BMC Medical Genomics

RESEARCH Open Access

Prediction of circRNA-disease associations based on inductive matrix completion

Menglu Li[1†], Mengya Liu[2†], Yannan Bin[1,2] and Junfeng Xia[1,2*]

From The International Conference on Intelligent Biology and Medicine (ICIBM) 2019
Columbus, OH, USA. 9-11 June 2019

Abstract

Background: Currently, numerous studies indicate that circular RNA (circRNA) is associated with various human complex diseases. While identifying disease-related circRNAs in vivo is time- and labor-consuming, a feasible and effective computational method to predict circRNA-disease associations is worthy of more studies.

Results: Here, we present a new method called SIMCCDA (Speedup Inductive Matrix Completion for CircRNA-Disease Associations prediction) to predict circRNA-disease associations. Based on known circRNA-disease associations, circRNA sequence similarity, disease semantic similarity, and the computed Gaussian interaction profile kernel similarity, we used speedup inductive matrix completion to construct the model. The proposed SIMCCDA method obtains an area under ROC curve (AUC) of 0.8465 with leave-one-out cross validation in the dataset, which is obtained by the combination of the three databases (circRNA disease, circ2Disease and circR2Disease). Our method surpasses other state-of-art models in predicting circRNA-disease associations. Furthermore, we conducted case studies in breast cancer, stomach cancer and colorectal cancer for further performance evaluation.

Conclusion: All the results show reliable prediction ability of SIMCCDA. We anticipate that SIMCCDA could be utilized to facilitate further developments in the field and follow-up investigations by biomedical researchers.

Keywords: CircRNA-disease associations, CircRNA sequence similarity, Disease semantic similarity, Inductive matrix completion

Background

As endogenous noncoding RNA, circular RNA (circRNA) is extremely distinct from linear RNA. The largest difference is that the circRNA does not possess a terminal structure (i.e., 5′ caps and 3′ polyA tails) and is covalently closed to form a loop structure [1]. Such a loop structure facilitates the resistance of the circRNA to the degradation of RNA exonuclease and offers a stable biological effect compared with the corresponding linear structure [2, 3].

Although circRNA was discovered as early as the 1970s, it was considered 'junk' RNA [4]. Recently, circRNA has been re-recognized and has gradually gained attention. CircRNA is involved in numerous important biological functions, especially regulatory functions [5]. Accumulating evidence has clearly demonstrated that changes in circRNA plays an important role in developing various pathological conditions and exhibits a significant correlation with diseases, especially cancer. For example, the circRNA CDR1as is an inhibitor of miR-7, which is known to be involved in various diseases, such as neurodegenerative diseases, atherosclerosis and breast cancer [6]. Therefore, circRNA is thought to be a promising disease biomarker and treatment target [5]. Analysis of existing circRNA-disease associations is necessary to help predict other potential associations and help us understand the molecular mechanisms of human disease and identify biomarkers for disease diagnosis, treatment, and prevention at the circRNA level [7].

* Correspondence: jfxia@ahu.edu.cn
†Menglu Li and Mengya Liu contributed equally to this work.
[1] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, Anhui, China
[2] Institutes of Physical Science and Information Technology, Anhui University, Hefei 230601, Anhui, China

To date, an increasing number of experimentally verified or reported databases are available for the circRNA-disease associations, such as circR2Disease [7], circRNA disease [8], circ2Disease [9], and circ2Traits [10]. However, experimental methods are too expensive and time-consuming to obtain a large validated circRNA-disease association data. Developing computational methods to predict novel circRNA-disease associations has attracted considerable attention as they can effectively decrease the time and cost of biological experiments. In addition, few methods are available for predicting the circRNA-disease associations using computational methods. Lei et al. [11] developed the method of predicting circRNA-disease associations based on a path weighted model, and Fan et al. [12] proposed the KATZHCDA method using the KATZ model on heterogeneous networks. However, these methods predict potential associations using a single database, which is not enough to illustrate the stability of the model. Moreover, it remains challenging to achieve significant performance for predicting circRNA-disease associations.

In this work, we proposed a new method called SIMCCDA (Speedup Inductive Matrix Completion for CircRNA-Disease Associations prediction), which considers the prediction of circRNA-disease associations as a recommendation system problem. To the best our knowledge, we are the first to apply the recommendation system approach inductive matrix completion (IMC) [13–15] to predict circRNA-disease associations. This method has been applied for various bioinformatics problems, such as drug-target interactions [16], drug repositioning [17], lncRNA (long non coding RNA)-disease [18] and miRNA (microRNA)-disease associations [19]. We model the circRNA-disease association prediction problem as a recommendation task and solve it using speedup IMC [20]. Three databases, including circRNA disease, circ2Disease and circR2Disease, are used as our raw data in this study. We then perform data screening, generate corresponding three sub-datasets (Dataset-1, Dataset-2 and Dataset-3), and combine them into a total dataset (named TotalCircRD-1). We first calculate circRNA sequence similarity and disease semantic similarity in these four datasets. Next, these two types of similarities are combined into a Gaussian interaction profile kernel to generate new circRNA similarity and disease similarity. Primary feature vectors of the similarity matrix are extracted by principal component analysis (PCA). The final model based on IMC is built for predicting circRNA-disease associations.

Leave-one-out cross validation (LOOCV) is used to examine the performance of our method. The optimal AUC on TotalCircRD-1 is 0.8465. The AUC results on the three datasets are 0.8682 (Dataset-1), 0.8303 (Dataset-2) and 0.8509 (Dataset-3), respectively. To further evaluate the performance of the proposed method, we rank and select the top 30 predictions of each dataset to determine the number of results that existed in verified associations. We also conduct case studies in breast cancer, stomach cancer and colorectal cancer to support our predictions. Finally, we compare our method with KATZHCDA, and the prediction results indicate that our method outperforms the previous method in predicting circRNA-disease associations. In summary, the proposed SIMCCDA method has the ability to predict associations in circRNA-disease and offers a guiding significance for future biomedical clinical experiments.

Methods

Model overview

Here, we apply IMC with feature vectors to build the model called SIMCCDA. In addition, we add a linear Bregman iteration to speed up the process of calculating the final score matrix. The flowchart is presented in Fig. 1. $A_{ij} = 1$ indicates that circRNA $circ_i$ and disease d_j are associated, whereas $A_{ij} = 0$ indicates that their association is currently in an unknown state. Given a known circRNA-disease association matrix $A \in \mathbb{R}^{m \times n}$ with circRNA sequences and disease DOIDs (disease ontology identities), we obtain circRNA and disease similarity, respectively. Then, PCA is employed to extract primary feature vectors from acquired similarity. Finally, we construct the model with IMC based on the above information to predict circRNA-disease associations.

Human circRNA-disease associations data

We use three databases, including circRNA disease, circ2Disease and circR2Disease, all of which include known human circRNA-disease associations. All data were downloaded before September 2018. The initial information regarding each downloaded dataset is as follows: the first database circRNA disease contains 354 circRNA-disease associations (including 330 circRNAs and 48 diseases), the second database circ2Disease includes 273 circRNA-disease associations (including 249 circRNAs and 61 diseases) and the third database circR2Disease includes 739 associations (including 661 circRNAs and 100 diseases). The sequence information of circRNA and disease DOID matching are applied to the circBase [21] and Disease Ontology [22] (DO) databases. Based on the above data processing, we generate the final three datasets (Dataset-1, Dataset-2 and Dataset-3). These datasets are merged to obtain TotalCircRD-1 without duplicated redundancy. Table 1 lists the detailed statistics of the four datasets.

