OXFORD

Briefings in Bioinformatics, 2022, **23(5)**, 1–11
https://doi.org/10.1093/bib/bbac379
Problem Solving Protocol

GraphCDA: a hybrid graph representation learning framework based on GCN and GAT for predicting disease-associated circRNAs

Qiguo Dai, Ziqiang Liu, Zhaowei Wang, Xiaodong Duan and Maozu Guo

Corresponding author: Xiaodong Duan, SEAC Key Laboratory of Big Data Applied Technology, Dalian Minzu University, 116600, Dalian, China, E-mail: duanxd@dlnu.edu.cn; Maozu Guo, School of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, 100044, Beijing, China. E-mail: guomaozu@bucea.edu.cn

Abstract

Motivation: CircularRNA (circRNA) is a class of noncoding RNA with high conservation and stability, which is considered as an important disease biomarker and drug target. Accumulating pieces of evidence have indicated that circRNA plays a crucial role in the pathogenesis and progression of many complex diseases. As the biological experiments are time-consuming and labor-intensive, developing an accurate computational prediction method has become indispensable to identify disease-related circRNAs. Results: We presented a hybrid graph representation learning framework, named GraphCDA, for predicting the potential circRNA–disease associations. Firstly, the circRNA–circRNA similarity network and disease–disease similarity network were constructed to characterize the relationships of circRNAs and diseases, respectively. Secondly, a hybrid graph embedding model combining Graph Convolutional Networks and Graph Attention Networks was introduced to learn the feature representations of circRNAs and diseases simultaneously. Finally, the learned representations were concatenated and employed to build the prediction model for identifying the circRNA–disease associations. A series of experimental results demonstrated that GraphCDA outperformed other state-of-the-art methods on several public databases. Moreover, GraphCDA could achieve good performance when only using a small number of known circRNA–disease associations as the training set. Besides, case studies conducted on several human diseases further confirmed the prediction capability of GraphCDA for predicting potential disease-related circRNAs. In conclusion, extensive experimental results indicated that GraphCDA could serve as a reliable tool for exploring the regulatory role of circRNAs in complex diseases.

Keywords: circRNA–disease association, graph convolutional network, graph attention network, representation learning

Introduction

Circular RNA (circRNA) is a kind of special single-stranded cyclic endogenous noncoding RNA. For a long time, it has been regarded as the product of abnormal shear, which was first discovered in plant viruses as early as 1976 [1]. In recent years, with the development of high-throughput sequencing techniques, a large number of circRNAs have been found in many organisms. The main characteristic of circRNA is that it has a closed-loop structure without a 5' and 3' polyadenylated tails [2]. Existing pieces of evidence have found that circRNAs play crucial roles in cell activity and gene regulation. For example, circRNAs could regulate the radiation sensitivity of esophageal squamous cell carcinomas through microRNA pathway [3]. Moreover, circRNAs are involved in the occurrence and development of many complex diseases, such as cancers, cardiovascular and nervous system diseases [4]. By collecting verified associations between circRNAs and diseases, some critical databases have been proposed, such as circRNADisease [5], circAtlas [6], circ2Disease [7], circFunbase [8], circR2Disease [9] and circR2Disease2.0 [10]. Nevertheless, there are a lot of circRNA–disease associations that have not been verified [11]. Therefore, it is necessary to conduct in-depth research to discover novel disease-associated circRNAs, which could help elucidate the pathogenesis and development mechanism of the diseases.

Recently, many computational methods have been proposed to identify potential circRNA–disease associations. Some of the previous studies applied the machine learning model as the classifier for prediction, such as GBDT [13], RWRKNN [14], etc. These methods predicted disease-related circRNAs based on the features that were extracted from a set of similarity networks of circRNAs and

Qiguo Dai. He received the B.S., M.S. and Ph.D. degrees in computer science and technology from Hubei University of Automotive Technology in 2006, Beijing University of Technology in 2010 and Harbin Institute of Technology in 2015, respectively. Currently, he is an associate professor at the School of Computer Science and Engineering, Dalian Minzu University. His research interests include bioinformatics and data mining.
Ziqiang Liu. He received the B.S. degree in Software engineering from Dalian Jiaotong University in 2020. He is currently pursuing the M.S. degree in the School of Computer Science and Engineering, Dalian Minzu University. His research interests include deep learning and bioinformatics.
Zhaowei Wang. He received the B.S. and M.S. degrees from the School of Computer Science and Engineering, Dalian Minzu University, in 2019 and 2022, respectively. He is currently working toward the Ph.D. degree in the School of Computer Science and Technology, Dalian Unversity of Technology. His research interest includes bioinformatics and machine learning.
Xiaodong Duan. He received the B.S. degrees in computer science and technology from Nankai University, Tianjin, China in 1985, and received the M.S. and Ph.D. degrees in applied mathematics and computer software and theory from Northeastern University, Shenyang, China in 1988 and 2001, respectively. Currently, he is a professor at the School of Computer Science and Engineering, Dalian Minzu University. His research interests include pattern recognition and data mining.
Maozu Guo. He is a professor at the College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing, China. He received the Ph.D. degree in Computer Science and Technology from Harbin Institute of Technology. His research interests include bioinformatics, machine learning and data mining.
Received: May 3, 2022. **Revised:** July 18, 2022. **Accepted:** August 9, 2022
© The Author(s) 2022. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

diseases. Although the works had made great progress, there are still some challenges in learning the feature representation of the circRNA–disease associations.

In the last decade, deep learning has become popular since it has achieved outstanding performances in the fields of image recognition [15], speech recognition [16] and natural language processing [17]. Deep-learning-based models can learn feature representations automatically from the input data without relying on traditional hand-crafted features. Currently, it has been also widely used in bioinformatics. For instance, Wang *et al*. [18] applied convolutional neural network (CNN) to extract high-level features from multi-source information, and then make the prediction of circRNA–disease associations. Deepthi *et al*. [19] employed a deep autoencoder for feature representation of circRNA–disease similarities, which were fed into the deep neural network to predict disease-associated circRNAs. Although the above works based on deep learning have achieved good performance, the deep learning models such as CNN and autoencoder were not suitable for graph data, like circRNA similarity network and disease similarity network. Graph Neural Networks (GNNs) could model the adjacency relationship between nodes by message passing [20]. Graph Convolutional Network (GCN) is a graph-based deep learning method, which utilizes the structural information within the graph to aggregate the information of neighboring nodes. It has achieved excellent performance in the fields of text classification and knowledge graphs [21]. In terms of identifying disease-related noncoding RNA, GCN had been used to predict disease-related miRNAs and long noncoding RNAs (lncRNAs) [22, 23]. With respect to circRNA–disease association prediction, Wang *et al*. [24] employed FastGCN to yeild a unified descriptor by fusing the similarity information of diseases and circRNAs. The extracted high-level features were used to predict the associations between circRNAs and diseases by ForestPA classifier [25]. Although GCN had achieved favorable results, there were still some limitations. For a certain node, the feature importance of its different neighbors were various. The attention mechanism could pay more attention to the important features and suppressed other ones. In graph learning, the graph attention network (GAT) [26] used a multi-head self-attention mechanism to learn the feature representation of nodes on the graph by assigning various weights to different nodes in the neighborhood. It has been employed to identify the miRNA–disease and lncRNA–disease associations with good performances [27, 28]. In the previous works [29, 30], GAT had been also employed to learn the node representations that were used to characterize the circRNA–disease associations. In general, GCN and GAT had been applied to predict circRNA–disease associations separately and they could characterize the representations of nodes from different perspectives. Therefore, more reasonable representations of circRNAs and diseases could be learned by effectively combining the two approaches.

