Briefings in Bioinformatics, 21(4), 2020, 1425–1436

doi: 10.1093/bib/bbz080
Advance Access Publication Date: 14 October 2019
Problem solving protocol

Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods

Xiangxiang Zeng*, Yue Zhong*, Wei Lin* and Quan Zou

Corresponding author: Quan Zou, Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu, China. Tel.: 0592-2580033; Fax: +86-592-2580258; E-mail: zouquan@nclab.net
*These authors contribute equally to this paper.

Abstract

Identification of disease-associated circular RNAs (circRNAs) is of critical importance, especially with the dramatic increase in the amount of circRNAs. However, the availability of experimentally validated disease-associated circRNAs is limited, which restricts the development of effective computational methods. To our knowledge, systematic approaches for the prediction of disease-associated circRNAs are still lacking. In this study, we propose the use of deep forests combined with positive-unlabeled learning methods to predict potential disease-related circRNAs. In particular, a heterogeneous biological network involving 17 961 circRNAs, 469 miRNAs, and 248 diseases was constructed, and then 24 meta-path-based topological features were extracted. We applied 5-fold cross-validation on 15 disease data sets to benchmark the proposed approach and other competitive methods and used Recall@k and PRAUC@k to evaluate their performance. In general, our method performed better than the other methods. In addition, the performance of all methods improved with the accumulation of known positive labels. Our results provided a new framework to investigate the associations between circRNA and disease and might improve our understanding of its functions.

Key words: circular RNAs; disease; biological networks; topological features; deep forests; positive-unlabeled learning

Introduction

Circular RNA (circRNA) is a type of recently 'rediscovered' RNA molecules that abundantly exist in various organisms [1–5]. It is believed that most RNAs are linear structures in the past 20 years, and circRNAs with a nonlinear structure is considered as the product of error transcription of RNA [6–8]. However, the existence of circRNAs has now been confirmed in human cells by deep sequencing of RNA [9]. CircRNAs are characterized by their noncollinearity, in which a splice donor attacks an upstream acceptor, forming a covalently closed circular structure [1, 3, 9–13]. Due to this characteristic, circRNAs can escape the digestion of exonuclease and are therefore more stable than linear RNAs [14]. This feature in combination with their ubiquity in cancer tissues, saliva, blood, and exosomes suggests that circRNAs are promising as biomarkers for diseases. Some studies have shown that circRNAs are transcribed by RNA polymerase II, and their biogenesis is likely mediated by the spliceosome [15]. It indicates that circRNAs affect gene regulation by competing with linear splicing during the cotranscription process, leading to a change in the level of gene expression [15, 16]. Another function of circRNAs is their capability to act as microRNA sponges [2, 3]. Specifically, one circRNA molecule can sequester multiple miRNAs from binding to their target mRNAs, thus affecting the

Xiangxiang Zeng is a professor in Hunan University. His research interests include biocomputing and bioinformatics.
Yue Zhong is a graduate student in Xiamen University. Her research interest is bioinformatics.
Wei Lin is a graduate student in Xiamen University. His research interest is classification of proteins in bioinformatics.
Quan Zou is a professor in University of Electronic Science and Technology of China. His research interest is bioinformatics.
Submitted: 26 March 2019; **Received (in revised form):** 23 May 2019

activity of miRNAs. In short, circRNAs have now been discovered to play vital roles in biological processes.

With the facilitation of high-throughput experiments, such as next-generation sequencing technologies, accumulating evidence has demonstrated the associations between aberrant expression of circRNAs and diseases [17–19]. Circular ANRIL is a species of circRNAs. Its production is associated with polycomb group-mediated repression of the human INK4a/ARF locus and is correlated with the risk of human atherosclerosis [14, 20]. UBE2A, an autophagic, phagocytic protein essential in the sporadic Alzheimer's disease (AD) and other progressive inflammatory degenerations of the human CNS, is depleted in AD brain [21, 22]. CircRNA is enriched in mammalian brain tissue and forms a miRNA-7–circRNA system with miRNA. This system may cause downregulation of gene expression, such as decrease UBE2A in brain. Therefore, circRNAs have the risk of causing AD. As microRNA sponges, circRNAs provide a novel mechanism for regulating the expression of microRNA. Cdr1as is a type of circRNA. Overexpression of Cdr1as in islet cells can inhibit miR-7 function to improve insulin secretion [23]. Similar effects of circRNAs also exist in some tumor cells. Expression of Cdr1as may reduce the miR-7 activity. Accordingly, the proliferative activity and invasiveness of glioma cells, breast cancer cells, and gastric cancer cells are significantly inhibited [24]. The above conclusions are all verified by large numbers of experiments. However, compared with the vast number of catalogued circRNAs [25–28], the number of experimentally validated disease-associated circRNAs is still scarce because the procedures of laboratory experiments are always costly and time-consuming. Computational approaches provide an efficient way to explore the associations on a large scale.

Based on the assumption and data from several related databases, Ghosal *et al*. [29] developed a statistical method that calculates the likelihood of a circRNA being associated with a disease and further compiled the potential associations between circRNAs and diseases into a database called Circ2Traits. However, the method of Ghosal *et al*. [29] has two limitations. First, their analysis focused only on a small portion of the currently identified circRNAs. Second, they used Bonferroni correction to deal with the problem of multiple testing, which tends to be conservative and may lead to a high false negative rate, especially when applying it on a vast number of circRNAs. Nevertheless, Circ2Traits can provide reliable association data between circRNAs and diseases for our experiments.

As increasing numbers of circRNA species and functions are discovered, the application and research of circRNA is more extensive. Due to the features of circRNA in human cells, circRNA is considered a potential biomarker for various diseases [30]. Looking for associations between circRNAs and diseases can provide a new perspective on understanding disease mechanisms. Systematic approaches for the prediction of disease-associated circRNAs can effectively target large numbers of circRNAs that have been discovered. However, only a small portion of the known associations has been verified by experiments. How to solve large numbers of unknown associations becomes a difficult. There is still a lack of systematic computational approaches until now. Therefore, we proposed a practical model to predict the associations between circRNAs and diseases.

In the present study, we performed the prediction of disease-associated circRNAs with a positive-unlabeled learning strategy that exploits deep forests. In particular, a heterogeneous biological network involving 17 961 circRNAs, 469 miRNAs, and 248 diseases was constructed, and then 24 meta-path-based topological features were extracted. Potential disease-associated circRNAs were downloaded from Cic2Traits, and those with P-value of <0.05 were further taken as positive labels. Nested 5-fold cross-validation was performed on 15 disease data sets to benchmark the proposed approaches. A series of comparative experiments were implemented with the existing methods to evaluate the predictive performance. The Recall@k, Precision@k and PRAUC@k metrics show the superiority of our method. Specifically, the average PRAUC@k of our method was ≈ 0.0075 for all disease data sets, while the second best method was less than 0.0060. All these results demonstrated that our method can provide a powerful and useful tool in predicting unknown associations between circRNAs and diseases.

Materials and methods

Heterogeneous biological network construction

A heterogeneous information network involving circRNAs, miRNAs, and diseases was constructed to predict disease-associated circRNAs. This network includes five types of biological networks, namely, a circRNA co-expression network, a miRNA–miRNA functional similarity network, a miRNA–circRNA interaction network, a disease–disease similarity network, and a miRNA–disease association network.

