BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 13 2010, pages 1644–1650
doi:10.1093/bioinformatics/btq241

Data and text mining

Advance Access publication 3 May 2010

Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases

Dong Wang[1,2], Juan Wang[1,2], Ming Lu[1,2], Fei Song[1,3] and Qinghua Cui[1,2,*]

[1]Department of Biomedical Informatics, Peking University Health Science Center, 38 Xueyuan Road, Beijing 100191, [2]MOE Key Laboratory of Molecular Cardiology, Peking University, 38 Xueyuan Road, Beijing 100191 and [3]Science Department, Northwest A & F University, Shaanxi Key Lab Mol Biol Agr, Yangling 712100, Shaanxi, China

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: It is popular to explore meaningful molecular targets and infer new functions of genes through gene functional similarity measuring and gene functional network construction. However, little work is available in this field for microRNA (miRNA) genes due to limited miRNA functional annotations. With the rapid accumulation of miRNAs, it is increasingly needed to uncover their functional relationships in a systems level.

Results: It is known that genes with similar functions are often associated with similar diseases, and the relationship of different diseases can be represented by a structure of directed acyclic graph (DAG). This is also true for miRNA genes. Therefore, it is feasible to infer miRNA functional similarity by measuring the similarity of their associated disease DAG. Based on the above observations and the rapidly accumulated human miRNA-disease association data, we presented a method to infer the pairwise functional similarity and functional network for human miRNAs based on the structures of their disease relationships. Comparisons showed that the calculated miRNA functional similarity is well associated with prior knowledge of miRNA functional relationship. More importantly, this method can also be used to predict novel miRNA biomarkers and to infer novel potential functions or associated diseases for miRNAs. In addition, this method can be easily extended to other species when sufficient miRNA-associated disease data are available for specific species.

Availability: The online tool is available at http://cmbi.bjmu.edu.cn/misim

Contact: cuiqinghua@hsc.pku.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 11, 2010; revised on April 22, 2010; accepted on April 28, 2010

1 INTRODUCTION

microRNAs (miRNAs) are endogenous small non-coding RNA molecules that can regulate gene expression at the post-transcriptional level by binding with 3′ untranslated regions (UTRs) of the target mRNAs through base pairing. This results in the cleavage or translation inhibition of target mRNAs (Berezikov *et al.*, 2006). miRNAs are considered to represent one of the most important components of the cell. They play critical roles in many important biological processes, and are therefore associated with various diseases (Esquela-Kerscher and Slack, 2006; Latronico *et al.*, 2007; Lu, 2008). Thus far, thousands of miRNAs have been identified. For example, among humans, more than 700 miRNAs have been reported in miRBase (Griffiths-Jones, 2004). In order to better understand miRNAs, it is increasingly necessary to measure their functional similarity and to further construct a network based on such. For protein-coding genes, measuring gene functional similarity and the construction and analysis of gene functional networks have obtained important results (Du *et al.*, 2009; Horvath *et al.*, 2006; Lee *et al.*, 2004; Lin *et al.*, 2007; Lord *et al.*, 2003; Pesquita *et al.*, 2008, 2009; Wang *et al.*, 2007). For example, Horvath *et al.* developed a gene functional network construction method based on gene expression similarity and identified an important molecular target (ASPM) of glioblastoma after applying their method to glioblastoma gene expression data (Horvath *et al.*, 2006). One class of widely used methods related to gene functional similarity and the construction of a gene functional network is by measuring their sequence or expression similarities (Horvath *et al.*, 2006; Lin *et al.*, 2007). Another class of methods to infer functional similarity of protein-coding genes is based on gene ontology (GO) data (Du *et al.*, 2009; Lee *et al.*, 2004; Lord *et al.*, 2003; Pesquita *et al.*, 2008, 2009; Wang *et al.*, 2007). For miRNAs, although sequence or expression similarities can interpret part of the functional similarity, like protein-coding genes, the correlation between gene functional similarities and gene sequence or gene expression similarities does not always exist (Wang *et al.*, 2007). The functional similarity of two miRNAs may be indirectly inferred based on their targets. However, it is difficult for this method to achieve high reliability because miRNA targets are mostly obtained by *in silico* prediction, which shows high false positives and false negatives (Bartel, 2009). Furthermore, although methods of measuring protein-coding gene functional similarity based on GO can achieve better results (Du *et al.*, 2009), these methods are not applicable for miRNA genes because the function of most miRNAs remains unknown and no such function annotation database is available. Therefore, a new method is required to measure miRNA functional similarity and to further construct a miRNA functional network for this purpose.

