Liu *et al. BMC Medical Genomics* 2020, **13**(Suppl 10):153
https://doi.org/10.1186/s12920-020-00783-0

BMC Medical Genomics

RESEARCH Open Access

Predicting miRNA-disease associations using a hybrid feature representation in the heterogeneous network

Minghui Liu[1†], Jingyi Yang[1†], Jiacheng Wang[1] and Lei Deng[1,2*]

From The 18th Asia Pacific Bioinformatics Conference
Seoul, Korea. 18-20 August 2020

Abstract

Background: Studies have found that miRNAs play an important role in many biological activities involved in human diseases. Revealing the associations between miRNA and disease by biological experiments is time-consuming and expensive. The computational approaches provide a new alternative. However, because of the limited knowledge of the associations between miRNAs and diseases, it is difficult to support the prediction model effectively.

Methods: In this work, we propose a model to predict miRNA-disease associations, MDAPCOM, in which protein information associated with miRNAs and diseases is introduced to build a global miRNA-protein-disease network. Subsequently, diffusion features and HeteSim features, extracted from the global network, are combined to train the prediction model by eXtreme Gradient Boosting (XGBoost).

Results: The MDAPCOM model achieves AUC of 0.991 based on 10-fold cross-validation, which is significantly better than that of other two state-of-the-art methods RWRMDA and PRINCE. Furthermore, the model performs well on three unbalanced data sets.

Conclusions: The results suggest that the information behind proteins associated with miRNAs and diseases is crucial to the prediction of the associations between miRNAs and diseases, and the hybrid feature representation in the heterogeneous network is very effective for improving predictive performance.

Keywords: miRNA-disease association, HeteSim measure, Diffusion feature, eXtreme gradient boosting

Background

MicroRNAs(miRNAs) are a kind of small single-stranded endogenous non-coding RNAs with a length about 22 nucleotides, which play an important role in regulating the gene expression during the post-transcriptional level [1, 2]. Many studies have shown that the dysregulation of miRNAs is involved in multiple human diseases like cancers [3], cardiovascular diseases [4] and Alzheimer's diseases [5] etc., and the prediction of miRNAs-diseases associations is crucial to understand the diseases pathogenesis [6]. Furthermore, George Adrian, et al. found that the miR15 and miR16 are deleted in a lot B cell chronic lymphocytic leukemias (B-CLL) [7], T. Sredni et al. demonstrated that miR-129 and miR-25 express abnormally in all pediatric brain tumor types [8]. Besides, Jun Lu et al. successfully classified poorly differentiated tumours using miRNA expression profiles [9], which demonstrated the potential of miRNAs as biomarkers. Therefore, Pre-

*Correspondence: leideng@csu.edu.cn
†Minghui Liu and Jingyi Yang contributed equally to this work.
[1]School of Computer Science and Engineering,Central South University, 410075, Changsha, China
[2]School of Software, Xinjiang University, 830008, Urumqi, China

dicting miRNA-disease associations is very meaningful. However, a lot of miRNA-disease associations remain unknown and experimental approaches for predicting the associations are time-consuming and expensive. Therefore, a lot computational methods have been developed to predict the miRNA-disease associations.

Computational methods can be grouped into two categories: network-based methods and machine learning-based methods. Network-based methods usually use similarity measurement to predict the associations. For example, Jiang et al. [10] presented a computational method to predict the associations between miRNAs and diseases by prioritizing entire human microRNAome according to the disease of interest. The higher the rank is, the more possibly the miRNA can associate with the disease. In 2010, the model was improved by introducing genomic data [11]. However, the performance of the model was still not satisfactory because the known target genes of miRNAs are too rare to support the methods effectively. Chen et al. develpoed a method called RWRMDA [12], the author ran random walk with restart algorithm on a miRNA functional similarity network to obtain a score for every miRNA, and the miRNA with a higher score is more likely to associate with a certain disease. Shi et al. [13] extended random walk with restart algorithm (RWR), they used proteins associated with diseases and miRNAs as seed nodes to calculate the ES score by RWR respectively, and then used the *P*-value to predict whether the disease and miRNA are related. PRINCE [14] is another algorithm optimized based on RWR, it proposed a novel method to initial probability of miRNAs. However, these methods, based on RWR, are dependent on known associations between miRNAs and a given disease, so it couldn't be applied to predict the relationships between miRNAs and a new disease, without any associations with miNRAs. Furthermore, defining a proper similarity calculation model is challengeable in this category.

The prediction models in another category are based on machine learning. For example, Xu et al. [15] extracted features from a miRNA-disease network, and then used the features to train a prediction model by support vector machine (SVM), the method can discover positive samples from massive negative samples. Chen et al. [16] presented a semi-supervised and global method RLSMDA, the method calculated possibilities of being associated with a given disease for each miRNA by a continuous classification function, and it could predict the associations of diseases and miRNAs without any known association between them. However, the method didn't integrate the information related to miRNAs and diseases completely since the continuous classification function is established for the miRNA network and the disease network separately.

Recently, more computational methods are proposed. Zheng et al. [17] developed a machine leaning-based method MLMDA, which used a variety of information including miRNA sequence information, miRNA functional similarity, disease semantic similarity and Gaussian interaction profile kernel similarity information to train their model by applying random forest classifier. The classifier achieved promising performance, but it might take a lot of effort to prepare the required data. What's more, the knowledge of deep learning was also applied in this field. Peng et al. [18] utilized a convolutional neural network to predict miRNA-disease association, input data was reduced miRNA-disease interaction features which were captured from a three-layer network. The similarity metric is essential in order to predict associations between miRNAs and diseases, where Yang et al. [19] used a novel method miRGOFS to measure functional similarities of miRNAs, and the method considered both common ancestors and descendants of GO terms when it was used to calculate similarities among GO sets in an asymmetric manner, so it can help predict the miRNA-disease associations. Chen et al. [20] presented the first decision tree, learning-based model, whose informative feature vectors were extracted from miRNA functional similarities, the disease semantic similarities, and known miRNA-disease associations. Yin et al. [21] put forward LWPCMF, they used weighted profile (WP) , collaborative matrix factorization (CMF) and logistic function to optimize their model.

