RSC Advances

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *RSC Adv.*, 2020, **10**, 11634

Heterogeneous graph inference based on similarity network fusion for predicting lncRNA–miRNA interaction†

Yongxian Fan, * Juan Cui and QingQi Zhu

Received 31st December 2019
Accepted 14th March 2020

DOI: 10.1039/c9ra11043g

rsc.li/rsc-advances

LncRNA and miRNA are two non-coding RNA types that are popular in current research. LncRNA interacts with miRNA to regulate gene transcription, further affecting human health and disease. Accurate identification of lncRNA–miRNA interactions contributes to the in-depth study of the biological functions and mechanisms of non-coding RNA. However, relying on biological experiments to obtain interaction information is time-consuming and expensive. Considering the rapid accumulation of gene information and the few computational methods, it is urgent to supplement the effective computational models to predict lncRNA–miRNA interactions. In this work, we propose a heterogeneous graph inference method based on similarity network fusion (SNFHGILMI) to predict potential lncRNA–miRNA interactions. First, we calculated multiple similarity data, including lncRNA sequence similarity, miRNA sequence similarity, lncRNA Gaussian nuclear similarity, and miRNA Gaussian nuclear similarity. Second, the similarity network fusion method was employed to integrate the data and get the similarity network of lncRNA and miRNA. Then, we constructed a bipartite network by combining the known interaction network and similarity network of lncRNA and miRNA. Finally, the heterogeneous graph inference method was introduced to construct a prediction model. On the real dataset, the model SNFHGILMI achieved AUC of 0.9501 and 0.9426 ± 0.0035 based on LOOCV and 5-fold cross validation, respectively. Furthermore, case studies also demonstrate that SNFHGILMI is a high-performance prediction method that can accurately predict new lncRNA–miRNA interactions. The Matlab code and readme file of SNFHGILMI can be downloaded from https://github.com/cj-DaSE/SNFHGILMI.

Introduction

Non-coding RNAs are a class of RNA that does not encode proteins, so for a long time, it was considered a "dark matter" on the genome.[1] However, with the accumulation of research, scientists found that only about 2% of the RNA encoded proteins in the human genome and the remaining 98% of the RNA no longer continues to be translated into proteins, but functions in various biological processes as non-coding RNA.[2] In the vast majority of human transcript expression, the length of non-coding RNA ranges from 22 nucleotides to hundreds of kb. According to the length of non-coding RNA, the length of more than 200 nucleotides is defined as long non-coding RNA (lncRNA),[3] which are involved in various cellular processes such as cell differentiation, growth, death, *etc.*[4] It is also involved in chromatin modification, transcription complex targeting, mRNA splicing and protein translation.[5,6] And more and more studies have shown that it is associated with diseases such as gastric cancer, lung cancer, prostate cancer and so on.[7,8] MicroRNA (miRNA) is a short non-coding RNA with a length of approximately 22 nucleotides,[9] which usually binds to target mRNA or proteins and acts as a negative regulator at the post-transcriptional level. Studies have shown that miRNAs play important roles in biological life processes such as cell growth, tissue differentiation, embryo development and apoptosis, cross-species gene expression variation, and coordinated regulation with transcription factors.[10] Both lncRNA and miRNA are closely related to human health and disease.[11–15] However, only a few lncRNA/miRNA annotations have experimental information, and there is still a lot of improvement for their research.

LncRNAs have been found can exert biological functions by interacting with DNA, RNA, and proteins.[16] Such interactions between lncRNA and biomolecule are very important for the regulation of life activities. Recently, it has been found that lncRNA can act as a sponge of miRNA and reduce the regulation effect of miRNA on mRNA by playing the function of competitive endogenous RNA, thus acting as a decoy to regulate the behavior of miRNA.[17,18] Similarly, miRNAs also play an important role in the molecular mechanism of lncRNA.[19] Mutual regulation of miRNA and lncRNA has a great influence on the pathological process of human diseases.[20] For example, the

School of Computer and Information Security, Guilin University of Electronic Technology, Guilin 541004, China. E-mail: yongxian.fan@gmail.com

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c9ra11043g

regulatory network of lncRNA–miRNA in prostate cancer,[21] gastric cancer,[22] and vascular disease[23] has been constructed. However, it is very limited to identify lncRNA–miRNA interaction through biological experiments. To understand the role of lncRNA–miRNA interactions in pathophysiology in more detail and to discover new biomarkers and therapeutic. A reasonable and efficient method for predicting lncRNA–miRNA interaction is necessary.

Recently, there are various computational methods have been developed to predict lncRNA–miRNA interactions. These prediction methods were developed based on the known lncRNA–miRNA heterogeneous network, the similarity network between lncRNAs, and the similarity network between miRNAs. Huang *et al.* proposed a two-way diffusion model called EPLMI,[24] which predicts potential lncRNA–miRNA interactions based on known lncRNA–miRNA interactions and lncRNA/miRNA expression profile information. This is also the first model dedicated to the prediction of lncRNA–miRNA interactions and has made a significant contribution to future research. INLMI is a new model proposed by Huang and Hu based on previous research.[25] This model uses network integration on two similarity networks of lncRNA/miRNA, and they predict new lncRNA–miRNA interactions by a non-negative matrix factorization method with integrated network and known lncRNA–miRNA interactions. Zhang *et al.* proposed a sequence-derived linear neighborhood propagation model (SLNPM).[26] This model proposed a novel similarity measure named fast linear neighborhood similarity approach (FLNS) and used the FLNS to calculate the lncRNA–lncRNA similarity and miRNA–miRNA similarity. And then, the label propagation algorithm is introduced to predict the potential lncRNA–miRNA interactions based on the known lncRNA–miRNA interaction, integrated lncRNA similarity, and integrated miRNA similarity. Subsequently, based on previous study, Zhang *et al.* proposed a sequence-derived linear neighborhood propagation method with information combination for lncRNA–miRNA interaction prediction.[27] Compared with SLNPM, this model proposed two strategies of similarity-based information combination and interaction profile-based information combination. Construct two editions of SLNPM based two strategies, and the weighted averages of their results are adopted as final scores. Wong *et al.* proposed LNRLMI model,[28] which constructed a binary network based on known associated lncRNA–miRNA interaction network, lncRNA similarity, and miRNA similarity, and used a semi-supervised learning model called linear neighbor representation method to predict lncRNA–miRNA interaction. Recently, Liu *et al.* predicted lncRNA–miRNA interactions based on logical matrix factorization with neighborhood regularized.[29] Neighborhood regularization takes into account the strong correlation between lncRNA–lncRNA and miRNA–miRNA, and the logical matrix factorization algorithm can use latent vectors to predict potential relationships globally. The combination of the two algorithms further improves the prediction accuracy of the model. In addition, the rich association prediction research in bioinformatics also provide meaningful references for our work (*e.g.*, lncRNA-disease,[30–32] miRNA-disease,[12,13,33–40] lncRNA–protein,[41–43] drug-target,[44–47] and drug-disease[48–50] interaction prediction). However, although previous work of lncRNA–miRNA interaction prediction has achieved excellent predictions, it is not difficult to find that there are some issues to improve the lncRNA–miRNA interaction prediction. It is noted that the fusion of heterogeneous data is a key factor to improve the accuracy of interaction prediction. Recently, prediction models based on the integration of diverse information and similarity have been proposed and show excellent performance.[43,51,52]