The uncompleted associations in the datasets include circRNAs without sequences or diseases without DOIDs. Given that the calculation of circRNA sequence

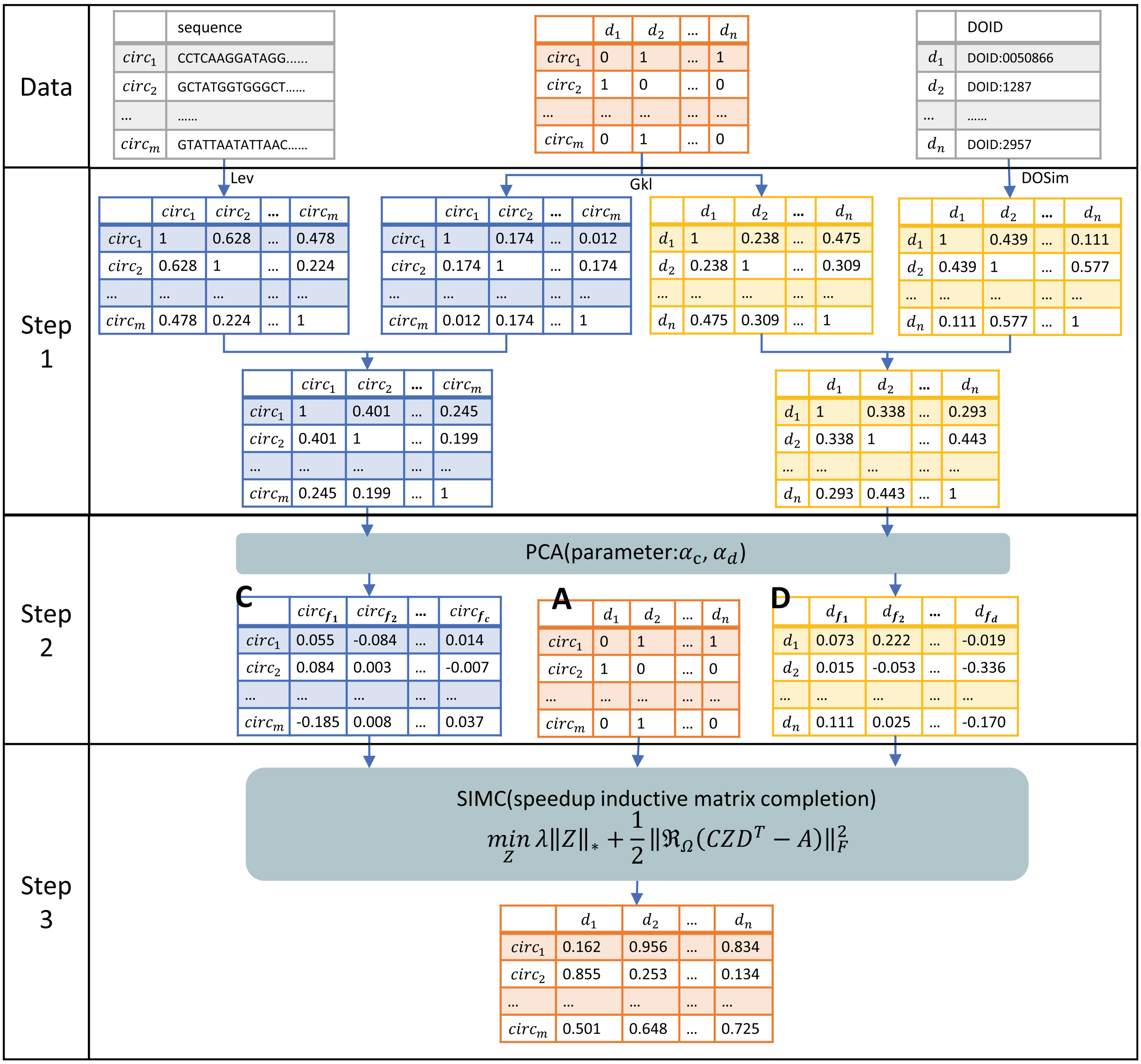
Data
Step 1
Step 2
Step 3
sequence
circ1 CCTCAAGGATAGG......
circ2 GCTATGGTGGGCT......
circm GTATTAATATTAAC......
DOID
d1 DOID:0050866
d2 DOID:1287
dn DOID:2957
Lev
Gkl
DOSim
PCA(parameter:αc, αd)
C
A
D
SIMC(speedup inductive matrix completion)
min_Z λ‖Z‖* + 1/2 ‖ℜΩ(CZD^T − A)‖²F

Fig. 1 The overall procedure of SIMCCDA. Step 1: compute circRNA similarity and disease similarity. Step 2: extract primary feature vectors. Step 3: predict the circRNA-disease association matrix with speedup IMC. Lev: Levenshtein distance, Gkl: Gaussian interaction profile kernel

similarity requires the circRNA sequence and the disease similarity requires the disease DOID information, the preceding datasets exclude the uncompleted associations. We wanted to assess whether these uncompleted associations would influence the prediction performance, so we add several uncompleted associations to form four new datasets (Dataset-4, Dataset-5, Dataset-6 and TotalCircRD-2) based on Dataset-1, Dataset-2, Dataset-3 and TotalCircRD-1, respectively (Additional file 1: Table S1).

Table 1 Details of four datasets

Dataset	Number of circRNAs	Number of diseases	Number of associations	Matrix density
Dataset-1	223	34	241	0.032
Dataset-2	215	46	240	0.024
Dataset-3	389	61	445	0.019
TotalCircRD-1	512	71	609	0.017

CircRNA sequence similarity

The sequence information of all the corresponding circRNAs in the aforementioned databases is obtained from circBase, and Levenshtein distance [23] is used to calculate the similarity between each two circRNA sequences. As a string metric for measuring the difference between two strings, the Levenshtein distance between two strings is the minimum cost of single-character edits (insertions, deletions or replacements) required to change one string into the other. In the present study, both editing costs of insertion and deletion are 1, and the replacement editing cost is 2. Formula (1) is the calculation of similarity for two circRNA sequences:

$$Sim_{lev}\left(circ_i, circ_j\right) = 1 - \frac{dist}{len(circ_i) + len\left(circ_j\right)} \qquad (1)$$

where *dist* represents the minimum editing cost of converting the circRNA $circ_i$ sequence to the circRNA $circ_j$ sequence, and $len(\cdot)$ represents the length of circRNA sequence.

Disease semantic similarity

We use DOSim [24] in DO-based DOSE (R package) to calculate the disease semantic similarity with Wang measure [25]. The detailed formula is displayed as follow:

$$Sim_{Wang}\left(d_i, d_j\right) = \frac{\sum_{t\in T_{d_i}\cap T_{d_j}}\left(S_{d_i}(t) + S_{d_j}(t)\right)}{\sum_{t\in T_{d_i}} S_{d_i}(t) + \sum_{t\in T_{d_j}} S_{d_j}(t)} \qquad (2)$$

For a given disease d_i, T_{d_i} is the ancestor term set of term d_i (including d_i itself). $S_{d_i}(t)$ is defined as the contribution score of disease t ($t\in T_{d_i}$) to disease d_i. It can be expressed by the following formula:

$$\begin{cases} S_{d_i}(d_i)=1 \\ S_{d_i}(t)= \max\left\{w_e\times S_{d_i}\left(t'\right)|t'\in childrenof(t)\right\}\ if\ \ t\neq d_i \end{cases} \qquad (3)$$

Here, w_e is the semantic contribution factor of edge e, where e belongs to the set of edges connecting d_i and its ancestor T_{d_i}. In DOSim, we set $w_e = 0.7$.