CircRNA co-expression network

We constructed a circRNA co-expression network as follows. First, we downloaded 80 total RNA-Seq tissue samples (Table 1) from ENCODE project [31, 32]. After data cleaning for each sample, we used CIRI [33] to detect circRNA candidates and obtained their expression data. Candidates not catalogued in CircBase [25] or not being detected in three samples or more were further filtered, and 17 961 candidates remained. The expression data of these 17 961 circRNAs were further normalized with sequencing depth and processed with WGCNA [34] package for co-expression analysis. As described in WGCNA, the co-expression similarity adjacency was further transformed with topological overlap measure, which is proportional to the number of common neighbors that a pair of nodes share and used in our following analysis. To validate our circRNA co-expression analysis, we found that the circRNA coexpression network, as other biological networks, exhibited a high degree of scale-free property with a fit index reaching 0.85 (Figure 1).

miRNA–miRNA functional similarity network

To construct a miRNA–miRNA functional similarity network, we extracted a target gene list for each miRNA deposited in miRTarBase 7.0 [35]. Then, tf-idf (term frequency-inverse document frequency) transformation was applied to obtain the miRNA representation matrix. Finally, the cosine similarity between two miRNAs was computed as their functional similarities. As evidence to support the validity of our method, we observed that the functional similarities of miRNAs from the same miRNA family or cluster were significantly higher than the miRNAs drawn from different families/clusters and/or randomly (Figure 2 and Table 2).

miRNA–circRNA interaction network

CircRNAs can sequester miRNAs. To measure the probability of a circRNA to act as miRNA sponges, we downloaded the spliced sequences of circRNAs from CircBase, mature miRNA sequences from MirBase [36], and mRNA sequences from GENCODE [37]. The transcript with the longest 3′ UTR (Untranslated Region) was

Table 1. Details of the 80 total RNA samples used in circRNA co-expression analysis

No.	SRA Id	Organ	Sex	Age	#Clean	%Clean
1	SRR4421868	Prostate gland	M	37	38.6	100
2	SRR4422158	Prostate gland	M	54	66.1	100
3	SRR4422592	Body of pancreas	F	53	48.4	99.9
4	SRR4422136	Body of pancreas	F	51	48.3	99.9
5	SRR4421506	Liver	F	53	41.2	99.9
6	SRR4421874	Liver	F	53	52.8	99.9
7	SRR4421791	Thoracic aorta	M	37	47.6	98.8
8	SRR4421792	Thoracic aorta	M	37	40.7	98.7
9	SRR4422345	Thoracic aorta	M	54	51.9	98.8
10	SRR4422192	Ascending aorta	F	53	51.3	100
11	SRR4422152, SRR4422153	Thyroid gland	F	51	82.0	100
12	SRR4421528, SRR4421529	Thyroid gland	F	53	73.5	100
13	SRR4421334	Spleen	M	54	28.7	100
14	SRR4421642	Spleen	M	37	45.6	100
15	SRR4421779	Lung	F	51	47.7	99.9
16	SRR4421758	Lung	F	53	47.1	99.9
17	SRR4422347	Gonad of ovary	F	53	54.0	99.9
18	SRR4422625	Gonad of ovary	F	51	64.2	99.9
19	SRR4422587, SRR4422588	Gonad of testis	M	37	91.5	100
20	SRR4421667, SRR4421668	Gonad of testis	M	54	83.8	100
21	SRR4422339	Adrenal gland	F	51	51.5	100
22	SRR4422204	Adrenal gland	F	53	58.3	100
23	SRR4421966	Right atrium auricular region	F	53	55.9	99.9
24	SRR4421817	Right atrium auricular region	F	51	43.2	99.9
25	SRR4422373	Stomach	F	51	60.4	100
26	SRR4422210	Stomach	F	53	57.7	100
27	SRR4421946	Skin of lower leg	F	51	52.8	100
28	SRR4422571	Skin of lower leg	F	53	52.3	100
29	SRR4421313	Large intestine of sigmoid colon	F	53	43.9	98.9
30	SRR4422344	Large intestine of sigmoid colon	F	51	40.3	98.9
31	SRR4422293	Large intestine of transverse colon	F	53	77.6	98.9
32	SRR4421756	Large intestine of transverse colon	F	51	74.8	100
33	SRR4421314	Esophagus muscularis mucosa	F	53	50.1	99.9
34	SRR4422046	Esophagus muscularis mucosa	F	51	52.8	99.9
35	SRR4421678	Esophagus squamous epithelium	F	53	33.7	98.4
36	SRR4421886	Esophagus squamous epithelium	F	51	39.0	98.4
37	SRR4422023	Gastrocnemius medialis	F	51	50.9	99.9
38	SRR4422107	Gastrocnemius medialis	F	53	60.9	99.9
39	SRR4422603	Gastroesophageal sphincter	F	51	51.8	98.6
40	SRR4422217	Gastroesophageal sphincter	F	53	40.4	98.5
41	SRR3192461	Uterus	F	28	87.0	99.4
42	SRR3192462	Uterus	F	24	70.5	99.4
43	SRR3192433	Heart	F	19	75.0	97.1
44	SRR3192434	Heart	F	28	93.7	97.1
45	SRR3192463	Temporal lobe	F	24	141.3	99.5
46	SRR3192464	Temporal lobe	F	20	30.7	99.5
47	SRR3192429	Urinary bladder	F	24	104.9	98.5
48	SRR3192430	Urinary bladder	F	20	109.8	99.2
49	SRR3192455	Thyroid gland	F	40	87.6	99.7
50	SRR3192456	Thyroid gland	F	37	77.1	99.6
51	SRR3192465	Eye	F	24	73.1	98.1
52	SRR3192466	Eye	F	20	70.0	97.6
53	SRR3192424	Frontal cortex	M	22	114.1	99.4
54	SRR3192425	Front cortex	F	20	139.1	99.3
55	SRR3192447	Skin of body	F	24	113.1	99.5
56	SRR3192448	Skin of body	M	22	79.4	99.5
57	SRR3192453	Skeletal muscle	F	19	92.8	99.6
58	SRR3192454	Skeletal muscle	M	22	100.7	99.4
59	SRR3192439	Liver	F	20	85.7	99.5
60	SRR3192440	Liver	M	22	103.0	99.4

(Continued)

Table 1. Continued

No.	SRA Id	Organ	Sex	Age	#Clean	%Clean
61	SRR3192431	Diencephalon	M	22	78.0	99.6
62	SRR3192432	Diencephalon	F	20	67.2	99.6
63	SRR3192445	Parietal lobe	F	24	101.8	99.5
64	SRR3192446	Parietal lobe	M	22	102.3	99.5
65	SRR3192451	Spinal cord	F	24	110.9	99.3
66	SRR3192452	Spinal cord	M	22	100.1	99.3
67	SRR3192437	Metanephros	F	24	99.5	99.5
68	SRR3192438	Metanephros	F	20	95.0	99.5
69	SRR3192441	Lung	F	24	113.2	97.6
70	SRR3192442	Lung	F	20	112.0	**96.7**
71	SRR3192443	Occipital lobe	F	20	89.8	98.9
72	SRR3192444	Occipital lobe	M	22	110.7	98.8
73	SRR3192427	Cerebellum	F	19	120.9	99.1
74	SRR3192428	Cerebellum	F	37	86.2	98.9
75	SRR3192458	Tongue	F	20	66.1	98
76	SRR3192457	Tongue	F	24	74.6	97.9
77	SRR3192449	Stomach	M	36	63.2	99.7
78	SRR3192450	Stomach	F	40	68.9	99.7
79	SRR3192459	Umbilical cord	M	31	167.7	98
80	SRR3192460	Umbilical cord	M	20	**12.0**	97.7

Note. On the 'Age' column, since the first 40 samples are from adult tissues, the unit for them is 'years', while the last 40 samples are from fetus tissues and the unit is 'months'. On the '#Clean' column, the unit is 'million read pairs'. We can see that, as highlighted in bold, the lowest sequencing depth was SRR3192460 sample, with 12.0 million read pairs after data cleaning. '%Clean' column was the percentage of read pairs left after data cleaning process. Also as highlighted in bold, the minimum was sample SRR3192442, with 96.7% data remained after cleaning, which indicated the high quality of these sequencing data. M: male; F: female.

taken as representative sequence for each protein coding gene. Then, target prediction was performed with TargetScan [38], and the miRNA binding density for each circRNA or mRNA was calculated. Finally, the proportion of mRNAs whose miRNA-binding density was smaller than that of a circRNA was computed as the possibility of a circRNA to act as miRNA sponge.

miRNA–disease association network and disease–disease similarity network

We constructed a miRNA–disease association network by collecting data from miR2Disease [39] and HMDD v2.0 [40], whereas disease–disease similarity data were downloaded from the supplementary data of Sun *et al*. [41], which was calculated based on text mining result of disease-associated symptoms.

Meta-path-based topological features

A meta-path [42], which represents a distinct semantic relation, is a composite of various links between two nodes in the heterogeneous network. Meta-path-based features were widely used in link prediction, and long meta-paths were considered to contribute only limited information to it. In this work, we enumerated all the 12 meta-paths between circRNA and disease of which the length was $\leq$4 (Figure 3) and then calculated PathCount and RandomWalk measures [43] to obtain the topological features. While PathCount counted the number of path instances between two nodes, RandomWalk measured the probability of arriving at each terminal node (disease) when we walked from a specific start node (circRNA).

