CircWalk: A novel approach to predict CircRNA-Disease association based on heterogeneous network representation learning

Morteza Kouhsar
Sharif University of Technology
Esra Kashaninia
Sharif University of Technology
Behnam Mardani
Institute for Advanced Studies in Basic Sciences (IASBS)
Hamid R. Rabiee (rabiee@sharif.edu)
Sharif University of Technology

Research Article

Keywords: circRNA, ceRNA, Disease, Biological Network, Representation Learning

Posted Date: March 18th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1447028/v1

Abstract

Background: Several types of RNA in the cell are usually involved in biological processes with multiple functions. Generally, coding RNAs translate to proteins, and non-coding ones regulate this translation in the gene regulatory networks. Some single-strand RNAs can create a circular shape via the back splicing process and convert into a new type called circular RNA (circRNA). circRNAs are among the most important non-coding RNAs in the cell that involve multiple disorders. One of the key functions of circRNAs is to regulate the expression of other genes through sponging micro RNAs (miRNAs) in diseases. This mechanism, known as the competing endogenous RNA (ceRNA) hypothesis, and additional information obtained from biological datasets can be used by computational approaches to predict novel associations between disease and circRNAs.

Results: We applied multiple classifiers to validate the extracted features from the heterogeneous network and selected the best one based on some evaluation criteria. Then, the XGBoost is utilized in our pipeline to generate a novel approach, called CircWalk, to predict CircRNA-Disease associations. Our results demonstrate that CircWalk has reasonable accuracy and AUC compared with other state-of-the-art algorithms. We also use CircWalk to predict novel circRNAs associated with lung, gastric, and colorectal cancers as a case study. The results show that our approach can accurately predict novel circRNAs related to these diseases.

Conclusions: In this study, considering the ceRNA hypothesis, we integrate multiple resources to construct a heterogeneous network from circRNAs, mRNAs, miRNAs, and diseases. Next, the DeepWalk algorithm is applied to the network to extract feature vectors for circRNAs and diseases. The extracted features are used to learn a classifier and generate a model to predict novel CircRNA-Disease associations. Our approach uses the concept of the ceRNA hypothesis and the miRNA sponge effect of circRNAs to predict their associations with diseases. Our results show that this outlook could help anticipate CircRNA-Disease associations more accurately.

Background

Non-coding RNAs are essential cells players who manipulate and control many biological processes. About 80 to 90 percent of human cell transcripts are non-protein-coding [1]. There are multiple types of non-coding RNAs, and each of them has specific functions in the complex system of gene regulation. One of the most important non-coding RNAs that researchers have recently noticed is Circular RNAs (circRNAs). circRNAs created from other transcripts through a non-canonical splicing event called back splicing. In this process, the transcript's 5’ and 3’ splice sites bind together and reconstruct a circular shape called circRNA [2]. This circular structure makes the circRNAs more stable than other RNAs [3, 4] and makes them attractive as a biomarker in complex diseases [5, 6].

Multiple functions have been identified for the circRNAs in the cell [7]. They can act as enhancers for the role of other proteins or as scaffolds to mediate complex formation for some enzymes [2]. circRNAs also

regulate the RNA Binding Proteins (RBP) function by decoying them [2]. One of the most important functions for circRNAs is trapping miRNAs based on their sequence and miRNA response elements (MREs) [8]. By this function, circRNAs can regulate the expression of the coding RNAs through sponging shared miRNAs [8]. This mechanism is known as the competing endogenous RNA (ceRNA) hypothesis [9], which is involved in multiple complex diseases such as cancer [10].

circRNAs are involved in many human diseases based on previous research [5, 11, 12]. For instance, circRNA Cdr1as affect insulin secretion in the pancreatic islet cells via decoying miR-7 miRNA. Consequently, this circRNA is a therapeutic target for diabetes [13]. hsa_circ_0054633 is another circRNA that is overexpressed in patients with type 2 Diabetes Mellitus [14]. Recently, two other circRNAs (hsa_circ_0063425 and hsa_circ_0056891) have been introduced as novel biomarkers to predict type 2 Diabetes Mellitus in the early stages [15]. In cardiovascular diseases, circRNA HRCR sponge miR-233 and prevent heart failure [16]. circFndc3b is another important cardio-related circRNA that has recently been detected. It is involved in cardiac repair pathways [17]. Alzheimer's disease (AD) is another disease in which the role of proteins has been proven [18]. For instance, a circular RNA created from the IGF2R transcript is associated with AD pathology [19]. Many circRNAs have also been involved in multiple cancer types [20]. For example, in glioma, circRNA 0001445 promotes tumor progression through the miRNA-127-5p/SNX5 signaling pathway [21]. hsa_circ_0062019 promotes prostate cancer cell proliferation, migration, and invasion through upregulating HMGA2 by decoying miR-195-5p [22]. Many other studies demonstrate the role of circRNAs in multiple cancer types such as thyroid, gastric, bladder, breast, and colon cancer [23–27].

Developing high-throughput technology such as RNA Sequencing (RNA-Seq) and public databases to store them has provided a valuable resource for researchers to create novel computational algorithms to mine the biological data. In circRNA-related studies, computational algorithms such as deep learning and machine learning-based methods can help predict more accurate CircRNA-Disease associations and deeply understand disease mechanisms. Many computational approaches have been developed to predict CircRNA-Disease associations in recent years. One of the most recent algorithms has been introduced by Lei et al. [28]. They reconstructed a heterogeneous network based on known circRNAs and disease relationships, circRNA-circRNA, and disease-disease similarities. After that, a novel weighted biased meta-structure search algorithm was applied to the network to predict CircRNA-Disease associations. A heterogeneous network was reconstructed in a similar approach by Zhang et al. [29]. They used multiple resources to create circRNA and disease similarity networks. In their novel algorithm, entitled PCD_MVMF, the metapath2vec + + method was applied on meta paths in the heterogeneous network. Then the matrix factorization algorithm was used to predict the novel association between circRNAs and diseases. A combination of deep learning and matrix factorization methods was also used in another study. The DMFCDA (Deep Matrix Factorization CircRNA-Disease Association) algorithm was developed to solve the problem [30]. Lu et al. developed a deep learning-based algorithm called CDASOR to predict CircRNA-Disease associations based on sequence and ontology representations with convolutional and recurrent neural networks [31]. In another study, Deng et al. proposed the KATZCPDA algorithm based on a previously developed algorithm (KATZ) [32, 33]. The KATZCPDA algorithm

integrated circRNA-protein and protein-disease association data with circRNA similarity and disease similarity data to reconstruct a heterogeneous network. Subsequently, a KATZ measure [32] was applied to extract unknown CircRNA-Disease associations by measuring the similarities between circRNAs and diseases [33].

