Journal of Biomedical Informatics 112 (2020) 103624

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Potential circRNA-disease association prediction using DeepWalk and network consistency projection

Guanghui Li [a,*], Jiawei Luo [b], Diancheng Wang [a], Cheng Liang [c], Qiu Xiao [d], Pingjian Ding [e], Hailin Chen [f]

[a] School of Information Engineering, East China Jiaotong University, Nanchang, China
[b] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
[c] College of Information Science and Engineering, Shandong Normal University, Jinan, China
[d] College of Information Science and Engineering, Hunan Normal University, Changsha, China
[e] School of Computer Science, University of South China, Hengyang, China
[f] School of Software, East China Jiaotong University, Nanchang, China

ARTICLE INFO

Keywords:
circRNA-disease association
DeepWalk
Network consistency projection
Similarity learning

ABSTRACT

A growing body of experimental studies have reported that circular RNAs (circRNAs) are of interest in pathogenicity mechanism research and are becoming new diagnostic biomarkers. As experimental techniques for identifying disease-circRNA interactions are costly and laborious, some computational predictors have been advanced on the basis of the integration of biological features about circRNAs and diseases. However, the existing circRNA-disease relationships are not well exploited. To solve this issue, a novel method named DeepWalk and network consistency projection for circRNA-disease association prediction (DWNCPCDA) is proposed. Specifically, our method first reveals features of nodes learned by the deep learning method DeepWalk based on known circRNA-disease associations to calculate circRNA-circRNA similarity and disease-disease similarity, and then these two similarity networks are further employed to feed to the network consistency projection method to predict unobserved circRNA-disease interactions. As a result, DWNCPCDA shows high-accuracy performances for disease-circRNA interaction prediction: an AUC of 0.9647 with leave-one-out cross validation and an average AUC of 0.9599 with five-fold cross validation. We further perform case studies to prioritize latent circRNAs related to complex human diseases. Overall, this proposed method is able to provide a promising solution for disease-circRNA interaction prediction, and is capable of enhancing existing similarity-based prediction methods.

1. Introduction

Circular RNA (circRNA) is a newly discovered endogenous non-coding RNA that is generated through back-splicing processes, thus presenting as a closed loop structure. For decades, circRNA was once thought to result from splicing artifacts [1]. Only recently have more and more circRNAs been found in eukaryotic cells [2–4]. Further, studies have shown that certain circRNAs display the role of miRNA sponges [5] and show differential abundance between tumor and normal cells [6,7]. For example, Vo et al. [6] detected more than 160,000 differentially expressed circRNAs through capture sequencing across over 2,000 tumor specimens and identified prostate cancer-related circRNAs in urine. Combined, deregulated circRNA expression is correlated with initiation and progression of disease, including cancer. Importantly, circRNAs are an ideal biomarkers because they are more stable compared with other linear non-coding RNA molecules [8–10]. Therefore, the identification of disease-circRNA relationships is facilitated to elucidate the mechanisms of circRNA functional roles as well as develop new drugs. In the past few years, many novel bioinformatics algorithms have been advanced to predict latent interactions between diseases and non-coding RNAs, such as disease-miRNA [11–13] and disease-lncRNA [14–16]. However, the research on disease-circRNA association prediction is just the beginning compared to linear non-coding RNA-disease.

Recently, several databases of disease-circRNA associations collected from experimental literatures have emerged [17–19], which provide

* Corresponding author.
E-mail address: ghli16@hnu.edu.cn (G. Li).

https://doi.org/10.1016/j.jbi.2020.103624
Received 6 July 2020; Received in revised form 8 November 2020; Accepted 12 November 2020
Available online 18 November 2020

valuable resources for the exploration of disease-circRNA correlations via computational methods. For example, as the first model of disease-circRNA association prediction, PWCDA first constructed a heterogeneous network based on circRNA and disease similarity data and known disease-circRNA associations, and then applied path weighted method to the constructed network to infer latent disease-associated circRNAs [20]. Another heterogeneous network-based model named KATZHCDA was proposed by Fan et al. [21]. They integrated known disease-circRNA correlations and different types of disease and circRNA similarity into a heterogeneous graph and implemented KATZ algorithm on the constructed graph to rank candidates. Meanwhile, Yan et al. [22] brought out a method called DWNN-RLS to simultaneously uncover underlying candidates for all diseases by employing regularized least squares built with Kronecker product kernel. By making full use of the inherent features of diseases and circRNAs, Xiao et al. [23] devised a computational framework in which association probability of each circRNA-disease pair were updated by low-rank approximation algorithm. Considering that there are many missing associations in the known datasets, Wei et al. [24] first derived neighbor interaction profiles based on disease and circRNA similarity and then employed matrix factorization algorithm to excavate unobserved disease-circRNA relationships. Later, label propagation algorithm was applied to the prediction of circRNA-disease interactions, which can efficiently transfer association label information on circRNA and/or disease similarity networks. For example, CD-LNLP [25] introduced linear neighborhood similarity and LLCDC [26] employed locality-constrained linear coding for constructing the similarity network when applying label propagation. By integrating known disease-protein associations and protein-circRNA interactions, Deng et al. [27] constructed an inferred disease-circRNA interaction network and then adopted KATZ method to measure the relevance for each disease-circRNA pair. Li et al. [28] introduced network consistency projection algorithm to effectively discover potential circRNAs for diseases. Lei et al. [29] presented a computational model termed ICFCDA using collaboration filtering algorithm based on known disease-circRNA associations and multiple types of disease and circRNA similarity data. Recently, Lei et al. [30] further proposed a model for the prediction task, which extracted the weighted features using random walk with restart to train k-nearest neighbors classifier.