In this work, we proposed a hybrid graph representation learning framework, named GraphCDA, which combined GCN and GAT for predicting circRNA–disease associations. Firstly, two similarity networks were constructed to characterize the relationships of circRNAs and diseases, including circRNA–circRNA similarity network and disease–disease similarity network. Secondly, a hybrid graph embedding model integrating GCN and GAT was put forward to learn the feature representations of circRNAs and diseases simultaneously. Thirdly, the learned representations were merged and employed to build the prediction model that identified disease-associated circRNAs. To validate the performance of the proposed GraphCDA, a series of experimental tests were conducted on several public databases under 5-fold cross-validation. As a result, the area under the receiver operating characteristic curve (AUROC) and the precision-recall curve (AUPR) of GraphCDA achieved 0.9882 and 0.9895, respectively, which were superior to other compared state-of-the-art methods. Moreover, several case studies were carried out on three important human diseases. The results showed that only using a small number of known circRNA–disease associations as the training set, the method's predictions could be well matched to the test database. In addition to CircR2Disease database, many prediction results of the methods were also confirmed by other relevant literature, which demonstrated that the proposed method had good generalization ability. Accordingly, some unconfirmed associations with high prediction probability have potential research value for further experimental verification.

Materials and Methods

In this study, a hybrid graph representation learning framework named GraphCDA was proposed to predict the circRNA–disease associations. The feature representations of circRNAs and diseases were learned, respectively, by using the combinations of GCN and GAT, which were employed to build the classification model to yield the final prediction results. As shown in Figure 1, the proposed framework consisted of the following modules: (i) The circRNA similarity network and disease similarity network were constructed from the CircR2Disease database and the disease ontology; (ii) The feature representations of circRNAs and diseases were learned from the corresponding similarity networks through the hybrid model integrating GCN and GAT, and the outputs of different GCN layers were combined and then taken as the features of circRNAs and diseases, respectively; (iii) The learned features were further concatenated to yield the descriptors of circRNA–disease associations, which were finally used to train a Random Forest classifier for predicting the circRNA–disease associations.

Data preparation and processing

In this study, the widely used CircR2Disease database [9] was employed to build and test GraphCDA. Based on the database, the circRNA similarity and disease similarity were computed and used to construct circRNA–circRNA integrated similarity network and disease–disease integrated similarity network.

The verified circRNA–disease associations

The circRNA–disease association information from the circRNA–Disease [5], circAtlas [6], circ2Disease [7], circFunbase [8], circR2Disease [9] and circR2Disease2.0 [10] databases was downloaded for constructing a circRNA–disease association matrix. The nonhuman and duplicate data were removed from these databases and the number of circRNAs, diseases and association pairs of each database were shown in Table 1. Let R be the circRNA–disease association matrix, in which $R_{i,j} = 1$ if circRNA c_i was confirmed to associate with disease d_j in the database; $R_{i,j} = 0$ otherwise.

Disease Semantic Similarity

To construct the disease similarity network, semantic similarities between different diseases were computed based on disease ontology (DO), in which the diseases were organized with a directed acyclic graph (DAG) [31, 32]. The DO terms of each disease in the database could be retrieved from https://disease-ontology.org/. The semantic similarity of two diseases could be calculated by using the *doSim* function in the DOSE software package [33] and

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

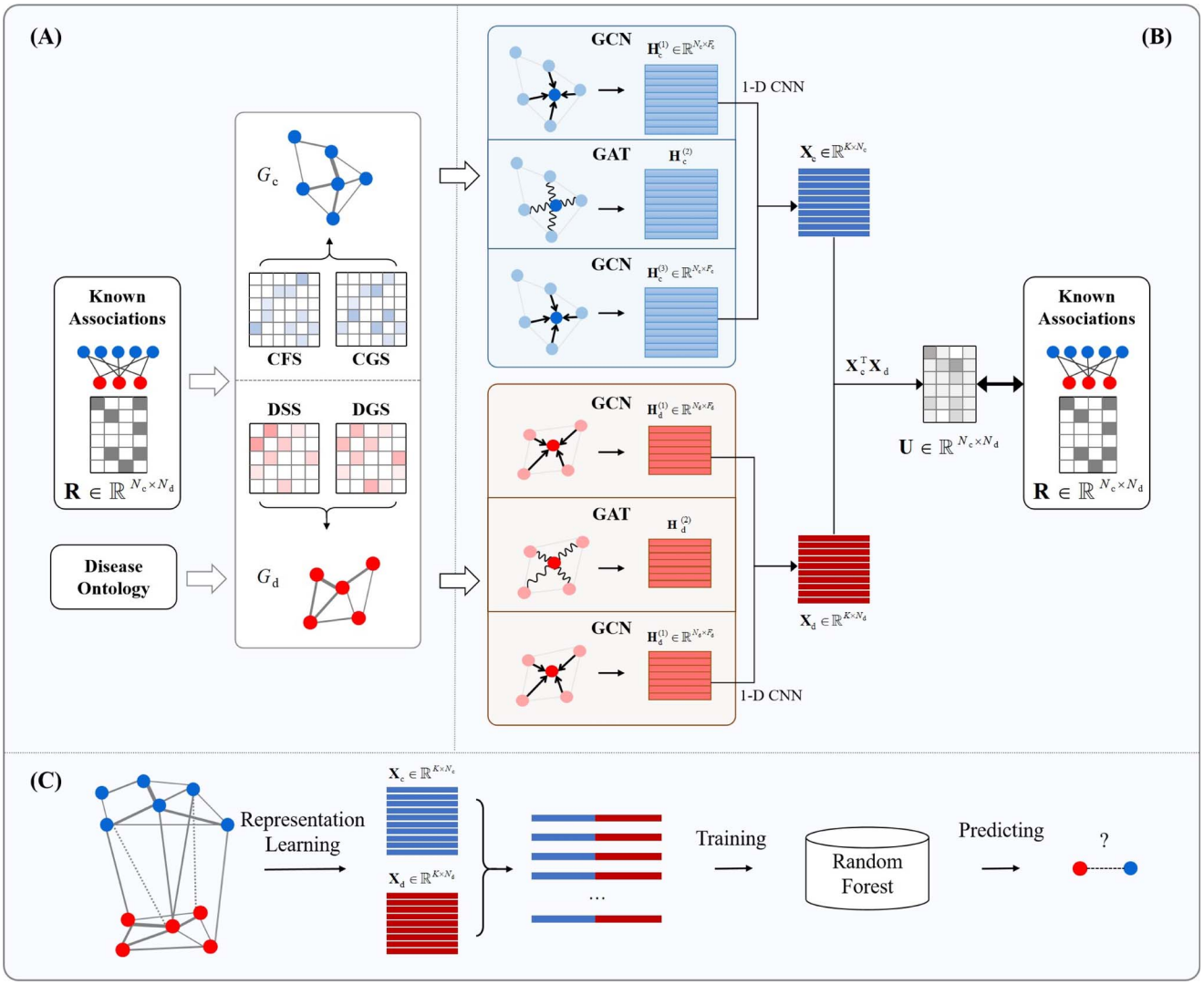

Figure 1. The schematic process of the proposed GraphCDA framework for predicting circRNA–disease associations.

defined as follows:

$$DSS(d_i, d_j) = \frac{\sum_{x \in N_{d_i} \cap N_{d_j}} (S_{d_i}(x) + S_{d_j}(x))}{\sum_{x \in N_{d_i}} S_{d_i}(x) + \sum_{x \in N_{d_j}} S_{d_j}(x)}, \quad (1)$$

where N_{d_i} consists of disease d_i and all its ancestral diseases in the DAG(d_i). $S_{d_i}(x)$ represents the semantic contribution value of disease x to disease d_i, as follows:

$$\begin{cases} S_{d_i}(x) = \max\{\mu * S_{d_i}(x') | x' \in children\ of\ d_i\} & \text{if } x \neq d_i \\ S_{d_i}(d_i) = 1 & \text{otherwise} \end{cases} \quad (2)$$

Similar to related works [31], the semantic contribution factor μ was set to 0.5.

