A Framework for integrating multiple biological networks to predict microRNA-disease associations

Wei Peng *, Wei Lan, Zeng Yu, Jianxin Wang, *Senior Member,* and Yi Pan *, *Senior Member*

***Abstract*—MicroRNAs have close relationship with human diseases. Therefore, identifying disease related MicroRNAs plays an important role in disease diagnosis, prognosis and therapy. However, designing an effective computational method which can make good use of various biological resources and predict the associations between MicroRNA and disease correctly is still a big challenge. Previous researchers have pointed out that there are complex relationships among microRNAs, diseases and environment factors. There are inter-relationships between microRNAs, diseases or environment factors based on their functional similarity or phenotype similarity or chemical structure similarity and so on. There are also intra-relationships between microRNAs and disease, microRNAs and environment factors,diseases and environment factors. Moreover, functionally similar microRNAs tend to associate with common diseases and common environment factors. The diseases with similar phenotypes are likely caused by common microRNAs and common environment factors. In this work, we propose a framework namely ThrRWMDE which can integrate these complex relationships to predict microRNA-disease associations. In this framework, microRNA similarity network(MFN), disease similarity network(DSN) and environmental factor similarity network(ESN) are constructed according to certain biological properties. Then, an unbalanced three random walking algorithm is implemented on the three networks so as to obtain information from neighbors in corresponding networks. This algorithm not only can flexibly infer information from different levels of neighbors with respect to the topological and structural differences of the three networks, but also in the course of working the functional information will be transferred from one network to another according to the associations between the nodes in different networks. The results of experiment show that our method achieves better prediction performance than other state-of-the-art methods.**

***Index Terms*—MicroRNA, Disease, Environment Factor, association, Random walking**

I. Introduction

MicroRNA(MiRNA) is a class of single-stranded, non-coding RNA molecules with approximate 22 nucleotides in length. They bind the 3'UTR of targeted mRNA to negatively regulate its expression and prevent protein production. MicroRNAs involve in almost every biological process, including cell cycle, growth, death, stem cell differentiation and stress response. A large number of evidences show that dysregulation of microRNA is associated with a wide range of diseases, including cancers. Analyzing the alteration of expression profiles of disease related microRNAs can help us to predict the stage and progress, prognosis and response to treatment of certain disease. For example, a novel screening assay that is based on microRNA expression profile is undergoing a clinical trial to detect early-stage colorectal cancer [1]. With the development of experiment techniques, plenty of human microRNAs have been detected and sequenced. However, many associations between these microRNAs and diseases are still undetected.

Recently, some computational methods have been proposed to determine the associations between microRNAs and diseases. Most of these methods are based on the idea that the candidate microRNAs can be inferred from the known disease microRNA according to their similarities. These methods usually exploit network to describe the relationship between microRNAs or diseases, or them both. After that, some algorithms are designed based on the networks to predict new microRNA-disease associations. The early methods usually integrate multiple biological sources into a single network and implement prediction. Jiang et al. [2] have proposed first microRNA-disease association prediction method, which uses naive bayes to integrate gene function information, gene expression profile and protein domain interaction to construct a network where the nodes are genes and the edges are their functional associations. Then they calculate a score for each microRNA for an interested disease according to the functional associations between the known disease genes and the target genes of the microRNAs. Xuan et al. [3] have constructed a microRNA similarity network according to the similarity of disease the microRNA associated. Then they selected the weighted k most similar neighbors to predict disease-related microRNAs. Considering previous method only use local network information to predict microRNA-disease association, Chen et al. [4] have constructed a microRNA-microRNA functional similarity network and implemented a Random walk with Restart method (RWRMDA) on the whole network to predict microRNA-disease associations.

The single network-based methods usually integrate all biological information into one network. However, they either ignore the associations between diseases or ignore the difference between the networks that are constructed according to different biological information. Recently, some methods have been designed based on two networks. Shi et al. [5] have mapped microRNA targeted genes and disease related genes to protein-protein interaction (PPI)network respectively. They obtained two ranked list of genes by random walk with restart algorithm with different seeds. Then they used the p-value to

Wei Peng was with Computer center, Kunming University of Science and Technology,Kunming,650050,China,e-mail: (weipeng1980@gmail.com).

Wei Lan and Jianxin Wang are with School of Information Science and Engineering, Central South University, Changsha,410083, China.

Zeng Yu and Yi Pan are with Department of Computer Science, Georgia State University,Atlanta,GA,30302-4110,USA.

*Corresponding author

Manuscript received April 19, 2005; revised August 26, 2015.

measure the significant that a microRNA is associated with a disease. Chen et al. [6] have constructed two microRNA similarity networks from two different perspectives. The one is on the basis of the functional similarity of microRNA, the other is on the basis of the phenotype similarity of the microRNA associated diseases. They calculated the Pearson correlation scores between the two types of similarities to infer the associations between microRNA and disease. After that, the same group [7] have constructed two networks, microRNA functional similarity network and disease semantic similarity network. A semi-supervised and global method, namely RLSMDA is implemented on the two networks simultaneously to predict potential microRNA-disease associations. As we all known, the key problem for the network-based methods is the way to construct network and the algorithm implemented on the network. There are different ways to calculate similarity and construct networks, i.e. microRNA functional similarity, microRNA sequence similarity, disease phenotype similarity, disease function similarity, disease semantic similarity. [8] and [9] have discussed the main similarity computation methods in predicting microRNA-disease associations.