In this paper, we explored a heterogeneous graph inference method based on similarity network fusion technology, named is SNFHGILMI, to predict lncRNA–miRNA interaction. The similarity network fusion technology was originally proposed by Wang *et al.*,[53] and applied to patient similarity network fusion, which can obtain shared and complementary information from various kernel matrices so that the integrated matrix reveals lncRNA/miRNA information as much as possible. The heterogeneous graph inference algorithm is an iterative process to find the optimal solution based on the similarity information of the global network. All the information about nodes and interactions are expressed on the heterogeneous graph, and only the links of different types of nodes need to be derived to predict lncRNA–miRNA interaction.[39,54,55] Based on these two algorithms, the classification model SNFHGILMI was constructed. Specifically, SNFHGILMI firstly integrated the similarity of lncRNA/miRNA based on sequence and interaction profile. Then, based on the integrated lncRNA–lncRNA similarity network, miRNA–miRNA similarity network and known lncRNA–miRNA interaction, a heterogeneous lncRNA–miRNA network was constructed. Finally, a heterogeneous graph inference algorithm was introduced to calculate the lncRNA–miRNA interaction score. The experimental results show that our approach has an excellent performance over other existing methods.

Materials and methods

Datasets

The data used in this work is from the Huang's paper,[24] which includes information on known lncRNA–miRNA interactions and the ID of lncRNA/miRNA. The specific information refers to the following three public databases. The database lncRNASNP provides comprehensive information about lncRNA–miRNA interaction, and these known lncRNA–miRNA interactions confirmed by laboratory studies (http://bioinfo.life.hust.edu.cn/lncRNASNP).[56] LNCipedia is a comprehensive human lncRNA database (https://lncipedia.org/) containing human lncRNA transcript sequences and annotation information, and lncRNA in this database with a uniform nomenclature system.[57] The miRBase is a publicly available database, which provides information of miRNA sequence, annotations and known gene targets (http://www.mirbase.org/index.shtml).[58]

First, 8091 known lncRNA–miRNA interaction data were collected from the lncRNASNP database. After deleting the duplicate data, and 5118 lncRNA–miRNA entries were obtained. These entries involved 780 lncRNAs and 275 miRNAs. Then, the sequence of each lncRNA was extracted from the LNCipedia database, and only 770 lncRNA sequence information was available. Similarly, the sequence of each miRNA was extracted from the miRBase database. Finally, we removed interaction whose lncRNA sequence is unavailable and 4966 lncRNA–miRNA interactions were obtained as the experimental dataset. Constructing an adjacency matrix A of lncRNA–miRNA interaction. We defined $A(i,j) = 1$ if lncRNA l_i interacts with miRNA m_j, and $A(i,j) = 0$ if there is an unknown interaction.

Problem formalization

The visualization of the lncRNA–miRNA interaction problem is shown in Fig. 1. Heterogeneous graph consists of nodes and edges. Defined $L = \{l_1, \ldots, l_i, \ldots, l_{nl}\}$ and $M = \{m_1, \ldots, m_j, \ldots, m_{nm}\}$ to represent the nodes of the nl lncRNA and nm miRNA, respectively. Inside the lncRNA (miRNA) network, the weight of the edge is determined by the similarity value of two lncRNA (miRNA) corresponding to two nodes. Between the lncRNA–miRNA networks, if there is an interaction between l_i and m_j, the edge exists and has a weight of 1. Based on the above description, the heterogeneous graph can be represented as $G = \{\{L, M\},\{E_{ll}, E_{mm}, E_{lm}\},\{W_{ll}, W_{mm}, W_{lm}\}\}$, where E_{ll}, E_{mm} and E_{lm} represent lncRNA–lncRNA, miRNA–miRNA, and lncRNA–miRNA edges, respectively. And W_{ll}, W_{mm} and W_{lm} represent the values of the weights on these three kinds of edges. It can be seen that there are missing edges between the lncRNA node and the miRNA node, and these missing edges are unknown lncRNA–miRNA interactions. Our goal is to elucidate new potential lncRNA–miRNA interactions based on lncRNA–lncRNA similarity, miRNA–miRNA similarity and known lncRNA–miRNA interaction.

Construct lncRNA/miRNA sequence similarity

Similar sequences may have similar structures, and similar structures may have similar functions.[59] Based on the hypothesis that functionally similar lncRNA/miRNAs tend to interact with similar miRNA/lncRNA,[33] we calculated the lncRNA/miRNA sequence similarity. Here, the sequence global alignment algorithm Needleman–Wunsch is used to calculate and construct the lncRNA/miRNA sequence similarity matrix.[60] Referring to previous studies, the parameters are set in the Needleman–Wunsch algorithm as follows, the identification score is 2, the gap-open penalty is -0.5 and the gap-open extending penalty is -0.1. Finally, the sequence similarity matrix of lncRNA is defined as SL, the sequence similarity matrix of miRNA is defined as SM.