Generally, in many computational approaches to predict CircRNA-Disease associations, interaction data between circRNAs and diseases from multiple resources integrate with circRNA similarity and disease similarity data to reconstruct a heterogeneous network in which the association between circRNAs and disease is hidden and should be mine. The basic concept in these methods is that similar circRNAs may be associated with similar diseases. In these approaches, more accurate data integration causes more accurate results. Similarly, this article proposed a novel algorithm called CircWalk to accurately extract potential CircRNA-Disease associations from a heterogeneous network based on a network representation algorithm. One of the most important circRNAs functions is acting as a miRNA sponge based on the ceRNA hypothesis. Many circRNAs are associated with diseases based on this mentioned mechanism. Our proposed method tried to integrate data based on the ceRNA hypothesis to reconstruct the heterogeneous network. Our results demonstrated that this strategy could predict more accurate CircRNA-Disease associations compared with other algorithms.

Methods

Our approach consists of three stages: In the first step, we merged data from multiple sources to reconstruct an informative heterogeneous network (Network reconstruction step). Next, we used the DeepWalk [34] algorithm to convert each circRNA and disease in this graph to a feature vector (Feature extraction step). At this stage, we have two feature vectors and a label (0 for unrelated and 1 for related pairs) for each CircRNA-Disease pair. We then train a classifier on this labeled dataset to create a model to predict CircRNA-Disease relationships accurately. Figure 1 shows the overall process of our algorithm.

Data sources

Considering the ceRNA hypothesis and miRNA sponge activity of circRNAs, we merged multiple bipartite networks from multiple experimentally validated datasets to reconstruct the above-mentioned heterogeneous network. Seven types of bipartite networks were used to reconstruct the final network:

1. CircRNA-Disease: the data in Circ2Disease [35], CircR2Disease [36], CTD [37], and CircAtlas [38] were merged to generate CircRNA-Disease interactions.

2. circRNA-circRNA: we calculated the alignment scores between every two circRNAs in our data and regarded them as a similarity measure among circRNAs. Next, we set the average score of all pairs as a cutoff threshold. After that, the circRNA pairs whose similarity score was more significant than this threshold was considered circRNA-circRNA networks for further analysis. Human circRNA sequence data were downloaded from the CircBase database [39], and built-in functions calculated the similarity scores from the BioPython package [40].

3. circRNA-miRNA: we extracted the circRNA-miRNA interaction data by combining the pairs from experimentally validated data in RAID [41] and StarBase [42].

4. miRNA-disease: the experimentally validated data in Circ2Disease [35], HMDD [43], and Mir2Disease [44] were used to generate miRNA-disease interactions.

5. miRNA-mRNA: miRTarbase [45], Circ2Disease [35], and StarBase [42] were used to extract miRNA-mRNA bipartite network.

6. mRNA-disease: The experimentally validated data in DisGeNet [46] was used to obtain mRNA-disease associations.

7. Disease-disease: Finally, we use the tree structure of diseases in the MeSH [47] database for the disease-disease similarity network. We calculate the semantic similarity between each pair of diseases in our data, and all the similarities above than specific threshold (0.8) were considered disease-disease pairs. The method proposed by Wang et al. [48] in the pyMeSHSim python library [49] was utilized to calculate the semantic similarity.

Network reconstruction

To reconstruct our final network, we merged all the above-mentioned bipartite networks. One of the important points in this step is to unify the genes and diseases identifiers in all the networks before merging. Different disease datasets may use different names for the same disease, e.g., “hepatocellular cancer” and “hepatocellular carcinoma." Such variations, however slight, matter since we organize and refer to our entries using their names. Therefore, we checked and unified the disease names in CircRNA-Disease data and set them as a reference for other datasets. Similarly, circRNAs have multiple naming systems in various datasets. To avoid duplication, we used the CircBase dataset [39] as a reference to unify circRNA names in our data (the circRNAs that have not been specified in CircBase were eliminated from the data). The mRNA and miRNA identifiers were the same in all data sources and didn’t need to unify.

Finally, we generated a heterogeneous network in which nodes are circRNA, mRNA, miRNA, and disease, and the edges are their relationships based on previously mentioned data sources. Note that, based on the ceRNA hypothesis and the sponge effect of circRNAs, the purpose of adding mRNAs and miRNAs to this network is to factor in the paths through which circRNAs indirectly influence diseases, in contrast to their direct associations. Consequently, if a miRNA or mRNA does not lie on any path from a circRNA to a disease, the edges that involve it may be left out with no side effects.

Feature extraction

The DeepWalk model takes a graph as its nodes' input and outputs k-dimensional embeddings [34]. We employed this method to extract features for our circRNAs and diseases. Therefore, we applied DeepWalk on the network generated in the previous step and extracted a k-dimensional feature vector for each circRNA and Disease in the network. Since DeepWalk uses random walks (paths) on the graph to learn

the embeddings of the nodes, we believe that adding new paths through mRNAs and miRNAs to the CircRNA-Disease graph can improve the performance of CircRNA-Disease associations prediction.

Binary classification of CircRNA-Disease Pairs

As a result of the previous step, we have a feature vector with a size of 2k for each pair of CircRNA-Disease. Besides, there is a class label for each pair: 0 means the circRNA is unrelated to the disease, and 1 implies the circRNA is related to the disease. Consequently, we can define a dataset and learn a classifier to predict the label of each input pair. To this end, we generated a benchmark dataset (see the result section). We applied 5-fold cross-validation based on multiple classifiers to evaluate the performance of extracted features from the heterogeneous network to predict the disease-related circRNAs. Six classification algorithms were used in this step: Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), AdaBoost with Random Forest base classifier (ABRF), XGBoost (XGB), and Multilayer Perceptron (MP) [50]. All classifiers were applied to the data using the scikit-learn Python package [51] (For non-default classifier hyperparameters, see supplementary Table 1).

Results

Evaluation metrics

The following evaluation metrics with 5-fold cross-validation were used to evaluate the performance of our algorithm and compare it with some other state-of-the-art algorithms. For simplicity, we use the abbreviations TP, FP, TN, and FN for true positive, false positive, true negative, and false negative, respectively.

