

Research Article

BRWSP: Predicting circRNA-Disease Associations Based on Biased Random Walk to Search Paths on a Multiple Heterogeneous Network

Xiujuan Lei  and Wenxiang Zhang

School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi 710119, China

Correspondence should be addressed to Xiujuan Lei; xjlei@snnu.edu.cn

Received 15 September 2019; Accepted 4 November 2019; Published 30 November 2019

Academic Editor: Manlio De Domenico

Copyright © 2019 Xiujuan Lei and Wenxiang Zhang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The circular RNAs (circRNAs) have significant effects on a variety of biological processes, the dysfunction of which is closely related to the emergence and development of diseases. Therefore, identification of circRNA-disease associations will contribute to analysing the pathogenesis of diseases. Here, we present a computational model called BRWSP to predict circRNA-disease associations, which searches paths on a multiple heterogeneous network based on biased random walk. Firstly, BRWSP constructs a multiple heterogeneous network by using circRNAs, diseases, and genes. Then, the biased random walk algorithm runs on the multiple heterogeneous network to search paths between circRNAs and diseases. Finally, the performance of BRWSP is significantly better than the state-of-the-art algorithms. Furthermore, BRWSP further contributes to the discovery of novel circRNA-disease associations.

1. Introduction

circRNAs are a special type of endogenous noncoding RNAs (ncRNAs), which widely exist in the gene expression of various organisms. The discovery of circRNAs could date back to the nineteen seventies. Sanger et al. [1] first observed circRNAs in the process of studying plant viruses by using electron microscopy. circRNAs were gradually found in different species and cells after the following decades, such as yeast [2], zebrafish [3], and mouse [4]. Because of the low abundance of circRNAs and the lack of known function, circRNAs have not got more attention for a very long time.

With the rise and development of high-throughput sequencing technologies, a large number of circRNAs have been found and identified [5, 6]. Along with gradually penetrating to the study of circRNAs, more and more circRNAs have been identified and published. Therefore, various circRNA databases with different emphases have been constructed, such as CircR2Disease [7], circBase [8],

exoRBase [9], PlantcircBase [10], circAtlas 2.0 [11], and CSCD [12]. In addition, the biological function of circRNAs has also been gradually revealed, such as acting as miRNA sponges [13], interacting with RNA-binding proteins (RBPs) [14], participating in transcriptional regulation [15] and so forth.

Complex diseases seriously threaten human health [16–18]. Therefore, studies on complex diseases have been a hot topic in the field of medicine and bioinformatics [19, 20]. As more and more biological functions of circRNAs have been revealed, massive evidence has indicated that circRNAs play an important role in the emergence and development of complex diseases. According to the reports of Liu et al. [21], the function of circRNAs was also versatile to function as microRNA (miRNA) sponges [5, 13] and protein sponges [22, 23]. For example, the circSMARCA5 [24] and circCFH [25] have been found to be expressed in a glioma-specific pattern which may be used as the tumor biomarkers. CircNFI [26] and circNT5E [27] have been found that they

play oncogenic roles in glioma, whereas circFBXW7 and circSHPRH have been reported to function as the tumor suppressors. Furthermore, circRNAs might become an ideal choice for gene/protein delivery in future brain cancer therapies [21].

The above methods of predicting circRNA-disease associations are time-consuming and costly. The disadvantage can be properly overcome by adopting computational methods to identify circRNA-disease associations. Due to the low number of known circRNA-disease associations in the past years, machine learning methods are not widely used in the identification of circRNA-disease associations. However, the research progression of prediction in miRNA-disease association and lncRNA-disease association would benefit the development of computational models for circRNAs [28–30]. Recently, Fan et al. [7] constructed CircR2Disease database by using the method of literature retrieval, which provides 661 circRNAs, 100 diseases, and 725 circRNA-disease associations. Another similar database is circRNA-Disease [31], which provides an opportunity to identify circRNA-disease associations by using computational methods. Lei et al. [32] employed a method called depth-first search to search paths between circRNAs and diseases in heterogeneous network composed of circRNAs and diseases and then used the path weighted method to infer the probability of circRNA-disease based on searched paths. Fan et al. [33] built a heterogeneous network by using circRNA similarity network, disease similarity network, and circRNA-disease associations, and then they employed the KATZ method to predict circRNA-disease associations. Xiao et al. [34] utilized a manifold regularization learning framework to predict human disease-related circRNAs based on a heterogeneous circRNA-disease bilayer network. Zhao et al. [35] proposed a novel computational algorithm to identify circRNA-disease associations, which is based on the bipartite network projection and KATZ algorithm. Wei et al. [36] employed an improved matrix factorization identification algorithm to identify circRNA-disease associations. Yan et al. [37] utilized a DWNN-RLS algorithm based on regularized least squares of Kronecker product kernel to identify circRNA-disease associations.

In this paper, we propose a new computational method, named BRWSP, to identify circRNA-disease associations based on biased random walk to search paths on a multiple heterogeneous network. Specifically, BRWSP first establishes a multiple heterogeneous network by using circRNA coexpression similarity network, gene similarity network, disease similarity network, circRNA-gene associations, circRNA-disease associations, and gene-disease associations. Containing multiple types of biological data can facilitate a comprehensive analysis of circRNA-disease associations. Next, a biased random walk runs on this multiple heterogeneous network to search paths between a specific circRNA and a specific disease. BRWSP then calculates the score of specific circRNA-disease association by using those searched paths. Compared with state-of-the-art algorithms, BRWSP obtains better performance in the identification of circRNA-disease associations. The overall framework of BRWSP is depicted in Figure 1.

2. Materials and Methods

2.1. Motivations

- (1) Ba-Alawi et al. [38] used depth-first search algorithm to traverse all simple paths between a specific drug and a specific target protein and then aggregated the score from these searched paths to infer drug-target interactions. Then this algorithm was extended to identify miRNA-disease associations [39], lncRNA-disease associations [40], circRNA-disease associations [32], and microbe-disease associations [41] and obtained satisfactory performance. However, this algorithm needs to search for all paths between a specific circRNA and a specific disease. If the network is very enormous, this type of algorithm cannot handle it well. Therefore, this type of algorithm cannot be well extended to a multiple heterogeneous network constructed by using many different types of biological networks. Being inspired by [42], a biased random walk is proposed to search paths. Compared with depth-first search algorithm, it chooses the paths according to the probabilities (such as Figure 1(c)). Therefore, if the probability of one path is very smaller than other paths, it is very likely that the walker will not select this path in the process of selecting the next path.
- (2) Recently, many methods [32–34] have been proposed based on a heterogeneous network to identify circRNA-disease associations. However, these methods use fewer biological data and depend greatly on the known circRNA-disease associations, which lead to insufficient analysis of circRNA-disease associations from a variety of biological perspectives. Therefore, gene similarity networks and gene-disease associations are imported to build a multiple heterogeneous network which contains circRNA coexpression network, circRNA-disease associations, and disease similarity network.

2.2. Materials and Preprocessing

2.2.1. circRNA-Disease Associations. The datasets of circRNA-disease associations are downloaded from the CircR2Disease database (<http://bioinfo.snnu.edu.cn/>) [7]. The CircR2Disease database contains 725 circRNA-disease associations consisting of 661 circRNAs and 100 diseases. In order to ensure the accuracy of data, we only extract circRNAs with circBase IDs and gene symbols. Finally, 427 circRNA-disease associations, consisting of 372 circRNAs, 330 gene symbols, and 77 diseases, are remained.