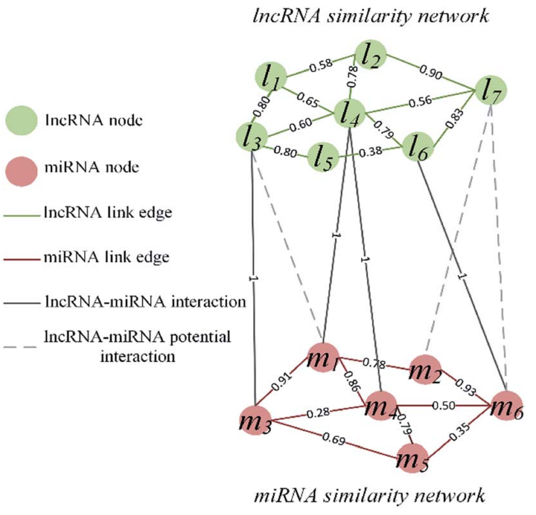

Fig. 1 Visualization of lncRNA–miRNA interaction heterogeneous network.

Construct lncRNA/miRNA Gaussian interaction profile kernel similarity

Considering that a single similarity network can not provide comprehensive bioinformatics, the Gaussian interaction profile kernel similarity is added here. Given the adjacency matrix $A_{nl\times nm}$ of the known lncRNA–miRNA interactions, nl is the number of lncRNA and nm is the number of miRNAs. The interaction profile (IP) of each lncRNA is represented by a row in the adjacency matrix $A_{nl\times nm}$, that is, $IP(l_i)$ is defined as all elements of the ith row of the adjacency matrix. Similarly, the IP of each miRNA is represented by a column in the adjacency matrix $A_{nl\times nm}$, that is, $IP(m_j)$ is defined as all elements of the jth column of the adjacency matrix. Therefore, the Gaussian interaction profile kernel similarity of lncRNA l_i and lncRNA l_j is defined as follows.

$$KL(l_i, l_j) = \exp(-\gamma_l\|IP(l_i) - IP(l_j)\|^2) \tag{1}$$

$$\gamma_l = \gamma_l^{'} \Big/ \left(\frac{1}{nl}\sum_{i=1}^{nl}\|IP(l_i)\|^2\right) \tag{2}$$

Here, the parameter γ_l is used to constrain the bandwidth of the kernel, which is affected by the new bandwidth parameter $\gamma_l^{'}$. The parameter $\gamma_l^{'}$ is the average number of associations of each lncRNA with the miRNA. Referring to the parameter settings of the previous paper,[61] we set $\gamma_l^{'} = 1$. Similarly, the same method is used to construct the Gaussian interaction profile kernel similarity of miRNA.

$$KM(m_i, m_j) = \exp(-\gamma_m\|IP(m_i) - IP(m_j)\|^2) \tag{3}$$

$$\gamma_m = \gamma_m^{'} \Big/ \left(\frac{1}{nm}\sum_{i=1}^{nm}\|IP(m_i)\|^2\right) \tag{4}$$

Integrate the similarity matrix of lncRNA/miRNA

Using the above similarity calculation method, four similarity matrices are obtained, lncRNA sequence similarity matrix SL, lncRNA Gaussian kernel interaction similarity matrix KL, miRNA sequence similarity matrix SM, miRNA Gaussian kernel interaction similarity matrix KM. It is a key issue that how to use a reasonable and effective method to fuse the two similarity matrices of lncRNA/miRNA so that the information after fusion can better improve the predictive performance of the model. Two matrix fusion methods are given here.

One is a simple linear fusion method called SLF,[62] the calculation process is as follows. α is the weight parameter (α is the values between 0 and 1).

$$WL(i,j) = \alpha \times SL(i,j) + (1 - \alpha) \times KL(i,j) \tag{5}$$

$$WM(i,j) = \alpha \times SM(i,j) + (1 - \alpha) \times KM(i,j) \tag{6}$$

The other is a nonlinear combination method called Similarity Network Fusion (SNF), which is used in this paper. This method utilizes the nonlinear iterative method of information to fuse the similarity matrix and achieves good performance.

Considering that the use of discrete data by SNF is not ideal, the lncRNA/miRNA sequence similarity matrix is calculated as follows, so that the values in the matrix are converted to continuous values between 0 and 1.

$$W(i,j) = \frac{B(i,j)}{\sum_{j} B(i,j)} \tag{7}$$

$$W_{\text{sym}} = (W + (\mathrm{W})^T)/2 \tag{8}$$

Here, B represents the sequence similarity matrix of the lncRNA or miRNA $(W)^T$ denotes the transpose of W and making them symmetric using the formula (8). Because of the self-similarity of the diagonal elements of the matrix, the similarity matrix is positively semi-determined by adding a small multiple of identity matrix.[45] Finally, the original SL and SM are converted into SL_{kner} and SM_{kner}.

Given four kernel matrices, SL_{kner}, SM_{kner}, KL and KM. Here we take lncRNA as an example to fuse matrix. First, we normalize the kernel matrix by dividing by the sum of the rows, so that the sum of all the elements in each row is 1. Then the normalized matrix is symmetrized by formula (8), and the resulting matrices are defined as $P^{(1)}$ and $P^{(2)}$, corresponding to matrix SL_{kner} and KL, respectively. Next, the K nearest neighbor method is used to calculate the local similarity of the matrix for each P.

$$S(i,j) = \begin{cases} \dfrac{P(i,j)}{\sum_{k\in N_i} P(i,j)}, j\in N_i \\ 0,\ \text{others} \end{cases} \tag{9}$$

Here, N_i represents the set of neighbors nearest to lncRNA l_i (including lncRNA l_i). k is the number of nearest neighbors, and the value of k is set by the user, where $k = 3$. For neighbors that are not in the nearest neighbor, the value converted to 0. The generated matrices $S^{(1)}$ and $S^{(2)}$ correspond to matrices $P^{(1)}$ and $P^{(2)}$, respectively. This process strengthens the strong links and eliminates weak links in the network, this greatly reduces noise interference.