To further improve the prediction performance, researchers design effective methods to integrate multiple similarities to get better prediction performance, i.e kernelized Bayesian matrix factorization(KBMFMDI) method [10], [11] or introduce more biological known knowledge, i.e environment factors. Recent studies show that microRNA expression can be regulated by environment factors (EF) [12], such as drug, alcohol, diet, stress, etc. Some environment factors may also cause diseases [8]. Qiu et al. [13] make use of microRNA-EF interaction patterns to predict new EF-disease associations. Chen et al. [14] have incorporated environmental factors to predict EF-microRNA associations. Li et al. [15]have employed EF to predict microRNA networks. Ha et al. [16] have combined EF data to predict microRNA-disease associations. They construct three different microRNA networks, according to microRNA functional similarity, whether or not share common diseases and whether or not share common Environment Factors, respectively. A simple strategy is adopted to integrate the three networks into a single network and a random walk model is implemented on the network to make prediction. This method is a single network-based method that ignores the associations between Environment Factors, the associations between diseases and the associations between environment factors and disease.

In fact, there are complex associations among microRNA, disease and EF. As Fig. 1 shown, there are inter-relationships between microRNAs, diseases or EFs based on their functional similarity or phenotype similarity or chemical structure similarity and so on. There are also intra-relationships between microRNAs and disease, microRNAs and environment factors, diseases and environment factors. Moreover, previous researches indicate following observations: 1) functionally similar microRNAs tend to associate with common diseases and common environment factors [6], 2) The diseases with similar phenotypes are likely caused by common microRNAs and common environment factors [13]. Consequently, in this work, we propose a framework namely ThrRWMDE which can integrate these complicated relationships to predict microRNA-disease associations. In this framework, three types of different biological networks are constructed. They are microRNA similarity network(MFN), disease similarity network(DSN) and environmental factor similarity network(ESN), where nodes are microRNAs, diseases and environmental factors respectively, and edges correspond to the similarities between the nodes. Each type of biological network can be constructed by a single similarity computation method or by integrating different ones. Based on the three networks, an unbalanced three random walking algorithm [17], [18]is implemented so as to obtain information from neighbors in corresponding networks. This algorithm not only can flexibly infer information from different levels of neighbors with respect to the topological and structural differences of the three networks, but also in the course of working the functional information will be transferred from one network to another according to the associations between the nodes in different networks. The results of experiment show that our method achieves better prediction performance than other three state-of-the-art methods KBMFMDI [10], RLSMDA [7] and Ha's method [16] in terms of area under the curves(AUC).

II. METHODS

A. Experimental data

The known human microRNA-disease association data is from [19], which is also downloaded from HMDD database [20]. The dataset includes 271 microRNAs, 137 diseases and 1395 miRNA-disease interactions.

The microRNA similarity network (namely microRNA functional similarity network) can be constructed based on microRNA functional similarity data which is downloaded from http://www.cuilab.cn/misim.zip [19]. The microRNA similarity network (namely microRNA sequence similarity network)can also be constructed based on microRNA sequence data which is downloaded from miRBase database [21]. The sequence similarity between two miRNAs is calculated by Emboss-Needle tool with default parameter values (Needleman-Wunsch alignment algorithm) [22].

The disease similarity network (namely disease functional similarity network) can be constructed based on disease function similarity which is calculated by SemFunSim [23]. This method assumed that similar diseases tend to be related to genes with similar functions. The gene functional similarity data is extracted from HumanNet dataset(http://www.functionalnet.org/humannet/) [24], which has assigned a functional similarity score for each pair of gene. The mapping between disease and gene are obtained from SIDD database [25]. The disease similarity network (namely disease semantic similarity network) can be constructed based on the semantic associations between different diseases which are downloaded from Disease Ontology database [26]. The methods in [10], [11] are adopted to calculate the semantic similarity between different diseases.

The environmental factor similarity network is constructed based on the chemical structure similarity of environmental factor. The chemical structure similarities are downloaded

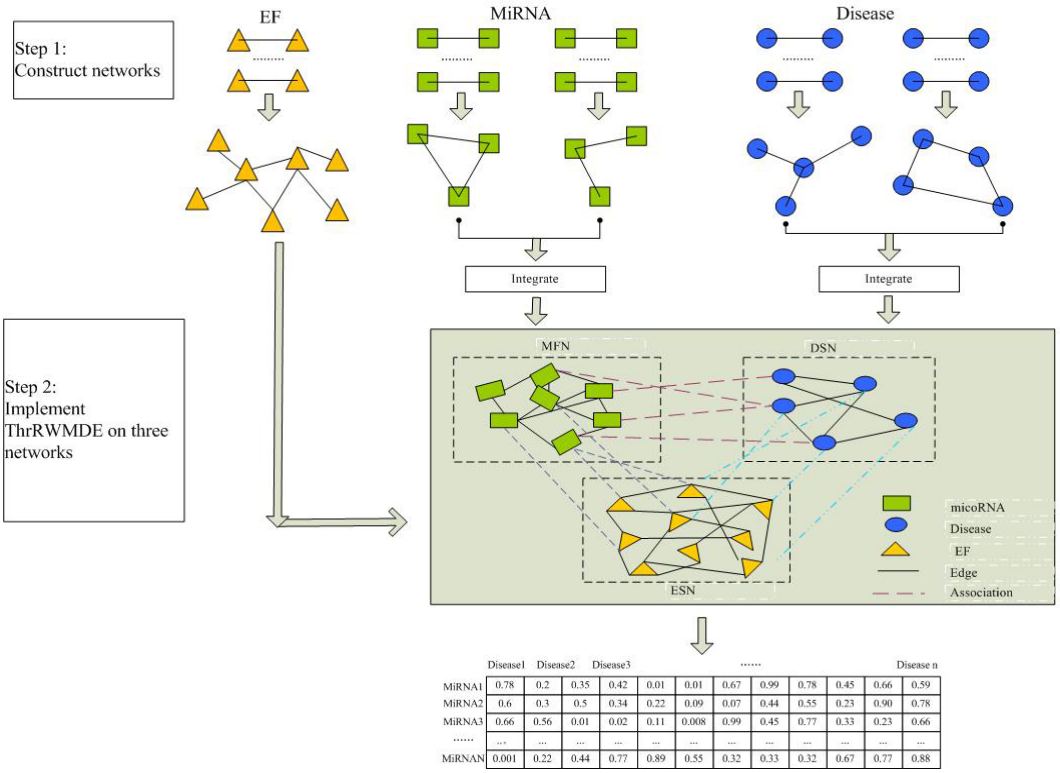


Fig. 1: The workflow of ThrRWMDE

from supplemental material of [14], which are calculated by SIMCOMP [27]. There are 138 EFs in EF similarity network. The association between the EFs and microRNAs, EFs and diseases are downloaded from miREnironment database. There are 1019 associations between the 138 EFs and the 271 microRNAs. There are 978 associations between the 138 EFs and the 137 diseases.