CircRNA Functional Similarity

In the previous related works [4, 40], functionally similar circRNAs were assumed more likely to be associated with phenotypically similar diseases and vice versa. Therefore, the circRNA functional similarity was also employed in our study, which was obtained from the disease semantic similarity and circRNA–disease association data. Similar to the previous methods [31, 34], the functional similarity between circRNA c_i and c_j could be computed as follows:

$$CFS(c_i, c_j) = \frac{\sum_{1 \le q \le |D_i|} DS(d_q, D_j) + \sum_{1 \le r \le |D_j|} DS(d_r, D_i)}{|D_i| + |D_j|}; \quad (3)$$

$$DS(d_q, D_j) = \max_{1 \le t \le |D_j|} (DSS(d_q, d_t)), \quad (4)$$

where D_i represents the disease set associated with circRNA c_i, and $DS(d_r, D_i)$ represents the semantic similarity between the disease d_r and D_i.

Gaussian interaction profile kernel similarity of circRNAs and diseases

Since there were some diseases that were not well annotated in DO, the corresponding disease semantic similarity and the functional similarity of their associated circRNAs could not be calculated using aforementioned method. The missed disease semantic similarity and circRNA functional similarity could be

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

Table 1. Details of the circRNA–disease association databases used in this work.

Database	circRNAs	Diseases	Associations
circRNAdisease	331	48	354
circAtlas	776	110	854
circ2Disease	237	60	273
CircFunbase	2597	64	2984
circR2Disease	585	88	650
circR2Disease2.0	3077	307	4201

replaced by Gaussian interaction profile kernel similarity (GIP) [31, 35]. Let $IP(c_i)$ denote the binary interaction profile vector of circRNA c_i, which corresponds to the i-th row in the adjacency matrix R. The GIP similarity between circRNA c_i and c_j can be calculated as follows:

$$\mathrm{CGS}(c_i, c_j) = \exp(-\rho||IP(c_i) - IP(c_j)||^2); \tag{5}$$

$$\rho = \frac{1}{\frac{1}{N_c}\sum_{i=1}^{N_c}||IP(c_i)||^2}, \tag{6}$$

where ρ is the parameter controlling the kernel bandwidth and N_c is the number of rows of the adjacency matrix R.

Similarly, the GIP kernel similarity between disease d_i and d_j could be denoted as

$$\mathrm{DGS}(d_i, d_j) = \exp(-\rho||IP(d_i) - IP(d_j)||^2); \tag{7}$$

$$\rho = \frac{1}{\frac{1}{N_d}\sum_{i=1}^{N_d}||IP(d_i)||^2}, \tag{8}$$

where $IP(d_i)$ denotes the binary interaction profile vector of disease d_i, which corresponds to the i-th column in the adjacency matrix R.

Therefore, the integrated similarity matrices of circRNA (C) and disease (D) could be obtained and the elements of them were denoted as follows:

$$C_{i,j} = \begin{cases} CGS(c_i, c_j) & \text{if } CFS(c_i, c_j) = 0 \\ CFS(c_i, c_j) & \text{otherwise;} \end{cases} \tag{9}$$

$$D_{i,j} = \begin{cases} DGS(d_i, d_j) & \text{if } DSS(d_i, d_j) = 0 \\ DSS(d_i, d_j) & \text{otherwise,} \end{cases} \tag{10}$$

where the similarity between circRNA i and j in the matrix $C \in R^{N_c \times N_c}$ was taken as the edge weight of nodes i and j in the circRNA similarity network G_c. The similarity between disease i and j in the matrix $D \in R^{N_d \times N_d}$ was the edge weight of disease i and j in the similarity network G_d.

Graph representation learning of circRNAs and diseases

Two graph learning modules consisting of a set of GCN and GAT layers were employed to learn the feature representations from the similarity networks of circRNAs and diseases, respectively. As shown in Figure 1(B), each graph learning module contained two graph convolutional layers with a graph attention layer between them; and a fusion layer was used to integrate the outputs of the different graph convolutional layers. The details are illustrated in the following sections.

Graph Convolutional Network

GCN was employed to obtain the feature representations of circRNAs and diseases from circRNA similarity network and disease similarity network, respectively. Given a network G, its adjacency matrix and input node representations could be denoted as $S \in R^{N \times N}$ and $H \in R^{N \times F}$, where N was the number of nodes and F was the dimension of node feature. The output node representations H^{new} could be obtained by a GCN layer as follows:

$$H^{new} = GCN(S, H) \tag{11}$$

$$GCN(S, H) = \sigma\left(\mathrm{A}^{-\frac{1}{2}}\widetilde{\mathrm{S}}\mathrm{A}^{-\frac{1}{2}}HQ\right), \tag{12}$$

where $\widetilde{S} = I + S$; $\mathrm{A} = \sum_j \widetilde{S}_{i,j}$ is the degree matrix; $Q \in R^{F \times F}$ is the trainable weight matrix; and $\sigma(\cdot)$ is the ReLU activation function.

Graph Attention Network

GAT [26] is a type of neural network that employs the multi-head attention to assign different weights to adjacent nodes according to their importances. In GraphCDA, the GAT layer was introduced between two GCN layers which aim to assist the following GCN layer to extract high-level features of circRNAs and diseases.

For the network G, the output node representations H^{new} of a GAT layer was as follows:

$$H^{new} = GAT(S, H) \tag{13}$$

$$\vec{H}_i^{new} = \sigma\left(\frac{1}{K}\sum_{k=1}^{K}\sum_{j \neq i}\phi_{ij}^k W_k \vec{H}_i\right), \tag{14}$$

where $\vec{H}_i^{new}$ denotes the feature representation of node i in H^{new}; K is the number of attention mechanisms in multi-head attention; W_k is the weight matrix of the k-th attention mechanism; $\vec{H}_i$ is the input feature vector of the circRNA node i; ϕ_{it}^k is denoted as the k-th attention coefficients between nodes i and t.

$$\phi_{ij}^k = \frac{\exp(\mathrm{LR}(a_k^T[W_k\vec{H}_i||W_k\vec{H}_j||B_kS_{ij}]))}{\sum_{t \neq i}\exp(\mathrm{LR}(a_k^T[W_k\vec{H}_i||W_k\vec{H}_t||B_kS_{it}]))}, \tag{15}$$

where $||$ is a concatenation operation; LR is the LeaklyReLU activation function; $a_k \in R^{2F+1}$ is the weight vector of the k-th attention mechanism; and B_k is the weight of the edge S_{ij} that is to be learned.

Representation learning modules of circRNAs and diseases

Based on the above GCN and GAT layers, representation learning modules for circRNAs and diseases could be constructed, which were used to construct feature representations of the nodes from corresponded similarity networks, respectively.

For the input circRNA similarity network G_c, its adjacency matrix was denoted as C, in which $C_{i,j} \in C$ was as in eq. (9). And $H_c^{(0)} \in R^{N_c \times F_c}$ was the input node features of the network, where F was the dimension of circRNA feature. The aforementioned GCN

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

and GAT layers were introduced alternately to obtain the graph feature representation for different levels of circRNA nodes as follows:

$$\begin{cases} H_c^{(1)} = GCN\left(C, H_c^{(0)}\right) \\ H_c^{(2)} = GAT\left(C, H_c^{(1)}\right) \\ H_c^{(3)} = GCN\left(C, H_c^{(2)}\right) \end{cases} \tag{16}$$

To combine the output features $H_c^{(1)}$ and $H_c^{(3)}$ of different GCN layers, a 1D CNN was employed to yield the circRNA representations X_c.