Finally, after a nonlinear iterative process, the fusion matrices is implemented. The specific process is as follows.

$$P_{t+1}^{(1)} = S^{(1)} \times P_t^{(2)} \times (S^{(1)})^T \tag{10}$$

$$P_{t+1}^{(2)} = S^{(2)} \times P_t^{(1)} \times (S^{(2)})^T \tag{11}$$

Here, $P_{t+1}^{(1)}$ is the state matrix transformed by lncRNA sequence kernel matrix after t iteration, and $P_{t+1}^{(2)}$ is the state matrix transformed by lncRNA Gaussian kernel similarity matrix after t iteration. Each iteration exchanges information of different original networks. After t iterations (in this work $t = 20$), the final state matrix of lncRNA is calculated as follows.

$$WL = (P_t^{(1)} + P_t^{(2)})/2 \tag{12}$$

Then, the matrix WL is normalized and further transformed as follows.

$$WL = (WL + (WL)^T + I)/2 \tag{13}$$

Here, $(WL)^T$ denotes the transpose of WL, and I is an identity matrix. Similarly, the fusion of miRNA also uses the above steps. In addition, the fusion method for three or more similarity matrices is not introduced here. For details, please refer to Wang's literature.[53]

Constructing prediction model

Combining the lncRNA similarity matrix, the miRNA similarity matrix, and the known lncRNA–miRNA interaction matrix, a SNFHGILMI computational model was established to predict potential lncRNA–miRNA interactions. The SNFHGILMI method proposed in this paper combines SNF and heterogeneous graph inference (HGI) algorithms to improve the performance of lncRNA–miRNA interaction prediction. Fig. 2 shows that the lncRNA–miRNA interaction network can combine independent lncRNA and miRNA internal similarities to predict new interactions between lncRNA–miRNAs. That is, if there is no known association between lncRNA l_i and miRNA m_j, then the potential association probability values between them can be solved by the following definition.

$$W(l_i, m_j) = \sum_{nl}^{nl} \sum_{m=1}^{nm} WL(l_n, l_i) \times A(l_n, m_m) \times WM(m_m, m_j) \tag{14}$$

The above formula indicates that the potential interaction probability between lncRNA l_i and miRNA m_j is calculated by summing all paths of length three (l_i and l_n, m_j and m_m, l_n and m_m) in the heterogeneous graph. To obtain the potential interaction probabilities for all unlabeled lncRNA and miRNA pairs in one iteration, we convert the above formula into a multiplication calculation between matrices as follows.

$$W_{i+1} = \lambda WL \times W_i \times WM + (1 - \lambda) W_0 \tag{15}$$

Here, λ is a decay factor with a value between 0 and 1. In this work, we set $\lambda = 0.1$. W_0 is the initial interaction adjacency matrix of lncRNA–miRNA. Each iteration will generate a new interaction matrix. When the matrix difference between W_{i+1} and W_i is less than a certain threshold, the matrix converges, which means that the information propagation tends to be stable. According to the previous literature,[54] the lncRNA–miRNA correlation probability matrix will converge when the matrices WL and WM are properly normalized using the following formula, respectively.

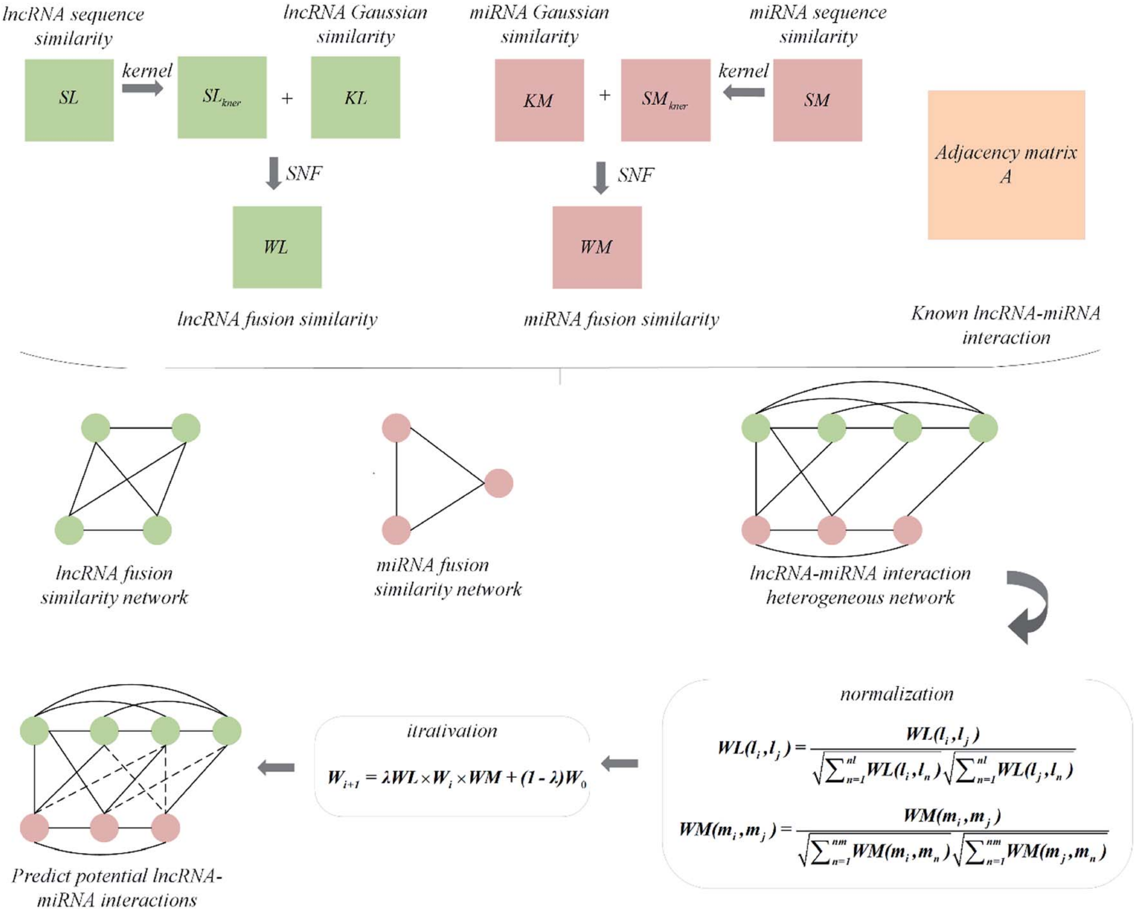
lncRNA sequence similarity
SL
kernel
SL_kner
+
lncRNA Gaussian similarity
KL
SNF
WL
lncRNA fusion similarity
miRNA Gaussian similarity
KM
+
SM_kner
kernel
miRNA sequence similarity
SM
SNF
WM
miRNA fusion similarity
Adjacency matrix A
Known lncRNA-miRNA interaction
lncRNA fusion similarity network
miRNA fusion similarity network
lncRNA-miRNA interaction heterogeneous network
normalization
WL(l_i,l_j) = WL(l_i,l_j) / √∑_{n=1}^{nl} WL(l_i,l_n) √∑_{n=1}^{nl} WL(l_j,l_n)
WM(m_i,m_j) = WM(m_i,m_j) / √∑_{n=1}^{nm} WM(m_i,m_n) √∑_{n=1}^{nm} WM(m_j,m_n)
itrativation
W_{i+1} = λWL×W_i×WM+(1-λ)W_0
Predict potential lncRNA-miRNA interactions