Existing computational approaches make use of additional biological information the circRNAs and diseases have. However, the known circRNA-disease association network is not well exploited and features of vertices in the network are rarely taken into account. Network embedding methods can extract embeddings of nodes in a network, which can be adapted to compute topological similarities between nodes [31,32]. Consequently, we can utilize the similarities to make predictions by adapting the existing similarity-based approaches.

In this paper, a computational method named DeepWalk and network consistency projection for circRNA-disease association prediction (DWNCPCDA) is proposed. An important innovation of DWNCPCDA is that it adopts DeepWalk, a network embedding method, to learn embeddings of nodes in the known circRNA-disease association network, which can be incorporated with similarity-based methods to provide more flexibility for circRNA-disease prediction. Cross validation and case studies are carried out to comprehensively evaluate the prediction performance of DWNCPCDA, and the simulation results indicate that our proposed method could be served as a promising tool and be beneficial to the field of disease-circRNA interaction prediction.

2. Materials and methods

2.1. Datasets

There are several databases about circRNA-disease associations collected from experimental literatures, such as Circ2Disease [17], CircR2Disease [18] and circRNADisease [19]. Among these three databases, CircR2Disease is the latest and owns the most association datasets, including 661 circRNAs, 100 diseases and 739 interactions. Then we remove non-human and redundant correlations, and eventually obtain 650 associations involving 585 circRNAs and 88 diseases. Therefore, we select the CircR2Disease database as the benchmark dataset for cross validation.

Considering that the CircR2Disease database contains more known association pairs than another two datasets (i.e., Circ2Disease and circRNADisease) and there are still very lack of known associations currently, we implement DWNCPCDA on the circRNADisease dataset for case studies and confirm the predicted relationships in CircR2Disease. We compile the dataset from circRNADisease, which consists of 332 interactions between 40 diseases and 312 circRNAs.

The known association dataset between m diseases and n circRNAs can be modelled as a bipartite graph, in which diseases and circRNAs are considered as vertices and their relationships are considered as edges. This graph can be further encoded as a $m \times n$ adjacency matrix M. Concretely, if the i-th disease is connected to the j-th circRNA, $M(i, j) = 1$; otherwise, $M(i, j) = 0$.

2.2. DWNCPCDA method

As mentioned above, most existing computational approaches depend on additional biological information the circRNAs and diseases have, such as circRNA expression profiles and disease ontology. However, topology features of vertices learned from known disease-circRNA interaction network are rarely taken into account, which can preserve the structural property of the network efficiently. In biomedical research, DeepWalk [31], a network embedding method, has been successfully applied to predict protein function [33], detect drug target interactions [34] and excavate links between diseases and miRNAs [35]. Therefore, we first adopt DeepWalk to disclose feature vectors of vertices in the known circRNA-disease association network for calculating the circRNA-circRNA similarity and disease-disease similarity, and then these two similarity networks are used as input to the network consistency projection method to obtain the predicted disease-circRNA relationships. The flowchart of DWNCPCDA is shown in Fig. 1. DWNCPCDA can be found at https://github.com/ghli16/DWNCPCDA.

2.2.1. Similarity learning with DeepWalk

DeepWalk is a deep unsupervised learning model, which can generate low-dimensional embedding vectors of vertices in the bipartite circRNA-disease network. This approach utilizes truncated random walks and SkipGram [36] to learn the vector-based representation of vertices. For starters, truncated random walks that γ random walks of length t for each vertex v_i are performed. Then, the SkipGram model is implemented to learn the embedding vector of vertex v_i for each random walk. SkipGram maximally compute the co-occurrence likelihood among the nodes that come into view within a path of window w, and its objective function is as follows:

$$\min_{\Phi} -logPr(\{v_{i-w}, \cdots, v_{i+w}\} \backslash v_i | \Phi(v_i)) \tag{1}$$

Here Φ means the vector-based representation related to each node v_i, which is a $(m + n) \times d$ low-dimensional space matrix, where d indicates the dimension of node representation. SkipGram further approximates the above conditional probability using the assumption as follows:

$$Pr(\{v_{i-w}, \cdots, v_{i+w}\} \backslash v_i | \Phi(v_i)) = \prod_{j=i-w, j \neq i}^{i+w} Pr(v_j | \Phi(v_i)) \tag{2}$$

To reduce the computational time of calculating $\Pr(v_j|\Phi(v_i))$, Hierarchical Softmax [37] is used to factorize the conditional probability by assigning the vertices of a walking sequence to the leaves of a binary tree as equation (3) shows:

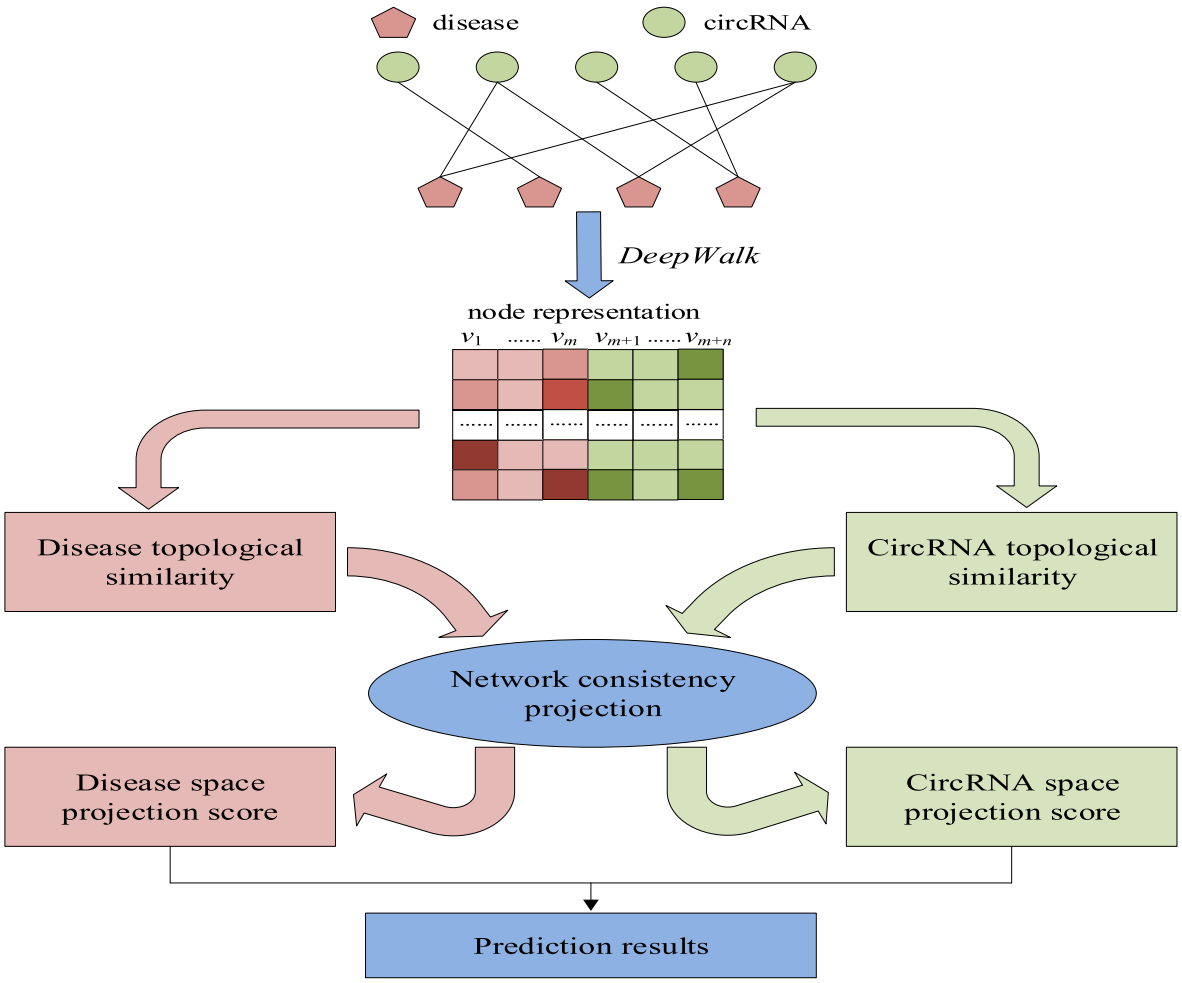

Fig. 1. The flowchart of DWNCPCDA method.

$$Pr\left(v_j\middle|\Phi(v_i)\right)=\prod_{l=1}^{\lceil log|V|\rceil}1\Big/\left(1+e^{-\Phi(v_i)\cdot\psi(b_l)}\right) \tag{3}$$

where (b_0, b_1,…,$b_{\log|v|}$) represents a sequence of tree nodes starting at the root and ending at v_j (i.e., b_0 = root and b&tnqh_0250C$_{\log|v|}$ &tnqh_02510= v_j). $\psi(b_l)$ denotes the embedding vectors associated with the parent of tree node b_l.

More details about DeepWalk are sketched in Ref. [31].

After applying DeepWalk to the circRNA-disease association network, we can obtain the embedding matrix Φ. Each row of Φ is a d-dimensional embedding vector corresponding to the latent topological representation of each disease node and circRNA node. Therefore, the cosine similarity between two embedding vectors can be calculated as the similarity of two disease (or circRNA) nodes, and the similarity between two nodes, v_i and v_j, is formulated as follows:

$$Sim\left(v_i,v_j\right)=\frac{\sum_{k=1}^{d}\Phi(v_i,k)\Phi\left(v_j,k\right)}{\sqrt{\sum_{k=1}^{d}\Phi(v_i,k)^2}\sqrt{\sum_{k=1}^{d}\Phi\left(v_j,k\right)^2}} \tag{4}$$

where $\Phi(v_i, k)$ and $\Phi(v_j, k)$ are the k-th components of embedding vector $\Phi(v_i)$ and $\Phi(v_j)$, respectively. According to equation (4), we can construct a circRNA topological similarity matrix Sim_c and a disease topological similarity matrix Sim_d, respectively.

2.2.2. Network consistency projection

Based on the above two similarity matrices, we capitalize on a simple and efficient network-based method, network consistency projection [38], to discover the potential disease-circRNA linkages in the view of circRNA space projection and disease space projection. From the view of circRNA space projection, the association score can be calculated by projecting circRNA similarity network on the known association network. The projection of vector form is formulated as follows:

$$CSP(i,j)=\frac{M(i,:)\times Sim_c(:,j)}{|M(i,:)|} \tag{5}$$

where $M(i, :)$ is a $1 \times n$ vector extracted from the i-th row of adjacency matrix M; $Sim_c(:, j)$ is a $n \times 1$ vector generated from j-th column of circRNA similarity matrix Sim_c; $|M(i, :)|$ represents the modulus of vector $M(i, :)$. As the adjacency matrix M is very sparse and most of the non-associated disease-circRNA pairs are actually missing interactions, we use a small positive value (10^{-30}) to replace 0 in M, which will not affect the prediction results and avoid the denominator being 0. According to equation (5), we can obtain the projection score of Sim_c on M, which is denoted as the circRNA space projection matrix $CSP \in R^{m\times n}$, where the element $CSP(i, j)$ is the projection value of $Sim_c(:, j)$ on $M(i, :)$.