Since miREnironment database collects the associations between EFs and disease phenotypes. We map the phenotypes in miREnironment to the disease in HMDD database by using the information downloaded from Disease Ontology (http://aber-owl.net/aber-owl/diseasephenotypes/data/) and Medical Subject Headings(MeSH, http://www.nlm.nil.gov)

B. Three random walk algorithm on three types of biological networks

As Fig. 1 shown, ThrRWMDE method mainly takes two steps to predict the associations between microRNA and disease. Firstly, three types of different biological networks are constructed. They are microRNA similarity network(MSN), disease similarity network (DSN) and environmental factor similarity network (ESN). Secondly several random walk steps are taken in ESN, DSN and MSN iteratively so as to obtain the information of level-k neighbors in corresponding network. Moreover, the parameters in our method can control the walking steps on the three networks with respect to their difference topology and structure. In the course of iteration, some potential associations between microRNAs and diseases can not only be explored from the inter- and intra- associations between microRNAs and diseases but also be inferred according to the associations between microRNAs and EFs and the associations between EFs and diseases. To formally define our method, some variables are introduced.

Let $M(m*m)$, $D(d*d)$ and $E(e*e)$ be the adjacency matrixes of MSN, DSN and ESN respectively. Let matrix $Y1(m*d)$, $Y2(e*m)$ and $Y3(e*d)$ store known microRNA-disease associations, known EF-microRNA associations and EF-disease associations respectively. The values of elements in these matrixes are 1, if there exist associations between corresponding nodes, 0 otherwise. Matrix $Rmd(m*d)$ and $Red(e*d)$ denote the predicted microRNA-disease associations and predicted EF-disease associations respectively.

Our work aims to get matrix Rmd according to matrix M, D, E, $Y1$, $Y2$ and $Y3$. The values in matrix Rmd can be updated through three ways. Firstly, several random walk steps (l_1) are taken in MSN network to get disease related information from level- l_1 neighbors of microRNA (see Formula 1). Secondly, several random walk steps (r_1) are taken in DSN to get microRNA related information from level-r_1 neighbors of disease(see Formula 2). Thirdly, the disease association of EFs are passed through the known microRNA-EF associations (see Formula 3).

$$MSN : Rmd^t = \alpha M * Rmd^{t-1} + (1-\alpha) * Y1 \quad (1)$$

$$DSN : Rmd^t = \alpha Rmd^{t-1} * D + (1-\alpha) * Y1 \quad (2)$$

$$MicroRNA{-}EF\ associations : Rmd^t = Y2{*}Red^{t-1} \quad (3)$$

Similarly, the predicted EF-disease associations can also be got. The values in matrix Red are also updated in three

ways. some potential EF-disease association can be explored by extending EF path and disease path in ESN and DSN respectively (see Formula 4 and 5). The predicted EF-disease associations can also be updated by transfering the microRNA-disease association to EF through the association between microRNA and EF (see Formula 6).

$$ESN : Red^t = \alpha E * Red^{t-1} + (1-\alpha) * Y3 \quad (4)$$

$$DSN : Red^t = \alpha Red^{t-1} * D + (1-\alpha) * Y3 \quad (5)$$

$$EF - microRNA\ associations : Red^t = Y2^{'} * Rmd^{t-1} \quad (6)$$

In summary, Algorithm 1 outlines the algorithm of ThrRWMDE. Parameters l_1, r_1, l_2 and r_2 are used to control the steps walking in the three networks in the course of iteration. Their values can be easily set with respect to the difference among the three networks.

Algorithm 1 ThrRWMDE

Input: Matrix M, D, E, $Y1$,$Y2$, $Y3$, parameter α, iteration steps l_1, r_1, l_2 and r_2, testing set S;
Output: predicted association matrix Rmd, Red ;
Clear the values i, j of matrix $Y1$, if i, j in S
Rmd^0 = $Y1$ = $\frac{Y1}{sum(Y1)}$
Red^0 = $Y2$ = $\frac{Y2}{sum(Y2)}$
for ($t = 1$ to max(l_1 , r_1, l_2 , r_2)) **do**
  $\lambda_{m1} = \lambda_{d1} = \lambda_{e1} = \lambda_{m2} = \lambda_{d2} = \lambda_{e2} = 0$;
  if ($t <= l_2$) **then**
    $Red_e^t = \alpha * E * Red^{t-1} + (1-\alpha) * Y2$
    $Red_m^t = Y3 * Rmd^{t-1}$
    $\lambda_{e2} = 1$
    $\lambda_{m2} = 1$
  end if
  if ($t<=r_2$) **then**
    $Red_d^t = \alpha * Red^{t-1} * D + (1-\alpha) * Y2$
    $\lambda_{d2} = 1$
  end if
  $Red^t = ((\lambda_{e2} * Red_e^t + \lambda_{d2} * Red_d^t + \lambda_{m2} * Red_m^t)/(\lambda_{e2} + \lambda_{d2} + \lambda_{m2}))$
  if ($t<=l_1$) **then**
    $Rmd_m^t = \alpha * M * Rmd^{t-1} + (1-\alpha) * Y1$
    $Rmd_e^t = Y2^{'} * Red^{t-1}$
    $\lambda_{m1} = 1$
    $\lambda_{e1} = 1$
  end if
  if ($t<=r_1$) **then**
    $Rmd_d^t = \alpha * Rmd^{t-1} * D + (1-\alpha) * Y1$
    $\lambda_{d1} = 1$
  end if
  $Rmd^t = ((\lambda_{m1} * Rmd_m^t + \lambda_{d1} * Rmd_d^t + \lambda_{e1} * Rmd_e^t)/(\lambda_{m1} + \lambda_{d1} + \lambda_{e1}))$
end for
return (Rmd,Red)