In this work, we present a computational method named MDAPCOM to predict the associations between miRNAs and diseases by combined features. First, we construct a miRNA-protein-disease global network by merging six subnetworks, which are miRNA-miRNA Similarity Network, Protein-Protein Interaction Network, Disease-Disease Similarity Network, miRNA-Target Interaction Network, miRNA-Disease Relationship Network and Protein-Disease Association Network respectively. Subsequently, we extract diffusion features for each node and a 39-dimensional HeteSim feature for each miRNA-disease pair in the global network. The diffusion features are extracted by random walk with restart algorithm and then reduced in dimension using the singular value decomposition algorithm (SVD). Finally, we integrate these two features to train the miRNA-disease association prediction model using eXtreme Gradient Boosting (XGBoost) algorithm. We apply the MDAPCOM method under 10-fold cross-validation and achieve an AUC of 0.991. MDAPCOM also performs better when compared with other two previous methods RWRMDA [12] and PRINCE [14], which also used network features for prediction. Furthermore, our method performs well on three unbalanced data sets with positive and negative samples ratios 1:2, 1:5 and 1:10, respectively.

Results

Data sources

We collect six different types of data from the Internet, which are the miRNA-miRNA similarity data, miRNA-Protein interactions, miRNA-Disease relationships, PPI data (Protein-Protein interactions), Protein-Disease association data, Disease-Disease similarity data, respectively, containing 2588 miRNAs, 18143 proteins and 5080 different kinds of diseases totally.

miRNA-miRNA similarity network

We obtain miRNA expression data from miRmine database [22]. In this database, the researchers analyzed overall expression profile of human, obtained from different miRNA-seq databases. It contains 2822 different precursor miRNAs where more than two of them consist one mature miRNA, so we can derive the expression values of every mature miRNA from the average values of its precursors'. In this way, we obtain 2588 miRNA expression profiles. Moreover, the Pearson Correlation Coefficient (PCC) scores are calculated to preform similarities of the expression profiles between two miRNAs [23]. The higher the PPC score is, the more likely these two miRNAs are similar. The miRNA-miRNA Similarity network is also built. In the network, every miRNA is a node and the PPC scores present the edges, and the negative edges are cut down.

Protein-protein interaction network

We derive data from the STRING database V10.0 [24]. The database offers data which is obtained from the results of biochemical experiments, biophysical or genetic techniques. We get 7,866,428 PPI entries from 18,143 proteins in the database and use them to build our Protein-Protein Interaction Network. In PPI network, each of the entry comprises a protein node A, a protein node B, and the predicted relationship's score between them. The highest score means the two proteins can interact with each other with the biggest possibility and vice versa. Last, we utilize the predicted score to present the value of each edge between two protein nodes to construct our Protein-Protein Interaction Network.

Disease-disease similarity network

To build the Disease-Disease Similarity Network, we obtain data from the MimMine database. [25] It is mapped from OMIM database, containing more than 5000 human genetic disease phenotypes. It is worthy to point out that we normalize disease-disease similarities' values into [0,1] in MimMiner database. Subsequently, we receive 5080 kinds of diseases and get the similarities between them. Finally, we construct the Disease-Disease Similarity Network where each node presents a kind of disease, and the weight is similarity between them.

miRNA-target interactions network

We download miRNA-target interactions from the miRTarBase database of release 7.0 [26], miRNA-Target Interaction Network can be built. It should be point out all data is validated in this database. Moreover, we map the genes onto protein entries, and remove invalid entries (miRNA or protein), which are repeated and out-of-range. Finally, we extract miRNAtarget interactions between 2,588 miRNAs and 18,143 proteins. Then, miRNA-Target Interaction network is constructed based on these data.

miRNA-disease relationship network

We get miRNA-disease data from HMDD v3.0 database [27], which is a reliable online database containing 1102 gene on miRNA, 850 different types of diseases and 32281 associations between miRNA and disease, and they are all based on literature. Furthermore, we receive the relationships between 2588 miRNAs and 5080 diseases which are mentioned above. Lastly, we build the miRNA-Disease Relationship network using these validated data.

Protein-disease association network

We obtain data from DisGeNET database [25] which collects data on genotype-phenotype relationships. In this work, we map genes into proteins and unify the name of diseases, so 18,143 proteins ,5080 diseases and the associations between them are extracted. Then, we construct a Protein-Disease Association Network from these data.

Global heterogeneous network

We integrate the aforementioned networks to build the global heterogeneous network:

$$T = \begin{bmatrix} M & B & C \\ B^T & P & W \\ C^T & W^T & D \end{bmatrix}$$

where T represents our global heterogeneous network, M, P, D present similarity of miRNA-miRNA, protein-protein and disease-disease respectively, B presents the miRNA-Target Interaction Network, C indicates miRNA-Disease Relationship Network, and W shows the Protein-Disease Association Network. Obviously, the B^T, C^T and W^T are transposed matrices of B, C and W, and the edges with value less than 0.5 are removed from the network.

There are 2588 miRNAs and 5080 diseases in our miRNA-protein-disease global network, so we can get a total of 13147040 (2588×5080) miRNA-disease pairs. We extract a 639-dimensional combined feature vector for each miNRA-disease pair in the global network, in which 11824 feature vectors are positive samples while the other 13135216 feature vectors are negative samples. We randomly select 11824 feature vectors from 13135216 negative samples to construct a standard dataset together

with 11824 positive samples, subsequently, we execute 10-fold cross-validation on the standard dataset. The positive and negative samples are randomly divided into 10 subsamples equalled in size(the size of the tenth subsample is 1186 because 11824 is't divisible by 10), one of which is retained as the validation set and the other 9 subsamples are regarded as the training set. Then the procedure iterates 10 times with each one in the 10 subsamples as the validation set, before each iteration, the associations occurred in the validation set are removed from the original global network, and then all feature vectors are re-extracted from the new global network. Furthermore, another three unbalanced data sets are obtained in the same way except the size of the selected negative samples, and the size of negative samples in three unbalanced data sets is 23648, 59120 and 118240, respectively.