1. AUC: AUC or area under Receiver Operating Characteristic (ROC) curve was the primary scoring metric we used in comparing models against each other. To obtain this, we need to calculate the area under a plot with points whose x coefficients are the false positive rates (FPR) of the model examined and whose y coefficients are the true positive rates (TPR) of that same model for different classification thresholds. The formulas to obtain FPR and TPR are shown below.

$$FPR = \frac{FP}{FP + TN}$$

$$TPR = \frac{TP}{TP + FN}$$

2. Accuracy (Acc): The ratio of correctly classified samples to all samples can be calculated as follows:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

3. Precision (Pre): The ratio of true positive samples to all samples labeled as positive. It can be calculated as follows:

$$Pre = \frac{TP}{TP + FP}$$

4. Sensitivity (Sen): The ratio of true positive samples to all ground truth (also known as Recall). It can be calculated based on the TPR formula.

5. F1 Score: It is the geometric mean of Precision and Recall and calculated as follows:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

6. Specificity (Spe): The ratio of true negative samples to all ground truth negative samples. It is calculated as follows:

$$Spe = \frac{TN}{TN + FP}$$

Benchmark dataset

To evaluate our method, we need a labeled set of CircRNA-Disease pairs as a benchmark dataset, wherein the label is 1 if the pair are associated and 0 otherwise. The labels will later be used for supervised binary classification. To create the benchmark dataset, we adopted the approach in [52], in which an equal number with the positive samples were randomly selected from unknown pairs as negative samples. Our dataset has 575 known circRNA-Disease pairs reconstructed from 474 unique circRNAs and 64 unique diseases. Hence, there are 474 × 64 = 30336 possible CircRNA-Disease combinations, 474 × 64–575 = 29761 of which are possibly unrelated. We randomly select 575 pairs from them as our negative samples (label = 0). This approach allows us to have a balanced dataset and reduces the probability of having false negatives (i.e., CircRNA-Disease pairs that are really associated but whose associations have not been discovered yet) by a factor of $\frac{575}{29761} \cong 1.93\%$.

Evaluate classification methods

For each classifier, the evaluation statistics depend on the number of features of circRNAs and diseases in the CircRNA-Disease dataset fed to them as input. Therefore, using DeepWalk, we created a set of feature vectors with different vector sizes (a multiple of 10 ranging from 10 through 200). We obtained the classification results on the benchmark dataset for each classifier and found the optimal number of features in terms of AUC. Overall, the best result we produced was achieved by the XGB and ABRF classifiers. Figure 2 shows how the average AUC of each classifier changes with the number of features extracted by DeepWalk. The optimal number of features for each classifier was considered for further evaluation.

Table 1 shows the values of the evaluation metrics for our six classifiers based on their optimal number of features. This table shows that SVM and LR have the minimum performance in our experiment with an average accuracy of 72 and 71, respectively. Overall, it seems that the boosting algorithms enjoy better

performance compared with the others. Random forest shows the appropriate performance as well. In terms of accuracy, the random forest has the best result after XGBoost, but if we consider AUC, its effect is very close to AdaBoost. We employed AdaBoost to improve the random forest model results, but as you can see in the table, the results of these two approaches are very close. The XGBoost algorithm obtained the best result. We chose this algorithm as the classifier in our final pipeline. It is noteworthy, however, that the training time of the XGBoost classifier is by far longer than AdaBoost and random forest.

Table 1
The values of evaluation metrics for different classifiers based on their optimal number of features.

Classifier (optimal feature vector size)	Fold	Acc (%)	F1 (%)	Pre (%)	Sen (%)	Spe (%)	AUC (%)
ABRF (10)	1	89.57	89.09	93.33	85.22	93.91	96.79
	2	88.70	88.29	91.59	85.22	92.17	95.58
	3	90.43	90.43	90.43	90.43	90.43	97.40
	4	90.87	90.58	93.52	87.83	93.91	96.19
	5	89.13	88.48	94.12	83.48	94.78	96.96
	Ave	89.74	89.37	92.60	86.44	93.04	96.58
LR (80)	1	71.30	70.54	72.48	68.70	73.91	77.64
	2	71.30	71.05	71.68	70.43	72.17	77.04
	3	74.78	75.63	73.17	78.26	71.30	81.07
	4	70.00	69.33	70.91	67.83	72.17	75.26
	5	69.13	70.04	68.03	72.17	66.09	76.42
	average	71.30	71.32	71.25	71.48	71.13	77.48
MP (10)	1	90.87	90.83	91.23	90.43	91.30	96.56
	2	90.43	90.83	87.20	94.78	86.09	93.44
	3	87.83	87.72	88.50	86.96	88.70	96.43
	4	88.26	88.41	87.29	89.56	86.96	94.82
	5	90.43	90.18	92.66	87.83	93.04	96.47
	Ave	89.56	89.59	89.37	89.91	89.22	95.54
RF (10)	1	89.13	88.69	92.45	85.21	93.04	97.20
	2	88.70	88.39	90.83	86.09	91.30	95.74
	3	91.74	91.63	92.86	90.43	93.04	96.54
	4	90.43	90.09	93.46	86.957	93.913	95.558
	5	90.43	90.09	93.46	86.96	93.91	97.17
	Ave	90.09	89.78	92.61	87.13	93.04	96.44
XGB (20)	1	90.87	90.83	91.23	90.4	91.30	97.99
	2	93.04	92.92	94.59	91.30	94.78	97.76

Classifier (optimal feature vector size)	Fold	Acc (%)	F1 (%)	Pre (%)	Sen (%)	Spe (%)	AUC (%)
	3	90.00	90.21	88.33	92.17	87.83	96.82
	4	92.17	92.10	92.92	91.30	93.04	97.49
	5	94.35	94.32	94.74	93.91	94.78	98.78
	Ave	92.09	92.078	92.36	91.82	92.35	97.77
SVM (90)	1	70.87	71.73	69.67	73.91	67.83	75.86
	2	73.04	74.17	71.20	77.39	68.70	79.10
	3	77.83	79.52	73.88	86.09	69.56	81.65
	4	69.13	70.04	68.03	72.17	66.09	72.77
	5	69.56	70.83	68.00	73.91	65.22	72.66
	Ave	72.09	73.26	70.16	76.69	67.48	76.41

The permutation of the samples in the cross-validation folds was identical for all six classifiers. Figure 3 shows a comparison of ROC curves of different classifiers for each fold of the data in the 5-fold CV. Other algorithms outperformed SVM and logistic regression with an approximate gap of 20% in terms of all six metrics. Not to mention that the SVM took the longest training time of all models. The multilayer perceptron was superior to SVM and logistic regression but missed out on the others by a margin of about 1%. XGBoost had the best AUC in 4 of the 5-folds of the dataset.

