Yan *et al. BMC Bioinformatics* 2018, **19**(Suppl 19):520
https://doi.org/10.1186/s12859-018-2522-6

BMC Bioinformatics

RESEARCH Open Access

DWNN-RLS: regularized least squares method for predicting circRNA-disease associations

Cheng Yan[1,2], Jianxin Wang[1*] and Fang-Xiang Wu[3]

From 29th International Conference on Genome Informatics
Yunnan, China. 3-5 December 2018

Abstract

Background: Many evidences have demonstrated that circRNAs (circular RNA) play important roles in controlling gene expression of human, mouse and nematode. More importantly, circRNAs are also involved in many diseases through fine tuning of post-transcriptional gene expression by sequestering the miRNAs which associate with diseases. Therefore, identifying the circRNA-disease associations is very appealing to comprehensively understand the mechanism, treatment and diagnose of diseases, yet challenging. As the complex mechanism between circRNAs and diseases, wet-lab experiments are expensive and time-consuming to discover novel circRNA-disease associations. Therefore, it is of dire need to employ the computational methods to discover novel circRNA-disease associations.

Result: In this study, we develop a method (DWNN-RLS) to predict circRNA-disease associations based on Regularized Least Squares of Kronecker product kernel. The similarity of circRNAs is computed from the Gaussian Interaction Profile(GIP) based on known circRNA-disease associations. In addition, the similarity of diseases is integrated by the mean of GIP similarity and sematic similarity which is computed by the direct acyclic graph (DAG) representation of diseases. The kernels of circRNA-disease pairs are constructed from the Kronecker product of the kernels of circRNAs and diseases. DWNN (decreasing weight k-nearest neighbor) method is adopted to calculate the initial relational score for new circRNAs and diseases. The Kronecker product kernel based regularised least squares approach is used to predict new circRNA-disease associations. We adopt 5-fold cross validation (5CV), 10-fold cross validation (10CV) and leave one out cross validation (LOOCV) to assess the prediction performance of our method, and compare it with other six competing methods (RLS-avg, RLS-Kron, NetLapRLS, KATZ, NBI, WP).

Conlusion: The experiment results show that DWNN-RLS reaches the AUC values of 0.8854, 0.9205 and 0.9701 in 5CV, 10CV and LOOCV, respectively, which illustrates that DWNN-RLS is superior to the competing methods RLS-avg, RLS-Kron, NetLapRLS, KATZ, NBI, WP. In addition, case studies also show that DWNN-RLS is an effective method to predict new circRNA-disease associations.

Keywords: CircRNA, CircRNA-disease association, Gaussian interaction profile, Kron-RLS

*Correspondence: jxwang@mail.csu.edu.cn
[1]School of Information Science and Engineering, Central South University, 932 South Lushan Rd, 410083 ChangSha, China
Full list of author information is available at the end of the article

Background

Circular RNAs (circRNAs) are a class of endogenous non-coding RNAs with distinct properties and diverse cellular functions, unlike the linear RNAs with 5' and 3' termini which reflect start and stop of the RNA polymerase on the DNA template, and are generated by back splicing (3'-5') or lariat introns [1–4]. The circRNAs are not easy to be degraded by exoribonucleases because they lack free ends [5, 6]. As forming a circRNA is usually considered a rare event in cells, it was suggested that they may be considered errors of normal splicing process [4, 7]. Therefore, despite their existence in both unicellular and multicellular organisms, they have been previously even disregarded as transcriptional noise or artifacts [8]. Nevertheless, with the advances of high-throughput deep sequencing and functional genomics, the knowledge of circRNAs has recently been learned substantially [9, 10].

To date, circRNAs have been found in various tissues and cell lines of plants, animals and so on [4, 11, 12]. Some circRNAs can be translated in some tissues or translated into a protein under splicing-dependent, cap-independent manner or other certain conditions [13]. Furthermore, circRNAs are expected to have other functions independent of their host genes because they have much longer half-life than other linear RNA transcripts [10]. Many circRNAs can regulate gene expression because they have strong potential to act as miRNA sponges or decoys [14]. In addition, some circRNAs can also function as protein sponges or decoys, and the best example is that protein MBL is prevented to bind to other targets when being tethered to a circRNA [15]. CircRNA circFoxo3 can also act as a protein scaffold, which binds to sites for mouse MDM2 and p53 [16]. Unlike the above functions of circRNAs are based on the fact that they are located to the cytoplasm, some circRNAs such as exon-intron circRNAs are retained in the nucleus and they may promote with transcription [17].

Through the understanding of functions of circRNAs, many evidences have shown that circRNAs play an important role in occurrence of human complex diseases, such as cancer [18]. CircRNA ciRS-7 has significant implications for diseases through efficiently regulating the activity of miRNA miR-7 [19]. Likewise, by sponging the miR-7, miR-17 and miR-214, cir-ITCH can increase the level of ITCH which further inhibits the Wnt pathway that is frequently aberrant in cancers [20, 21]. SRY can affect the proliferation, migration and invasion of cholangiocarcinoma cells, which is the sponge of miR-138 and can strongly suppress its level [22, 23]. CircRNA-MYLK level is elevated and correlated with BC (bladder carcinoma) progression and plays an oncogenic role in BC in vitro and vivo [24]. Circ-Foxo3 was minimally expressed in patient tumor samples and in a panel of cancer cells and its expression was found to be significantly increased during the cancer cell apoptosis [16, 25]. Circular RNA MTO1 can suppress hepatocellular carcinoma progression by acting as the sponge of miR-9 [26]. In addition, the aberrant expression of circCCDC66 also is associated with a late-stage diagnosis and metastases [27].