Gaussian interaction profile kernel similarity for circRNA and disease

By considering the assumption that similar circRNAs tend to be bound with similar diseases, Gaussian interaction profile kernel similarity is computed based on the known circRNA-disease association datasets. Inspired by van Laarhoven et al. [26], we calculate the circRNA and disease similarity using the Gaussian interaction profile kernel on four datasets. Equations (4) and (5) determine the similarity between $circ_i$ and $circ_j$, where m is circRNA number, $IP(circ_i)$ is the associated disease set corresponding to the $circ_i$, and γ_c is the regulation parameter of kernel bandwidth.

$$Gkl\left(circ_i, circ_j\right) = \exp\left(-\gamma_c\left\|\mathrm{IP}(circ_i) - IP\left(circ_j\right)\right\|^2\right) \qquad (4)$$

$$\gamma_c = \frac{1}{\frac{1}{m}\sum_{i=1}^{m}\left\|IP(circ_i)\right\|^2} \qquad (5)$$

The Gaussian interaction profile kernel similarity of diseases d_i and d_j is similar to the defined equations (6) and (7), where n is the number of diseases:

$$Gkl\left(d_i, d_j\right) = \exp\left(-\gamma_d\left\|IP(d_i) - IP\left(d_j\right)\right\|^2\right) \qquad (6)$$

$$\gamma_d = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\left\|IP(d_i)\right\|^2} \qquad (7)$$

Integrated similarity for circRNA and disease

Based on the previously defined circRNA sequence similarity, disease semantic similarity and Gaussian interaction profile kernel similarities, the integrated circRNA similarity matrix *CS* and the disease similarity matrix *DS* are calculated using the following equations (8) and (9):

$$CS\left(circ_i, circ_j\right) = \frac{Sim_{lev}\left(circ_i, circ_j\right) + Gkl\left(circ_i, circ_j\right)}{2} \qquad (8)$$

$$DS\left(d_i, d_j\right) = \frac{Sim_{Wang}\left(d_i, d_j\right) + \mathrm{Gkl}\left(d_i, d_j\right)}{2} \qquad (9)$$

Extract primary feature vectors

To remove the similarity redundancy, we use principal component analysis (PCA) to extract the primary feature vectors from integrated similarity, *CS* and *DS*. In this method, based on the dominating energy strategy [27], we use singular value decomposition (SVD) to perform PCA and formulas (10) and (11) to obtain the primary feature vectors of circRNA and disease similarity.

$$\arg\min_{f_c}\left\{\frac{\sum_{i=1}^{f_c}(S_c)_{ii}}{\sum_{j=1}^{m}(S_c)_{jj}} \geq \alpha_c\right\} \qquad (10)$$

$$\arg\min_{f_d}\left\{\frac{\sum_{i=1}^{f_d}(S_d)_{ii}}{\sum_{j=1}^{n}(S_d)_{jj}} \geq \alpha_d\right\} \qquad (11)$$

In the above formulas, S_c and S_d are the singular values of circRNA and the disease similarity matrix, respectively. α_c and α_d are adjusted parameters to obtain optimal results. In this study, Dataset-2, Dataset-3 and TotalCircRD-1 share the parameters α_c =0.6 and α_d = 0.9, whereas the parameters of Dataset-1 are α_c =0.7 and α_d =0.9. Detailed adjustment work of α_c and α_d is discussed in the Results section.

Model construction

In this study, we formulate circRNA-disease association prediction as a recommendation system problem. Generally, a recommendation system is an information filtering system that seeks to predict the user's preference of a certain item based on partial known preference information. We here use the recommendation system method IMC [15] to identify circRNAs for a disease that is dependent on validated circRNA-disease associations.

Observing the matrix density of the last column in Table 1, we find that the association matrix is very sparse. As we know, there are a small amount of experimental data of associations due to the structural complexity of circRNAs and ignored biological functions. The available data scale is in the primary stage. As a result, we cover the unknown associations of circRNAs and diseases through IMC to enhance the quality of our data. The advantage is that IMC can solve matrix completion problems using a relatively small set of known information. The detailed process of IMC is described below. First, based on the assumption that the human circRNA-disease association matrix is A, the row vectors in A lie in the subspace spanned by the column vectors in D (disease feature vectors), and the column vectors in A lie in the subspace spanned by the column vectors in C (circRNA feature vectors). The problem can be defined as:

$$\min_{Z\in\mathfrak{R}^{f_c\times f_d}} \lambda\|Z\|_* + \frac{1}{2}\left\|\mathfrak{R}_\Omega\left(CZD^T-A\right)\right\|_F^2 \tag{12}$$

where Z is the objective matrix to complete A, CZD^T is the final scoring matrix based on the association matrix and the similarity matrix, Ω represents known association sets, $\|\cdot\|_*$ is the nuclear norm defined as the sum of the singular values, λ is the regularization parameter controlling the extent of the nuclear norm (here we set λ to 1), and $\|\cdot\|_F$ is the Frobenius norm of the matrix.

Representing $f(Z)$ as $\frac{1}{2}\|\mathfrak{R}_\Omega(CZD^T-A)\|_F^2$, the formula (12) can be expressed as:

$$\min_{Z\in\mathfrak{R}^{f_c\times f_d}} \lambda\|Z\|_* + f(Z) \tag{13}$$

For any given $Y\in R^{f_c\times f_d}$, the following quadratic approximation of $f(Z)$ at Y can be considered as:

$$\begin{aligned} f(Z) \approx f_\tau(Z,Y) &= f(Y) + \langle\nabla f(Y), Z-Y\rangle \\ &\quad + \frac{\tau}{2}\|Z-Y\|_F^2 + P(Z) \\ &= \frac{\tau}{2}\left\|Z-\left(Y-\frac{1}{\tau}\nabla f(Y)\right)\right\|_F^2 \\ &\quad + f(Y) - \frac{1}{2\tau}\|\nabla f(Y)_F^2\| \end{aligned} \tag{14}$$

where $\nabla f(Y) = C^T\mathfrak{R}_\Omega(CYD^T-A)D$ is the gradient of $f(Z)$ at Y, $\langle\cdot\rangle$ represents matrix inner product, and τ is a proximal parameter for estimating the second-order gradient $\nabla^2 f(Y)$. Accordingly, the above formula (13) calculates the minimum model, which can be converted into the following formula:

$$\min_{Z\in\mathfrak{R}^{f_c\times f_d}} \lambda\|Z\|_* + \frac{\tau}{2}\left\|Z-\left(Y-\frac{1}{\tau}\nabla f(Y)\right)\right\|_F^2 \tag{15}$$

Then, we use the accelerated proximal gradient singular value thresholding algorithm [28] with iterate h times to obtain Z [29].

Algorithm 1 accelerated proximal gradient singular value thresholding algorithm

```
1.  Initialize threshold $\delta$, $t_1 = t_2 \in (0,1]$, $Z_1 = Z_2$, $\tau, \omega > 1$, $h$, $k = 2$
2.  do
3.      $Y_k = Z_k + ((t_{k-1}-1)/t_k)(Z_k - Z_{k-1})$
4.      $Z_{k+1} = D_{\frac{\lambda}{\tau}}\left(Y_k - \frac{1}{\tau}C^T(CY_kD^T - A)D\right) = \sum_i^{\sigma_i \ge \frac{\lambda}{\tau}}(\sigma_i - \frac{\lambda}{\tau})u_i v_i^T$
5.      % obtained by recasting the linearize Breg man iterations as a special form of Uzawa's algorithm [29]
6.      while $f(Z_{k+1}) \ge f(Z_{k+1}, Y_k)$
7.          $\tau = \tau \times \omega$
8.          $Z_{k+1} = D_{\frac{\lambda}{\tau}}\left(Y_k - \frac{1}{\tau}C^T(CY_kD^T - A)D\right) = \sum_i^{\sigma_i \ge \frac{\lambda}{\tau}}(\sigma_i - \frac{\lambda}{\tau})u_i v_i^T$
9.      end while
10.     $t_{k+1} = (1 + \sqrt{1 + 4 * t_k^2})/2$
11.     $k = k + 1$
12. while $|f(Z_{k+1}) - f(Z_k)| \ge \delta$ and $k < h$
13. Output: $Z_{k+1}$
```

Note: For above step 4, $D_{\frac{\lambda}{\tau}}(\cdot)$ denotes the matrix shrinkage operator based on SVD on the operand matrix. $\frac{\lambda}{\tau}$ is the threshold. Here, u_i and v_i are the left and right singular vector of $Y_k - \frac{1}{\tau}C^T(CY_kD^T - A)D$ corresponding to σ_i, respectively.

In order to see the relationship between the objective function value and the number of iterations, we divide the circRNAs into several categories according to their chromosomal location and then select randomly one from each class to view the trend of the curve.

