

OPEN

Integrating random walk with restart and k-Nearest Neighbor to identify novel circRNA-disease association

Xiujuan Lei* & Chen Bian

CircRNA is a special type of non-coding RNA, which is closely related to the occurrence and development of many complex human diseases. However, it is time-consuming and expensive to determine the circRNA-disease associations through experimental methods. Therefore, based on the existing databases, we propose a method named RWRKNN, which integrates the random walk with restart (RWR) and k-nearest neighbors (KNN) to predict the associations between circRNAs and diseases. Specifically, we apply RWR algorithm on weighting features with global network topology information, and employ KNN to classify based on features. Finally, the prediction scores of each circRNA-disease pair are obtained. As demonstrated by leave-one-out, 5-fold cross-validation and 10-fold cross-validation, RWRKNN achieves AUC values of 0.9297, 0.9333 and 0.9261, respectively. And case studies show that the circRNA-disease associations predicted by RWRKNN can be successfully demonstrated. In conclusion, RWRKNN is a useful method for predicting circRNA-disease associations.

CircRNA, as a star molecule in the recent years, is a kind of non-coding endogenous RNA with single-stranded, closed and circular structure^{1,2}. Unlike the linear RNA, circRNA is the result of “back-splice” or derived from linear RNA. Hence, they lack 5'-3' ends representing the RNA transcription's start and stop³⁻⁶. The first circRNA was discovered by electron microscopy in RNA viruses⁷ and afterwards in eukaryotic cells⁸. Unfortunately, researchers regarded circRNA initially as a by-product of abnormal splicing without regulatory potential. Thus, circRNA did not attract much scientific attention⁹.

With the increasing researches on circRNAs, lots of circRNAs have been found in viruses, animals and plants^{6,10-12}. So far, circRNA has been confirmed to regulate multiple major biological processes, like cell invasion, proliferation as well as apoptosis^{13,14}. And circRNA is an important part in process of transcription¹⁵, mRNA splicing¹⁶, RNA translation and decay¹⁷. Thus, the regulatory mechanism of circRNA is closely related to the occurrence of disease, which was identified by advanced biotechnology. For instance, the expression level of hsa_circ_0001982 in breast cancer tissues is significantly high¹⁸. In addition, there are some circRNAs (Hsa_circ_0014717¹⁹, CircMTO1²⁰, Circ-PRKCI²¹) that act as miRNA's sponge to regulate tumorigenesis. Therefore, it can provide new ideas for the treatment of diseases with acquisition and utilization of information related to circRNAs and diseases.

In recent years, some circRNA-disease related databases have also been proposed to further investigate the associations between circRNAs and diseases, involving CircR2Disease²², circRNADisease²³ and Circ2Disease²⁴. The effective calculation methods based on these databases will effectively reduce the time consumption caused by the methods in biological experiments. Thus, it is urgent to use computational methods for exploring disease-related circRNA. Fan *et al.*²⁵ raised a similarity-based method with KATZ measure called KATZHCDA on a heterogeneous network. Yan *et al.*²⁶ advanced a kernel-based method with regularized least squares. Lei *et al.*²⁷ proposed a path-weighted method (PWCDA) integrating disease functional similarities and circRNA semantic similarities. Xiao *et al.*²⁸ put forward a model (MRLDC) using a weighted manifold regularized-based algorithm. Wei *et al.*²⁹ proposed a factorization Machine (FM) based method called iCircDA-MF using matrix factorization. Zhao *et al.*³⁰ developed a method (IBNPKATZ) integrating the KATZ measure and bipartite network projection. Zhang *et al.*³¹ proposed a label propagation method (CD-LNLP) based on linear neighborhood. However, these methods above rely on the information of circRNA-disease, circRNAs or diseases, and the number of datasets

School of Computer Science, Shaanxi Normal University, Xi'an, 710119, China. *email: xjlei@snnu.edu.cn