Similarly, we also utilized GCN and GAT to learn multi-level node features $H_d^{(1)}$, $H_d^{(2)}$ and $H_d^{(3)}$ from disease similarity network G_d. The adjacency matrix of G_d was denoted as D, in which $D_{i,j} \in D$ was as in eq. (10). The initial features of the disease network was $H_d^{(0)} \in R^{N_d \times F_d}$.

$$\begin{cases} H_d^{(1)} = GCN\left(D, H_d^{(0)}\right) \\ H_d^{(2)} = GAT\left(D, H_d^{(1)}\right) \\ H_d^{(3)} = GCN\left(D, H_d^{(2)}\right) \end{cases} \tag{17}$$

As mentioned above, a 1D CNN was also used to integrate the outputs of GCN layers $\{H_d^{(1)}, H_d^{(3)}\}$ to obtain the disease representations X_d.

Model training of Representation Learning

Based on the representations of X_c and X_d, the predicted preference matrix U of circRNA–disease could be obtained as

$$U = X_c{}^{\mathrm{T}} X_d \tag{18}$$

The higher the U_{ij} in U, the greater the possibility that circRNA i was related to disease j. The binary cross-entropy (BCE) was employed to measure the difference between the preference matrix U and known adjacency matrix R, which was taken as the loss function for training the graph representation learning model. By minimizing the loss function above on training database of circRNA–disease associations, the graph representation matrices of X_c and X_d could be learned, which were used for predicting novel circRNA–disease associations.

Predicting circRNA–disease associations based on GraphCDA

To build a prediction model identifying circRNA–disease associations, the training set was constructed from the CircR2Disease database, in which the known associated pairs were taken as positive samples, and the equal number of unconfirmed circRNA–disease associations pairs were randomly selected as negative samples.

The feature representations of circRNAs and diseases learned by GraphCDA were concatenated to form the fusion descriptors for circRNA–disease pairs. The fusion descriptor of a pair of circRNA i and disease j was as follows:

$$z_{ij} = [X_c(i), X_d(j)], \tag{19}$$

where $X_c(i)$ represents the i-th row in the circRNA feature matrix X_c and $X_d(j)$ represents the j-th row in the feature matrix X_d.

Random Forest (RF) was a powerful ensemble classifier, which utilized multiple decision trees to alleviate the overfitting problem against the training data [44, 46]. It had been widely employed to solve the prediction problems in the field of bioinformatics [47, 48]. In our previous study [49], RF was also used to predict miRNA–disease associations. Therefore, in this work, RF was also trained based on the above fusion descriptors to predict disease-related circRNAs.

Results

To validate the performance of the proposed GraphCDA, a series of experimental tests have been conducted on several public databases. Firstly, the experimental settings and performance evaluation metrics were introduced. Secondly, ablation experiments were carried out to test the effectiveness of the hybrid graph representation learning model in GraphCDA. Thirdly, the prediction performance of the GraphCDA using different circRNA similarities was evaluated on several public databases. Fourthly, the prediction performance of GraphCDA using different classifiers were also tested. Lastly, the proposed method was compared with other state-of-the-art methods, including the methods based on GNNs.

Experimental settings and performance evaluation

In the following experiments, 5-fold cross-validation was used to evaluate the performance and several evaluation metrics including Accuracy (Acc.), Precision (Pre.), Sensitivity (Sen.), F1-Score (F1), Matthews Correlation Coefficient (MCC), the Area Under the Receiver Operating Characteristics Curve (AUROC) and the Precision-Recall Curve (AUPR) were taken as in related works [50–52]. Unless otherwise noted, the following experiments were tested on the human circRNA–disease associations in CircR2Disease database under 5-fold cross-validation. To ensure the accuracy of the experimental results, the following test results were the average value of 100 runs. The experimental results of GraphCDA were summarized in Table 2, in which the performance of 5-fold cross-validation was shown. With respect to the average of the 5 folds, GraphCDA achieved 0.9458, 0.9456, 0.9462, 0.9457 and 0.8917, in terms of accuracy, precision, sensitivity, F1-score and MCC on the tested database, respectively. The mean values of AUROC and AUPR reached 0.9882 and 0.9895. In addition, the cross-validation experiments with other different folds were carried out, and the detailed results were listed in Supplementary Tables S2–S3. These experimental results demonstrated that GraphCDA performed well on the tested database and could effectively predict potential circRNA–disease associations.

Ablation experiments

In order to verify the effectiveness of the hybrid graph learning strategies introduced in GraphCDA, we disassembled GraphCDA into two models, namely GraphCDA-nogat and GraphCDA-last. GraphCDA-nogat referred to the method that removed the GAT between the two GCN layers. GraphCDA-last was the version that only used the output of the last GCN layer for feature extraction without using the GAT layer and 1D CNN combiner. The obtained node feature matrices of circRNAs and diseases after training were fed into Random Forest classifier for prediction, and 100 times of 5-fold cross-validation were carried out on circRNADisease [5], circAtlas [6], circ2Disease [7], circFunbase [8], circR2Disease [9] and circR2Disease2.0 [10]. The average testing results were shown in Table 3 and the confusion matrices

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

Table 2. The 5-fold cross-validation testing results of GraphCDA on CircR2Disease

Validation set	Accuracy	Precision	Sensitivity	F1-Score	MCC	AUROC	AUPR
1	0.9398	0.9399	0.9391	0.9392	0.8798	0.9848	0.9865
2	0.9438	0.9455	0.9433	0.9441	0.8879	0.9877	0.9892
3	0.9488	0.9474	0.9495	0.9482	0.8978	0.9892	0.9901
4	0.9479	0.9465	0.9496	0.9479	0.8958	0.9903	0.9914
5	0.9486	0.9486	0.9495	0.9488	0.8974	0.9891	0.9904
Avg.	**0.9458 ± 0.35**	**0.9456 ± 0.30**	**0.9462 ± 0.43**	**0.9457 ± 0.36**	**0.8917 ± 0.70**	**0.9882 ± 0.19**	**0.9895 ± 0.17**

Table 3. Ablation experiment results of GraphCDA on all six databases

Database	Model	Accuracy	Precision	Sensitivity	F1-Score	MCC	AUROC	AUPR
circRNADisease	GraphCDA	**0.9297**	**0.9367**	**0.9223**	**0.9288**	**0.8600**	**0.9823**	**0.9850**
	GraphCDA-nogat	0.9284	0.9351	0.9213	0.9275	0.8574	0.9818	0.9845
	GraphCDA-last	0.9165	0.9198	0.9128	0.9157	0.8334	0.9774	0.9808
circAtlas	GraphCDA	**0.9487**	**0.9466**	**0.9512**	**0.9487**	**0.8976**	**0.9892**	**0.9907**
	GraphCDA-nogat	0.9466	0.9437	0.9500	0.9467	0.8934	0.9887	0.9903
	GraphCDA-last	0.9257	0.9134	0.9407	0.9266	0.8519	0.9829	0.9848
circ2Disease	GraphCDA	**0.9203**	**0.9251**	**0.9153**	**0.9195**	**0.8412**	**0.9757**	**0.9798**
	GraphCDA-nogat	0.9155	0.9165	0.9150	0.9150	0.8316	0.9739	0.9782
	GraphCDA-last	0.8933	0.8903	0.8978	0.8932	0.7872	0.9631	0.9689
circFunbase	GraphCDA	**0.9787**	**0.9806**	**0.9768**	**0.9787**	**0.9575**	**0.9984**	**0.9984**
	GraphCDA-nogat	0.9782	0.9805	0.9759	0.9782	0.9565	0.9983	0.9983
	GraphCDA-last	0.8771	0.8716	0.8850	0.8780	0.7547	0.9502	0.9525
circR2Disease	GraphCDA	**0.9458**	**0.9456**	**0.9462**	**0.9457**	**0.8917**	**0.9882**	**0.9895**
	GraphCDA-nogat	0.9453	0.9449	0.9459	0.9451	0.8907	0.9880	0.9892
	GraphCDA-last	0.9171	0.9131	0.9220	0.9173	0.8343	0.9786	0.9812
circR2Disease2.0	GraphCDA	**0.9377**	**0.9375**	**0.9381**	**0.9377**	**0.8755**	**0.9871**	**0.9881**
	GraphCDA-nogat	0.9362	0.9356	0.9369	0.9362	0.8724	0.9870	0.9879
	GraphCDA-last	0.8655	0.8623	0.8700	0.8660	0.7311	0.9423	0.9430

