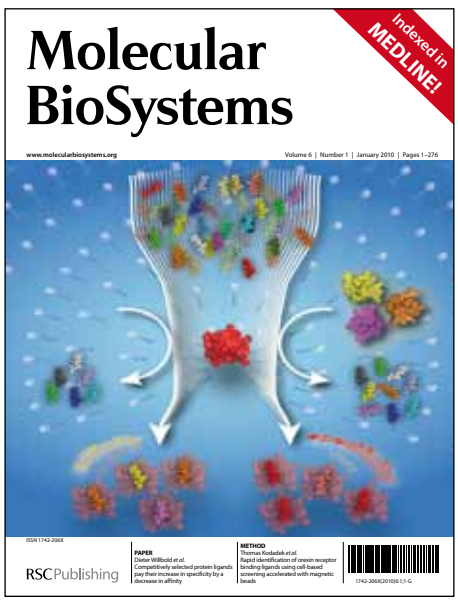
Molecular Biosystems

Accepted Manuscript



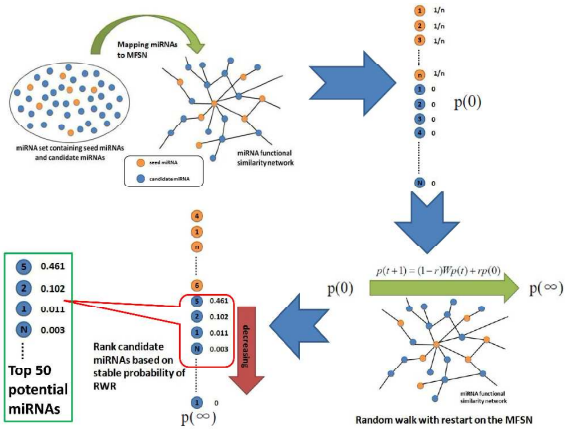
This is an *Accepted Manuscript*, which has been through the RSC Publishing peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, which is prior to technical editing, formatting and proof reading. This free service from RSC Publishing allows authors to make their results available to the community, in citable form, before publication of the edited article. This *Accepted Manuscript* will be replaced by the edited and formatted *Advance Article* as soon as this is available.

To cite this manuscript please use its permanent Digital Object Identifier (DOI®), which is identical for all formats of publication.

More information about *Accepted Manuscripts* can be found in the **Information for Authors**.

Please note that technical editing may introduce minor changes to the text and/or graphics contained in the manuscript submitted by the author(s) which may alter content, and that the standard **Terms & Conditions** and the **ethical guidelines** that apply to the journal are still applicable. In no event shall the RSC be held responsible for any errors or omissions in these *Accepted Manuscript* manuscripts or any consequences arising from the use of any information contained in them.



RWRMDA adopt global network similarity to infer potential miRNA-disease interactions by implementing random walk on the miRNA functional similarity network.

View Online

Molecular BioSystems
Cite this: DOI: 10.1039/c0xx00000x

www.rsc.org/molecularbiosystems

PAPER

RWRMDA: predicting novel human microRNA-disease associations†‡

Xing Chen,[a,b] Ming-Xi Liu,[b,c] and Gui-Ying Yan*[a,b]

Received (in XXX, XXX) Xth XXXXXXXXX 20XX, Accepted Xth XXXXXXXXX 20XX
DOI: 10.1039/b000000x

Recently, more and more evidences have shown that microRNAs (miRNAs) play critical roles in the development and progression of various diseases, but it is not easy to predict potential human miRNA-disease associations from vast amount of biological data. Computational methods for predicting potential disease-miRNA associations have been paid greater attention based on their feasibility, guidance and effectiveness. Different from traditional local network similarity measures, we adopted global network similarity measures and developed Random Walk with Restart for MiRNA-Disease Association (RWRMDA) to infer potential miRNA-disease interactions by implementing random walk on the miRNA-miRNA functional similarity network. We tested RWRMDA on 1616 known miRNA-disease associations based on leave-one-out cross-validation and achieved an area under the ROC curve of 86.17%, which significantly improves previous methods. The method was also applied to three cancers for accuracy evaluation. As a result, 98% (Breast cancer), 74% (Colon cancer), and 88% (Lung cancer) of top 50 predicted miRNAs are confirmed by published experiments. These results suggest that RWRMDA will represent an important bioinformatics resource in biomedical research of both miRNAs and diseases.

Introduction

MicroRNAs (miRNAs) are a class of small (~22nt) non-coding regulatory RNAs normally suppressing the expression of the target mRNA at post transcriptional level by binding to the 3'-UTRs of target mRNA through sequence-specific base pairing [1-4]. However, some reports have pointed out miRNAs may also function as positive regulators in some cases [5, 6]. Caenorhabditis elegans (C. elegans) lin-4 and let-7, the first two known miRNAs, were identified by conventional forward genetic screens [7-9]. Over the past few years, thousands of miRNAs have been discovered in eukaryotic organisms ranging from nematodes to humans [10]. The newest version of miRBase has contained 16772 entries and more than 1000 miRNAs have been discovered in human (miRBase, Release 17) [11].

Plenty of studies reveal miRNA is one of the most important components in the cell and plays critical roles in diverse fundamental biological processes, such as cell development, cell proliferation, cell differentiation, cell apoptosis, signal transduction, viral infection, and so on [12-17]. Therefore, the miRNA related dysfunction is associated with various diseases [3, 18-24]. An exciting example is mir-375 can regulate insulin secretion [25, 26]. Also dysregulation of numerous miRNAs is associated with the initiation and progression of various cancers [27]. Nowadays, miRNAs have taken centre stage in the field of human molecular oncology [28]. Therefore, a large scale search for the relationship between miRNAs and diseases has become an important goal of biomedical research [29], which will accelerate the understanding of the disease pathogenesis at the molecular level, more importantly, benefit the prognosis, diagnosis, evaluation, treatment and prevention of disease and promote the human medical improvement [18, 21, 30-33]. However, current knowledge about the relation between miRNAs and diseases is relatively limited. Experimental identification of disease-related miRNAs by existing techniques is expensive and time-consuming [29]. Fortunately vast amount of biological data about miRNAs has been generated, so there is a strong motivation to develop powerful computational methods which can effectively uncover potential disease-miRNA associations in a large scale. Computational methods can select most promising miRNAs for further analysis and hence decrease the number of the experiments, benefit the understanding of miRNAs function. However, the difficulty of prediction task lies in the rarity of known disease-miRNA interactions.