In recent years, some databases about circRNAs have been developed to further study the function mechanism of circRNAs. CircBase is the first database about circRNAs, which merges and unifies data sets of circRNAs and provides the interface to access, download, and browse the evidence supporting their expression within the genomic context [28]. CircRNADb is a comprehensively annotated human circular RNAs database, which containes 32,914 human exonic circRNAs from diversified sources and provides the genomic information, exon splicing, genome sequence, internal ribosome entry site (IRES), open reading frame (ORF) and references of these circRNAs [29]. PlantcircBase is a database of plant circRNAs, which also provided other functions such as visualization of the structures of circRNA based on their genomic position [12]. Likewise, PlantCircNet also is a database of plant circRNAs, which has the main feature of plantCircNet to provide visualized plant circRNA-miRNA-mRNA regulatory networks and can identify metabolic effects of circRNAs [30]. ExoRBase is a web-accessible database, which provides the circRNA, lncRNA and mRNA information by RNA-seq data analyses of human blood exosomes [31]. CircNet provides tissue-specific circRNA expression profiles and circRNA-miRNA-gene regulatory networks by utilizing sequencing datasets to systematically identify the expression of circRNAs in RNA-seq samples [32]. TSTD also provides the tissue-specific circRNAs and further characterizes the functions of these circRNAs [33]. The cancer somatic mutations that alter miRNA targeting and functioning are provided by SomamiR 2.0 database which also collects the associations between miRNA and other competing endogenous RNAs such as mRNAs, circRNAs and lncRNAs [34]. The CSCD is also a cancer-specific circRNAs database which identifies the cancer-specific circRNAs by analyzed the RNA-seq samples and further predicts the miRNA response element sites and RNA binding protein sites of each circRNA [35]. Circ2Traits is the circRNA-disease associations database, which is constructed by circRNA-miRNA associations, miRNA-disease associations and disease-SNPS associations [18]. To our knowledge, CircR2Disease is the first manually curated database about circRNA-disease associations by reviewing existing literatures and provides the important foundation to study the associations of circRNAs and diseases [36].

In general, we have obtained some significant progresses in understanding features and functions of circRNAs. In addition, some databases about circRNAs have also been constructed. However, current studies of circRNA-disease

associations mainly focus on biomedical experimentations that are notoriously expensive and time-consuming. Therefore, there is a very urgent need to predict circRNA-disease associations by computational methods. To our knowledge, the development of computational approach is very limited because the databases of circRNA-disease associations are incomplete. However, circR2Disease provides the chance to effectively predict novel circRNA-disease associations through developing computational methods.

In this study, we develop a novel method (call DWNN-RLS) to predict new circRNA-disease associations. Firstly, DWNN-RLS computes the Gaussian interaction profile (GIP) kernel similarities of circRNAs and diseases based on the known circRNA-disease associations. By considering their direct acyclic graph(DAG) representation, the sematic similarity of diseases is also calculated. We further obtain the final similarity of diseases with the mean of GIP similarity and sematic similarity. Then the association possibility scores of circRNA-disease pairs are predicted by Kronecker product kernel based Regularized Least Squares approach. The kernels of circRNA-disease pairs are calculated by the Kronecker product of kernels of circRNAs and diseases. Furthermore, the decreasing weight k-nearest neighbor (DWNN) method is used to calculate the initial relational scores of new circRNAs and new diseases. In order to assess the prediction performance of DWNN-RLS and compare with other competing methods, we conduct 5-fold cross validation (5CV), 10-fold cross validation (10CV) and leave-one-out cross validation (LOOCV). The experiment results demonstrate that DWNN-RLS outperforms other six competing methods (RLS-avg, RLS-Kron, NetLapRLS, KATZ, NBI, WP) in terms of AUC (area under the ROC curve) values. Specifically, the AUC values of DWNN-RLS in 5CV, 10CV and LOOCV reach 0.8854, 0.9205 and 0.9701, respectively, which are superior to the second best results (KATZ: 0.8224 and 0.8343, RLS-avg: 0.9169). Furthermore, the prediction ability of DWNN-RLS also is illustrated by the case studies.

Methods

Materials

In this study, we download the known circRNA-disease associations data from the CircR2Disease database (http://bioinfo.snnu.edu.cn/CircR2Disease/). These circRNA-disease associations were curated circRNA-disease associations from the existing literature prior to 31 March 2018. After removing the duplicated data, we obtain the benchmark dataset that includes 725 circRNA-disease associations, 676 circRNAs and 100 diseases. In addition, the Mesh database [37] (https://www.nlm.nih.gov/bsd/disted/meshtutorial/themeshdatabase/) is used to compute the sematic similarity of diseases.

Similarity of circRNAs

As the successful application of GIP kernel similarity in other relative areas [38–42], we also use it to calculate the similarities of circRNAs. The GIP kernel was computed from the known circRNA-disease associations. Let $C = \{c_1, c_2, ..., c_{N_c}\}$ be the set of N_c circRNAs and $D = \{d_1, d_2, ..., d_{N_d}\}$ be the set of N_d diseases. Let matrix $Y \in R^{N_c \times N_d}$ represents known circRNA-disease associations, in which the value of y_{ij} is 1 if circRNA i and disease j exists a known association, otherwise 0. Then the GIP similarity of circRNA c_i and circRNA c_j can be computed as follows:

$$S_c(c_i, c_j) = G_c(c_i, c_j) = exp\left(-\gamma_c ||y_{c_i} - y_{c_j}||^2\right) \quad (1)$$

$$\gamma_c = 1/\left(\frac{1}{N_c}\sum_{i=1}^{N_c} ||y_{c_i}||^2\right), \quad (2)$$

where $y_{c_i} = \{y_{i1}, y_{i2}, ..., y_{iN_d}\}$ and $y_{c_j} = \{y_{j1}, y_{j2}, ..., y_{jN_d}\}$ are the association profiles of circRNA c_i and circRNA c_j, respectively. Since the GIP kernel is computed by a decaying function of the distance between the vectors, this function is of the form of a bell-shaped curve. In addition, since a larger value of γ_c yields a narrower bell while a smaller value of γ_c yields a wider bell, the parameter γ_c can be used to regulate the bandwidth of kernel. In this study, parameter γ_c is computed as the reciprocal of average number of associations per circRNA.