were listed in Supplementary Table S1. The AUROC values of GraphCDA-nogat were 0.9818, 0.9887, 0.9739, 0.9983, 0.9980 and 0.9870, respectively. The AUROC values of GraphCDA-last were 0.9774, 0.9829, 0.9631, 0.9502, 0.9786 and 0.9423, respectively. It could be found that GraphCDA-nogat significantly outperformed GraphCDA-last on all metrics, which demonstrated that combining the node features outputted from different layers of GCN could integrate various levels of representations effectively. For the comparison of GraphCDA and GraphCDA-nogat, it could be seen that GraphCDA outperformed GraphCDA-nogat on all metrics. It indicated that the GAT layer in GraphCDA could improve the representation learning ability for circRNAs and diseases in GraphCDA effectively.

Effectiveness of different circRNA similarity for GraphCDA

To investigate the impact of different circRNA similarities on the performance of GraphCDA, we compared our method with the model using only circRNA functional similarity (GraphCDA-func) and the model using only circRNA GIP kernel similarity (GraphCDA-gip). The 100 times of 5-fold cross-validation experiments were implemented on circRNADisease, circAtlas, circ2Disease, circFunbase, circR2Disease and circR2Disease2.0, respectively, and the average test results were shown in Table 4. The results showed that the performance of GraphCDA achieved better performance than GraphCDA-func and GraphCDA-gip on most databases, which demonstrated that integrating the two types of circRNA similarities was conducive to represent the relationships among circRNAs and predict circRNA–disease associations. In addition, it could be also found that the performance of GraphCDA-func was slightly better than that of GraphCDA on circFunBase and circR2Disease2.0. The reason might be that there were a large number of circRNAs in the two databases, resulting in less missing values in the functional similarity of circRNAs. It suggested that on large data sets, better performance could be achieved by using functional similarity alone. Generally, in this work, in order to adapt to most cases, we utilized the integrated similarity to calculate the similarities among circRNAs.

Comparison of GraphCDA using different classifiers

In order to determine the most suitable classifier for GraphCDA, various machine-learning-based models were tested in this section, such as Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), Adaptive Boosting (AB) and Naive Bayesian (NB). Specifically, we validated the performances of GraphCDA frameworks using different classifiers for circRNA–disease association prediction. All of these algorithms were implemented with scikit-learn toolkit [45] and the important parameters of these algorithms were tuned. For RF, the number of estimators ranged from 100 to 1000 with the step of 100 and the max depth was from 2 to 20 with step size of 2. For SVM [46], C ranged in {$1e-5, 1e-4, \ldots, 10\,000$} and gamma was in {$1e-7, 1e-6, \ldots, 0.1$}. For DT, the parameter of minimum samples of a split was from 2 to 10, and the max depth was tuned from 2 to 20. For LR, the parameter of C was {$1e-5, 1e-4, \ldots, 10\,000$}; and the max iterations was range from 50 to 500 with an interval of 50. For AB,

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

Table 4. The comparison results of GraphCDA using different circRNA similarities on all six databases

Database	Model	Accuracy	Precision	Sensitivity	F1-Score	MCC	AUROC	AUPR
circRNADisease	GraphCDA	**0.9297**	0.9367	**0.9223**	**0.9288**	**0.8600**	**0.9823**	**0.9850**
	GraphCDA-func	0.9269	**0.9381**	0.9148	0.9257	0.8544	0.9806	0.9234
	GraphCDA-gip	0.9207	0.9299	0.9109	0.9195	0.8424	0.9791	0.9822
circAtlas	GraphCDA	**0.9487**	**0.9466**	**0.9512**	**0.9487**	**0.8976**	**0.9892**	**0.9907**
	GraphCDA-func	0.9458	0.9443	0.9476	0.9458	0.8917	0.9887	0.9902
	GraphCDA-gip	0.9408	0.9423	0.9392	0.9405	0.8818	0.9860	0.9881
circ2Disease	GraphCDA	**0.9203**	**0.9251**	**0.9153**	**0.9195**	**0.8412**	**0.9757**	**0.9798**
	GraphCDA-func	0.9165	0.9206	0.9122	0.9157	0.8336	0.9731	0.9780
	GraphCDA-gip	0.9138	0.9160	0.9121	0.9132	0.8284	0.9658	0.9736
circFunbase	GraphCDA	0.9787	0.9806	0.9768	0.9787	0.9575	**0.9984**	**0.9984**
	GraphCDA-func	**0.9797**	**0.9819**	**0.9774**	**0.9796**	**0.9594**	0.9983	**0.9984**
	GraphCDA-gip	0.9768	0.9782	0.9754	0.9768	0.9536	0.9981	0.9981
circR2Disease	GraphCDA	**0.9458**	0.9456	**0.9462**	**0.9457**	**0.8917**	**0.9882**	**0.9895**
	GraphCDA-func	0.9448	**0.9482**	0.9412	0.9444	0.8898	0.9868	0.9887
	GraphCDA-gip	0.9311	0.9369	0.9249	0.9305	0.8626	0.9821	0.9851
circR2Disease2.0	GraphCDA	0.9377	0.9375	0.9381	0.9377	0.8755	0.9871	0.9881
	GraphCDA-func	**0.9408**	**0.9390**	**0.9429**	**0.9409**	**0.8816**	**0.9881**	**0.9891**
	GraphCDA-gip	0.9364	0.9290	0.9452	0.9370	0.8731	0.9862	0.9872

Table 5. The comparison results of GraphCDA using different classifiers on circR2Disease

Classifiers	Accuracy	Precision	Sensitivity	F1-Score	MCC	AUROC	AUPR
Random forest	**0.9458**	**0.9456**	0.9462	**0.9457**	**0.8917**	**0.9882**	**0.9895**
SVM	0.9413	0.9238	**0.9618**	0.9422	0.8835	0.9844	0.9742
Decision Tree	0.8906	0.8604	0.9321	0.8944	0.7842	0.8967	0.8451
Logistic Regression	0.7288	0.7161	0.7592	0.7360	0.4595	0.7791	0.6957
Adaptive Boosting	0.6947	0.6993	0.6860	0.6910	0.3913	0.7495	0.6921
Naive Bayes	0.6412	0.6529	0.6082	0.6274	0.2842	0.6762	0.6623

the number of estimators was from 100 to 1000, and the step size was 100; the learning rate ranged from 0.1 to 1 with an interval of 0.1. For LR, the parameter of C was $\{1e-5, 1e-4, ..., 10\,000\}$; and the max iterations was range from 50 to 500 with an interval of 50. For NB, the GaussianNB model was employed. All of these algorithms were tested 100 times under 5-fold cross-validation on the CircR2Disease database.