From the view of disease space projection, another relevance score can be calculated by projecting disease similarity network on the known association network. The projection of vector form is formulated as follows:

$$DSP(i,j)=\frac{Sim_d(i,:)\times M(:,j)}{|M(:,j)|} \tag{6}$$

where $M(:, j)$ is a $m \times 1$ vector extracted from the j-th column of adjacency matrix M and $Sim_d(i, :)$ is a $1 \times m$ vector generated from i-th

row of disease similarity matrix Sim_d. According to equation (6), we can obtain the projection score of Sim_d on M, which is represented as the disease space projection matrix $DSP \in R^{m \times n}$, where the element $DSP(i, j)$ is the projection value of $Sim_d(i, :)$ on $M(:, j)$.

Finally, we take the integration and normalization of the above two projection score matrices CSP and DSP as the final output:

$$NCP(i,j) = \frac{CSP(i,j) + DSP(i,j)}{|Sim_d(i,:)| + |Sim_c(:,j)|} \quad (7)$$

where NCP denotes the final prediction score matrix for potential disease-circRNA associations.

Network consistency projection is a simple and efficient network-based method that has already shown excellent performance in link prediction research [28,38]. The model does not need negative instances and can simultaneously excavate candidate circRNAs for all diseases on a large scale. Moreover, by adopting network consistency projection, we could handle limited association data efficiently.

3. Results and discussion

3.1. Evaluation metrics

To evaluate how well the constructed model can predict, the most popular approach is K-fold cross validation. In the framework of K-fold cross validation, the entire dataset (i.e., all labeled circRNA-disease pairs), is divided into K folds evenly, of which one-fold of data is set as testing sample and the other folds are used as training set in turns. We can obtain a receiver operating characteristic (ROC) curve for each fold, and then an AUC (i.e., area under the ROC curve) value can be calculated. Finally, the AUC values for all K folds are averaged to obtain the final metric. Here K is set at 5 (i.e., 5-fold cross validation) and the number of known circRNA-disease associations (i.e., leave-one-out cross validation), respectively. Considering that the circRNA-circRNA and disease-disease similarities computed with DeepWalk depend on labeled circRNA-disease pairs, we need to recalculate them in each fold.

3.2. Parameter analysis

DeepWalk contains five parameters: γ, t, w, d, and α, where γ is the number of random walks started at each node, t is the walk length, w is the window size of SkipGram model, d represents the dimension of node representation, and α denotes the ratio of learning data available. To preserve local structural property, we fix the number of walks, the walk length, and the window size, which are set as the default values. The default values of γ, t, and w are 10, 40, and 5, respectively. The other two parameters are considered from the following combinations: {32, 64, 128} for d and {0.01, 0.05, 0.09} for α, which are the same as previous works [31,35]. Then we perform leave-one-out cross validation on two datasets (i.e., CircR2Disease and circRNADisease) to investigate the impact of these two parameters on the performance of DWNCPCDA. Fig. 2(a) illustrates the influence of varying the dimension and learning ratio to our model on CircR2Disease dataset, from which we can see that the performance has a small fluctuation for the change of α. On the other hand, the AUC scores improve as d increases and become relatively stable at $d = 128$. The results on circRNADisease dataset are shown in Fig. 2(b), which reveals that the AUC values almost have no change for the change of d and the predictive performance has a small fluctuation for the change of α. The prediction accuracy of DWNCPCDA on above two datasets is at a solid level, and the difference between the maximal and minimal AUC is 2.44% on two datasets, which shows the robustness of our method.

3.3. Comparison with other methods

In order to illustrate the predictive ability of DWNCPCDA, the performance of the proposed method is compared with several prevalent predictors, including NCPCDA [28], CD-LNLP [25], DWNN-RLS [22], KATZHCDA [21], and PWCDA [20]. All of the 6 predictors are trained based on known disease-circRNA interactions downloaded from CircR2Disease database by using five-fold cross validation and leave-one-out cross validation.

As shown in Fig. 3, DWNCPCDA performs best among all 6 predictors in leave-one-out cross validation, achieving an AUC of 0.9647, while the AUCs of NCPCDA, CD-LNLP, DWNN-RLS, KATZHCDA, and PWCDA are 0.9541, 0.9012, 0.9180, 0.8672, and 0.9000, respectively. The comparison of ROC curves among the methods considered here in terms of five-fold cross validation is shown in Fig. 4. As a result, the average AUC of DWNCPCDA is up to 0.9599, which is obviously better than that of the others (NCPCDA: 0.9201; CD-LNLP: 0.7996; DWNN-RLS: 0.6503; KATZHCDA: 0.8632; PWCDA: 0.8900;). Furthermore, we capitalize on the paired t-test to detect the differences between DWNCPCDA and the other five models, and the results in Table 1 show the superiority of the proposed method statistically.

