left(p_k\right) = \left(\prod_{t=1}^{n} W_{e_{k_t}}\right)^{\alpha \times \exp(\mathrm{len}(p_k))} \tag{1}$$

Table 1. Databases recording circRNA-disease associations

Database	Number of circRNAs	Number of diseases	Number of associations	URL
Circ2Traits [59]	1951	105	Unknown	http://gyanxet-beta.com/circdb/
Circ2Disease [60]	237	54	273	http://bioinformatics.zju.edu.cn/Circ2Disease/index.html
CircR2Disease [61]	661	100	725	http://bioinfo.snnu.edu.cn/CircR2Disease/
CircRNADisease [33]	330	48	354	http://cgga.org.cn:9091/circRNADisease/
Circad [62]	1338	720	1338	http://clingen.igib.res.in/circad/

Table 2. Databases providing annotation resources for circRNAs

Database	Number of circRNAs	Short description	URL
circBase [63]	92 375	Provides information of circRNAs including the genomic position, gene symbols, evidence for the occurrence	http://www.circbase.org/
CircNet [64]	34 000	Provides the information of circRNA expression profiles, circRNA-miRNA sponge regulatory network, circRNA-gene-miRNA regulatory network	http://circnet.mbc.nctu.edu.tw/
deepBase v2.0 [65]	14 867	Provides comprehensive expression and evolution profiles of circRNAs.	http://biocenter.sysu.edu.cn/deepBase/
circRNADb [66]	32 914	Provides the information of protein-coding potential of circRNAs	http://reprod.njmu.edu.cn/circrnadb
TSCD [67]	302 853	Provides the genomic location and conservation of tissue specific circRNAs	http://gb.whu.edu.cn/TSCD
CSCD [68]	272 152	Records the function and regulation of cancer-associated circRNAs	http://gb.whu.edu.cn/CSCD
CIRCpedia v2 [69]	262 782	Records the information of location, strand, isoform, expression value, sequencing type and conservation of circRNAs	https://www.picb.ac.cn/rnomics/circpedia/
exoRBase [70]	58 330	Provides circRNA expression profile, expression rank, gene symbol and spliced length	http://www.exorbase.org/
CircFunBase [71]	7059	Provides the information of circRNA function, GO annotations and circRNA-associated miRNAs	http://bis.zju.edu.cn/CircFunBase/
TRCirc [72]	92 375	Contains more than 765 000 transcription factor-circRNA relationships	http://www.licpathway.net/TRCirc
circbank [73]	140 790	Develops a new naming system based on the host genes of circRNAs	http://www.circbank.cn/
CircRiC [74]	92 589	Provides the modules of integrative analysis, drug response, biogenesis, and expression landscape	https://hanlab.uth.edu/cRic/
MiOncoCirc [75]	227 056	Records circRNAs from metastases, primary tumors, and very rare cancer types	https://nguyenjoshvo.github.io/
VirusCircBase [76]	11 924	Provides the information of the location, genes involved in the viral circRNA, the abundance, the detection method	http://www.computationalbiology.cn/ViruscircBase/home.html

where $W_{e_{k_t}}$ is the weight of the edge e_{k_t} in the path p_k. Besides, α is a constraint factor and len(p_k) denotes the length of p_k. The decaying function $\alpha \times \exp(\text{len}(p_k))$ is used to further reduce the CS of long paths. Then, the final association score between c_i and d_j is defined as follows:

$$AS(c_i, d_j) = \sum_{k=1}^{m_{i,j}} CS(p_k) \quad (2)$$

In PCWDA, only paths within three steps are used to decrease the noisy information. However, the decaying function in PCWDA is relatively simple.

BRWSP

Lei *et al.* [79] proposed a computational model (see Figure 2) of Biased Random Walk to Search Paths on a multiple heterogeneous network (BRWSP) to predict circRNA-disease associations. Firstly, they construct the multi-layer heterogeneous network by

Table 3. List of different types of circRNA-disease prediction models

Model name	Core algorithm	Model type	Source code
PCWCDA	DFS algorithm	Network algorithm-based model	Unavailable
BRWSP	Biased random walk algorithm	Network algorithm-based model	Unavailable
KATZHCDA	KATZ	Network algorithm-based model	Unavailable
KATZCPDA	KATZ	Network algorithm-based model	Unavailable
IBNPKATZ	Bipartite network projection algorithm and KATZ	Network algorithm-based model	Unavailable
NCPCDA	Network consistency projection	Network algorithm-based model	Unavailable
DWNCPCDA	DeepWalk and network consistency projection	Network algorithm-based model	Unavailable
LLCDC	LLC and label propagation algorithm	Network algorithm-based model	Unavailable
CD-LNLP	Label propagation algorithm	Network algorithm-based model	Unavailable
DWNN-RLS	Regularized least squares of kronecker product kernel	The first type of machine learning-based model	Unavailable
RWRLCDA	Random work and logistic regression	The first type of machine learning-based model	Unavailable
MRLDC	Manifold regularization-learning	The first type of machine learning-based model	Unavailable
iCircDA-MF	Matrix factorization	The first type of machine learning-based model	Unavailable
GMCDA	Graph-based multi-label learning	The first type of machine learning-based model	Unavailable
iCDA-CMG	Collective Matrix completion	The first type of machine learning-based model	Unavailable
NMFIBAC	Non-negative matrix factorization	The first type of machine learning-based model	Unavailable
SIMCCDA	Speedup inductive matrix completion	The first type of machine learning-based model	https://github.com/bioinformaticsAHU/SIMCCDA
PreCDA	PersonalRank algorithm	The first type of machine learning-based model	https://github.com/wythit/PreCDA
ICFCDA	Collaboration filtering	The first type of machine learning-based model	Unavailable
RWRKNN	Random walk with restart and KNN	The second type of machine learning-based model	Unavailable
iCDA-CGR	SVM	The second type of machine learning-based model	Unavailable
GBDTCDA	GBDT	The second type of machine learning-based model	Unavailable
DFPUCDA	DF	The second type of machine learning-based model	https://github.com/xzenglab/DeepDCR
CNNCDA	CNN	The second type of machine learning-based model	Unavailable
GCNCDA	Graph Convolutional Network	The second type of machine learning-based model	Unavailable
AE-DNN	Autoencoder and DNN	The second type of machine learning-based model	Unavailable
AE-RF	Autoencoder and RF	The second type of machine learning-based model	https://github.com/Deepthi-K523/AE-RF

utilizing the information of circRNA similarity matrix CS, disease similarity matrix DS, gene similarity matrix GS, circRNA-disease association matrix CD, circRNA-gene interaction matrix CG as well as gene-disease association matrix GD. The heterogeneous network is represented as follows:

$$A^* = \begin{pmatrix} CS & CG & CD \\ CG^T & GS & GD \\ CD^T & GD^T & DS \end{pmatrix} \quad (3)$$

To avoid the biases caused by larger values in A^*, a normalized multi-layer heterogeneous network denoted by $NMH = D^{-(1/2)}A^*D^{-(1/2)}$ is further established, where D is the degree matrix of A^*.

Secondly, a biased random walk algorithm is employed to search paths between circRNAs and diseases in the heterogeneous network. Specifically, the random walker starts from the investigated circRNA node u and first randomly moves to one neighbor of u. Then, the walker continues to walk to the next node. Here, c_k is employed to denote the node accessed by the

Table 4. The main symbols utilized throughout the Computational models section

Symbol	Definition and description
A^*	Adjacency matrix of heterogeneous network
W_{cg}	Weight matrix of circRNA graph
W_{dg}	Weight matrix of disease graph
L_c	Laplacian matrix of circRNA graph
L_d	Laplacian matrix of disease graph
CS	CircRNA Similarity matrix
CSS	CircRNA Semantic Similarity matrix
CFS	CircRNA Functional Similarity matrix
CES	CircRNA Expression Similarity matrix
CTS	CircRNA Topological Similarity matrix
RCS	Reconstructed CircRNA Similarity matrix
DS	Disease Similarity matrix
DSS	Disease Semantic Similarity matrix
DTS	Disease Topological Similarity matrix
RDS	Reconstructed Disease Similarity matrix
KC	GIP Kernel similarity matrix of CircRNA
KD	GIP Kernel similarity matrix of Disease
GS	Gene Similarity matrix
CD	CircRNA-Disease association matrix
CG	CircRNA-Gene interaction matrix
GD	Gene-Disease association matrix
AS	The predicted circRNA-disease Association Score matrix
c_i	CircRNA *i*
d_j	Disease *j*
N_c	The number of circRNAs
N_d	The number of diseases
$N(c_i)$	The neighbors of c_i
$N(d_j)$	The neighbors of d_j

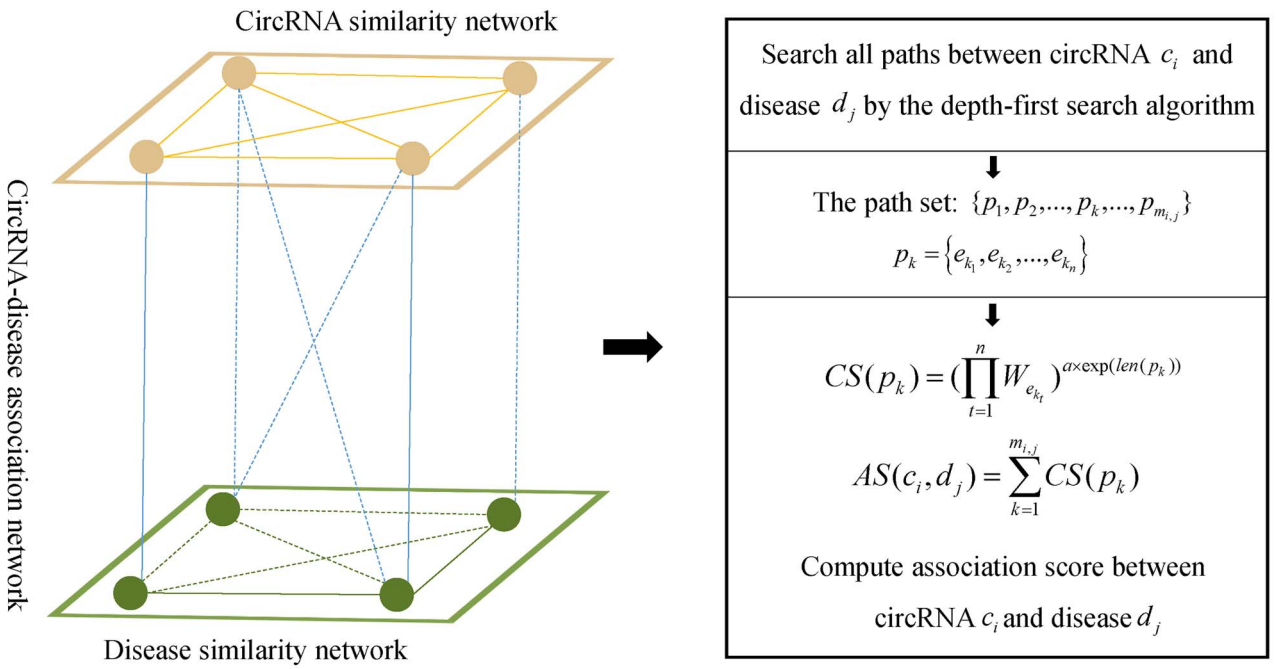

Figure 1. The workflow of PWCDA to infer potential circRNA-disease associations based on DFS algorithm to search paths on a heterogeneous network.

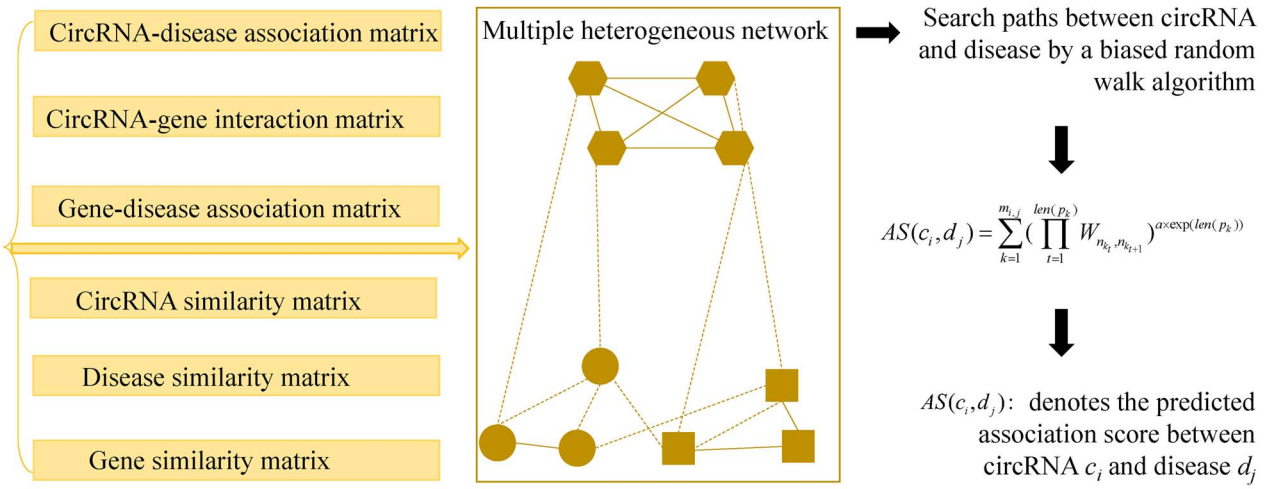

Figure 2. The flowchart of BRWSP to predict circRNA-disease associations based on biased random walk to search paths on a multiple heterogeneous network.

walker on its kth move. The strategy of selecting the next node is described as follows:

$$P(c_{k+1}=x|c_k=v,c_{k-1}=t)=\begin{cases}\dfrac{\Phi(t,v,x)*\mathrm{NMH}(v,x)}{\sum_{i\in\mathrm{Nei}(v)}\mathrm{NMH}(v,i)}, & \text{if } x\in\mathrm{Nei}(v) \text{ and } x\notin\{c_0,c_1,\ldots,c_k\}\\ 0, & \text{otherwise}\end{cases} \tag{4}$$

$$\Phi(t,v,x)=\begin{cases}q, & \text{if } x\in\mathrm{Nei}(v) \text{ and } x\in\mathrm{Nei}(t)\\ 1-q, & \text{otherwise}\end{cases} \tag{5}$$

where $P(c_{k+1}=x|c_k=v,c_{k-1}=t)$ represents the transition probability from the current node v to the next node x when the last visited node is the node t. Besides, Nei(v) and Nei(t) denote the neighbors of the current node v and the last visited node t in the heterogeneous network, respectively. For the parameter q, if q is assigned a larger value, the biased random walk algorithm tends to select the nodes near the investigated node. Otherwise, the biased random walk algorithm tends to select the nodes away from the investigated node. It can be seen from Eq. (4) that the next accessed node will be chosen from the neighbors of the current nodes based on their probability. The random walker keeps moving until the investigated disease node is accessed. $p_k=\{n_{k_1},n_{k_2},\ldots n_{k_i},\ldots,n_{k_{L+1}}\}$ is used to denote one path between circRNA n_{k_1} and disease $n_{k_{L+1}}$, where n_{k_i} represents the node (circRNA, disease or gene) of p_k and L is the length of p_k. To search more paths between investigated circRNA and disease, the above process will be repeated. Only paths with lengths less than L will be left. The set $\{p_1,p_2,\ldots,p_{m_{i,j}}\}$ is utilized to denote the searched paths, where $m_{i,j}$ is the number of paths between circRNA c_i and disease d_j.

Finally, the association score $AS(c_i,d_j)$ between circRNA c_i and disease d_j can be computed as follows:

$$AS(c_i,d_j)=\sum_{k=1}^{m_{i,j}}\left(\prod_{t=1}^{\mathrm{len}(p_k)}W_{n_{k_t},n_{k_{t+1}}}\right)^{\alpha\times\exp(\mathrm{len}(p_k))} \tag{6}$$

where $W_{n_{k_t},n_{k_{t+1}}}$ denotes the weight of the edge connecting the node n_{k_t} and $n_{k_{t+1}}$. In addition, α is a decay factor and len(p_k) is the length of p_k.