Methods

By integrating this heterogeneous information network, we extracted 24 meta-path-based topological features to predict potential associations between circRNAs and diseases. We obtained sample information of diseases and circRNAs from the public database, and most of the associations are unknown. Labels are appended to train machine learning models.

Deep forests combined with positive-unlabeled learning algorithm

To predict potential associations between circRNAs and diseases, we computed samples from the unlabeled data that are likely to have the positive labels. We observed that positive samples and unlabeled samples were imbalanced in our data sets. Using traditional classification methods to predict our data sets may get unsatisfactory results. Therefore, we proposed an effective positive-unlabeled learning algorithm to solve this imbalance problem. This positive-unlabeled learning strategy we used was divided into two steps. First, each disease data set contains many unlabeled samples and a small number of positive samples—sets U and P, respectively. We randomly extracted one-fifth of the samples from the unlabeled set U as a subset, assuming it is a negative sample set N. Trained a classifier with set N and a positive sample set P, the specific parameters of which are given in the following paragraph. Then, use this classifier to predict the category probability of the unlabeled sample set U and select the top 1% of the sample with the positive probability of the category. After training for five times, we can obtain reliable negative sample set RN with the above combined data set. Next, we retrained a new classifier with positive sample set P and reliable negative sample set RN. We used 5-fold cross-validation to verify the performance of this classifier. Finally, we exploited this classifier to predict positive samples in our data sets (Figure 4). Among them, we used deep forests as the classifiers in our positive-unlabeled learning strategy. All hyperparameters of our model were chosen by empirical evaluations, including number of samples extracted and training times, etc.

Deep forests [44] outperform other classifiers on low-dimensional data. The advantage of deep forests is the ability to ensure the diversity of ensemble learning and the representation of feature information on layer-by-layer. Compared with the

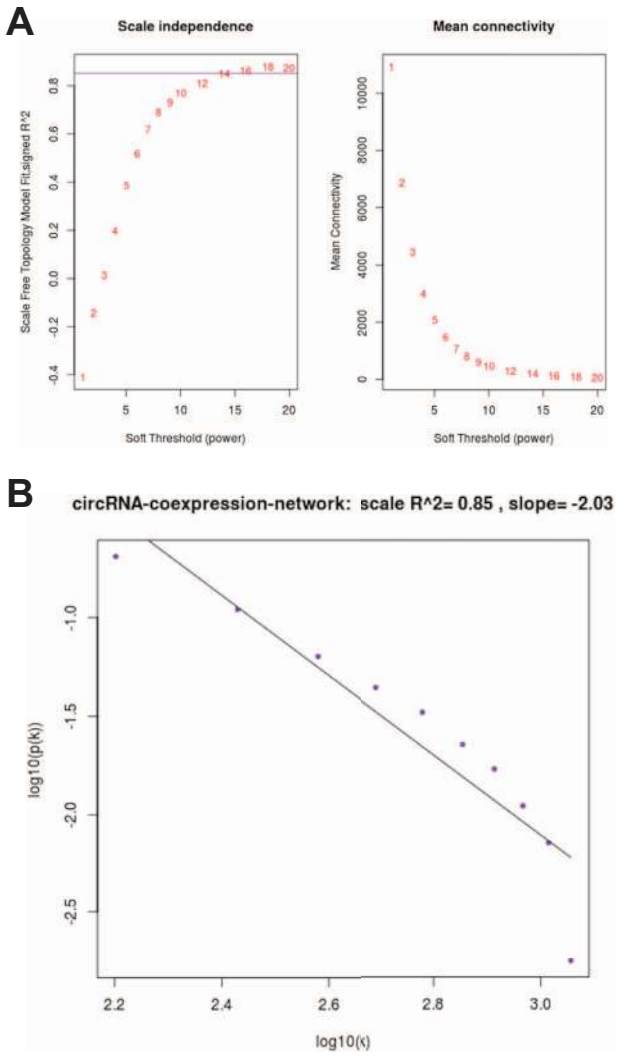

Figure 1. The scale-free property of circRNA co-expression network. (**A**) When we set the power of adjacency function [34] to 14, we achieved the best fitting result with the fitting index reaching a plateau of 0.85 (the horizon purple line). (**B**) Detail fitting line for degree distribution of circRNA co-expression network.

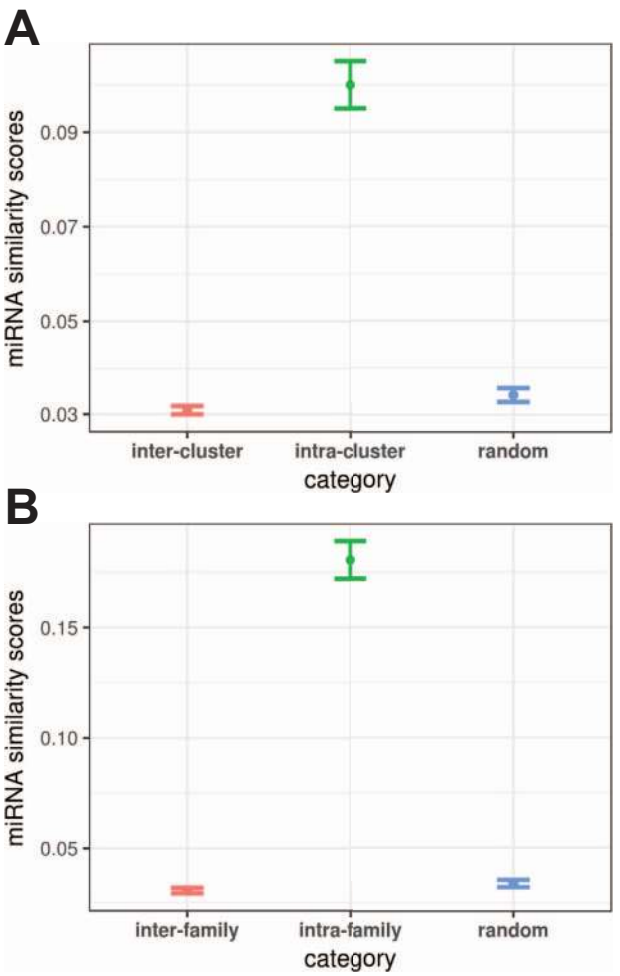

Figure 2. Comparison of miRNA functional similarity among different category (the error bars on the plots were mean ± SEM for each category). (**A**) The similarity of miRNAs in the 'intrafamily' category was much higher than miRNAs in 'interfamily' or 'random'. (**B**) The similarity of miRNAs in the 'intracluster' category was much higher than miRNAs in 'intercluster' or 'random'.

Table 2. *P*-value for miRNA–miRNA functional similarity comparison

Category	Intra versus inter	Intra versus random
Family-wise	2.872e−33	2.914e−28
Cluster-wise	1.197e−15	1.527e−12

Note. miRNA family data were downloaded from MiRBase version 21 database [1]. For miRNA clusters, according to Hansen et al. [2], miRNAs whose distances are within 50 kb were treated as in the same cluster. We randomly sampled equivalent miRNA pairs from 'interfamily' or 'random' category to compare with all the miRNA pairs from 'intrafamily'. The test used was one-tailed Wilcoxon rank sum test, with H_0: 'The functional similarity of miRNAs from "intrafamily" was smaller than or equal to that of miRNAs from "interfamily" or "random"'. We can see that the functional similarity scores of miRNAs within the 'intrafamily' are significantly higher than miRNA pairs from 'interfamily' or 'random'. Similar results were obtained from the cluster-wise analysis. These results indicate the validity of our method used to calculate the miNRA–miRNA functional similarity.

popular deep neural network, deep forests have less cost and stronger interpretable. Also, deep forests only require very few hyperparameters. We employed the cascade structure of deep forests, where each level consists of two completely random tree forests and two random forests [45]. Each forest contains 100 trees. Each tree selects the number of feature and generates leaf nodes according to their own type. In our data sets, there are 24 meta-path-based topological features as input for the first level. Each level of cascade takes feature information from the previous layer as input. The number of cascaded levels is determined by accuracy and does not need extra hyperparameters. By experiments, we demonstrated that deep forests represent excellent performance as classifiers for our positive-unlabeled learning algorithm. In order to facilitate the researchers to reconstruct the methods in this article, we provided the source code and data for all methods at https://github.com/xzenglab/DeepDCR.