Fig. 2 The flowchart of prediction process of SNFHGILMI.

$$WL(l_i,l_j) = \frac{WL(l_i,l_j)}{\sqrt{\sum_{n=1}^{nl} WL(l_i,l_n)}\sqrt{\sum_{n=1}^{nl} WL(l_j,l_n)}} \tag{16}$$

$$WM(m_i,m_j) = \frac{WM(m_i,m_j)}{\sqrt{\sum_{n=1}^{nm} WM(m_i,m_n)}\sqrt{\sum_{n=1}^{nm} WM(m_j,m_n)}} \tag{17}$$

We adopt the threshold as 10^{-6}, which means that when the difference between W_{i+1} and W_i measured by the L1 norm is less than 10^{-6}, the iterative process stops and the formula (15) convergence. Moreover, we have given the details of the proof about convergence in the S1 text.†

Results and discussion

Convergence performance

Convergence is a very important problem for iterative algorithms. Therefore, we analyzed the number of iterations of the SNF. Specifically, the experiment defines the relative change index as $E_t = \frac{\|P_{t+1} - P_t\|}{\|P_t\|}$ in the iterative process, and the range of t was set to $t \in \{1, 2, 3, \ldots, 28, 29, 30\}$. The change of relative change index with the number of iterations is shown in Fig. 3. We find that after 15 iterations, the state matrix tends to be stable, which also indicates that the SNF algorithm has good convergence. In this paper, we set the number of iterations $t = 20$ to ensure sufficient convergence.

Parameter settings

The parameters have a great influence on the performance of the algorithm. There are two parameters in SNFHGILMI, *i.e.* k and λ. In this section, we will discuss the influence of parameters on the performance of the SNFHGILMI model in a 5-fold cross-validation experiment. K is defined as the number of neighbors. For the selection of parameter k, we set the range $k \in \{1, 2, 3, \ldots, 10\}$. The AUC value of model SNFHGILMI keeps a small fluctuation range from 0.001 to 0.005, we finally choose parameter $k = 3$. λ is a decay factor, and the range of λ is set to $\lambda \in \{0.1, 0.2, 0.3, \ldots, 0.9\}$. The change of the AUC value of the model SNFHGILMI with the parameter λ is shown in Fig. 4. When $\lambda = 0.1$, the model has high prediction performance.

Performance evaluation for SNFHGILMI

5-foldCV (5-fold cross validation) and LOOCV (Leave One Out cross validation) are two popular methods to evaluate the performance of the model. In this paper, these two methods are

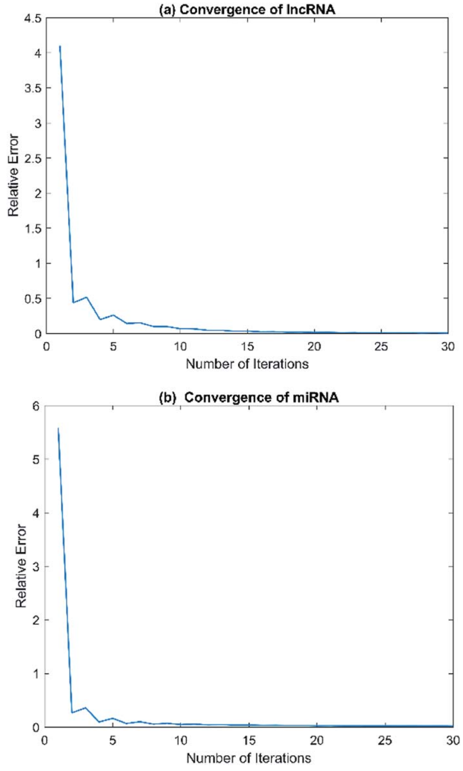

Fig. 3 The relative errors of SNF model with different numbers of iteration t. (a) Convergence of lncRNA. (b) Convergence of miRNA.

also used to evaluate the performance of our proposed computational model SNFHGILMI. Area under the Receiver Operator Characteristics curve (AUC) is an important metrics to evaluate classification models. In addition, there are several popular metrics, *i.e.* sen, acc, F 1-score.

For LOOCV, each known lncRNA–miRNA interaction is taken as a test sample, and all remaining known interactions are used as training samples to construct a prediction model. We compare and rank the test sample prediction scores with the all unknown interaction lncRNA–miRNA in the dataset. If the score is higher than the given threshold, we will consider the test sample to be a positive sample. For each of the different thresholds set in the experiment, we obtained the corresponding true positive rate (TPR, sensitivity) and false positive rate (FPR, specificity). As shown in Fig. 5, we plot the TPR and FPR values under different thresholds to obtain the ROC curve (Receiver Operating Curves), and we calculate the value of AUC. Implementing LOOCV on the SNFHGILMI, we have an AUC value of 0.9501. For 5-foldCV, the lncRNA–miRNA interaction was randomly divided into equal 5 subsets. Then, one subset was selected as the test sample in turn, and the remaining four subsets were used as the training samples to construct the prediction model. In order to avoid the error caused by randomly dividing data, the calculation model SNFHGILMI was verified by 5-foldCV for 100 times, and the average of AUC was used as the final prediction performance. As shown in Fig. 5, SNFHGILMI's average AUC values are 0.9426 ± 0.0015.

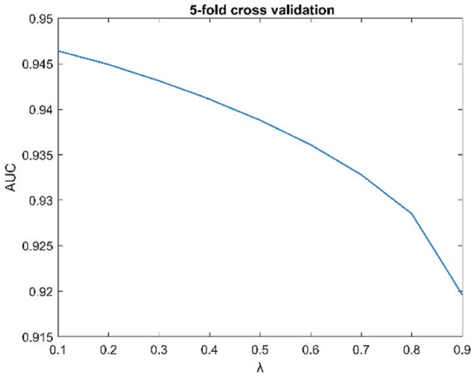

Fig. 4 The influence of parameters λ on model performance.