KATZHCDA

Fan *et al.* [80] established a calculation model (see Figure 3) of KATZ-based Human CircRNA-Disease Association prediction (KATZHCDA). KATZ measure is a network-based method, which computes similarity of nodes in a heterogeneous network to solve the problem of association prediction [81, 22]. In KATZHCDA, the authors first compute the integrated similarity for circRNAs and diseases, which are denoted by the matrices of CS and DS, respectively. Besides, the association matrix CD is employed to denote the information of circRNA-disease associations, and $CD(i,j)$ is equal to 1 if circRNA c_i is associated with disease d_j, otherwise 0. Secondly, circRNA similarity network, disease similarity network as well as circRNA-disease association network are combined to construct a heterogeneous network whose adjacency matrix can be described as follows:

$$A^*=\begin{bmatrix}CS & CD\\ CD^T & DS\end{bmatrix} \tag{7}$$

The number of walks between circRNA nodes and disease nodes, as well as the length of walks are two key similarity metrics in the heterogeneous network. Because the contribution of longer walks is lower than that of shorter walks, the parameter γ is utilized to control the contribution of walks with different lengths. The final association score between c_i and d_j can be defined as follows:

$$AS(c_i,d_j)=\sum_{L=1}^{K}\gamma^L A^{*L}(i,j) \tag{8}$$

where the variable L denotes the length of walk and the variable K is the user specified parameter. Equation (8) can be transformed into the matrix form

$$AS=\sum_{L\geq 1}\gamma^L A^{*L}=(I-\gamma A^*)^{-1}-I \tag{9}$$

where AS can be used to predict potential circRNA-disease associations. As walks with longer length may be insignificant, the variable K is normally set as 2, 3 and 4, respectively. One

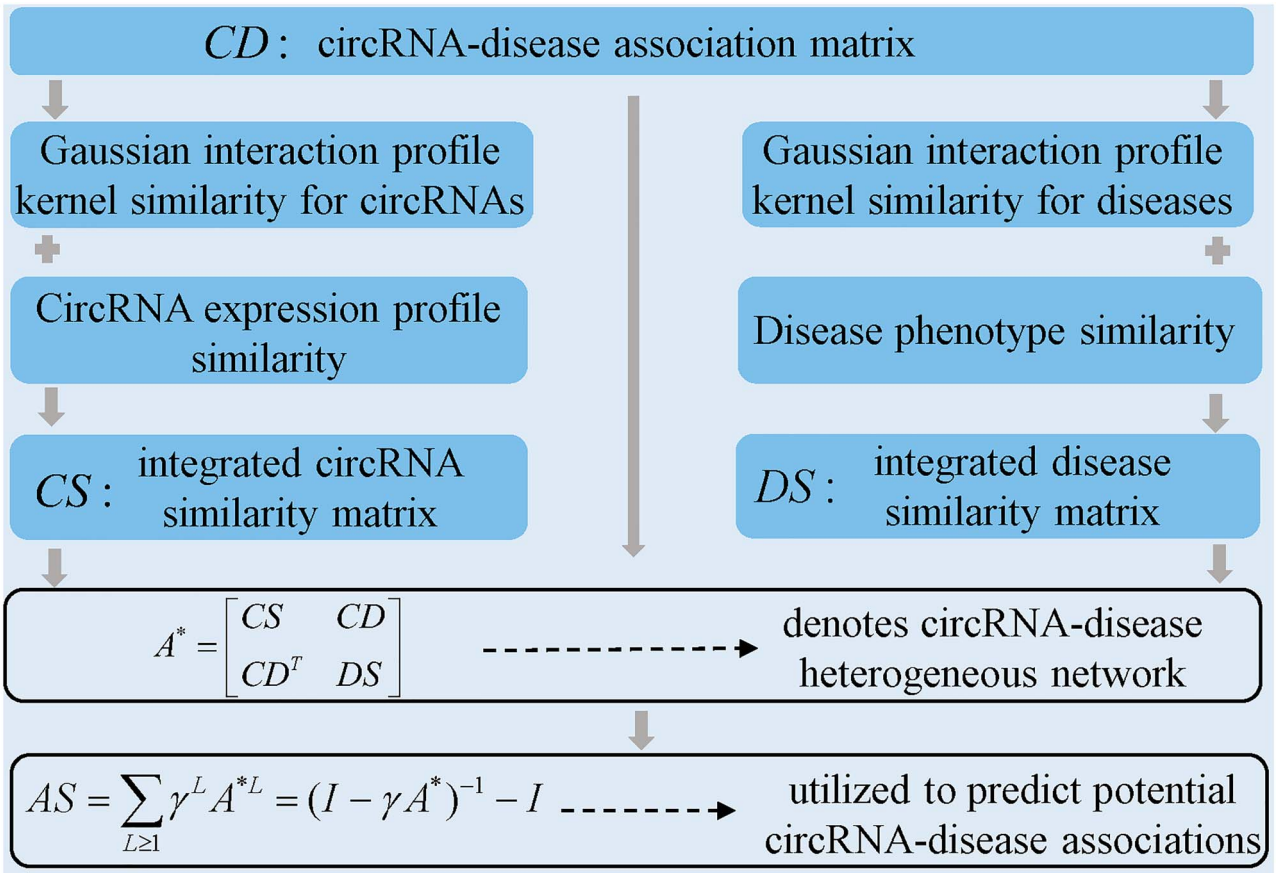

Figure 3. The flow diagram of KATZHCDA to predict human circRNA-disease associations based on KATZ algorithm.

advantage of KATZHCDA lies that it can predict circRNA-disease association scores for all diseases simultaneously. Besides, KATZHCDA can predict associated circRNAs for new diseases without any known associations.

KATZCPDA

Deng *et al.* [83] developed the model of KATZCPDA based on the KATZ method and the information of circRNA, protein and disease. Because the number of circRNA-disease associations validated by experiments is insufficient, they first obtain inferred circRNA-disease relationships by utilizing protein-circRNA association network and protein-disease association network based on the principle of gilt-by-association, that is biological objects are more likely to be associated if they have the same or related behavior [84]. Then, they construct a heterogeneous network by integrating the circRNA similarity network denoted by matrix *CS*, the disease similarity network denoted by matrix *DS* and the circRNA-disease association network denoted by *CD*, which combines the experimentally confirmed circRNA-disease associations and inferred circRNA-disease associations. The heterogeneous network can be represented as follows:

$$A^* = \begin{bmatrix} CS & CD \\ CD^T & DS \end{bmatrix} \tag{10}$$

Next, the final circRNA-disease association matrix is obtained in the similar way as KATZHCDA. KATZCPDA introduces the bridge of protein to obtain inferred circRNA-disease relationships, which increases the number of associations and the quantity of heterogeneous network.

IBNPKATZ

Zhao *et al.* [85] raised a novel circRNA-disease association prediction model (see Figure 4) by Integrating Bipartite Network Projection algorithm and KATZ measure (IBNPKATZ). Firstly, in the bipartite network projection algorithm, resource scores of circRNAs are used to be the association scores for a given disease. Specifically, a hierarchical clustering algorithm is utilized to construct circRNAs' bias ratings which denote the association degree between diseases and their associated circRNAs from circRNAs' perspective. For disease d_i, the bias rating of its related circRNA c_j can be computed as follows:

$$r(d_i, c_j) = \frac{n_{cr}(c_j)}{T(d_i)} \tag{11}$$

where $n_{cr}(c_j)$ is the number of circRNAs in the cluster cr including c_j and $T(d_i)$ denotes the number of circRNAs related with d_i. For d_i, the initial resource score of its related circRNA c_j can be calculated by normalizing the bias rating of c_j as follows:

$$\hat{r}(d_i, c_j) = \frac{r(d_i, c_i)\, T(d_i)}{\sum_{k=1}^{N_c} r(d_i, c_k)} \tag{12}$$

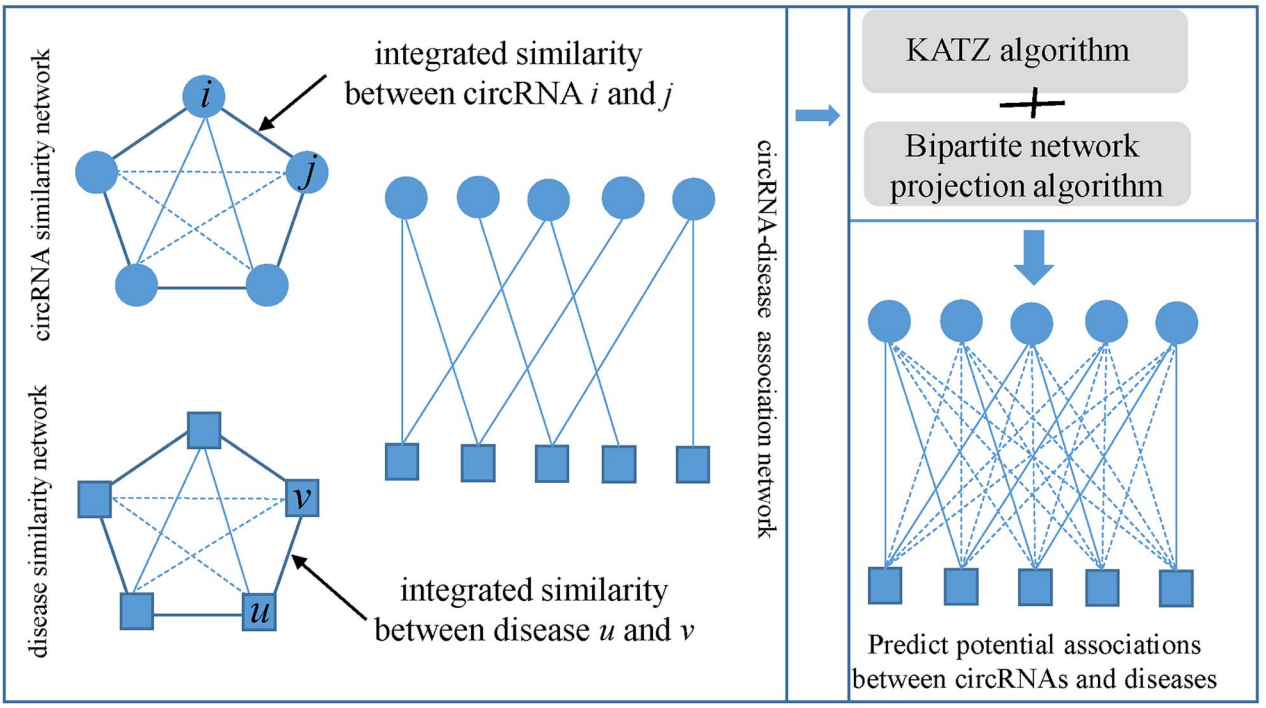

Figure 4. The workflow of IBNPKATZ to infer circRNA-disease associations based on bipartite network projection algorithm and KATZ algorithm.

where N_c is the number of circRNAs. Then, circRNAs associated with d_i allocate their resource score to their associated diseases as follows:

$$R_{cd}\left(d_i, c_j\right) = \frac{\hat{r}\left(d_i, c_j\right)}{\sum_{k=1}^{N_d} \hat{r}\left(d_k, c_j\right)} \times \hat{r}\left(d_i, c_j\right) \tag{13}$$

where N_d is the number of diseases. Next, the diseases distribute their received resource score to their associated circRNAs as follows:

$$R_{dc}\left(d_i, c_j\right) = \frac{r\left(d_i, c_j\right)}{\sum_{k=1}^{N_c} r\left(d_i, c_k\right)} \times \sum_{k=1}^{N_c} R_{cd}\left(d_i, c_k\right) \tag{14}$$

The final resource score of cricRNA c_j for given disease d_i can be computed as follows:

$$R_{\text{fin}}\left(c_j|d_i\right) = \sum_{k=1}^{N_d} R_{dc}\left(d_k, m_j\right) \tag{15}$$

Similarly, the final resource score $R_{\text{fin}}(d_i|c_j)$ of disease d_i for circRNA c_j could be obtained. Finally, the predicted circRNA-disease association score based on the bipartite network projection algorithm is defined as

$$S_{\text{BNP}}\left(d_i, c_j\right) = \frac{R_{\text{fin}}\left(c_j|d_j\right) + R_{\text{fin}}\left(d_i|c_j\right)}{2} \tag{16}$$

Secondly, the authors utilize KATZ measure on the heterogeneous network, constructed by using information of integrated circRNA similarity, integrated disease similarity and known circRNA-disease relationships, to predict circRNA-disease association score $S_{\text{KATZ}}(d_i, c_j)$ in the similar way as KATZHCDA. Finally, the circRNA-disease association scores of $S_{\text{BNP}}(d_i, c_j)$ and $S_{\text{KATZ}}(d_i, c_j)$ are integrated as the final association score

$$AS\left(d_i, c_j\right) = \frac{S_{\text{BNP}}\left(d_i, c_j\right) + S_{\text{KATZ}}\left(d_i, c_j\right)}{2} \tag{17}$$

Combination of two different prediction algorithms contributes to the ideal predictive performance of IBNPKATA.

NCPCDA

Li *et al.* [86] raised a calculation model (see Figure 5) of Network Consistency Projection for inferring CircRNA-Disease Association (NCPCDA). In NCPCDA, the binary matrix *CD* denotes the circRNA-disease associations. Besides, *CS* and *DS* represent integrated similarity matrices of circRNAs and diseases, respectively. The circRNA similarity and disease similarity are defined as follow:

$$CS\left(c_i, c_j\right) = \begin{cases} KC\left(c_i, c_j\right) & \text{if } CFS\left(c_i, c_j\right) = 0 \\ CFS\left(c_i, c_j\right) & \text{otherwise} \end{cases} \tag{18}$$

$$DS\left(d_i, d_j\right) = \begin{cases} KD\left(d_i, d_j\right) & \text{if } DSS\left(d_i, d_j\right) = 0 \\ DSS\left(d_i, d_j\right) & \text{otherwise} \end{cases} \tag{19}$$

where *KC* and *KD* denote the Gaussian interaction profile (GIP) kernel similarity matrices of circRNAs and diseases, respectively. Besides, the matrices *CFS* and *DSS* are circRNA functional similarity matrix and disease semantic similarity matrix, respectively. NCPCDA is made up of circRNA space projection *CSP* and disease space projection *DSP*, which are defined as

$$CSP\left(i, j\right) = \frac{CS\left(i, :\right) \times CD\left(:, j\right)}{\left\|CD\left(:, j\right)\right\|_2} \tag{20}$$

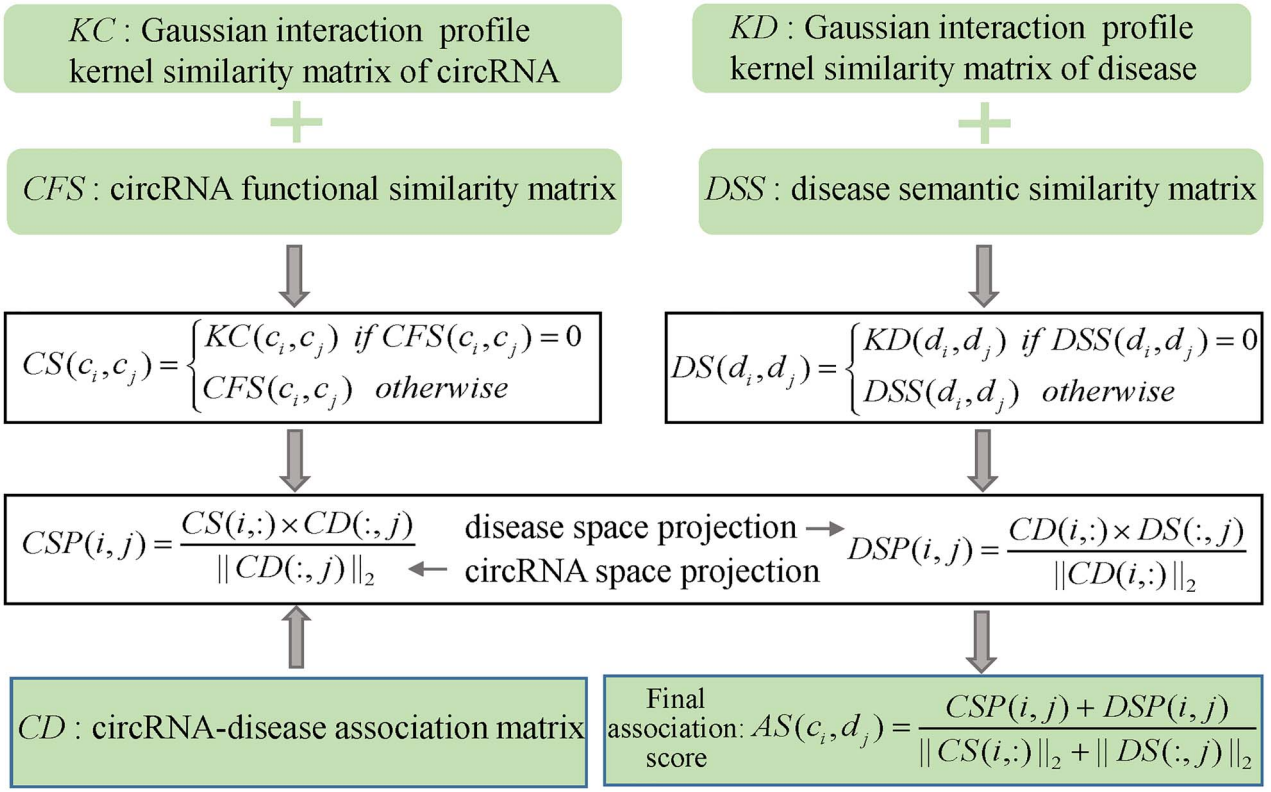

Figure 5. The flowchart of NCPCDA for circRNA-disease association prediction based on the circRNA space projection and disease space projection.

$$DSP(i,j)=\frac{CD(i,:)\times DS(:,j)}{\|CD(i,:)\|_2} \tag{21}$$

where $CS(i,:)$ and $CD(i,:)$ are the ith rows of CS and CD, respectively. Besides, $DS(:,j)$ and $CD(:,j)$ are the jth columns of DS and CD, respectively. In the end, the final associations score $AS(c_i, d_j)$ between circRNA c_i and disease d_j can be calculated by integrating and normalizing CSP and DSP as follows:

$$AS(c_i,d_j)=\frac{CSP(i,j)+DSP(i,j)}{\|CS(i,:)\|_2+\|DS(:,j)\|_2} \tag{22}$$

No parameters appear in NCPCDA, which reduces the complexity of prediction process. However, the similarity of circRNA is calculated only based on known circRNA-disease associations, which leads to the failure of NCPCDA for predicting associated diseases for cirRNAs without any known related diseases.