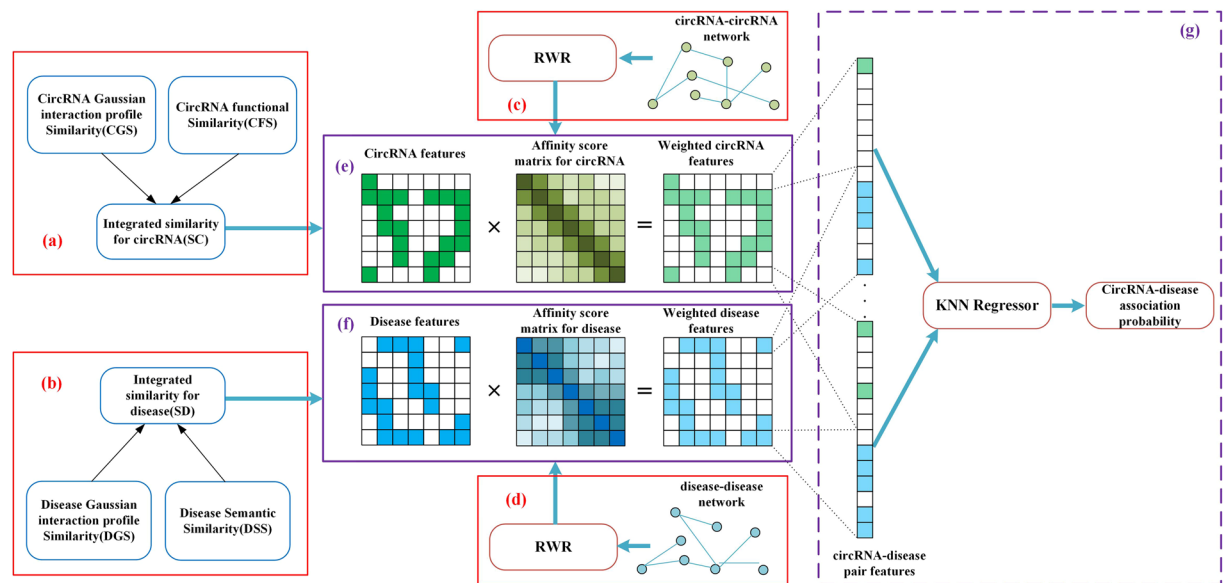


Figure 1. The flowchart of the computational method RWRKNN.

is relatively limited. Therefore, it is not very suitable to discover the relationship of new diseases and novel circRNAs. To solve the problem further, Deng *et al.*³² proposed a KATZ-based method (KATZCPDA) integrating the information of circRNAs, diseases and proteins. Due to bioinformatics analysis of protein information, KATZCPDA could predict potential association that cannot be inferred when only using information of circRNAs and diseases.

Inspired by Lee *et al.*³³, a model weighting the features of circRNAs and diseases in the global network topology is put forward. In this work, all features of circRNA-disease pairs are weighted using the random walk with restart (RWR) algorithm. Firstly, we construct circRNA-disease associations, and calculate circRNA functional similarity, Gaussian interaction profile (GIP) kernel similarity of circRNAs, disease semantic similarity and GIP kernel similarity of diseases. Secondly, based on these similarities, we further construct two matrixes, *i.e.*, circRNA-circRNA matrix, disease-disease matrix. Next, the RWR is performed on all nodes in circRNA-circRNA matrix and disease-disease matrix respectively. With affinity scores of all circRNA and disease nodes from RWR, features of circRNAs (diseases) consisting of integrated circRNA (disease) similarity are weighted. In the end, negative circRNA-disease pairs are selected randomly and a k-Nearest Neighbor (KNN) model get trained with weighted features (See Fig. 1).

Results

Performance evaluation. Leave-one-out cross validation (LOOCV), 5-fold cross-validation (5CV) and 10-fold cross-validation (10CV) are utilized to evaluate the prediction performance of our model. For LOOCV, each positive sample is left out in turn as a testing sample, and the other positive samples are used to train the model with the negative samples. Different from the LOOCV, 5CV and 10CV randomly divide the positive samples into 5 equal parts and 10 equal parts respectively, and take out one part of them as testing samples while the rest of samples are regarded as training samples in turn. Next, the predicted scores are sorted in descending order. Further, we draw the receiver operating characteristics (ROC) curve via plotting the true positive rate (TPR) versus the false positive rate (FPR) at different score thresholds. TPR (FPR) refers to the percentage of positive (negative) cases that are correctly identified. Generally, the area under the ROC curve (AUC) is calculated and employed to evaluate the prediction performance. Specifically, the closer the AUC value is to 1, the better the prediction performance. As a result, in LOOCV, RWRKNN achieves an AUC of 0.9297. And concerning 5CV and 10CV, RWRKNN yields the average AUCs of 0.9333 and 0.9261 respectively. The results are shown in the Fig. 2.

Adjustment of parameters. RWRKNN model involves four parameters: DA's threshold value α , CA's threshold value β , neighbors' number k and distance metric p . The value of α and β might affect the weighted feature matrixes of circRNAs and diseases. The value of k and p probably influence KNN's classification performance. Let α and β both range between 0.5 and 0.9. Let k be an integer value between 1 and 5 and $p \in \{1, 2, 3\}$. As a result, among these four parameters, RWRKNN ($\alpha = 0.6$, $\beta = 0.8$, $k = 5$ and $p = 1$) gains the highest AUCs of 0.9333 in 5CV as shown in the Supplement. Specifically, $p = 1$ means the Manhattan distance metric.

Compared with other methods. To analyze the performance of RWRKNN model in predicting circRNA-disease associations, RWRKNN ($\alpha = 0.6$, $\beta = 0.8$, $k = 5$ and $p = 1$) is compared with four methods. Firstly, to show the importance of weighting features, we compared RWRKNN with a model with unweighted features called raw_KNN ($k = 5$, $p = 1$). And in order to highlight the classification performance of KNN, Support Vector Machine (SVM) is compared with our model. In the end, we compare RWRKNN with KATZHCDA²⁵ and DWNN-RLS²⁶ previously mentioned. The ROC curves of each method using LOOCV are shown in Fig. 3. In