Meanwhile, we conduct the experiment by using all known associated pairs in our benchmark dataset as the training samples to count the percentage of correctly identified known circRNAs for each disease in various top rankings. Fig. 5 displays that DWNCPCDA can detect more true interactions than other computational models in the top 5, top 10,

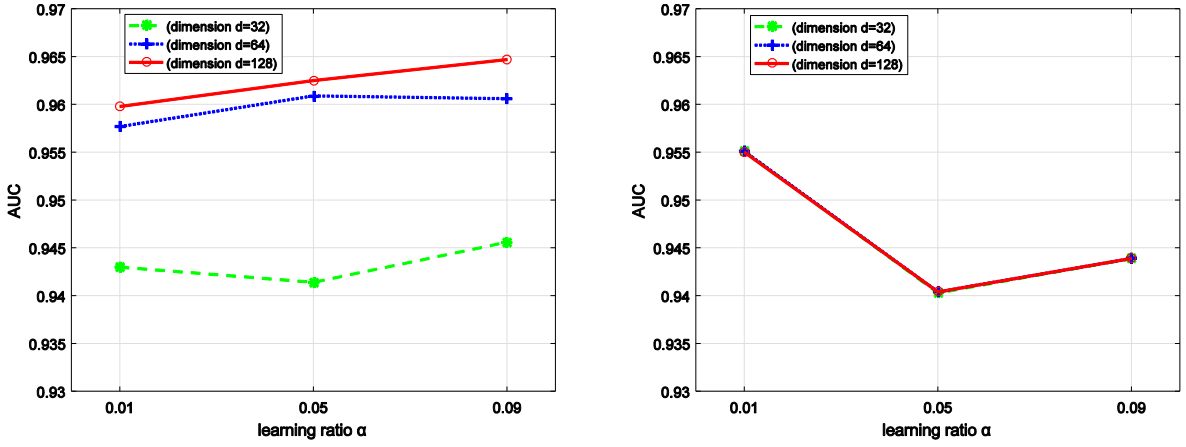

Fig. 2. The effect of dimension and learning ratio on the performance of DWNCPCDA. (a) The AUC scores on CircR2Disease dataset. (b) The AUC scores on circRNADisease dataset.

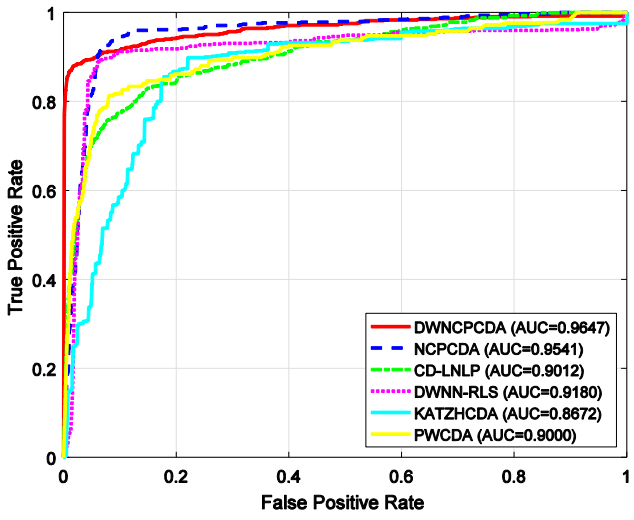

Fig. 3. The ROC curves of six predictors via leave-one-out cross validation.

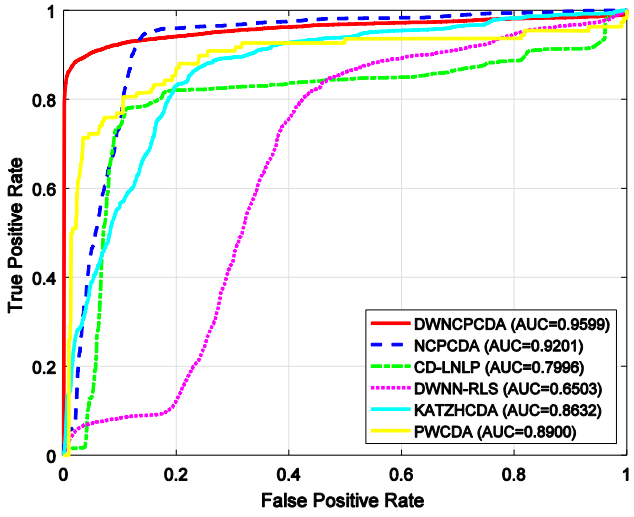

Fig. 4. The ROC curves of six predictors via five-fold cross validation.

Table 1
Results of paired *t*-test between DWNCPCDA and the other five methods.

DWNCPCDA vs.	NCPCDA	CD-LNLP	DWNN-RLS	KATZHCDA	PWCDA
p-value	3.626e-02	3.50e-03	3.790e-04	1.501e-04	1.971e-05

top 15, top 20, top 25, and top 30 predictions. For instance, among the 650 known associated pairs, there are 602 (or 92.62%) true positives successfully retrieved in the top 30 ranking for each disease by DWNCPCDA.

These six predictors are all network-based methods, which focus on constructing disease and/or circRNA similarity networks. NCPCDA, DWNN-RLS, KATZHCDA, and PWCDA fuse additional information sources to overcome the incompleteness and sparseness of current disease-circRNA association datasets. NCPCDA performs better because it is also based on network consistency projection, which has the advantage of simplicity and effectiveness on the disease and/or circRNA similarity networks. To further demonstrate the advantage of DeepWalk-based similarity measure, we implement another classic topology similarity measure in our circRNA-disease association network,

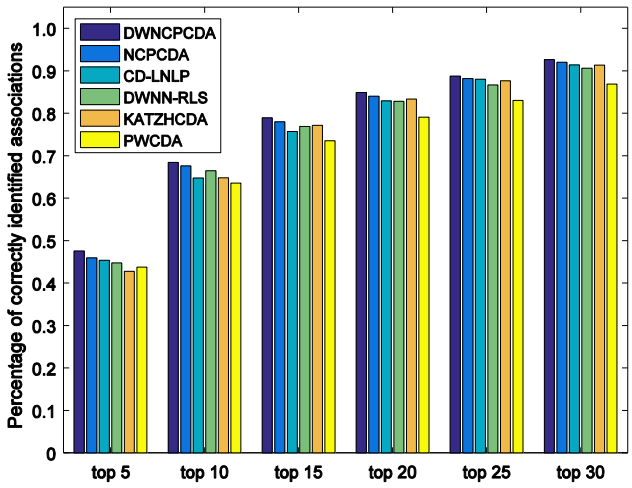

Fig. 5. Percentage of successfully identified true positives in various top rankings.