DWNCPCDA

Li *et al.* [87] developed the DeepWalk and Network Consistency Projection-based algorithm to predict CircRNA-Disease Association (DWNCPCDA). In most of circRNA-disease association prediction models, the circRNA similarity and disease similarity are usually calculated by multiple biological information of circRNAs and diseases. In this study, the authors construct circRNA topological similarity matrix CTS and disease topological similarity matrix DTS only based on circRNA-disease association network. More formally, the DeepWalk algorithm [88] is utilized to learn circRNA representations stored by the matrix CR and disease representations stored by the matrix DR based on the circRNA-disease association network. DeepWalk obtains local information of input graph by truncated random walk and utilizes them to learn latent representations of vertices in the input graph [88]. Then, similarity between circRNAs or diseases can be computed as follows:

$$CTS(c_i,c_j)=\frac{\sum_{k=1}^{d}CR(c_i,k)\times CR(c_j,k)}{\sqrt{\sum_{k=1}^{d}CR(c_i,k)^2}\sqrt{\sum_{k=1}^{d}CR(c_j,k)^2}} \tag{23}$$

$$DTS(d_i,d_j)=\frac{\sum_{k=1}^{d}DR(d_i,k)\times DR(d_j,k)}{\sqrt{\sum_{k=1}^{d}DR(d_i,k)^2}\sqrt{\sum_{k=1}^{d}DR(d_j,k)^2}} \tag{24}$$

where the variable d is the dimension of representations of circRNAs and diseases.

After obtaining CTS and DTS, network consistency projection method, which have been used in the prediction model of NCPCDA, is adopt to calculate circRNA-disease association matrix AS. Although similarity of circRNA and disease is computed only based on the circRNA-disease association network, DWNCPCDA still achieves good predictive accuracy, which demonstrates the excellent ability of DeepWalk in learning latent representations of circRNAs and diseases.

LLCDC

Ge *et al.* [89] proposed a computational model of LLCDC (see Figure 6) to predict potential circRNA-disease associations

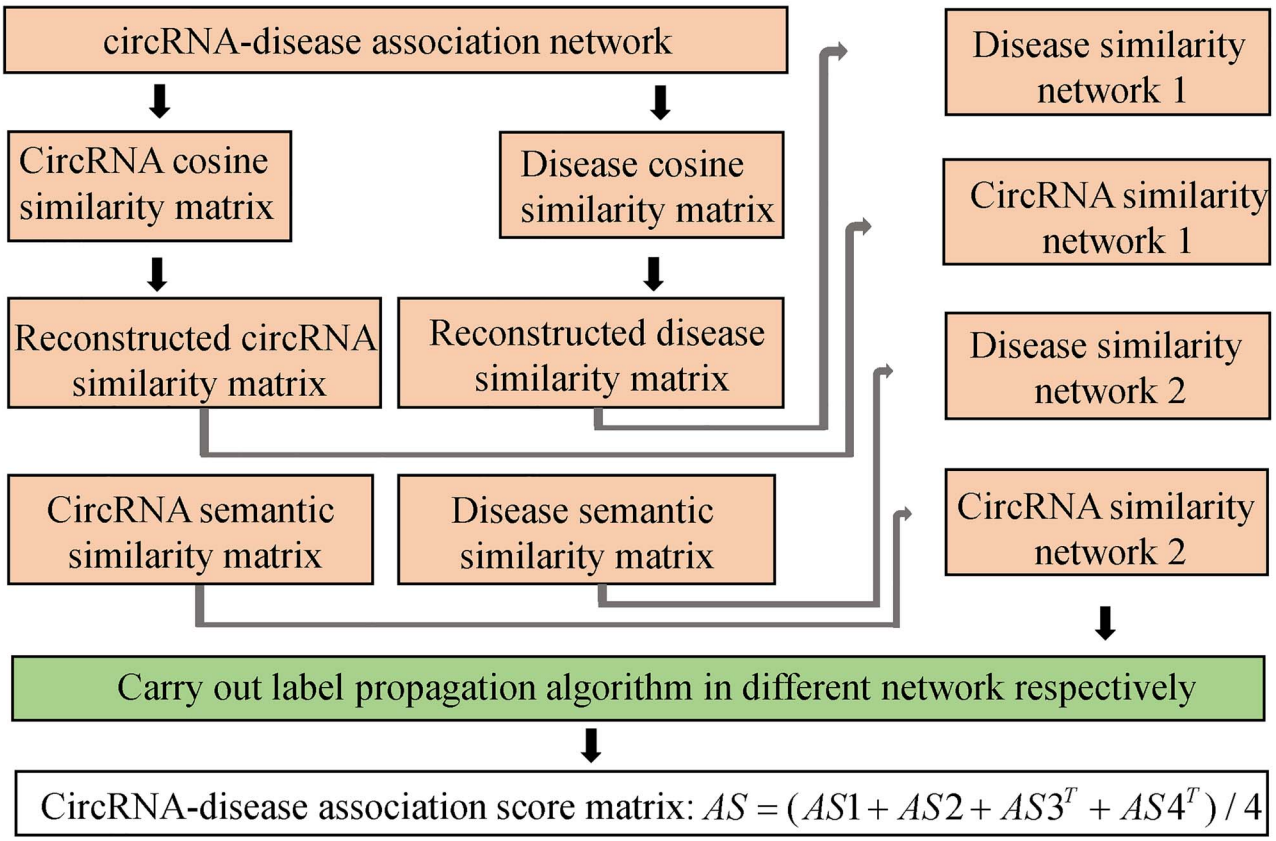

Figure 6. The framework of LLCDC to predict potential circRNA-disease associations based on LLC and label propagation algorithm.

based on locality-constrained linear coding (LLC) and label propagation algorithm. Firstly, they calculate circRNA semantic similarity matrix *CSS* based on GO terms of circRNA-related genes. Besides, disease semantic matrix *DSS* is calculated based on MeSH descriptors of diseases. Secondly, they also calculate cosine similarity matrices of circRNAs and diseases based on circRNA-disease association information and further utilized LLC to obtain reconstructed circRNA similarity matrix *RCS* and reconstructed disease similarity matrix *RDS* based on above two cosine similarity matrices. Thirdly, label propagation algorithm is employed to obtain the initial predicted circRNA-disease association matrix *AS1* based on circRNA semantic similarity network by the following iterative equation:

$$AS1(t+1) = \theta \times CSS \times AS1(t) + (1-\theta)\,CD \tag{25}$$

where $AS1(0) = CD$ and θ are used to control the utilization of similarity and association information. $AS1(t)$ denotes the association matrix obtained in the tth iteration. The iterative equation will be conducted until *AS1* converges. In a similar way, label propagation algorithm is carried out based on *DSS*, *RCS* and *RDS* to obtain association matrices *AS2*, *AS3* and *AS4*, which are combined as the finally predicted association matrix *AS* as follows:

$$AS = \frac{1}{4}\left(AS1 + AS2 + AS3^T + AS4^T\right) \tag{26}$$

CD-LNLP

Zhang *et al.* [90] put forward a computational method to infer CircRNA-Disease associations based on a Linear Neighborhood similarity measure and Label Propagation algorithm (CD-LNLP). The information of associations between N_c circRNAs and N_d diseases is recorded in the binary matrix *CD*. In CD-LNLP, linear neighborhood similarity (LNS) measure is utilized to construct circRNA similarity matrix *CS* and disease similarity matrix *DS*. In LNS, the ith row vector of *CD* is considered as the feature profile of circRNA c_i. The basic idea of LNS is that each feature profile of circRNA can be reconstructed by the linear combination of feature profiles of neighbors of the circRNA, which can be formulated as follows:

$$\begin{aligned} &\min_{CS} \tfrac{1}{2}\|CD - (C * CS)\,CD\|_F^2 + \tfrac{\mu}{2}\sum_{i=1}^{N_c}\|(C * CS)_{i.}\|_1^2 \\ &s.t. \quad (C * CS)\,e = e, \quad CS \geq 0 \end{aligned} \tag{27}$$

where $*$ is the Hadamard product. The matrix C with the size of $N_c \times N_c$ is an indicator matrix, whose element $C(i,j)$ is equal to 1 if circRNA c_j is one of the K nearest neighbors (by Euclidean distance) of circRNA c_i; otherwise, $C(i,j) = 0$. Besides, $(C * CS)_{i.}$ is the ith row of $C * CS$. In addition, e is a $N_c \times 1$ vector and all elements in e are 1. The first item of above formula is the loss function of LNS. The second item is used to achieve row sparsity of $C * CS$. The constraint condition is used to ensure that the sum of similarity values between any circRNA and its neighbors is equal to 1. By utilizing Lagrange multiplier method to solve the

optimization problem, they obtain the update rule for CS

$$CS(i,j) = \begin{cases} CS(i,j)\dfrac{\left(\mathbf{CDCD}^T + ee^T\right)_{i,j}}{\left((C*CS)\mathbf{CDCD}^T + \mu(C*CS)ee^T\right)_{ij}} & i \neq j \\ 0 \quad i = j \end{cases} \tag{28}$$

In a similar way, disease similarity matrix DS can be obtained. Next, a label process [91] is employed to predicted potential circRNA-disease relationships, which can be formulated as follows:

$$AS_{\text{cicRNA}} = (1-\alpha)(I - \alpha CS)^{-1}CD \tag{29}$$

$$AS_{\text{disease}} = (1-\alpha)(I - \alpha DS)^{-1}CD^T \tag{30}$$

where the $N_c \times N_d$ matrix AS_{circRNA} and the $N_d \times N_c$ matrix AS_{disease} are the predicted association matrix based on circRNA similarity and disease similarity, respectively. Finally, the integrated association scores between circRNAs and diseases can be computed as follows:

$$AS = \rho AS_{\text{circRNA}} + (1-\rho)AS^T_{\text{disease}} \tag{31}$$

where the parameter ρ is utilized to regulate the weight of AS_{circRNA} and AS_{disease}. The application of LNS measure contributes to the effectiveness of CD-LNLP. However, the similarity of diseases and circRNAs is calculated only based on circRNA-disease association network.

Machine learning-based models

Machine learning algorithms have been successfully used in many fields of association prediction [92–101]. In the last few years, researchers utilized different machine learning methods to construct prediction models for the identification of potential circRNA-disease associations. These machine learning-based models can be further roughly divided into two types. The first type of models can obtain the predictive association matrix by directly solving specific optimization problem, such as regularized least squares, manifold regularization learning, matrix decomposition and inductive matrix completion algorithm-based models. In addition, the second type of models train classifier to infer circRNA-disease association, such as logistic regression-, K-Nearest Neighbor (KNN)-, Support Vector Machines (SVM)-, Random Forest (RF)-, Gradient Boosting Decision Tree (GBDT)-, Deep Forest (DF)-, Convolutional Neural Network (CNN)-, Graph Neural Network (GNN)- and Deep Neural Network (DNN)-based models. When feature vector of a sample is input into classifier, the classifier can output an association score for the sample. Furthermore, some prediction models combine different algorithms to improve the prediction accuracy.

The first type of machine learning-based models

DWNN-RLS

Yan *et al.* [102] developed a computational model, called as DWNN-RLS (see Figure 7) to infer potential circRNA-disease associations based on regularized least squares of kronecker product kernel (RLS-kron). In DWNN-RLS, the matrix CD is utilized to denote the information of known circRNA-disease relationships. In addition, the disease similarity matrix DS is obtained by integrating disease GIP kernel similarity matrix KD and disease semantic similarity matrix DSS. In this study, the authors first utilize DWNN (decreasing weight KNN) method to calculate the initial association score between new circRNA c_i and disease d_j as follows:

$$AS_{\text{initial}}(c_i, d_j) = \frac{\sum_{c_l \in N(c_i)} KC(c_i, c_l) \times CD(c_l, d_j)}{\sum_{c_l \in N(c_i)} KC(c_i, c_l)} \tag{32}$$

where the new circRNA c_i means that c_i has no known associated disease. In addition, $N(c_i)$ is the set of all neighbors of c_i. Similarly, the initial association score between new disease d_j and circRNA c_i can be calculated as follows:

$$AS_{\text{initial}}(c_i, d_j) = \frac{\sum_{d_l \in N(d_j)} KD(d_i, d_l) \times CD(c_i, d_l)}{\sum_{d_l \in N(d_j)} KD(d_i, d_l)} \tag{33}$$

where $N(d_j)$ represents the set of all neighbors of d_i. Then, they employ the RLS-kron method to infer new associations between circRNAs and diseases as follows:

$$\text{vec}\left(AS^T\right) = K(K + \lambda I)^{-1}\text{vec}\left(CD^T\right) \tag{34}$$

where the kernel $K = KC \otimes DS$ is the Kronecker product of KC and DS. As KC and DS are real symmetric matrices, the two matrices can be decomposed as follows:

$$KC = \vee_c \wedge_c \vee_c^T \tag{35}$$

$$DS = \vee_d \wedge_d \vee_d^T \tag{36}$$

where the columns of the matrices of $\vee_c$ and $\vee_d$ are the eigenvectors of KC and DS, respectively. Besides, $\wedge_c$ and $\wedge_d$ are diagonal matrices whose diagonal elements are the eigenvalues of KC and DS, respectively. Thus, the finally predicted circRNA-disease association matrix can be computed as follows:

$$AS = \vee_c Z^T \vee_d^T \tag{37}$$

$$\text{vec}(Z) = (\wedge_c \otimes \wedge_d)(\wedge_c \otimes \wedge_d + \lambda I)^{-1}\text{vec}\left(\vee_d^T CD^T \vee_c\right) \tag{38}$$

RWLRCDA

Ding *et al.* [103] built a computational model based on Random Walk and Logistic Regression to infer CircRNA-Disease Associations (RWLRCDA). Specifically, they first calculate the circRNA similarity matrix CS and construct circRNA similarity network where vertex c_i and c_j are connected by an edge with the weight value of $CS(c_i, c_j)$. Subsequently, aiming to obtain the global relationship information of each circRNA, the authors treat each circRNA as seed node in turn and utilize the random walk with restart algorithm on circRNA similarity network to obtain related circRNAs for the seed node with corresponding probability. Next, they extract three features, namely *pos*, *neg* and *label*, for each pair of circRNA c_i and disease d_j. Specifically, $C_k(i)$ denotes the set of top-k circRNAs related with c_i. The *pos* value is the sum of probability of circRNAs which are in $C_k(i)$ and related with d_j. Similarly, The *neg* value is the sum of probability of circRNAs which are in $C_k(i)$ and not related with d_j. The *label* value is 1 or

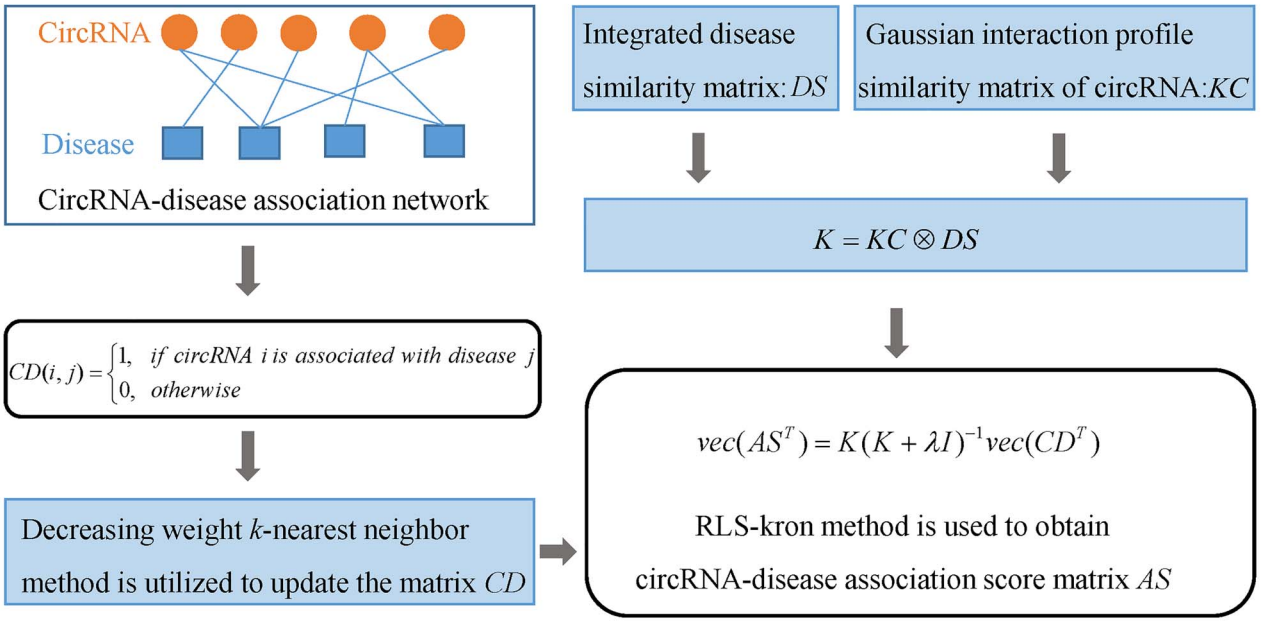

Figure 7. The workflow of DWNN-RLS to infer potential circRNA-disease associations based on regularized least squares of kronecker product kernel.