Performance measures

We apply 10-fold cross-validation, and obtain the average performance of our model through the performance evaluation. In terms of performance evaluation, we select precision(PRE), recall(REC), F-score(FSC), accuracy(ACC) and the area under the receiver operating characteristic curve(AUC):

$$PRE = \frac{TP}{TP + FP},$$
$$REC = \frac{TP}{TP + FN},$$
$$ACC = \frac{TP + TN}{TP + TN + FP + FN},$$
$$FSC = \frac{2 \times PRE \times REC}{PRE + REC},$$

TP and FP are the amount of correctly predicted positive and negative samples, FP and FN are the numbers of positive and negative samples predicted by mistakes. Simultaneously, we calculate the area under ROC curve (AUC) to measure the overall performance.

Excellent combined feature

In our method, we extract two different features from a global heterogeneous network, a global matrix of nine different data, and combine them to construct our training dataset. Firstly, with the help of random walk with restart algorithm, we extract diffusion feature of each node from our global network, so we can get a 20588*20588 feature matrix, where a row represents a feature vector of one node. For example, the first row shows the miRNA1's feature vector, the 2589 th row is the protein1's feature vector, and the 20732 th row is the disease1's feature vector. In the next step, we apply SVD algorithm on this feature matrix to reduce the dimension of it from 20588 to 300, here our feature matrix is 20588*300. After obtaining reduced feature vectors of each node, we combine each miRNA's feature vector with each disease's, so we get a (2588*5080) * 600 miRNA-disease feature matrix, where a row shows the feature vector of a pair of miRNA-disease. Secondly, we calculate HeteSim scores of each miRNA-disease pair, and get a (2588*5080) * 39 HeteSim matrix. Finally, in order to construct our training data, we joint the SVD feature and HeteSim score, so we get a (2588*5080) * 639 feature vector, where a row is the combined feature vector of a miRNA-disease pair. To show excellent performance of our method, we use diffusion features, the HeteSim feature and the combined feature to train the prediction model using 10-fold cross-validation under the standard data set, respectively, and the result shows in Fig. 1. The AUC value of training model with the diffusion feature and the HeteSim feature reach 0.970 and 0.986,

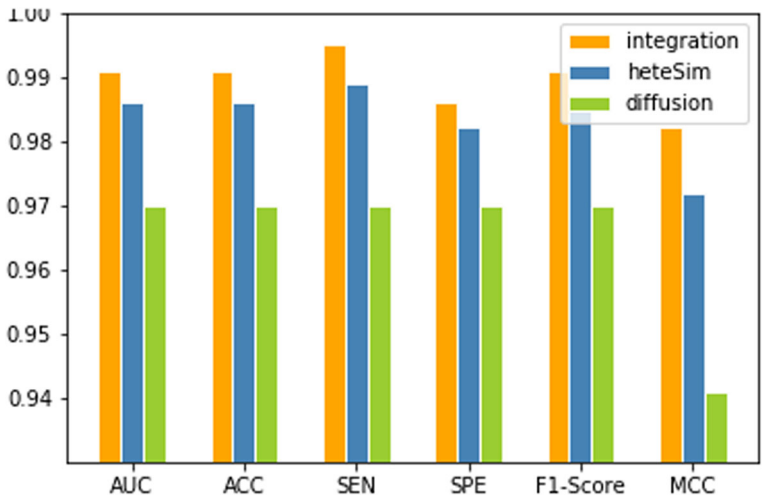

Fig. 1 Performance comparison of different feature groups (Diffusion, HeteSim and combined feature)

respectively, and we get an AUC of 0.991 using combined feature.

Superiority of XGBoost algorithm

In this work, we apply eXtreme Gradient Boosting(XGBoost) [28] algorithm to train our model. We compare XGBoost algorithm with other machine learning algorithm to present that the eXtreme Gradient Boosting(XGBoost) algorithm is the most suitable method for us. To achieve the goal, we obtain other classifiers from python toolkits scikit-learn and apply 10-fold cross-validation. We compare XGBoost algorithm with random forest (RF) [29], support vector machine (SVM) [30]and gradient tree boosting (GTB) [31] algorithm. In random forest algorithm, we set the minimize samples split to 42, maximize depth of tree to 11 and the resting parameter values to default. In the support vector machine algorithm, we use RBF kernel setting the C value to 100, gamma value to 0.0001. In gradient tree boosting algorithm, we set the minimize samples split to 110, the maximize depth of tree to 9. The results perform in Fig. 2.

Performance comparison with existing methods

We implement RWRMDA [12] and PRINCE [14] under a standard dataset and three unbalanced datasets, applying 10-fold cross-validation to calculate their AUC values and compare theirs with MDAPCOM's. For PRINCE, we set α=0.95, d=log (9999), c=-15 and then apply the random walk with restart 10 times. The probability of restarting in RWRMDA is set to 0.5. To visually describe and compare the performance of the three methods, we plot the Receiver Operating Characteristic (ROC) curve with its horizontal axis representing false positive rate (FPR) and the vertical axis representing true positive rate (TPR). Subsequently, we use the area under the ROC curve (AUC) to accurately compare the performance of the three methods. Figures 3, 4, 5 and 6 show the performance of the three methods under four datasets with different positive and negative ratios, respectively. Among three methods, MDAPCOM significantly outperforms the other two methods, achieving an amazing AUC score 0.99. Furthermore, the AUC scores of our method are all above 0.99 under four data sets, which proves its stability.