Similarity of diseases

Firstly, we also compute the GIP similarity of disease d_i and disease d_j as follows:

$$G_d(d_i, d_j) = exp\left(-\gamma_d ||y_{d_i} - y_{d_j}||^2\right) \quad (3)$$

$$\gamma_d = 1/\left(\frac{1}{N_d}\sum_{i=1}^{N_d} ||y_{d_i}||^2\right), \quad (4)$$

where $y_{d_i} = \{y_{1i}, y_{2i}, ..., y_{N_c i}\}^T$ is the association profiles of disease d_i while $y_{d_j} = \{y_{1j}, y_{2j}, ..., y_{N_c j}\}^T$ is the association profiles of disease d_j. In addition, the parameter γ_d is used to regulate the bandwidth of kernel.

Secondly, we use the Mesh descriptions of diseases to compute the sematic similarity. Specifically, for disease A which can be represented by a DAG $(DAG_A, DAG_A = T_A, E_A)$ in mesh database. Set T_A includes the parent diseases nodes of A and itself while set E_A includes the direct edges between disease nodes within T_A. The similarity of diseases A and B can be calculated as follows:

$$D_{semsim}(A, B) = \frac{\sum_{t \in T_A \cap T_B} (SV_A(t) + SV_B(t))}{Sem(A) + Sem(B)}, \quad (5)$$

where $SV_A(t)(SV_B(t))$ is the sematic value between disease $A(B)$ and t which is the all common ancestors of diseases A and B. In addition, $Sem(A)$ and $Sem(B)$ are the sematic values of diseases A and B, respectively. For disease A, the $Sem(A)$ and $SV_A(t)$ can be calculated as follows:

$$Sem(A) = \sum_{t \in T_A} SV_A(t), \tag{6}$$

$$SV_A(t) = \begin{cases} 1, t = A \\ \Delta^w, t = \textit{the smallest w layer ancestor node of } A \end{cases} \tag{7}$$

where Δ is the layer contribution factor between disease node and its direct ancestor disease nodes in DAG. The value of Δ is set to 0.5 in this study [37].

After computing the GIP similarity and sematic similarity of diseases, we integrate the final similarity of diseases with their mean as follows:

$$S_d = \frac{G_d + D_{semsim}}{2}, \tag{8}$$

DWNN for new circRNAs and diseases

The good performance of prediction method largely depends on the quality of known circRNA-disease associations. In fact, new circRNAs (or new diseases) have no any association with diseases (or circRNAs). In this study, we use the DMNN to compute the initial association score based on similarities of circRNAs and diseases. Specifically, the initial association score between new circRNA c_i and disease d_j can be calculated as follows:

$$y\left(c_i, d_j\right) = \frac{\sum G_c^{il} y_{lj}}{\sum G_c^{il}}, c_l \in N(c_i) \tag{9}$$

where $N(c_i)$ is the set of k_{c_i} nearest neighbors of new circRNA c_i. The parameter k_{c_i} is calculated as follows:

$$k_{c_i} = \begin{cases} max(k), if \ \dfrac{1 - simset(c_i)_l}{l} \le \epsilon^l, 1 \le l \le k \\ 0, otherwise \end{cases} \tag{10}$$

where $simset(c_i)_l$ is the l-th similarity value of the ranked vector based on similarity between circRNA c_i and other circRNAs from high to low. Furthermore, the parameter ϵ is used to control the range of ϵ^l that is used to select k nearest neighbors for each new circRNA and disease. In this study, the value of ϵ is set to 1, so the value of ϵ^l is 1 and all neighbors are used to calculate initial association score.

Similarly, we also compute the initial association scores of new disease d_j and circRNA c_i as follows:

$$y\left(c_i, d_j\right) = \frac{\sum G_d^{jl} y_{il}}{\sum G_d^{jl}}, d_l \in N(d_j) \tag{11}$$

where $N(d_j)$ is the set of k_{d_j} nearest neighbors of new disease d_j. The parameter k_{d_j} is also calculated as follows:

$$k_{d_j} = \begin{cases} max(k), if \ \dfrac{1 - simset(d_j)_l}{l} \le \epsilon^l, 1 \le l \le k \\ 0, otherwise \end{cases} \tag{12}$$

where $simset(d_j)_l$ is the l-th similarity value of the ranked vector based on similarity between disease d_j and other diseases from high to low. Parameter ϵ is also used to control the range for selecting neighbors.

Kronecker product kernel based regularized least squares(RLS-Kron)

In this study, we use RLS-Kron method to predict new circRNA-disease associations [38, 39, 43]. Based on the kernel K , the predicted circRNA-disease associations matrix has a simple closed-form solution as follows:

$$vec\left(\hat{Y}^T\right) = K(K + \sigma I)^{-1} vec\left(Y^T\right) \tag{13}$$

in which the parameter σ is a regularizations parameter and is set to 0.2 in this study. Kron-RLS has no any prediction ability when σ is set to 0. The kernel K is calculated from the Kronecker product $K_c \otimes K_d$ of the circRNA kernel and disease kernel, which is defined as follows:

$$K\left(\left(c_i, d_j\right), \left(c_u, d_v\right)\right) = K_c(d_i, d_u) K_d(t_j, t_v) \tag{14}$$

where matrices K_c and K_d are the similarity matrices of circRNAs and diseases, respectively. In addition, in order to calculate the predicted matrix, Kron-RLS needs to compute the inverse of an $N_c N_d \times N_c N_d$ matrix. Therefore, we also use an effective method based on matrix eigenvalue decomposition. According to the matrix theory, the eigenvalues (vectors) of a kronecker product are the Kronecker product of eigenvalues (vectors). Specifically, the kernal can be calculated as follows:

$$K = K_c \otimes K_d = \vee \wedge \vee^T \tag{15}$$

where $\wedge = \wedge_c \otimes \wedge_d$ and $\vee = \vee_c \otimes \vee_d$ are all derived from the eigenvalues decompositions of the two kernel matrices K_c and K_d. As K_c and K_d are real symmetric matrices, their specific eigenvalues decompositions process are defined as follows:

$$K_c = \vee_c \wedge_c \vee_c^T \tag{16}$$

$$K_d = \vee_d \wedge_d \vee_d^T \tag{17}$$

where $\vee_c$ and $\vee_d$ are orthogonal matrices whose columns are the eigenvectors of K_c and K_d, respectively. $\wedge_c$ and $\wedge_d$ are diagonal matrices whose diagonal entries are the eigenvalues of K_c and K_d, respectively. Therefore, the final predicted circRNA-disease associations matrix $\hat{Y}$ can be calculated as follows:

$$\hat{Y} = \vee_c Z^T \vee_d^T \tag{18}$$

$$vec(Z) = (\wedge_c \otimes \wedge_d)(\wedge_c \otimes \wedge_d + \sigma I)^{-1} vec\left(\vee_d^T Y^T \vee_c\right) \quad (19)$$