0. Finally, logistic regression is utilized to predict the association score for circRNA-disease pair as follows:

$$AS(x) = \frac{\exp(w \times x)}{1 + \exp(w \times x)} \tag{39}$$

where x is the feature vector consisting of three features (*pos*, *neg* and *label*) of the circRNA-disease pair and w is the weight vector which can be trained by maximizing the posterior association probability of circRNA-disease training samples as follows:

$$w = \arg\max\left(\prod_{i=1}^{m} p(y_i|x_i)\right) \tag{40}$$

$$p(y_i = 1|x_i) = \frac{\exp(w \times x_i)}{1 + \exp(w \times x_i)} \tag{41}$$

$$p(y_i = 0|x_i) = \frac{1}{1 + \exp(w \times x_i)} \tag{42}$$

where m is the number of training samples. Besides, x_i and y_i are the feature vector and label of the ith circRNA-disease sample. RWLRCDA can predict associations for new diseases or new circRNAs. However, RWLRCDA utilizes too little information of diseases.

MRLDC

Xiao *et al.* [103] developed a manifold regularization-learning framework, called MRLDC, for predicting human disease-associated circRNAs (see Figure 8). They construct a circRNA-disease bilayer heterogeneous network by connecting circRNA-circRNA, disease-disease and circRNA-disease through edges weighted by the matrices CS, DS and CD, respectively. Besides, they construct circRNA graph and disease graph to inspect the geometrical structure of circRNA data and disease data. The weight matrix W_{cg} of circRNA graph is formulated as follows:

$$W_{cg}(i,j) = W_c(i,j) \cdot CS(i,j) \tag{43}$$

$$W_c(i,j) = \begin{cases} 1, & \text{if } c_i \in C_k \text{ and } c_j \in C_k \\ 0, & \text{otherwise} \end{cases} \tag{44}$$

where C_k represents the kth cluster obtained by using ClusterONE [105] based on circRNA similarity network. Besides, D'_c is a diagonal matrix, where $(D'_c)_{ii} = \sum_j W_{cg}(i,j)$. The matrix $L_c = D'_c - W_{cg}$ denotes the graph Laplacian matrix of circRNA graph. Similarly, the graph Laplacian matrix L_d of disease graph can be obtained. Then, to obtain the low-rank feature matrices of circRNAs and diseases, namely P and Q, which can be used for predicting circRNA-disease associations, they formulate the weighted dual-manifold regularization learning-based calculation model of MRLDC as follows:

$$\begin{aligned} \min_{P,Q\geq 0} f(P,Q) = &\|I * (CD - PQ)\|_F^2 + \lambda_1 Tr(P^T L_c P) \\ &+ \lambda_2 Tr(QL_d Q^T) + \lambda_3 \left\|PP^T - CS\right\|_F^2 \\ &+ \lambda_4 \left\|Q^T Q - DS\right\|_F^2 + \lambda_5 \left(\|P\|_F^2 + \|Q\|_F^2\right) \end{aligned} \tag{45}$$

where P and Q are the low-rank feature matrices of circRNAs and diseases in the bilayer heterogeneous network, which can be obtained by solving above formula. Besides, I is an indicator weighted matrix where $I(i,j)$ is equal to 1 if circRNA c_i is associated with disease d_j, otherwise $I(i,j) = 0$. In addition, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and λ_5 are regulation parameters. The second item and the third item in above formula are the manifold regularization terms of circRNA and disease space, respectively. The fourth item (fifth item) is utilized to achieve the purpose that the similarity of circRNAs (diseases) should approximate the inner product of their feature vectors. The last item is to ensure the smoothness of P and Q. Next, the Lagrange multiplier method is employed to optimize above objective function and the following updating rules can be obtained:

$$\begin{aligned} P_{ik} \leftarrow P_{ik} &\frac{(I_{i\cdot} * CD_{i\cdot})(Q^T)_{\cdot k} + \lambda_1 (W_{cg}P)_{ik} + 0.5\lambda_3 \left(CS^T P\right)_{ik}}{(I_{i\cdot} * (P_{i\cdot}Q))(Q^T)_{\cdot k} + \lambda_1 (D'_c P)_{ik} + U} \\ &\left(U = 0.5\lambda_3 \left(PP^T P\right)_{ik} + \lambda_5 P_{ik}\right) \end{aligned} \tag{46}$$

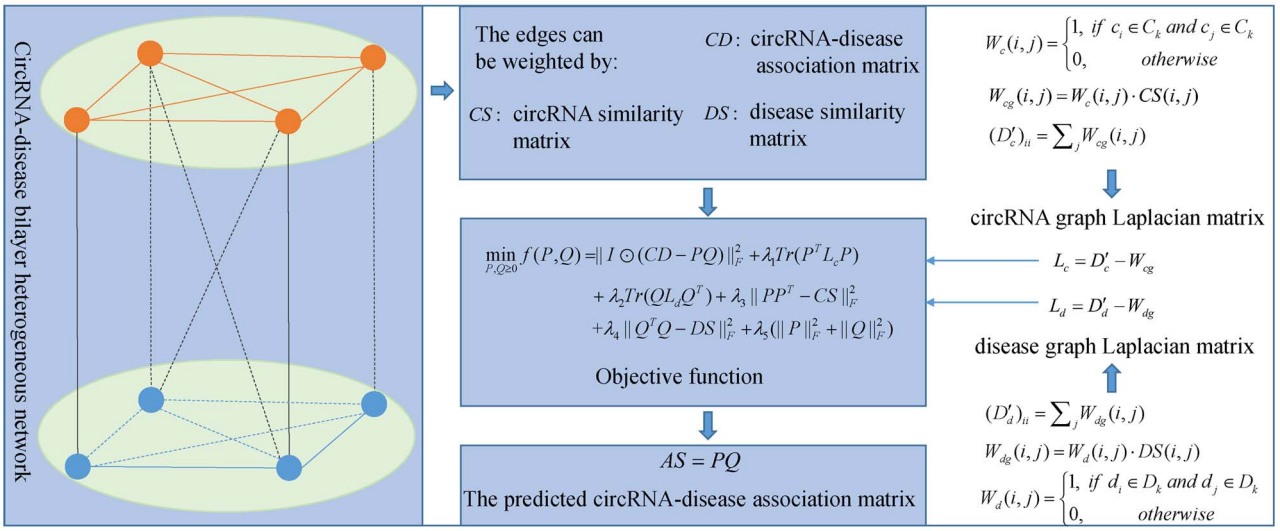

Figure 8. The flowchart of MRLDC for predicting human disease-associated circRNAs based on a manifold regularization-learning framework.

$$Q_{kj} \leftarrow Q_{kj} \frac{\left(P^T\right)_{k\cdot}\left(I_j * CD_{\cdot j}\right) + \lambda_2\left(QW_{dg}\right)_{kj} + 0.5\lambda_4\left(QDS^T\right)_{kj}}{\left(P^T\right)_{k\cdot}\left(I_j * \left(PQ_{\cdot j}\right)\right) + \lambda_2\left(QD'_d\right)_{kj} + V} \quad \left(V = 0.5\lambda_4\left(QQ^TQ\right)_{kj} + \lambda_5 P_{jk}\right) \tag{47}$$

Finally, the predicted circRNA-disease association matrix $AS = PQ$. The parameters in MRLDC are hard to select. Besides, MRLDC is inappropriate for new disease without any observed associations.

iCircDA-MF

Wei *et al.* [106] proposed a calculation method (see Figure 9) to identify CircRNA-Disease Associations based on Matrix Factorization (iCircDA-MF). In the model of iCircDA-MF, the authors first construct circRNA similarity matrix *CS* by integrating circRNA GIP kernel similarity and circRNA-related gene-based similarity, and disease similarity matrix *DS* by integrating disease GIP kernel similarity and disease semantic similarity. Besides, the collected circRNA-disease associations are denoted by the matrix *CD*. However, many false negative associations are assigned as zero in *CD*. To reduce the noise, the authors reformulate the matrix *CD* to CD_d and CD_c from the vertical direction and the horizontal direction by utilizing the interaction profiles of top-*k* neighbors of investigated disease and circRNA as follows:

$$CD_d\left(:, d_i\right) = \frac{1}{W_{d_i}} \sum_{j=1}^{k} DS\left(d_i, d_j\right) \times CD\left(:, d_j\right) \tag{48}$$

$$CD_c\left(c_m, :\right) = \frac{1}{W_{c_m}} \sum_{n=1}^{k} CS\left(c_m, c_n\right) \times CD\left(c_n, :\right) \tag{49}$$

where $CD_d(:, d_i)$ and $CD(:, d_j)$ are the ith column of CD_d and the *j*th column of *CD*. Besides, $W_{d_i} = \sum_{1 \le j \le k} DS(d_i, d_j)$. In addition, $CD_c(c_m, ;)$ and $CD(c_n, :)$ denote the *m*th row of CD_c and the *n*th row of *CD*. Moreover, $W_{c_m} = \sum_{1 \le n \le k} CS(c_m, c_n)$. The final reformulated circRNA-disease association matrix is as follows:

$$CD' = \max\left(CD, \left(CD_c + CD_d\right)/2\right) \tag{50}$$

Next, matrix factorization method is utilized to predict potential circRNA-disease associations, which can be formulated as follows:

$$\min_{P \ge 0, Q \ge 0} \left\| CD' - PQ^T \right\|_F^2 + \alpha \left\| PQ^T \right\|_F^2 + \beta\left(Tr\left(P^T L_c P\right) + Tr\left(Q^T L_d Q\right)\right) \tag{51}$$

where *P* and *Q* represent two low-dimension feature matrices of circRNAs and diseases, respectively. In addition, $L_c = D'_c - CS$ and $L_d = D'_d - DS$ are two graph Laplacian matrices of circRNA and disease space. Here, D'_c and D'_d are two diagonal matrices, where $D'_c(i, i) = \sum_j CS(i, j)$ and $D'_d(i, i) = \sum_j DS(i, j)$. The first item in Eq. (51) is the loss function of matrix factorization method. The second item is used to avoid overfitting and ensure the smoothness of circRNA and disease space. Besides, the last item can restrict the geometrical structure of target space and reduce noise [107, 108]. In addition, α and β are regulation parameters.

Finally, the predicted circRNA-disease association matrix *AS* can be calculated as $AS = PQ^T$ after solving Eq. (51). This work can effectively deal with noise data.

GMCDA

Xiao *et al.* [109] designed a Graph-based Multi-label learning for CircRNA-Disease Association prediction (GMCDA). The integrated similarity matrices of *CS* and *DS* are obtained by fusing directed acyclic graphs of diseases and circRNA-disease associations. The authors aim to generate an expected association matrix *AS* to restore the missing values in the original circRNA-disease association matrix *CD*. To achieve the aim, the multi-label learning-based framework is proposed and formulated by an objective function with three constraints as follows:

$$\begin{aligned} &\min_{AS \ge 0} \|I * (AS - CD)\|_F^2 \\ &\quad + \lambda\left(\left\|AS \times CS \times AS^T - CS\right\|_F^2 + \left\|AS^T \times DS \times AS - DS\right\|_F^2\right) \\ &\quad + \gamma\left(Tr\left(AS^T \times L_c \times AS\right) + Tr\left(AS \times L_d \times AS^T\right)\right) + \mu\|AS\|_{1,2}^2 \end{aligned} \tag{52}$$

where *I* is an indicator matrix (*I*=*CD*). Besides, the graph Laplacian matrices of L_c and L_d can be computed by the same way used in the previous model of MRLDC. In addition, λ, γ and μ are constants used to control the contributions of different terms.

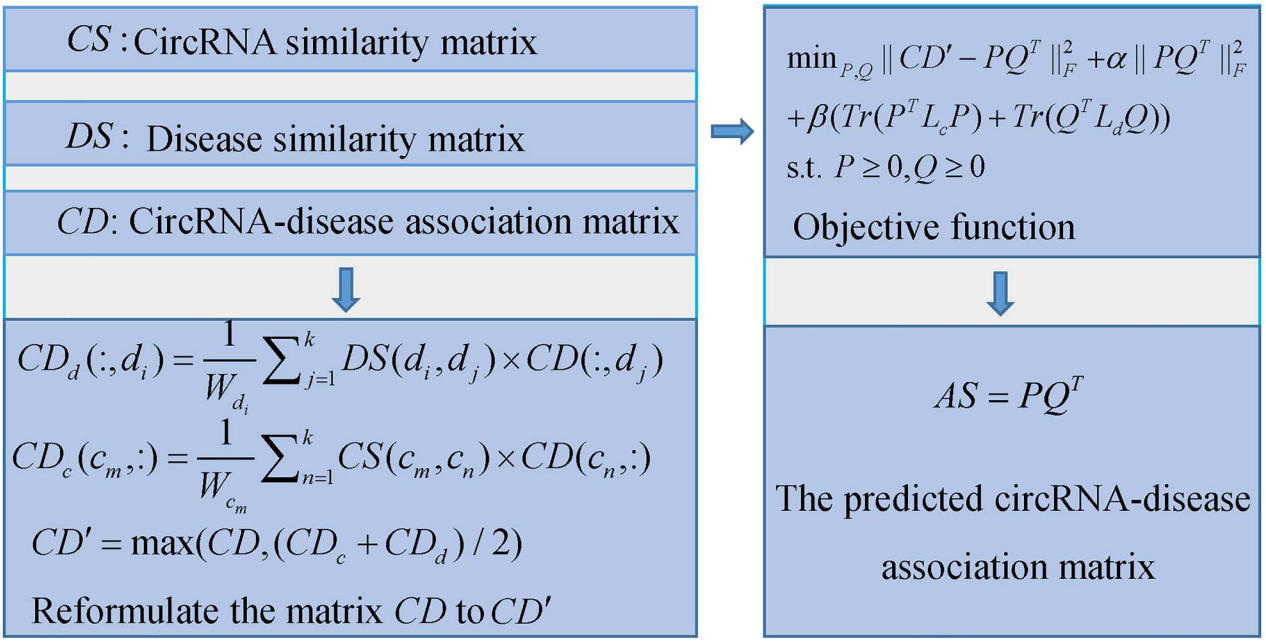

Figure 9. The flowchart of iCircDA-MF to identify circRNA-disease associations based on matrix factorization.

The first item in above formula is the loss function of GMCDA. The second item means that the expected similarity values of circRNA pairs and disease pairs should be approximate to the original similarities. The third item is used to capture geometrical structures of data. The last item is utilized to increase the sparsity of *AS* and reduce noisy. The local optimal solution of this objective function can be obtained by an iterative method.

iCDA-CMG

Xiao *et al.* [110] proposed the algorithm of identifying CircRNA-Disease Associations by using Collective Matrix completion with Graph learning (iCDA-CMG). First, the circRNA similarity matrix *CS* is obtained based on circRNA-disease association information. Besides, the disease similarity matrix *DS* fuses the data of directed acyclic graphs of diseases and circRNA-disease associations. Then, the DWNN method, in the same way as that used in the model of iCircDA-MF, is adopt to reconstruct circRNA-disease association matrix *CD* to the matrix *CD′*.

Next, the similarity matrices of *CS* and *DS* are reconstructed to the sparse similarity matrices of *CS′* and *DS′* by utilizing the structure information of circRNA graph (circRNA similarity network) and disease graph (disease similarity network). Subsequently, the objective function of iCDA-CGM is formulated to obtain the latent circRNA feature matrix $P \in R^{K\times N_c}$ and the latent disease feature matrix $Q \in R^{K\times N_d}$ as follows:

$$\begin{aligned}\min_{P\geq 0,Q\geq 0}&\left\|CD'-P^TQ\right\|_F^2+\lambda_c\sum_{i,j=1}^{N_c}\left\|P(:,i)-P(:,j)\right\|_F^2CS'_{i,j}\\&+\lambda_d\sum_{i,j=1}^{N_d}\left\|Q(:,i)-Q(:,j)\right\|_F^2DS'_{i,j}\\&+\delta_c\sum_{i=1}^{N_c}\left\|P(:,i)\right\|_1^2+\delta_d\sum_{i=1}^{N_d}\left\|Q(:,i)\right\|_1^2\end{aligned}\tag{53}$$

where the parameters of λ_c, λ_d, δ_c and δ_d are utilized to control the contributions of different regulation terms. The first item in above formula is the loss function of collective matrix completion. The second item (third item) is employed to achieve the purpose that the latent feature vectors of similar circRNAs (diseases) should be similar. The last two items are used to ensure the sparsity of *P* and *Q*. Finally, an alternating method with Lagrange multipliers is used to solve the objective function, and the predicted circRNA-disease association matrix is $AS = P^TQ$.