C. Evaluation metrics

To evaluate the performance of a method, five-fold cross validation is adopted, which randomly divides all known microRNA-disease associations into five parts, four of the five parts are regarded as training set and the left one is regarded as testing set. The predicting method calculates a likelihood score for each testing microRNA-disease association. For each disease, the testing microRNA-disease associations ranked in top t are regarded as predicted ones. The accuracy of prediction depends on how well the predicted associations match the real ones, which is measured by two widely used statistic metrics, true positive rate (TPR), false positive rate (FPR). TPR measures the percentage of predicted associations that match the known ones in testing set. FPR, on the other hand, defines the percent that negative MicroRNA-disease associations in the testing set are incorrectly predicted as positive ones. With different values of t selected, different pairs of FPR and TPR are calculated. Using FPR and TPR as x and y axes respectively, we can draw ROC curve. The area under the ROC curve (AUC) can be utilized to evaluate the results. The higher the AUC value is, the better the method performs. We repeat the five-fold cross validation 1000 times, the average AUC value over all diseases is adopted to evaluate the overall prediction performance of each method.

III. Results

In order to assess the effectiveness of ThrRWMDE, we compare it with other three methods, KBMFMDI [10], [11], RLSMDA [7] and Ha's method [16]. RLSMDA is a recent method which constructs two networks, microRNA functional similarity network and disease semantic similarity network to predict the associations between microRNAs and diseases. KBMFMDI currently has the best prediction performance, which combines multiple microRNA networks and disease networks to make prediction. Ha's method uses environment factors in a different way to predict the associations between microRNA and disease. For fair comparison, the parameters in RLSMDA, KBMFMDI and Ha's method are all selected according to the authors recommendation.

In this section, we first show the performance of ThrRWMDE, where MSN is constructed based on microRNA functional similarity, DSN is constructed based on disease functional similarity and ESN is constructed based on chemical structure similarity of environmental factor. Firstly, the effect of parameters on the performance of ThrRWMDE is discussed. Then the effectiveness of ThrRWMDE is evaluated by comparing with the other three existing methods. After that, some cases of certain disease are shown to further verify our method. Finally, we show the prediction performance when ThrRWMDE combines different similarity computation methods to construct the three types of biological networks.

A. Effect of parameters

ThrRWMDE introduces five parameters(α, l_1, r_1, l_2 and r_2). Parameter α controls the weight of the regulation of known microRNA-disease associations, EF-disease associations and EF-microRNA associations in the course of iteration. When α

TABLE II: Comparison of AUC scores over all diseases with respect to different l_1, r_1, l_2 and r_2 values.

$(l1, r1, l2, r2)$	AUC	$(l1, r1, l2, r2)$	AUC	$(l1, r1, l2, r2)$	AUC	$(l1, r1, l2, r2)$	AUC
$(1, 1, 1, 1)$	0.8279	$(2, 1, 1, 1)$	0.8293	$(3, 1, 1, 1)$	0.8239	$(100, 1, 1, 1)$	0.8449
$(1, 2, 1, 1)$	0.7533	$(2, 2, 1, 1)$	0.7961	$(3, 2, 1, 1)$	0.8313	$(1, 100, 1, 1)$	0.7291
$(1, 3, 1, 1)$	0.7407	$(2, 3, 1, 1)$	0.7474	$(3, 3, 1, 1)$	0.7971	$(100, 100, 1, 1)$	0.8178
$(1, 1, 1, 1)$	0.8279	$(1, 1, 2, 1)$	0.8284	$(1, 1, 3, 1)$	0.8272	$(1, 1, 100, 1)$	0.8286
$(1, 1, 1, 2)$	0.8282	$(1, 1, 2, 2)$	0.8287	$(1, 1, 3, 2)$	0.8279	$(1, 1, 1, 100)$	0.8275
$(1, 1, 1, 3)$	0.8281	$(1, 1, 2, 3)$	0.8287	$(1, 1, 3, 3)$	0.8278	$(1, 1, 100, 100)$	0.8274
$(1, 100, 100, 1)$	0.7295	$(100, 1, 1, 100)$	0.8267	$(100, 1, 100, 1)$	0.8691	$(1, 100, 1, 100)$	0.7303
$(1, 100, 100, 100)$	0.729	$(100, 1, 100, 100)$	0.8641	$(100, 100, 1, 100)$	0.8152	$(100, 100, 100, 1)$	0.8189
$(100, 100, 100, 100)$	0.8193						

TABLE I: Comparison of AUC scores over all diseases with respect to different α values

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AUC	0.827	0.829	0.828	0.827	0.828	0.828	0.828	0.828	0.828

is set to 1, ThrRWMDE explores potential MicroRNA-disease associations and EF-disease associations without considering known ones. In order to test the effect of parameter α on performance of ThrRWMDE, we set α to different values ranging from 0.1 to 0.9 and the values of parameter l_1, r_1, l_2 and r_2 are set to 1. The average AUC values of ThrRWMDE with respect to different α are listed in Table 1. The results in Table 1 show that the change of α value has little influence on the predict performance. Consequently, in this work, we set the value of parameter α to 0.9.