Comparison among different similarity network integrate methods

The fusion of similarity networks is a very important part of the model construction process. In order to prove the superiority of the SNF method, the model was constructed using SLF and SNF, respectively. The performance of the above two models on the benchmark dataset was evaluated using 5-foldCV. As shown in Fig. 6, the model based on the SNF is significantly better than the model based on the SLF, which shows that SNF can share and complement the kernel matrix information more effectively and improve the prediction performance of the model. In addition, model performance based on SNF and SLF methods is superior to model performance based on single sequence information or single Gaussian interaction profile kernel information, which also illustrates the importance of similarity network fusion from another aspect.

Compared with other methods

To further demonstrate the good performance of the model SNFHGILMI. We compared SNFHGILMI method with other state-of-the-art methods and some network-based methods. Among them, EPLMI calculated the expression profiles similarity of lncRNA and miRNA, and constructs a model based on the network two-way diffusion algorithm. INLMI integrates sequence similarity and expression profile similarity of lncRNA and miRNA, and constructs

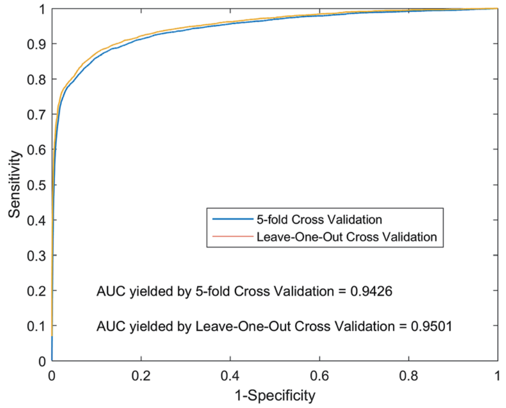

Fig. 5 Performance comparison by using SNFHGILMI in the framework of 5-foldCV and LOOCV.

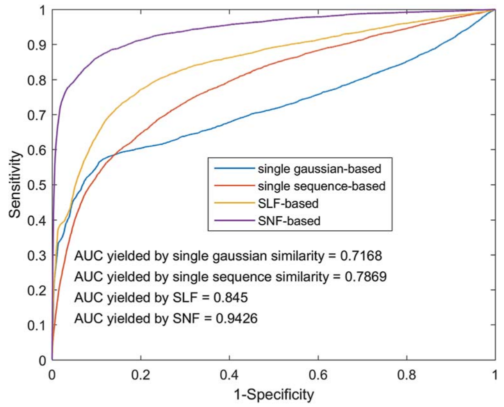

Fig. 6 Performance comparison among different similarity network integrates methods by using heterogeneous graph.

a model using non-negative matrix factorization. In addition, we also compared SNFHGILMI method with some classical network propagation algorithms such as RWR (Random Walk with Restart),[63] LncCF (LncRNA-based collaborative filtering) and MiCF (MiRNA-based collaborative filtering) that were widely used in bioinformatics.[64] RWR simulates a random walker that randomly transitions from the current seed node to the neighbor nodes, and can return to the start node with a given restart probability. LncCF and MiCF are based on the same assumption that if a lncRNA (miRNA) interacts with a miRNA (lncRNA), similar lncRNAs (miRNAs) tend to interact with this miRNA (lncRNA). On the benchmark dataset, 5foldCV was used for evaluation of all model's performance, and the results are shown in Table 1. Our method achieved an AUC value of 0.9457, which was significantly higher than EPLMI, INLMI, RWR, LncCF, and MiCF. The results show that SNFHGILMI has reliable and effective prediction performance.

Case studies

We conducted case studies in known lncRNA–miRNA interaction to further validate the predictive capability of the model. Here, lnc-CHSY1-5:1 and hsa-miR-17-5p were selected as candidate prediction objects, respectively. The lncRNA lnc-CHSY1-5:1 is associated with breast cancer, which is the leading cause of cancer death in women worldwide.[65] The miRNA hsa-miR-17-5p is associated with a variety of cancers and can act as both oncogene and suppressor genes in different cellular environments.[66] As shown in Table 2, for lncRNA lnc-CHSY1-5:1, all miRNAs are sorted according to the interaction score predicted by the model, and the top miRNAs will have a greater probability of interacting with lnc-CHSY1-5:1. We selected the top 10 potential miRNAs, and six of them have been confirmed by biochemical experiments to be searched in starBasev2.0 and lncRNASNP2. Similarly, as shown in Table 3, for miRNA hsa-mir-17-5p, seven of the top 10 potential lncRNAs have been confirmed by biochemical experiments. This shows that our model can effectively predict new potential interactions.

Table 1 Performance comparison among different methods by using SNFHGILMI in the framework of 5-foldCV

Method	AUC	SEN	ACC	F1
EPLMI	0.8451	0.1343	0.9940	0.1078
INLMI	0.8523	0.1541	0.9938	0.1085
RWR	0.9231	0.3841	0.9951	0.4283
LncCF	0.7847	0.3051	0.9963	0.4403
MiCF	0.8727	0.2823	0.9932	0.2824
SNFHGILMI	0.9457	0.4882	0.9959	0.5318

Table 2 Top 10 predictions for lnc-CHSY1-5:1 by SNFHGILMI

No.	LncRNA	Confirmed?
1	lnc-SLTM-3:1	YES
2	lnc-CPT2-3:1	
3	lnc-LRIG1-2:1	
4	lnc-FAS-1:1	YES
5	lnc-CALCOCO2-3:1	YES
6	lnc-MYC-2:16	
7	lnc-KB-1507C5.2.1-3:3	YES
8	lnc-ACER2-1:1	YES
9	lnc-RPGRIP1L-1:1	YES
10	lnc-PIGM-1:1	YES

Table 3 Top 10 predictions for hsa-miR-17-5p by SNFHGILMI

No.	MiRNA	Confirmed?
1	hsa-miR-424-5p	YES
2	hsa-miR-195-5p	YES
3	hsa-miR-15a-5p	YES
4	hsa-miR-15b-5p	YES
5	hsa-miR-485-5p	
6	hsa-miR-24-3p	
7	hsa-miR-421	
8	hsa-miR-27a-3p	YES
9	hsa-miR-155-5p	
10	hsa-miR-27b-3p	YES