Comparison with existing methods

We compared CircWalk with three state-of-the-art algorithms based on the benchmark dataset: DMFCDA (Deep Matrix Factorization CircRNA-Disease Association) [53], GCNCDA [54], and SIMCCDA (Speedup Inductive Matrix Completion for CircRNA-Disease Associations prediction) [55].

Table 2 shows the evaluation process results for the selected algorithms and our method based on the benchmark dataset. As shown in this table, CircWalk is the most outperforming algorithm in our experiment, and its average values for all evaluation metrics are larger than 90%. After CircWalk, DMFCDA obtained higher accuracy compared with the others. GCNCDA is the most similar algorithm to our method among these comparison methods. Although this approach shows lower accuracy than CircWalk and DMFCDA, it is more stable and shows approximately the same results in all folds. SIMCCDA has acceptable performance in all metrics except precision and F1. This algorithm accurately predicted the negative class (unassociated CircRNA-Disease pairs), but its true positive rate was very low.

Table 2
evaluation metrics values for our algorithm and three other state-of-the-art algorithms based on the benchmark dataset.

Algorithm	Fold	Acc (%)	F1 (%)	Pr (%)	Se (%)	Sp (%)	AUC (%)
CircWalk	1	90.87	90.83	91.23	90.43	91.30	97.99
	2	93.04	92.92	94.59	91.30	94.78	97.76
	3	90.00	90.21	88.33	92.17	87.83	96.82
	4	92.17	92.10	92.92	91.30	93.04	97.49
	5	94.35	94.32	94.74	93.91	94.78	98.78
	Ave	92.09	92.08	92.36	91.83	92.35	97.77
DMFCDA	1	77.83	77.83	71.92	91.30	64.35	77.83
	2	80.00	80.00	79.49	80.87	79.13	80.00
	3	88.26	88.26	87.29	89.57	86.96	88.26
	4	86.84	86.84	85.59	88.60	85.09	86.84
	5	85.53	85.53	83.47	88.60	82.46	85.53
	Ave	83.69	83.69	81.55	87.79	79.60	83.69
GCNCDA	1	73.48	73.59	73.28	73.91	73.04	82.17
	2	76.09	76.79	74.59	79.13	73.04	82.92
	3	73.91	73.45	74.77	72.17	75.65	82.45
	4	74.35	74.24	74.56	73.91	74.78	84.62
	5	74.78	76.42	71.76	81.74	67.83	81.42
	Ave	74.52	74.90	73.79	76.17	72.87	82.72
SIMCCDA	1	78.04	11.88	06.41	81.82	77.97	68.82
	2	82.80	16.73	09.30	83.20	82.79	71.30
	3	86.96	18.33	10.25	86.41	86.97	76.80
	4	84.00	18.44	10.35	84.64	83.98	73.91
	5	85.00	16.64	09.20	86.67	84.97	75.45
	Ave	83.36	16.40	09.10	84.54	83.34	73.30

Figure 4 compares the ROC curve of each algorithm in each fold of the validation. As shown in this figure, CircWalk obtained an AUC of more than 96% (about 97% on average). GCNCDA and DMFCDA have

almost the same results, and SIMCCDA has the worst results in our experiment (because of its low true positive rate).

Case study

This step aims to evaluate the performance of CircWalk in the prediction of novel CircRNA-Disease associations in some selected common diseases. To this end, we selected three common cancer (lung, gastric, and colorectal) that are the target of many circRNA-related kinds of research. We train our model on the feature vectors of the positive pairs and a third of the negative pairs. As we pointed out earlier, the negative pairs (i.e., associations) are a subset of unverified CircRNA-Disease associations, which means there may be positive associations among them. As a result, we decided to train our model on a few negative pairs as possible to reduce learning from these false negatives. However, we could not completely omit them as there must be at least two classes in the dataset for XGBoost to be trained on it. Then, we make a list of all CircRNA-Disease pairs whose circRNA is present in our initial CircRNA-Disease dataset and whose disease is one of the three diseases we selected in this part. After that, filter out the CircRNA-Disease pairs present in the data, which our model was trained on in this part. We give this list of CircRNA-Disease associations as input to our trained model. Instead of labeling them as positive (1) or negative (0), we use our model to calculate the probability of association in each pair. Finally, for each disease, we find the circRNAs that are most probable to be associated with that disease and investigate the existing literature in PubMed to check if empirical studies have already confirmed that CircRNA-Disease association. Table 3 shows the result of this investigation.

Table 3
Predicted CircRNA-Disease relations with the highest probability for some selected diseases.

Disease	circRNA	Probability	Related article (PMID)
Lung cancer	hsa_circ_0007534	0.996	30017736
	hsa_circ_0001946	0.995	31249811
	hsa_circ_0002874	0.992	33612481
	hsa_circ_0014130	0.991	29440731, 31241217, 31818066, 32060230, 32616621, 34349347
	hsa_circ_0002702	0.990	32962802
	hsa_circ_0007874	0.988	30975029
	hsa_circ_0074930	0.985	32962802
	hsa_circ_0086414	0.983	30777071
	hsa_circ_0079530	0.972	29689350
	hsa_circ_0007385	0.972	29372377, 32602212, 32666646
	hsa_circ_0016760	0.968	29440731
	hsa_circ_0012673	0.960	29366790, 32141553
	hsa_circ_0067934	0.954	33832139
	hsa_circ_0000567	0.950	32328186, 33768996, 34435479
	hsa_circ_0072088	0.941	32308427, 34135596
	hsa_circ_0001727	0.934	32010565
	hsa_circ_0008305	0.901	30261900
gastric cancer	hsa_circ_0001313	0.999	32253030
	hsa_circ_0004771	0.998	29098316
	hsa_circ_0002874	0.998	34388244
	hsa_circ_0000615	0.998	34049561
	hsa_circ_0006404	0.977	32445925
	hsa_circ_0001982	0.977	33000178
	hsa_circ_0032683	0.910	33449227
	hsa_circ_0014130	0.819	32190005
colorectal cancer	hsa_circ_0006054	0.995	30585259