2.2.2. Disease Semantic Similarity. The similarity between diseases can be calculated by a directed acyclic graph (DAG). Firstly, we search DOID corresponding to 77 diseases, being extracted in Section 2.2.1, from the Disease Ontology database (<http://www.disease-ontology.org/>) [43]. After deleting diseases without DOID, the dataset contains 55 diseases with DOID, 291 circRNAs, 261 gene symbols, and

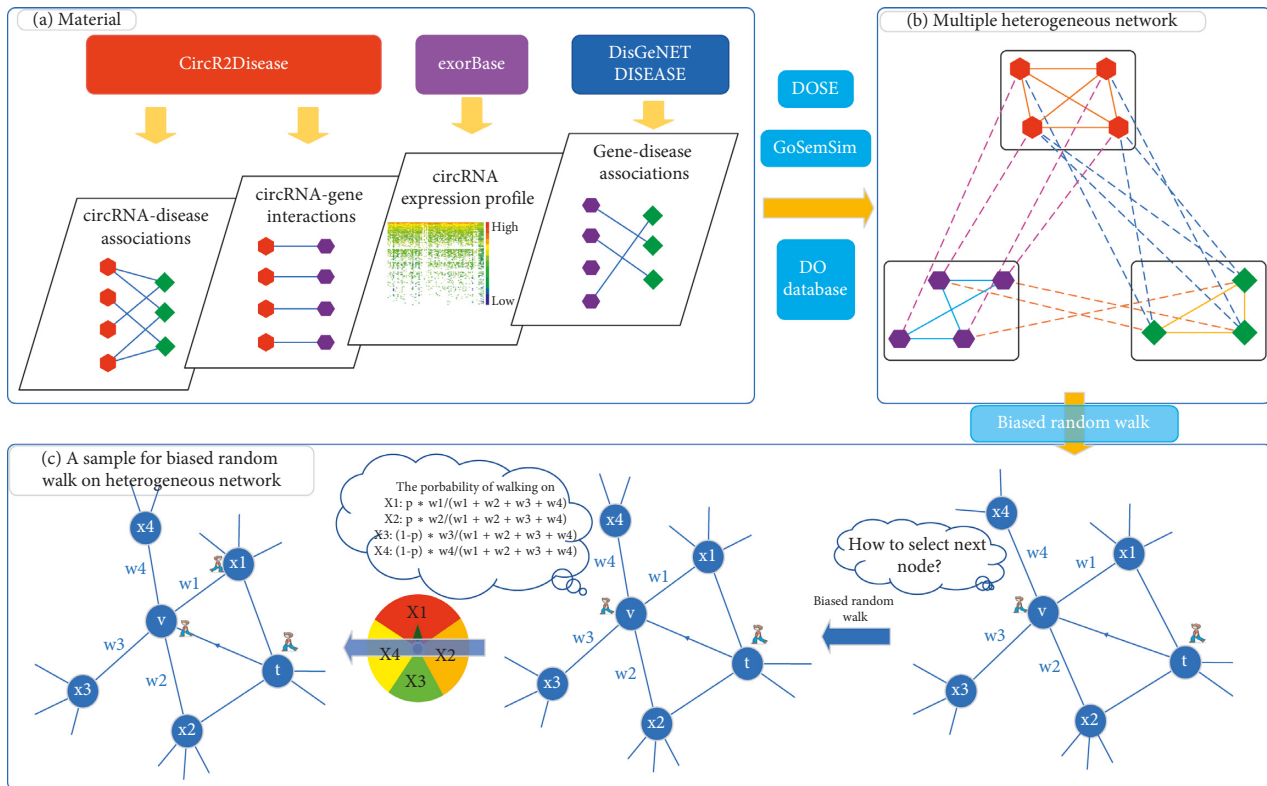


FIGURE 1: The framework of BRWSP. (a) Some original data are downloaded from corresponding databases. (b) A multiple heterogeneous network is constructed. (c) Biased random walk algorithm runs on a heterogeneous network to find paths between a specific circRNA and a specific disease.

340 circRNA-disease associations. Based on disease ontology, Yu et al. [44] created a DOSE package of R, which can calculate disease semantic similarity by *doSim* function based on Wang's method [45]. In this study, we adopt this DOSE package to calculate disease semantic similarity.

2.2.3. circRNA Expression Profile. To calculate the circRNA coexpression similarity network, the circRNA expression profile is downloaded from the database exorBase (<http://www.exorbase.org/>) [9]. After converting *exor_circ_ID* to circBase ID, we eliminate some circRNAs without expression profile among 291 circRNAs. The final data contain expression profile data of 154 circRNAs on 90 samples, 192 circRNA-disease associations consisting of 154 circRNAs (corresponding to 140 gene symbols) and 48 diseases (being shown in Figure 2).

2.2.4. Gene-Disease Associations. In order to detect associations between 48 diseases and 140 genes (corresponding to circRNAs), we download the integrated gene-disease associations from the *human_disease_textmining_full.tsv* file of the DISEASE Database [46]. A confidence score is given to evaluate associations in this database. In order to ensure the reliability of data, we only select the gene-disease associations whose confidence score is greater or equal to 2 according to previous research [47]. In total, among 48 diseases and 140 gene symbols, we obtain sufficiently 80

gene-disease associations consisting of 29 diseases and 34 genes.

Besides, we also extract some genes associated with the 48 diseases mentioned above from the DISEASE database [46] and DisGeNET database [48]. Similarly, we only extract gene-disease associations with confidence score greater or equal to 2 for the *human_disease_experiments_full.csv* file of the DISEASE database [46]. And for the DisGeNET database, the gene-disease associations are extracted from the *curated_gene_disease_associations.tsv.gz* file. Finally, among the 48 diseases mentioned above, 2193 disease-gene associations are extracted, which contain 37 diseases and 1607 disease-related genes.

2.2.5. Constructing Multiple Heterogeneous Network. In this paper, we extract 140 gene symbols (corresponding to circRNAs) from CircR2Disease. According to these gene symbols, gene similarity network is constructed by mapping gene products to GO annotations [49]. Genes are annotated by cellular component (CC), molecular function (MF), and biological process (BP). Herein, we use the biological process (BP) to measure gene semantic similarity value, which has been proven to embrace better performance in previous papers [50]. Finally, the adjacency matrix GS is utilized to represent the gene similarity network, and the value $GS(i, j)$ represents a functional similarity value between gene i and gene j , which can be calculated by the function of *geneSim* in the GoSemSim package of R [49].

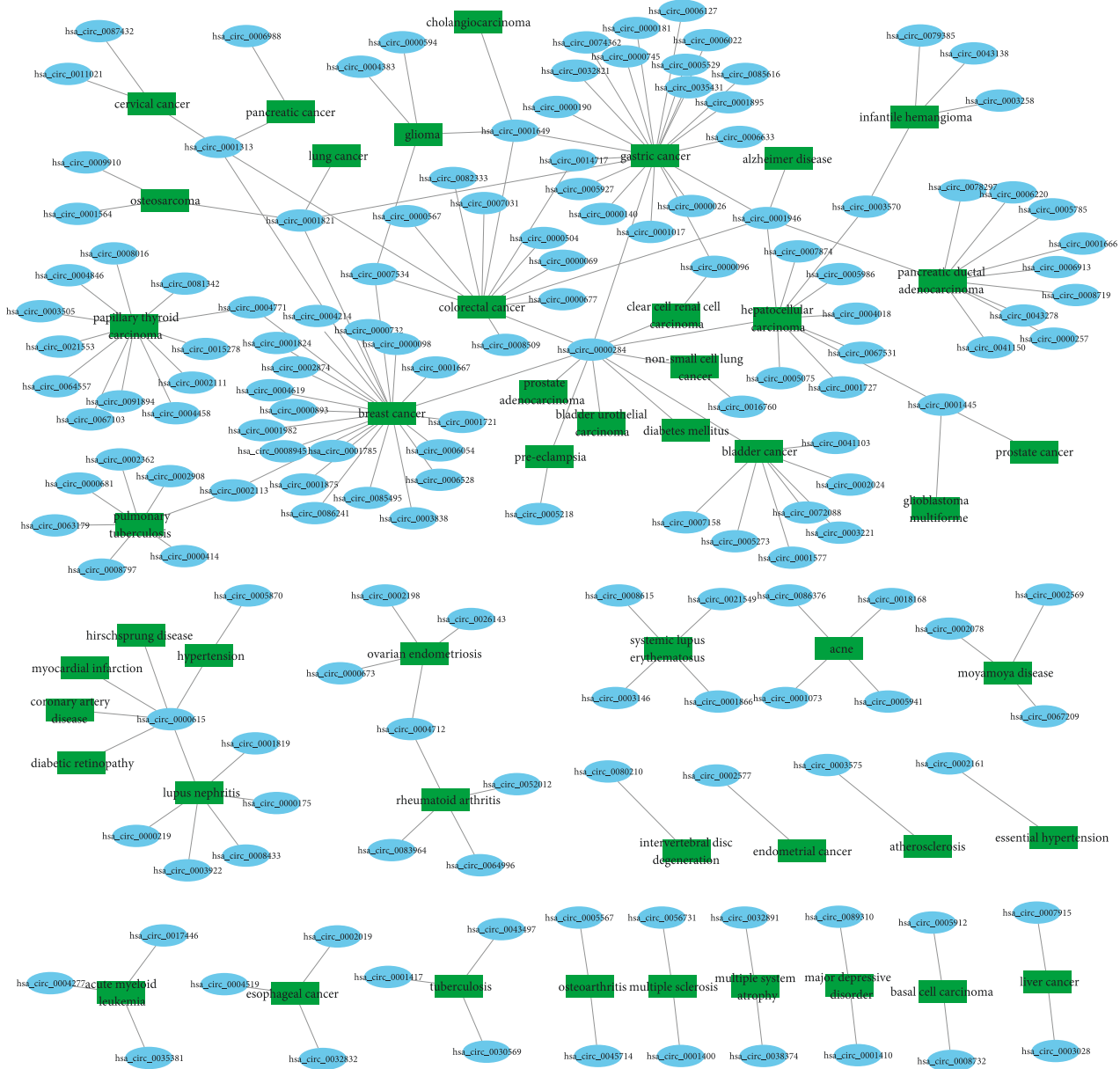


FIGURE 2: The bipartite graph of circRNA-disease associations.