Some important conclusions and computational methods about disease-related miRNAs prediction have been proposed. Lu et al [3] analyzed the human microRNA-disease association data and proposed many important patterns between miRNAs and human diseases, which laid a solid foundation for current disease-related miRNA research and provided powerful support to the research about the diseases at the miRNA level. Based on the assumption that phenotypical similar diseases tend to be associated with functional related miRNAs proposed by Lu et al [3], Zhang et al [34] developed the first miRNA-disease association prediction method, which identified potential cardiovascular disease related miRNAs by miRNAs set, family analysis and Gene Ontology. However, the fact that this method strongly depend on miRNA sets has limited its application. Jiang et al [21] developed a computational method based on the hypergeometric distribution to infer disease-related miRNAs by integrating miRNAs functional interactions network, disease similarity network and known phenome-microRNAome network consisting of 270 experimentally verified disease-miRNAs associations obtained from miR2Disease [35]. Although miRNA functional network has been constructed in this paper, only the neighbor information of each miRNA was used in the scoring system. Making fully use of global network similarity information would improve the accuracy of the algorithm. Another limitation is that this method highly depends on the predicted miRNA target. It is known that the current in-silico prediction tool for miRNA target prediction has a high false positive and high false negative. As a result, the prediction accuracy of this method is not high. Jiang et al [22] further proposed an approach for prioritizing candidate miRNAs based on genomic data integration by Naïve Bayes model, which strongly relies on datasets of disease-gene associations and miRNA-target interactions. As we all known, the molecular bases for about 60% of human diseases are still unknown [36]. Also, the problem of high false positive and high false negative in the miRNA-target interactions predicted by

current different algorithms still exists in this method. Jiang et al [29] proposed an approach for distinguishing positive disease miRNAs from negative disease miRNAs based on Support Vector Machine by extracting the feactures based on microRNA-target data and phenotype similarity data. Under the assumption that miRNAs implicated in a specific disease will show aberrant regulations of their target mRNAs, Xu et al [28] introduced a network-centric method to prioritize candidate disease miRNAs by constructing four topological features to distinguish prostate cancer (PC) miRNAs and non-PC miRNAs. The common problem of aforementioned two methods is the selection of negative samples, because there are no verified negative miRNA-disease associations. Compiling a list of negative disease miRNAs is currently difficult or even impossible [28]. Besides the above methods which based on similar train of thought, Rossi et al proposed a method called OMiR to identify potential relationship between miRNAs and OMIM diseases, which achieved their aim through calculating the significance of the overlap between miRNAs' loci and OMIM diseases' loci, without utilizing known miRNA-disease associations and other information, such as miRNA target, disease pathogeny information and so on [37].

Taken together, the above mentioned methods for miRNA-disease association prediction have various limitations. Therefore, novel methods are urgently needed. In this paper, we have investigated the hypothesis that global network similarity measures are better suited to capture the associations between diseases and miRNAs than traditional local network similarity measures such as neighbor information. Based on the global network similarity measure and the assumption that functionally related miRNAs tend to be associated with phenotypical similar diseases [21], the method of Random Walk with Restart for MiRNA-Disease Association (RWRMDA) has been developed to infer potential miRNA-disease associations by implementing random walk on the miRNA functional similarity network to prioritize candidate miRNAs for disease of interest. Random walk has been widely applied in bioinformatics, especially for disease gene identification and drug target interactions prediction [38-41]. Cross validation and case studies about three kinds of cancers have illustrated RWRMDA is superior to previous predictive method based on local network similarity measure.

Methods

The human miRNA-disease association data

Considering many studies have produced a large number of miRNA-disease associations, Lu et al [3] collected the miRNA-disease associations and constructed a human miRNA-associated disease database (HMDD), which contains 3760 miRNA-disease associations and the information of 493 miRNAs and 295 diseases from 1688 publications. Jiang et al [35] also constructed a manually curated miRNA-disease relationships database (miR2Disease), which aims at providing a comprehensive resource of experimentally confirmed miRNA-disease associations. After recent update, 3273 miRNA-disease associations about 349 miRNAs and 163 diseases have been collected in the database. Yang et al [42] constructed a publicly available DataBase of Differentially Expressed MiRNAs in human Cancers (dbDEMC) with the aim to provide potential cancer-related miRNAs by insilco computing. The current version of dbDEMC includes 607 differentially expressed miRNAs (590 mature miRNAs and 17 precursor miRNAs) in 14 cancers from 48 microarray experiments in peer-reviewed publications.

The human miRNA-disease association data used for prediction accuracy evaluation was downloaded from the supplementary data of [43]. This dataset consists of 1616 distinct high-quality experimentally verified human miRNA-disease associations, which were obtained from HMDD in September, 2009. The operations of merging the records of different miRNA copies that produce the same mature miRNA into one group, unifying the name of different mature miRNAs as one miRNAgene, and curating the disease name based on standard MeSH disease terms were implemented. These associations were used as the benchmark dataset for the performance evaluation of our model in the cross validation schema and the seed dataset for predicting potential human miRNA-disease associations (See Table S1). We did not use the latest version data of HMDD data, because potential human miRNA-disease associations predicted by our model can be evaluated by new associations introduced to HMDD after September, 2009. MiRNA-Disease Association Network (MDAN) was constructed, where vertices set $M = \{m_1, m_2, \ldots, m_n\}$ denotes the set of n miRNAs and $D = \{d_1, d_2, \ldots, d_k\}$ denotes the set of k diseases. Vertex m_i and d_j are linked by an edge in the MDAN if miRNA i is associated with disease j in our datasets. The weights of all edges are set to 1. Actually, MDAN is a bipartite graph containing two sets of vertices corresponding to miRNAs and diseases, respectively.