Conclusions

In this work, we present a prediction method based on machine learning to predict the associations between miRNAs and diseases, MDAPCOM. We build a miRNA-protein-disease global network, then extract dimensional reduced RWR diffusion feature and HeteSim feature from the network, the diffusion feature reflects the node topological information in the heterogeneous network and the HeteSim feature extracts the correlation of node pairs. Subsequently, the two features are combined to train the miRNA-disease association prediction model using 10-fold cross-validation by eXtreme Gradient Boosting (XGBoost). The MDAPCOM shows better performance than other two previous methods, based on network feature. The excellent performance suggests that the information behind proteins which are associated with miNRAs and diseases is crucial to predict associations between miRNAs and diseases. Furthermore, the two features extract network information from different perspectives and the combination of them integrates network information effectively, which also contributes to the excellent performance of the method.

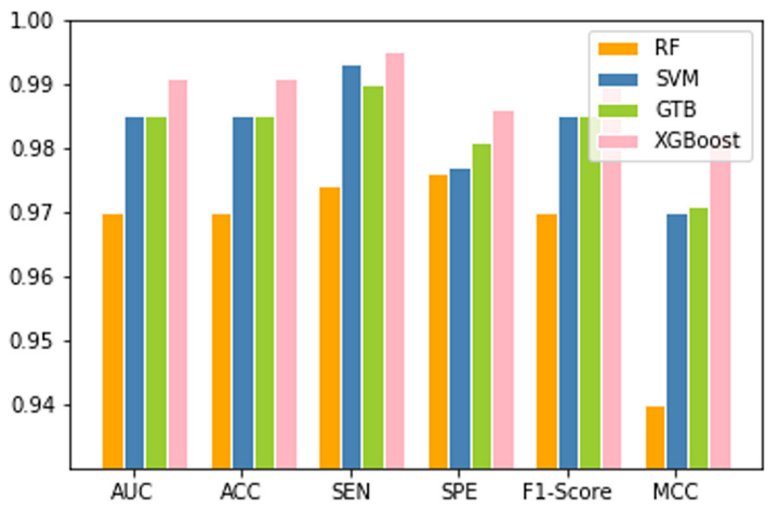

Fig. 2 Comparison result between XGBoost and other machine learning algorithms including PF, SVM and GTB

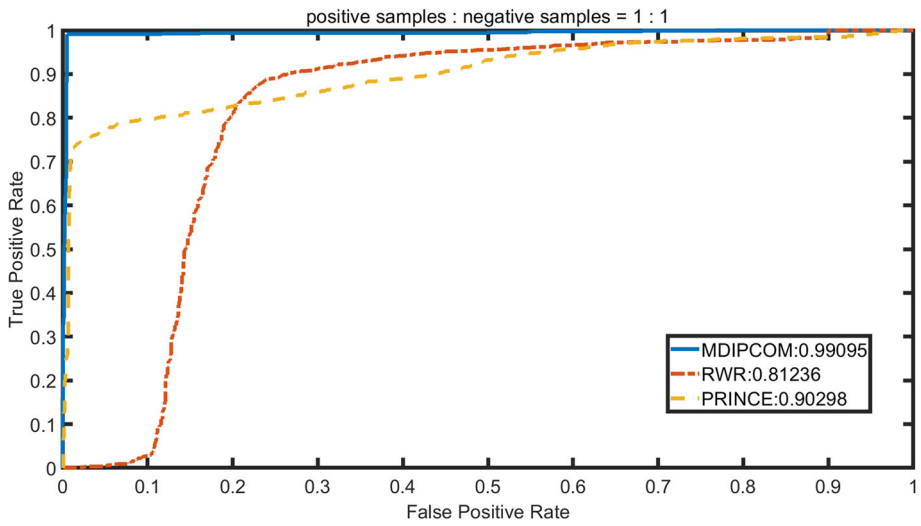

Fig. 3 The ROC curves of MDAPCOM and previous approaches containing PRINCE and RWRMDA on the data sets. The ratio of positive and negative sample 1:1

Methods

Overview of MDAPCOM

Our method is displayed in Fig. 7, which is built through following steps: (A) Collect six types of data sources and remove invalid and repeated data. (B) Merge the six networks to build a global miRNA-protein-disease heterogeneous network. (C) Run random walks with restart (RWR) algorithm in the global network to calculate a diffusion feature for every node, which reflects the relevance of one node with all other nodes (miRNAs, proteins and diseases) in the network (D) Run the singular value decomposition (SVD) algorithm to reduce dimension of the diffusion feature, obtaining a 300-dimensional feature vector for every node. (E) Use HeteSim measure to estimate the correlation between two nodes and get a 39-dimensional HeteSim feature for each miRNA-disease pair. (F) Integrate the 600-dimension diffusion feature(300-dimensional for miRNA and 300-dimensional for disease) and 39-dimensional HeteSim feature to train a miRNA-disease association prediction model by eXtreme Gradient Boosting (XGBoost).

Diffusion feature of reduced dimension

To predict the miRNA-disease associations, we transform the problem to obtain possibility that a miRNA can associate with a disease. The Random Walk with