Results

Performance evaluation

In this study, we conduct 5CV, 10CV and LOOCV to evaluate the performance of DWNN-RLS for predicting new circRNA-disease associations. AUC (area under the ROC curve) value is used as the evaluation metric.

We perform 10 repetitions of 10CV and 5CV. That is, under 10CV, the known circRNA-disease associations data are divided into 10 folds, and each fold takes in turn as the test set and the rest as the train set at each time. Similarly, the data set are randomly divided into 5 folds and each fold takes in turn as the test data and the rest as the train set on each time. In LOOCV, each known circRNA-disease association is in turn chosen as the test set while the rest known circRNA -disease associations as the train set. The larger AUC values show the better prediction ability of the method, while if AUC value is less than or equal to 0.5, the prediction method has no prediction ability.

Comparison with other methods

As there is no competing computational method for predicting circRNA-disease associations in the literature, to assess the performance of our method, we also compare DWNN-RLS against other six effective methods in other relevant prediction issues. These methods include RLS-avg [38], RLS-Kron [38], NetLapRLS [44], KATZ [45, 46], NBI [47] and WP [47, 48]. We briefly review them here. RLS-avg use the average of the output values which are computed from two kernels, respectively. RLS-Kron compute the prediction scores by Kronecker product kernel based on regularised least squares approach. NetLapRLS is used to predict circRNA-disease associations by exploiting information on similarities of links and nodes. KATZ is a network-based method which considers the number of walks between network nodes and lengths in a heterogeneous network to predict associations. NBI is also a network-based method to infer new associations, which only uses cricRNA-disease bipartite network topology similarity. WP and DBSI are recommendation models which directly use the similarities of circRNAs and diseases.

Figure 1 shows the AUC curves of seven prediction methods on CircR2Disease data set in terms of 5CV. The AUC value of DWWN-RLS is the highest among the seven methods, indicating that the prediction performance of DWWN-RLS is better than other methods.

Figure 2 shows the AUC curves of seven prediction methods in terms of 10CV on CircR2Disease dataset. The AUC value of DWWN-RLS reaches 0.9205, which is better than other methods (RLS-avg: 0.7477, RLS-Kron: 0.8103, NetLapRLS: 0.6744, KATZ: 0.8343, NBI: 0.6648, WP: 0.6198).

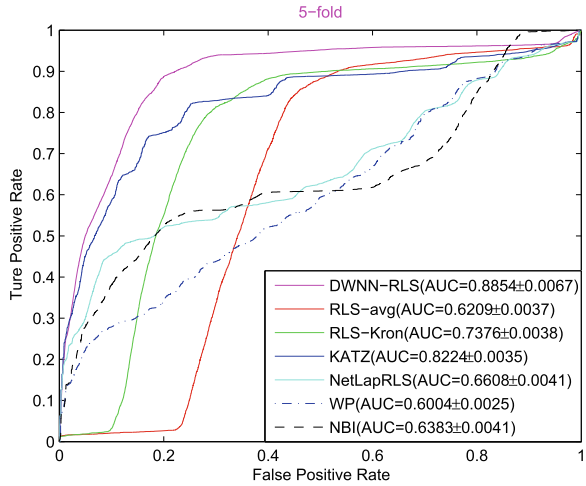

Fig. 1 The AUC curves of seven methods in the 5CV

Figure 3 shows the prediction comparison result between DWWN-RLS and other six methods in terms of LOOCV on CircR2Disease data set. We can see from the Fig.3 that the prediction performance of DWWN-RLS (0.9701) is superior to other methods in terms of AUC values (RLS-avg: 0.9169, RLS-Kron: 0.9088, NetLapRLS: 0.6905, KATZ: 0.8432, NBI: 0.699, WP: 0.6362).

Note that the advantage of prediction performance is more obvious in 10CV and LOOCV than 5CV, indicating that DWWN-RLS can achieve good result based on more known circRNA-disease associations. In addition, the sematic similarity of diseases can improve the prediction performance of DWWN-RLS. When only the GIP