The miRNA-miRNA functional similarity network

The miRNA-miRNA functional similarity scores were downloaded from http://cmbi.bjmu.edu.cn/misim/ [43] in January 2010 (See Table S2). In this dataset, functional similarity score for each miRNA pair is calculated based on the observation that genes with similar functions are often associated with similar diseases. miRNA functional similarity matrix is defined as S, where the entity $S(i,j)$ in row i column j is the functional similarity score between miRNA i and j. Based on the miRNA similarity matrix, miRNA functional similarity network (MFSN) is constructed, where vertices set $M = \{m_1, m_2, \ldots, m_n\}$ denotes the set of n miRNAs. Vertex m_i and m_j are linked by an edge in the network if the functional similarity between miRNA i and j is more than zero. The functional similarity score between miRNA i and j is used as the weight of this edge.

Random Walk with Restart for MiRNA-Disease Association (RWRMDA)

In this paper, based on the observation that functionally related miRNAs are often associated with phenotypically similar diseases [3], Random Walk with Restart for MiRNA-Disease Association (RWRMDA) was developed to uncover the potential associations between human miRNAs and diseases. The code and readme file in both MATLAB and R can be downloaded from http://asdcd.amss.ac.cn/Software/Details/3. Random walk method simulates a random walker's transition from its current nodes randomly to neighbors in the network starting at some given seed nodes [39]. Hence RWRMDA consists of three steps as follows: (1) decide the initial probability of each miRNA, (2) implement random walk on the MFSN, and (3) obtain stable probability of random walk and rank candidate miRNAs (See Figure 1).

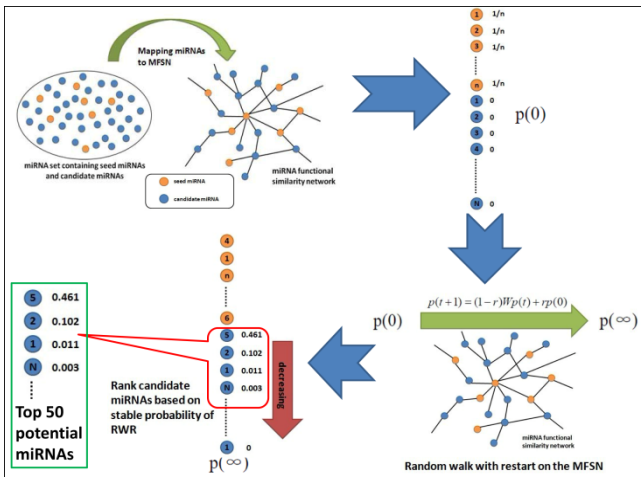


Fig.1 Flowchart of RWRMDA. This flowchart is a brief description of RWRMDA. After mapping all the miRNAs (containing seed miRNAs and candidate miRNAs) to miRNA functional similarity network (MFSN), each node in MFSN is assigned corresponding initial probability *p(0)*. Then random walk with restart would be implemented until the termination condition occurs and we get stable probability *p(∞)*. The last step is to rank all candidate miRNAs based on *p(∞)* to select potential disease-related miRNAs for experimental validation.

If we want to predict potential miRNAs for a given disease *d* of interest, all the miRNAs which have already been confirmed to be associated with this disease will be considered as seed miRNAs. Other non-seed miRNAs will be considered as candidate miRNAs. The initial probability *p(0)* is formed such that equal probabilities are assigned to seed miRNAs, with the sum equal to 1, while the initial probabilities of non-seed miRNAs are zero. Here we allow the restart of random walk in every time step at source nodes with probability *r* ($0<r<1$). *p(t)* is defined as a vector in which the i-th element holds the probability of finding the random walk at node *i* at step *t*. The column-normalized miRNA-miRNA functional similarity matrix was denoted as W. Therefore, the random walk is defined as:

$$p(t+1) = (1-r)Wp(t) + rp(0)$$

After some steps, the random walk is stable (the change between *p(t)* and *p(t+1)* measured by L_1 norm is less than a cutoff, here we took the cutoff as 10^{-6}). Selecting this criterion to stop random walk is frequently used in the research about applying random walk related methods to solve problems such as disease gene prioritization, drug target prediction[38-41]. The stable probability is defined as *p(∞)*. Candidate miRNAs are ranked according to *p(∞)* to select potential miRNAs of the given disease. The high-scored miRNAs can be expected to have a high probability to be associated with the given disease, which will have priority to be tested in the biological experiments. Hence the cost to identify true miRNA-disease associations can be significantly reduced.

Results

Global properties of MDAN and MFSN

There are 1395 miRNA–disease associations between 271 miRNAs and 137 diseases in the MDAN. We found that 64.96% of the diseases were associated with at least two miRNAs, and approximately 70% miRNAs were associated with two or more diseases. MDAN is shown in Figure S1, indicating most edges in the network are connected and form a large connecting subnetwork. This figure fully demonstrates the complexity of multifactorial or polygenic diseases which are likely to be associated with the effects of multiple miRNA or genes in combination with lifestyle and environmental factors. The degree distribution of miRNAs and diseases were evaluated and power-law distributions were observed (Figure S2 and S3). Therefore, MDAN displays scale-free characteristics like many other large-scale networks. There are 688 miRNA-miRNA functional interactions (the functional similarity between these two miRNAs is more than 0.7) between 271 miRNAs in the MFSN. It can be observed that approximately half of the miRNAs interact with at least two miRNAs (Figure S4).