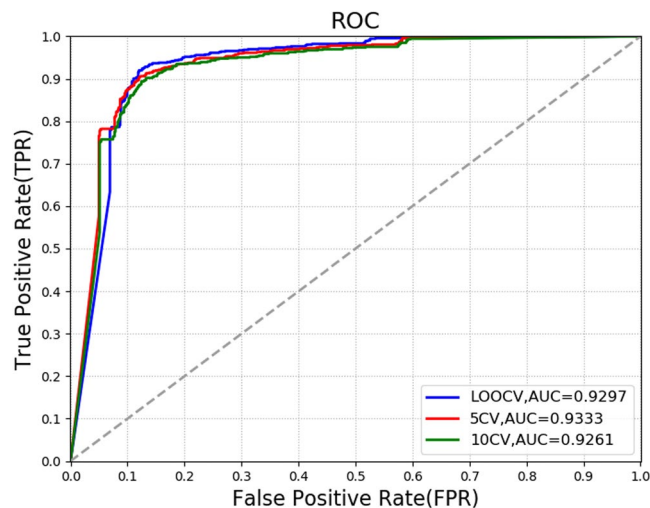


Figure 2. The ROC curves and AUCs of RWRKNN in LOOCV, 5CV and 10CV.

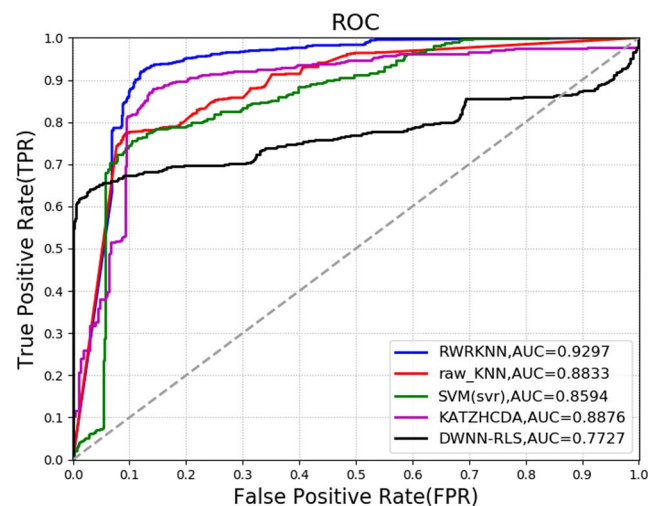


Figure 3. The ROC curves and AUCs of five methods using LOOCV.

addition, we also compared RWRKNN with other four methods in other evaluation criteria (see Fig. 4) including accuracy (ACC), F1-Score, Matthews Correlation Coefficient (MCC). And Precision-Recall (PR) curves and area under the PR curves (AUPRs) are also adopted to reflect the performance of these five methods (see Fig. 5). We can see that RWRKNN gets the satisfactory and optimal performance.

Case study. To further evaluate the prediction performance of RWRKNN ($\alpha = 0.6$, $\beta = 0.8$, $k = 5$ and $p = 1$), we also carry out case studies on three common diseases, *i.e.*, breast cancer, bladder cancer and colorectal cancer. Breast cancer is one of the most common cancer affecting women, and its incidence and mortality rates are expected to increase significantly the next years³⁴. Bladder cancer is a kind of cancer with high incidence, morbidity and mortality³⁵. Colorectal cancer is also one of the most common cancers worldwide³⁶. However, the complex biology of the three types of diseases remains uncertain and unexplored. Therefore, it is necessary to explore the biological characteristics of these diseases by using computational methods. In this work, all known associations between the investigated disease and circRNAs are assumed to be unknown. Through the calculation of the model, the circRNAs with the top 10 scores are selected among all the predicted associations between the investigated disease and circRNAs. Through searching related literatures or databases, some circRNAs are confirmed to be related to the investigated disease. The results of the case studies of these diseases (breast cancer, bladder cancer and the colorectal cancer) are shown in Tables 1, 2 and 3, respectively.

Conclusion

At circRNA level, identifying unknown associations of circRNA-disease get crucial for the study of biomarkers for disease diagnosis. In this study, a computational method (RWRKNN) is proposed, which integrates RWR and KNN regression. The existing circRNA-disease association from CircR2Disease is used to assign labels to circRNA-disease pairs. In view of constructing feature of circRNA-disease pairs and circRNA-circRNA

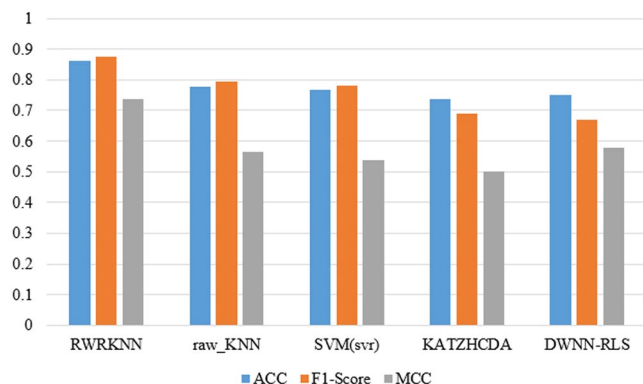


Figure 4. Comparison of five methods in ACC, F1-Score, MCC (LOOCV).