Parameter l_1 and r_1 control the walking steps in MSN and DSN, which means that potential microRNA-disease associations are inferred from up to l_1 level neighbors in MSN and up to r_1 neighbors level in DSN. Similarly, parameter l_2 and r_2 decide the number of walking steps in ESN and DSN when exploring EF-disease associations. In ThrRWMDE, some potential EF-disease associations are deduced from up to l_2 level neighbors in ESN and up to r_2 level neighbors in DSN. Therefore, we can infer new potential microRNA-disease associations through EF-disease associations and EF-microRNA associations. To investigate how variation of l_1, r_1, l_2 and r_2 affects the prediction performance, we set them different values ranging from 1 to 3. We also test the performance when the values of l_1, r_1, l_2 and r_2 are extremely large, i.e. 100. As Table II shown, when fixing the values of r_1, l_2, r_2, the AUC values rise slightly with the increase of l_1. For example, the AUC value rises from 0.8279 to 0.8449 when l_1 changes from 1 to 100 and the values of r_1, l_2, r_2 are fixed to 1. On the other hand, when fixing the values of l_1, l_2, r_2, the AUC values drop with the increase of r_1. For example, the AUC value drop from 0.8279 to 0.7291 when r_1 changes from 1 to 100 and the values of l_1, l_2, r_2 are fixed to 1. As for l_2, r_2, changing their values has little effect on the performance of ThrRWMDE. As far as we see from the testing results, ThrRWMDE achieves the best performance when l_1, r_1, l_2 and r_2 are 100, 1, 100 and 1. Whenever r_1 rises to 100, the performance of ThrRWMDE drops sharply. This suggests that global information of MSN and ESN and local information of DSN are helpful to explore potential microRNA-disease associations. The reason may be the three networks have different topology and structure. MSN consists of 271 nodes and 28008 edges. Its path length is 1.23 and network density is 0.766. ESN consists of 80 nodes and 1997 edges. Its path length is 1.386 and network density is 0.632. DSN consists of 137 nodes and 9316 edges. Its path length is 1 and network density is 1. Obviously, DSN is very dense and is almost full-connected. Therefore, the local information of DSN is more useful for us. The difference may also come from the underlying biological properties. i.e. a cluster of microRNA tends to associate with a common disease [10], [11]. In the following comparison, ThrRWMDE is compared on the basis of its best performance.

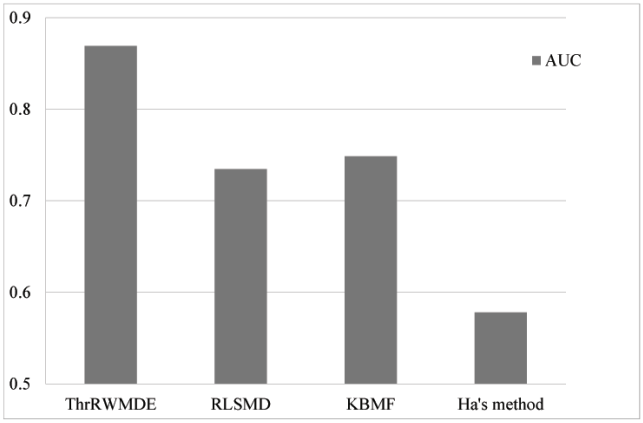


Fig. 2: Comparison of each method performance in terms of AUC values over all diseases.

B. Five-fold cross validation

We compare our method with RLSMD, KBMF and Ha's method to evaluate its effectiveness. All comparing methods are implemented on our experimental material mentioned above and undergo five-fold cross validation. In original Ha's method, gene expression profiles are adopted as initial values. However, we can not obtain suitable gene expression profiles for their methods. For fair comparison, we utilize known

microRNA-disease associations in training set to initialize the values of Ha's method and control its iteration. The same strategy is taken in our method. Fig. 2 shows the average AUC values of all comparing methods over all diseases. ThrRWMDE has the highest AUC value, which is 0.8691. Compared with KBMF(AUC is 0.7484), RLSMD(AUC is 0.7344) and Ha's method (AUC is 0.578), ThrRWMDE achieves 11%, 13%, 50% improvement than KBMF, RLSMD and Ha's method respectively.

We also focus on some diseases to further assess our method. Fig. 3 shows the AUC values comparison between each method based on some diseases, including Breast cancer, Hepatocellular carcinoma, Renal cell carcinoma, Glioblastoma, Acute T cell leukemia, Lung cancer, Multiple Myeloma, Germ cell cancer, Ovarian carcinoma, Pancreatic cancer and prostatic cancer. These disease have got common attentions from previous researches [7], [10], [28]. As the figure shown, ThrRWMDE has the highest AUC values for all selected diseases among all comparing methods. Notably, the AUC values of THrRWMDE for Glioblastoma, Acute T cell leukemia, Multiple Myeloma and Germ cell cancer was 0.9408, 0.9336, 0.9295, 0.9539, respectively. Its AUC values for Renal cell carcinoma, Hepatocellular carcinoma,Pancreatic cancer and prostatic cancer are 0.8937, 0.8531, 0.8608 and 0.856, respectively. As for Breast cancer, Lung cancer and Ovarian carcinoma, the average AUC values of ThrRWMDE is more than 0.75.

The outperformance of ThrRWMDE suggests it successfully integrate the inter-associations and intra-associations between micronRNAs, diseases and environment factors. Although Ha's method uses environment factor information, it ignores the relationship between the environment factors and between diseases. RLSMD and KBMF make good use of the inter- and intra-relationships between disease and miroRNA, however, they ignore the environment factors.