Baseline and competitive methods

The baseline is the deep forest with the same structure as our method. To prove the performance of our proposed algorithm, we compared four strategies, of which a total of eight methods were applied on our data sets. These four strategies differed in the way they treated the unlabeled data during the training phase. Katz and one-class support vector machine (SVM) ignored their information, whereas weighted strategy took them all as negative data, and bagging strategy randomly sampled a portion

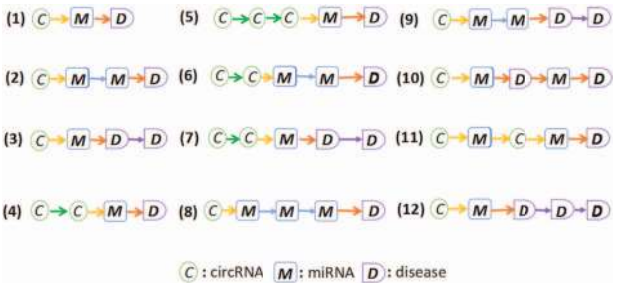

Figure 3. Details of the 12 meta-paths used. Each meta-path denoted a distinct semantic relationship between circRNA and disease. It enumerated all the 12 meta-paths between circRNA and disease of which the length was ≤4.

of them as the counterpart of the positive labels. Below are the details of our description of these four strategies.

One-class SVM. When using a binary classifier, the imbalance of training data may cause the classifier to prefer the one with a large number of category, resulting in the bias of the model. One-class SVM [46] ignored the potential useful information hidden in the unlabeled data set and tried to solve the positive-unlabeled problem by learning a function that captured the distribution of the positive data points in the training set, and then, the label of new test data was determined by whether it was located in the projected 'positive' region.

$$\min_{w\in H} \frac{||w||^2}{2} + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \rho \qquad (1)$$
$$\text{s.t.}\left(w \bigodot \varnothing(x_i)\right) \geq \rho - \xi_i, \xi_i \geq 0,$$

where w is the normal vector to the hyperplane of SVM, ξ is the slack variable for computing the cost function and l is the total number of positive samples, while v refers to the percentage of noise allowed in the positive set and ρ relates to the distance from the origin to the separating hyperplane.

Weighted strategy. For positive-unlabeled learning problem, there are separate penalties on different misclassified labels (if we take the unlabeled samples as negative, then intuitively, misjudged losses on positive samples should be penalized more heavily than losses on unlabeled samples). Then, penalties are used to adjust classifier weights. The goal of this strategy is to minimize losses.

$$L(x) = \text{argmin}\left(\sum_{j\in\{+,-\}} C^{(+,-)} p\left(j|x\right)\right) \qquad (2)$$

where C^+ and C^- are the penalty factors for the misjudgment of positive sample and negative sample, respectively, and $p\left(j|x\right)$ is the category probability of classifier to the sample x.

Methods with weighted strategy, such as weighted SVMs [47], were proven to be state-of-the-art in applications such as text classification [47], disease gene identification [48], and compound–protein interactions [49]. In the present study, weighted logit, weighted SVM and weighted RandomForest were applied in our work.

Bagging strategy. Bagging strategy has been successfully applied to solve several PU (Positive-Unlabeled) learning problems [50, 51]. This strategy balanced the training data set by undersampling unlabeled samples. Undersampling is a popular method in dealing with imbalance problems, which uses only a subset of the majority category and thus is very efficient [52]. By the bagging of ensemble learning, avoid the loss of data information during undersampling. Since bagging algorithm conducts sampling to train the model every time, it has a strong generalization ability and plays a significant role in reducing the variance of the model. To train the base classifier, we used bagging strategy to randomly sample a portion of the unlabeled data as a counterpart of the positive data set, and then, the remaining unlabeled samples were scored by the trained classifier. This process iterated for user-defined times, and finally, the average score of each unlabeled sample was calculated to infer its label.

$$U_{\text{score}} = \text{avg}\left(\sum_{i=1}^{t}\sum_{j=1}^{m} S_{ij}(u)\right) \qquad (3)$$

where $S_{ij}(u)$ is the output result of each subclassifier to the sample u, t is the number of subset to be sampled from unlabeled data set and m is the number of subclassifiers.

In our work, we combined bagging strategy with three classifiers, namely, logistic regression, SVM and AdaBoost.

Katz. Katz [53] is a method that integrates different meta-paths information to calculate the similarity between two nodes in the heterogeneous information network. It was proved to be effective for link prediction in social network [54]. Recently, it has also been successfully applied in disease gene prediction [51]. Katz only considered PathCount as effective similarity metrics. Similarly, we focused on the meta-path with the length of less than 4 for Katz, as shown below:

$$\text{Score}\ (d,c) = \sum_{l=1}^{4} \beta^l \left(\text{paths}^l_{d,c}\right) \qquad (4)$$

where the contribution of longer meta-paths was exponentially damped by the factor β. The association score between a disease (d) and a circRNA (c) was calculated by summarizing the total number of path instances of a specific length l, of which was then multiplied by β^l.

Results

With topological features extracted from the constructed heterogeneous biological network and the positive-unlabeled learning methods applied, we described our experiment setup and compared the results of our method and other competitive methods in this section.

Experimental data

We downloaded the association data between circRNAs and diseases from the Circ2Traits database. An association between circRNA and disease would be considered as a positive instance only when both the circRNA and the disease were within our heterogeneous biological network, and of which the Bonferroni-corrected P-value was <0.05. Too small a training set will make overfitting much harder to avoid, and outliers become much more dangerous. To avoid the randomness of the results caused by extremely sparse data sets, we had to ensure that there are sufficient positive samples for a disease. Thus, we only considered diseases associated with ≥60 circRNAs. As shown in Table 3, 15 diseases satisfied the criteria. Among them, breast neoplasms had the most positive samples, with 233 associated circRNAs, but

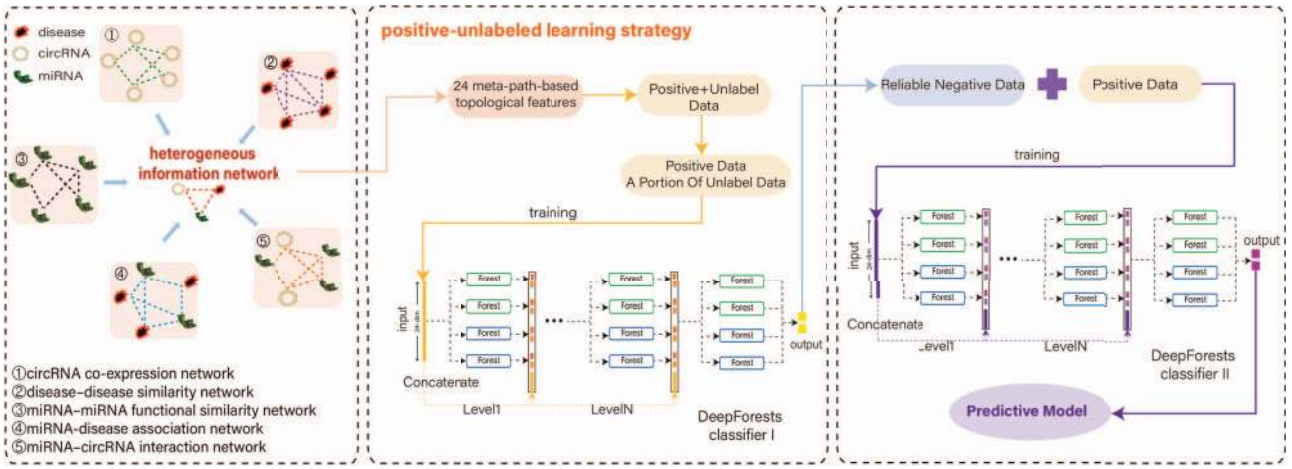

Figure 4. Flowchart of deep forests combined with positive-unlabeled learning algorithm. Schematic of the proposed method. First, we constructed a heterogeneous biological network using different sources and extracted 24 meta-path based topological features in the network. Then, the positive-unlabeled learning strategy we used was divided into two steps. First, we randomly extracted a portion from the unlabeled samples and trained with positive samples. After training for five times, we found reliable negative samples with the above combined data set. Then, we retrained a new classifier with positive samples and reliable negative samples. Finally, we exploited this classifier to predict positive samples in our data sets. Among them, we used deep forests as the classifiers.