As shown in Table 5, the average AUROC values for RF, SVM, DT, LR, AB and NB were 0.9882, 0.9844, 0.8967, 0.7791, 0.7495 and 0.6762, respectively. It could be found that Random Forest achieved the best results in terms of Accuracy, Precision, F1-score, MCC, AUROC and AUPR. In terms of overall performance, Random Forest outperformed the other classification models. As shown in Figure 2, the GraphCDA using RF as the final classifier was also significantly better than other methods with respect to the distribution of the AUROC values. It indicated that the Random Forest classifier was more suitable for the GraphCDA framework than other tested algorithms for predicting the circRNA–disease associations. Therefore, the Random Forest was taken as the classifier in the proposed GraphCDA framework.

Comparison of GraphCDA with state-of-the-art methods

In this section, the proposed GraphCDA was compared with other six state-of-the-art methods for predicting circRNA–disease associations, including NCPCDA [40], iCircDA-MF [42], AE-RF [4], GATNNCDA [30], CRPGCN [55] and GMNN2CD [20]. Among them, CRPGCN, GMNN2CD and GATNNCDA all used graph representation learning methods in their models, and the rest models used other machine learning methods. The performance advantages of GraphCDA could be fully demonstrated by selecting various models using different machine learning methods to compare with GraphCDA. The performances of these methods were reported in the corresponding literature, which were all yielded by conducting on the CircR2Disease database under 5-fold cross-validation. Although the evaluation metrics reported in different literature were various, the evaluation results against to AUROC were provided by most methods, which was considered to be an comprehensive performance metric for predicting circRNA–disease associations. Therefore, we compared the AUROC values between GraphCDA and different methods to evaluate GraphCDA. For CRPGCN, the testing results was obtained from its report [55]. The comparison results between GraphCDA and other methods on human data in CircR2Disease are shown in Table 6.

It could be found that the proposed GraphCDA outperformed other tested methods in term of AUROC. It should be noted that GATNNCDA used the GAT for learning the node representations, and it achieved an AUROC of 0.9613 which was 0.0269 less than that of the GraphCDA. Among these methods, CRPGCN employed GCN but not using GAT, which achieved AUROC of 0.9720. In addition, we compared our method with GMNN2CD [20] on five databases, and the results were shown in Table 7. It could be seen that the AUROC values of GraphCDA in the five databases are all higher than that of GMNN2CD.

In general, the above comparisons showed that GraphCDA outperformed the other six prediction methods on the tested databases, which indicated that the proposed GraphCDA had excellent performance in predicting circRNA–disease

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

Table 6. The comparison results of GraphCDA with other state-of-the-art methods on CircR2Disease

Methods	GraphCDA	CRPGCN	GATNNCDA	AE-RF	NCPCDA	iCirDA-MF
AUROC	**0.9882**	0.9720	0.9613	0.9486	0.9201	0.9178

Table 7. The comparison results of GraphCDA with GMNN2CD on five databases

Database	AUROC	
	GraphCDA	GMNN2CD
circR2Disease	**0.9882**	0.9634
circ2Disease	**0.9757**	0.9616
circRNAdisease	**0.9823**	0.9773
circAtlas	**0.9892**	0.9428
CircFunbase	**0.9984**	0.9674

associations. Moreover, compared with other methods based on GNN, it could be found that the combination of GCN and GAT employed in GraphCDA could achieve better prediction performance than using them alone.

Case Studies

How to utilize the limited experimentally verified circRNA–disease association data to identify novel disease-related circRNAs accurately is a hot issue at present. Here, in order to evaluate the generalization ability of the proposed GraphCDA method when using a small amount of known circRNA–target disease association data, we conducted some case studies on three common human diseases including breast cancer, glioma and gastric cancer. For each investigated disease, we employed 0%, 5%, 10%, 15% and 20% of the associations related to the disease and the samples of other diseases in CircR2Disease database as the training set. Then, the candidate circRNAs were ranked according to the predicted probability scores in a descending order, and the top 15 candidates of each investigated disease were selected, respectively. The proportion of the known associated circRNAs in the top 15 candidates of each investigated disease is shown in Figure 3, which is the average of the results obtained by conducting 100 experiments. As shown, GraphCDA could effectively identify novel circRNA–disease associations, even if only a few known association data were used as training set. For example, when only using 5% of the known associations for training, 82.93, 79.40 and 90.73 percentages of the top 15 predictions for breast cancer, glioma and gastric cancer were matched to the rest known associations in CircR2Disease. Even without using any known associations for a specific disease, 22.60%, 12.67% and 29.13% of the top 15 predicted results for the three diseases were in the database. It demonstrated that GraphCDA performed good generalization ability when the number of known associations was limited.

In addition, we also validate the ability of GraphCDA to predict novel circRNA–disease associations. To this end, all of the associations in the CircR2Disease were used for training the GraphCDA and the top 15 circRNAs related to breast cancer, glioma and gastric cancer predicted by the framework were studied, which are shown in Table 8. Among the top 15 predicted circRNAs for each disease, 9, 4 and 9 circRNAs had been validated to be associated with breast cancer, glioma and gastric cancer by relevant literature, respectively. For example, the gene of circPTK2 was confirmed to be associated with breast cancer in 13 cell lines by literature [56]. Furthermore, the top 50 circRNAs for each disease predicted by GraphCDA were uploaded to https://github.com/Ziqiang-Liu/Predict. For those circRNAs that have not been confirmed to be associated with the target disease but with high predictive probabilities, it is worthy to be taken as candidate

Table 8. Prediction of the top 15 candidate circRNAs related to breast cancer, glioma and gastric cancer by GraphCDA

Disease	circRNA	Probability	PMID
Breast cancer	circGFRA1	0.95	29037220
	hsa:circ_0001649	0.92	34285509
	hsa:circ_0000518	0.90	–
	hsa:circ_0001946	0.90	30884120
	Cir-ITCH	0.85	30509108
	circZFR	0.83	32831653
	hsa:circ_0000096	0.76	–
	hsa:circ_0001445	0.75	31446897
	hsa:circ_0000520	0.73	32002039
	circPRKCI	0.72	–
	circPTK2	0.70	33436041
	hsa:circ_0041103	0.70	31446897
	hsa:circ_0007158	0.70	–
	hsa:circ_0082582	0.70	–
	hsa:circ_0061265	0.70	–
Glioma	hsa:circ_0001946	0.88	26683098
	hsa:circ_0000518	0.86	–
	hsa:circ_0001313	0.84	–
	hsa:circ_0002113	0.83	–
	circHIPK3	0.81	30576808
	Cir-ITCH	0.80	–
	circSMARCA5	0.80	26873924
	hsa:circ_0003570	0.75	–
	hsa:circ_0072088	0.73	–
	circPRKCI	0.70	–
	hsa:circRNA_104075	0.68	31112718
	circC3P1	0.68	–
	hsa:circ_0067531	0.68	–
	hsa:circ_0085154	0.68	–
	hsa:circRNA_0007874	0.68	–
Gastric cancer	circCCDC66	0.98	32253030
	hsa:circ_0007534	0.95	31446897
	Cir-ITCH	0.94	33499704
	circZFR	0.94	29361817
	hsa:circ_0000518	0.94	–
	hsa:circ_0002113	0.91	–
	hsa:circ_0004771	0.88	33116589
	hsa:circ_100721	0.87	31446897
	hsa:circ_0013339	0.87	–
	circFUT8	0.86	28831102
	hsa:circ_0007915	0.86	–
	hsa:circ_0006528	0.84	27986464
	hsa:circ_0093859	0.84	31446897
	circRNA-000911	0.84	–
	circRNA-001283	0.84	–