NMFIBAC

Wang *et al.* [111] developed a Non-negative Matrix Factorization algorithm (NMF)-based model to Identify Breast cancer Associated CircRNAs (NMFIBAC), which integrated multiple biological data including mRNA, miRNA, circRNA and pathway-related data. Firstly, they search DE circRNAs and miRNAs from RNA-seq data involving disease samples and normal samples. Then, they construct circRNA-mRNA association matrix X_1 based on DE circRNAs and co-expressed mRNAs, miRNA-mRNA association matrix X_2 based on DE miRNAs and miRNA target genes, as well as pathway-mRNA association matrix X_3. Subsequently, NMF algorithm is utilized to establish *K* circRNA modules by the following objective function *F*:

$$F(W,H)=\sum_{I=1}^{3}\|X_I-WH_I\|\tag{54}$$

where *W* is a matrix with the size of $M\times K$ (*M* denotes the number of mRNAs) representing the basis vector. In addition, the matrix $H_I(I \in (1,2,3))$ denotes the coefficient vector. After solving the objective function, the matrix *W* and $H_I(I \in (1,2,3))$ are utilized to determine the members (including miRNAs, mRNAs, circRNAs and pathways) of the *K* circRNA modules based on a previous method [112]. Finally, in each module, circRNAs connecting with more than four members are considered to be associated with breast cancer.

SIMCCDA

Li *et al.* [113] raised a model (see Figure 10) of Speedup Inductive Matrix Completion for CircRNA-Disease Association prediction

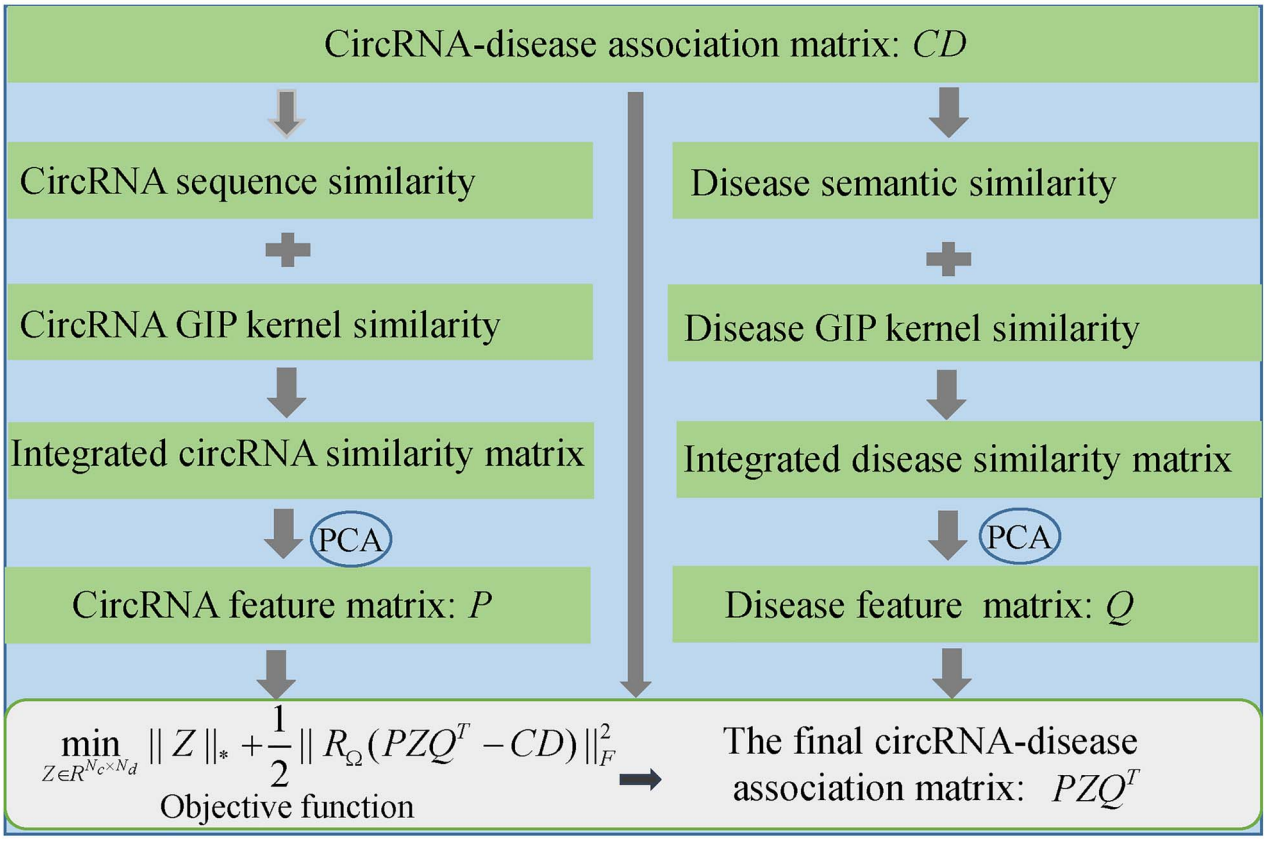

Figure 10. The framework of SIMCCDA for circRNA-disease association prediction based on inductive matrix completion.

(SIMCCDA). In SIMCCDA, *CS* and *DS* are calculated by combining circRNA sequence similarity, circRNA GIP kernel similarity, disease semantic similarity and disease GIP kernel similarity. Besides, principal component analysis is utilized to extract primary feature vectors of the matrices *CS* and *DS*. The extracted feature vectors are used to construct the circRNA feature matrix *P* and disease feature matrix Q. The objective function of inductive matrix completion can be defined as

$$\min_{Z \in R^{N_c \times N_d}} \|Z\|_* + \frac{1}{2} \left\| R_\Omega \left(PZQ^T - CD \right) \right\|_F^2 \tag{55}$$

where *Z* is the target matrix to complete *CD* and $\| \cdot \|_*$ denotes the nuclear norm. Besides, PZQ^T is the final circRNA-disease association matrix. In addition, Ω denotes known association sets. The first item in Eq. (55) is the constraint of low rank. The second item is employed to cater to the hypothesis that the row (or column) vectors in *CD* are located in the subspace spanned by the column vectors in Q (or *P*). The solution of *Z* can be obtained by using an accelerated proximal gradient algorithm [114].

PreCDA

Wang *et al.* [115] developed a calculation model named PreCDA to infer underling circRNA-disease associations (see Figure 11). They compute circRNA expression similarity matrix *CES* by Spearman correlation coefficient based on circRNA expression profile in 78 human cell types or tissues. Besides, the circRNA functional similarity matrix *CFS* is calculated based on known circRNA-disease associations. Then, they construct a circRNA association network, where the weight between circRNA c_i and c_j is defined as

$$\text{CicWeight}(i,j) = \begin{cases} (CFS(i,j) + CES(i,j))/2 & \text{if } CES(i,j) > 0 \\ CFS(i,j) & \text{otherwise} \end{cases} \tag{56}$$

To infer potential disease-associated circRNAs, the information of circRNA-disease associations is introduced into the circRNA association network. Based on the new network composed of circRNAs and diseases, PersonalRank algorithm is employed to identify disease-related circRNAs. Specifically, *PR*(i) is used to denote the possibility value that node *i* is accessed. In the beginning, *PR*(i) is equal to 1 if the node *i* is the target disease node *t*, otherwise 0. Then, the target node *t* randomly moves to neighbor nodes. In each move, the probability of returning to node *t* is $(1-\alpha)$. The following formula is defined to update *PR*(i) after each move:

$$PR(i) = (1-\alpha) r_i + \alpha \sum_{j \in in(i)} \frac{PR(j)}{|\, out(j) \,|} \tag{57}$$

$$r_i = \begin{cases} 1 & \text{if } i = t \\ 0 & \text{if } i \neq t \end{cases} \tag{58}$$

where in(i) and out(*j*) are the in-degree of node *i* and out-degree of node *j*, respectively; *d* is the transfer probability; *t* denotes the target node. After enough moves, the possibility value that node *i* is accessed will be stable. Finally, the probability value that a circRNA node is accessed can be used as the association score between the target disease *t* and this circRNA. The main

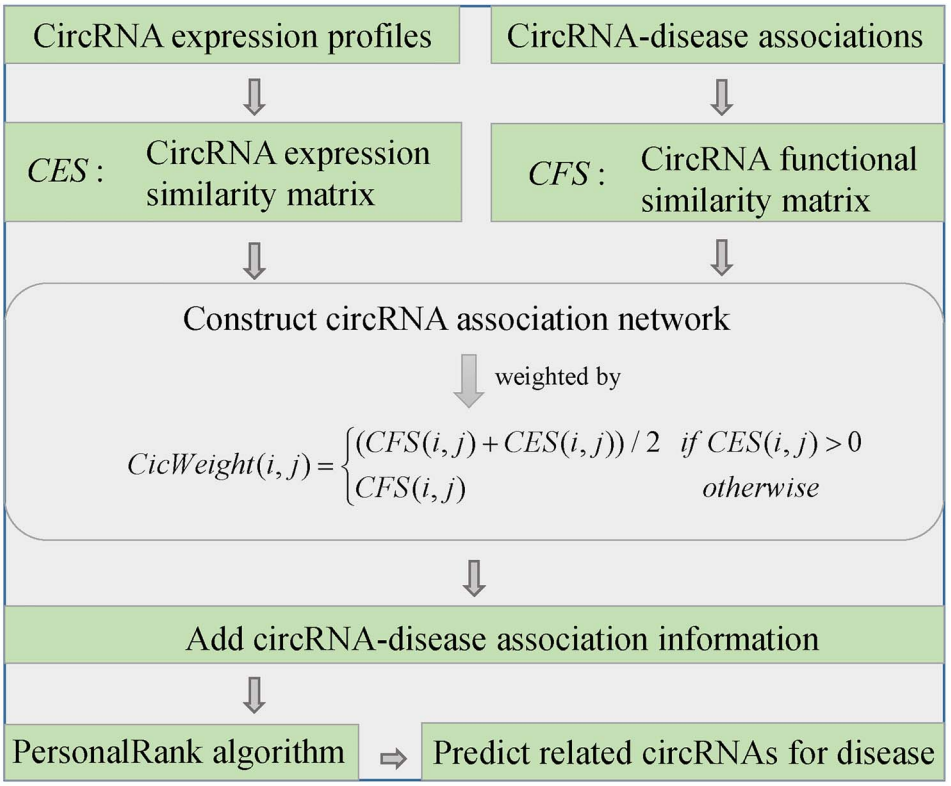

Figure 11. The workflow of PreCDA to infer underling circRNA-disease associations based on PersonalRank algorithm.

limitation of PreCDA lies in the invalid application for disease without any known related circRNAs.

ICFCDA

Lei *et al.* [116] raised an improved collaboration filtering recommendation system-based model named ICFCDA to predict circRNA-disease associations (see Figure 12). They construct circRNA similarity matrix *CS* by integrating circRNA functional annotation semantic similarity, circRNA sequence similarity as well as circRNA GIP kernel similarity. Besides, the disease similarity matrix *DS* can be obtained by integrating disease functional similarity, disease semantic similarity and disease GIP kernel similarity. To calculate recommendation score between circRNA c_i and disease d_j, the top k similar neighbors $N(c_i)$ of c_i and the top k similar neighbors $N(d_j)$ of disease d_j are selected according to similarity matrices of circRNA and disease. Then, circRNA-based recommendation score between c_i and d_j can be computed based on the matrices of *CD* and *CS* as follows:

$$CRS(i,j)=\frac{1}{k}\left(\sum_{c_n\in N(c_i)} CD(n,j)\times CS(n,i)\right) \tag{59}$$

Similarly, disease-based recommendation score between c_i and d_j is defined as follows:

$$DRS(i,j)=\frac{1}{k}\left(\sum_{d_n\in N(d_j)} CD(i,n)\times DS(n,j)\right) \tag{60}$$

Finally, the two recommendation scores are integrated as the predicted association score between c_i and d_j as follows:

$$AS(i,j)=\lambda DRS(i,j)+(1-\lambda)\,CRS(i,j) \tag{61}$$

where the parameter λ is a balance factor.

The second type of machine learning-based models

RWRKNN

Lei *et al.* [117] put forward a method named Random Walk with Restart and KNNs (RWRKNN) (see Figure 13) to predict novel circRNA-disease associations. Firstly, they construct disease similarity matrix *DS* by integrating disease semantic similarity and GIP kernel similarity, and circRNA similarity matrix *CS* by integrating circRNA functional similarity and GIP kernel similarity. The matrices of *DS* and *CS* are considered to be the feature matrices of disease and circRNA. Secondly, the matrices of *DA* and *CA* are utilized to represent disease-disease association network and circRNA-circRNA association network, respectively. These two matrices can be defined as follows:

$$DA(i,j)=\begin{cases}1 & \text{if } DS(i,j)\geq\alpha\\ 0 & \text{otherwise}\end{cases} \tag{62}$$

$$CA(i,j)=\begin{cases}1 & \text{if } CS(i,j)\geq\beta\\ 0 & \text{otherwise}\end{cases} \tag{63}$$

where α and β are different threshold values.

Thirdly, the affinity scores between a disease (circRNA) node and all disease (circRNA) nodes can be calculated by utilizing

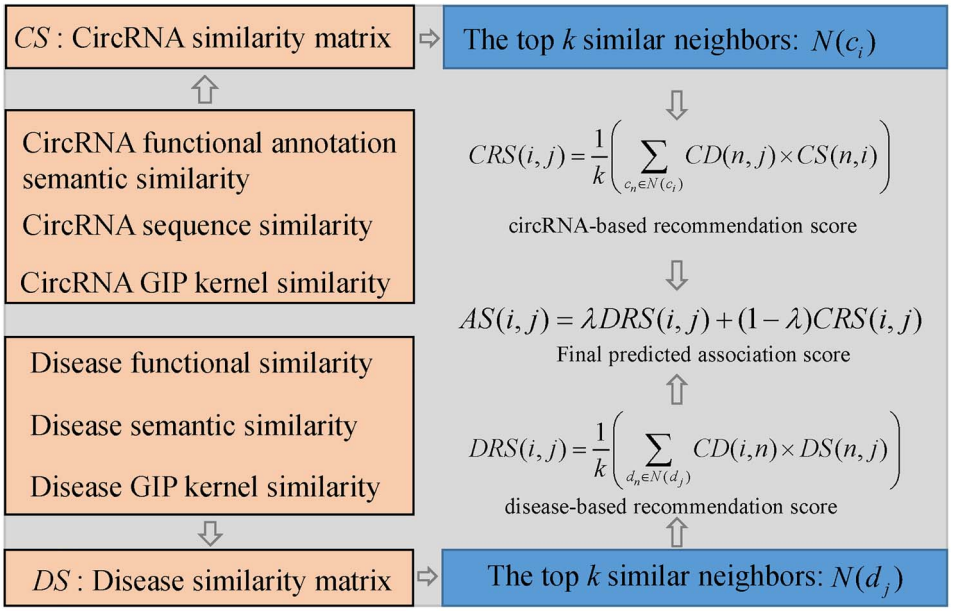

Figure 12. The workflow of ICFCDA to predict circRNA-disease associations based on improved collaboration filtering recommendation system.

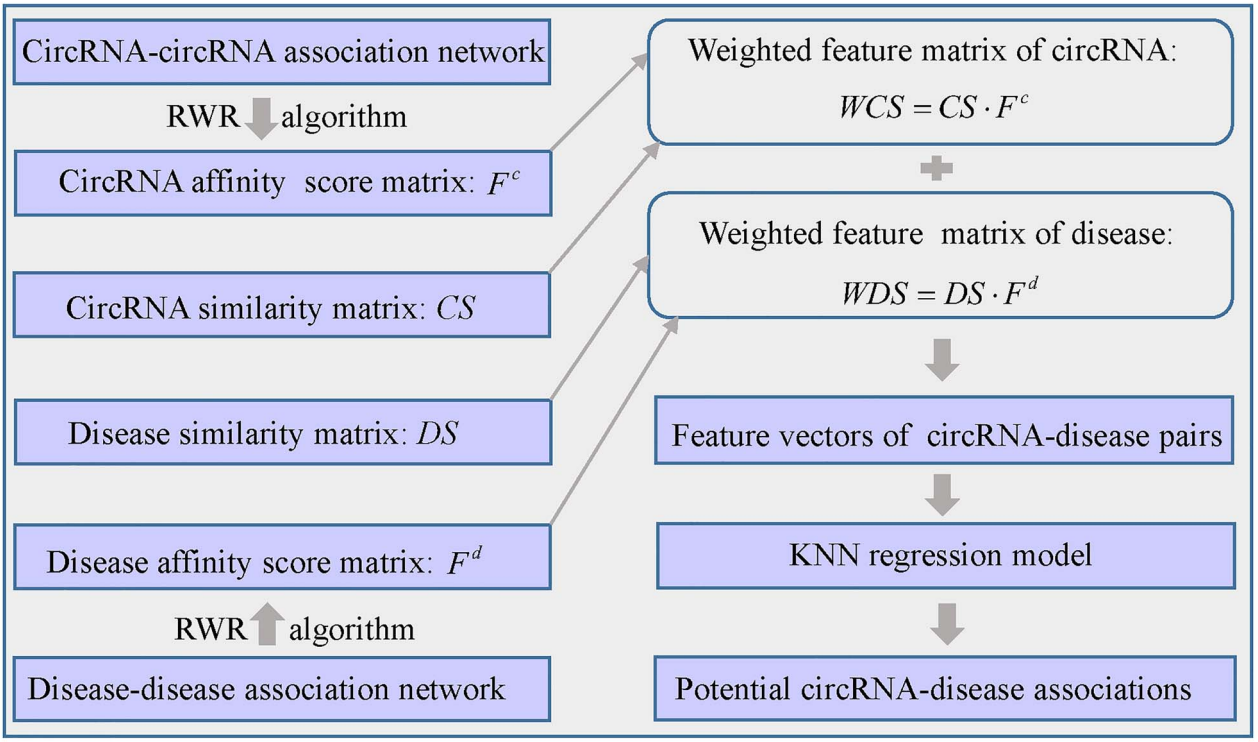

Figure 13. The workflow of RWRKNN to predict novel circRNA-disease associations based on random walk with restart and KNN.