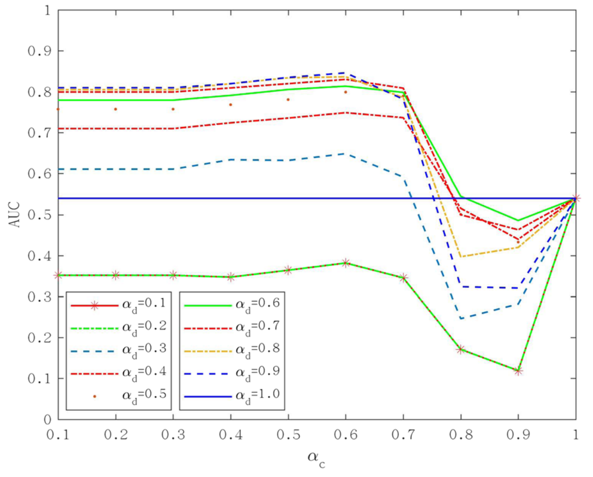

Fig. 2 Adjusted parameters α_c and α_d with their impact on TotalCircRD-1 dataset

Additional file 1: Figure S1 shows that the value of the objective function decreases as the number of iterations increases. When the gap of objective function values between two iterations is particularly small, i.e. $1-\frac{objective\ value_k}{objective\ value_{k-1}} < 10^{-5}$, the iterative process will end.

Results

LOOCV

To assess the predictive accuracy of SIMCCDA, we performed the following method using the leave-one-out cross validation (LOOCV) framework on the known circRNA-disease associations. The reason why LOOCV is used in this study is that the current common practice in this field (prediction of lncRNA/miRNA/circRNA-disease associations) [30–32] is to use LOOCV to measure the performance of the model. For a disease d_i, each known circRNA association corresponding to the disease was left as a test sample. Other known associations were used as training samples, and an initial non-association was regarded as a candidate sample. In the candidate samples and test sample set, the test sample was deemed

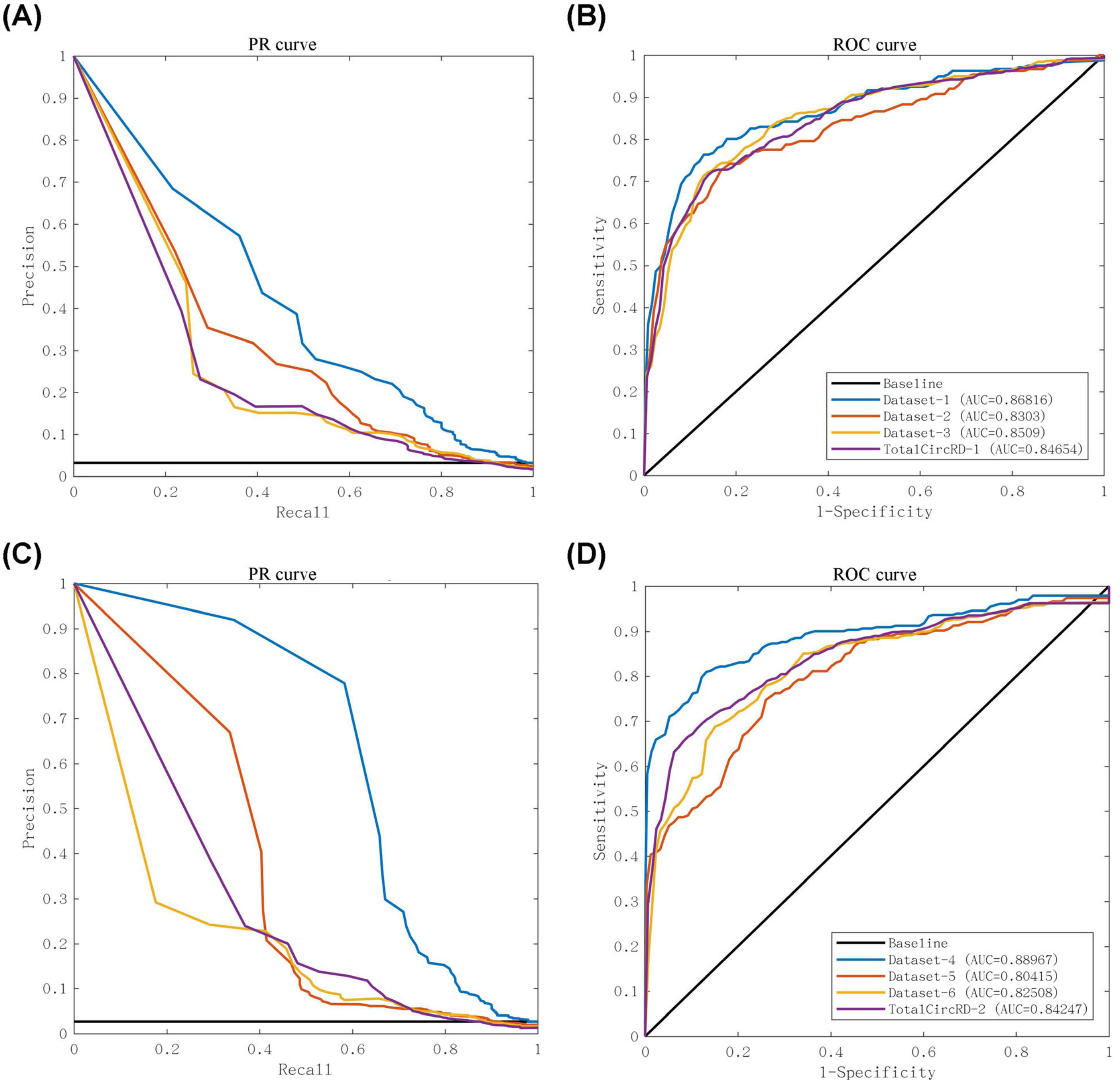

Fig. 3 ROC curve and PR curve using LOOCV of eight datasets under optimal parameters. **a** PR curve in four datasets (Dataset-1, Dataset-2, Dataset-3 and TotalCircRD-1). **b** ROC curve in four datasets (Dataset-1, Dataset-2, Dataset-3 and TotalCircRD-1). **c** PR curve in four datasets (Dataset-4, Dataset-5, Dataset-6 and TotalCircRD-2). **d** ROC curve in four datasets (Dataset-4, Dataset-5, Dataset-6 and TotalCircRD-2)

as a positive sample, and the others were negative samples. After running the model, the probabilities of associations between candidate samples and disease d_i were calculated. We took the highest values as the final score of the candidate sample among probabilities. Finally, we calculated the sensitivity and specificity as follows:

$$sensitivity = \frac{TP}{TP + FN} \tag{16}$$

$$specificity = \frac{TN}{TN + FP} \tag{17}$$

where TP indicate true positives, FP is false positives, TN refer to true negatives, and FN represent false negatives.

A Receiver Operating Characteristics (ROC) curve is drawn based on the LOOCV result. The X-axis of the ROC graph is the 1-specificity, and the Y-axis is the sensitivity. From the ROC curve, the Area Under ROC Curve (AUC) can be calculated as an evaluation measure for the model.

The effect of adjusting parameters on the prediction result

In the PCA section of the Methods, two parameters α_c and α_d were included, which represent the percentage of singular values of circRNA and disease similarity matrix, respectively. We tried to take values between 0.1 and 1 for α_c and α_d, and the step size was 0.1. The results of the parameterization of TotalCircRD-1 are presented in Fig. 2, and results for Dataset-1, Dataset-2 and Dataset-3 are presented in Additional file 1: Figures S2-S4. As noted in Fig. 2, as α_c increases, AUC is initially stable and the generally declines. The results are consistent when α_d =0.1 or α_d =0.2. As α_d increases, the AUC gradually increases, but the growth rate is slow. The optimal parameters of the three datasets of Dataset-2, Dataset-3, TotalCircRD-1 are all α_c =0.6 and α_d =0.9, whereas the optimal parameters for Dataset-1 are α_c =0.7 and α_d = 0.9. LOOCV-based AUC results for four datasets with optimal parameters are shown in Fig. 3b. The results of our model on the four datasets are at a solid level, and the gap between the maximum and minimum values is 3% in four datasets, which reveals that our model exhibits better robustness. Figure 3a shows the PR (Precision-Recall) curves on the four datasets, respectively, which have the same trend as the ROC curve. Figure 4 presents the number of experimental validated associations predicted by our model from the top 30 predicted associations from our four datasets. Additional file 1: Table S2 shows the predicted results of the top 10, 30, 50 and 100. It can be observed that whether it is top 10, top 30, top 50, or top 100, the ultimate trends are similar. For the sake of convenience, we only show the results of top 30 in this work. Based on the above optimal parameters, we predicted 30 known circRNA-disease associations from Dataset-2, Dataset-3 and TotalCircRD-1, and 26 known associations from the Dataset-1. This shows that our results are optimal under these parameters, and four unknown associations in Dataset-1 may be potential associations based on subsequent analysis.