Conclusions

LncRNA–miRNA interactions can reveal various biological functions and mechanisms. Therefore, the study of lncRNA–miRNA interactions is an important direction. At present, only some experimental data and a few calculation methods still have great limitations on the study of lncRNA–miRNA. It is urgent to propose more computational methods to predict novel lncRNA–miRNA interactions. Compare with other state-of-the-art methods, the SNFHGILMI method proposed in this work has the following advantages: (1) it integrates similarity data using a nonlinear network fusion technology, which not only characterizes the valuable information of each individual network but also shows better performance when combining two or more networks without obvious linear relationship; (2) all the information about lncRNA, miRNA and lncRNA–miRNA interactions are expressed on the heterogeneous graph, and only the links of different types of nodes need to be derived; (3) the heterogeneous graph inference algorithm can be used to predictions for new lncRNAs which have no known related miRNAs and miRNAs without any known associated lncRANs; (4) it converted iterative process into matrix multiplications, which can predict potential interactions score for all lncRNA–miRNA pairs simultaneously. The performance of the model was evaluated by LOOCV and 5-foldCV on the benchmark dataset. The experimental results show that our method has excellent performance in predicting lncRNA–miRNA interactions.

Although SNFHGILMI showed good performance in lncRNA–miRNA prediction as described above, it still has some limitations. SNFHGILMI used two lncRNA-related and two miRNA-related information to integrate the lncRNA/miRNA similarity matrices, we hope that more information (*e.g.*, lncRNA/miRNA functional annotations, lncRNA/miRNA expression profile) will be used to integrate similarity matrices. In addition, the SNFHGILMI relies on known interactions, we expect lncRNA and miRNA are well studied further to better predict lncRNA–miRNA interactions.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 61762026, 61462018), the Guangxi Provincial Natural Science Foundation (Grant No. 2017GXNSFAA198278), the Innovation Project of GUET Graduate Education (Grant No. 2018YJCX47 and 2019YCXS056), the Guangxi Key Laboratory of Trusted Software (Grant No. kx201403), the Guangxi Colleges and Universities Key Laboratory of Intelligent Processing of Computer Images and Graphics (Grant No. GIIP201502).

References

1 K. Adelman and E. Egan, *Nature*, 2017, **543**, 183.
2 S. Yamamura, M. Imai-Sumida, Y. Tanaka and R. Dahiya, *Cell. Mol. Life Sci.*, 2018, **75**, 467–484.
3 Q. Liao, D. Bu, L. Sun, H. Luo and Y. Zhao, in *Health Informatics Data Analysis*, Springer, 2017, pp. 51–60.
4 J. Li, H. Tian, J. Yang and Z. Gong, *DNA Cell Biol.*, 2016, **35**, 459–470.
5 J. M. Engreitz, J. E. Haines, E. M. Perez, G. Munson, J. Chen, M. Kane, P. E. McDonel, M. Guttman and E. S. Lander, *Nature*, 2016, **539**, 452.
6 L. Chen, Y.-H. Zhang, X. Pan, M. Liu, S. Wang, T. Huang and Y.-D. Cai, *Int. J. Mol. Sci.*, 2018, **19**, 3416.
7 Y.-A. Huang, X. Chen, Z.-H. You, D.-S. Huang and K. C. Chan, *Oncotarget*, 2016, **7**, 25902–25914.
8 X. Pan, L. J. Jensen and J. Gorodkin, *Bioinformatics*, 2018, **35**, 1494–1502.
9 B. Panwar, G. S. Omenn and Y. Guan, *Bioinformatics*, 2017, **33**, 1554–1560.
10 L. F. Gebert and I. J. MacRae, *Nat. Rev. Mol. Cell Biol.*, 2019, **20**, 21–37.
11 X. Pan and H.-B. Shen, *bioRxiv*, 2019, 666719.
12 X. Chen, D. Xie, L. Wang, Q. Zhao, Z.-H. You and H. Liu, *Bioinformatics*, 2018, **34**, 3178–3186.
13 X. Chen, D. Xie, Q. Zhao and Z.-H. You, *Briefings Bioinf.*, 2019, **20**, 515–539.
14 X. Chen, C. C. Yan, X. Zhang and Z.-H. You, *Briefings Bioinf.*, 2017, **18**, 558–576.
15 X. Chen, Y.-Z. Sun, N.-N. Guan, J. Qu, Z.-A. Huang, Z.-X. Zhu and J.-Q. Li, *Briefings Funct. Genomics*, 2019, **18**, 58–82.
16 M. K. Atianand and K. A. Fitzgerald, *Trends Mol. Med.*, 2014, **20**, 623–631.
17 G. Militello, T. Weirick, D. John, C. Döring, S. Dimmeler and S. Uchida, *Briefings Bioinf.*, 2016, **18**, 780–788.
18 D. Li, J. Ainiwaer, I. Sheyhiding, Z. Zhang and L. Zhang, *Eur. Rev. Med. Pharmacol. Sci.*, 2016, **20**, 2285–2295.
19 M. D. Paraskevopoulou and A. G. Hatzigeorgiou, in *Long Non-Coding RNAs*, Springer, 2016, pp. 271–286.
20 J. Liz and M. Esteller, *Biochim. Biophys. Acta, Gene Regul. Mech.*, 2016, **1859**, 169–176.
21 G. Ma, M. Tang, Y. Wu, X. Xu, F. Pan and R. Xu, *Am. J. Transl. Res.*, 2016, **8**, 5141.
22 Y. Mao, R. Liu, H. Zhou, S. Yin, Q. Zhao, X. Ding and H. Wang, *Cancer Gene Ther.*, 2017, **24**, 267.
23 M. Li, L. Duan, Y. Li and B. Liu, *Life Sci.*, 2019, **233**, 116440.
24 Y.-A. Huang, K. C. Chan and Z.-H. You, *Bioinformatics*, 2017, **34**, 812–819.
25 P. Hu, Y.-A. Huang, K. C. Chan and Z.-H. You, *International Conference on Intelligent Computing*. Springer, Cham, 2018.
26 W. Zhang, G. Tang, S. Wang, Y. Chen, S. Zhou and X. Li, *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018.
27 W. Zhang, G. Tang, S. Zhou and Y. Niu, *BMC Genomics*, 2019, **20**, 1–12.
28 L. Wong, Y. A. Huang, Z. H. You, Z. H. Chen and M. Y. Cao, *J. Cell. Mol. Med.*, 2020, **24**, 79–87.
29 H. Liu, G. Ren, H. Chen, Q. Liu, Y. Yang and Q. Zhao, *Knowl. Base Syst.*, 2020, **191**, 105261.
30 J. Yu, Z. Xuan, X. Feng, Q. Zou and L. Wang, *BMC Bioinf.*, 2019, **20**(1), 396.
31 G. Fu, J. Wang, C. Domeniconi and G. Yu, *Bioinformatics*, 2018, **34**, 1529–1537.
32 C. Lu, M. Yang, F. Luo, F.-X. Wu, M. Li, Y. Pan, Y. Li and J. Wang, *Bioinformatics*, 2018, **34**, 3357–3364.
33 X. Chen, L. Y. Wang and L. Huang, *J. Cell. Mol. Med.*, 2018, **22**, 2884–2895.
34 X. Chen, Y. Gong, D. Zhang, Z. You and Z. e. Li, *J. Cell. Mol. Med.*, 2018, 472–485.
35 X. Chen, Q.-F. Wu and G.-Y. Yan, *RNA Biol.*, 2017, **14**, 952–962.
36 C. Liang, S. Yu and J. Luo, *PLoS Comput. Biol.*, 2019, **15**, e1006931.
37 X. Chen, L. Huang, D. Xie and Q. Zhao, *Cell Death Dis.*, 2018, **9**, 1–16.
38 X. Chen and L. Huang, *PLoS Comput. Biol.*, 2017, **13**, e1005912.