which is Gaussian interaction profile (GIP) kernel, and then assemble the similarity measure with network consistency projection association discovery model. We compare it with the DeepWalk-based results. As shown in Fig. 6, the model using DeepWalk consistently outperforms the model using GIP kernel in both validations. In summary, all the above results indicate that both a reasonable similarity metric for diseases and circRNAs and effective association prediction model for excavating latent disease-circRNA relationships is pivotal to improve prediction accuracy.

3.4. Case studies

As discussed in the Datasets section, we select the circRNADisease dataset as training samples for case studies, and then verify our findings in CircR2Disease. Furthermore, the predicted potential disease-circRNA relationships are confirmed by searching literatures on PubMed. Here, we analyze the top 30 candidate circRNAs for hepatocellular carcinoma and lung cancer yielded by DWNCPCDA, and list the rank of those novel verified associations in Table 2, from which we can see the following: (1) CircRNAs hsa_circ_0067934 and hsa_circ_0072088 are observed to be associated with above two investigated cancers, suggesting that these two circRNAs may function as novel diagnostic biomarkers for cancers. (2) There are eight circRNAs supported to be associated with hepatocellular carcinoma and seven circRNAs validated to be correlated with lung cancer among the top 30 prioritized candidates, respectively. For example, compared with the non-tumorous tissues, the expression levels of hsa_circ_0067934, hsa_circ_0002768, hsa_circ_0072088, hsa_circ_0005273, circMYLK, hsa_circ_0003221, hsa_circ_0003028, and hsa_circ_0007915 in hepatocellular carcinoma tissues are all shown to be markedly up-regulated [39–42], which suggest that these eigth circRNAs might be promising biomarkers and therapeutic targets in patients with this cancer. In lung adenocarcinoma tissues, Qiu et al. [43] reported that hsa_circ_0067934 is overexpressed and is crucial for lung adenocarcinoma tumorigenesis. (3) The hsa_circRNA_104912 ranked at top 29 in lung cancer is the fifteenth ranked circRNA by NCPCDA, which indirectly demonstrates that it is probably involved in lung neoplasms. The case studies further suggest that DWNCPCDA can provide valuable candidate circRNAs for drug development. The circRNA candidates for each disease ranked by DWNCPCDA are provided (shown in Supplementary Table S1). However, there are still many candidate circRNAs whose molecular mechanism has not been unveiled. They may also be validated by using biological experiments in the future.

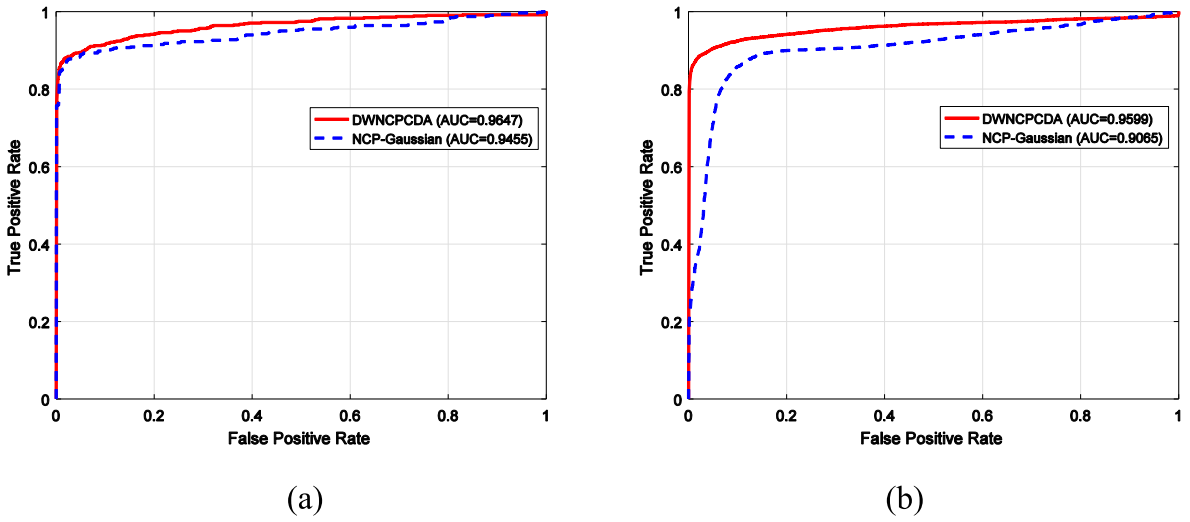

Fig. 6. Comparison of ROC curves using different similarity measures in two validations. (a) leave-one-out cross validation. (b) five-fold cross validation.

Table 2
The novel verified circRNAs for hepatocellular carcinoma and lung cancer and their rank predicted by DWNCPCDA.