Disease	circRNA	Probability	Related article (PMID)
	hsa_circ_0000745	0.990	28974900
	hsa_circ_0044556	0.989	32884449
	hsa_circ_0005075	0.964	31081084, 31476947, 34015582
	hsa_circ_0040809	0.958	34438465
	hsa_circ_0004771	0.945	31737058, 32419229
	hsa_circ_0007874	0.924	32419229
	hsa_circ_0080210	0.914	34222420

As shown in Table 3, all the predicted pairs (except gastric cancer) had a probability of larger than 90 percent. There is much experimentally validated evidence in the result of this step. For instance, CircWalk predicted an association between hsa_circ_0001313 and gastric cancer with a probability of almost 100 percent. Based on a recent study by Zhang et al. [56], this circRNA is a key regulator in drug resistance in gastric cancer. CircRNA hsa_circ_0007534 (predicted by a probability of 99.6 percent) is an important oncogene in lung cancer related to cancer cell proliferation and apoptosis [57]. Another example is the association between hsa_circ_0044556 and colorectal cancer (predicted by a probability of 98.9 percent). Knocking down this circRNA cause inhibit proliferation, migration, and invasion of colorectal cancer cells [58]. These results represent the power of CircWalk to predict truly novel CircRNA-Disease associations.

Discussion

This study tried to integrate multiple data from multiple resources about genes and disease interactions to predict more significant CircRNA-Disease associations. Although biological data generation technologies have been advanced in recent years, this data type is primarily incomplete and has false positives. Data Integration can be a key approach to reducing noise and false positives. On the other, biological events are closely related to each other and work as a system. The cause of many disorders in the human body only can be explained using this systematic view of the biological processes. Therefore, our main idea to solve the problem of our study is integrating multiple data in a complex network and trying to extract associations between the circRNAs and diseases through the features of the network. Another key point in our approach involves the concept of the ceRNA hypothesis and miRNA sponge effect of circRNAs to predict their associations with diseases. The results of our study demonstrated that this point of view could be helpful to predict CircRNA-Disease associations more accurately.

One of the most challenging steps in our study was preparing the data. Each dataset uses its identifier for circRNAs, and converting these identifiers sometimes can be impossible. So, defining a standard for naming this type of RNA and creating a comprehensive database is primarily needed. Another challenge that can significantly affect the result of algorithms is the lack of validated negative class (un-associated pairs) for the CircRNA-Disease associations. As we mentioned in previous sections, we generated the

negative class by randomly selected circRNAs and Diseases that are un-associated in our data. But there is no guarantee for confirming that there is no association between these selected negative pairs in reality. This misinformation can affect the learning process of the classifiers and lead to generating inappropriate models. Therefore, creating standard datasets and benchmarks to validate the models accurately is one of the ideal future works in this field of study.

It is necessary to keep in mind that the sponging effect of circRNAs is not the only biological aspect that can help predict their association with a disease. The other biological information can be involved to solve the problem. For instance, their expression data, their exonic or intronic structure, the miRNA response elements information related to their sequence, and any other information about their structure and function can be helpful to associate them to a disease provided that the related data be accessible. Furthermore, novel machine learning approaches such as deep learning and graph convolutional neural networks can integrate multiple data and extract meaningful features from them.

Declarations

Acknowledgments

Not applicable

Authors' contributions

MK: conceptualization, funding acquisition, data curation, result analysis, methodology, writing, review & editing. EK: formal analysis, writing original draft, programming, visualization. BM: result analysis. HRR: conceptualization, supervision, project administration, funding acquisition, review & editing.

Funding

This work has been supported by Iran National Science Foundation (INSF) Grant No. 96006077.

Availability of data and materials

The datasets generated and/or analyzed during the current study are available in the CircWalk repository, at https://github.com/bcb-sut/CircWalk, along with the source code. As for the raw data, the twelve datasets we used can be found at the following web addresses, respectively: Circ2Disease at http://bioinformatics.zju.edu.cn/Circ2Disease, CircR2Disease at http://bioinfo.snnu.edu.cn/CircR2Disease, CTD at https://ctdbase.org, circAtlas at http://circatlas.biols.ac.cn/, circBase at http://www.circbase.org/, RAID at http://www.rna-society.org/404.shtml, starBase at http://www.sysu.edu.cn/403.html, HMDD at http://www.cuilab.cn/hmdd, miR2Disease at http://www.mir2disease.org/, miRTarBase at

https://mirtarbase.cuhk.edu.cn/~miRTarBase/miRTarBase_2022/php/index.php, DisGeNET at https://www.disgenet.org/, and MeSH at https://www.nlm.nih.gov/mesh/meshhome.html.

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests.

References

1. Huang, M., et al., Comprehensive analysis of differentially expressed profiles of lncRNAs and circRNAs with associated co-expression and ceRNA networks in bladder carcinoma. Oncotarget, 2016. 7(30): p. 47186–47200.
2. Kristensen, L.S., et al., *The biogenesis, biology and characterization of circular RNAs*. Nature Reviews Genetics, 2019. 20(11): p. 675–691.
3. Jeck, W.R., et al., Circular RNAs are abundant, conserved, and associated with ALU repeats. Rna, 2013. 19(2): p. 141–157.
4. Memczak, S. et al., Circular RNAs are a large class of animal RNAs with regulatory potency. Nature, 2013. 495(7441): p. 333–338.
5. Zhang, Z., T. Yang, and J. Xiao, *Circular RNAs: promising biomarkers for human diseases*. EBioMedicine, 2018. 34: p. 267–274.
6. Lei, B., et al., Circular RNA: a novel biomarker and therapeutic target for human cancers. Int J Med Sci, 2019. 16(2): p. 292–301.
7. Geng, X., et al., Circular RNA: Biogenesis, degradation, functions and potential roles in mediating resistance to anticarcinogens. Epigenomics, 2020. 12(3): p. 267–283.
8. Mitra, A., K. Pfeifer, and K.S. Park, *Circular RNAs and competing endogenous RNA (ceRNA) networks*. Transl Cancer Res, 2018. 7(Suppl 5): p. S624-S628.
9. Salmena, L., et al., A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? Cell, 2011. 146(3): p. 353–8.
10. Zhong, Y., et al., Circular RNAs function as ceRNAs to regulate and control human cancer progression. Molecular cancer, 2018. 17(1): p. 1–11.