C. Case study

After fully evaluate the performance of ThrRWMDE, we use all known microRNA-disease associations as training set and implement ThrRWMDE to predict new microRNA-disease associations. The top 50 candidate mincroRNAs of breast cancer are listed in Table III. Breast cancer is the leading type female cancer and comprises of 25% of all cancers in women [29]. Among the top 50 candidate breast cancer related micro-RNA, 42 microRNAS (84%) can be confirmed that they have associations with breast cancer from HMDD, mir2disease, Pubmed and the supplementary materia of [4].

D. Prediction performance when combining different similarity computation methods to construct biological networks

In ThrRWMDE, the first step is constructing three types of different biological networks. Each type of biological network can be constructed by a single similarity computation method or by integrating different ones. After showing the outperformance of ThrRWMDE based on networks constructed by single similarity computation method, here, we present the performance of ThrRWMDE based on the networks that are constructed by integrating multiple similarity computation methods. A simple and intuitive method is adopted to combine the different networks. Let $M_1(m*m)$ and $M_2(m*m)$ be the adjacency matrixes of the networks constructed by microRNA functional similarity and microRNA sequence similarity respectively. Let $D_1(d*d)$ and $D_2(d*d)$ be the adjacency matrixes of the networks constructed by disease functional similarity and disease semantic similarity respectively. Let $M(m*m)$ and $D(d*d)$ be the adjacency matrixes of MSN and DSN. Therefore,

$$M = b * M\_norm\_1 + (1 - b) * M\_norm\_2 \quad (7)$$

$$D = c * D\_norm\_1 + (1 - c) * D\_norm\_2 \quad (8)$$

Where M_norm_1 and M_norm_2 are row normalization of M_1 and M_2, respectively. Parameters b and c control the weights of the two networks in the integrated networks. Here, we do not state our way of integrating networks is the most advanced. More efficient methods can be designed to integrate multiple networks. We set b and c to 0.9 and 0.5, respectively. After M and D are calculated, Algorithm 1 is implemented with the parameters as the above recommendation. After five-fold cross validation, the average AUC value of ThrRWMDE over all disease is 0.8712. Compared with the AUC value (0.8691) when it using one similarity computation method to construct network, the performance of ThrRWMDE gets slight improvement. It suggests that the performance of our method can be further improved if more efficient method is utilized to integrate multiply networks constructed by different similarity methods

IV. Conclusion

In this work, we propose a framework ThrRWMDE that integrates multiple biological data sources to predict microRNA-disease associations. It consists of two steps. The first step is constructing three types of networks, including MSN, DSN and ESN. Each type of the network can be constructed either by one similarity computation method or by integrating multiple similarity networks. In this step, some useful biological information, such as microRNA function, microRNA sequence, disease phenotype, environment factor chemical structure, can been merged into the networks. The second step is implementing unbalanced three random walking on the three types of network to explore new microRNA-disease associations or EF-disease associations by inferring information from the neighbors in the same networks or the nodes in the other networks. Compared to previous methods, our method makes good use of the inter- and intra-relationships between the microRNAs, diseases and environmental factors simultaneously. Moreover, our method is flexible to deal with the structure and topological difference of the three types of networks. Through analyzing the effect of parameters on the performance of ThrRWMDE, we observe that global information in MSN and local information in DSN are helpful to detect new microRNA-disease associations correctly. The experimental results show that ThrRWMDE outperforms other three existing methods and achieves higher AUC values over all diseases and some selected diseases. The potential breast

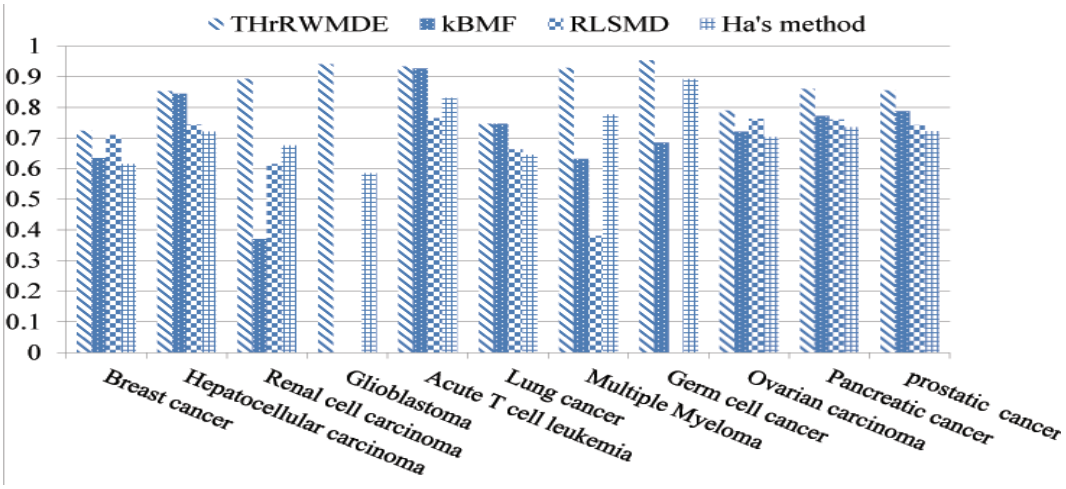


Fig. 3: Comparison of each method performance in terms of AUC values based on some diseases.