The adjacency matrix CD is constructed to represent circRNA-disease associations and $CD(i, j)$ is equal to 1 when circRNA $c(i)$ is associated with disease $d(j)$; similarly, the adjacency of CG and GD is used to describe circRNA-gene interactions and gene-disease associations, respectively. Besides, we employ the adjacency matrix DS to describe disease semantic similarity, in which the $DS(i, j)$ indicates the semantic similarity between disease $d(i)$ and disease $d(j)$. For circRNA coexpression similarity CS , $CS(i, j)$ represents the similarity value between circRNA $c(i)$ and circRNA $c(j)$, which is calculated by using the Pearson correlation coefficients based on circRNA expression profile.

In the process of predicting circRNA-disease associations, the performance of the algorithm largely depends on the known circRNA-disease associations. However, the existing known circRNA-disease associations are still

limited, which will affect the accuracy of the algorithm for predicting circRNA-disease associations. In order to solve this problem, we calculate the initial score for circRNA-disease associations based on the gene-disease associations. The initial score of the association between circRNA i and disease k is as follows:

$$\text{Initial}_{\text{Score}(i,k)} = \max(\text{genSim}(g_i, dg_j^k)), \quad (1)$$

where g_i is the gene corresponding to circRNA $c(i)$ and dg_j^k represents the gene associating with disease $d(k)$. $\text{genSim}(g_i, dg_j^k)$ represents the semantic similarity value between gene g_i and gene dg_j^k calculated by the GoSemSim package of R [49]; $\text{Initial}_{\text{Score}(i,k)}$ represents the initial score of the association of circRNA $c(i)$ and disease $d(k)$. If $CD(i, j)$ is equal to 0, $\text{Initial}_{\text{Score}(i,k)}$ will be assigned $CD(i, j)$ as a new value.

Next, a multiple heterogeneous network is constructed by using circRNA coexpression network, disease similarity network, gene functional similarity network, and their association information, which is represented as follows:

$$H = \begin{bmatrix} CS & CG & CD \\ CG^T & GS & GD \\ CD^T & GD^T & DS \end{bmatrix}, \quad (2)$$

where CG^T , CD^T , and DS^T are the transposed matrices of CG , CD , and DS , respectively. To avoid the biases caused by larger values in the multiple heterogeneous network, H is utilized to construct a normalized multiple heterogeneous network $NMH = D^{-(1/2)}HD^{-(1/2)}$, and D is a degree matrix of H .

The overall framework of BRWSP is depicted in Figure 3.

$$P(c_{k+1} = x | c_k = v, c_{k-1} = t) = \begin{cases} \frac{\Phi(t, v, x) * NMH(v, x)}{\sum_{i \in \text{Nei}(v)} NMH(v, i)}, & \text{if } x \in \text{Nei}(v), x \notin \text{Path}, \\ 0, & \text{otherwise,} \end{cases} \quad (3)$$

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

where $P(c_{k+1} = x | c_k = v, c_{k-1} = t)$ represents the transition probability of selecting node x the next biased random walk, and the currently visited node and the last visited node are v and t , respectively. $\text{Nei}(v)$ and $\text{Nei}(t)$ represent the neighbourhoods of v and t , respectively. For parameter q , if q is assigned a larger value, the nodes of Path are highly interconnected and belong to communities or similar network clusters (similar to BFS algorithm). Otherwise, the nodes of Path can more exactly describe a macroview of the neighbourhood (similar DFS algorithm). In other words, we can integrate the strategies of DFS and BFS by adjusting the value of the parameter q . Finally, each neighbourhood of v can obtain a probability of being visited in the next biased random walk. A roulette selection algorithm, a simple random choice based on probability, is employed to randomly select the next node from the neighbourhood of v based on their probability. Then the selected node is added to corresponding Path. If k is equal to 1, the next node is randomly selected from the neighbourhoods of the last node based on their probability.

In the process of biased random walk to search paths between circRNA $c(i)$ and disease $d(j)$, the path from $c(i)$ to $d(j)$ will be saved if its length is less than or equal to L . Otherwise, the current biased walk fails to search for a corresponding path. In order to search for more possible paths between circRNA $c(i)$ and disease $d(j)$, we will repeat the above steps *maxiter* times. Therefore, after the biased random walk, we can get a lot of paths from circRNA $c(i)$ to disease $d(j)$.

2.3. BRWSP Methods

2.3.1. Biased Random Walk to Search Paths. In the paper [42], DFS can search for more different types of nodes because it explored a network as deeply as possible. The breadth-first search (BFS) can search the neighbourhoods of source node. Being inspired by it, a biased random walk algorithm is designed to search paths between circRNAs and diseases, which combines the advantages of DFS and BFS by adjusting the BRWSP's parameter (being explained as follows).

Formally, let $\text{Path} = \{p_1, p_2, \dots, p_{L+1}\}$ represents one path between circRNA p_1 and disease p_{L+1} . In this Path, p_i represents the node (circRNA or disease) of Path and L represents the length of Path. Let c_k indicate the node accessed by the k th biased random walk. The strategy of selecting the next node is described as follows:

2.3.2. Calculating circRNA-Disease Score Based on Paths.

It is known that circRNA $c(i)$ and disease $d(j)$ are possibly associated with each other if many paths with higher weight and shorter length are found among them. Therefore, an exponential decay function for circRNA $c(i)$ and disease $d(j)$ is utilized to give more support for paths with high weight and short length as follows:

$$\text{score}(c(i), d(j)) = \sum_{\text{Path}_i \in \text{All_Path}} \left(\prod_{e=1}^{\text{len}(\text{Path}_i)} W_e(\text{Path}_i) \right)^{\alpha \times \text{len}(\text{Path}_i)}, \quad (4)$$

where $\text{score}(c(i), d(j))$ represents the score of predicted association score between circRNA $c(i)$ and disease $d(j)$. $\text{All_Path} = \{\text{Path}_1, \text{Path}_2, \dots, \text{Path}_n\}$ represents all paths we have searched between circRNA $c(i)$ and disease $d(j)$, where Path_i represents the i th searched path. $W_e(\text{Path}_i)$ represents the weight of the e th edge in Path_i . $\text{len}(\text{Path}_i)$ is the length of Path_i and the parameter α represents a decay factor.

3. Results and Discussion

3.1. Evaluation Metrics. In this paper, the leave-one-out cross-validation (LOOCV) is utilized to analyse the performance of BRWSP in the process of predicting circRNA-disease associations. According to the results of LOOCV, the receiver operating characteristic (ROC) curve is plotted and the area under of ROC curve (AUC) is calculated as evaluation criteria.