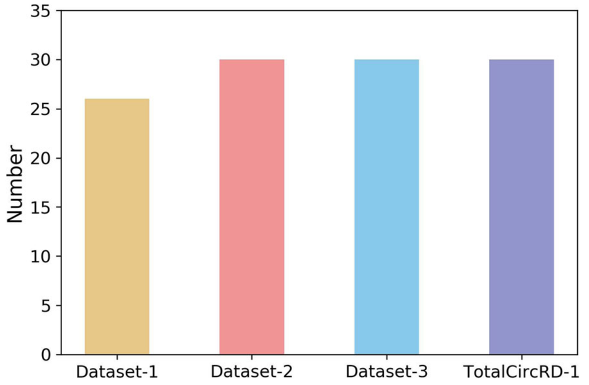

Fig. 4 The number of associations validated by our model for the top 30 on four datasets (Dataset-1, Dataset-2, Dataset-3, and TotalCircRD-1)

In addition, we added weights to each part of the integration similarity to see how the performance could be impacted. We added weights (range from 0 to 1) to $Sim_{lev}(circ_i, circ_j)$ and $Gkl(circ_i, circ_j)$ in equation (8) and (9), respectively. For different weights circRNAs and diseases similarity, the final results were obtained by combining the two pairs. The Additional file 1: Figure S5 shows that the combinations of different similarity weights have similar results for the models obtained on different datasets. So, in the end, our model used equation (8) and (9) to respectively calculate the circRNAs similarity and diseases similarity.

The effect of uncompleted associations

The α_c and α_d were adjusted in the same manner as described above, and the optimal parameters were selected to calculate the AUC in Dataset-4, Dataset-5, Dataset-6 and TotalCircRD-2 datasets, as presented in Fig. 3c and d. The AUC scores of new-added datasets (Fig. 3d) are slightly reduced compared with the initial datasets (Fig. 3b). Given that most of the newly added circRNA only involved in one disease, thus making the final association matrix sparser than previous one. For example, circ-BANP is only associated with colorectal cancer and is not associated with other diseases. Increasing association data are noted between circ-BANP and colorectal cancer, and the unknown associations of circ-BANP with other diseases also increase, as observed from the matrix density columns of Additional file 1: Table S1. In summary, uncompleted associations exhibit a minimal effect on the results and only slightly reduce the performance of predictions.

The above results show that the sparseness of the data set has little effect on the prediction results. But if the correlation matrix is too sparse, it will still affect the final prediction results. So our method has a premise that the association matrix cannot be too sparse. We conducted the following experiment to explore how the varying sparsity of datasets affect the overall performance. Since the final result has a certain relationship with the dataset, we performed sparsity processing on each dataset (0.002 was the step size, and the sparsity was reduced by 0.002 each time), respectively. The Additional file 1: Figure S6 shows that the result is not much changed when the sparsity of Dataset-1 is 0.015. But when the sparsity is 0.013, the performance starts to drop significantly. Similarly, for the other three datasets (Dataset-2, Dataset-3, TotalCircRD-1), the performance starts to drop significantly when the sparsity is 0.013, 0.009, and 0.007, respectively.

Compared with the other method

Two methods are currently available for predicting circRNA-disease associations: PWCDA [11] and KATZHCDA [12]. Given that the PWCDA method

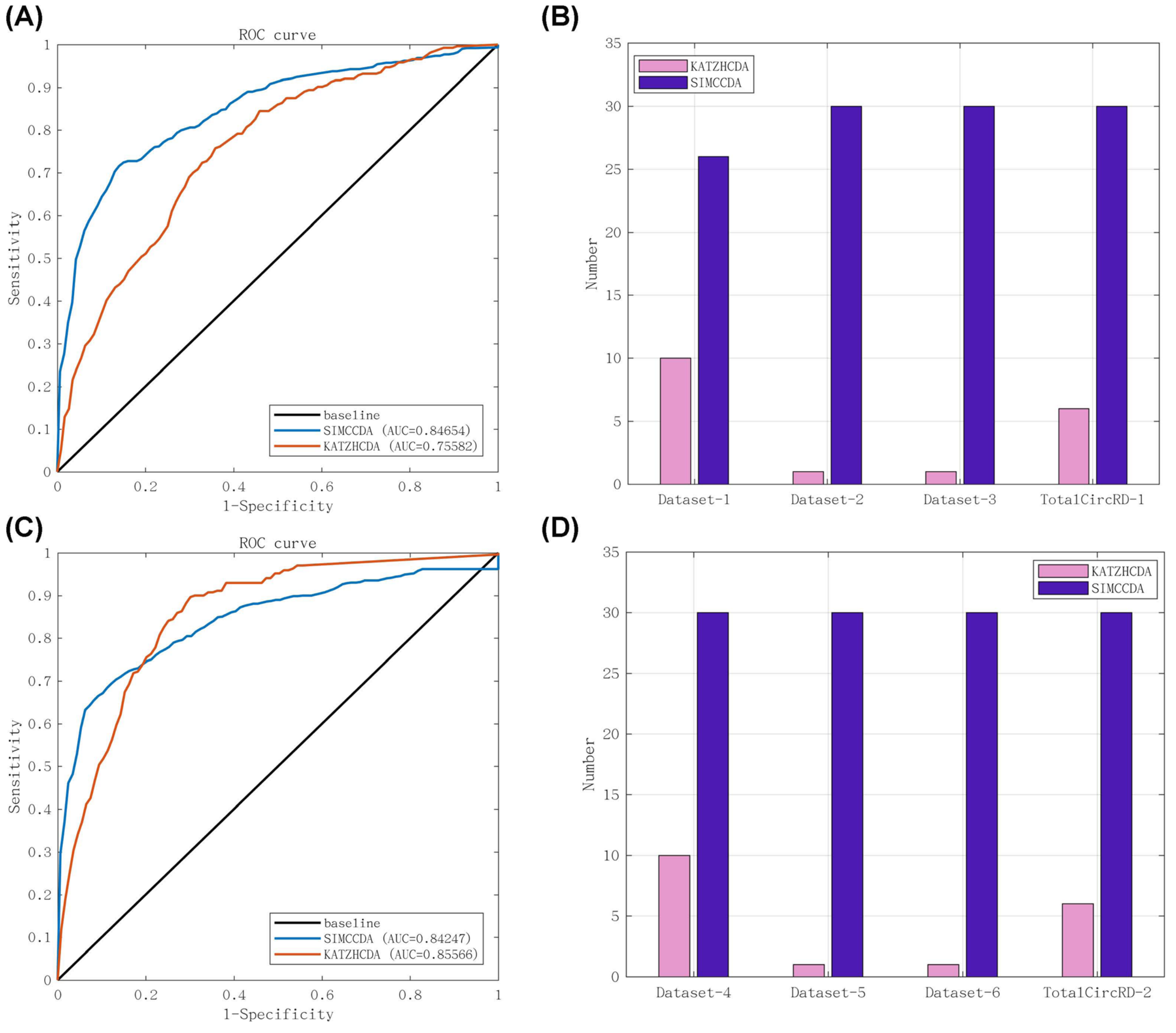

Fig. 5 Comparison of SIMCCDA with KATZHCDA on the TotalCircRD-1 and Total CircRD-2 dataset. **a** Performance of methods in terms of ROC curve using LOOCV in TotalCircRD-1 dataset. **b** The number of experimental validations of the top 30 predicted circRNA-disease associations from four datasets (Dataset-1, Dataset-2, Dataset-3 and TotalCircRD-1). **c** Performance of methods in terms of ROC curve using LOOCV in TotalCircRD-2 dataset. **d** The number of experimental validations of the top 30 predicted circRNA-disease associations from four datasets (Dataset-4, Dataset-5, Dataset-6 and TotalCircRD-2)