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

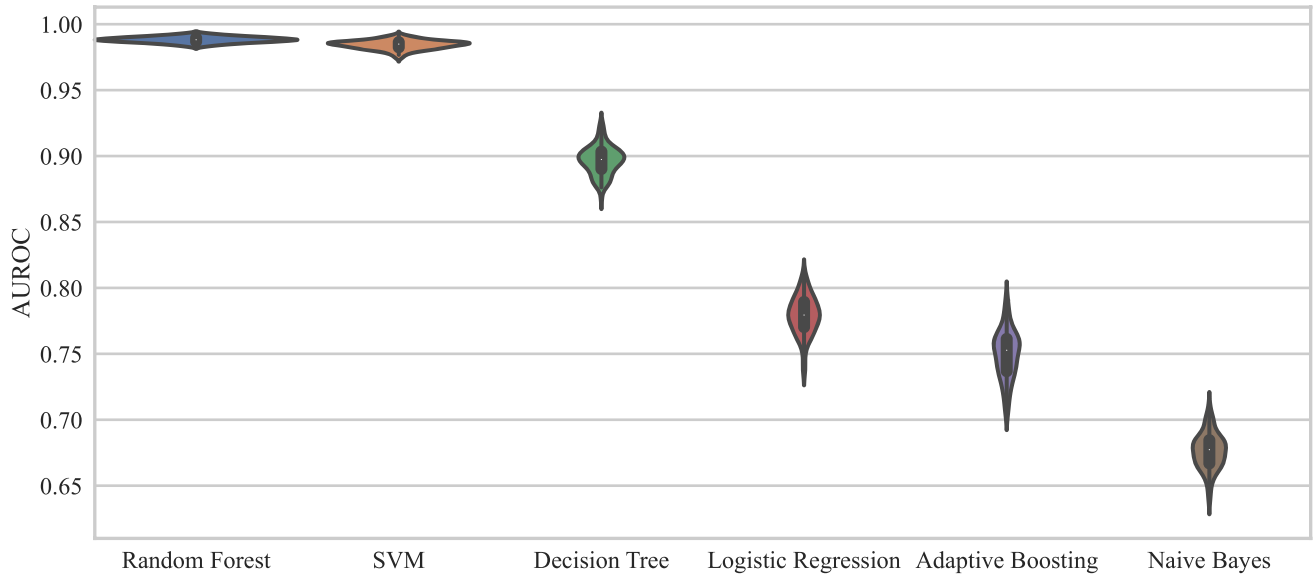

Figure 2. Violin plot of the AUROC results for GraphCDAs using different classifiers.

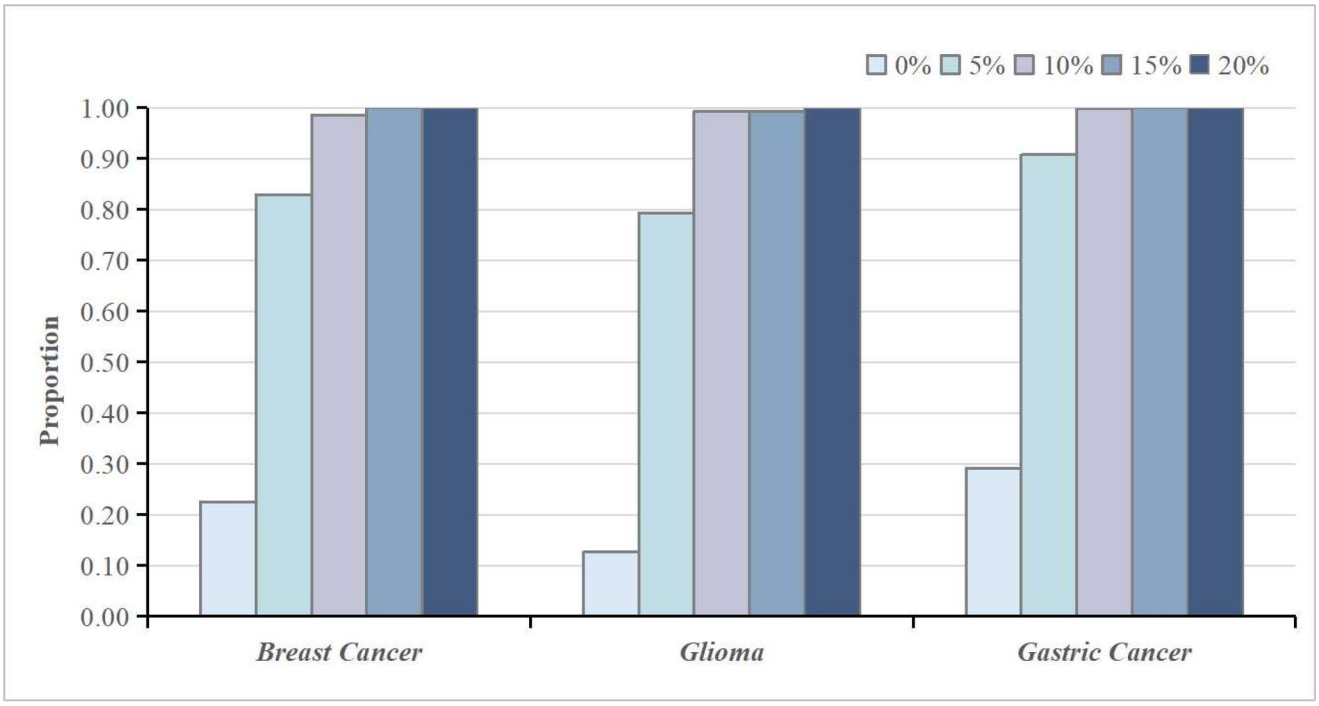

Figure 3. Proportions of known associated circRNAs of the investigated diseases in top 15 predictions of GraphCDA.

associated ones for conducting further experimental verification. In sum, the above studies showed that the proposed GraphCDA had excellent predictive performance in predicting novel circRNA–disease associations.

Conclusion

Predicting circRNA–disease associations using computational methods is of great importance to understand the roles of circRNAs in pathological mechanisms, diagnosis and treatment of human complex diseases. In this study, a hybrid graph representation learning framework by combining GCN and GAT was put forward to identify the potential disease associated circRNAs, named GraphCDA. Firstly, the circRNA similarity network and disease similarity network were built by integrating various similarities. Secondly, GCN and GAT were combined to learn the feature representations of circRNAs and diseases effectively. The node features learned by different GCN layers were integrated with a 1D CNN. Finally, based on the learned representations of circRNAs and diseases, the random forest based model was trained for predicting the circRNA–disease associations. In order to validate the performance of GraphCDA, a series of experiments were carried on several public databases under 5-fold cross-validation. We compared GraphCDA with other state-of-the-art methods on several public databases and the results showed that the proposed method achieved better performance than the other state-of-the-art methods. Moreover, case studies on three common diseases demonstrated

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

that GraphCDA performed good generalization capability when using only a few known related circRNAs as the training set. In general, the GraphCDA proposed in this work was an effective and accurate approach for predicting potential circRNA–disease associations.

Key Points

- Predicting disease-related circRNAs facilitated the diagnosis and treatment of complex human diseases.
- We proposed a hybrid graph representation learning framework by combining GCN and GAT (GraphCDA) to obtain the feature representations of circRNAs and diseases, which were used to predict potential circRNA–disease associations.
- Experimental results on several public databases showed that the GraphCDA achieved better performance than the other state-of-the-art methods. Case studies of three human diseases further confirmed that our proposed method had great potential for predicting novel circRNA–disease associations.

Acknowledgments

The authors would like to appreciate Xiujuan Lei and his colleagues for providing data and valuable help for our research.

Funding

This work has been supported by National Natural Science Foundation of China (Grant No. 61701073 and 62031003) and the High-level Talent Innovation Support Program of Dalian City (No. 2020RQ059).

Data availability

The source code and databases are available at (https://github.com/Ziqiang-Liu/GraphCDA).