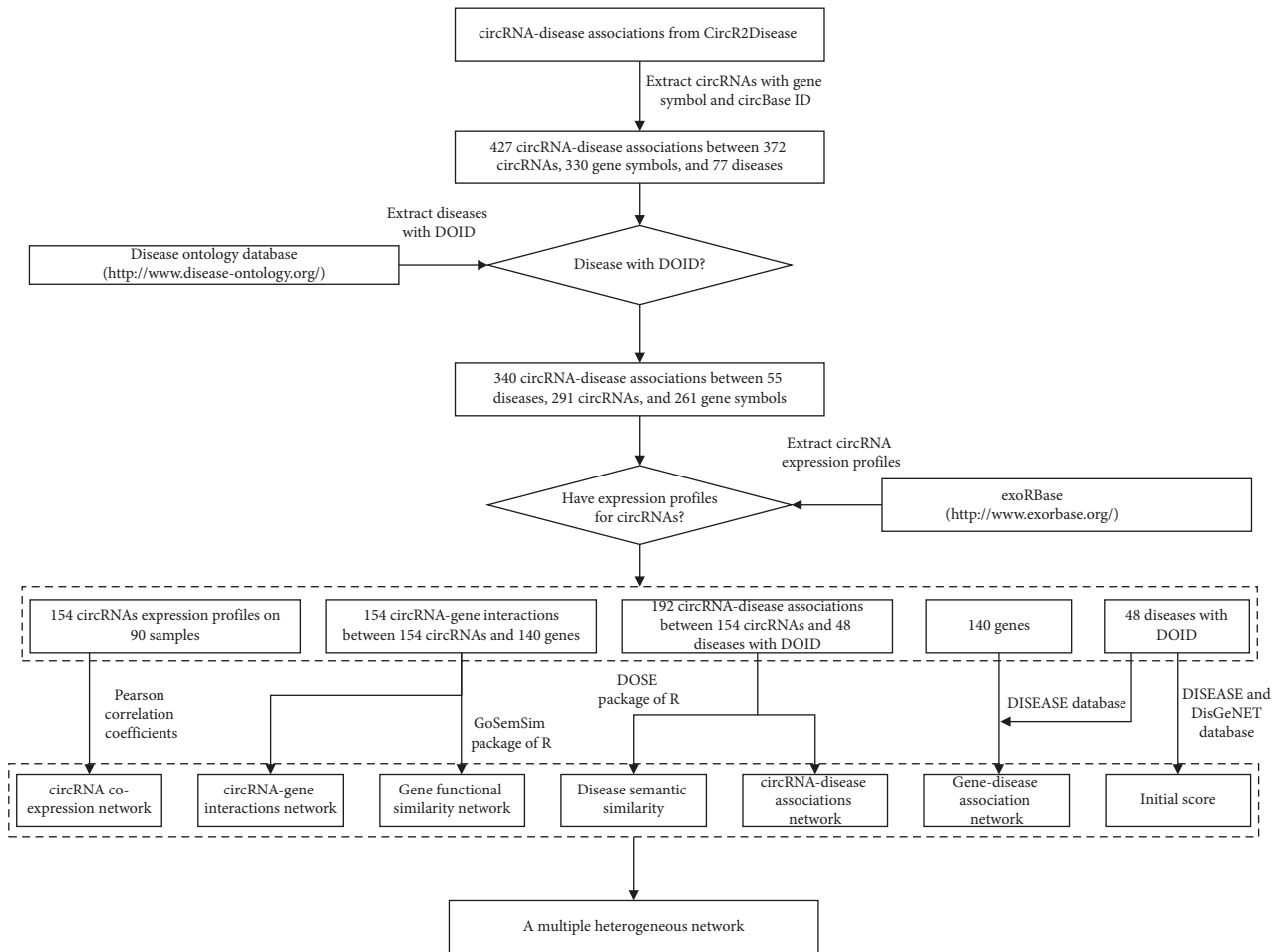


FIGURE 3: The framework of materials and preprocessing.

In the process of predicting circRNA associated with disease k , the positive samples are those known circRNAs associated with disease k . Reliable negative samples are required in the process of evaluation. However, there is no prior information about the negative samples (non-disease-related circRNAs). All unknown genes can be regarded as negative samples. However, there are two disadvantages to this approach. Firstly, there is no evidence to prove that the unknown circRNAs are related or unrelated to diseases currently. It is not scientific to make that all unknown genes are regarded as negative samples. Secondly, this approach will lead to class-imbalance problem since the number of known circRNAs is much fewer than the number of unknown circRNAs. This phenomenon has also been widely discussed in identifying disease-related genes, miRNAs and lncRNAs [47,51–53]. Therefore, it is not scientific to regard all unknown genes as negative samples. To overcome these problems and extract reliable negative samples, we first calculate all initial scores of the associations between all circRNAs and disease k according to equation (1) and arrange them in ascending order. The circRNAs whose number is same with the number of positive samples are selected as negative samples from the front of the results of ascending order. If all initial scores are equal to 0, we randomly select some circRNAs as negative samples from

unknown circRNAs associated with disease k , in which the number of negative samples is equal to the number of positive samples. Finally, we can get all predicted scores for positive samples and negative samples.

3.2. Effects of Parameters. There are four parameters in the BRWSP algorithm. Among them, we set the path length L is equal to 3 based on the previous studies [38–41]. However, the values of q , $maxiter$, and decay factor α are undefined. Therefore, we set $maxiter = 300$, $q \in \{0.02, 0.04, 0.06, 0.08, 0.1, 0.12, 0.14, 0.16, 0.18\}$ and $\alpha \in \{0.5, 1, 1.5, 2, 3, 5\}$. The experimental results after combining different values of q and α are listed in Figure 4. Figure 4 shows that the BRWSP algorithm will get the best AUC value (0.8675) when $q = 0.12$ and $\alpha = 1$.

3.3. Comparison with Other Methods. In order to analyse the performance of the BRWSP algorithm in predicting circRNA-disease associations, BRWSP ($L = 3$, $q = 0.12$, $maxiter = 300$, and $\alpha = 1$) is compared with KATZHCDA [33], iCircDA-MF [36], RLS-Kron [37, 54], and DFSPW [38–41]. Herein, for DFSPW algorithm, it first searches all paths between circRNAs and diseases and then calculates the score between circRNAs and diseases based on paths by