RWR algorithm on the disease-disease (circRNA-circRNA) association network. The matrices of F^c and F^d denote the affinity scores for circRNA and disease, respectively. Next, the weighted feature matrices of circRNA and disease, namely *WCS* and *WDS*, are defined as follows:

$$WCS = CS \cdot F^c \tag{64}$$

$$WDS = DS \cdot F^d \tag{65}$$

The feature vectors of circRNA-disease pairs can be obtained by splicing the row vector of *WCS* and *WDS*. Finally, KNN regression model is adopted to predict potential circRNA-disease associations.

iCDA-CGR

Zheng *et al.* [118] proposed the method of identification of CircRNA-Disease Associations based on Chaos Game Representation (iCDA-CGP). The matrix of *DS* is constructed by integrating disease semantic similarity and GIP kernel similarity, while the matrix *CS* is constructed by integrating circRNA-related gene-based similarity, circRNA sequence-based similarity and circRNA GIP kernel similarity. The model of iCDA-CGP can be roughly divided into three steps. First of all, they construct training sample set including the same number of positive and negative samples. The positive samples are gathered from benchmark database of circRNA-disease associations, while the negative samples are selected from unlabeled circRNA-disease pairs. Secondly, the descriptor of each circRNA-disease pair in the training sample set can be formed based on the matrices of *CS* and *DS*

$$F\left(d_i, c_j\right) = \left(DS\left(i,:\right), CS\left(j,:\right)\right) \tag{66}$$

where $F(d_i, c_j)$ denotes the descriptor of the pair of d_i and c_j. Besides, $DS(i,:)$ and $CS(j,:)$ are the ith row of *DS* and the jth row of *CS*. Finally, based on SVM, the descriptors of training samples are utilized to train prediction model which is used to infer novel circRNA-disease associations. The model of iCDA-CGP has one main limitation, that is the negative samples used in the model are not reliable.

GBDTCDA

Lei *et al.* [119] developed a prediction model of GBDT with multiple biological data to predict CircRNA-Disease Association (GBDTCDA) (see Figure 14). Specifically, they compute circRNA sequence similarity, circRNA functional annotation semantic similarity as well as circRNA expression profile similarity, and combine them into the matrix *CD* by a similarity network fusion algorithm [120]. In addition, they integrate disease semantic and functional similarity as the matrix *DS* by endowing different weights for the two types of similarity. Secondly, four types of features of each circRNA-disease pair are extracted from the data of collected circRNA-disease associations, integrated similarity of circRNAs and diseases as well as circRNA nucleic acid sequence. The feature vector of the pair of circRNA c_i and disease d_j can be denoted as follows:

$$F\left(c_i, d_j\right) = \left[F_1\left(c_i, d_j\right), F_2\left(c_i, d_j\right), F_3\left(c_i, d_j\right), F_4\left(c_i, d_j\right)\right] \tag{67}$$

where F_i represents the ith type of features. Finally, they utilize GBDT regression to train the training samples and obtain predictive model for potential circRNA-disease association identification. In GBDTCDA, the authors make full use of multiple biological data and extract various kind of features, which facilitates the reliable performance of GBDTCDA.

DFPUCDA

Zeng *et al.* [121] raised a computational model of DF combined with Positive-Unlabeled learning based CircRNA-Disease Association prediction (DFPUCDA). In the first step of DFPUCDA, the authors construct a heterogeneous biological network, which contains a disease similarity network, a miRNA functional similarity network, a circRNA co-expression network, a miRNA-circRNA interaction network and a miRNA-disease association network. Then, they extract 24 meta-path-based features to represent circRNA-disease samples by PathCount and RandomWalk measures [122, 123]. Next, a positive-unlabeled learning algorithm is exploited to select reliable negative samples from unlabeled samples. Subsequently, DF algorithm is employed to train a classifier with collected positive samples and reliable negative samples. Finally, they utilize the classifier to infer positive circRNA-disease samples. It is difficult to obtain negative circRNA-disease samples and the number of positive samples is far less than that of unlabeled samples. In DFPUCDA, the positive-unlabeled algorithm can make full use of the information of unlabeled samples and solve the problem of data imbalance to some extent.

CNNCDA

Wang *et al.* [124] put forward a CNN-based method to predict CircRNA-Disease Associations (CNNCDA). Firstly, they construct the matrix *DS* through merging disease semantic similarity and disease GIP kernel similarity. Besides, the matrix *CS* is constructed based on circRNA GIP kernel similarity. Secondly, the authors define the circRNA-disease fusion descriptor $F(c_i, d_j)$ between circRNA c_i and disease d_j as follows:

$$F\left(c_i, d_j\right) = \left[CS\left(i,:\right), DS\left(j,:\right)\right] \tag{68}$$

where $CS(i,:)$ and $DS(j,:)$ denote the ith row and jth row of *CS* and *DS*, respectively.

Next, CNN, composed of input layer, convolution layer, subsampling layer, full connection layer and the output layer, is utilized to extract hidden deep features from circRNA-disease fusion descriptor. Finally, the extreme learning machine algorithm [125, 126] is used to train prediction model based on positive circRNA-disease samples and negative samples. However, the circRNA similarity is computed only based on known circRNA-disease associations, which would reduce the prediction performance.

GCNCDA

Wang *et al.* [127] further proposed a Graph Convolutional Network-based algorithm to infer CircRNA-Disease Associations (GCNCDA) whose flow diagram is shown in Figure 15. Firstly, the circRNA similarity matrix *CS* is constructed based on circRNA GIP kernel similarity, and the disease similarity matrix *DS* is constructed based on disease GIP kernel similarity and disease semantic similarity. Secondly, each circRNA-disease pair can be denoted by a feature descriptor which can be obtained in the same way as that in CNNCDA (i.e. Eq. (68)). Then, the Fast learning with Graph Convolutional Networks (FastGCN) [128] is utilized to further extract high-level features from original feature descriptors to construct new descriptors. Compared with GCN, FastGCN can make the training process more efficient. Next, the Forest by Penalizing Attributes (Forest PA) algorithm [129] is used to train classifier. Forest PA generates the training data set for trees by bootstrap sampling. The decision trees are built by using an improved CART algorithm [130]. The only difference between original CART algorithm and the improved CART algorithm is that the merit values is employed to instead of classification capacities (e.g. Gini Index) to select splitting attributes. Finally, the Forest PA classifier can be used to predict potential circRNA-disease associations.

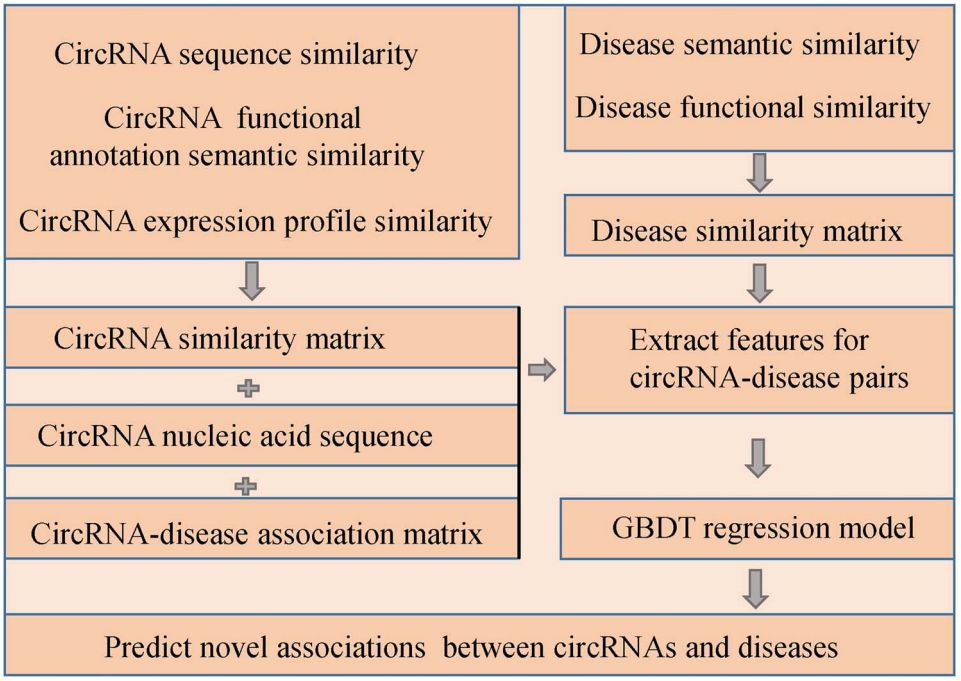

Figure 14. The workflow of GBDTCDA to predict potential circRNA-disease associations based on GBDT algorithm.

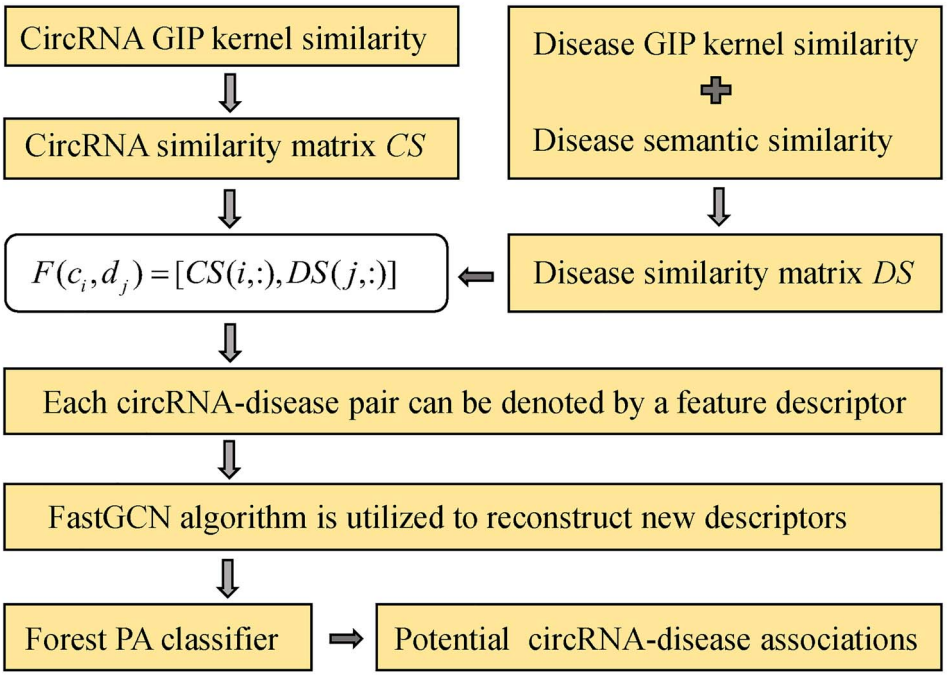

Figure 15. The flow diagram of GCNMDA to predict potential circRNA-disease associations based on Graph Convolutional Network.

AE-DNN

Deepthi *et al.* [131] devised an ensemble method to predict circRNA-disease associations based on AutoEncoder and DNN (AE-DNN). First, the circRNA similarity matrix is constructed by integrating circRNA sequence similarity and circRNA GIP similarity, while the disease similarity matrix *DS* is computed by integrating disease semantic similarity as well as disease GIP similarity. Then, they construct training sample set which contains both positive and negative samples. The positive samples are obtained from the CircR2Disease database and the negative samples are randomly selected from unlabeled circRNA-disease pairs. For each training sample (c_i, d_j), the feature vector is the splicing of the vectors of $CS(i,:)$ and $DS(j,:)$. Next, the autoencoder consisting of encoder and decoder is utilized to extract the

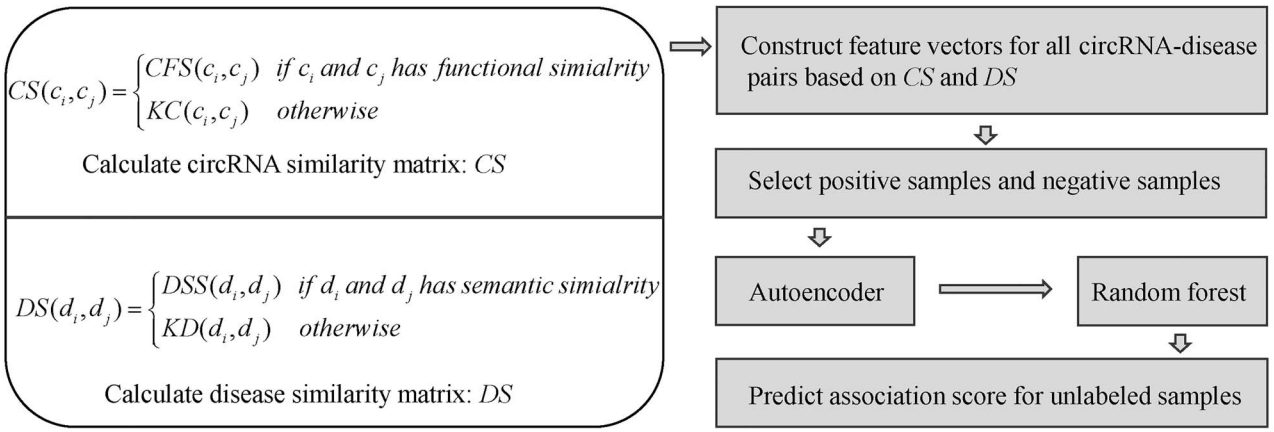

Figure 16. The flow diagram of AE-RF to predict potential circRNA-disease associations based on AutoEnconder and RF.

high-level features and reduce the dimension of feature vectors. Autoencoder [132] is a special neural network structure, which can learn the latent features of input data. Finally, the high-level feature vectors of training samples are used to train a three-layer feed-forward DNN. After training, the DNN can predict association probability for unlabeled circRNA-disease pair.

AE-RF

Deepthi *et al.* [133] proposed an ensemble method of circRNA-disease association prediction based on a deep AntoEncoder and RF classifier (AE-RF) whose flow diagram is shown in Figure 16. They first construct circRNA similarity matrix *CS* and disease similarity matrix *DS* by combing multiple types of similarity of circRNA and disease as follows:

$$CS(c_i,c_j)=\begin{cases}CFS(c_i,c_j) & \text{if } c_i \text{ and } c_j \text{ has functional simialrity}\\ KC(c_i,c_j) & \text{otherwise}\end{cases} \tag{69}$$

$$DS(d_i,d_j)=\begin{cases}DSS(d_i,d_j) & \text{if } d_i \text{ and } d_j \text{ has semantic simialrity}\\ KD(d_i,d_j) & \text{otherwise}\end{cases} \tag{70}$$

Then, the feature vector of circRNA-disease pair (c_i, d_j) is constructed by splicing the vector $CS(c_i,:)$ and vector $DS(d_j,:)$. Next, the training set consisting of equal positive and negative samples is utilized to train an autoencoder which is also used in the prediction model of AE-DNN. After training, the autoencoder can be used to reconstructed the feature vectors of samples in training set and remaining unlabeled circRNA-disease pairs. Subsequently, the training samples are utilized to train the RF classifier. The trained classifier can be used to predict association score for unlabeled samples. The innovative of this study lies in the combined application of autoencoder and RF where autoencoder can help reduce noise data and extract high-level features, while RF has good generalization ability. However, the false negative problem of randomly selected negative samples still exists.

Algorithm evaluation methods

To evaluate the predictive performance of computational models, researchers usually report their AUC values based on distinct cross validation including LOOCV, 5-fold and 10-fold cross validation (collectively called *K*-fold cross-validation). LOOCV and *K*-fold cross validation have been widely utilized to evaluate the performance of not only the circRNA-disease association prediction models but also other biological association prediction models, such as miRNA-disease association prediction models [92, 99, 134], lncRNA-disease association prediction models [95, 135], lncRNA-miRNA interaction and lncRNA-protein interaction prediction models [136–138]. In this section, we will introduce LOOCV and *K*-fold cross validation in detail. In addition to cross validation, we also introduced two types of case studies, which have been frequently utilized to evaluate the prediction performance of different circRNA-disease association prediction algorithms.