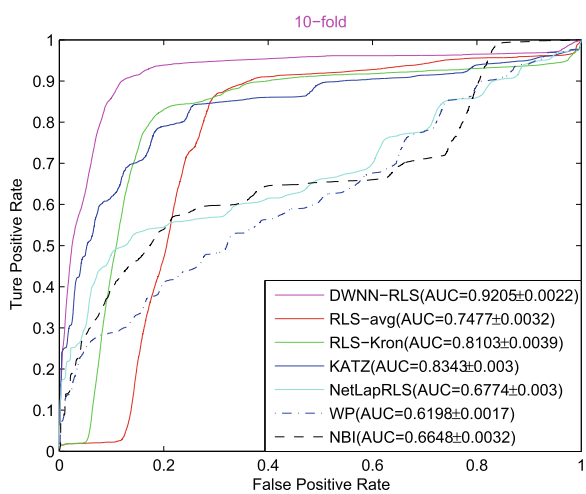

Fig. 2 The AUC curves of seven methods in the 10CV

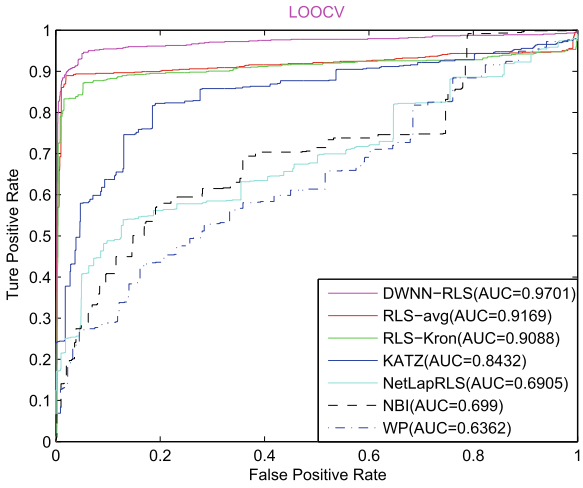

Fig. 3 The AUC curves of seven methods in the LOOCV

similarity is used, the AUCs of DWNN-RLS are 0.8368, 0.8819 and 0.9423 in 5CV, 10CV and LOOCV, respectively. When the GIP similarity combined with the disease sematic similarity, DWNN-RLS obtains the increased AUCs of 0.8854, 0.9205 and 0.9701 in 5CV, 10CV and LOOCV. By comparing with RLS-Kron method, the DMNN method also can improve the prediction performance. Comparing with KATZ, NBI and WP methods, we think that DWNN-RLS is a machining learning model and has the objective function and solution process that is beneficial to obtain better prediction performance.

Parameter analysis for ϵ and σ

To further understand the robustness of DWWN-RLS method, we analyze the influence of parameters ϵ and σ on the prediction performance in 10CV. The parameter ϵ is used to control the range for selecting k nearest neighbors of cicrRNAs and diseases. The parameter σ is the regularization parameter of DWWN-RLS method. The value of ϵ is set to be 1.0 when analyzing parameter σ. Furthermore, we also set the default value of σ to be 0.2 when analyzing parameter ϵ. With parameter σ of 0.2, Table 1 demonstrates the prediction performance of DWWN-RLS method in 10CV when ϵ ranges from 0.1 to 1.0 with 0.1 increments. The prediction performance of DWWN-RLS method is best when ϵ is set to be 1.0, indicating that all neighbors of circRNAs and diseases are involved in calculating their initial associations scores.

Furthermore, Table 2 describes the prediction performances of DWWN-RLS with different values of σ when ϵ is set to be 1.0. We can see from Table 2 that DWWN-RLS obtains the best prediction performance when σ is set to be 0.2. Therefore, in this study, we set the default value of σ to be 0.2.

Case studies

After confirming the prediction performance and robustness of DWWN-RLS method in 10CV, 5CV and LOOCV, we further analyze the prediction ability of DWWN-RLS in discovering new circRNA-disease associations. In predicting new circRNA-disease associations, all known circRNA-disease associations on CircR2Disease dataset are chosen as the train set and all other circRNA-disease pairs are the candidate circRNA-disease associations. We adapt DWWN-RLS to compute the prediction scores for these candidate circRNA-disease pairs. Here, we analyze the prediction results of Atherosclerotic vascular disease and Breast cancer.

Atherosclerotic vascular disease is responsible for the majority of cases of CVD (Cardiovascular disease) in both developing and developed countries, which encompasses coronary heart disease, cerebrovascular disease, and peripheral arterial disease, and which also result the CVD, the leading cause of death and disability all over the world [49, 50]. Table 3 shows that 2 of top 10 predicted associations are confirmed in the previous literature. Elevated cANRIL expression could lead to worse EC (endothelial cells) inflammation, exacerbating AS (atherosclerosis) [51]. CANRIL is transcribed at a locus of atherosclerotic cardiovascular disease on chromosome 9p21, and induces nucleolar stress and apoptosis, and inhibits the proliferation in smooth muscle cells and macrophages [52]. The cZNF292 also associates with atherosclerotic cardiovascular disease by stimulating angiogenesis through vascular sprouting and cell proliferation [53].

There is approximately 1 in 12 women developing breast cancer in Western Europe and the United States, and which is characterized by a distinct metastatic pattern

Table 1 The 10CV prediction performance of various parameter values of ϵ ranging from 0.1 to 1.0 with 0.1 increments, the best result is in bold face

ϵ	0.1	0.2	0.3	0.4	0.5
AUC	0.7927±0.0048	0.7927±0.0035	0.7922±0.0042	0.7902±0.0034	0.7920±0.0035
ϵ	0.6	0.7	0.8	0.9	1.0
AUC	0.7922±0.0042	0.7889±0.0032	0.7903±0.0047	0.7897±0.0044	**0.9205±0.0022**

Table 2 The 10CV prediction performance of various parameter values of σ ranging from 0.1 to 1.0 with 0.1 increments, the best result is in bold face

σ	0.1	0.2	0.3	0.4	0.5
AUC	0.9200±0.0024	**0.9205±0.0022**	0.9182±0.0023	0.9154±0.0018	0.9136±0.0021
σ	0.6	0.7	0.8	0.9	1.0
AUC	0.9110±0.0025	0.9078±0.0033	0.9042±0.0020	0.9041±0.0020	0.9010±0.0025

involving the regional lymph nodes, bone marrow, lung and liver [54, 55]. Table 4 shows the validation results of top 10 new circRNA-disease associations predicted by DWNN-RLS. There is 3 out of top 10 predicted associations that can be validated in previous studies. CircRNAs circGFRA1 and GFRA1 act as ceRNAs in triple negative breast cancer by regulating miR-34a [56]. The human breast cancer cell line MDA-MB-231 are stably transfected with circ-Foxo3, the ectopic expression of the Foxo3 circular RNA could suppress tumor growth, cancer cell proliferation and survival [25]. CDR1as contains more than 70 selectively conserved target sites of miR-7 which can directly downregulate oncogenes in cancers such as breast cancer [57].

Above case studies demonstrate that there are a number of prediction results that have not been confirmed by previous literature. To our knowledge, a possible reason is that the database Circ2Disease are still limited and the new studies have not been published yet. In summary, these predicted circRNA-disease associations deserve being studied and considered in the future.