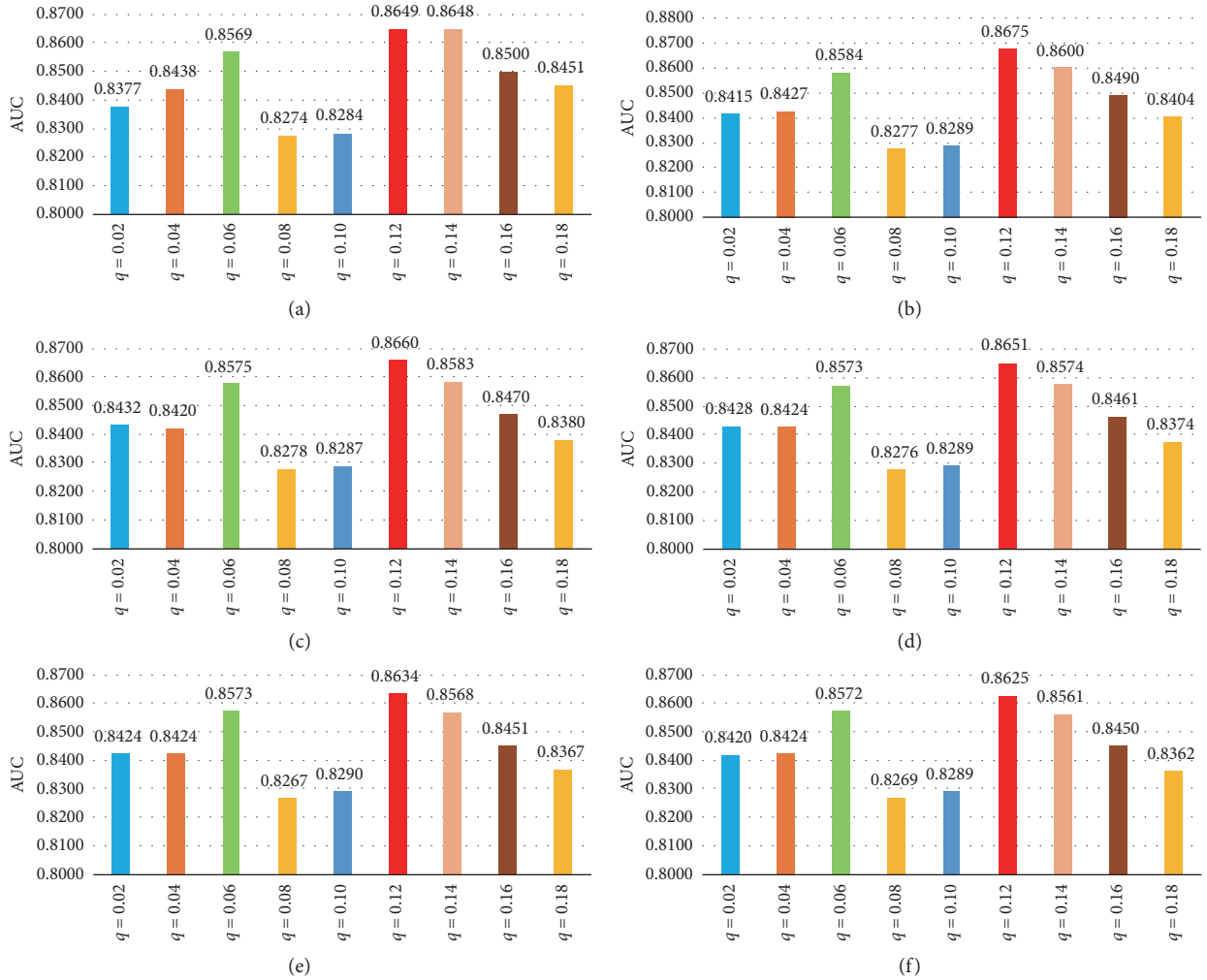


FIGURE 4: The influence of α and q on AUC. (a) $\alpha = 0.5$, (b) $\alpha = 1$, (c) $\alpha = 1.5$, (d) $\alpha = 2$, (e) $\alpha = 3$, and (f) $\alpha = 5$.

formula (4). For DFSPW algorithm's parameters, the maximum length of path and the decay factor are equal to 3 and 2.26, respectively, based on the previous study [38–41]. For the convenience of comparison, we apply these computational methods on the same dataset in this paper.

The comparison results of BRWSP and other algorithms are shown in Figures 5–7. Obviously, we can observe clearly from Figure 5 that the AUC value of BRWSP is 0.8675, which improves the prediction precision by 6.49%, 19.36%, 21.65%, and 22.81% compared to the KATZHCDA, RLS-Kron, iCircDA-MF, and DFSPW algorithm, respectively. The precision and recall are listed at each top 100 circRNAs in Figure 7, in which we can find BRWSP get excellent performance. In addition, we calculate the number of circRNAs with each disease. Then, we arrange them in ascending order and select the top 4 cancer diseases (breast cancer, stomach cancer, colorectal cancer, and papillary thyroid carcinoma) to analysis. The four common diseases are associated with 24 circRNAs, 22 circRNAs, 13 circRNAs, and 12 circRNAs, respectively. Figure 7 shows the performance of each algorithm on the four cancer diseases. In a word, we can see that BRWSP gets the satisfactory performance from Figures 5–7.

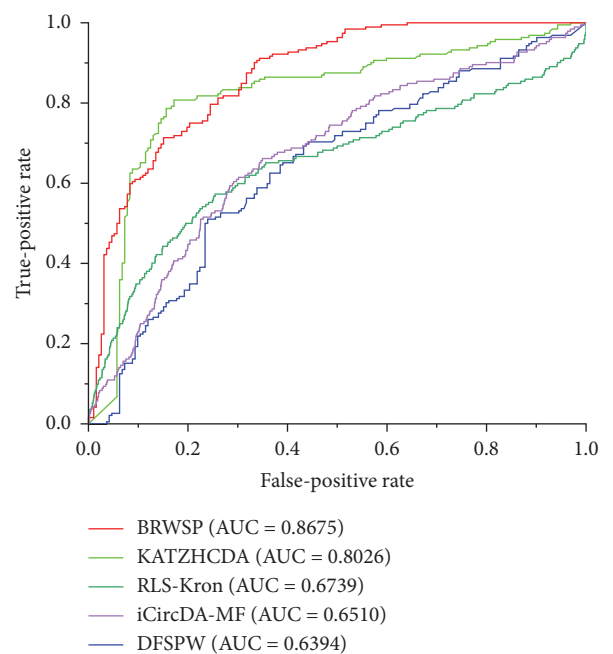


FIGURE 5: Comparison between BRWSP and other algorithms.

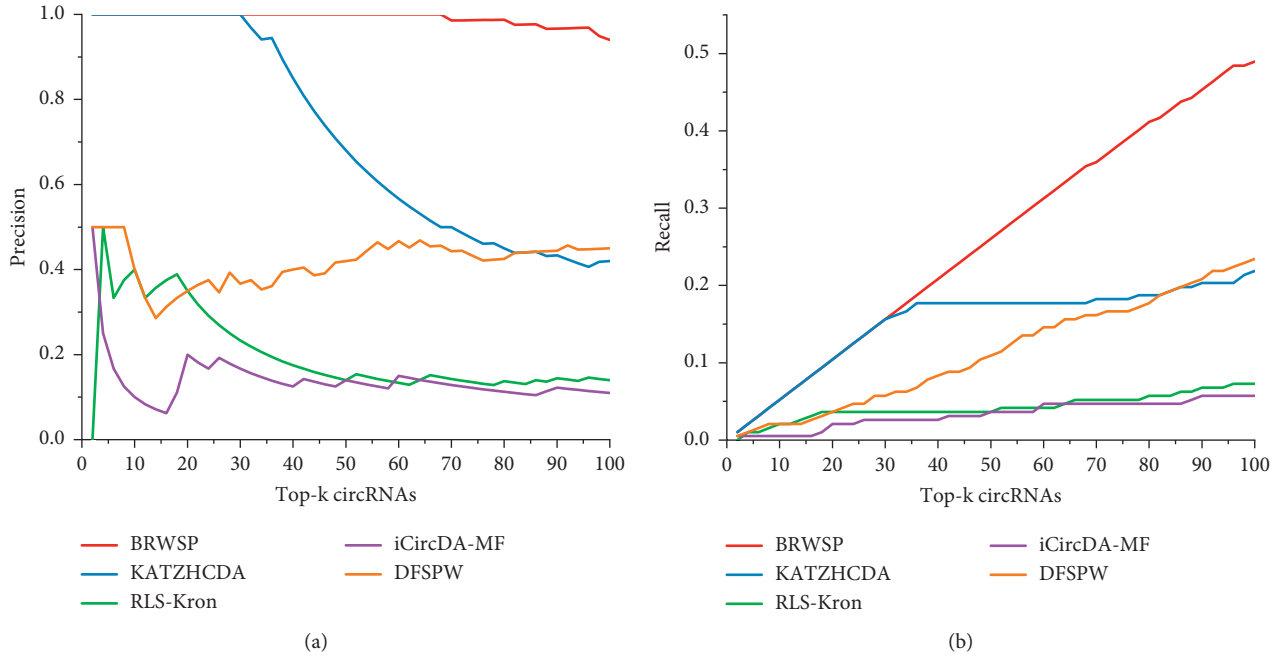


FIGURE 6: The precision and recall at top-k circRNAs.

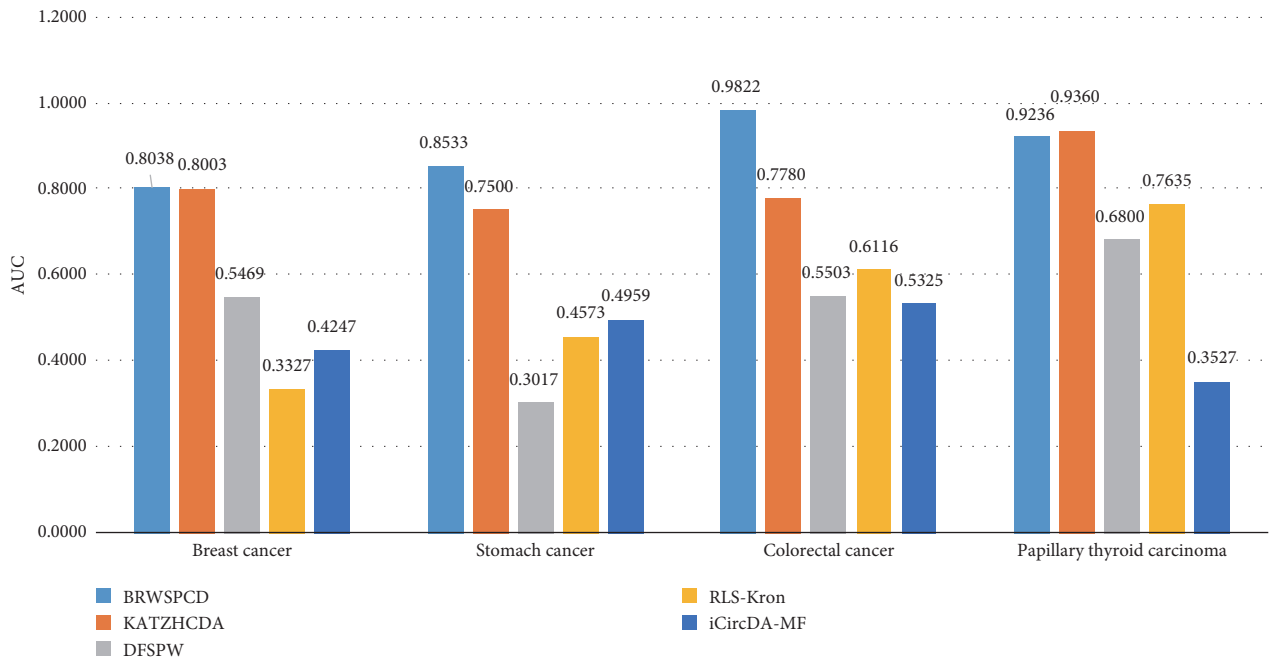


FIGURE 7: The AUC performance of different algorithms for 4 common diseases.