needs to set the circRNA similarity and disease similarity < 0.5 part to 0 and most of the similarities on our datasets are less than 0.5, we only compared our method with KATZHCDA. KATZHCDA is a computational model of KATZ measures and constructs heterogeneous networks by employing the circRNA expression profiles, disease phenotype similarity and Gaussian interaction profile kernel similarity. Here, we used the same eight datasets in KATZHCDA model and obtain predicted results. The results of six datasets (Dataset-1, Dataset-2, Dataset-3, Dataset-4, Dataset-5 and Dataset-6) are presented in Additional file 1: Figure S7, and TotalCircRD-1 and TotalCircRD-2 results are presented in Fig. 5a and c. As shown in Additional file 1: Figure S7, both the PR curve and the ROC curve indicate that our model performance is superior to KATZHCDA. The AUC scores of four datasets (Dataset-1, Dataset-2, Dataset-3 and TotalCircRD-1) are 0.7604, 0.7458, 0.7442 and 0.7558, respectively. According to the comparison of two methods, our model obtains an average AUC of 0.8490, which is 9% higher than KATZHCDA. The resulting top 30 predicted associations are also analyzed, demonstrating that our predicted top 30 results are superior to KATZHCDA (Fig. 5b, d).

In addition, we compared our method with KATZHCDA by using Dataset-1 as the training set and Dataset-2, Dataset-3 as the test set. As can be seen from Additional file 1: Figure S8, our performance is slightly better than KATZHCDA. Specifically, the early stage of KATZHCDA prediction effect is better than ours, but its accuracy is reduced in the prediction of later stages. A comprehensive look at the above two results, our model is superior to KATZHCDA on the whole.

Case study

Analysis of predicted circRNA-disease associations with experimental evidence from the TotalCircRD-1 dataset

To further measure the performance of SIMCCDA, case studies of three diseases, including breast cancer, stomach cancer and colorectal cancer, from the TotalCircRD-1 dataset were analyzed in detail. The top 30 predicted disease-related circRNAs by SIMCCDA and supporting evidence from PubMed are presented in Tables 2, 3 and 4.

Breast cancer is the most common cancer and remains the leading cause of cancer death among women worldwide [33]. Among top 30 predicted candidate circRNA for breast cancer, 29 are associated with breast cancer in related studies (Table 2). For instance, hsa_circ_0001875 (top 1) is upregulated in breast cancer tissues compared with the normal breast tissue [34]. In addition, circRNA hsa_circ_0006054 (top 2) expression is significantly downregulated in breast cancer tissues compared with non-breast cancer tissues [34].

Gastric cancer is the second disease to lead cancer-related mortality and the fourth most frequent cancer globally [35]. Using the SIMCCDA method, we successfully predicted 30 of top 30 candidate circRNAs for gastric cancer (Table 3). Among them, CircRNA hsa_circ_0084606 (top 1) is one of the top 10 upregulated circRNAs in stomach cancer tissues [36], whereas hsa_circ_0000140 (top 2), a typical circular RNA, is

Table 2 Top 30 candidate circRNAs for breast cancer

Rank	circRNA	Evidence (PMID)	Rank	circRNA	Evidence (PMID)
1	hsa_circ_0001875	28484086	16	hsa_circ_0000911	28744405
2	hsa_circ_0006054	28484086	17	hsa_circ_0092276	28803498
3	hsa_circ_0000098	28744405	18	hsa_circ_0008945	28744405
4	hsa_circ_0107327	29221160	19	hsa_circ_0003838	28803498
5	hsa_circ_0001785	29045858	20	hsa_circ_0004619	28484086
6	hsa_circ_0103038	29221160	21	hsa_circ_0033144	29221160
7	hsa_circ_0002874	28803498	22	hsa_circ_0001283	28744405
8	hsa_circ_0002220	29221160	23	hsa_circ_0057129	29221160
9	hsa_circ_0006528	28803498	24	hsa_circ_0001824	28484086
10	hsa_circ_0008717	28744405	25	hsa_circ_0085495	28803498
11	hsa_circ_0000893	28744405	26	hsa_circ_0000732	28744405
12	hsa_circ_0068033	29045858	27	hsa_circ_0086241	28803498
13	hsa_circ_0011946	29593432	28	hsa_circ_0003221	unconfirmed[a]
14	hsa_circ_0001982	28933584	29	hsa_circ_0018293	28744405
15	hsa_circ_0001667	28803498	30	hsa_circ_0093859	29593432

[a] without the evidence reported in literatures

Table 3 Top 30 candidate circRNAs for stomach cancer

Rank	circRNA	Evidence (PMID)	Rank	circRNA	Evidence (PMID)
1	hsa_circ_0084606	28544609	16	hsa_circ_0076304	28831102
2	hsa_circ_0000140	25689795	17	hsa_circ_0057104	28831102
3	hsa_circ_0008383	28761361, 28206972	18	hsa_circ_0138960	28980874
4	hsa_circ_0074362	28544609, 29240459	19	hsa_circ_0013048	28657541, 28206972
5	hsa_circ_0003159	28618205	20	hsa_circ_0003789	28544609
6	hsa_circ_0006022	28639908	21	hsa_circ_0035445	28544609
7	has_circ_0031027	28206972	22	hsa_circ_0058766	28831102
8	hsa_circ_0050547	28544609	23	hsa_circ_0001895	28443463
9	hsa_circ_0001546	28544609	24	hsa_circ_0005927	28737829
10	hsa_circ_0063809	28544609	25	hsa_circ_0076305	28831102
11	hsa_circ_0084720	28831102	26	hsa_circ_0006633	28656881
12	hsa_circ_0032821	28737829	27	hsa_circ_0000154	28544609
13	hsa_circ_0001539	28184940	28	hsa_circ_0006470	28544609
14	hsa_circ_0003707	28639908	29	hsa_circ_0001017	29098316
15	hsa_circ_0006127	28974900	30	hsa_circ_0003222	28893265

significantly increased in stomach cancer tissues compared with paired adjacent non-tumorous tissues [37].

Colorectal cancer is the third most common cancer worldwide with 1.36 million people diagnosed in 2012 [38]. The inferred results cover 23 experimental verified associations out of the top 30 ranked predictions (Table 4). The evidence in the literature reveals that circRNA hsa_circ_0000523 exhibits significantly reduced expression in cancer compared with normal colorectal tissues. In colorectal cancer cells, the well-validated circRNA hsa_circ_0000504 is upregulated [39].

Analysis of predicted circRNA-disease associations without experimental evidence from four datasets

Given that the top 30 well-validated associations were successfully investigated by our method using Dataset-2, Dataset-3 and TotalCircRD-1 dataset, here we concentrated on four new predicted potential circRNA-disease

Table 4 Top 30 candidate circRNAs for colorectal cancer

Rank	circRNA	Evidence (PMID)	Rank	circRNA	Evidence (PMID)
1	hsa_circ_0000523	25624062	16	hsa_circ_0006174	28656150
2	hsa_circ_0000504	28656150	17	hsa_circ_0001724	29207676
3	hsa_circ_0002138	25624062	18	hsa_circ_0001451	26884878
4	hsa_circ_0000567	29333615	19	hsa_circ_0074806	28656150
5	hsa_circ_0007006	28656150	20	hsa_circ_0003221	unconfirmed[a]
6	hsa_circ_0024169	25624062	21	hsa_circ_0001577	unconfirmed[a]
7	hsa_circ_0082333	26138677	22	hsa_circ_0008509	28656150
8	hsa_circ_0087862	28656150	23	hsa_circ_0022080	28656150
9	hsa_circ_0005949	28656150	24	hsa_circ_0002024	unconfirmed[a]
10	hsa_circ_0007031	28656150	25	hsa_circ_0000172	unconfirmed[a]
11	hsa_circ_0008494	28656150	26	hsa_circ_0000677	27058418
12	hsa_circ_0000069	28003761	27	hsa_circ_0002768	unconfirmed[a]
13	hsa_circ_0048232	28656150	28	hsa_circ_0091017	unconfirmed[a]
14	hsa_circ_0003098	28103507	29	hsa_circ_0002702	27058418
15	hsa_circ_0074930	28656150	30	hsa_circ_0128454	unconfirmed[a]

[a] without the evidence reported in literatures

associations from Dataset-1 (as shown in Fig. 4). We employed circRNA-miRNA and miRNA-disease associations to construct corresponding circRNA-miRNA-mRNA networks for the four new circRNA-disease associations.