Performance evaluation

In order to evaluate the performance of RWRMDA to infer potential miRNA-disease associations, leave-one-out cross validation was implemented on 1394 known experimentally verified miRNA-disease associations. For given disease *d*, each known disease-related miRNA was left out in turn as test miRNA and other known disease-related miRNA were taken as seed miRNAs. The candidate miRNA set consisted of all the miRNAs which have no evidence to show their association with disease *d*. When each known disease-related miRNA was left out as test miRNA, how well this miRNA ranked relative to the candidate set of this disease was assessed. If the rank of test miRNA exceeds the given threshold, the model was considered to successfully predict this miRNA-disease association. For simplicity, we just choose *r*=0.2, which can be better selected by further cross-validation. Actually, it has been demonstrated the predictive result is robust to the restart probability in the research about disease-related genes identification [38, 39]. Since this is the first time to introduce random walk algorithm into the research about disease-miRNA association prediction, we will discuss the effect of restart probability in the next section.

Receiver-operating characteristics (ROC) curve plots true positive rate (sensitivity) versus false positive rate (1-specificity) at different thresholds. Sensitivity refers to the percentage of the test microRNAs whose ranking is higher than a given threshold, namely the ratio of the successfully predicted miRNA-disease associations to the total number of known miRNA-disease associations. Specificity refers to the percentage of miRNAs that are below the threshold. The area under ROC curve (AUC) was calculated. AUC=1 indicates perfect performance and AUC=0.5 indicates random performance. Here we compared RWRMDA with the computational method based on the hypergeometric distribution [21] using the same miRNA similarity network. As mentioned before, RWRMDA is based on the global network similarity measure, while the method based on the hypergeometric distribution [21] is based on the local network similarity measure, only using the neighbor information. The AUC of RWRMDA was 0.8617, while hypergeometric distribution method only had an AUC of 0.7783

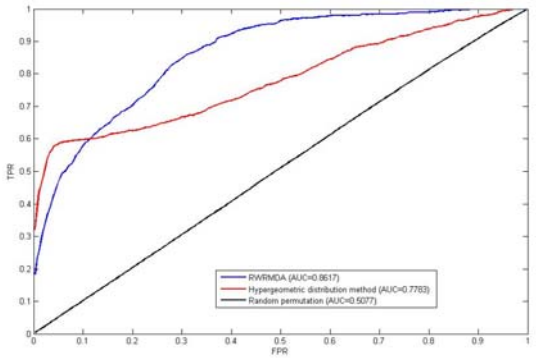


Fig.2 Method comparison. Comparison between RWRMDA and hypergeometric distribution method proposed by Jiang et al [21] in terms of ROC curve and AUC based on leave-one-out cross validation on 1394 known experimentally verified miRNA-disease associations is shown to confirm the performance advantage of RWRMDA.

(Figure 2). In fact, functionally related miRNA network used in [21] was constructed based on the assumption that two miRNAs are functionally related if the overlap between their target genes is statistically significant. In that network, implementing leave-one-out cross validation for hypergeometric distribution method only obtained the AUC of 0.7580, even less than 0.7783 here. Considering network construction is also part of biological methods, the comparison here actually overestimates hypergeometric distribution method's predictive ability. Even so, the comparison here still shows the superior performance of RWRMDA to previous methods. Excellent performance indicates RWRMDA can recover known experimentally verified miRNA-disease associations and hence has the potential to uncover potential miRNA-disease associations. The advantage of global network similarity to local similarity has also been demonstrated. To evaluate whether the results of cross validation by RWRMDA were likely to be obtained by chance, seed miRNAs are randomly chosen from candidate miRNAs 100 times. For each time, RWRMDA was implemented under the schema of leave-one-out cross validation, and AUC were recalculated with overall ROC curve and AUC shown in Figure 2. It was apparent AUC under random circumstances were much lower than the original values, demonstrating that observed good performance of RWRMDA can't be achieved by chance, and hence prediction results by RWRMDA would be of biological significance.

Effects of restart probability

To investigate the selection of restart probability r for the performance of RWRMDA, we set various value of r ranging from 0.1 to 0.9 and calculated AUC in the framework of leave-one-out cross validation. Table 1 shows the effects of restart probability on the cross validation result in human miRNA-disease association dataset. It could be observed that the performance of RWRMDA is superior to previous methods based on any selection of restart probability.

Case studies

Recent studies show that about half of miRNAs are located in cancer associated genomic regions or fragile sites [10, 44] and many miRNAs are related to the development of various human malignancies [10, 19, 24, 44-46]. For example, mir-143 and mir-145 are downregulated in colon cancer [10, 47]; mir-99 is overexpressed in pancreatic cancer [10]. To evaluate the performance of RWRMDA on independent dataset, case studies about three cancers (Breast cancer, colonic cancer, and lung cancer) were implemented. Prediction results were validated by various databases and literature.

Breast cancer (malignant breast neoplasm) is one of the most commonly occurring female cancers and comprises 22% of all cancers in women [21]. In the golden standard data, 78 miRNAs have shown their associations with breast cancer. Candidate miRNAs were prioritized based on RWRMDA. Among the top 50 predicted breast cancer-related miRNAs, 49 miRNAs have been confirmed to be associated with breast cancer by HMDD [3], dbDEMC [42], mir2disease [35], and literature [21, 48]. The reason for using HMDD to confirm predictive result is that HMDD has been updated many times since September 2009 when we downloaded the seed miRNA-disease associations. The top 50 miRNAs and evidences for their associations with breast cancer were listed (see Table 2). As a result, 98% of the top 50 predictions were confirmed to be true.