3.4. The Effect of Gene Network. One of the highlights of our paper is that the gene similarity network is utilized to construct a multiple heterogeneous network with circRNA coexpression similarity network, disease semantic similarity, and associations among them. In this section, we analyse its impact on predicting circRNA-disease associations. In other words, we run our algorithm on a heterogeneous network (constructed by circRNA coexpression similarity network, disease semantic similarity, initial score, and their association information) and a multiple heterogeneous network

(constructed by circRNA coexpression similarity network, gene similarity network, disease semantic similarity, initial score, and their association information).

Obviously, we can clearly see from Figure 8 that our algorithm on a multiple heterogeneous network (*Mul_Het_Net*) gets better performance than that on heterogeneous network (*Het_Net*). The difference between *Mul_Het_Net* and *Het_Net* is that *Mul_Het_Net* introduces gene similarity network. Therefore, the introduction of gene similarity network is helpful to identify circRNA-disease associations.

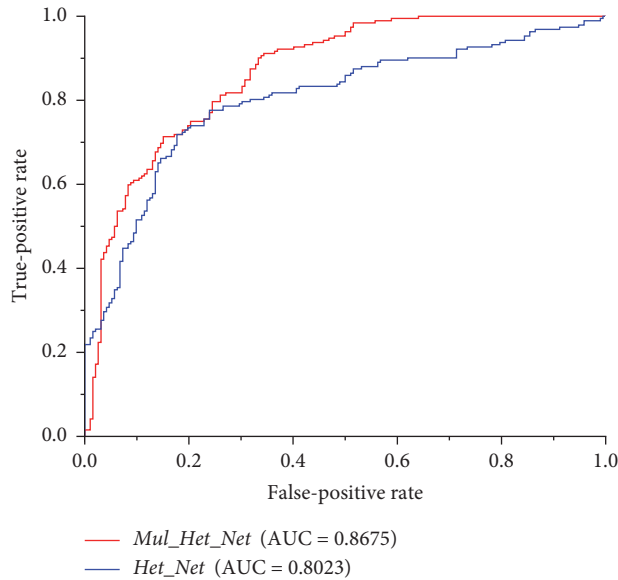


FIGURE 8: The effect of BRWSP on different networks.

TABLE 1: The known 13 circRNAs for colorectal cancer in the CircR2Disease database.

circRNA name	circBase ID	Gene symbol	PMID
hsa_circ_0000284/circHIPK3	hsa_circ_0000284	HIPK3	29549306/27050392
hsa_circ_0000567	hsa_circ_0000567	SETD3	29333615
hsa_circ_0001313/circCCDC66	hsa_circ_0001313	CCDC66	28249903
hsa_circ_0001649	hsa_circ_0001649	SHPRH	29421663
hsa_circ_0000504	hsa_circ_0000504	TUBGCP3	28656150
CDR1as/ciRS-7/hsa_circ_0001946	hsa_circ_0001946	CDR1	28174233/28435295
hsa_circ_0014717	hsa_circ_0014717	CCT3	29571246
hsa_circ_0007534	hsa_circ_0007534	DDX42	29364478
hsa_circ_0000069	hsa_circ_0000069	STIL	28003761
hsa_circ_0007031	hsa_circ_0007031	TUBGCP3	28656150
circ-KLDHC10/hsa_circ_0082333	hsa_circ_0082333	KLHDC10	26138677
hsa_circ_0000677/hsa_circ_001569/circABCC	hsa_circ_0000677	ABCC1	27058418
hsa_circ_0008509	hsa_circ_0008509	NAV3	28656150

3.5. Case Study. To further demonstrate the effectiveness of BRWSP ($L = 3$, $q = 0.12$, $maxiter = 300$, and $\alpha = 1$) in predicting new circRNA-disease associations, a case study is performed for colorectal cancer, which is associated with 13 circRNAs (being shown in Table 1). In the process of experiment, 13 circRNAs associating with colorectal cancer are still assigned as training data and other circRNAs act as candidate samples. At the end of the prediction, we rank the score of candidate samples in descending order, and then the top 20 candidate samples (circRNAs) are selected. The literature mining method and interaction network method are utilized to analyse associations between them and colorectal cancer.

The result of the literature validation method is shown in Table 2. For the fourth column in Table 2, if there is a corresponding literature indicating that the gene corresponding to circRNA is associated with colorectal cancer, and the corresponding position in the fourth column is set the corresponding literature’s PMID, otherwise “-”. Obviously, we can clearly see that there are 12 literature studies to support our result from Table 2.

Interaction network method is to show the host gene of circRNA interacts with disease genes in PPI network and Pathway network. If host gene of predicted circRNA interacts with disease genes, this phenomenon indicates that the predicted circRNA is likely to be associated with the corresponding disease. Genes associating with colorectal cancer are extracted from the DISEASE database [46] and DisGeNET database [48]; protein-protein interaction (PPI) network and Pathway network are extracted from the research [55]. Then, we extract the interaction between genes associating with colorectal cancer and genes corresponding top 20 circRNAs in PPI network and Pathway network. The final analysis result is shown in Figure 9. We can clearly observe that 11 genes corresponding to circRNAs interact with colorectal cancer genes. The gene *POLD1* is not just colorectal gene and also associated with *hsa_circ_0052012*. In addition, three sets of connected graphs are constructed by predicted circRNAs, the host gene of predicted circRNAs, and colorectal cancer genes. The first set of connected graph contains *hsa_circ_0067531*, *hsa_circ_0002362*, *hsa_circ_0091894*, *hsa_circ_0000893*, *hsa_circ_0052012*,

TABLE 2: Literature validation of the top 20 results.

Rank	circRNA	Gene symbol	PMID	Rank	circRNA	Gene symbol	PMID
1	hsa_circ_0067531	PIK3CB	26747178	11	hsa_circ_0001875	FAM120A	26682245
2	hsa_circ_0002362	ASAP2	—	12	hsa_circ_0064996	SNRK	—
3	hsa_circ_0021553	MPPED2	30846004	13	hsa_circ_0052012	POLD1	26133394
4	hsa_circ_0026143	TROAP	30021381	14	hsa_circ_0005218	FAM120A	26682245
5	hsa_circ_0021549	MPPED2	30846004	15	hsa_circ_0008615	PPP1R13L	—
6	hsa_circ_0041150	RPH3AL	21536019	16	hsa_circ_0000893	DHPS	—
7	hsa_circ_0072088	ZFR	31010678	17	hsa_circ_0091894	FLNA	31002357
8	hsa_circ_0001400	RELL1	—	18	hsa_circ_0001727	ZKSCAN1	—
9	hsa_circ_0002577	WDR26	—	19	hsa_circ_0006022	LYN	29188362
10	hsa_circ_0001666	FAM120B	—	20	hsa_circ_0008719	AKT2	28837154