References

1. 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* 1976;**73**(11):3852–6.
2. Qu S, Yang X, Li X, *et al.* Circular RNA: a new star of noncoding RNAs. *Cancer Lett* 2015;**365**(2):141–8.
3. Liu J, Xue N, Guo Y, *et al.* CircRNA_100367 regulated the radiation sensitivity of esophageal squamous cell carcinomas through miR-217/Wnt3 pathway. *Aging (Albany NY)* 2019;**11**(24):12412.
4. Deepthi K, Jereesh AS. Inferring potential CircRNA-disease associations via deep autoencoder-based classification. *Mol Diagn Ther* 2021;**25**(1):87–97.
5. Zhao Z, Wang K, Wu F, *et al.* circRNA disease: amanually curated database of experimentally supported circRNA-disease associations. *Cell Death Dis* 2018;**9**:1–2.
6. Wu W, Ji P, Zhao F. CircAtlas: An integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. *Genome Biol* 2020;**21**:14.
7. Yao D, Zhang L, Zheng M, *et al.* Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Sci Rep* 2018;**8**(1):11018.
8. Meng X, Hu D, Zhang P, *et al.* CircFunBase: a database for functional circular RNAs. *Database* 2019;**2019**:baz003.
9. 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.
10. Fan C, Lei X, Tie J, *et al.* CircR2Disease v2.0: An Updated Web Server for Experimentally Supported CircRNA-Disease Associations and Its Application. *Genomics Proteomics Bioinformatics* 2021. 10.1016/j.gpb.2021.10.002.
11. Wang C, Han C, Zhao Q, *et al.* Circular RNAs and complex diseases: from experimental results to computational models. *Brief Bioinform* 2021;**22**(6):bbab286.
12. Zhang W, Yu C, Wang X, *et al.* Predicting CircRNA-disease associations through linear neighborhood label propagation method. *Ieee Access* 2019;**7**:83474–83.
13. 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**(13):2911.
14. Lei X, Bian C. Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease association. *Sci Rep* 2020;**10**(1):1–9.
15. Stoimchev M, Ivanovska M, Štruc V. Learning to Combine Local and Global Image Information for Contactless Palmprint Recognition. *Sensors (Basel)* 2021;**22**(1):73.
16. Lu X, Shi D, Liu Y, *et al.* Speech depression recognition based on attentional residual network. *Frontiers in bioscience (Landmark edition)* 2021;**26**(12):1746–59.
17. Tsuruoka Y. Deep learning and natural language processing. *Brain Nerve* 2019;**71**(1):45–55.
18. 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**(13):4038–46.
19. Deepthi K, Jereesh AS. An ensemble approach for CircRNA-disease association prediction based on autoencoder and deep neural network. *Gene* 2020;**762**:145040.
20. Deng L, Zhang W, Shi Y. GMNN2CD: identification of circRNA-disease associations based on variational inference and graph Markov neural networks. *Bioinformatics* 2022;btac079.
21. Spinelli I, Scardapane S, Uncini A. Adaptive Propagation Graph Convolutional Network. *IEEE Trans Neural Netw Learn Syst* 2021;**32**(10):4755–60.
22. Zhu R, Ji C, Wang Y, *et al.* Heterogeneous graph convolutional networks and matrix completion for miRNA-disease association prediction. *Front Bioeng Biotechnol* 2020;**8**:901.
23. Xuan P, Pan S, Zhang T, *et al.* Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. *Cell* 2019;**8**(9):1012.
24. 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.
25. Adnan MN, Islam MZ. Forest PA: Constructing a decision forest by penalizing attributes used in previous trees. *Expert Systems with Applications* 2017;**89**:389–403.
26. Veličković P, Cucurull G, Casanova A, *et al.* Graph attention networks. In: *Proceedings of International Conference on Learning Representations*, 2018.
27. Xuan P, Cao Y, Zhang T, *et al.* Dual Convolutional Neural Networks With Attention Mechanisms Based Method for Predicting Disease-Related lncRNA Genes. *Front Genet* 2019;**10**:416.

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024

28. Tang X, Luo J, Shen C, *et al.* Multi-view multichannel attention graph convolutional network for miRNA-disease association prediction. *Brief Bioinform* 2021;**22**(6):bbab174.
29. Bian C, Lei XJ, Wu FX. GATCDA: Predicting circRNA-Disease Associations Based on Graph Attention Network. *Cancer* 2021;**13**(11):2595.
30. Ji C, Liu Z, Wang Y, *et al.* GATNNCDA: A Method Based on Graph Attention Network and Multi-Layer Neural Network for Predicting circRNA-Disease Associations. *Int J Mol Sci* 2021;**22**(16):8505.
31. Wang D, Wang J, Lu M, *et al.* Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 2010;**26**(13):1644–50.
32. Schriml LM, Arze C, Nadendla S, *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res* 2012;**40**(D1):D940–6.
33. Yu G, Wang LG, Yan GR, *et al.* DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 2015;**31**(4):608–9.
34. Chen X, Yan CC, Luo C, *et al.* Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci Rep* 2015;**5**(1):1–12.
35. Xuan P, Han K, Guo M, *et al.* Prediction of microRNAs associated with human diseases based on weighted k most similar neighbors. *PloS one* 2013;**8**(8):e70204.
36. 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.
37. Yao DX, Zhang L, Zheng MY, et al. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Sci Rep* 2018; **8**:11018.
38. Deng L, Zhang W, Shi Y, *et al.* Fusion of multiple heterogeneous networks for predicting circRNA-disease associations. *Sci Rep* 2019;**9**(1):1–10.
39. 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.
40. Li G, Yue Y, Liang C, *et al.* NCPCDA: network consistency projection for circRNA-disease association prediction. *RSC Adv* 2019;**9**(57):33222–8.
41. Yan C, Wang J, Wu FX. DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. *BMC bioinformatics* 2018;**19**(19):73–81.
42. Wei H, Liu B. iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Brief Bioinform* 2020;**21**(4):1356–67.
43. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* 2018;**PP**(8):2011–23.
44. Breiman L. Random forests. *Machine learning* 2001;**45**(1):5–32.
45. Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 2011;**12**:2825–30.
46. Cortes C, Vapnik V. Support-vector networks. *Machine learning* 1995;**20**(3):273–97.
47. Achawanantakun R, Chen J, Sun Y, *et al.* LncRNA-ID: Long non-coding RNA IDentification using balanced random forests. *Bioinformatics* 2015;**31**(24):3897–905.
48. Lertampaiporn S, Thammarongtham C, Nukoolkit C, *et al.* Identification of non-coding RNAs with a new composite feature in the Hybrid Random Forest Ensemble algorithm. *Nucleic Acids Res* 2014;**42**(11):e93–3.
49. Dai Q, Wang Z, Liu Z, *et al.* Predicting miRNA-disease associations using an ensemble learning framework with resampling method. *Brief Bioinform* 2022;**23**(1):bbab543.
50. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;**240**(4857):1285–93.
51. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clin Chem* 1993;**39**(4):561–77.
52. Bradley AP. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 1997;**30**(7):1145–59.
53. Lei X, Fang Z, Chen L, *et al.* PWCDA: path weighted method for predicting circRNA-disease associations. *Int J Mol Sci* 2018;**19**(11):3410.
54. Deng L, Zhang W, Shi Y, et al. Fusion of multiple heterogeneous networks for predicting circRNA-disease associations. *Sci Rep* 2019; **9**(1): 1–10.
55. Ma Z, Kuang Z, Deng L. CRPGCN: predicting circRNA-disease associations using graph convolutional network based on heterogeneous network. *BMC Bioinformatics* 2021;**22**(1):1–23.
56. Ruan H, Xiang Y, Ko J, *et al.* Comprehensive characterization of circular RNAs in 1000 human cancer cell lines. *Genome Med* 2019;**11**(1):55.

Downloaded from https://academic.oup.com/bib/article/23/5/bbac379/6692549 by guest on 30 December 2024