Colonic cancer is a cancer characterized by neoplasia in the colon, rectum, or vermiform appendix. It is the third most commonly cancer in the world, but more than half of the people who die of colonic cancer are from developed countries (http://en.wikipedia.org/wiki/Colonic_cancer). Thirty-seven miRNAs are considered as seed miRNAs in the golden standard data. Among the top 50 colonic cancer-related miRNAs predicted by RWRMDA, 37 miRNAs have been confirmed to be related to colonic cancer by HMDD [3], dbDEMC [42], mir2disease [35], and literature [10, 45, 48, 49]. The top 50 miRNAs and evidence were listed in Table S3. As a result, 74% of the top 50 predictions were confirmed to be true.

Loss or amplification of a number of miRNAs has been found related to lung cancer, in which the cells of lung tissues grow uncontrollably and form tumors [10]. Seventy-two miRNAs are considered to be related to lung cancer in the benchmark dataset. Forty-four of top 50 predicted miRNAs were confirmed to be related to lung cancer by HMDD [3], dbDEMC [42], mir2disease [35], and literature [50]. The top 50 miRNAs and evidence were listed in Table S4. As a result, 88% of the top 50 predictions were confirmed to be true.

According to the miRNA-miRNA functional similarity scores, the functional relations among known disease-related miRNAs and top 50 potential miRNAs are shown for three kinds of cancers

Table 1 The effect of restart probability value on the cross validation result of RWRMDA.

Restart probability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
AUC	0.8112	0.8617	0.8930	0.9140	0.9299	0.9412	0.9496	0.9557	0.9608

Table 2 The top 50 potential miRNAs of breast cancer predicted by RWRMDA and the confirmation for their association by various databases and literature. Forty-nine of top 50 miRNAs have been confirmed to be related with breast cancer. The only unconfirmed has-mir-142 were ranked 46th in the prediction list of potential breast cancer related miRNAs.

Name	Evidence	Name	Evidence
let-7e	dbDEMC	mir-24	dbDEMC

let-7b	dbDEMC	mir-32	dbDEMC
let-7c	dbDEMC	mir-195	HMDD, miR2Disease, dbDEMC
let-7i	miR2Disease, dbDEMC, literature[21]	mir-181a	miR2Disease, dbDEMC
mir-126	miR2Disease, dbDEMC	mir-106a	dbDEMC
mir-520b	dbDEMC	mir-23b	dbDEMC
mir-30e	literature [48]	mir-148a	miR2Disease, dbDEMC
mir-27a	HMDD, miR2Disease, dbDEMC	mir-135a	dbDEMC
let-7g	dbDEMC	mir-22	miR2Disease, dbDEMC
mir-191	miR2Disease, dbDEMC	mir-99b	dbDEMC
mir-223	HMDD, dbDEMC	mir-182	HMDD, miR2Disease, dbDEMC
mir-130a	dbDEMC	mir-150	dbDEMC
mir-192	dbDEMC	mir-203	miR2Disease, dbDEMC
mir-18b	HMDD, dbDEMC	mir-29c	literature [21], miR2Disease, dbDEMC
mir-101	miR2Disease, dbDEMC	mir-107	HMDD, dbDEMC
mir-130b	dbDEMC	mir-100	dbDEMC
mir-92b	dbDEMC	mir-186	dbDEMC
mir-98	literature [21], miR2Disease, dbDEMC	mir-95	dbDEMC
mir-373	literature [21], miR2Disease, dbDEMC	mir-28	dbDEMC
mir-30a	miR2Disease, HMDD	mir-299	dbDEMC
mir-372	dbDEMC	**mir-142**	**unconfirmed**
mir-16	literature [21], dbDEMC	mir-128b	miR2Disease
mir-199b	miR2Disease, dbDEMC	mir-497	literature [21], HMDD, miR2Disease, dbDEMC
mir-26a	dbDEMC	mir-335	literature [21], miR2Disease, dbDEMC
mir-92a	HMDD	mir-452	dbDEMC

(Figure 3: Breast cancer; Figure S5: Colon cancer; Figure S6: Lung cancer). The edges indicating the functional similarity between two miRNAs is equal to or greater than 0.7 were retained. We also implemented functional enrichment analysis for these known cancer-related and predicted top 50 potential cancer-related miRNAs based on TAM [51]. TAM is a convenient online tool developed by Lu et al for annotations of human miRNAs, which evaluates the statistical significance of each miRNA category among lists of miRNAs using the hypergeometric test. Moreover, TAM can also be used to predict novel related miRNAs based on a list of given miRNAs.

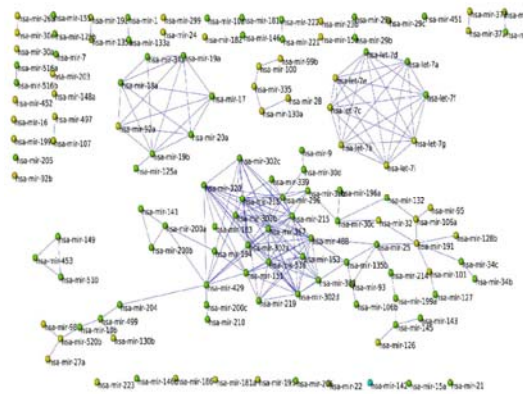


Fig.3 Functional relations between breast cancer-related miRNAs. The functional relations between known disease-related miRNAs and top 50 potential miRNAs are shown for breast cancer. Yellow nodes indicate confirmed top 50 disease-related miRNAs, blue nodes indicate unconfirmed top 50 miRNAs, and green nodes indicate known seed disease-related miRNAs. Disease-related miRNAs tends to form some functional modules and this observation coincides with the basic assumption of RWRMDA.