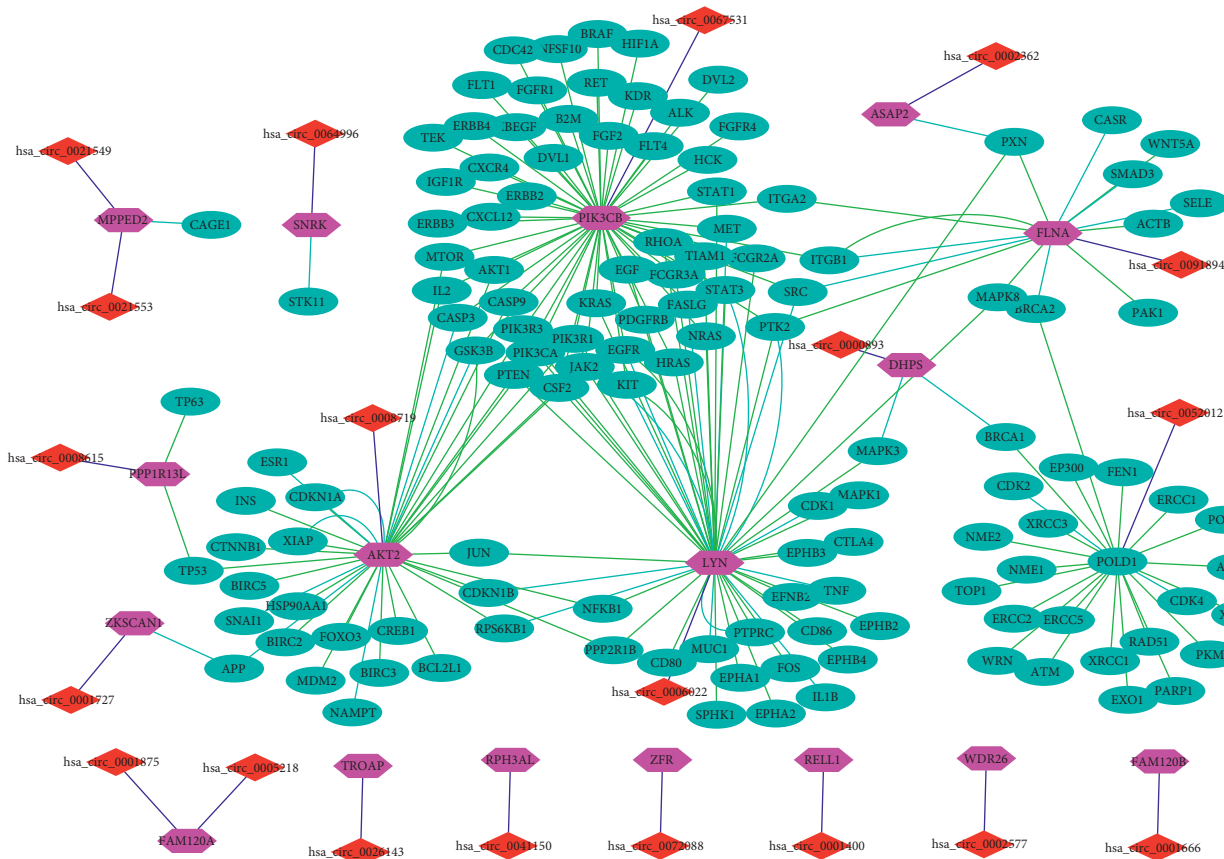


FIGURE 9: The interaction network method validated the top 20 results. Red nodes represent the circRNAs; pink nodes represent genes corresponding to circRNAs; dark turquoise nodes represent colorectal cancer genes; blue edges represent circRNA-gene associations; cyan edges represent protein-protein interactions; green edges represent Pathway associations.

hsa_circ_0006022, hsa_circ_0008719, hsa_circ_0008615, hsa_circ_0001727, the host gene of the predicted circRNAs, and corresponding disease genes. The second set of connected graph includes hsa_circ_0021549, hsa_circ_0021553, MPPED2, and CAGE1. Similarly, the hsa_circ_0064996, SNRK, and STK11 construct the third set of connected graph.

4. Conclusion

In this study, we propose a novel path weighted computational method, named BRWSP, to predict circRNA-disease associations. Highlights of BRWSP are to construct a multiple

heterogeneous network and to employ the biased random walk strategy to search paths between circRNAs and diseases. Firstly, BRWSP constructs a multiple heterogeneous network by using circRNA similarity network, gene similarity network, disease similarity network, and their associations, which can analyse the circRNA-disease associations from different biological perspectives. Secondly, the biased random walk is employed to search paths, which can eliminate some low probability paths. Experimental results show that BRWSP receives a satisfactory performance compared with other algorithms.

Although the BRWSP can effectively predict circRNA-disease associations, it still has several shortcomings. Firstly,

we only use a small amount of circRNA-disease associations and do not consider those circRNAs without gene symbol, circBase ID, and expression profile information. Besides, BRWSP has to consider four parameters (*maxiter*, *p*, *L*, and α). Therefore, it is a challenge about how to select optimal parameters in different situations. In a word, these limitations will encourage us to do further research studies in the future work.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

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

References

- [1] H. L. Sanger, G. Klotz, D. Riesner, H. J. Gross, and A. K. Kleinschmidt, "Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 73, no. 11, pp. 3852–3856, 1976.
- [2] Y. Matsumoto, R. Fishel, and R. B. Wickner, "Circular single-stranded RNA replicon in *Saccharomyces cerevisiae*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 19, pp. 7628–7632, 1990.
- [3] Y. Shen, X. Guo, and W. Wang, "Identification and characterization of circular RNAs in zebrafish," *FEBS Lett*, vol. 591, no. 1, pp. 213–220, 2017.
- [4] S. Werfel, S. Nothjunge, T. Schwarzmayr, T. M. Strom, T. Meitinger, and S. Engelhardt, "Characterization of circular RNAs in human, mouse and rat hearts," *Journal of Molecular and Cellular Cardiology*, vol. 98, pp. 103–107, 2016.
- [5] S. Memczak, M. Jens, A. Elefsinioti et al., "Circular RNAs are a large class of animal RNAs with regulatory potency," *Nature*, vol. 495, no. 7441, pp. 333–338, 2013.
- [6] M. Danan, S. Schwartz, S. Edelheit, and R. Sorek, "Transcriptome-wide discovery of circular RNAs in Archaea," *Nucleic Acids Research*, vol. 40, no. 7, pp. 3131–3142, 2012.
- [7] C. Y. Fan, X. J. Lei, Z. Q. Fang, Q. H. Jiang, and F. X. Wu, "CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases," *Database*, vol. 2018, 2018.
- [8] P. Glazar, P. Papavasileiou, and N. Rajewsky, "circBase: a database for circular RNAs," *RNA-Publ RNA Soc*, vol. 20, no. 11, pp. 1666–1670, 2014.
- [9] S. L. Li, Y. C. Li, B. Chen et al., "exoRBase: a database of circRNA, lncRNA and mRNA in human blood exosomes," *Nucleic Acids Research*, vol. 46, no. D1, pp. D106–D112, 2018.
- [10] Q. Chu, X. Zhang, X. Zhu et al., "PlantcircBase: a database for plant circular RNAs," *Molecular Plant*, vol. 10, no. 8, pp. 1126–1128, 2017.
- [11] P. Ji, W. Wu, S. Chen et al., "Expanded expression landscape and prioritization of circular RNAs in mammals," *Cell Reports*, vol. 26, no. 12, pp. 3444–3460.e5, 2019.
- [12] S. Xia, J. Feng, K. Chen et al., "CSCD: a database for cancer-specific circular RNAs," *Nucleic Acids Research*, vol. 46, no. D1, pp. D925–D929, 2018.
- [13] T. B. Hansen, T. I. Jensen, B. H. Clausen et al., "Natural RNA circles function as efficient microRNA sponges," *Nature*, vol. 495, no. 7441, pp. 384–388, 2013.
- [14] D. B. Dudekulay, A. C. Panda, I. Grammatikakis et al., "A web tool for exploring circular RNAs and their interacting proteins and microRNAs," *RNA Biology*, vol. 13, no. 1, pp. 34–42, 2016.
- [15] C. W. Chao, D. C. Chan, A. Kuo, and P. Leder, "The mouse formin (Fmn) gene: abundant circular RNA transcripts and gene-targeted deletion analysis," *Molecular Medicine (Cambridge, Mass)*, vol. 4, no. 9, pp. 614–628, 1998.
- [16] B. Liu, S. Feng, X. Guo, and J. Zhang, "Bayesian analysis of complex mutations in HBV, HCV, and HIV studies," *Big Data Mining and Analytics*, vol. 2, no. 3, pp. 145–158, 2019.
- [17] C. Sun, Q. Li, L. Cui, H. Li, and Y. Shi, "Heterogeneous network-based chronic disease progression mining," *Big Data Mining and Analytics*, vol. 2, no. 1, pp. 27–36, 2019.
- [18] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A. L. Barabási, "The human disease network," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 21, pp. 8685–8690, 2007.
- [19] C.-Y. Liu and C.-H. Chang, "An optimal algorithm for determining risk factors for complex diseases: depressive disorder, osteoporosis, and fracture in young patients with breast cancer receiving curative surgery," *Complexity*, vol. 2018, Article ID 7536731, 8 pages, 2018.
- [20] F.-X. Wu, J. Wang, M. Li, and H. Wang, "Biomolecular networks for complex diseases," *Complexity*, vol. 2018, Article ID 4210160, 3 pages, 2018.
- [21] J. Liu, K. Zhao, N. Huang, and N. Zhang, "Circular RNAs and human glioma," *Cancer Biology & Medicine*, vol. 16, no. 1, pp. 11–23, 2019.
- [22] K. Abdelmohsen, A. C. Panda, R. Munk et al., "Identification of HuR target circular RNAs uncovers suppression of PABPN1 translation by CircPABPN1," *RNA Biology*, vol. 14, no. 3, pp. 361–369, 2017.
- [23] R. Ashwal-Fluss, M. Meyer, N. R. Pamudurti et al., "circRNA biogenesis competes with pre-mRNA splicing," *Molecular Cell*, vol. 56, no. 1, pp. 55–66, 2014.
- [24] D. Barbagallo, A. Caponnetto, M. Cirnigliaro et al., "CircS-MARCA5 inhibits migration of glioblastoma multiforme cells by regulating a molecular Axis involving splicing factors SRSF1/SRSF3/PTB," *International Journal of Molecular Sciences*, vol. 19, no. 2, p. 480, 2018.
- [25] A. Bian, Y. Wang, J. Liu et al., "Circular RNA complement factor H (CFH) promotes glioma progression by sponging miR-149 and regulating AKT1," *Medical Science Monitor*, vol. 24, pp. 5704–5712, 2018.
- [26] H. Xu, Y. Zhang, L. Qi, L. Ding, H. Jiang, and H. Yu, "NFIX circular RNA promotes glioma progression by regulating miR-34a-5p via notch signaling pathway," *Frontiers in Molecular Neuroscience*, vol. 11, 2018.
- [27] R. Wang, S. Zhang, X. Chen et al., "CircNT5E acts as a sponge of miR-422a to promote glioblastoma tumorigenesis," *Cancer Research*, vol. 78, no. 17, pp. 4812–4825, 2018.