Table 3. Fifteen disease data sets used in our experiment

No.	MeSHId	Disease name	#Pos	%Pos	Ratio
1	D001943	Breast neoplasms	233	1.30	76
2	D008175	Lung neoplasms	196	1.09	91
3	D006528	Carcinoma, hepatocellular	183	1.02	97
4	D007889	leiomyoma	154	0.86	116
5	D013274	Stomach neoplasms	142	0.79	125
6	D015451	Leukemia, lymphocytic, chronic, B-cell	135	0.75	132
7	D005910	Glioma	115	0.64	155
8	D011471	Prostatic neoplasms	112	0.62	159
9	D002289	Carcinoma, non-small cell lung	103	0.57	173
10	D018281	Cholangiocarcinoma	80	0.45	224
11	D010051	Ovarian neoplasms	74	0.41	242
12	D015470	Leukemia, myeloid, acute	72	0.40	248
13	D015458	Leukemia, T-cell	66	0.37	271
14	D012174	Retinitis pigmentosa	62	0.35	289
15	D012559	Schizophrenia	60	0.33	298

Note. Fifteen disease data sets were used in our experiment. #Pos: number of positive labels in the data set. %Pos: percentage of positive labels in the data set. Ratio: ratio between unlabeled samples to positive samples.

they only accounted for 1.3% of the 17 961 circRNAs, and the ratio between unlabeled and positive samples was about 76. With such an extreme imbalanced class labels, it was a challenging PU learning task.

Experimental settings and evaluation metrics

To evaluate the performance of these methods on our data sets, we used the 5-fold cross validation to run our experiments. In specific, we divided our data into five bins and preserved the same percentage of positive and unlabeled samples across bins. Each time, we rotated to take one of them as the test set for evaluation while the other four bins as the training set. Our positive-unlabeled learning algorithm was trained five times to find reliable negative samples in the whole data sets, and then, we trained a new classifier with reliable negative samples and positive labeled samples, which was used to predict unlabeled samples. For all the competitive methods, we also used 5-fold cross-validation to select the optimal parameters in the training set. After we obtained the optimal parameters, the methods were retrained with these parameters and all the training data and finally evaluated on the test set (Figure 5). The overall performance was an average of the five different test sets. To reduce the influence of sampling randomness, we repeated the whole process for 10 times.

To evaluate the performance of a method, we only use the top-k predictions. The motivation is to evaluate the method's capability of recovering a positive association in the top-k predictions for a given disease [55]. Since we focused on the top-k predictions, we have redefined the performance metrics, including Recall@k, Precision@k and PRAUC@k ($k = 500$). Recall measured the proportion of true positives recovered to the total number of them in the hidden set.

$$\text{Recall@k} = \frac{TP}{P_k} \tag{5}$$

where TP indicates the number of true positives recovered in the top-k predictions. P_k indicates the number of positive samples

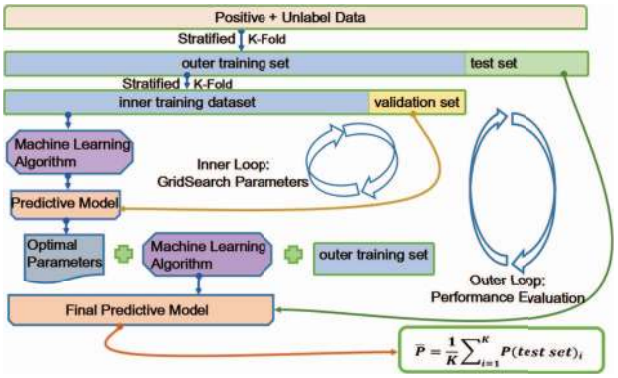

Figure 5. Nested k-fold cross-validation to evaluate the performance of methods ($k = 5$). Detailed experimental setting for all competitive methods. The inner 5-fold cross-validation to select the optimal parameters in the training set, and the outer 5-fold cross-validation retained the model with the optimal parameters and evaluated the performance.

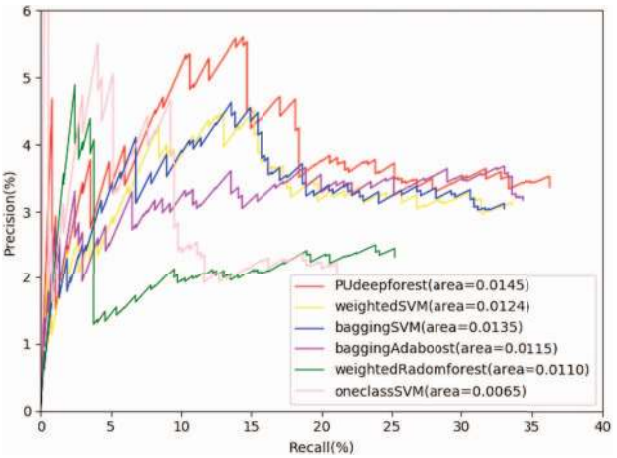

Figure 6. Precision–Recall curve of different methods. The plot shows precision versus recall rates of different methods for different values of k thresholds, ranging $1 \leq k \leq 500$. Precision is the fraction of true positive association recovered in the top-k predictions for a disease. Recall is the ratio of true positive association recovered in the top-k predictions to the total number of positive samples for the disease in the test set. PRAUC of the proposed method was still in the lead.

known in the top-k predictions. Precision metric is the fraction of true positives in the k predictions.

$$\text{Precision@k} = \frac{\text{TP}}{k} \tag{6}$$

PRAUC@k considered the area under Precision–Recall curve, which obtained by adjusting the threshold k. In our experiment, we plotted the Precision–Recall curve after we obtained results in different values of k ($1 \leq k \leq 500$) (Figure 6). Figure 7 showed the average Recall@k of various methods for all diseases in different values of k.

Experimental results

Overall performance

For each method, we calculated its average recall rate on every disease data set, and the results for all the 15 disease data sets

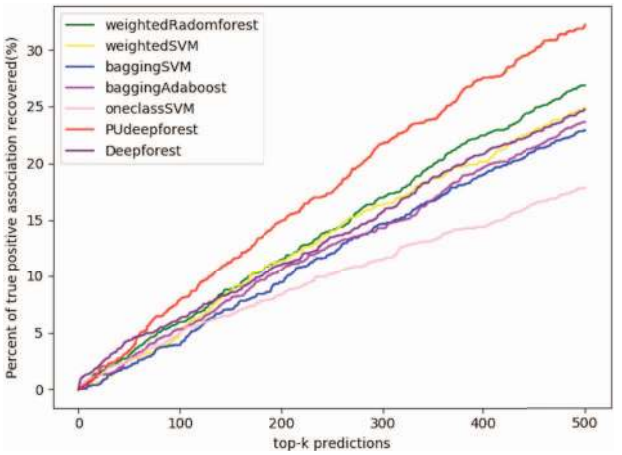

Figure 7. Comparison of disease–circRNA association prioritization methods. It shows the average recall of various methods for all diseases in different values of k. The vertical axis in the plots shows the probability that a positive association is recovered in different k values (shown on the horizontal axis) predictions for disease data sets. We observed that the proposed method consistently and significantly outperformed competitive methods by a large margin over almost all k values.

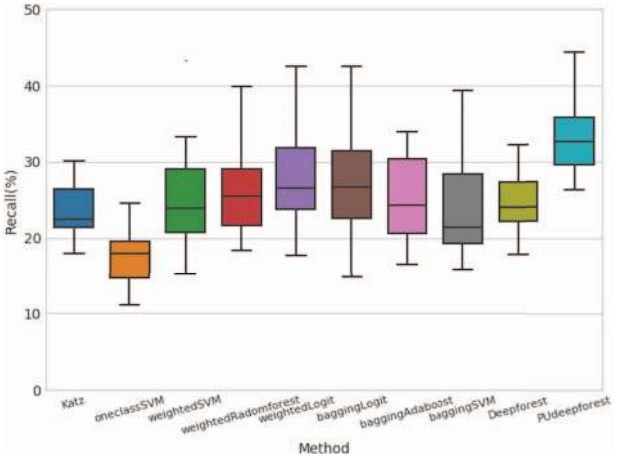

Figure 8. Recall@k scores of different methods. The box plot shows the distribution of recall values in 15 disease data sets for each method.

are presented in Figure 8. As shown in Figure 8, the performance of our method was considerably better than the other methods, irrespective of whether from the perspective of average or median of recall rate. Meanwhile, the average performance of one-class SVM was the worst. Intriguingly, other competitive methods, Katz, weightedLogit, weightedSVM, baggingLogit, baggingSVM, weightedRandomforest and baggingAdaBoost, albeit in different ways, performed similarly. The mean Recall@k and Precision@k score of each method on the 15 disease data sets were shown in Tables 4 and 5, respectively.