At present, TAM is freely available at http://cmbi.bjmu.edu.cn/tam [51]. Undoubtedly, as an effective tool to process high throughput data, it provides a new way for researchers to study the common rules or patterns behind a list of miRNAs. Functional enrichment analysis results for three kinds of important cancers are shown (Figure 4: Breast cancer; Figure S7: Colon cancer; Figure S8: Lung cancer;).

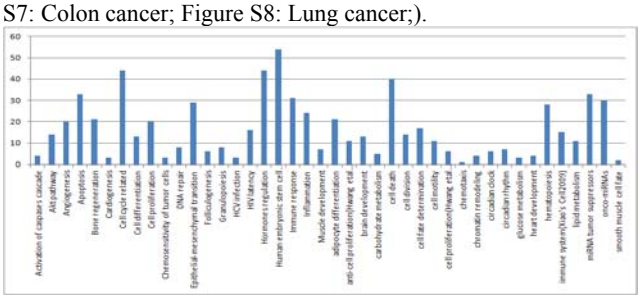


Fig.4 Functional enrichment analysis of breast cancer-related miRNAs. Functional enrichment analysis of breast cancer related miRNAs. The horizontal ordinates stand for 42 significant enriched functions of 128 miRNAs related to Breast cancer, which contain both seed miRNAs and potential ones predicted by our algorithm. The vertical coordinates stand for the number of miRNAs which are significant enriched in the corresponding functions among those 128 breast cancer-related miRNAs.

Because of space limitations, here we focused on the results for Breast cancer. The aim of functional enrichment analysis here is to investigate the function of known breast cancer-related miRNAs and top 50 potential miRNAs and confirm the reasonability of RWRMDA by the fact that the function of potential cancer-related miRNAs is associated with cancer development. We don't intend to identify cancer-related functions. Among 128 human miRNAs that related to Breast cancer (including known miRNAs and top 50 potential miRNAs), 54 miRNAs have a specific function called Human embryonic stem cell regulation, 44 miRNAs are related to the Cell cycle related function, 44 miRNAs are concerned with Hormones regulation, and in addition, 40, 33, 33 miRNAs are related to cell death, Apoptosis, and miRNA tumor suppressors, respectively. In the recent years, with the development of human cancer study, all the above functions have been confirmed separately to play important part in the whole process of cancer development. For instance, the most important function of Human embryonic stem cell is the differentiation function, and some researches indicate that Human

embryonic stem cells can be the cell-of-origin for a range of solid tumors and alterations in normal stem cells contribute to many types of cancer, including lung cancer, prostate cancer, skin cancer, stomach cancer, breast cancer and so on [52, 53]. It's worth noting that mir-142, the only one miRNA unconfirmed in our case study, also has the Human embryonic stem cell regulation function. Although we have not found evidence for this miRNA, it has a very high probability to be related to Breast cancer. As for the Cell cycle related function, many researchers have reached an agreement that cancer is a disease of the cell cycle. If the normal cell cycle changes, some of the body's cells would divide uncontrollably and may cause tumors [54-56]. Also there is a significant relationship between the risk for breast cancer and the abnormal level of human hormones[57]. Obviously, the other three enriched functions, such as cell death, Apoptosis and miRNA tumor suppressors, are also of great concern during the developing process of Breast Cancer. All in all, these cancer associated miRNAs would affect some crucial biological processes, such as cell division, cell differentiation and so on, which are all major causes for cancer. Through the above analysis, we could have a more clear cognition about the biological function of known breast cancer-related miRNAs and top 50 potential miRNAs.

Predicting novel human miRNAs-disease associations

After confirming the accuracy of RWRMDA by cross validation and case studies about three important cancers, we further predicted novel miRNAs associated with various diseases. Here all the known disease-related miRNAs in the golden standard data were used as seed miRNAs. For all the 137 diseases, top 50 potential miRNAs are publicly released to facilitate the discovery of human miRNA-disease associations (see Table S5). Most of top 50 cancer-related miRNAs have been confirmed as mentioned above, hence we have good reason to believe potential miRNAs associated with other diseases predicted by RWRMDA would also be confirmed by further experiments and attention paid to this disease.

Discussion

The success of RWRMDA can be largely attributed to several factors. First, miRNA functional similarly network was used to capture the functional relationship between each miRNA pair and biological characteristic of miRNAs tending to exert the same or similar functions. Second, known experimentally verified miRNA-disease associations were used as the benchmark dataset in the cross validation schema and the seed dataset for predicting potential human miRNA-disease associations. More importantly, global network information was used to capture the association between miRNAs and diseases, whose advantages over local network information methods have been confirmed in many previous studies of capturing potential disease-related genes. RWRMDA makes full use of global network information by the tool of random walk. RWRMDA represents a novel, important and useful resource for miRNA-disease association prediction.

Of course, limitations exist in the current version of RWRMDA. Our approach can be improved in the following directions. Firstly, the performance of RWRMDA can be further improved by more available experimentally verified human miRNA-disease associations. Secondly, more reliable construction of miRNA functional interaction network would improve RWRMDA. More information sources can be integrated to measure the functional similarity between two miRNAs, such as the information of their targets. Finally, RWRMDA does not work for disease which does not have any known associated miRNA. For this point, phenotype similarity information may be useful to improve RWRMDA.

Conclusions

Identifying novel disease-related miRNAs is an important goal of biomedical research. In this work, RWRMDA was developed to predict potential human miRNA-disease associations by integrating known human miRNA-disease associations and human miRNA-miRNA functional similarity information on a large scale. RWRMDA was motivated based on the observation that functionally related miRNAs tend to be associated with phonotypical similar disease and investigation that global network similarity measures are better suited to capture the association between diseases and miRNAs than traditional local network similarity measures. The results indicate that RWRMDA has a high performance of prediction, which suggests its ability to recover the known experimentally verified miRNA-disease associations. Potential miRNA-disease associations predicted for 137 diseases were also provided to guide future biological experiments. Moreover, case studies about three kinds of cancers (breast cancer, colonic cancer, and lung cancer) were also implemented and plenty of prediction results were confirmed by various databases and literature. These results fully demonstrated the superior performance of RWRMDA to previous methods.