- [28] X. Chen and G. Y. Yan, "Novel human lncRNA-disease association inference based on lncRNA expression profiles," *Bioinformatics*, vol. 29, no. 20, pp. 2617–2624, 2013.
- [29] X. Chen, D. Xie, Q. Zhao, and Z. H. You, "MicroRNAs and complex diseases: from experimental results to computational models," *Briefings in Bioinformatics*, vol. 20, no. 2, pp. 515–539, 2019.
- [30] X. Chen and L. Huang, "LRSSLMDA: laplacian regularized sparse subspace learning for miRNA-disease association prediction," *PLoS Computational Biology*, vol. 13, no. 12, Article ID e1005912, 2017.
- [31] Z. Zhao, K. Y. Wang, F. Wu et al., "circRNA disease: a manually curated database of experimentally supported circRNA-disease associations," *Cell Death & Disease*, vol. 9, no. 5, p. 475, 2018.
- [32] X. Lei, Z. Fang, L. Chen, and F.-X. Wu, "PWCD: path weighted method for predicting circRNA-disease associations," *International Journal of Molecular Sciences*, vol. 19, no. 11, 2018.
- [33] C. Fan, X. Lei, and F. X. Wu, "Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks," *International Journal of Biological Sciences*, vol. 14, no. 14, pp. 1950–1959, 2018.
- [34] Q. Xiao, J. Luo, and J. Dai, "Computational prediction of human disease-associated circRNAs based on manifold regularization learning framework," *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 6, pp. 2661–2229, 2019.
- [35] Q. Zhao, Y. Yang, G. Ren, E. Ge, and C. Fan, "Integrating bipartite network projection and KATZ measure to identify novel CircRNA-disease associations," *IEEE Transactions on Nanobioscience*, vol. 18, no. 4, pp. 578–584, 2019.
- [36] H. Wei and B. Liu, "iCircDA-MF: Identification of circRNA-disease associations based on matrix factorization," *Briefings in Bioinformatics*, 2019, <https://academic.oup.com/bib/advance-article/doi/10.1093/bib/bbz057/5510697?searchresult=1#>.
- [37] C. Yan, J. Wang, and F.-X. Wu, "DWNN-RLS: Regularized least squares method for predicting circRNA-disease associations," *BMC Bioinformatics*, vol. 19, no. 19, p. 520, 2018.
- [38] W. Ba-Alawi, O. Soufan, M. Essack, P. Kalnis, and V. B. Bajic, "DASPfind: new efficient method to predict drug-target interactions," *Journal of Cheminformatics*, vol. 8, no. 15, 2016.
- [39] Z. H. You, Z. A. Huang, Z. X. Zhu et al., "PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction," *PLoS Computational Biology*, vol. 13, no. 3, Article ID e1005455, 2017.
- [40] X. F. Xiao, W. Zhu, B. Liao et al., "BPLDA: predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network," *Frontiers in Genetics*, vol. 9, p. 411, 2018.
- [41] Z. A. Huang, X. Chen, Z. X. Zhu et al., "PBHMDA: path-based human microbe-disease association prediction," *Frontiers in Microbiology*, vol. 8, p. 233, 2017.
- [42] A. Grover and J. Leskovec, "node2vec: scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining—KDD'16*, pp. 855–864, San Francisco, CA, USA, August 2016.
- [43] W. A. Kibbe, C. Arze, V. Felix et al., "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic Acids Res*, vol. 43, no. D1, pp. D1071–D1078, 2015.
- [44] G. Yu, L.-G. Wang, G.-R. Yan, and Q.-Y. He, "DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis," *Bioinformatics*, vol. 31, no. 4, pp. 608–609, 2015.
- [45] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. F. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274–1281, 2007.
- [46] S. Pletscher-Frankild, A. Palleja, K. Tsafou, J. X. Binder, and L. J. Jensen, "DISEASES: text mining and data integration of disease-gene associations," *Methods*, vol. 74, pp. 83–89, 2015.
- [47] X. Pan, L. J. Jensen, and J. Gorodkin, "Inferring disease-associated long non-coding RNAs using genome-wide tissue expression profiles," *Bioinformatics (Oxford, England)*, vol. 35, no. 9, pp. 1494–1502, 2019.
- [48] J. Piñero, À Bravo, N. Queralt-Rosinach et al., "DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants," *Nucleic Acids Research*, vol. 45, no. D1, pp. D833–D839, 2017.
- [49] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, and S. Wang, "GOSemSim: an R package for measuring semantic similarity among GO terms and gene products," *Bioinformatics*, vol. 26, no. 7, pp. 976–978, 2010.
- [50] Y. Li and J. Li, "Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data," *BMC Genomics*, vol. 13, no. Suppl 7, p. S27, 2012.
- [51] B. Chen, M. Li, J. Wang, X. Shang, and F.-X. Wu, "A fast and high performance multiple data integration algorithm for identifying human disease genes," *BMC Medical Genomics*, vol. 8, no. S3, 2015.
- [52] T. Zhao, J. Xu, L. Liu et al., "Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features," *Molecular Biosystems*, vol. 11, no. 1, pp. 126–136, 2015.
- [53] J. Peng, W. Hui, Q. Li et al., "A learning-based framework for miRNA-disease association identification using neural networks," *Bioinformatics*, vol. 35, no. 21, pp. 4364–4371, 2019.
- [54] T. van Laarhoven, S. B. Nabuurs, and E. Marchiori, "Gaussian interaction profile kernels for predicting drug-target interaction," *Bioinformatics*, vol. 27, no. 21, pp. 3036–3043, 2011.
- [55] A. Valdeolivas, L. Tichit, C. Navarro et al., "Random walk with restart on multiplex and heterogeneous biological networks," *Bioinformatics*, vol. 35, no. 3, pp. 497–505, 2019.