We used the hsa_circ_0070963-stomach cancer association as an example for a detailed exposition. First, possible miRNA targets of hsa_circ_0070963 were predicted with the miRNA Target Sites tool of CircInteractome [40]. Their target genes with experimental verification were screened out from miRTarBase [41], and then, hsa_circ_0070963-miRNA-disease regulatory network was constructed using Cytoscape [42]. Finally, the corresponding experimentally verified miRNA-stomach cancer associations were obtained from HMDD [43] and added to the above network. As noted from the result (Fig. 6), hsa_circ_0070963 may be targeted by four miRNAs, including has-miR-223, has-miR-421, has-miR-610 and has-miR-526b. CircRNA can act as competing endogenous RNAs (ceRNAs) (also termed miRNA sponges) to buffer the target genes expression (i.e., mRNA) of miRNAs [36, 37], and miRNA has-miR-223 is linked the most number of targets. Thus, we hypothesize that hsa_circ_0070963 may function as a hsa-miR-223 sponge to interact with stomach carcinoma.

Other three predicted new associations (hsa_circ_0061893, hsa_circ_0071410, and hsa_circ_0054345 in stomach cancer) exhibit similar scenarios, which are presented in Additional file 1: Figures S9-S11.

Conclusions

Increasing evidence demonstrates that circRNA plays an important role in the development of various diseases. Understanding the underlying mechanisms of circRNA in disease is becoming an urgent problem worldwide. To date, the number of experimentally validated circRNA-disease associations is small, and few computational methods for predicting circRNA-disease associations are

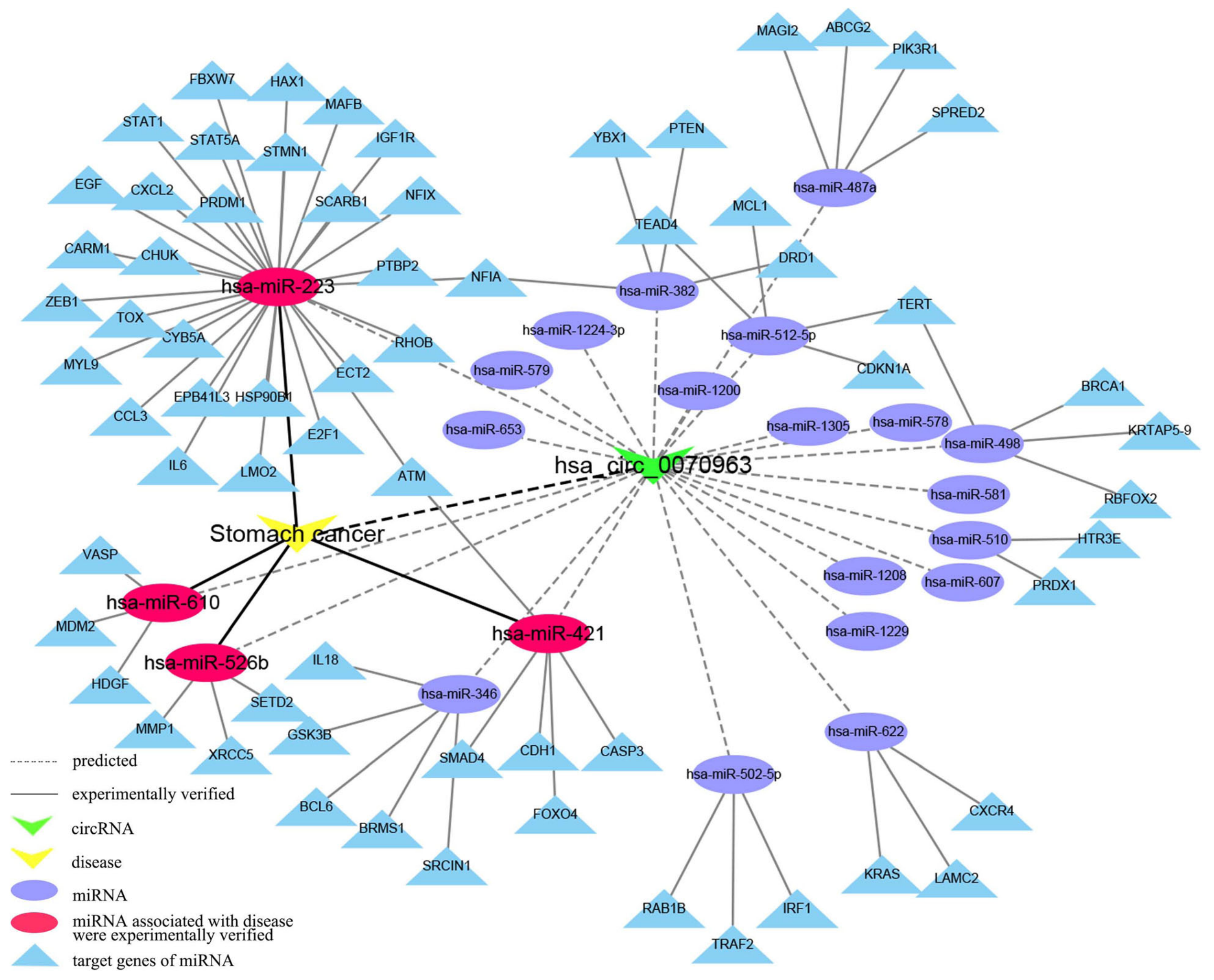

Fig. 6 hsa_circ_0070963-miRNA-mRNA regulatory network in stomach cancer

available. In this paper, we proposed a method called SIMCCDA for predicting circRNA-disease associations based on known circRNA-disease associations. Integrating data regarding circRNA similarity and disease similarity, we employed IMC to construct the model. LOOCV was applied to assess the accuracy of the SIMCCDA. We then compared our method with KATZHCDA. Further case studies were also performed on breast cancer, stomach cancer and colorectal cancer. Based on the prediction results, SIMCCDA performs well in cross validations on the four datasets we used. Simultaneously, the compared results indicate that our method can identify more associations between circRNA and disease.

The prominent performances of SIMCCDA may have been facilitated by the following factors. First, SIMCCDA was constructed based on the integrated circRNA and disease similarities, which can make a full use of various similarity data to characterize potential circRNA-disease associations. Second, SIMCCDA transformed circRNA-disease associations into a recommendation system problem and applied the IMC algorithm of the recommendation system to predict potential circRNA-disease associations. A decisive advantage of IMC is that it can supplement the missing values in the circRNA-disease association matrix to improve the performance. Third, the datasets used in this study were derived from various validated databases. Observing the results obtained on the four datasets, we found that the prediction ability of our model was better than the previous method.

However, our model also has some limitations. First, although we introduce the sequence similarity of circRNA and the semantic similarity of disease, the calculation of Gaussian interaction profile kernel similarity relies heavily on known circRNA-disease associations, thus causing inevitable bias towards well-investigated circRNAs and diseases. Second, SIMCCDA could not be applied to unknown circRNA and diseases. In our future work, we will extend our method to solve these limitations.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12920-020-0679-0.

Additional file 1. Supplementary file to this work (**Table S1-S2** and **Figures S1-S11**).