Acknowledgements

This work was supported by the National Natural Science of Foundation of China under Grant Nos. 10531070, 10721101, KJCX-YW-S7 and NCMIS.

Notes and references

[a] *National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100190, China. E-mail: xingchen@amss.ac.cn, yangy@amss.ac.cn.*
[b] *Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. E-mail: xingchen@amss.ac.cn, liumingxi@amss.ac.cn, yangy@amss.ac.cn.*
[c] *Graduate University of Chinese Academy of Sciences, Beijing100190, China. E-mail: liumingxi@amss.ac.cn.*

† Electronic Supplementary Information (ESI) available. See DOI: 10.1039/b000000x/

‡ **Authors' contributions.** XC conceived and developed the prediction method, conceived, designed and implemented the experiments, analyzed the result, and wrote the paper. MXL analyzed the result. GYY conceived the prediction method, analyzed the result, wrote the paper, and provided guidance and supervision. All authors read and approved the final manuscript.

1. V. Ambros, *Cell*, 2001, **107**, 823-826.
2. D. P. Bartel, *Cell*, 2004, **116**, 281-297.
3. M. Lu, Q. P. Zhang, M. Deng, J. Miao, Y. H. Guo, W. Gao and Q. H. Cui, *PLoS One*, 2008, **3**, e3420.
4. G. Meister and T. Tuschl, *Nature*, 2004, **431**, 343-349.
5. C. L. Jopling, M. K. Yi, A. M. Lancaster, S. M. Lemon and P. Sarnow, *Science*, 2005, **309**, 1577-1581.
6. S. Vasudevan, Y. C. Tong and J. A. Steitz, *Science*, 2007, **318**, 1931-1934.
7. R. C. Lee, R. L. Feinbaum and V. Ambros, *Cell*, 1993, **75**, 843-854.
8. A. E. Pasquinelli and G. Ruvkun, *Annu Rev Cell Dev Bi*, 2002, **18**, 495-513.