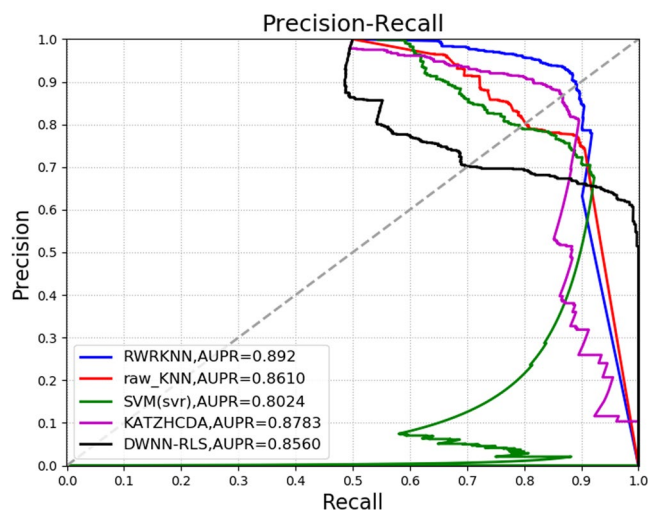


Figure 5. Comparison of five methods in PR curves and AUPRs (LOOCV).

Rank	CircRNA name	Evidences	Rank	CircRNA name	Evidences
1	hsa_circ_005239	PMID:29037220	6	hsa_circ_0108942	PMID:29045858
2	hsa_circ_0007534	PMID:30139516	7	hsa_circ_0001946	PMID:28049499
3	hsa_circ_0001982	PMID:28933584	8	hsa_circ_0006528	PMID:30520151
4	circRNA-000911	PMID:29431182	9	hsa_circ_0003575	unconfirmed
5	hsa_circ_0001785	PMID:29045858	10	circDENND4C	PMID:31488193

Table 1. Candidate circRNAs of breast cancer.

Rank	CircRNA name	Evidences	Rank	CircRNA name	Evidences
1	hsa_circ_0003221	PMID:29125888	6	hsa_circ_0007158	circRNADisease
2	hsa_circ_0091017	PMID:29151929	7	hsa_circ_0041103	circRNADisease
3	hsa_circ_0000284	circRNADisease	8	hsa_circ_0008732	unconfirmed
4	hsa_circ_0002768	circRNADisease	9	hsa_circ_0005941	unconfirmed
5	hsa_circ_0058058	unconfirmed	10	hsa_circ_0002024	PMID:30972190

Table 2. Candidate circRNAs of bladder cancer.

associations *CA* and disease-disease associations *DA*, we take use of circRNA-disease associations, GIP kernel similarities of circRNAs and diseases, circRNA functional similarity and disease semantic similarity. For every circRNA (disease), we complement RWR on the constructed *CA* (*DA*) matrix to obtain affinity scores, which are

Rank	CircRNA name	Evidences	Rank	CircRNA name	Evidences
1	hsa_circ_0007534	PMID:29364478	6	hsa_circ_0020397	PMID:28707774
2	hsa_circ_0001649	PMID:29421663	7	circ-BANP	PMID:28103507
3	hsa_circ_0014717	PMID:29571246	8	hsa_circ_0000069	PMID:28003761
4	hsa_circ_0000567	PMID:29333615	9	hsa_circRNA_104700	PMID:28349836
5	circRNA0003906	PMID:29123417	10	hsa_circRNA_103809	PMID:30249393

Table 3. Candidate circRNAs of colorectal cancer.

employed to weight the features of circRNAs (diseases). After obtaining the global feature vectors of circRNAs (diseases), KNN regression model could output the possibility of inquired circRNA-disease association pairs. In addition, both multiple performance evaluation criteria and case studies on breast cancer, bladder cancer and colorectal cancer have illustrated the reliable prediction ability of RWRKNN. However, RWRKNN also has limitations. It relies on prior information about circRNAs and diseases. Therefore, it is slightly inadequate in uncovering the relationship between new diseases and new circRNAs.

Materials and Methods

Human circRNA-disease associations. To acquire circRNA-disease associations verified by biological experiments, we download the circRNA-disease associations from circR2Disease database (<http://bioinfo.snnu.edu.cn/CircR2Disease/>)²². CircR2Disease provides association information between circRNAs and diseases supported by experiments, including 725 circRNA-disease associations between 661 circRNAs and 100 diseases. In this study, we extract all circRNA and disease associations in the database and then construct a matrix A to reflect the adjacency associations of circRNA-disease. If a disease i has been confirmed to have an association with a circRNA j , $A(i, j) = 1$, otherwise $A(i, j) = 0$. The dimension of A is $N_c \times N_d$, where N_c and N_d represent the number of the known circRNAs and the known diseases, respectively.