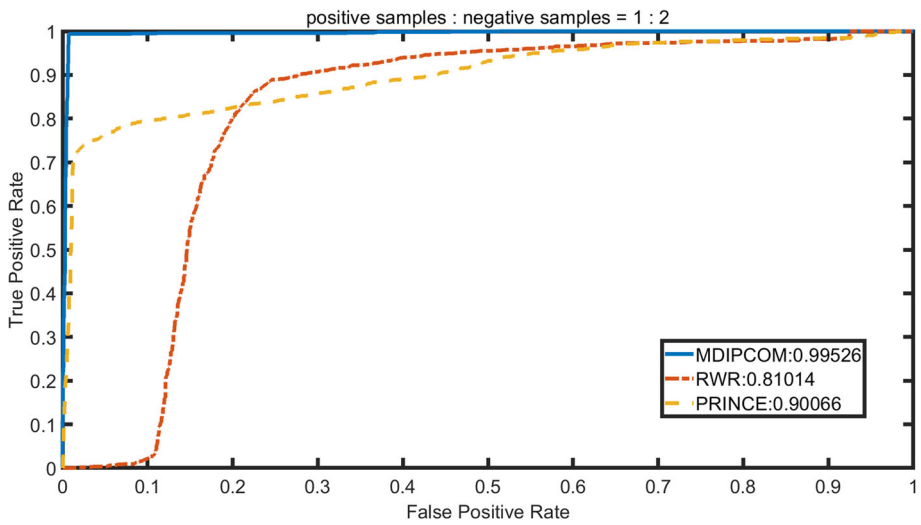

Fig. 4 The ROC curves of MDAPCOM and previous approaches containing PRINCE and RWRMDA on the data sets. The ratio of positive and negative sample 1:2

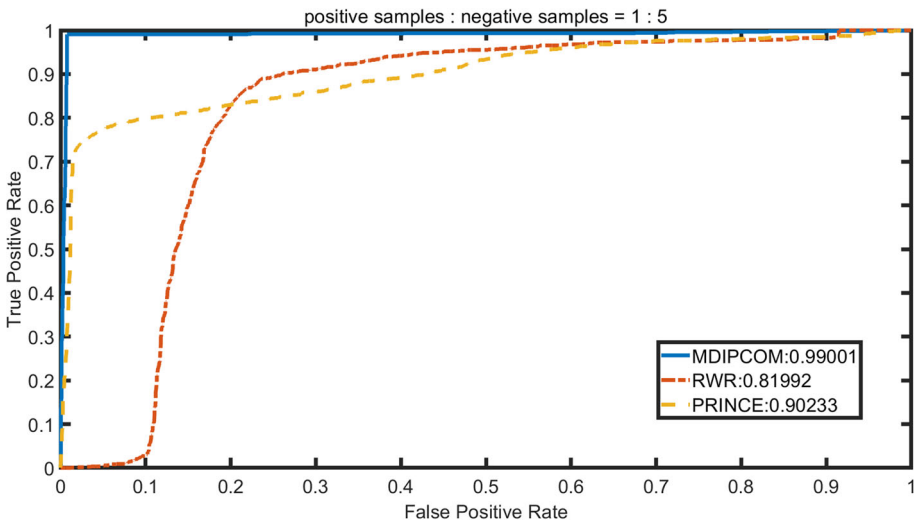

Fig. 5 The ROC curves of MDAPCOM and previous approaches containing PRINCE and RWRMDA on the data sets. The ratio of positive and negative sample 1:5

Restart algorithm can capture the relationships between two nodes and the global topological information of nodes in the network [32–34]. In this study, we run RWR algorithm on the global heterogeneous network and get a high-dimensional(25,811) vector for each node. The vector reveals the topological properties of the node in the network, which includes a set of possibilities that a node can access to other nodes. We use D to represent the adjacency matrix of our global heterogeneous network, and T, a normalized matrix, represents the transition probability from the node i to the node j, T is defined as

$$T_{ij} = \frac{D_{ij}}{\sum_k D_{ik}} \tag{1}$$

If a node i is connected with a node j, the value of D_{ij} is 1, otherwise the value is 0. The RWR can be regarded as an iterative process, which is expressed as

$$P_{t+1} = (1-\alpha)TP_t + \alpha P_0 \tag{2}$$

Where α is the restart rate of random walker which is in the range of [0,1], P_0 is the initial probability of the heterogeneous network, P_t is the state of the heterogeneous network when the process is in the t-th.

Here, we get a 25,811-dimensional feature for every node which reveals the topological relevance of a node to other nodes(2,588 miRNAs, 18,143 proteins and 5,080

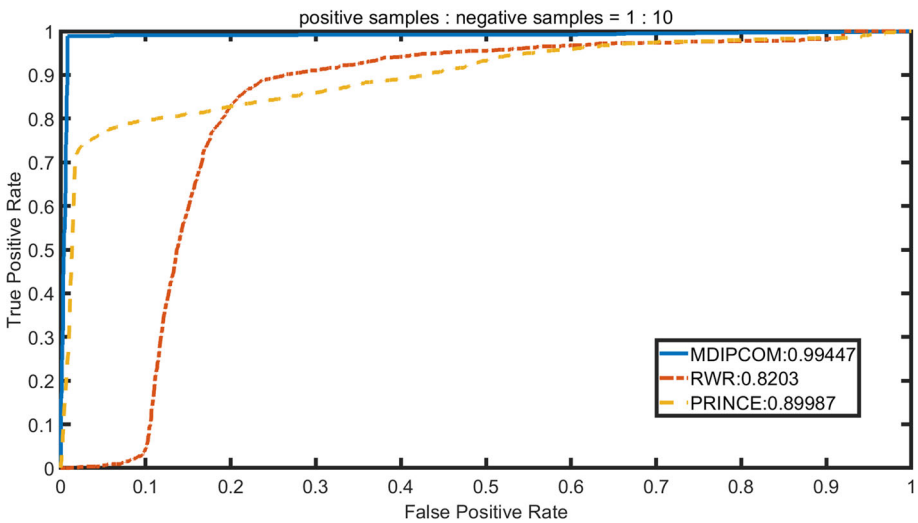

Fig. 6 The ROC curves of MDAPCOM and previous approaches containing PRINCE and RWRMDA on the data sets. The ratio of positive and negative sample 1:10