11. Verduci, L., et al., CircRNAs: role in human diseases and potential use as biomarkers. Cell Death Dis, 2021. 12(5): p. 468.
12. Altesha, M.A., et al., *Circular RNA in cardiovascular disease*. Journal of cellular physiology, 2019. 234(5): p. 5588–5600.
13. Xu, H., et al., The circular RNA Cdr1as, via miR-7 and its targets, regulates insulin transcription and secretion in islet cells. Scientific reports, 2015. 5(1): p. 1–12.
14. Liang, H., et al., Serum hsa_circ_0054633 Is Elevated and Correlated with Clinical Features in Type 2 Diabetes Mellitus. Annals of Clinical & Laboratory Science, 2021. 51(1): p. 90–96.
15. Lu, Y.-K., et al., Identification of circulating hsa_circ_0063425 and hsa_circ_0056891 as novel biomarkers for detection of type 2 diabetes. The Journal of Clinical Endocrinology & Metabolism, 2021.
16. Wang, K. et al., A circular RNA protects the heart from pathological hypertrophy and heart failure by targeting miR-223. European heart journal, 2016. 37(33): p. 2602–2611.
17. Garikipati, V.N.S., et al., Circular RNA CircFndc3b modulates cardiac repair after myocardial infarction via FUS/VEGF-A axis. Nature communications, 2019. 10(1): p. 1–14.
18. Zhang, Y., et al., *Exploring the regulatory roles of circular RNAs in Alzheimer's disease*. Translational Neurodegeneration, 2020. 9(1): p. 1–8.
19. Bigarré, I.M., et al., IGF2R circular RNA hsa_circ_0131235 expression in the middle temporal cortex is associated with AD pathology. Brain and Behavior, 2021. 11(4): p. e02048.
20. Kristensen, L., et al., Circular RNAs in cancer: opportunities and challenges in the field. Oncogene, 2018. 37(5): p. 555–565.
21. Han, Y., et al., Exosomal circRNA 0001445 promotes glioma progression through miRNA-127-5p/SNX5 pathway. Aging (Albany NY), 2021. 13.
22. Wang, P. et al., hsa_circ_0062019 promotes the proliferation, migration, and invasion of prostate cancer cells via the miR-195-5p/HMGA2 axis. Acta Biochim Biophys Sin (Shanghai), 2021.
23. Ding, H., et al., Higher circular RNA_0015278 correlates with absence of extrathyroidal invasion, lower pathological tumor stages, and prolonged disease-free survival in papillary thyroid carcinoma patients. J Clin Lab Anal, 2021: p. e23819.
24. Gao, H., et al., Depletion of hsa_circ_0000144 Suppresses Oxaliplatin Resistance of Gastric Cancer Cells by Regulating miR-502-5p/ADAM9 Axis. Onco Targets Ther, 2021. 14: p. 2773–2787.
25. Luo, L., et al., Circ-ZFR Promotes Progression of Bladder Cancer by Upregulating WNT5A Via Sponging miR-545 and miR-1270. Front Oncol, 2020. 10: p. 596623.
26. Sui, C., et al., Hsa_circ_0069094 knockdown inhibits cell proliferation, migration, invasion and glycolysis, while induces cell apoptosis by miR-661/HMGA1 axis in breast cancer. Anticancer Drugs, 2021.
27. Gao, C. et al., Circ_0055625 knockdown inhibits tumorigenesis and improves radiosensitivity by regulating miR-338-3p/MSI1 axis in colon cancer. World J Surg Oncol, 2021. 19(1): p. 131.

28. Lei, X.-J., C. Bian, and Y. Pan, *Predicting CircRNA-Disease Associations Based on Improved Weighted Biased Meta-Structure*. Journal of Computer Science and Technology, 2021. 36(2): p. 288–298.
29. Zhang, Y., et al., CircRNA-disease associations prediction based on metapath2vec + + and matrix factorization. Big Data Mining and Analytics, 2020. 3(4): p. 280–291.
30. Zeng, M., et al., Deep matrix factorization improves prediction of human circRNA-disease associations. IEEE Journal of Biomedical and Health Informatics, 2020.
31. Lu, C., et al., Improving circRNA-disease association prediction by sequence and ontology representations with convolutional and recurrent neural networks. Bioinformatics, 2020.
32. Chen, X., et al., A novel approach based on KATZ measure to predict associations of human microbiota with non-infectious diseases. Bioinformatics, 2017. 33(5): p. 733–739.
33. Deng, L., et al., Fusion of multiple heterogeneous networks for predicting circRNA-disease associations. Scientific reports, 2019. 9(1): p. 1–10.
34. Perozzi, B., R. Al-Rfou, and S. Skiena. Deepwalk: Online learning of social representations. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. 2014.
35. Yao, D., et al., Circ2Disease: a manually curated database of experimentally validated circRNAs in human disease. Scientific reports, 2018. 8(1): p. 1–6. https://doi.org/10.1038/s41598-018-29360-3
36. Fan, C., et al., CircR2Disease: a manually curated database for experimentally supported circular RNAs associated with various diseases. Database, 2018. 2018. https://doi.org/10.1093/database/bay044
37. Davis, A.P., et al., *Comparative Toxicogenomics Database (CTD): update 2021*. Nucleic acids research, 2021. 49(D1): p. D1138-D1143https://doi.org/10.1093/nar/gkaa891.
38. Wu, W., P. Ji, and F. Zhao, CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes. Genome biology, 2020. 21: p. 1–14 https://doi.org/10.1186/s13059-020-02018-y.
39. Glažar, P., P. Papavasileiou, and N. Rajewsky, *circBase: a database for circular RNAs*. Rna, 2014. 20(11): p. 1666–1670 https://doi.org/10.1261/rna.043687.113.
40. Cock, P.J., et al., Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 2009. 25(11): p. 1422–1423.
41. Yi, Y., et al., RAID v2. 0: an updated resource of RNA-associated interactions across organisms. Nucleic acids research, 2017. 45(D1): p. D115-D118 https://doi.org/10.1093/nar/gkw1052.
42. Yang, J.-H., et al., starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic acids research, 2011. 39(suppl_1): p. D202-D209 https://doi.org/10.1093/nar/gkq1056.
43. Huang, Z., et al., HMDD v3. 0: a database for experimentally supported human microRNA–disease associations. Nucleic acids research, 2019. 47(D1): p. D1013-D1017 https://doi.org/10.1093/nar/gky1010.