Disease similarity. The semantic similarity between diseases is calculated based on DAG (directed acyclic graph) topology. To be specific, the DAG of a disease d can be defined as $DAG(d) = (d, T(d), E(d))$, where $T(d)$ is an ancestor set of disease d and $E(d)$ includes the corresponding edges. According to Eqs. (1) and (2), the semantic value $DSV(d)$ of disease d can be obtained³⁷.

$$D_d(t) = \begin{cases} 1, & \text{if } t = d \\ \max \{\Delta_* D_d(t') \mid t' \in \text{children of } t\}, & \text{if } t \neq d \end{cases} \quad (1)$$

$$DSV(d) = \sum_{t \in T(d)} D_d(t) \quad (2)$$

where the disease $t \in T(d)$, Δ ($\Delta = 0.5$) is semantic contribution decay factor, and $D_d(t)$ denotes the contribution of ancestor node t to d . Next, the semantic similarity between d_i and d_j can be calculated as follows:

$$DSS(d_i, d_j) = \frac{\sum_{t \in T(d_i) \cap T(d_j)} (D_{d_i}(t) + D_{d_j}(t))}{DSV(d_i) + DSV(d_j)} \quad (3)$$

In the end, the semantic similarity matrix of diseases DSS is constructed.

GIP kernel similarity is calculated based on the topological structure of the association network of biological information nodes³⁸. An assumption supporting this approach is that more similar diseases tend to be associated with the similar circRNAs³⁸. Disease GIP kernel similarity is calculated with the known circRNA-disease associations, which is obtained according to the Eq. (4).

$$DGS(i, j) = \exp\left(-\sigma_d \|IP_{d(i)} - IP_{d(j)}\|^2\right) \quad (4)$$

where $IP_{d(i)}$ represents the interaction profile of disease i as a binary vector reflecting whether disease i is associated with each circRNA or not. $DGS(i, j)$ is the GIP kernel similarity between disease i and disease j . σ_d is influential in tuning the kernel bandwidth calculated by the Eq. (5).

$$\sigma_d = \sigma_d^* / \left(\frac{1}{N_d} \sum_{i=1}^{N_d} \|IP_{d(i)}\|^2 \right) \quad (5)$$

where N_d is the number of all diseases, and σ_d^* is set to 1 as the initial value following the previous study³⁸.

In order to make full use of the disease semantic similarity and the disease GIP kernel similarity, we construct a new disease similarity matrix SD (See Fig. 1b) by integrating DSS and DGS based on the Eq. (6).

$$SD(d(i), d(j)) = \begin{cases} \frac{DSS(d(i), d(j)) + DGS(d(i), d(j))}{2} & \text{if } DSS(d(i), d(j)) \neq 0 \\ DGS(d(i), d(j)) & \text{otherwise} \end{cases} \quad (6)$$

CircRNA similarity. Similar to the calculation method of disease GIP kernel similarity, we use the Eq. (7) to calculate circRNA GIP kernel similarity.

$$CGS(i, j) = \exp\left(-\sigma_c \|IP_{c(i)} - IP_{c(j)}\|^2\right) \quad (7)$$

where $IP_{c(i)}$ represents the interaction profile of circRNA i as a binary vector reflecting whether circRNA i is associated with each disease or not. $CGS(i, j)$ is the GIP kernel similarity between circRNA i and circRNA j . σ_c is influential in tuning the kernel bandwidth calculated by the Eq. (8).

$$\sigma_c = \sigma_c^* / \left(\frac{1}{N_c} \sum_{i=1}^{N_c} \|IP_{c(i)}\|^2 \right) \quad (8)$$

where N_c is the number of all circRNAs, and σ_c^* is set to 1 as the initial value following the previous study³⁸.

We adopt a similar method to Wang's method³⁷ for calculating circRNA functional similarity to improve the accuracy of the calculation model. To be specific, the functional similarity score between a circRNA U and a circRNA V is obtained by calculating the semantic similarity between the two groups of circRNA-related diseases. First, let dx be any given disease, and Dy be a group of diseases defined as $Dy = \{dy_1, dy_2, dy_3, \dots, dy_r\}$. Then, the semantic similarity between dx and Dy can be calculated as follows:

$$SS(dx, Dy) = \max_{1 \leq i \leq r} (DSS(dx, dy_i)) \quad (9)$$

Second, the functional similarity between circRNA U and circRNA V can be calculated as follows:

$$CFS(U, V) = \frac{\sum_{1 \leq i \leq |Du|} SS(Du_i, Dv) + \sum_{1 \leq j \leq |Dv|} SS(Dv_j, Du)}{|Du| + |Dv|} \quad (10)$$

where Du is a group of circRNA U -related diseases and Dv is another group of circRNA V -related diseases. $Du_i \in Du$ and $Dv_j \in Dv$. In the end, circRNA functional similarity matrix is constructed, which is symmetric and has all 1s on its diagonal. $CFS(i, j)$ represents the functional similarity between circRNA i and circRNA j .