Table 3 The validation results of predicted top 10 new circRNA-disease associations of Atherosclerotic vascular disease

Disease	CircRNA	Rank	Source
Atherosclerotic vascular disease	cANRIL	1	PMID:28683453, PMID:28946214
	hsa_circ_0003575	2	Unknown
	circSMARCA5/ hsa_circ_0001445	3	Unknown
	hsa_circ_0000284/ circHIPK3	4	Unknown
	hsa_circ_0004383/ cZNF292	5	PMID:27836747
	circRNA-chr19	6	Unknown
	CircDOCK1/ hsa_circ_100721	7	Unknown
	mmu-circRNA-015947	8	Unknown
	hsa-circRNA 2149	9	Unknown
	circRar1	10	Unknown

Discussion

With the advances of RNA-Seq, high-throughput sequencing and other techniques, we have achieved some important progresses in understanding characteristics and functions of cricRNAs. CricRNAs may play key roles in diseases as miRNA sponges or decoys, protein sponges or decoys and regulation gene transcription. Therefore, systematically understanding association between circRNAs and diseases has become an important issue of bioinformatics research, which is beneficial to disease diagnose and treatment. Although some databases about circRNA have been established in recent years, these databases rarely focused on the associations between circRNAs and diseases. The computation methods for predicting circRNA-disease associations are also lacking because of these limitations. To our knowledge, CircR2Disease is the first database about circRNA-disease associations, which provides the chance to develop effective methods to identify novel associations between circRNAs and diseases.

Conclusion

DWNN-RLS method is developed to predict new associations between circRNAs and diseases on CircR2Disease dataset. Firstly, DWNN-RLS computes the GIP similarities of circRNAs and diseases based on the known circRNA-disease associations. Secondly, we further compute the sematic similarity of disease and compute the final similarity of diseases with the mean of GIP similarity and sematic similarity. Finally, the Kron-RLS is used to predict novel circRNA-disease associations based on their similarities. 10CV, 5CV and LOOCV are used to evaluate the prediction performance of DWNN-RLS. In addition, we use the DWNN to calculate the initial associations scores for new circRNAs and diseases. We also compare our method with other six methods. In terms of 10CV, 5CV and LOOCV, DWNN-RLS all achieves the best prediction performance. In addition, we also show that DWNN-RLS method may achieves better prediction performance with the more known circRNA-disease associations. Case studies further illustrate the prediction performance of DWNN-RLS.

However, there still exist some limitations in DWNN-RLS. We all know that cricRNAs can function as miRNA sponges or decoys, protein sponges or decoys. In this

Table 4 The validation results of predicted top 10 new circRNA-disease associations of Breast cancer

Disease	CircRNA	Rank	Source
Breast cancer	circGFRA1/ hsa_circ_005239	1	PMID:29037220
	circUBAP2	2	Unknown
	circ-Foxo3/ hsa_circ_0006404	3	PMID:26657152
	Cir-ITCH/ hsa_circ_0001141/ hsa_circ_001763	4	Unknown
	hsa_circ_0001649	5	Unknown
	CDR1as/ciRS-7/ hsa_circ_0001946	6	PMID:28049499
	hsa_circ_0043256	7	Unknown
	hsa_circ_0016760	8	Unknown
	hsa_circ_0007385	9	Unknown
	hsa_circ_0014130	10	Unknown

study, we only use the GIP similarity of circRNAs. In the future, the similarity computation of circRNAs could consider more relevant biological network information, such as cricRNA-miRNA associations and sequence information. Similarly, the disease functional information also should be considered [58–60]. Other latest matrix factorization methods such as NRLMF [61], SRMF [62], DRRS [63] should be considered to predict cricRNA-disease association when we integrate more biological network information such as circRNA-miRNA associations, circRNA sequence information and disease functional information. Therefore, to further improve the prediction performance, we would develop a more effective approach to discover new circRNA-disease associations by reasonably integrating more biological network information.

Acknowledgements
The authors are very grateful to the anonymous reviewers for their constructive comments which have helped significantly in revising this work. The authors would like to express their gratitude for the support from the National Natural Science Foundation of China under Grant No. 61772552, No. 61420106009, No. 61622213 and No. 61732009.

Funding
Publication costs are funded by National Natural Science Foundation of China under Grant No. 61420106009.

Availability of data and materials
The used data is provided by Fan et al. [36]. Please download the data from http://bioinfo.snnu.edu.cn/CircR2Disease/ or contact the authors for data requests.

About this supplement
This article has been published as part of *BMC Bioinformatics Volume 19 Supplement 19, 2018: Proceedings of the 29th International Conference on Genome Informatics (GIW 2018): bioinformatics*. The full contents of the supplement are available online at https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-19.

Authors' contributions
JW conceived the project; CY designed the experiments; and CY performed the experiments; CY and FXW wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]School of Information Science and Engineering, Central South University, 932 South Lushan Rd, 410083 ChangSha, China. [2]School of Computer and Information,Qiannan Normal University for Nationalities, Longshan Road, 558000 DuYun, China. [3]Biomedical Engineering and Department of Mechanical Engineering, University of Saskatchewan, SKS7N5A9 Saskatoon, Canada.