LOOCV

In the process of LOOCV, each known circRNA-disease association is left out as the test sample in turn, and the remaining known associations are adopted as training samples. In addition, all unknown circRNA-disease pairs are candidate samples. Specifically, the prediction model based on the training samples can score for the investigated test sample and all candidate samples. Then, the test sample and candidate samples are ranked in descending order according to their association scores. Above process is repeated until every known circRNA-disease association is tested. According to the results of LOOCV, true positive rate (*TPR*) and false positive rate (*FPR*) can be calculated as follows:

$$TPR=\frac{TP}{TP+FN} \tag{71}$$

$$FPR=\frac{FP}{FP+TN} \tag{72}$$

where *TP* denotes the number of true positive samples which are test samples ranked higher than the given threshold; *FN* denotes the number of false negative samples, which are test samples ranked lower than the given threshold. In addition, *FP* represents the number of false positive samples, which are candidate samples ranked higher than the given threshold; *TN* represents the number of true negative sample, which are candidate samples ranked lower than the given threshold. The ROC (receiver operating characteristic) curve can be drawn by plotting the *TPR* against the *FPR* under a series of thresholds. Furthermore, the value of AUC can demonstrate the performance of prediction model and the higher the AUC, the better the prediction performance of the model.

K-fold cross validation

In *K*-fold cross validation, all known circRNA-disease associations are divided into *K* subsets with the same size. Then, one of the *K* subsets is left out as the test set and the remaining $K-1$ subsets are utilized as training set to train the prediction model. All unknown circRNA-disease pairs are candidate samples. The trained prediction model can score for the samples in the test set and candidate samples. Next, each sample in the test set is ranked with the candidate samples in descending order according to their association scores. When all the *K* subsets have been tested, the ROC curve and AUC value can be drawn and calculated in the same way used in the LOOCV.

Case study

Usually, one or several diseases would be investigated in case study. In addition, the types of case studies are also diverse. In the following, we will introduce two common types of case studies utilized to evaluate predictive performance of circRNA-disease association prediction model. The first type of case study aims to assess the prediction ability of calculation model in identifying novel circRNA-disease relationships [85, 102]. Specifically, the trained prediction model is used to compute the association scores for candidate samples involving investigated disease. Then, the result of case study for investigated disease can be obtained by inspecting how many associations in the top-*M* predicted results have been confirmed by other database or literature. The second type of case study aims to evaluate the prediction ability of calculation model in predicting associated circRNAs for novel disease without any known related circRNAs [86, 117]. To be more specific, the association information involving an investigated disease is removed from training sample set. Then, the trained model is utilized to infer associated circRNAs for this investigated disease. Finally, researches observe how many circRNAs in the top-ranked predictions have been confirmed by database or literature.

Discussion and conclusion

CircRNAs have caught much attention from scientists. More and more circRNAs were discovered by biological experiments and bioinformatics methods. Later, researchers found that circRNAs have important biological functions including acting as miRNA sponges, regulating the expression of parental genes as well as competing with pre-mRNA splicing. In addition, many experimental evidences indicate that circRNAs have close relationships with complex human diseases. The occurrence and development of many complex diseases are usually accompanied by abnormal expression of circRNA. Thus, studying associations between circRNAs and diseases could promote the understanding of the functions of circRNAs and the pathogenesis of complex diseases, which would further provide new ideas and strategies for detection, diagnosis and treatment of complex diseases. Identifying novel circRNA-disease associations is a critical step. However, it is inefficient to discover novel associations by traditionally experimental methods. Fortunately, massive biological data about circRNAs and circRNA-disease associations have been accumulated after conducting various biological experiments and RNA sequencing. Therefore, researchers have proposed effective computational methods to predict novel circRNA-disease relationships by mining useful information from biological data such as circRNA sequence, circRNA expression profile, disease directed acyclic graph, circRNA-gene interaction, disease-gene association and circRNA-disease association.

In this review, we first briefly summarized the general concepts and classification of circRNAs. Then, we introduced some common functions of circRNAs and associations between circRNAs and several important human diseases, since circRNAs may be a novel classes of biomarkers of complex diseases. Next, we presented two types of databases which can provide biological data about circRNAs and circRNA-disease associations. Proper application of these databases can promote the research of circRNA function and identification of novel circRNA-disease associations. Subsequently, we introduced 27 computational models for inferring novel circRNA-disease associations. According to the core algorithms used in these models, we divided the computational models into two classes, namely network algorithm-based models and machine learning-based models. Finally, we summarized several common measures for performance evaluation of circRNA-disease association prediction models.

In the following, we will discuss the advantages and limitations of aforementioned two types of computational models. First of all, in the network algorithm-based models, it is a key step to construct the circRNA-disease associations network, circRNA similarity network and disease similarity network. Generally, circRNA-circRNA similarity can be calculated based on circRNA sequences, circRNA-related genes, expression profiles of circRNAs- and circRNA-related diseases. In addition, disease-disease similarity can be computed based on disease related genes, phenotype descriptions of diseases, directed acyclic graphs of diseases and disease-related circRNAs. The different network algorithms, such as KATZ, label propagation and bipartite network projection, were utilized to infer novel circRNA-disease associations based on these networks. One advantage of network algorithm lies that these models can integrate multiple biological data to construct single layer network or heterogeneous network and make full use of topological information of circRNA-disease network. In addition to circRNA and disease, other biological object can also be introduced into heterogeneous networks. For example, in the model of BRWSP, the authors introduced gene similarity network, gene-disease association network and gene-circRNA interaction network into their constructed heterogeneous network. Another advantage of network algorithm lies in the wide choice for similarity calculation methods. Except for the full use of multiple data, similarity calculation method also plays an important role in network algorithm-based models. For example, in the model of CD-LNLP, the authors utilized LNS measure to calculate circRNA similarity and disease similarity. As a result, CD-LNLP obtains impressive performance even though only circRNA-disease association data are used to calculate similarity. Therefore, reliable similarity calculation method would contribute to the predictive performance of network algorithm-based models. However, most of network algorithm-based models cannot predict associations for diseases without any known related circRNAs. Besides, it is difficult to determine the weights of distinct types of similarity in the process of similarity integration. Therefore, how to construct different circRNA similarity networks and disease similarity networks, and reasonably integrate the similarity from different biological source information is an important topic worthy of further study.

Machine learning-based circRNA-disease association prediction models could be further divided into two classes. Specifically, regularized least squares, logistic regression and manifold regularization learning, matrix decomposition and inductive matrix completion algorithm-based calculation methods belong

to the first category, which usually transform the problem of circRNA-disease association prediction into solving diverse optimization models based on circRNA-disease adjacency matrix, circRNA similarity matrix and disease similarity matrix. One advantage of the first class of machine learning-based models is that negative samples are not necessary. Actually, negative circRNA-disease associations are hard to collect due to the fact that experimentally validated negative circRNA-disease relationships are usually not reported in literature or database. Besides, different regulation terms can be added into objective functions of the first types of machine learning-based models. For example, in the models of MRLDC and iCircRA-MF, graph regularization term is introduced into their objective functions to restrict the geometrical structure of target space and reduce noise. However, the parameters in the objective functions are hard to determine. In addition, how to choose suitable optimization algorithm to solve different objective functions is worth considering. In the second type of machine learning-based circRNA-disease association prediction models, the algorithms of KNN, SVM, RF, GBDT, DF, CNN, GNN and DNN are utilized to construct different classifiers. Besides, distinct feature construction methods are employed in the second type of machine learning-based models. One advantage of these models lies that they could make full use of the prior information of known circRNA-disease associations since all know positive samples are utilized to train the prediction models. In addition, most of the second type of machine learning-based models can be employed to predict associated circRNAs for novel disease without any known related circRNAs. However, negative samples are necessary in these prediction models. As mentioned above, negative circRNA-disease samples are difficult to collect and randomly selecting unlabeled samples as negative samples is a common strategy in these models, which would reduce the prediction accuracy to some extent. Furthermore, the second type of machine learning-based models belong to supervised learning models, so the class imbalance problem of circRNA-disease samples is one of main obstacles in these prediction models. Semi-supervised learning methods work well dealing with the class-imbalance data. Therefore, researchers can utilize semi-supervised learning algorithms to establish new prediction models in the future.

Overall, circRNA plays an important role in the development of various complex diseases and is a novel biomarker of complex diseases. Accumulation of experimental data about circRNAs and diseases makes it possible to predict new circRNA-disease associations by computational methods. However, the number of current known circRNA-diseases associations is too less, which limits the predictive accuracy of existing computational models. Thus, collection and accumulation of experimentally verified circRNA-disease associations remains an important mission in the future study. Besides, researchers can consider utilizing the information of other biological objects, such as pathway and protein, to help circRNA-disease association prediction, since biological objects are usually closely interdependent. In terms of calculation model, new effective algorithms should be proposed since the current methods have different limitations. In this paper, we mainly reviewed the research of circRNA-disease association from distinct aspects. Actually, the studies of miRNA-disease association and lncRNA-disease association are also hot research fields [134, 135]. MiRNAs and lncRNAs also play important roles in the occurrence and development of many human diseases. However, the studies of associations between circRNAs, miRNAs, lncRNAs and human diseases were conducted independently. The joint research of associations between circRNAs, miRNAs, lncRNAs and human diseases may be an important future direction. In the end, scientists have demonstrated that non-coding RNA can be of drug targets [101]. Specially, some works have been implemented to identify miRNAs as drug targets [139–141]. CircRNA is also an important class of non-coding RNA. Therefore, identifying circRNAs as drug targets could be a promising future direction.

Key Points

- CircRNAs play a growing important role in a large number of life activities and are thus closely related to various human complex diseases.
- Studying associations between circRNAs and diseases could promote the understanding of the functions of circRNAs and the pathogenesis of complex diseases.
- We listed some publicly accessible databases about circRNAs and circRNA-disease associations.
- Computational models could effectively predict potential circRNA-disease associations for further experimental verification, which would save many resources.
- Computational models of circRNA-disease prediction were divided into two categories, namely network algorithm and machine learning-based model.
- We introduced several methods of algorithm evaluation to estimate the predictive performance of calculation models.
- The advantages and limitations of various existing computational models were analyzed.

Data availability

The source code of SIMCCDA is available at https://github.com/bioinformaticsAHU/SIMCCDA. The source code of PreCDA is available at https://github.com/wyt-nwpu/PreCDA. The source code of DFPUCDA is available at https://github.com/xzenglab/DeepDCR. The source code of AE-RF is available at https://github.com/Deepthi-K523/AE-RF.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib

Funding

National Natural Science Foundation of China (grant no. 61972399 and 11931008 to X.C.).

References

1. Kristensen LS, Andersen MS, Stagsted LVW, *et al*. The biogenesis, biology and characterization of circular RNAs. *Nat Rev Genet* 2019; **20**: 675–91.
2. Sanger HL, Klotz G, Riesner D, *et al*. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proc Natl Acad Sci USA* 1976; **73**: 3852–6.
3. Gross HJ, Domdey H, Lossow C, *et al*. Nucleotide sequence and secondary structure of potato spindle tuber viroid. *Nature* 1978; **273**: 203–8.

4. Hsu MT, Coca-Prados M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature* 1979; **280**: 339–40.
5. Grabowski PJ, Zaug AJ, Cech TR. The intervening sequence of the ribosomal RNA precursor is converted to a circular RNA in isolated nuclei of Tetrahymena. *Cell* 1981; **23**: 467–76.
6. Danan M, Schwartz S, Edelheit S, *et al*. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res* 2012; **40**: 3131–42.
7. Cocquerelle C, Mascrez B, Hetuin D, *et al*. Mis-splicing yields circular RNA molecules. *FASEB J* 1993; **7**: 155–60.
8. Lu T, Cui L, Zhou Y, *et al*. Transcriptome-wide investigation of circular RNAs in rice. *RNA* 2015; **21**: 2076–87.
9. Broadbent KM, Broadbent JC, Ribacke U, *et al*. Strand-specific RNA sequencing in *Plasmodium falciparum* malaria identifies developmentally regulated long non-coding RNA and circular RNA. *BMC Genomics* 2015; **16**: 454.
10. Wang PL, Bao Y, Yee MC, *et al*. Circular RNA is expressed across the eukaryotic tree of life. *PLoS One* 2014; **9**: e90859.
11. Barrett SP, Wang PL, Salzman J. Circular RNA biogenesis can proceed through an exon-containing lariat precursor. *Elife* 2015; **4**: e07540.
12. Zhang XO, Wang HB, Zhang Y, *et al*. Complementary sequence-mediated exon circularization. *Cell* 2014; **159**: 134–47.
13. Ivanov A, Memczak S, Wyler E, *et al*. Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell Rep* 2015; **10**: 170–7.
14. Fang Y. Circular RNAs as novel biomarkers with regulatory potency in human diseases. *Future Sci OA* 2018; **4**: Fso314.
15. Starke S, Jost I, Rossbach O, *et al*. Exon circularization requires canonical splice signals. *Cell Rep* 2015; **10**: 103–11.
16. Li Z, Huang C, Bao C, *et al*. Exon-intron circular RNAs regulate transcription in the nucleus. *Nat Struct Mol Biol* 2015; **22**: 256–64.
17. Zhang Y, Zhang XO, Chen T, *et al*. Circular intronic long noncoding RNAs. *Mol Cell* 2013; **51**: 792–806.
18. Guo JU, Agarwal V, Guo H, *et al*. Expanded identification and characterization of mammalian circular RNAs. *Genome Biol* 2014; **15**: 409.
19. Kelly S, Greenman C, Cook PR, *et al*. Exon Skipping Is Correlated with Exon Circularization. *J Mol Biol* 2015; **427**: 2414–7.
20. Jeck WR, Sorrentino JA, Wang K, *et al*. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 2013; **19**: 141–57.
21. Memczak S, Jens M, Elefsinioti A, *et al*. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013; **495**: 333–8.
22. Liang D, Wilusz JE. Short intronic repeat sequences facilitate circular RNA production. *Genes Dev* 2014; **28**: 2233–47.
23. Chen LL. The biogenesis and emerging roles of circular RNAs. *Nat Rev Mol Cell Biol* 2016; **17**: 205–11.
24. Hansen TB, Jensen TI, Clausen BH, *et al*. Natural RNA circles function as efficient microRNA sponges. *Nature* 2013; **495**: 384–8.
25. Zheng Q, Bao C, Guo W, *et al*. Circular RNA profiling reveals an abundant circHIPK3 that regulates cell growth by sponging multiple miRNAs. *Nat Commun* 2016; **7**: 11215.
26. Li F, Zhang L, Li W, *et al*. Circular RNA ITCH has inhibitory effect on ESCC by suppressing the Wnt/β-catenin pathway. *Oncotarget* 2015; **6**: 6001–13.
27. Wang K, Long B, Liu F, *et al*. A circular RNA protects the heart from pathological hypertrophy and heart failure by targeting miR-223. *Eur Heart J* 2016; **37**: 2602–11.
28. Wan L, Zhang L, Fan K, *et al*. Circular RNA-ITCH suppresses lung cancer proliferation via inhibiting the Wnt/beta-catenin pathway. 2016; **2016**: 1579490.
29. Ashwal-Fluss R, Meyer M, Pamudurti NR, *et al*. circRNA biogenesis competes with pre-mRNA splicing. *Mol Cell* 2014; **56**: 55–66.
30. Khodor YL, Menet JS, Tolan M, *et al*. Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. *RNA* 2012; **18**: 2174–86.
31. Abdelmohsen K, Panda AC, Munk R, *et al*. Identification of HuR target circular RNAs uncovers suppression of PABPN1 translation by CircPABPN1. *RNA Biol* 2017; **14**: 361–9.
32. Yang Y, Gao X, Zhang M, *et al*. Novel role of FBXW7 circular RNA in repressing glioma tumorigenesis. *J Natl Cancer Inst* 2018; **110**: 304–15.
33. Zhao Z, Wang K, Wu F, *et al*. circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. *Cell Death Dis* 2018; **9**: 475.
34. Zhang M, Zhao K, Xu X, *et al*. A peptide encoded by circular form of LINC-PINT suppresses oncogenic transcriptional elongation in glioblastoma. *Nat Commun* 2018; **9**: 4475.
35. Rawla P, Barsouk A. Epidemiology of gastric cancer: global trends, risk factors and prevention. *Prz Gastroenterol* 2019; **14**: 26–38.
36. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin* 2019; **69**: 7–34.
37. Li R, Jiang J, Shi H, *et al*. CircRNA: a rising star in gastric cancer. *Cell Mol Life Sci* 2020; **77**: 1661–80.
38. Li T, Shao Y, Fu L, *et al*. Plasma circular RNA profiling of patients with gastric cancer and their droplet digital RT-PCR detection. *J Mol Med (Berl)* 2018; **96**: 85–96.
39. Schwarz B-A, Bange R, Vahlenkamp TW, *et al*. Detection and quantitation of group A rotaviruses by competitive and real-time reverse transcription-polymerase chain reaction. *J Virol Methods* 2002; **105**: 277–85.
40. Rački N, Morisset D, Gutierrez-Aguirre I, *et al*. One-step RT-droplet digital PCR: a breakthrough in the quantification of waterborne RNA viruses. *Anal Bioanal Chem* 2014; **406**: 661–7.
41. Huang YS, Jie N, Zou KJ, *et al*. Expression profile of circular RNAs in human gastric cancer tissues. *Mol Med Rep* 2017; **16**: 2469–76.
42. Lai Z, Yang Y, Yan Y, *et al*. Analysis of co-expression networks for circular RNAs and mRNAs reveals that circular RNAs hsa_circ_0047905, hsa_circ_0138960 and has-circRNA7690-15 are candidate oncogenes in gastric cancer. *Cell Cycle* 2017; **16**: 2301–11.
43. Waks AG, Winer EP. Breast cancer treatment: a review. *JAMA* 2019; **321**: 288–300.
44. Tang YY, Zhao P, Zou TN, *et al*. Circular RNA hsa_circ_0001982 promotes breast cancer cell carcinogenesis through decreasing miR-143. *DNA Cell Biol* 2017; **36**: 901–8.
45. Chen X, Yin J, Qu J, *et al*. MDHGI: matrix decomposition and heterogeneous graph inference for miRNA-disease association prediction. *PLoS Comput Biol* 2018; **14**: e1006418.
46. He R, Liu P, Xie X, *et al*. circGFRA1 and GFRA1 act as ceRNAs in triple negative breast cancer by regulating miR-34a. *J Exp Clin Cancer Res* 2017; **36**: 145.
47. Komaroff AL. Harvard medical school family health guide. New York, NY: *Simon and Schuster* 2005.