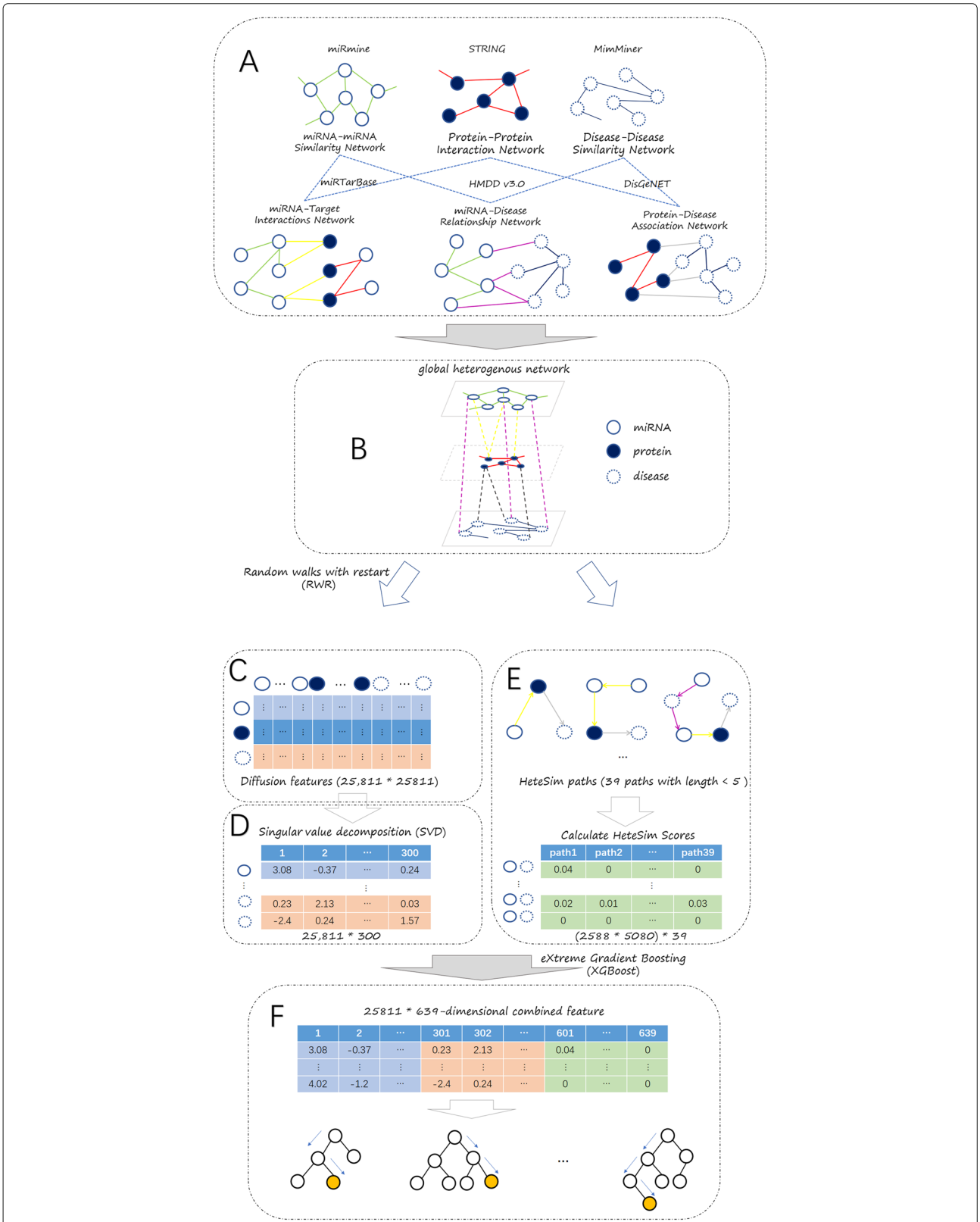

Fig. 7 The flowchart of MDAPCOM: **a** Obtain six kinds of data from online databases. **b** Merge these data to build a global heterogeneous network **c** Utilize RWR algorithm to get the diffusion feature. **d** Apply SVD to reduce dimension of the diffusion feature. **e** Use HeteSim measure to obtain HeteSim feature. **f** Integrate reduced diffusion feature and HeteSim feature and then apply XGBoost algorithm to train the model using the combined feature

diseases) in the network. Using such tremendous features directly to train the model is pretty time-consuming and unnecessary, since they contain some noise. Therefore, we reduce the 25,811-dimensional diffusion feature to 300-dimension by singular value decomposition (SVD) algorithm [35, 36].

HeteSim measure

The HeteSim measure performs well in measuring the correlation of nodes in the heterogeneous biological network [37]. It's a self-maximum and symmetric measure, using an uniform framework to measure the correlation of nodes based on specified path [38]. In this paper, we use HeteSim scores of miNRA-disease pairs to extract network information.

Definition 1 (**Transition probability matrix** [38]) *A and B are two types of nodes in the heterogeneous network.* $(M_{AB})_{m*n}$ *is an adjacency matrix indicating the relation between A and B, if there is an association between a node i belonging to A and a node j belonging to B,* $M_{AB}(i,j) = 1$, *otherwise* $M_{AB}(i,j) = 0$. *The transition probability matrix* T_{AB} *is defined as follows*

$$T_{AB}(x,y) = \frac{M_{AB}(x,y)}{\sum_{i=1}^{n} M_{AB}(x,i)} \tag{3}$$

Definition 2 (**Reachable probability matrix** [38]) R_ρ *represents the reachable probability matrix based on the path* $\rho = P_1P_2P_3 \ldots P_{n+1}$, *where* P_i *represents any types of nodes of the heterogeneous network.* R_ρ *can be calculated as*

$$R_\rho = T_{P_1P_2} T_{P_2P_3} \ldots T_{P_nP_{n+1}} \tag{4}$$

Based on the above 2 definitions, we can calculate the HeteSim score in 3 steps [38].