39 X. Chen, J. Yin, J. Qu and L. Huang, *PLoS Comput. Biol.*, 2018, **14**, e1006418.

40 X. Chen, L. Wang, J. Qu, N.-N. Guan and J.-Q. Li, *Bioinformatics*, 2018, **34**, 4256–4265.

41 Q. Zhao, Y. Zhang, H. Hu, G. Ren, W. Zhang and H. Liu, *Front. Genet.*, 2018, **9**, 239.

42 W. Zhang, Q. Qu, Y. Zhang and W. Wang, *Neurocomputing*, 2018, **273**, 526–534.

43 W. Zhang, X. Yue, G. Tang, W. Wu, F. Huang and X. Zhang, *PLoS Comput. Biol.*, 2018, **14**, e1006616.

44 N. Zong, H. Kim, V. Ngo and O. Harismendy, *Bioinformatics*, 2017, **33**, 2337–2344.

45 X. Chen, C. C. Yan, X. Zhang, X. Zhang, F. Dai, J. Yin and Y. Zhang, *Briefings Bioinf.*, 2016, **17**, 696–712.

46 A. Ezzat, M. Wu, X.-L. Li and C.-K. Kwoh, *Briefings Bioinf.*, 2019, **20**, 1337–1357.

47 R. S. Olayan, H. Ashoor and V. B. Bajic, *Bioinformatics*, 2018, **34**, 1164–1173.

48 W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang and F. Liu, *BMC Bioinf.*, 2018, **19**, 233.

49 W. Zhang, X. Yue, F. Huang, R. Liu, Y. Chen and C. Ruan, *Methods*, 2018, **145**, 51–59.

50 W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang and F. Liu, *BMC Bioinf.*, 2018, **19**, 1–12.

51 W. Zhang, K. Jing, F. Huang, Y. Chen, B. Li, J. Li and J. Gong, *Inf. Sci.*, 2019, **497**, 189–201.

52 W. Zhang, Z. Li, W. Guo, W. Yang and F. Huang, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2019, DOI: 10.1109/TCBB.2019.2931546.

53 B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains and A. Goldenberg, *Nat. Methods*, 2014, **11**, 333.

54 X. Chen, C. C. Yan, X. Zhang, Z.-H. You, Y.-A. Huang and G.-Y. Yan, *Oncotarget*, 2016, **7**, 65257–65269.

55 J. Yin, X. Chen, C.-C. Wang, Y. Zhao and Y.-Z. Sun, *Mol. Pharmaceutics*, 2019, **16**, 3157–3166.

56 J. Gong, W. Liu, J. Zhang, X. Miao and A.-Y. Guo, *Nucleic Acids Res.*, 2014, **43**, D181–D186.

57 P.-J. Volders, K. Verheggen, G. Menschaert, K. Vandepoele, L. Martens, J. Vandesompele and P. Mestdagh, *Nucleic Acids Res.*, 2014, **43**, D174–D180.

58 A. Kozomara and S. Griffiths-Jones, *Nucleic Acids Res.*, 2013, **42**, D68–D73.

59 J.-L. Chen, M. A. Blasco and C. W. Greider, *Cell*, 2000, **100**, 503–514.

60 V. Likic, *Lecture given at the 7th Melbourne Bioinformatics Course,*Bi021 Molecular Science and Biotechnology Institute, University of Melbourne, 2008, pp. 1–46.

61 Z.-H. You, Z.-A. Huang, Z. Zhu, G.-Y. Yan, Z.-W. Li, Z. Wen and X. Chen, *PLoS Comput. Biol.*, 2017, **13**, e1005455.

62 T. van Laarhoven, S. B. Nabuurs and E. Marchiori, *Bioinformatics*, 2011, **27**, 3036–3043.

63 I. Lee and H. Nam, *BMC Bioinf.*, 2018, **19**, 208.

64 X. Zheng, Y. Wang, K. Tian, J. Zhou, J. Guan, L. Luo and S. Zhou, *BMC Bioinf.*, 2017, **18**, 420.

65 Q. Yao, L. Wu, J. Li, L. guang Yang, Y. Sun, Z. Li, S. He, F. Feng, H. Li and Y. Li, *Sci. Rep.*, 2017, **7**, 39516.

66 N. Cloonan, M. K. Brown, A. L. Steptoe, S. Wani, W. L. Chan, A. R. Forrest, G. Kolle, B. Gabrielli and S. M. Grimmond, *Genome Biol.*, 2008, **9**, R127.