44. Jiang, Q., et al., miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic acids research, 2009. 37(suppl_1): p. D98-D104 https://doi.org/10.1093/nar/gkn714.
45. Huang, H.-Y. et al., miRTarBase 2020: updates to the experimentally validated microRNA–target interaction database. Nucleic acids research, 2020. 48(D1): p. D148-D154 https://doi.org/10.1093/nar/gkz896.
46. Piñero, J., et al., DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic acids research, 2016: p. gkw943 https://doi.org/10.1093/nar/gkw943.
47. Lipscomb, C.E., *Medical subject headings (MeSH)*. Bulletin of the Medical Library Association, 2000. 88(3): p. 265.
48. Wang, J.Z., et al., *A new method to measure the semantic similarity of GO terms*. Bioinformatics, 2007. 23(10): p. 1274–1281.
49. Luo, Z.-H., et al., pyMeSHSim: an integrative python package for biomedical named entity recognition, normalization, and comparison of MeSH terms. BMC bioinformatics, 2020. 21(1): p. 1–14.
50. Adnan, M.N. and M.Z. Islam, Forest PA: Constructing a decision forest by penalizing attributes used in previous trees. Expert Systems with Applications, 2017. 89: p. 389–403.
51. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python*. The Journal of Machine Learning Research, 2011. 12: p. 2825–2830.
52. Wang, L., et al., GCNCDA: A new method for predicting circRNA-disease associations based on Graph Convolutional Network Algorithm. PLOS Computational Biology, 2020. 16(5): p. e1007568.
53. Lu, C. et al., *Deep matrix factorization improves prediction of human circRNA-disease associations*. IEEE Journal of Biomedical and Health Informatics, 2020. 25(3): p. 891–899.
54. Wang, L., et al., GCNCDA: A new method for predicting circRNA-disease associations based on Graph Convolutional Network Algorithm. PLoS Comput Biol, 2020. 16(5): p. e1007568.
55. Li, M., et al., Prediction of circRNA-disease associations based on inductive matrix completion. BMC medical genomics, 2020. 13(5): p. 1–13.
56. Zhang, Q., et al., CircRNACCDC66 regulates cisplatin resistance in gastric cancer via the miR-618/BCL2 axis. Biochem Biophys Res Commun, 2020. 526(3): p. 713–720.
57. Qi, Y. et al., Upregulation of circular RNA hsa_circ_0007534 predicts unfavorable prognosis for NSCLC and exerts oncogenic properties in vitro and in vivo. Gene, 2018. 676: p. 79–85.
58. Jing, L., et al., Identification of circular RNA hsa_circ_0044556 and its effect on the progression of colorectal cancer. Cancer Cell Int, 2020. 20: p. 427.

Figures

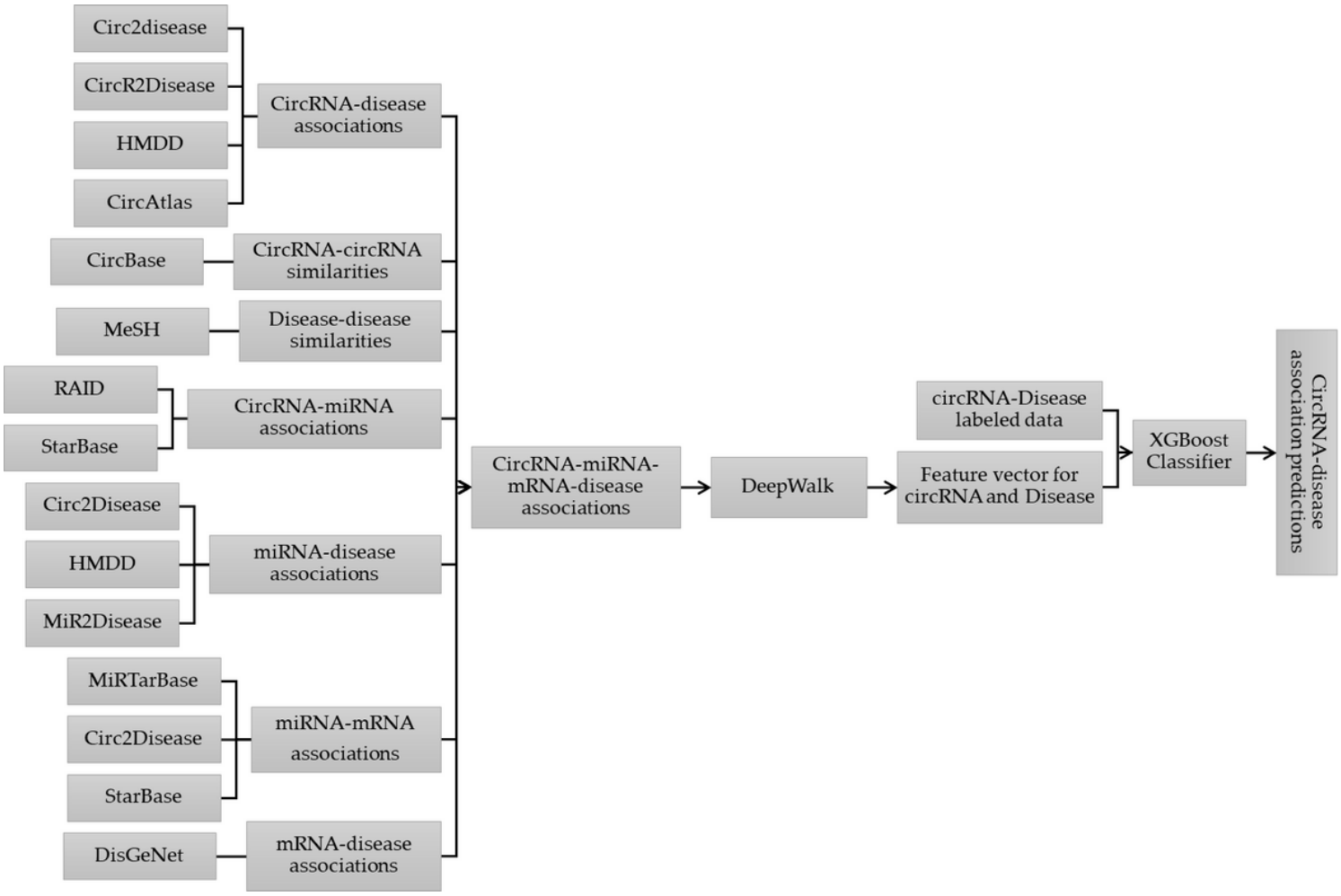

Figure 1

The overall workflow of the proposed algorithm.