1 Separate the path ρ from the middle into ρ_L and ρ_R. When the path length is even, ρ_L and ρ_R are equal in length, and R_{ρ_L} and R_{ρ_R} can be directly calculated. When the path length is odd, there are two intermediate nodes, take each one of them as intermediate node respectively to obtain ρ_{L_1} , ρ_{L_2}, ρ_{R_1} and ρ_{R_2}, then R_{ρ_L} , R_{ρ_R} can be calculated as

$$R_{\rho_L} = \frac{R_{\rho_{L_1}} + R_{\rho_{L_2}}}{2}$$

$$R_{\rho_R} = \frac{R_{\rho_{R_1}} + R_{\rho_{R_2}}}{2}$$

2 Calculate the R_{ρ_L} and $R_{\rho_R^{-1}}$, where ρ_R^{-1} represents the reverse of ρ_R, for example, if $\rho_R = ABC$, then $\rho_R^{-1} = CBA$.

Table 1 All paths less than 5 in length starting at miRNA and ending at disease. M is miRNA, P is protein and D is disease, for example, path1 MMD is the path miRNA-miRNA-disease

id	path	id	path	id	path
1	MMD	14	MMMPD	27	MPPDD
2	MPD	15	MMMDD	28	MPDMD
3	MDD	16	MMPMD	29	MPDPD
4	MMMD	17	MMPPD	30	MPDDD
5	MMPD	18	MMPDD	31	MDMMD
6	MMDD	19	MMDMD	32	MDMPD
7	MPMD	20	MMDPD	33	MDMDD
8	MPPD	21	MMDDD	34	MDPMD
9	MPDD	22	MPMMD	35	MDPPD
10	MDMD	23	MPMPD	36	MDPDD
11	MDPD	24	MPMDD	37	MDDMD
12	MDDD	25	MPPMD	38	MDDPD
13	MMMMD	26	MPPPD	39	MDDDD

3 Achieve the HeteSim measure as

$$HeteSim\,(a,b|\rho) = \frac{R_{\rho_L}(a,:)\left(R_{\rho_R^{-1}}(b,:)\right)^T}{\|R_{\rho_L}(a,:)\|_2 \times \|R_{\rho_R^{-1}}(b,:)\|_2} \tag{5}$$

Using the above method, we can derive 39 HeteSim scores for each miRNA-disease pair(i.e. a 39-dimensional HeteSim feature vector for each miRNA-disease pair) based on all paths less than 5 in length starting at miRNA and ending at disease. The detailed paths are listed in Table 1.

The eXtreme gradient boosting (XGBoost) algorithm

The eXtreme Gradient Boosting is an end-to-end system extended by tree boosting, and it's used widely in machine learning [28]. The algorithm can be obtained from python toolkits scikit-learn. In this study, a 600-dimensional diffusion feature(300-dimensional for miRNA and 300-dimensional for disease) and a 39-dimensional HeteSim feature are extracted for each miNRA-disease pair in the global network. Subsequently, the two features are combined, forming a 639-dimensional feature, to train the prediction model by XGBoost, where the optimal learning rate is 0.15, the number of iterations is 650, the max depth of tree is 4 and default values set for the other parameters.

Abbreviations

XGBoost: eXtreme Gradient Boosting; miRNA: MicroRNA; RWR: random walk with restart algorithm; SVM: support vector machine; GO: gene ontology; WP: weighted profile; CMF: collaborative matrix factorization; SVD: singular value decomposition; PPI: Protein-Protein interaction; PCC: Pearson correlation coefficient; ROC: receiver operating characteristic curve; AUC: area under the receiver operating characteristic curve; PRE: precision; REC: recall; FSC: F-score; ACC: accuracy; RF: random forest; GTB: gradient tree boosting; FPR: false positive rate; TPR: true positive rate; RBF: Radial Basis Function

Acknowledgements
We would like to thank the Experimental Center of School of Computer Science and Engineering of Central South University, for providing computing resources.

About this supplement
This article has been published as part of *BMC Medical Genomics Volume 13 Supplement 10, 2020: Selected articles from the 18th Asia Pacific Bioinformatics Conference (APBC 2020): medical genomics*. The full contents of the supplement are available online at https://bmcmedgenomics.biomedcentral.com/articles/supplements/volume-13-supplement-10.

Authors' contributions
ML, JY and LD designed the study and conducted experiments. ML, JY and LD performed statistical analyses. JY and ML drafted the manuscript. ML and JY prepared the experimental materials and benchmarks. All author(s) have read and approved the final manuscript.

Funding
This work was supported by National Natural Science Foundation of China under grants No. 61972422 and No. 61672541. Publication costs are funded by National Natural Science Foundation of China under grant No. 61972422.

Availability of data and materials
The datasets used during the current study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Published: 22 October 2020