Similar to the method of disease similarity integration, circRNA functional similarity and circRNA GIP kernel similarity are integrated to constitute a new circRNA similarity matrix SC (See Fig. 1a) based on the Eq. (11).

$$SC(c(i), c(j)) = \begin{cases} \frac{CFS(c(i), c(j)) + CGS(c(i), c(j))}{2} & \text{if } CFS(c(i), c(j)) \neq 0 \\ CGS(c(i), c(j)) & \text{otherwise} \end{cases} \quad (11)$$

Human disease-disease associations. In order to get the disease association adjacency matrix DA , a threshold value α is set for the integrated disease similarity SD as shown in Eq. (12). If the similarity value is greater or equal to α , the corresponding position in DA has a value of 1, otherwise 0.

$$DA(i, j) = \begin{cases} 1 & \text{if } SD(i, j) \geq \alpha \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

Human circRNA- circRNA associations. In order to get the circRNA association adjacency matrix CA , a threshold value β is set for the integrated circRNA similarity SC as shown in Eq. (13). If the similarity value is greater or equal to β , the corresponding position in CA has a value of 1, otherwise 0.

$$CA(i, j) = \begin{cases} 1 & \text{if } SC(i, j) \geq \beta \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

RWRKNN. After having constructed four matrixes, *i.e.*, the integrated disease similarity matrix, the integrated circRNA similarity matrix, disease-disease association matrix and circRNA-circRNA association matrix, RWRKNN will do the following three steps, *i.e.*, RWR for every circRNA and disease (Fig. 1c,d), feature weighting (Fig. 1e,f) and training KNN model (Fig. 1g).

Considering the input requirements of the KNN regression model, we transform the features of circRNA-disease pairs into vectors. Firstly, for diseases, we take each row of the integrated disease similarity SD as the feature vector of diseases with 100 dimensions. Similarly, with respect to circRNAs, we take each row of the integrated circRNA similarity SC as the feature vector of circRNAs with 661 dimensions.

To make predictions of circRNA-disease associations from a global network perspective, we could obtain affinity scores between a circRNA (disease) node and all circRNA (disease) nodes using the RWR algorithm on the CA (DA) network. RWR estimates affinity level (affinity score) between two nodes by repeatedly exploring the overall structure of a network. Starting at a seed node, the random walker diffuses its resources by (1) moving to a neighbor node with probability $1-c$ and (2) restarting from the seed node with restarting probability c . This process is iterated repeatedly until all nodes are traversed. At this time, the probability vector obtained contains

the affinity scores of all nodes and the seed node. The affinity scores of all nodes during each step are represented by the Eq. (14).

$$p = (1 - c)Wp + cq \quad (14)$$

where q is the starting vector whose seed node is set to 1 while the others are set to 0, and W is the normalized adjacency matrix, and p would finally reach a steady state after multiple iterations, and c is set to 0.7 according to Park *et al.*³⁹'s work. Consequently, by multiplying the adjacency matrix, it diffuses its resources throughout the network. By the iteration of the p_i value, *i.e.* the result of RWR for the seed node i , the affinity score matrix F could be obtained, whose element F_{ij} refers to how closely node j is connected to seed node i . Finally, the $(N_c \times N_c)$ circRNA affinity score matrix F^c and the $(N_d \times N_d)$ disease affinity score matrix F^d are constructed, where N_c is the number of circRNAs and N_d is the number of diseases.

Next, we utilize the affinity scores to weight the circRNA and disease features. As regards the circRNA features, they are weighted using the Eq. (15).

$$WSC(c(i)) = F_i^c{}^T \times SC(c(i)) \quad (15)$$

where F_i^c means the affinity score of circRNA $c(i)$, which is a row vector. $F_i^c{}^T$ is the transpose of F_i^c . And $SC(c(i))$ denotes the integrated similarity of circRNA $c(i)$, which is also a row vector. WSC is the weighted feature matrix of circRNAs and $WSC(c(i))$ represents the weighted feature of circRNA $c(i)$.

In the case of the disease features, Eq. (16) is used.

$$WSD(d(i)) = F_i^d{}^T \times SD(d(i)) \quad (16)$$

where F_i^d means the affinity score of disease $d(j)$, which is a row vector. $F_i^d{}^T$ is the transpose of F_i^d . And $SC(d(j))$ denotes the integrated similarity of disease $d(j)$, which is also a row vector. $WSD(d(j))$ represents the weighted feature of disease $d(j)$ and WSD is the weighted feature matrix of diseases.

In the whole, the features of circRNAs and diseases are weighted by means of adding each feature of all nodes to a certain seed node via affinity scores from the RWR. The weighting can be conducted by multiplying feature matrix by affinity score matrix as depicted in Fig. 1e,f.

With the weighted features of the circRNAs and diseases, we link each feature vector of diseases and circRNAs together to compose a 761-dimensional feature vector for each circRNA-disease pair as the input of KNN regressor model. To train the KNN regressor model, we prepare positive samples and negative samples. The known circRNA-disease association pairs are used as positive samples. To get negative samples, the following steps are taken: (1) A circRNA i is selected at first, and then (2) calculate the number nd_i of diseases associated with the circRNA i . (3) Next, select nd_i diseases unassociated to the circRNA i . (4) Until all circRNAs are traversed, we end up with the same number of negative samples as positive samples. In RWRKNN model, the KNN regression could find k neighbors closest to a certain circRNA-disease pair based on the Minkowski distance metric (defined as Eq. (17)), which is a set of distance definitions.

$$d(x, y) = \left(\sum_{i=1}^n |a_i - b_i|^p \right)^{\frac{1}{p}} \quad (17)$$

Here, different values of p represent different distance metrics for calculating the distance between vectors a and b with n -dimension, and we set $p = 1$, which represents the Manhattan distance is used as a metric between vectors. In addition, considering the closer neighbors should have more weight, we take the inverse of the distance as the weight.