The average recall rate of different methods for a specific k was also presented in Figure 7. The vertical axis in the plots showed the probability that a positive association is recovered in different k values (shown on the horizontal axis) predictions for disease data sets. We observed that the proposed method consistently and significantly outperformed competitive methods by a large margin over almost all k values. Our method had more than 30% chance of recovering a positive association in the top 500 predictions for disease data sets, whereas the

Table 4. Recall@k score of each method on 15 disease data sets

Data set	dF	pu-dF	bAda	bSVM	bLog	I-SVM	wRF	wSVM	wLog	Katz
D001943	23.47	33.26	**33.60**	33.26	33.50	18.97	33.03	31.85	33.13	28.15
D002289	22.74	**31.00**	20.88	19.01	19.81	19.48	21.37	16.73	20.02	18.06
D005910	**27.31**	26.26	18.35	16.87	18.08	13.39	19.74	18.61	17.74	21.65
D006528	22.84	**27.75**	24.15	21.36	26.18	16.40	22.78	25.09	25.97	23.45
D007889	20.20	**27.29**	20.40	19.75	23.33	17.97	20.99	23.19	24.56	22.41
D008175	28.12	**36.28**	31.60	29.53	31.38	19.24	32.26	33.32	32.56	27.30
D010051	32.30	**44.38**	33.91	39.35	42.57	23.20	39.84	43.38	42.57	30.10
D011471	25.51	**34.27**	28.73	19.49	23.13	24.54	28.11	19.58	23.93	22.12
D012174	20.64	**35.23**	31.35	27.16	32.27	15.77	27.78	29.65	32.81	28.31
D012559	17.83	**32.17**	16.50	15.83	14.83	11.17	18.33	15.17	22.00	17.99
D013274	24.10	**29.09**	25.21	24.79	26.71	19.57	24.88	23.09	26.50	21.00
D015451	21.70	**30.07**	21.48	19.93	26.82	16.81	21.93	23.93	27.56	22.52
D015458	25.80	**36.84**	29.46	25.81	29.09	13.43	25.48	27.00	28.52	24.90
D015470	28.11	**32.67**	20.42	17.71	22.03	20.47	27.84	21.90	23.70	21.20
D018281	27.38	**36.50**	24.25	31.13	31.38	13.50	30.00	28.50	31.00	25.50
Average	24.54	**32.87**	25.35	24.07	26.74	17.59	26.29	25.40	27.50	23.64

Note. The bold data in Table 4 indicates the best performance of Recall@k on each data set. On each data set, every method was run for 10 times with different random seed sets. The mean Recall@k score displayed on this table was multiplied by 100. The prefix 'b' character of method names denotes 'bagging' strategy, whereas 'w' represents 'weighted' strategy, and 'I-SVM' is one-class SVM. Ada: AdaBoost; RF: Randomforest; Log: Logistic regression; dF: deep forest; pu-dF: PU learning strategy of deep forest.

Table 5. Precision@k score of each method on 15 disease data sets

Data set	dF	pu-dF	bAda	bSVM	bLog	I-SVM	wRF	wSVM	wLog	Katz
D001943	2.19	3.10	**3.13**	3.10	3.13	1.77	3.08	2.97	3.09	2.62
D002289	0.93	**1.28**	0.86	0.78	0.82	0.80	0.88	0.69	0.82	0.74
D005910	**1.26**	1.21	0.84	0.80	0.83	0.62	0.91	0.86	0.82	1.00
D006528	1.67	**2.03**	1.77	1.56	1.92	1.20	1.67	1.84	1.90	1.72
D007889	1.24	**1.68**	1.26	1.22	1.44	1.11	1.29	1.43	1.51	1.38
D008175	2.20	**2.84**	2.48	2.32	2.46	1.51	2.53	2.61	2.55	2.14
D010051	0.95	**1.31**	1.00	1.16	1.26	0.74	1.18	1.28	1.26	0.89
D011471	1.14	**1.54**	1.29	0.87	1.04	1.10	1.26	0.88	1.07	0.99
D012174	0.51	**0.88**	0.78	0.67	0.80	0.39	0.69	0.74	0.81	0.70
D012559	0.43	**0.77**	0.40	0.38	0.36	0.27	0.44	0.36	0.53	0.43
D013274	1.36	**1.65**	1.43	1.41	1.52	1.11	1.41	1.31	1.50	1.19
D015451	1.17	**1.62**	1.16	1.08	1.45	0.91	1.18	1.29	1.49	1.22
D015458	0.68	**0.97**	0.78	0.68	0.77	0.36	0.67	0.71	0.75	0.66
D015470	0.80	**0.94**	0.59	0.51	0.64	0.59	0.80	0.63	0.68	0.61
D018281	0.88	**1.17**	0.78	1.00	1.00	0.43	0.96	0.91	0.99	0.82
Average	1.16	**1.53**	1.24	1.17	1.30	0.86	1.26	1.23	1.32	1.14

Note. The bold data in Table 5 indicates the best performance of Precision@k on each data set. On each data set, every method was run for 10 times with different random seed sets. The mean Precision@k score displayed on this table was multiplied by 100. The prefix 'b' character of method names denotes 'bagging' strategy, whereas 'w' represents 'weighted' strategy, and 'I-SVM' is one-class SVM. Ada: AdaBoost; RF: Randomforest; Log: Logistic regression; dF: deep forest; pu-dF: PU learning strategy of deep forest.

second best performed method, weightedRandomforest, had only round 25%.

In Figure 6, we present Precision–Recall curve of different methods for a disease. The plot showed precision versus recall rates for different values of k thresholds, ranging $1 \leq k \leq 500$. We also calculated the PRAUC@k metric for each method, which also considered the precision metric during computation. As shown in Figure 9, our method still showed the best performance on this metric. The baseline method also performed well, which only used the deep forest model. This result demonstrated the ability of deep forest models to extract valid features. Obviously, one-class SVM still demonstrated underperformance, but the shorter interquartile range of this method indicated that its performance was more consistent among different data sets than those of the other methods. We found that Katz achieved suboptimal results on this metric, and other methods still showed a similar performance. Finally, the mean PRAUC score of each method on the 15 disease data sets is shown in Table 6. As shown in Table 6, the performance of our method was the best in most data sets. Meanwhile, on D001943 and D008175, the two data sets of which were with the most positive labels, most of the methods in our experiment achieved better performance than the other methods.