TABLE III: Top 50 potential breast cancer related micrRNA predicted by ThrRWMDE

Rank	*Name*	*Evidence*	*Rank*	*Name*	*Evidence*
1	$hsa-mir-532$	$PMID:24866763$	26	$hsa-mir-29c$	$HMDD$
2	$hsa-mir-521$		27	$hsa-mir-217$	
3	$hsa-let-7i$	$HMDD$	28	$hsa-mir-190$	$PMID:24865188$
4	$hsa-mir-92b$	$PMID:26878388$	29	$hsa-mir-128b$	$mir2disease$
5	$hsa-let-7g$	$HMDD$	30	$hsa-mir-335$	$HMDD$
6	$hsa-mir-101$	$PMID:25059472$	31	$hsa-mir-371$	
7	$hsa-mir-16$	$PMID:23741392$	32	$hsa-mir-26b$	$HMDD$
8	$hsa-let-7b$	$HMDD$	33	$hsa-mir-455$	
9	$hsa-mir-196b$		34	$hsa-mir-203$	$HMDD$
10	$hsa-mir-92a$	$literature$[4]	35	$hsa-mir-137$	$HMDD$
11	$hsa-mir-191$	$HMDD$	36	$hsa-mir-323$	
12	$hsa-mir-124$	$mir2disease$	37	$hsa-mir-372$	$PMID:25333260$
13	$hsa-mir-100$	$HMDD$	38	$hsa-mir-514$	
14	$hsa-mir-31$	$HMDD$	39	$hsa-mir-378$	$PMID:25120807$
15	$hsa-mir-126$	$HMDD$	40	$hsa-mir-182$	$HMDD$
16	$hsa-mir-99b$	$literature$[4]	41	$hsa-mir-224$	$HMDD$
17	$hsa-let-7c$	$HMDD$	42	$hsa-mir-491$	$PMID:25299770$
18	$hsa-mir-18b$	$HMDD$	43	$hsa-mir-181a$	$mir2disease$
19	$hsa-mir-373$	$HMDD$	44	$hsa-mir-183$	$HMDD$
20	$hsa-let-7e$	$HMDD$	45	$hsa-mir-421$	$PMID:23526361$
21	$hsa-mir-223$	$HMDD$	46	$hsa-mir-24$	$PMID:25120807$
22	$hsa-mir-106a$	$PMID:25541910$	47	$hsa-mir-128a$	$PMID:20054641$
23	$hsa-mir-15b$	$PMID:25783158$	48	$hsa-mir-96$	$HMDD$
24	$hsa-mir-122$	$HMDD$	49	$hsa-mir-144$	$PMID:26252024$
25	$hsa-mir-498$		50	$hsa-mir-32$	$PMID:26276160$

cancer related microRNAs predicted by our method also have been verified by database and literatures. In the future, combining more efficient methods that integrate multiple similarity networks can further improve the prediction performance of ThrRWMDE.

ACKNOWLEDGMENT

This work is supported in part by the National Natural Science Foundation of China under grant No.61502214, No.31560317, No.61472133, No.61502166, No.81460007 and No.81560221. Natural Science Foundation of Yunnan Province of China(No.2016FB107).

REFERENCES

[1] B. S. Nielsen, S. Jørgensen, J. U. Fog, R. Søkilde, I. J. Christensen, U. Hansen, N. Brünner, A. Baker, S. Møller, and H. J. Nielsen, "High levels of microrna-21 in the stroma of colorectal cancers predict short disease-free survival in stage ii colon cancer patients," *Clinical & experimental metastasis*, vol. 28, no. 1, pp. 27–38, 2011.

[2] Q. Jiang, Y. Hao, G. Wang, L. Juan, T. Zhang, M. Teng, Y. Liu, and Y. Wang, "Prioritization of disease microrna through a human phenome-microrna interactome network," *BMC Systems Biology*, vol. 4, no. Suppl 1, p. S2, 2010.

[3] P. Xuan, K. Han, M. Guo, Y. Guo, J. Li, J. Ding, Y. Liu, Q. Dai, J. Li, Z. Teng *et al.*, "Prediction of microrna associated with human diseases based on weighted k most similar neighbors," *PloS one*, vol. 8, no. 8, p. e70204, 2013.

[4] X. Chen, M.-X. Liu, and G.-Y. Yan, "Rwrmda: predicting novel human microrna–disease associations," *Molecular BioSystems*, vol. 8, no. 10, pp. 2792–2798, 2012.

[5] H. Shi, J. Xu, G. Zhang, L. Xu, C. Li, L. Wang, Z. Zhao, W. Jiang, Z. Guo, and X. Li, "Walking the interactome to identify human mirna-disease associations through the functional link between mirna targets and disease genes," *BMC systems biology*, vol. 7, no. 1, p. 101, 2013.

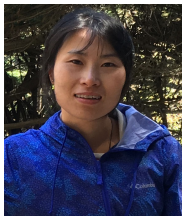
[6] H. Chen and Z. Zhang, "Similarity-based methods for potential human microrna-disease association prediction," *BMC medical genomics*, vol. 6, no. 1, p. 12, 2013.

[7] X. Chen and G.-Y. Yan, "Semi-supervised learning for potential human microrna-disease associations inference," *Scientific reports*, vol. 4, 2014.

[8] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microrna function and prioritizing disease-related microrna using biological interaction networks," *Briefings in bioinformatics*, p. bbv033, 2015.