Published: 31 December 2018

References

1. Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, Oliner JD, Kinzler KW, Vogelstein B. Scrambled exons. Cell. 1991;64(3):607–13.
2. Zhang Y, Zhang X-O, Chen T, Xiang J-F, Yin Q-F, Xing Y-H, Zhu S, Yang L, Chen L-L. Circular intronic long noncoding rnas. Mol Cell. 2013;51(6):792–806.
3. Knupp D, Miura P. Circrna accumulation: A new hallmark of aging? Mech Ageing Dev. 2018;173:71–9.
4. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, Maier L, Mackowiak SD, Gregersen LH, Munschauer M, et al. Circular rnas are a large class of animal rnas with regulatory potency. Nature. 2013;495(7441):333.
5. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. Circular rnas are abundant, conserved, and associated with alu repeats. Rna. 2013;19(2):141–57.
6. Enuka Y, Lauriola M, Feldman ME, Sas-Chen A, Ulitsky I, Yarden Y. Circular rnas are long-lived and display only minimal early alterations in response to a growth factor. Nucleic Acids Res. 2015;44(3):1370–83.
7. Cocquerelle C, Mascrez B, Hetuin D, Bailleul B. Mis-splicing yields circular rna molecules. FASEB J. 1993;7(1):155–60.
8. Lasda E, Parker R. Circular rnas: diversity of form and function. Rna. 2014;20(12):1829–42.
9. Ye C-Y, Chen L, Liu C, Zhu Q-H, Fan L. Widespread noncoding circular rnas in plants. New Phytol. 2015;208(1):88–95.
10. Kristensen L, Hansen T, Venø M, Kjems J. Circular rnas in cancer: opportunities and challenges in the field. Oncogene. 2018;37(5):555.
11. Danan M, Schwartz S, Edelheit S, Sorek R. Transcriptome-wide discovery of circular rnas in archaea. Nucleic Acids Res. 2011;40(7):3131–42.
12. Chu Q, Zhang X, Zhu X, Liu C, Mao L, Ye C, Zhu Q-H, Fan L. Plantcircbase: a database for plant circular rnas. Mol Plant. 2017;10(8):1126–8.
13. Legnini I, Di Timoteo G, Rossi F, Morlando M, Briganti F, Sthandier O, Fatica A, Santini T, Andronache A, Wade M, et al. Circ-znf609 is a circular rna that can be translated and functions in myogenesis. Mol Cell. 2017;66(1):22–37.
14. Qu S, Yang X, Li X, Wang J, Gao Y, Shang R, Sun W, Dou K, Li H. Circular rna: a new star of noncoding rnas. Cancer Lett. 2015;365(2):141–8.
15. Ashwal-Fluss R, Meyer M, Pamudurti NR, Ivanov A, Bartok O, Hanan M, Evantal N, Memczak S, Rajewsky N, Kadener S. circrna biogenesis competes with pre-mrna splicing. Mol Cell. 2014;56(1):55–66.
16. Du WW, Fang L, Yang W, Wu N, Awan FM, Yang Z, Yang BB. Induction of tumor apoptosis through a circular rna enhancing foxo3 activity. Cell Death Differ. 2017;24(2):357.