Deep forests combined with positive-unlabeled learning algorithm are important for the success of our method. Katz only calculated PathCount to metric the similarity between two nodes and performed well in PRAUC@k metric, which also confirmed the validity of the topology features we extracted. One-class SVM only focused on one type of data to solve the unknown label problem. However, due to the small number of positive samples, it cannot effectively capture the positive sample boundary. For this reason, the ability of one-class SVM

Table 6. PRAUC@k score of each method on 15 disease data sets

Data set	dF	pu-dF	bAda	bSVM	bLog	I-SVM	wRF	wSVM	wLog	Katz
D001943	8.0	12.7	**14.2**	13.2	13.4	5.2	11.6	12.6	13.6	12.5
D002289	3.4	**8.6**	3.1	2.2	2.1	2.7	3.9	1.6	2.1	3.9
D005910	**5.0**	3.9	2.6	2.6	2.1	1.6	3.0	2.9	2.6	4.2
D006528	7.4	6.7	6.5	4.3	6.4	2.8	5.6	5.9	6.1	**7.8**
D007889	**11.0**	6.9	4.2	3.6	5.3	2.9	4.2	4.2	5.8	5.9
D008175	8.2	**12.1**	9.4	9.2	10.3	5.5	9.6	11.5	11.1	9.2
D010051	6.5	7.9	4.3	6.9	7.4	3.9	5.8	7.6	7.2	**8.2**
D011471	3.0	**8.7**	5.5	2.3	3.0	5.4	5.7	2.4	3.4	7.3
D012174	2.0	5.5	8.7	6.5	9.4	1.7	6.2	8.8	**11.7**	6.4
D012559	1.3	**4.1**	2.3	1.0	0.8	3.7	1.4	1.9	1.8	1.8
D013274	6.3	**6.5**	5.7	4.0	5.2	3.1	5.7	4.7	5.6	5.8
D015451	5.1	**6.8**	3.7	3.8	5.8	2.3	4.6	4.3	5.4	4.1
D015458	2.2	**6.8**	3.7	3.3	4.1	2.6	3.6	3.2	4.0	3.6
D015470	**10.6**	6.1	1.5	1.9	2.1	2.1	5.9	2.7	2.3	2.3
D018281	7.3	**8.8**	2.5	4.4	4.3	1.9	8.1	3.9	4.5	4.4
Average	5.82	**7.47**	5.19	4.61	5.45	3.16	5.66	5.21	5.81	5.83

Note. The bold data in Table 6 indicates the best performance of PRAUC@k on each data set. On each data set, every method was run for 10 times with different random seed sets. The mean PRAUC@k score displayed on this table was multiplied by 1000. The prefix 'b' character of method names denotes 'bagging' strategy, whereas 'w' represents 'weighted' strategy, and 'I-SVM' is one-class SVM. Ada: AdaBoost; RF: Randomforest; Log: Logistic regression; dF: deep forest; pu-dF: PU learning strategy of deep forest.

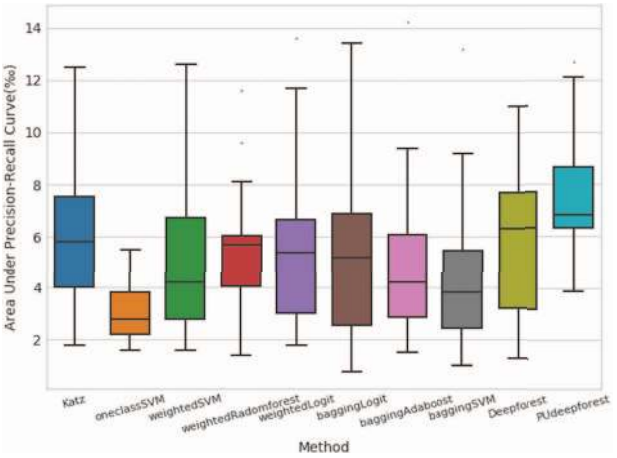

Figure 9. PRAUC@k scores of different methods. The boxplot shows the distribution of PRAUC values in 15 disease data sets for each method.

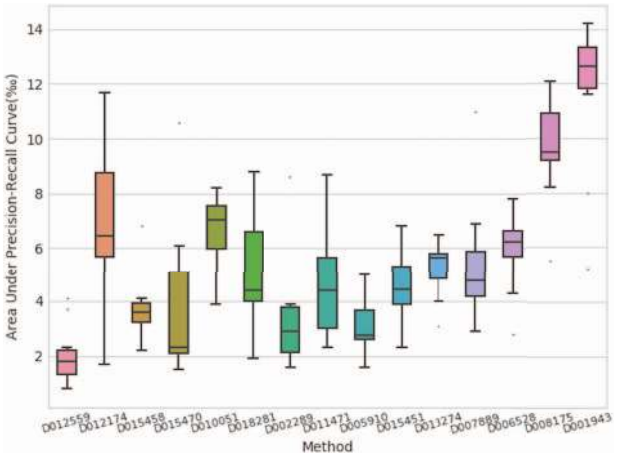

Figure 10. Relation between PRAUC and the number of positive labels in each disease data set. The box plot shows the distribution of PRAUC values in each disease data set. We found a significant positive correlation (Pearson $r = 0.666$, one-tailed P-value = 6.67e−3) between performance and the number of known positive labels in the data sets.

to predict the association between circRNAs and diseases was poor. For weighted strategies, the selection of penalty factors depended on the training of data. However, the lack of known information limited the function of the penalty factors. The major shortcoming of the bagging strategy is its much higher computational complexity.

Performance comparison among data sets

After we obtained the PRAUC@k scores of each method on the 15 disease data sets, we determined whether a performance difference exists among the data sets. We found a significant positive correlation (Pearson $r = 0.666$, one-tailed P-value = 6.67e−3) between performance and the number of known positive labels in the data sets (Figure 10). Interestingly, fewer known circRNAs were associated with diseases D012174 and D010051. Most of these methods achieved relatively high performance on these two data sets, which suggested that further investigation of these two data sets could provide insights to improve the performance of these classifiers.

Compare the computational complexity

In this section, we further evaluated the computational complexity of each method. Table 7 shows the running time of each algorithm on the same disease data. As a result, Katz ran fastest, followed by one-class SVM and the proposed method, and bagging strategy took the longest time. Theoretically, Katz only needs to consider the similarity between two nodes and calculate the number of paths between two nodes (a disease and a circRNA) and the number of path length in the network, so the computational complexity is low. One-class SVM learnt a function that captured the distribution of the positive data points in training set, and the proposed method used a limited number of negative samplings to balance the data sets, effectively reducing computational complexity. For weighted strategies, the computational complexity greatly depended on the classifier it used. Bagging algorithms required frequent sampling and training the

Table 7. Running time for each method

Method	pu-dF	bAda	bSVM	bLog	I-SVM	wRF	wSVM	wLog	Katz
Time(s)	396	6148	5184	2330	389	1173	2708	897	0.743

This table shows the running time of each algorithm on the same disease data. The prefix. 'b' character of method names denotes "bagging" strategy, whereas 'w' represents 'weighted' strategy, and 'I-SVM' is one-class SVM. Ada, Adaboost; RF, random forest; Log, logistic regression; dF, is deep forest; pu-dF, pu-learning strategy of deep forest.

classifier for each new sample set. Thus, bagging algorithms have the highest computational complexity.

Conclusion

Identification of disease-associated circRNAs not only enables us to further understand the vital roles they take part in biological processes but also promotes improvement in disease diagnosis and treatment. We proposed a systematic computational method to predict large numbers of unknown associations by deep forests joint positive-unlabeled learning algorithm. To our knowledge, it is the first computational model proposed for predicting potential associations between circRNAs and diseases. In this work, we first constructed a heterogeneous information network from five correlative biological networks. Then, we extracted 24 meta-path-based topological features in this heterogeneous information network by PathCount and RandomWalk. In addition, we proposed a positive-unlabeled learning strategy with deep forest methods to predict circRNA–disease associations. Our strategy extracted valid samples from data sets. The problem of sample imbalance is avoided to some extent. The deep forest model of the proposed method was shown to be reliable with 24 meta-path-based topological features, aside from less parameter tuning. We compared a baseline method and four strategies, Katz, one-class SVM, weighted, and bagging, with a total of eight methods on 15 disease data sets. Our method can achieve superior prediction performance over other methods on almost all disease data sets. The experimental results also show that the performance of these methods significantly correlated with the number of known positive labels in the data sets, which suggested that, with the accumulation of experimentally validated data [56, 57], positive-unlabeled learning based methods would be much more effective in the future. Moreover, we demonstrated that the proposed method was effective for all diseases in heterogeneous bioinformatics networks. Therefore, we believed that the proposed method could provide a useful and effective computational tool for biomedical researches.

Key Point

- Identification of disease-associated circRNAs provides new ideas for disease diagnosis and treatment.
- Integrate topological features from known associated networks to further understand the role of circRNAs in biological processes
- The method of combining deep forests and positive-unlabeled learning strategy performs better than other traditional methods for data imbalance problems.

Funding

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61472333, 61772441, 61472335, 61672033, 61425002, 61872309 and 61771331), Project of Marine Economic Innovation and Development in Xiamen (No. 16PFW034SF02), Natural Science Foundation of the Higher Education Institutions of Fujian Province (No. JZ160400), Natural Science Foundation of Fujian Province (No. 2017J01099) and President Fund of Xiamen University (No. 20720170054).