48. Ezzati M, Henley SJ, Lopez AD, *et al*. Role of smoking in global and regional cancer epidemiology: current patterns and data needs. *Int J Cancer* 2005; **116**: 963–71.
49. Alberg AJ, Samet JM. Epidemiology of lung cancer. *Chest* 2003; **123**: 21s–49.
50. Zhu X, Wang X, Wei S, *et al*. hsa_circ_0013958: a circular RNA and potential novel biomarker for lung adenocarcinoma. *FEBS J* 2017; **284**: 2170–82.
51. Yao JT, Zhao SH, Liu QP, *et al*. Over-expression of CircRNA_100876 in non-small cell lung cancer and its prognostic value. *Pathol Res Pract* 2017; **213**: 453–6.
52. Chu LC, Goggins MG, Fishman EK. Diagnosis and detection of pancreatic cancer. *Cancer J* 2017; **23**: 333–42.
53. Walter FM, Mills K, Mendonça SC, *et al*. Symptoms and patient factors associated with diagnostic intervals for pancreatic cancer (SYMPTOM pancreatic study): a prospective cohort study. *Lancet Gastroenterol Hepatol* 2016; **1**: 298–306.
54. Pham A, Forsmark C. Chronic pancreatitis: review and update of etiology, risk factors, and management. *F1000Res* 2018; **7**: F1000 Faculty Rev-607.
55. Guo S, Xu X, Ouyang Y, *et al*. Microarray expression profile analysis of circular RNAs in pancreatic cancer. *Mol Med Rep* 2018; **17**: 7661–71.
56. Chen G, Shi Y, Zhang Y, *et al*. CircRNA_100782 regulates pancreatic carcinoma proliferation through the IL6-STAT3 pathway. *Onco Targets Ther* 2017; **10**: 5783–94.
57. Shang X, Li G, Liu H, *et al*. Comprehensive circular RNA profiling reveals that hsa_circ_0005075, a new circular RNA biomarker, is involved in hepatocellular crcinoma development. *Medicine (Baltimore)* 2016; **95**: e3811.
58. Chen X, Sun Y-Z, Liu H, *et al*. RNA methylation and diseases: experimental results, databases, web servers and computational models. *Brief Bioinform* 2017; **20**: 896–917.
59. Ghosal S, Das S, Sen R, *et al*. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front Genet* 2013; **4**: 283.
60. Yao D, Zhang L, Zheng M, *et al*. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Sci Rep* 2018; **8**: 11018.
61. Fan C, Lei X, Fang Z, *et al*. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. *Database (Oxford)* 2018; **2018**: bay044.
62. Rophina M, Sharma D, Poojary M, *et al*. Circad: a comprehensive manually curated resource of circular RNA associated with diseases. *Database (Oxford)* 2020; **2020**: baaa019.
63. Glažar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. *RNA* 2014; **20**: 1666–70.
64. Liu YC, Li JR, Sun CH, *et al*. CircNet: a database of circular RNAs derived from transcriptome sequencing data. *Nucleic Acids Res* 2016; **44**: D209–15.
65. Zheng LL, Li JH, Wu J, *et al*. deepBase v2.0: identification, expression, evolution and function of small RNAs, LncRNAs and circular RNAs from deep-sequencing data. *Nucleic Acids Res* 2016; **44**: D196–202.
66. Chen X, Han P, Zhou T, *et al*. circRNADb: a comprehensive database for human circular RNAs with protein-coding annotations. *Sci Rep* 2016; **6**: 34985.
67. Xia S, Feng J, Lei L, *et al*. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes. *Brief Bioinform* 2017; **18**: 984–92.
68. Xia S, Feng J, Chen K, *et al*. CSCD: a database for cancer-specific circular RNAs. *Nucleic Acids Res* 2018; **46**: D925–d929.
69. Dong R, Ma XK, Li GW, *et al*. CIRCpedia v2: an updated database for comprehensive circular RNA annotation and expression comparison. *Genomics Proteomics Bioinformatics* 2018; **16**: 226–33.
70. Li S, Li Y, Chen B, *et al*. exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes. *Nucleic Acids Res* 2018; **46**: D106–d112.
71. Meng X, Hu D, Zhang P, *et al*. CircFunBase: a database for functional circular RNAs. *Database (Oxford)* 2019; **2019**: baz003.
72. Tang Z, Li X, Zhao J, *et al*. TRCirc: a resource for transcriptional regulation information of circRNAs. *Brief Bioinform* 2019; **20**: 2327–33.
73. Liu M, Wang Q, Shen J, *et al*. Circbank: a comprehensive database for circRNA with standard nomenclature. *RNA Biol* 2019; **16**: 899–905.
74. Ruan H, Xiang Y, Ko J, *et al*. Comprehensive characterization of circular RNAs in ∼1000 human cancer cell lines. *Genome Med* 2019; **11**: 55.
75. Vo JN, Cieslik M, Zhang Y, *et al*. The landscape of circular RNA in cancer. *Cell* 2019; **176**: 869–881.e813.
76. Cai Z, Fan Y, Zhang Z, *et al*. VirusCircBase: a database of virus circular RNAs. *Brief Bioinform* 2021; **22**: 2182–90.
77. Lei X, Fang Z, Chen L, *et al*. PWCDA: path weighted method for predicting circRNA-disease associations. *Int J Mol Sci* 2018; **19**: 3410.
78. You ZH, Huang ZA, Zhu Z, *et al*. PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. *PLoS Comput Biol* 2017; **13**: e1005455.
79. Lei X, Zhang W. BRWSP: predicting circRNA-disease associations based on biased random walk to search paths on a multiple heterogeneous network. *Complexity* 2019; **2019**: 5938035.
80. Fan C, Lei X, Wu FX. Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks. *Int J Biol Sci* 2018; **14**: 1950–9.
81. Chen X, Huang YA, You ZH, *et al*. A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. *Bioinformatics* 2017; **33**: 733–9.
82. Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction. *Sci Rep* 2015; **5**: 16840.
83. Deng L, Zhang W, Shi Y, *et al*. Fusion of multiple heterogeneous networks for predicting circRNA-disease associations. *Sci Rep* 2019; **9**: 9605.
84. Zhang J, Zhang Z, Chen Z, *et al*. Integrating multiple heterogeneous networks for novel LncRNA-disease association inference. *IEEE/ACM Trans Comput Biol Bioinform* 2019; **16**: 396–406.
85. Zhao Q, Yang Y, Ren G, *et al*. Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations. *IEEE Trans Nanobioscience* 2019; **18**: 578–84.
86. Li G, Yue Y, Liang C, *et al*. NCPCDA: network consistency projection for circRNA–disease association prediction. *RSC Adv* 2019; **9**: 33222–8.
87. Li G, Luo J, Wang D, *et al*. Potential circRNA-disease association prediction using DeepWalk and network consistency projection. *J Biomed Inform* 2020; **112**: 103624.
88. Perozzi B, Al-Rfou R, Skiena S. 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: Association for Computing Machinery, 2014, 701–10.
89. Ge E, Yang Y, Gang M, *et al*. Predicting human disease-associated circRNAs based on locality-constrained linear coding. *Genomics* 2020; **112**: 1335–42.
90. Zhang W, Yu C, Wang X, *et al*. Predicting CircRNA-disease associations through linear neighborhood label propagation method. *IEEE Access* 2019; **7**: 83474–83.
91. Zhang W, Yue X, Huang F, *et al*. Predicting drug-disease associations and their therapeutic function based on the drug-disease association bipartite network. *Methods* 2018; **145**: 51–9.
92. Chen X, Sun L-G, Zhao Y. NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion. *Brief Bioinform* 2020; **22**: 485–96.
93. Chen X, Wang L, Qu J, *et al*. Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics* 2018; **34**: 4256–65.
94. Zhao Y, Chen X, Yin J. Adaptive boosting-based computational model for predicting potential miRNA-disease associations. *Bioinformatics* 2019; **35**: 4730–8.
95. Chen X, Yan G-Y. Novel human lncRNA–disease association inference based on lncRNA expression profiles. *Bioinformatics* 2013; **29**: 2617–24.
96. Chen X, Yan CC, Zhang X, *et al*. Drug-target interaction prediction: databases, web servers and computational models. *Brief Bioinform* 2016; **17**: 696–712.
97. Wang C-C, Zhao Y, Chen X. Drug-pathway association prediction: from experimental results to computational models. *Brief Bioinform* 2020; **22**: bbaa061.
98. Zhao Y, Wang C-C, Chen X. Microbes and complex diseases: from experimental results to computational models. *Brief Bioinform* 2020; **22**: bbaa158.
99. Chen X, Zhu C-C, Yin J. Ensemble of decision tree reveals potential miRNA-disease associations. *PLoS Comput Biol* 2019; **15**: e1007209.
100. Chen X, Huang L. LRSSLMDA: laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput Biol* 2017; **13**: e1005912.
101. Chen X, Guan N-N, Sun Y-Z, *et al*. MicroRNA-small molecule association identification: from experimental results to computational models. *Brief Bioinform* 2018; **21**: 47–61.
102. Yan C, Wang J, Wu FX. DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. *BMC Bioinformatics* 2018; **19**: 520.
103. Ding Y, Chen B, Lei X, *et al*. Predicting novel CircRNA-disease associations based on random walk and logistic regression model. *Comput Biol Chem* 2020; **87**: 107287.
104. Xiao Q, Luo J, Dai J. Computational prediction of human disease- associated circRNAs based on manifold regularization learning framework. *IEEE J Biomed Health Inform* 2019; **23**: 2661–9.
105. Nepusz T, Yu H, Paccanaro A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat Methods* 2012; **9**: 471–2.
106. Wei H, Liu B. iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Brief Bioinform* 2020; **21**: 1356–67.
107. Liu X, Zhai D, Zhao D, *et al*. Progressive image denoising through hybrid graph Laplacian regularization: a unified framework. *IEEE Trans Image Process* 2014; **23**: 1491–503.
108. Cai D, He X, Han J, *et al*. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans Pattern Anal Mach Intell* 2011; **33**: 1548–60.
109. Xiao Q, Yu H, Zhong J, *et al*. An in-silico method with graph-based multi-label learning for large-scale prediction of circRNA-disease associations. *Genomics* 2020; **112**: 3407–15.
110. Xiao Q, Zhong J, Tang X, *et al*. iCDA-CMG: identifying circRNA-disease associations by federating multi-similarity fusion and collective matrix completion. *Mol Genet Genomics* 2021; **296**: 223–33.
111. Wang S, Xia P, Zhang L, *et al*. Systematical identification of breast cancer-related circular RNA modules for deciphering circRNA functions based on the non-negative matrix factorization algorithm. *Int J Mol Sci* 2019; **20**: 919.
112. Wang L, Wang Y, Hu Q, *et al*. Systematic analysis of new drug indications by drug-gene-disease coherent subnetworks. *CPT Pharmacometrics Syst Pharmacol* 2014; **3**: e146.
113. Li M, Liu M, Bin Y, *et al*. Prediction of circRNA-disease associations based on inductive matrix completion. *BMC Med Genomics* 2020; **13**: 42.
114. Toh K-C, Yun S. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization* 2010; **6**: 615–40.
115. Wang Y, Nie C, Zang T, *et al*. Predicting circRNA-disease associations based on circRNA expression similarity and functional similarity. *Front Genet* 2019; **10**: 832.
116. Lei X, Fang Z, Guo L. Predicting circRNA-disease associations based on improved collaboration filtering recommendation system with multiple data. *Front Genet* 2019; **10**: 897.
117. Lei X, Bian C. Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease association. *Sci Rep* 2020; **10**: 1943.
118. Zheng K, You ZH, Li JQ, *et al*. iCDA-CGR: identification of CircRNA-disease associations based on chaos game representation. *PLoS Comput Biol* 2020; **16**: e1007872.
119. Lei X, Fang Z. GBDTCDA: Predicting circRNA-disease associations based on gradient boosting decision tree with multiple biological data fusion. *Int J Biol Sci* 2019; **15**: 2911–24.
120. Wang B, Mezlini AM, Demir F, *et al*. Similarity network fusion for aggregating data types on a genomic scale. *Nat Methods* 2014; **11**: 333–7.
121. Zeng X, Zhong Y, Lin W, *et al*. Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief Bioinform* 2020; **21**: 1425–36.
122. Sun Y, Norick B, Han J, *et al*. PathSelClus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Trans Knowl Discov Data* 2013; **7**: 11.
123. Sun Y, Barber R, Gupta M, *et al*. Co-author relationship prediction in heterogeneous bibliographic networks. In: *2011 International Conference on Advances in Social Networks Analysis and Mining*, 2011, 121–8. Kaohsiung, Taiwan: IEEE.
124. Wang L, You ZH, Huang YA, *et al*. An efficient approach based on multi-sources information to predict circRNA-disease associations using deep convolutional neural network. *Bioinformatics* 2020; **36**: 4038–46.
125. Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: theory and applications. *Neurocomputing* 2006; **70**: 489–501.

126. Huang G-B, Wang DH, Lan Y. Extreme learning machines: a survey. *Int J Mach Learn Cybern* 2011; **2**: 107–22.
127. Wang L, You ZH, Li YM, *et al*. GCNCDA: a new method for predicting circRNA-disease associations based on graph convolutional network algorithm. *PLoS Comput Biol* 2020; **16**: e1007568.
128. Chen J, Ma T, Xiao C. Fastgcn: fast learning with graph convolutional networks via importance sampling. arXiv preprint arXiv:1801.10247 2018.
129. Adnan MN, Islam MZ. Forest PA: Constructing a decision forest by penalizing attributes used in previous trees. *Expert Syst Appl* 2017; **89**: 389–403.
130. Breiman L, Friedman J, Stone CJ, *et al*. *Classification and Regression Trees*. CRC Press, Boca Raton, Florida 1984.
131. Deepthi K, Jereesh AS. An ensemble approach for CircRNA-disease association prediction based on autoencoder and deep neural network. *Gene* 2020; **762**: 145040.
132. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006; **313**: 504–7.
133. Deepthi K, Jereesh AS. Inferring potential CircRNA-disease associations via deep autoencoder-based classification. *Mol Diagn Ther* 2021; **25**: 87–97.
134. Chen X, Xie D, Zhao Q, *et al*. MicroRNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2017; **20**: 515–39.
135. Chen X, Yan CC, Zhang X, *et al*. Long non-coding RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2016; **18**: 558–76.
136. Liu H, Ren G, Chen H, *et al*. Predicting lncRNA–miRNA interactions based on logistic matrix factorization with neighborhood regularized. *Knowl Based Syst* 2020; **191**: 105261.
137. Zhao Q, Yu H, Ming Z, *et al*. The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. *Mol Ther Nucleic Acids* 2018; **13**: 464–71.
138. Hu H, Zhang L, Ai H, *et al*. HLPI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. *RNA Biol* 2018; **15**: 797–806.
139. Wang C-C, Chen X. A unified framework for the prediction of small molecule–MicroRNA association based on cross-layer dependency inference on multilayered networks. *J Chem Inf Model* 2019; **59**: 5281–93.
140. Qu J, Chen X, Sun YZ, *et al*. In silico prediction of small molecule-miRNA associations based on the hetesim algorithm. *Mol Ther Nucleic Acids* 2019; **14**: 274–86.
141. Zhao Y, Chen X, Yin J, *et al*. SNMFSMMA: using symmetric nonnegative matrix factorization and Kronecker regularized least squares to predict potential small molecule-microRNA association. *RNA Biol* 2020; **17**: 281–91.