Cancer	circRNAs	Rank	Evidences
Hepatocellular carcinoma	hsa_circ_0067934	1	CircR2Disease
	hsa_circ_0002768	6	PMID: 31,413,665
	hsa_circ_0072088	10	PMID: 28,727,484
	hsa_circ_0005273	15	PMID: 28,969,099
	circMYLK	16	PMID: 31,413,665
	hsa_circ_0003221	17	PMID: 28,969,099
	hsa_circ_0003028	20	PMID: 28,727,484
	hsa_circ_0007915	21	PMID: 28,727,484
Lung cancer	hsa_circ_0067934	2	PMID: 29,588,350
	hsa_circ_0007874	9	PMID: 30,975,029
	hsa_circ_0001727	11	PMID: 32,010,565
	hsa_circ_0007158	17	PMID: 28,969,099
	hsa_circ_0082582	24	PMID: 28,969,099
	hsa_circ_0072088	27	PMID: 29,698,681
	circFoxo3	28	CircR2Disease
	hsa_circRNA_104912	29	NCPCDA

4. Conclusions

There is growing evidence that circRNAs are becoming new diagnostic and prognostic biomarkers. In this work, a novel computational method is proposed to recommend circRNA molecules for queried diseases, in which based on the network topological similarity obtained from DeepWalk. The resulting similarity data is further adopted as the input to the network consistency projection method to uncover unknown circRNA-disease interactions. Different from existing computational predictors which primarily make use of additional biological information the circRNAs and diseases have, DWNCPCDA extract the embeddings of vertices including diseases and circRNAs from the disease-circRNA association network to construct the prediction model. The comparative experiments reveal that DWNCPCDA has better performance than those models using biological features and models using other topological similarity. Moreover, DWNCPCDA illustrates the good prediction ability in the case study.

In future research, more biomedical linked data of circRNAs or diseases, like circRNA- miRNA associations and disease-miRNA associations, could be integrated for similarity learning, which may further boost the performance. In addition, DWNCPCDA assembles the similarity-based solution with the network consistency projection algorithm for circRNA-disease prediction. In fact, network consistency projection can be replaced by certain machine learning techniques based vectorial data, which may get more accurate prediction overall.

CRediT authorship contribution statement

Guanghui Li: Conceptualization, Methodology, Software, Writing - original draft, Funding acquisition. **Jiawei Luo:** Conceptualization, Writing - review & editing, Funding acquisition. **Diancheng Wang:** Data curation, Software. **Cheng Liang:** Software, Validation. **Qiu Xiao:** Investigation, Visualization, Funding acquisition. **Pingjian Ding:** Formal analysis, Supervision, Funding acquisition. **Hailin Chen:** Visualization, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (Grant Nos. 61862025, 61873089, 62002116, 11862006, 61862026, and 62002154), Natural Science Foundation of Jiangxi Province of China (Grant Nos. 20181BAB211016, 2018ACB21032, and 20181BAB202008), Natural Science Foundation of Hunan Province of China (Grant Nos. 2020JJ5373 and 2018JJ2024), and Scientific Research Startup Foundation of University of South China (Grant No. 190XQD096).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2020.103624.

References

[1] C. Cocquerelle, B. Mascrez, D. Hétuin, B. Bailleul, Mis-splicing yields circular RNA molecules, FASEB J. 7 (1993) 155–160.
[2] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, et al., Circular RNAs are a large class of animal RNAs with regulatory potency, Nature 495 (2013) 333.
[3] L. Chen, C. Huang, X. Wang, G. Shan, Circular RNAs in eukaryotic cells, Curr. Genomics 16 (2015) 312–318.
[4] Q. Chu, X. Zhang, X. Zhu, C. Liu, L. Mao, C. Ye, Q.-H. Zhu, L. Fan, PlantcircBase: a database for plant circular RNAs, Molecular plant 10 (2017) 1126–1128.