Data availability

Data can be obtained by sending an email to the corresponding author upon reasonable request.

Received: 7 October 2019; Accepted: 23 January 2020;

Published online: 06 February 2020

References

1. Fan, X. *et al.* Circular RNAs in Cardiovascular Disease: An Overview. *BioMed research international* **2017**, 5135781, <https://doi.org/10.1155/2017/5135781> (2017).
2. Greene, J. *et al.* Circular RNAs: Biogenesis, Function and Role in Human. *Diseases. Frontiers in molecular biosciences* **4**, 38, <https://doi.org/10.3389/fmolb.2017.00038> (2017).
3. Nigro, J. M. *et al.* Scrambled exons. *Cell* **64**, 607–613, [https://doi.org/10.1016/0092-8674\(91\)90244-s](https://doi.org/10.1016/0092-8674(91)90244-s) (1991).
4. Zhang, Y. *et al.* Circular intronic long noncoding RNAs. *Molecular cell* **51**, 792–806, <https://doi.org/10.1016/j.molcel.2013.08.017> (2013).
5. Knupp, D. & Miura, P. CircRNA accumulation: A new hallmark of aging? *Mechanisms of ageing and development* **173**, 71–79, <https://doi.org/10.1016/j.mad.2018.05.001> (2018).
6. Memczak, S. *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* **495**, 333–338, <https://doi.org/10.1038/nature11928> (2013).
7. Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J. & Kleinschmidt, A. K. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. *Proceedings of the National Academy of Sciences of the United States of America* **73**, 3852–3856, <https://doi.org/10.1073/pnas.73.11.3852> (1976).