Abbreviations

AUC: Area under ROC curve; circRNA: circular RNA; DO: Disease Ontology; DOIDs: Disease ontology identities; IMC: Inductive matrix completion; lncRNA: long non coding RNA; LOOCV: Leave-one-out cross validation; miRNA: microRNA; PCA: Principal component analysis; PR: Precision-Recall; ROC: Receiver Operating Characteristics; SIMCCDA: Speedup Inductive Matrix Completion for CircRNA-Disease Associations prediction; SVD: Singular value decomposition

Acknowledgements

The authors thank the members of our laboratory for their valuable contributions to SIMCCDA.

About this supplement

This article has been published as part of *BMC Medical Genomics Volume 13 Supplement 5, 2020: The International Conference on Intelligent Biology and Medicine (ICIBM) 2019: Computational methods and application in medical genomics (part 1)*. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-13-supplement-5 .

Authors' contributions

ML (Li) implemented the prediction system, performed the analysis, and drafted the manuscript. ML (Liu) collected the datasets, performed the analysis, and drafted the manuscript. YB participated in the design of study and drafted the manuscript. JX designed the study, performed the analysis, and drafted the manuscript. All authors read and approved the final manuscript.

Funding

Publication costs are funded by the National Natural Science Foundation of China (61672037, 21601001, U19A2064, 11835014), the Anhui Provincial Outstanding Young Talent Support Plan (gxyqZD2017005), the Young Wanjiang Scholar Program of Anhui Province, and the Key Project of Anhui Provincial Education Department (KJ2017ZD01), the China Postdoctoral Science Foundation Grant (2018 M630699) and the Anhui Provincial Postdoctoral Science Foundation Grant (2017B325).

Availability of data and materials

The data pertaining to the present study has been included in table and/or figure form in the present manuscript. And all datasets and computational code underlying this study are available in an online archive https://github.com/bioinformaticsAHU/SIMCCDA.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Published: 3 April 2020

References

1. 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.
2. Bahn JH, Zhang Q, Li F, Chan TM, Lin X, Kim Y, Wong DT, Xiao X. The landscape of microRNA, Piwi-interacting RNA, and circular RNA in human saliva. Clin Chem. 2015;61(1):221–30.
3. 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):333–8.
4. Zhang Z, Yang T, Xiao J. Circular RNAs: promising biomarkers for human diseases. EBioMedicine. 2018;34:267–74.
5. Fang Y. Circular RNAs as novel biomarkers with regulatory potency in human diseases. Future Sci OA. 2018;4(07):FSO314.
6. Peng L, Yuan XQ, Li GC. The emerging landscape of circular RNA ciRS-7 in cancer (review). Oncol Rep. 2015;33(6):2669–74.
7. Fan C, Lei X, Fang Z, Jiang Q, Wu FX. CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. Database. 2018;2018:bay044.
8. Zhao Z, Wang K, Wu F, Wang W, Zhang K, Hu H, Liu Y, Jiang T: circRNA disease: a manually curated database of experimentally supported circRNA-disease associations. Cell Death Dis 2018, 9(5):475.
9. Yao D, Zhang L, Zheng M, Sun X, Lu Y, Liu P. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. Sci Rep. 2018;8(1):11018.

10. 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.
11. Lei X, Fang Z, Chen L, Wu FX. PWCDA: Path Weighted Method for Predicting circRNA-Disease Associations. Int J Mol Sci. 2018;19(11):E3410.
12. Fan C, Lei X, Wu F-X. Prediction of CircRNA-disease associations using KATZ model based on heterogeneous networks. Int J Biol Sci. 2018;14(14):1950–9.
13. Shin D, Cetintas S, Lee K-C, Dhillon IS. Tumblr Blog Recommendation with Boosted Inductive Matrix Completion; 2015. p. 203–12.
14. Koren Y, Bell R, Volinsky C. Matrix factorization techniques for recommender systems. Computer. 2009;8:30–7.
15. Jain P, Dhillon IS. Provable inductive matrix completion. arXiv preprint arXiv: 13060626; 2013.
16. Zheng X, Ding H, Mamitsuka H, Zhu S. Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. Chicago: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2013. p. 1025–33.
17. 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):1904–12.
18. Lu C, Yang M, Luo F, Wu FX, Li M, Pan Y, Li Y, Wang J. Prediction of lncRNA-disease associations based on inductive matrix completion. Bioinformatics. 2018;34(19):3357–64.
19. Chen X, Wang L, Qu J, Guan NN, Li JQ. Predicting miRNA-disease association based on inductive matrix completion. Bioinformatics. 2018; 34(24):4256–65.
20. Xu M, Jin R, Zhou Z-H. Speedup matrix completion with side information: application to multi-label learning. In: Advances in neural information processing systems, vol. 2013; 2013. p. 2301–9.
21. Glazar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. RNA. 2014;20(11):1666–70.
22. Schriml LM, Arze C, Nadendla S, Chang YW, Mazaitis M, Felix V, Feng G, Kibbe WA. Disease ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012;40(Database issue):D940–6.
23. Levenshtein VI. Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady, vol. 1966; 1966. p. 707–10.
24. Li J, Gong B, Chen X, Liu T, Wu C, Zhang F, Li C, Li X, Rao S, Li X. DOSim: an R package for similarity between diseases based on disease ontology. BMC Bioinformatics. 2011;12:266.
25. Wang JZ, Du Z, Payattakool R, Yu PS, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10):1274–81.
26. van Laarhoven T, Nabuurs SB, Marchiori E. Gaussian interaction profile kernels for predicting drug-target interaction. Bioinformatics. 2011;27(21): 3036–43.
27. Ji H, Yu W, Li Y. A rank revealing randomized singular value decomposition (r3svd) algorithm for low-rank matrix approximations. arXiv preprint arXiv: 160508134; 2016.
28. Toh K-C, Yun S. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. Pacific J Optim. 2010; 6(615–640):15.
29. Cai JF, Candès EJ, Shen Z. A singular value thresholding algorithm for matrix completion. SIAM J Optim. 2010;20:1956–82.
30. Chen X, Yan GY. Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics. 2013;29(20):2617–24.
31. Chen X, Qu J, Yin J. TLHNMDA: triple layer heterogeneous network based inference for MiRNA-disease association prediction. Front Genet. 2018;9:234.
32. Chen X. KATZLDA: KATZ measure for the lncRNA-disease association prediction. Sci Rep. 2015;5:16840.
33. Nagini S. Breast cancer: current molecular therapeutic targets and new players. Anticancer Agents Med Chem. 2017;17(2):152–63.
34. Lü L, Sun J, Shi P, Kong W, Xu K, He B, Zhang S, Wang J. Identification of circular RNAs as a promising new class of diagnostic biomarkers for human breast cancer. Oncotarget. 2017;8(27):44096.
35. Ang TL, Fock KM. Clinical epidemiology of gastric cancer. Singap Med J. 2014;55(12):621.
36. Shao Y, Li J, Lu R, Li T, Yang Y, Xiao B, Guo J. Global circular RNA expression profile of human gastric cancer and its clinical significance. Cancer Med. 2017;6(6):1173–80.
37. Li P, Chen S, Chen H, Mo X, Li T, Shao Y, Xiao B, Guo J. Using circular RNA as a novel type of biomarker in the screening of gastric cancer. Clin Chim Acta. 2015;444:132–6.
38. Yiu AJ, Yiu CY. Biomarkers in colorectal cancer. Anticancer Res. 2016;36(3): 1093–102.
39. Xiong W, Ai Y-Q, Li Y-F, Ye Q, Chen Z-T, Qin J-Y, Liu Q-Y, Wang H, Ju Y-H, Li W-H. Microarray analysis of circular RNA expression profile associated with 5-fluorouracil-based chemoradiation resistance in colorectal cancer cells. Biomed Res Int. 2017;2017:8421614.
40. Dudekula DB, Panda AC, Grammatikakis I, De S, Abdelmohsen K, Gorospe M. CircInteractome: a web tool for exploring circular RNAs and their interacting proteins and microRNAs. RNA Biol. 2016;13(1):34–42.
41. Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. Nucleic Acids Res. 2017;46(D1):D296–302.
42. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11): 2498–504.
43. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. Nucleic Acids Res. 2018;47:D1013–7.

Publisher's Note

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