17. Li Z, Huang C, Bao C, Chen L, Lin M, Wang X, Zhong G, Yu B, Hu W, Dai L, et al. Exon-intron circular rnas regulate transcription in the nucleus. Nat Struct Mol Biol. 2015;22(3):256.
18. Ghosal S, Das S, Sen R, Basak P, Chakrabarti J. Circ2traits: a comprehensive database for circular rna potentially associated with disease and traits. Front Genet. 2013;4:283.
19. Li J, Zheng Y, Sun G, Xiong S. Restoration of mir-7 expression suppresses the growth of lewis lung cancer cells by modulating epidermal growth factor receptor signaling. Oncol Rep. 2014;32(6):2511–6.
20. Li F, Zhang L, Li W, Deng J, Zheng J, An M, Lu J, Zhou Y. Circular rna itch has inhibitory effect on escc by suppressing the wnt/β-catenin pathway. Oncotarget. 2015;6(8):6001.
21. Anastas JN, Moon RT. Wnt signalling pathways as therapeutic targets in cancer. Nat Rev Cancer. 2013;13(1):11.
22. Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, Kjems J. Natural rna circles function as efficient microrna sponges. Nature. 2013;495(7441):384.
23. Wang Q, Tang H, Yin S, Dong C. Downregulation of microrna-138 enhances the proliferation, migration and invasion of cholangiocarcinoma cells through the upregulation of rhoc/p-erk/mmp-2/mmp-9. Oncol Rep. 2013;29(5):2046–52.
24. Zhong Z, Huang M, Lv M, He Y, Duan C, Zhang L, Chen J. Circular rna mylk as a competing endogenous rna promotes bladder cancer progression through modulating vegfa/vegfr2 signaling pathway. Cancer Lett. 2017;403:305–17.
25. Yang W, Du W, Li X, Yee A, Yang B. Foxo3 activity promoted by non-coding effects of circular rna and foxo3 pseudogene in the inhibition of tumor growth and angiogenesis. Oncogene. 2016;35(30):3919.
26. Han D, Li J, Wang H, Su X, Hou J, Gu Y, Qian C, Lin Y, Liu X, Huang M, et al. Circular rna mto1 acts as the sponge of mir-9 to suppress hepatocellular carcinoma progression. Hepatology. 2017;66:1151–64.
27. Hsiao K-Y, Lin Y-C, Gupta SK, Chang N, Yen L, Sun HS, Tsai S-J. Noncoding effects of circular rna ccdc66 promote colon cancer growth and metastasis. Cancer Res. 2017;77(9):2339–50.
28. Glažar P, Papavasileiou P, Rajewsky N. circbase: a database for circular rnas. Rna. 2014;20(11):1666–70.
29. Chen X, Han P, Zhou T, Guo X, Song X, Li Y. circrnadb: a comprehensive database for human circular rnas with protein-coding annotations. Sci Rep. 2016;6:34985.
30. Zhang P, Meng X, Chen H, Liu Y, Xue J, Zhou Y, Chen M. Plantcircnet: a database for plant circrna–mirna–mrna regulatory networks. Database. 2017;2017:1–6.
31. Li S, Li Y, Chen B, Zhao J, Yu S, Tang Y, Zheng Q, Li Y, Wang P, He X, et al. exorbase: a database of circrna, lncrna and mrna in human blood exosomes. Nucleic Acids Res. 2017;46(D1):106–12.
32. Liu Y-C, Li J-R, Sun C-H, Andrews E, Chao R-F, Lin F-M, Weng S-L, Hsu S-D, Huang C-C, Cheng C, et al. Circnet: a database of circular rnas derived from transcriptome sequencing data. Nucleic Acids Res. 2015;44(D1):209–15.
33. Xia S, Feng J, Lei L, Hu J, Xia L, Wang J, Xiang Y, Liu L, Zhong S, Han L, et al. Comprehensive characterization of tissue-specific circular rnas in the human and mouse genomes. Brief Bioinform. 2016;18(6):984–92.
34. Bhattacharya A, Cui Y. Somamir 2.0: a database of cancer somatic mutations altering microrna–cerna interactions. Nucleic Acids Res. 2015;44(D1):1005–10.
35. Xia S, Feng J, Chen K, Ma Y, Gong J, Cai F, Jin Y, Gao Y, Xia L, Chang H, et al. Cscd: a database for cancer-specific circular rnas. Nucleic Acids Res. 2017;46(D1):925–9.
36. 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. 2018;2018:1–8.
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. 2010;26(13):1644–50.
38. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics. 2011;27(21):3036–43.
39. Yan C, Wang J, Lan W, Wu F-X, Pan Y. Sdtrls: Predicting drug-target interactions for complex diseases based on chemical substructures. Complexity. 2017;2017(Article ID 2713280):10.
40. Yan C, Wang J, Ni P, Lan W, Wu FX, Pan Y. Dnrlmf-mda:predicting microrna-disease associations based on similarities of micrornas and diseases. IEEE/ACM Trans Comput Bio Bioinforma. 2017(to be published). https://doi.org/10.1109/TCBB.2017.2776101.
41. Lan W, Li M, Zhao K, Liu J, Wu F-X, Pan Y, Wang J. Ldap: a web server for lncrna-disease association prediction. Bioinformatics. 2016;33(3):458–60.
42. Lu C, Yang M, Luo F, Wu F-X, Li M, Pan Y, Li Y, Wang J. Prediction of lncrna-disease associations based on inductive matrix completion. Bioinformatics. 2018;1:8.
43. Raymond R, Kashima H. Fast and scalable algorithms for semi-supervised link prediction on static and dynamic graphs. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Heidelberg: Springer; 2010. p. 131–47.
44. Xia Z, Wu LY, Zhou X, et al. Semi-supervised drug-protein interaction prediction from heterogeneous biological spaces. BMC Syst Biol. 2010;4:S6.
45. Chen X, Huang Y-A, You Z-H, Yan G-Y, Wang X-S. A novel approach based on katz measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics. 2016;33(5):733–9.
46. Qu Y, Zhang H, Liang C, Dong X. Katzmda: prediction of mirna-disease associations based on katz model. IEEE Access. 2018;6:3943–50.
47. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y. Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol. 2012;8(5):1002503.
48. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics. 2008;24(13):232–40.
49. Organization WH. The World Health Report 2002. http://www.who.int/whr/en. Accessed 24 Aug 2018.
50. Hackam DG, Anand SS. Emerging risk factors for atherosclerotic vascular disease: a critical review of the evidence. Jama. 2003;290(7):932–40.
51. Song C-L, Wang J-P, Xue X, Liu N, Zhang X-H, Zhao Z, Liu J-G, Zhang C-P, Piao Z-H, Liu Y, et al. Effect of circular anril on the inflammatory response of vascular endothelial cells in a rat model of coronary atherosclerosis. Cell Physiol Biochem. 2017;42(3):1202–12.
52. Li C-Y, Ma L, Yu B. Circular rna hsa_circ_0003575 regulates oxldl induced vascular endothelial cells proliferation and angiogenesis. Biomed Pharmacother. 2017;95:1514–9.
53. Devaux Y. Transcriptome of blood cells as a reservoir of cardiovascular biomarkers. Biochim Biophys Acta (BBA) - Mol Cell Res. 2017;1864(1):209–16.
54. Wooster R, Bignell G, Lancaster J, Swift S, Seal S, Mangion J, Collins N, Gregory S, Gumbs C, Micklem G, et al. Identification of the breast cancer susceptibility gene brca2. Nature. 1995;378(6559):789.
55. Müller A, Homey B, Soto H, Ge N, Catron D, Buchanan ME, McClanahan T, Murphy E, Yuan W, Wagner SN, et al. Involvement of chemokine receptors in breast cancer metastasis. nature. 2001;410(6824):50.
56. He R, Liu P, Xie X, Zhou Y, Liao Q, Xiong W, Li X, Li G, Zeng Z, Tang H. circgfra1 and gfra1 act as cernas in triple negative breast cancer by regulating mir-34a. J Exp Clin Cancer Res. 2017;36(1):145.
57. Dong Y, He D, Peng Z, Peng W, Shi W, Wang J, Li B, Zhang C, Duan C. Circular rnas in cancer: an emerging key player. J Hematol Oncol. 2017;10(1):2.
58. Cheng L, Li J, Ju P, Peng J, Wang Y. Semfunsim: a new method for measuring disease similarity by integrating semantic and gene functional association. PloS ONE. 2014;9(6):99415.
59. Lan W, Wang J, Li M, Liu J, Wu F-X, Pan Y. Predicting microrna-disease associations based on improved microrna and disease similaritiesOncol. IEEE/ACM Trans Comput Biol Bioinforma. 2016(to be published). https://doi.org/10.1109/TCBB.2016.2586190.
60. Ni P, Wang J, Zhong P, Li Y, Wu F, Pan Y. Constructing disease similarity networks based on disease module theory. IEEE/ACM Trans Comput Biol Bioinforma. 2018(to be published). https://doi.org/10.1109/TCBB.2018.2817624.
61. Liu Y, Wu M, Miao C, Zhao P, Li X.-L. Neighborhood regularized logistic matrix factorization for drug-target interaction prediction. PLoS computational biology. 2016;12(2):1004760.
62. Wang L, Li X, Zhang L, Gao Q. Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC cancer. 2017;17(1):513.
63. Luo H, Li M, Wang S, Liu Q, Li Y., Wang J. Computational drug repositioning using low-rank matrix approximation and randomized algorithms. Bioinformatics. 2018;34(11):1904–1912.