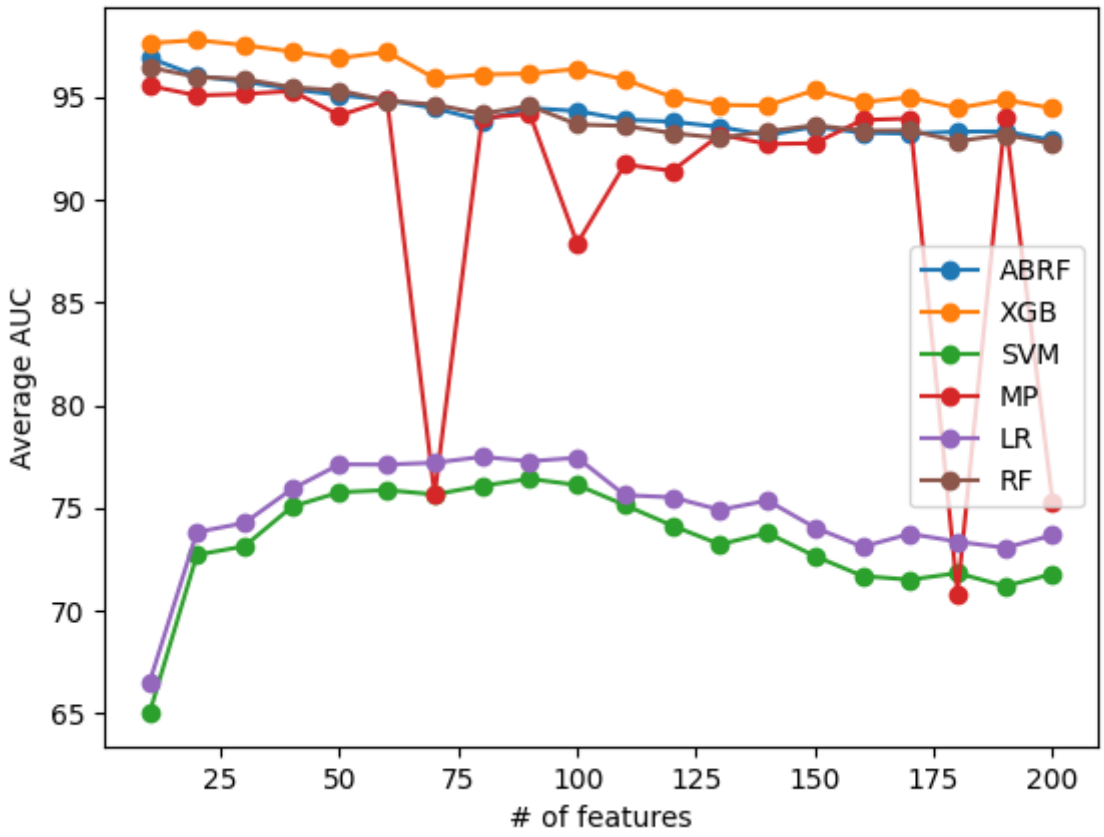

Figure 2

The average AUC of each classifier is based on the size of feature vectors extracted by DeepWalk.

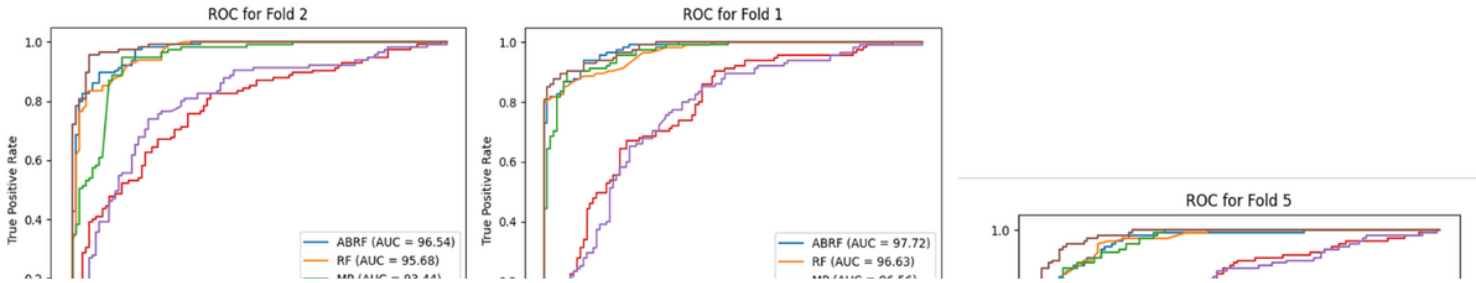

Figure 3

ROC curve and AUC for different classifiers in 5-fold cross-validation analysis (the size of extracted feature vector with Deepwalk set to the optimum value for each classifier).

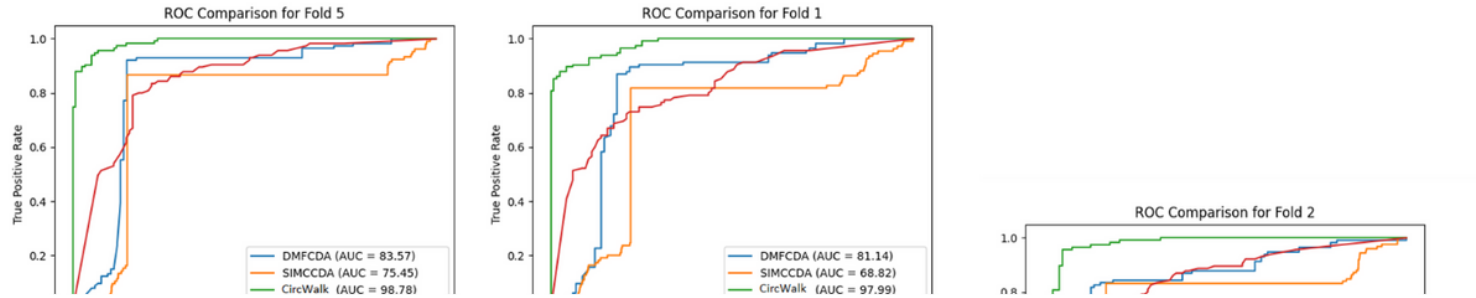

Figure 4

Comparison of ROC curve obtained by different algorithms in 5-fold cross-validation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.docx