References
1. Ambros V. The functions of animal micrornas. Nature. 2004;431(7006):350.
2. Liu H, Zhang W, Zou B, Wang J, Deng Y, Deng L. DrugCombDB: a comprehensive database of drug combinations toward the discovery of combinatorial therapy. Nucleic Acids Res. 2020;48(D1):D871–D881. https://doi.org/10.1093/nar/gkz1007.
3. Nagaraja AK, Creighton CJ, Yu Z, Zhu H, Gunaratne PH, Reid JG, Olokpa E, Itamochi H, Ueno NT, Hawkins SM, et al. A link between mir-100 and frap1/mtor in clear cell ovarian cancer. Mol Endocrinol. 2010;24(2):447–63.
4. Latronico MV, Catalucci D, Condorelli G. Emerging role of micrornas in cardiovascular biology. Circ Res. 2007;101(12):1225–36.
5. Nunez-Iglesias J, Liu C-C, Morgan TE, Finch CE, Zhou XJ. Joint genome-wide profiling of mirna and mrna expression in alzheimer's disease cortex reveals altered mirna regulation. PloS ONE. 2010;5(2):8898.
6. Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P. Modulation of hepatitis c virus rna abundance by a liver-specific microrna. Science. 2005;309(5740):1577–81.
7. Calin GA, Dumitru CD, Shimizu M, Bichi R, Zupo S, Noch E, Aldler H, Rattan S, Keating M, Rai K, et al. Frequent deletions and down-regulation of micro-rna genes mir15 and mir16 at 13q14 in chronic lymphocytic leukemia. Proc Natl Acad Sci. 2002;99(24):15524–9.
8. Sredni ST, Huang C-C, Bonaldo MdF, Tomita T. Microrna expression profiling for molecular classification of pediatric brain tumors. Pediatr Blood Cancer. 2011;57(1):183–4.
9. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, et al. Microrna expression profiles classify human cancers. Nature. 2005;435(7043):834.
10. Jiang Q, Hao Y, Wang G, Juan L, Zhang T, Teng M, Liu Y, Wang Y. Prioritization of disease micrornas through a human phenome-micrornaome network. BMC Syst Biol. 2010;4(1):2.
11. Jiang Q, Wang G, Wang Y. An approach for prioritizing disease-related micrornas based on genomic data integration. In: 2010 3rd International Conference on Biomedical Engineering and Informatics, vol. 6. Yantai: IEEE; 2010. p. 2270–4.
12. Chen X, Liu M-X, Yan G-Y. Rwrmda: predicting novel human microrna–disease associations. Mol BioSyst. 2012;8(10):2792–8.
13. Shi H, Xu J, Zhang G, Xu L, Li C, Wang L, Zhao Z, Jiang W, Guo Z, Li X. Walking the interactome to identify human mirna-disease associations through the functional link between mirna targets and disease genes. BMC Syst Biol. 2013;7(1):101.
14. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Comput Biol. 2010;6(1):1000641.
15. Xu J, Li C-X, Lv J-Y, Li Y-S, Xiao Y, Shao T-T, Huo X, Li X, Zou Y, Han Q-L, et al. Prioritizing candidate disease mirnas by topological features in the mirna target–dysregulated network: Case study of prostate cancer. Mol Cancer Ther. 2011;10(10):1857–66.
16. Chen X, Yan G-Y. Semi-supervised learning for potential human microrna-disease associations inference. Sci Rep. 2014;4:5501.
17. Zheng K, You Z-H, Wang L, Zhou Y, Li L-P, Li Z-W. Mlmda: a machine learning approach to predict and validate microrna–disease associations by integrating of heterogenous information sources. J Transl Med. 2019;17(1):260.
18. Peng J, Hui W, Li Q, Chen B, Hao J, Jiang Q, Shang X, Wei Z. A learning-based framework for mirna-disease association identification using neural networks. Bioinformatics. 2019;35(21):4364–71.
19. Yang Y, Fu X, Qu W, Xiao Y, Shen H-B. Mirgofs: a go-based functional similarity measurement for mirnas, with applications to the prediction of mirna subcellular localization and mirna–disease association. Bioinformatics. 2018;34(20):3547–56.
20. Chen X, Huang L, Xie D, Zhao Q. Egbmmda: extreme gradient boosting machine for mirna-disease association prediction. Cell death Dis. 2018;9(1):3.
21. Yin M-M, Cui Z, Gao M-M, Liu J-X, Gao Y-L. Lwpcmf: Logistic weighted profile-based collaborative matrix factorization for predicting mirna-disease associations. IEEE/ACM Trans Comput Biol Bioinforma. 2019. https://doi.org/10.1109/TCBB.2019.2937774.
22. Panwar B, Omenn GS, Guan Y. mirmine: a database of human mirna expression profiles. Bioinformatics. 2017;33(10):1554–60.
23. Zhang J, Zhang Z, Chen Z, Deng L. Integrating multiple heterogeneous networks for novel lncrna-disease association inference. IEEE/ACM Trans Comput Biol Bioinforma. 2017;16(2):396–406.
24. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, et al. String v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 2014;43(D1):447–52.
25. Van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. Eur J Hum Genet. 2006;14(5):535.
26. Chou C-H, Shrestha S, Yang C-D, Chang N-W, Lin Y-L, Liao K-W, Huang W-C, Sun T-H, Tu S-J, Lee W-H, et al. mirtarbase update 2018: a resource for experimentally validated microrna-target interactions. Nucleic Acids Res. 2017;46(D1):296–302.
27. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. Hmdd v3. 0: a database for experimentally supported human microrna–disease associations. Nucleic Acids Res. 2018;47(D1):1013–17.
28. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM; 2016. p. 785–94.
29. Liaw A, Wiener M, et al. Classification and regression by randomforest. R News. 2002;2(3):18–22.
30. Burges CJ. A tutorial on support vector machines for pattern recognition. Data Min Knowl Disc. 1998;2(2):121–67.
31. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189–232.
32. Wang F, Landau D. Determining the density of states for classical statistical models: A random walk algorithm to produce a flat histogram. Phys Rev E. 2001;64(5):056101.
33. Liu Y, Zeng X, He Z, Zou Q. Inferring microrna-disease associations by random walk on a heterogeneous network with multiple data sources. IEEE/ACM Trans Comput Biol Bioinforma. 2016;14(4):905–15.
34. Shang H, Liu Z-P. Prioritizing type 2 diabetes genes by weighted pagerank on bilayer heterogeneous networks. IEEE/ACM Trans Comput Biol Bioinforma. 2019. https://doi.org/10.1109/TCBB.2019.2917190.

35. Golub GH, Reinsch C. Singular value decomposition and least squares solutions. In: Linear Algebra. Berlin, Heidelberg: Springer; 1971. p. 134–151.
36. Wang S, Cho H, Zhai C, Berger B, Peng J. Exploiting ontology graph for predicting sparsely annotated gene function. Bioinformatics. 2015;31(12): 357–64.
37. Zeng X, Liao Y, Liu Y, Zou Q. IEEE/ACM Trans Comput Biol Bioinforma (TCBB). 2017;14(3):687–95.
38. Shi C, Kong X, Huang Y, Philip SY, Wu B. Hetesim: A general framework for relevance measure in heterogeneous networks. IEEE Trans Knowl Data Eng. 2014;26(10):2479–92.

Publisher's Note