[9] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation

strategies in the microrna-disease network: a survey," *Briefings in functional genomics*, vol. 15, no. 1, pp. 55–64, 2016.
[10] W. Lan, J. Wang, M. Li, J. Liu, and Y. Pan, "Predicting microrna-disease associations by integrating multiple biological information," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 183–188.
[11] W. Lan, J. Wang, M. Li, J. Liu, F. Wu, and Y. Pan, "Predicting microrna-disease associations based on improved microrna and disease similarities." *IEEE/ACM transactions on computational biology and bioinformatics/IEEE, ACM*, 2016.
[12] U. N. Das, "Obesity: genes, brain, gut, and environment," *Nutrition*, vol. 26, no. 5, pp. 459–473, 2010.
[13] C. Qiu, G. Chen, and Q. Cui, "Towards the understanding of microrna and environmental factor interactions and their relationships to human diseases," *Scientific reports*, vol. 2, 2012.
[14] X. Chen, M.-X. Liu, Q.-H. Cui, and G.-Y. Yan, "Prediction of disease-related interactions between micrornas and environmental factors based on a semi-supervised classifier," *PloS one*, vol. 7, no. 8, p. e43425, 2012.
[15] J. Li, Z. Wu, F. Cheng, W. Li, G. Liu, and Y. Tang, "Computational prediction of microrna networks incorporating environmental toxicity and disease etiology," *Scientific reports*, vol. 4, 2014.
[16] J. Ha, H. Kim, Y. Yoon, and S. Park, "A method of extracting disease-related micrornas through the propagation algorithm using the environmental factor based global mirna network," *Bio-Medical Materials and Engineering*, vol. 26, no. s1, pp. S1763–S1772, 2015.
[17] W. Peng, J. Wang, L. Chen, J. Zhong, Z. Zhang, and Y. Pan, "Predicting protein functions by using unbalanced bi-random walk algorithm on protein-protein interaction network and functional interrelationship network," *Current Protein and Peptide Science*, vol. 15, no. 6, pp. 529–539, 2014.
[18] W. Peng, M. Li, L. Chen, and L. Wang, "Predicting protein functions by using unbalanced random walk algorithm on three biological networks," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 99, p. 1, 2015.
[19] D. Wang, J. Wang, M. Lu, F. Song, and Q. Cui, "Inferring the human microrna functional similarity and functional network based on microrna-associated diseases," *Bioinformatics*, vol. 26, no. 13, pp. 1644–1650, 2010.
[20] Y. Li, C. Qiu, J. Tu, B. Geng, J. Yang, T. Jiang, and Q. Cui, "Hmdd v2. 0: a database for experimentally supported human microrna and disease associations," *Nucleic acids research*, p. gkt1023, 2013.
[21] A. Kozomara and S. Griffiths-Jones, "mirbase: annotating high confidence micrornas using deep sequencing data," *Nucleic acids research*, vol. 42, no. D1, pp. D68–D73, 2014.
[22] H. McWilliam, W. Li, M. Uludag, S. Squizzato, Y. M. Park, N. Buso, A. P. Cowley, and R. Lopez, "Analysis tool web services from the embl-ebi," *Nucleic acids research*, vol. 41, no. W1, pp. W597–W600, 2013.
[23] L. Cheng, J. Li, P. Ju, J. Peng, and Y. Wang, "Semfunsim: a new method for measuring disease similarity by integrating semantic and gene functional association," *PloS one*, vol. 9, no. 6, p. e99415, 2014.
[24] M. Ammad-Ud-Din, E. Georgii, M. Gonen, T. Laitinen, O. Kallioniemi, K. Wennerberg, A. Poso, and S. Kaski, "Integrative and personalized qsar analysis in cancer by kernelized bayesian matrix factorization," *Journal of chemical information and modeling*, vol. 54, no. 8, pp. 2347–2359, 2014.
[25] L. Cheng, G. Wang, J. Li, T. Zhang, P. Xu, and Y. Wang, "Sidd: a semantically integrated database towards a global view of human disease," *PloS one*, vol. 8, no. 10, p. e75504, 2013.
[26] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant *et al.*, "Disease ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic acids research*, p. gku1011, 2014.
[27] M. Hattori, Y. Okuno, S. Goto, and M. Kanehisa, "Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways," *Journal of the American Chemical Society*, vol. 125, no. 39, pp. 11 853–11 865, 2003.
[28] C. Xu, Y. Ping, X. Li, H. Zhao, L. Wang, H. Fan, Y. Xiao, and X. Li, "Prioritizing candidate disease mirnas by integrating phenotype associations of multiple diseases with matched mirna and mrna expression profiles," *Mol. BioSyst.*, vol. 10, no. 11, pp. 2800–2809, 2014.
[29] W. H. Organization, *Global status report on alcohol and health*. World Health Organization, 2014.

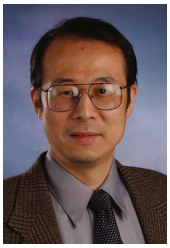


Wei Peng received the PhD degree in computer science from Central South University, China, in 2013. Currently, she is an Associate Professor of Kunming University of Science and Technology, China. She is also a visiting scholar of the Department of Computer Science, Georgia State University, USA. Her research interests include bioinformatics and data mining.

Wei Lan Wei Lan received his BSc and MSc degree in Henan Polytechnical University and Guangxi University, China in 2009 and 2012, respectively. He is currently a Ph.D Candidate in Bioinformatics at Central South University. His currently research interest including bioinformatics and data mining especially in disease gene and noncoding RNA.

Zeng Yu Zeng Yu received MS degrees from the Department of Mathematics, School of Sciences, China University of Mining and Technology 2011. He is currently a PhD candidate in the School of Information Science and Technology, Southwest Jiaotong University, China. He is also a visiting PhD student of the Department of Computer Science, Georgia State University, USA. His current research interests include data mining, bioinformatics, deep learning and cloud computing.

JianXin Wang Jianxin Wang received the PhD degree in computer science from Central South University, China, in 2001. He is the vice dean and a professor in School of Information Science and Engineering, Central South University, Changsha, Hunan, P.R. China. His current research interests include algorithm analysis and optimization, parameterized algorithm, bioinformatics and computer network. He has published more than 150 papers in various International journals and refereed conferences. He is a senior member of the IEEE.



Yi Pan received the PhD degree in computer science from the University of Pittsburgh, Pennsylvania, in 1991. He is a Regents' Professor of Computer Science and an Interim Associate Dean and Chair of Biology at Georgia State University, USA and a Changjiang Chair Professor in the Department of Computer Science at Central South University, P.R. China. His research interests include parallel and distributed computing, networks, and bioinformatics. He has published more than 100 journal papers with 50 papers published in various IEEE/ACM journals. He has served as the editor in chief or an editorial board member for 15 journals, including six IEEE Transactions. He has delivered more than 10 keynote speeches at many international conferences and is a speaker for several distinguished speaker series.