References

1. Jeck WR, Sharpless NE. Detecting and characterizing circular RNAs. *Nat Biotechnol* 2014;**32**:453–61.
2. Hansen TB, Jensen TI, Clausen BH, *et al*. Natural RNA circles function as efficient microRNA sponges. *Nature* 2013;**495**:384–8.
3. Memczak S, Jens M, Elefsinioti A, *et al*. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013;**495**:333–8.
4. Salzman J, Chen RE, Olsen MN, *et al*. Cell-type specific features of circular RNA expression. *PLoS Genet* 2013;**9**:e1003777.
5. Wang PL, Bao Y, Yee MC, *et al*. Circular RNA is expressed across the eukaryotic tree of life. *PLoS One* 2014;**9**:e90859.
6. Lasda E, Parker R. Circular RNAs: diversity of form and function. *RNA* 2014;**20**:1829–42.
7. Qu S, Yang X, Li X, *et al*. Circular RNA: a new star of noncoding RNAs. *Cancer Lett* 2015;**365**:141–8.
8. Chen LL. The biogenesis and emerging roles of circular RNAs. *Nat Rev Mol Cell Biol* 2016;**17**:205–11.
9. Salzman J, Gawad C, Wang PL, *et al*. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS One* 2012;**7**:e30733.
10. Nigro JM, Cho KR, Fearon ER, *et al*. Scrambled exons. *Cell* 1991;**64**:607–13.
11. Capel B, Swain A, Nicolis S, *et al*. Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell* 1993;**73**:1019–30.
12. Hansen TB, Wiklund ED, Bramsen JB, *et al*. miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO J* 2011;**30**:4414–22.
13. Zaphiropoulos PG. Exon skipping and circular RNA formation in transcripts of the human cytochrome P-450 2C18 gene in epidermis and of the rat androgen binding protein gene in testis. *Mol Cell Biol* 1997;**17**:2985–93.
14. Jeck WR, Sorrentino JA, Wang K, *et al*. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 2013;**19**:141–57.
15. Ashwal-Fluss R, Meyer M, Pamudurti Nagarjuna R, *et al*. circRNA biogenesis competes with pre-mRNA splicing. *Mol Cell* 2014;**56**:55–66.

16. Zhang X-O, Wang H-B, Zhang Y, *et al*. Complementary sequence-mediated exon circularization. *Cell* 2014;**159**: 134–47.
17. Dong Y, He D, Peng Z, *et al*. Circular RNAs in cancer: an emerging key player. *J Hematol Oncol* 2017;**10**(2).
18. Kristensen LS, Hansen TB, Veno MT, *et al*. Circular RNAs in cancer: opportunities and challenges in the field. *Oncogene* 2018;**37**:555–65.
19. Wang Y, Mo Y, Gong Z, *et al*. *Circular RNAs in human cancer* Mol Cancer 2017;**16**:25.
20. Burd CE, Jeck WR, Yan L, *et al*. Expression of linear and novel circular forms of an INK4/ARF-associated noncoding RNA correlates with atherosclerosis risk. *PLoS Genet* 2010;**6**:e1001233.
21. Irina L. Diminished parkin solubility and co-localization with intraneuronal amyloid-β are associated with autophagic defects in Alzheimer's disease. *Journal of Alzheimer's disease: JAD* 2013;**1**.
22. Lukiw WJ. Circular RNA (circRNA) in Alzheimer's disease (AD). *Front Genet* 2013;**4**:307.
23. Xu HY, Guo S, Li W, *et al*. The circular RNA Cdr1as, via miR-7 and its targets, regulates insulin transcription and secretion in islet cells. *Sci Rep* 2015;**5**:12.
24. Hansen TB, Kjems J, Damgaard CK, *et al*. miR-7 in cancer. *Cancer Res* 2013;**73**:5609–12.
25. Glazar P, Papavasileiou P, Rajewsky N. circBase: a database for circular RNAs. *RNA* 2014;**20**:1666–70.
26. Xia S, Feng J, Chen K, *et al*. CSCD: a database for cancer-specific circular RNAs. *Nucleic Acids Res* 2018;**46**:D925–9.
27. Zheng LL, Li JH, Wu J, *et al*. deepBase v2.0: identification, expression, evolution and function of small RNAs, LncRNAs and circular RNAs from deep-sequencing data. *Nucleic Acids Res* 2016;**44**:D196–202.
28. Chen X, Han P, Zhou T, *et al*. circRNADb: a comprehensive database for human circular RNAs with protein-coding annotations. *Sci Rep* 2016;**6**:34985.
29. Ghosal S, Das S, Sen R, *et al*. Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Front Genet* 2013;**4**:283.
30. Zhang Z, Yang T, Xiao J. Circular RNAs: promising biomarkers for human diseases. *EBioMedicine* 2018;**34**:267–74.
31. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature* 2012;**489**:57–74.
32. Sloan CA, Chan ET, Davidson JM, *et al*. ENCODE data at the ENCODE portal. *Nucleic Acids Res* 2016;**44**:D726–32.
33. Gao Y, Zhang J, Zhao F. Circular RNA identification based on multiple seed matching. *Brief Bioinform* 2018;**19**:803–10.
34. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;**9**:559.
35. Chou CH, Shrestha S, Yang CD, *et al*. miRTarBase update 2018: a resource for experimentally validated microRNA-target interactions. *Nucleic Acids Res* 2018;**46**:D296–302.
36. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* 2014;**42**:D68–73.
37. Harrow J, Frankish A, Gonzalez JM, *et al*. GENCODE: the reference human genome annotation for the ENCODE project. *Genome Res* 2012;**22**:1760–74.
38. Agarwal V, Bell GW, Nam JW, *et al*. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* 2015;**4**.
39. Jiang Q, Wang Y, Hao Y, *et al*. miR2Disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res* 2009;**37**:D98–104.
40. Li Y, Qiu C, Tu J, *et al*. HMDD v2.0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res* 2014;**42**:D1070–4.
41. Zhou X, Menche J, Barabasi AL, *et al*. Human symptoms-disease network. *Nat Commun* 2014;**5**:4212.
42. Sun Y, Norick B, Han J, *et al*. Pathselclus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Trans Knowl Discov Data* 2013;**7**:11.
43. Sun Y, Barber R, Gupta M, *et al*. Co-author relationship prediction in heterogeneous bibliographic networks. In: *Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2011, 121–8.
44. Zhou Z-H, Feng J. Deep forest: towards an alternative to deep neural networks. In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, 2017, 3553–9.
45. Breiman L. Random forests. *Mach Learn* 2001;**45**:5–32.
46. Schölkopf B, Williamson RC, Smola AJ, *et al*. Support vector method for novelty detection. In: *Advances in Neural Information Processing Systems*. NIPS, 2000, 582–8.
47. Liu B, Dai Y, Li X, *et al*. Building text classifiers using positive and unlabeled examples. In: *Third IEEE International Conference on Data Mining*. IEEE, 2003, 179–86.
48. Yang P, Li XL, Mei JP, *et al*. Positive-unlabeled learning for disease gene identification. *Bioinformatics* 2012;**28**:2640–7.
49. Cheng Z, Zhou S, Wang Y, *et al*. Effectively identifying compound-protein interactions by learning from positive and unlabeled examples. *IEEE/ACM Trans Comput Biol Bioinform* 2016;**15**:1832–43.
50. Mordelet F, Vert J-P. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recogn Lett* 2014;**37**:201–9.
51. Singh-Blom UM, Natarajan N, Tewari A, *et al*. Prediction and validation of gene–disease associations using methods inspired by social network analyses. *PLoS One* 2013; **8**:e58977.
52. Liu X, Wu J, Zhou Z. Exploratory undersampling for class-imbalance learning. In: *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. IEEE, 2009, 539–50.
53. Katz L. A new status index derived from sociometric analysis. *Psychometrika* 1953;**18**:39–43.
54. Wang P, Xu B, Wu Y, *et al*. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences* 2015;**58**:1–38.
55. Natarajan N, Dhillon IS. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics* 2014; **30**:i60–8.
56. Zhao Z, Wang K, Wu F, *et al*. circRNA disease: a manually curated database of experimentally supported circRNA–disease associations. *Cell Death Dis* 2018;**9**:475.
57. Yao D, Zhang L, Zheng M, *et al*. Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. *Sci Rep* 2018;**8**:11018.