9. B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz and G. Ruvkun, *Nature*, 2000, **403**, 901-906.
10. S. Bandyopadhyay, R. Mitra, U. Maulik and M. Q. Zhang, *Silence*, 2010, **1**, 6.
11. S. Griffiths-Jones, H. K. Saini, S. van Dongen and A. J. Enright, *Nucleic Acids Res*, 2008, **36**, D154-158.
12. A. M. Cheng, M. W. Byrom, J. Shelton and L. P. Ford, *Nucleic Acids Res*, 2005, **33**, 1290-1297.
13. Q. Cui, Z. Yu, E. O. Purisima and E. Wang, *Molecular systems biology*, 2006, **2**, 46.
14. X. Karp and V. Ambros, *Science*, 2005, **310**, 1288-1289.
15. E. A. Miska, *Current Opinion in Genetics Development*, 2005, **15**, 563-568.
16. P. Xu, M. Guo and B. A. Hay, *Trends in genetics : TIG*, 2004, **20**, 617-624.
17. Z. Yu, Z. Li, N. Jolicoeur, L. Zhang, Y. Fortin, E. Wang, M. Wu and S. H. Shen, *Nucleic Acids Res*, 2007, **35**, 4535-4541.
18. G. A. Calin and C. M. Croce, *Nature reviews. Cancer*, 2006, **6**, 857-866.
19. A. Esquela-Kerscher and F. J. Slack, *Nature reviews. Cancer*, 2006, **6**, 259-269.
20. Q. Huang, K. Gumireddy, M. Schrier, C. le Sage, R. Nagel, S. Nair, D. A. Egan, A. Li, G. Huang, A. J. Klein-Szanto, P. A. Gimotty, D. Katsaros, G. Coukos, L. Zhang, E. Pure and R. Agami, *Nature cell biology*, 2008, **10**, 202-210.
21. Q. H. Jiang, Y. Y. Hao, G. H. Wang, L. R. Juan, T. J. Zhang, M. X. Teng, Y. L. Liu and Y. D. Wang, *Bmc Syst Biol*, 2010, **4**, Suppl 1:S2.
22. Q. H. Jiang, G. H. Wang and Y. D. Wang, in *Biomedical Engineering and Informatics (BMEI),*, 2010, vol. 6, pp. 2270-2274.
23. M. V. Latronico, D. Catalucci and G. Condorelli, *Circulation research*, 2007, **101**, 1225-1236.
24. J. Lu, G. Getz, E. A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B. L. Ebert, R. H. Mak, A. A. Ferrando, J. R. Downing, T. Jacks, H. R. Horvitz and T. R. Golub, *Nature*, 2005, **435**, 834-838.
25. M. N. Poy, L. Eliasson, J. Krutzfeldt, S. Kuwajima, X. Ma, P. E. Macdonald, S. Pfeffer, T. Tuschl, N. Rajewsky, P. Rorsman and M. Stoffel, *Nature*, 2004, **432**, 226-230.
26. H. H. van Es and G. J. Arts, *Drug discovery today*, 2005, **10**, 1385-1391.
27. F. X. Xin, M. Li, C. Balch, M. Thomson, M. Y. Fan, Y. Liu, S. M. Hammond, S. Kim and K. P. Nephew, *Bioinformatics*, 2009, **25**, 430-434.
28. J. Xu, C. X. Li, J. Y. Lv, Y. S. Li, Y. Xiao, T. T. Shao, X. Huo, X. Li, Y. Zou, Q. L. Han, L. H. Wang and H. Ren, *Molecular cancer therapeutics*, 2011, **10**, 1857-1866.
29. Q. H. Jiang, G. H. Wang, T. J. Zhang and Y. D. Wang, in *Bioinformatics and Biomedicine (BIBM)*, 2010, pp. 467-472.
30. R. F. Duisters, A. J. Tijsen, B. Schroen, J. J. Leenders, V. Lentink, I. van der Made, V. Herias, R. E. van Leeuwen, M. W. Schellings, P. Barenbrug, J. G. Maessen, S. Heymans, Y. M. Pinto and E. E. Creemers, *Circulation research*, 2009, **104**, 170-178.
31. A. Markou, E. G. Tsaroucha, L. Kaklamanis, M. Fotinou, V. Georgoulias and E. S. Lianidou, *Clin Chem*, 2008, **54**, 1696-1704.
32. T. E. Miller, K. Ghoshal, B. Ramaswamy, S. Roy, J. Datta, C. L. Shapiro, S. Jacob and S. Majumder, *The Journal of biological chemistry*, 2008, **283**, 29897-29903.
33. M. S. Weinberg and M. J. Wood, *Human molecular genetics*, 2009, **18**, R27-R39.
34. F. Zhang, M. Lu, Q. Zhang, F. Zhang, W. Gao and Q. Cui, *Beijing da xue xue bao*, 2009, **41**, 112.
35. Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang and Y. Liu, *Nucleic Acids Res*, 2009, **37**, D98-104.
36. A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, *Nucleic Acids Res*, 2005, **33**, D514-517.
37. S. Rossi, A. Tsirigos, A. Amoroso, N. Mascellani, I. Rigoutsos, G. A. Calin and S. Volinia, *Genomics*, 2011, **97**, 71-76.
38. X. Chen, G. Y. Yan and X. P. Liao, *Omics : a journal of integrative biology*, 2010, **14**, 337-356.
39. S. Kohler, S. Bauer, D. Horn and P. N. Robinson, *American journal of human genetics*, 2008, **82**, 949-958.
40. Y. Li and J. C. Patra, *Bioinformatics*, 2010, **26**, 1219-1224.
41. X. Chen, M. X. Liu and G. Y. Yan, *Mol Biosyst*, 2012, **8**, 1970-1978.
42. Z. Yang, F. Ren, C. Liu, S. He, G. Sun, Q. Gao, L. Yao, Y. Zhang, R. Miao, Y. Cao, Y. Zhao, Y. Zhong and H. Zhao, *BMC genomics*, 2010, **11 Suppl 4**, S5.
43. D. Wang, J. Wang, M. Lu, F. Song and Q. Cui, *Bioinformatics*, 2010, **26**, 1644-1650.
44. G. A. Calin, C. Sevignani, C. Dan Dumitru, T. Hyslop, E. Noch, S. Yendamuri, M. Shimizu, S. Rattan, F. Bullrich, M. Negrini and C. M. Croce, *Proceedings of the National Academy of Sciences of the United States of America*, 2004, **101**, 2999-3004.
45. S. Volinia, G. A. Calin, C. G. Liu, S. Ambs, A. Cimmino, F. Petrocca, R. Visone, M. Iorio, C. Roldo, M. Ferracin, R. L. Prueitt, N. Yanaihara, G. Lanza, A. Scarpa, A. Vecchione, M. Negrini, C. C. Harris and C. M. Croce, *Proceedings of the National Academy of Sciences of the United States of America*, 2006, **103**, 2257-2261.
46. G. A. Calin, M. Ferracin, A. Cimmino, G. Di Leva, M. Shimizu, S. E. Wojcik, M. V. Iorio, R. Visone, N. I. Sever, M. Fabbri, R. Iuliano, T. Palumbo, F. Pichiorri, C. Roldo, R. Garzon, C. Sevignani, L. Rassenti, H. Alder, S. Volinia, C. G. Liu, T. J. Kipps, M. Negrini and C. M. Croce, *The New England journal of medicine*, 2005, **353**, 1793-1801.
47. M. Z. Michael, O. C. SM, N. G. van Holst Pellekaan, G. P. Young and R. J. James, *Molecular cancer research : MCR*, 2003, **1**, 882-891.
48. R. Baffa, M. Fassan, S. Volinia, B. O'Hara, C. G. Liu, J. P. Palazzo, M. Gardiman, M. Rugge, L. G. Gomella, C. M. Croce and A. Rosenberg, *The Journal of pathology*, 2009, **219**, 214-221.
49. Y. X. Wang, X. Y. Zhang, B. F. Zhang, C. Q. Yang, X. M. Chen and H. J. Gao, *J Digest Dis*, 2010, **11**, 50-54.

50. W. Roa, B. Brunet, L. H. Guo, J. Amanie, A. Fairchild, Z. Gabos, T. Nijjar, R. Scrimger, D. Yee and J. Xing, *Clin Invest Med*, 2010, **33**, E124-E132.
51. M. Lu, B. Shi, J. Wang, Q. Cao and Q. Cui, *BMC Bioinformatics*, 2010, **11**, 419.
52. F. Martin-Belmonte and M. Perez-Moreno, *Nature reviews. Cancer*, 2012, **12**, 23-38.
53. J. E. Visvader, *Nature*, 2011, **469**, 314-322.
54. M. Park and S. Lee, *Journal of biochemistry and molecular biology*, 2003, **36**, 60.
55. B. Clurman and J. Roberts, *Journal of the National Cancer Institute*, 1995, **87**, 1499.
56. K. Collins, T. Jacks and N. P. Pavletich, *Proceedings of the National Academy of Sciences of the United States of America*, 1997, **94**, 2776.
57. T. Key, P. Appleby, I. Barnes, G. Reeves and H. Endogenous, *Journal of the National Cancer Institute*, 2002, **94**, 606.