[5] T.B. Hansen, T.I. Jensen, B.H. Clausen, J.B. Bramsen, B. Finsen, C.K. Damgaard, et al., Natural RNA circles function as efficient microRNA sponges, Nature 495 (2013) 384–388.
[6] J.N. Vo, M. Cieslik, Y. Zhang, S. Shukla, L. Xiao, Y. Zhang, et al., The landscape of circular RNA in cancer, Cell 176 (2019) 869–881.e813.
[7] S. Chen, V. Huang, X. Xu, J. Livingstone, F. Soares, J. Jeon, et al., Widespread and functional RNA circularization in localized prostate cancer, Cell 176 (2019) 831–843.e822.
[8] F.J. Slack, A.M. Chinnaiyan, The role of non-coding RNAs in oncology, Cell 179 (2019) 1033–1055.
[9] Q. Shang, Z. Yang, R. Jia, S. Ge, The novel roles of circRNAs in human cancer, Molecular Cancer 18 (2019) 6.
[10] M. Zhang, Y. Xin, Circular RNAs: a new frontier for cancer diagnosis and therapy, Journal of Hematology & Oncology 11 (2018) 21.
[11] X. Chen, C.-C. Zhu, J. Yin, Ensemble of decision tree reveals potential miRNA-disease associations, PLoS Comput. Biol. 15 (2019), e1007209.
[12] Q. Xiao, J. Luo, C. Liang, J. Cai, P. Ding, A graph regularized non-negative matrix factorization method for identifying microRNA-disease associations, Bioinformatics 34 (2018) 239–248.
[13] G. Li, J. Luo, Q. Xiao, C. Liang, P. Ding, Predicting microRNA-disease associations using label propagation based on linear neighborhood similarity, J. Biomed. Inform. 82 (2018) 169–177.
[14] X. Chen, C.C. Yan, X. Zhang, Z.-H. You, Long non-coding RNAs and complex diseases: from experimental results to computational models, Briefings Bioinf. 18 (2016) 558–576.
[15] C. Lu, M. Yang, F. Luo, F. Wu, M. Li, Y. Pan, et al., Prediction of lncRNA-disease associations based on inductive matrix completion, Bioinformatics 34 (2018) 3357–3364.
[16] G. Li, J. Luo, C. Liang, Q. Xiao, P. Ding, Y. Zhang, Prediction of lncRNA-disease associations based on network consistency projection, IEEE Access 7 (2019) 58849–58856.
[17] D. Yao, L. Zhang, M. Zheng, X. Sun, Y. Lu, P. Liu, Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease, Sci. Rep. 8 (2018) 11018.
[18] C. Fan, X. Lei, Z. Fang, Q. Jiang, F.-X. Wu, CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases, Database 2018 (2018) 1–8.
[19] Z. Zhao, K. Wang, F. Wu, W. Wang, K. Zhang, H. Hu, et al., circRNA disease: a manually curated database of experimentally supported circRNA-disease associations, Cell Death Dis. 9 (2018) 475.
[20] X. Lei, Z. Fang, L. Chen, F.-X. Wu, PWCDA: path weighted method for predicting circRNA-disease associations, Int. J. Mol. Sci. 19 (2018) 3410.
[21] C. Fan, X. Lei, F.-X. Wu, Prediction of circRNA-disease associations using KATZ model based on heterogeneous networks, Int. J. Biol. Sci. 14 (2018) 1950–1959.
[22] C. Yan, J. Wang, F.-X. Wu, DWNN-RLS: regularized least squares method for predicting circRNA-disease associations, BMC Bioinf. 19 (2018) 520.
[23] Q. Xiao, J. Luo, J. Dai, Computational prediction of human disease-associated circRNAs based on manifold regularization learning framework, IEEE J. Biomed. Health. Inf. 23 (2019) 2661–2669.
[24] H. Wei, B. Liu, iCircDA-MF: identification of circRNA-disease associations based on matrix factorization, Briefings Bioinf. 21 (2020) 1356–1367.
[25] W. Zhang, C. Yu, X. Wang, F. Liu, Predicting circRNA-disease associations through linear neighborhood label propagation method, IEEE Access 7 (2019) 83474–83483.
[26] E. Ge, Y. Yang, M. Gang, C. Fan, Q. Zhao, Predicting human disease-associated circRNAs based on locality-constrained linear coding, Genomics 112 (2020) 1335–1342.
[27] L. Deng, W. Zhang, Y. Shi, Y. Tang, Fusion of multiple heterogeneous networks for predicting circRNA-disease associations, Sci. Rep. 9 (2019) 9605.
[28] G. Li, Y. Yue, C. Liang, Q. Xiao, P. Ding, J. Luo, NCPCDA: network consistency projection for circRNA-disease association prediction, RSC Adv. 9 (2019) 33222–33228.
[29] X. Lei, Z. Fang, L. Guo, Predicting circRNA–disease associations based on improved collaboration filtering recommendation system with multiple data, Front. Genet. 10 (2019) 897.
[30] X. Lei, C. Bian, Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease association, Sci. Rep. 10 (2020) 1–9.
[31] B. Perozzi, R. Alrfou, S. Skiena, DeepWalk: online learning of social representations, in, in: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, 2014, pp. 701–710.
[32] J. Tang, M. Qu, M. Wang, M. Zhang, J. Yan, Q. Mei, LINE: large-scale information network embedding, in, in: Proceedings of the 24th international conference on World Wide Web, 2015, pp. 1067–1077.
[33] M. Kulmanov, M.A. Khan, R. Hoehndorf, DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier, Bioinformatics 34 (2018) 660–668.
[34] N. Zong, H. Kim, V. Ngo, O. Harismendy, Deep mining heterogeneous networks of biomedical linked data to predict novel drug-target associations, Bioinformatics 33 (2017) 2337–2344.
[35] G. Li, J. Luo, Q. Xiao, C. Liang, P. Ding, B. Cao, Predicting microRNA-disease associations using network topological similarity based on DeepWalk, IEEE Access 5 (2017) 24032–24039.
[36] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv:1301.3781, 2013.
[37] A. Mnih, G.E. Hinton, A scalable hierarchical distributed language model, in: neural information processing systems, 2008, pp. 1081-1088.
[38] C. Gu, B. Liao, X. Li, K. Li, Network consistency projection for human miRNA-disease associations inference, Sci. Rep. 6 (2016) 36054.
[39] Q. Zhu, G. Lu, Z. Luo, F. Gui, J. Wu, D. Zhang, et al., CircRNA circ_0067934 promotes tumor growth and metastasis in hepatocellular carcinoma through regulation of miR-1324/FZD5/Wnt/β-catenin axis, Biochem. Biophys. Res. Commun. 497 (2018) 626–632.
[40] Z. Li, Y. Hu, Q. Zeng, H. Wang, J. Yan, H. Li, et al., Circular RNA MYLK promotes hepatocellular carcinoma progression by increasing Rab23 expression by sponging miR-362-3p, Cancer Cell International 19 (2019) 211.
[41] S. Ren, Z. Xin, Y. Xu, J. Xu, G. Wang, Construction and analysis of circular RNA molecular regulatory networks in liver cancer, Cell Cycle 16 (2017) 2204–2211.
[42] C. Han, N.A. Seebacher, F.J. Hornicek, Q. Kan, Z. Duan, Regulation of microRNAs function by circular RNAs in human cancer, Oncotarget 8 (2017) 64622–64637.
[43] M. Qiu, W. Xia, R. Chen, S. Wang, Y. Xu, Z. Ma, et al., The circular RNA circPRKCI promotes tumor growth in lung adenocarcinoma, Cancer Res. 78 (2018) 2839–2851.