8. Hsu, M. T. & Coca-Prados, M. Electron microscopic evidence for the circular form of RNA in the cytoplasm of eukaryotic cells. *Nature* **280**, 339–340, <https://doi.org/10.1038/280339a0> (1979).
9. Cocquerelle, C., Mascrez, B., Hetuin, D. & Bailleul, B. Mis-splicing yields circular RNA molecules. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **7**, 155–160, <https://doi.org/10.1096/fasebj.7.1.7678559> (1993).
10. Danan, M., Schwartz, S., Edelheit, S. & Sorek, R. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic acids research* **40**, 3131–3142, <https://doi.org/10.1093/nar/gkr1009> (2012).
11. Chu, Q. *et al.* PlantCircBase: A Database for Plant Circular RNAs. *Molecular plant* **10**, 1126–1128, <https://doi.org/10.1016/j.molp.2017.03.003> (2017).
12. Chen, L., Huang, C., Wang, X. & Shan, G. Circular RNAs in Eukaryotic Cells. *Current genomics* **16**, 312–318, <https://doi.org/10.2174/1389202916666150707161554> (2015).
13. Holdt, L. M., Kohlmaier, A. & Teupser, D. Molecular roles and function of circular RNAs in eukaryotic cells. *Cellular and molecular life sciences: CMLS* **75**, 1071–1098, <https://doi.org/10.1007/s00018-017-2688-5> (2018).
14. Salzman, J., Chen, R. E., Olsen, M. N., Wang, P. L. & Brown, P. O. Cell-type specific features of circular RNA expression. *PLoS genetics* **9**, e1003777, <https://doi.org/10.1371/journal.pgen.1003777> (2013).
15. Hansen, T. B. *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* **495**, 384–388, <https://doi.org/10.1038/nature11993> (2013).
16. Qu, S. *et al.* Circular RNA: A new star of noncoding RNAs. *Cancer letters* **365**, 141–148, <https://doi.org/10.1016/j.canlet.2015.06.003> (2015).
17. Wang, M. *et al.* Circular RNAs: A novel type of non-coding RNA and their potential implications in antiviral immunity. *International journal of biological sciences* **13**, 1497–1506, <https://doi.org/10.7150/ijbs.22531> (2017).
18. Tang, Y. Y. *et al.* Circular RNA hsa_circ_0001982 Promotes Breast Cancer Cell Carcinogenesis Through Decreasing miR-143. *DNA and cell biology* **36**, 901–908, <https://doi.org/10.1089/dna.2017.3862> (2017).
19. Wang, F., Wang, J., Cao, X., Xu, L. & Chen, L. Hsa_circ_0014717 is downregulated in colorectal cancer and inhibits tumor growth by promoting p16 expression. *Biomedicine & pharmacotherapy = Biomedicine & pharmacotherapie* **98**, 775–782, <https://doi.org/10.1016/j.biopha.2018.01.015> (2018).
20. Han, D. *et al.* Circular RNA circMTO1 acts as the sponge of microRNA-9 to suppress hepatocellular carcinoma progression. *Hepatology (Baltimore, Md.)* **66**, 1151–1164, <https://doi.org/10.1002/hep.29270> (2017).
21. Qiu, M. *et al.* The Circular RNA circPRKCI Promotes Tumor Growth in Lung Adenocarcinoma. *Cancer research* **78**, 2839–2851, <https://doi.org/10.1158/0008-5472.can-17-2808> (2018).
22. Fan, C., Lei, X., Fang, Z., Jiang, Q. & Wu, F. X. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. *Database: the journal of biological databases and curation* **2018**, <https://doi.org/10.1093/database/bay044> (2018).
23. Zhao, Z. *et al.* circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. *Cell death & disease* **9**, 475, <https://doi.org/10.1038/s41419-018-0503-3> (2018).
24. Yao, D. *et al.* Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Scientific reports* **8**, 11018, <https://doi.org/10.1038/s41598-018-29360-3> (2018).
25. Fan, C., Lei, X. & Wu, F. X. Prediction of CircRNA-Disease Associations Using KATZ Model Based on Heterogeneous Networks. *International journal of biological sciences* **14**, 1950–1959, <https://doi.org/10.7150/ijbs.28260> (2018).
26. Yan, C., Wang, J. & Wu, F. X. DWNN-RLS: regularized least squares method for predicting circRNA-disease associations. *BMC bioinformatics* **19**, 520, <https://doi.org/10.1186/s12859-018-2522-6> (2018).
27. Lei, X., Fang, Z., Chen, L. & Wu, F. X. PWCDA: Path Weighted Method for Predicting circRNA-Disease Associations. *International journal of molecular sciences* **19**, <https://doi.org/10.3390/ijms19113410> (2018).
28. Xiao, Q., Luo, J. & Dai, J. Computational Prediction of Human Disease-associated circRNAs based on Manifold Regularization Learning Framework. *IEEE journal of biomedical and health informatics*, <https://doi.org/10.1109/jbhi.2019.2891779> (2019).
29. Wei, H. & Liu, B. iCircDA-MF: identification of circRNA-disease associations based on matrix factorization. *Briefings in bioinformatics*, <https://doi.org/10.1093/bib/bbz057> (2019).
30. Zhao, Q., Yang, Y., Ren, G., Ge, E. & Fan, C. Integrating Bipartite Network Projection and KATZ Measure to Identify Novel CircRNA-Disease Associations. *IEEE transactions on nanobioscience*, <https://doi.org/10.1109/tnb.2019.2922214> (2019).
31. Zhang, W., Yu, C., Wang, X. & Liu, F. Predicting CircRNA-Disease Associations Through Linear Neighborhood Label Propagation Method. *IEEE Access* **7**, 83474–83483, <https://doi.org/10.1109/ACCESS.2019.2920942> (2019).
32. Deng, L., Zhang, W., Shi, Y. & Tang, Y. Fusion of multiple heterogeneous networks for predicting circRNA-disease associations. *Scientific reports* **9**, 9605, <https://doi.org/10.1038/s41598-019-45954-x> (2019).
33. Lee, I. & Nam, H. Identification of drug-target interaction by a random walk with restart method on an interactome network. *BMC bioinformatics* **19**, 208, <https://doi.org/10.1186/s12859-018-2199-x> (2018).
34. Anastasiadi, Z., Lianos, G. D., Ignatiadou, E., Harisis, H. V. & Mitsis, M. Breast cancer in young women: an overview. *Updates in surgery* **69**, 313–317, <https://doi.org/10.1007/s13304-017-0424-1> (2017).
35. Martinez Rodriguez, R. H., Buisan Rueda, O. & Ibarz, L. Bladder cancer: Present and future. *Medicina clinica* **149**, 449–455, <https://doi.org/10.1016/j.medcli.2017.06.009> (2017).
36. Brody, H. Colorectal cancer. *Nature* **521**, S1, <https://doi.org/10.1038/521S1a> (2015).
37. Wang, D., Wang, J., Lu, M., Song, F. & Cui, Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics (Oxford, England)* **26**, 1644–1650, <https://doi.org/10.1093/bioinformatics/btq241> (2010).
38. van Laarhoven, T., Nabuurs, S. B. & Marchiori, E. Gaussian interaction profile kernels for predicting drug-target interaction. *Bioinformatics (Oxford, England)* **27**, 3036–3043, <https://doi.org/10.1093/bioinformatics/btr500> (2011).
39. Park, K., Kim, D., Ha, S. & Lee, D. Predicting Pharmacodynamic Drug-Drug Interactions through Signaling Propagation Interference on Protein-Protein Interaction Networks. *PLOS ONE* **10** (2015).

Acknowledgements

We thank the financial support which comes from National Natural Science Foundation of China (61672334, 61972451, 61902230) and the Fundamental Research Funds for the Central Universities (No. GK201901010).

Author contributions

X.L. contributed to the conception of the study, the design of the method and writing of the manuscript. C.B. contributed to designing the method and carrying out the experiments. C.B. performed the data analyses and wrote the manuscript. X.L. revised the manuscript and polished the English expression. All the authors have read and approved the final manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-59040-0>.

Correspondence and requests for materials should be addressed to X.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020