It has been reported that genes with similar functions are often implicated in similar diseases and vice versa (Goh *et al.*, 2007). This observation also exists in miRNAs (Lu, 2008). Moreover, according to prior knowledge, the relationships of different diseases can be represented in a structure of directed acyclic graph (DAG). Therefore, the functional similarity of miRNAs can be evaluated by

*To whom correspondence should be addressed.

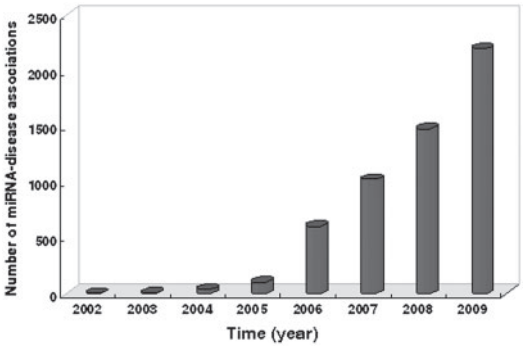

Fig. 1. The increasing pattern of miRNA–disease associations recorded in HMDD with the time.

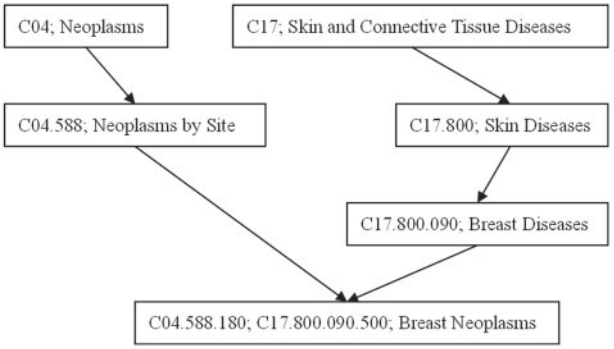

Fig. 2. The disease DAG of breast neoplasms. The addresses of its ancestors are shown in a DAG structure. The semantic value of disease 'Breast Neoplasms' is calculated by summing up the weighted contribution of other diseases to 'Breast Neoplasms' and the contribution to 'Breast Neoplasms' by 'Breast Neoplasms' itself.

quantitatively measuring the similarity of disease DAG associated with these miRNAs. This observation provides a chance to infer miRNA functional similarity through their associated disease DAGs. However, this method is only applicable if there are sufficient miRNA–disease association data. Fortunately, in recent years, more miRNA–disease associations have been reported (Fig. 1). For example, a human miRNA disease database (HMDD) has recently recorded 2205 human miRNA–disease associations (Lu, 2008).

Here, based on the miRNA–disease association data and disease DAG, we presented a method, MISIM (miRNA similarity), to measure the functional similarity of miRNAs and to further construct miRNA functional networks according to the calculated functional similarity. We validated our method by comparing it with other potential miRNA functional similarity inferring methods, such as miRNA family, miRNA cluster and miRNA expression similarity. Results show that our method is reliable. More importantly, our method allows the discovery of novel miRNA pairs with high functional similarity, and can predict novel function and associated disease by analyzing the miRNA functional network.

2 METHODS

2.1 MeSH disease DAG structure

We downloaded MeSH descriptors from the National Library of Medicine (http://www.nlm.nih.gov/). MeSH descriptors were organized into 16 categories: Category A for anatomic terms, Category B for organisms, Category C for diseases, Category D for drugs and chemicals and so on. We then obtained the relationship of various diseases based on the disease DAG from the MeSH descriptor of Category C (Supplementary Material 1).

Each MeSH descriptor showed a structure of a hierarchical DAG. All nodes in the DAG are connected by a direct edge from a more general term, we call it parent, to a more specific term, and we call it child. For example for the DAG of breast neoplasms (Fig. 2), 'Skin diseases' points to 'Breast Diseases'. The purposes of constructing all nodes in this DAG format are to let computers interpret this DAG in a quantitative way and to let it be readable by human.

Each node consists of a descriptor, which carries a unique ID that will not change, and tree numbers which consist of a list of its parent tree numbers separated by ';' from all general nodes. The benefit is that from any one node, we can easily parsing a single tree number without querying the whole DAG related to it when we need access all its ancestors. Typically, the whole DAG is saved into a database. Thus, this organization makes the computation much faster when we have a large amount of data to compute.

2.2 Human miRNA–disease association data

We downloaded human miRNA–disease association data from a human miRNA-associated disease database, HMDD, which recorded 1616 distinct human miRNA–disease associations (September 2009). For simplicity, all the records of different miRNA copies that produce the same mature miRNA (such as hsa-mir-376a-1 and hsa-mir-376a-2) were merged into one group. We further unified the name of different mature miRNAs as one miRNA gene. Finally, we curated the disease name using the standard MeSH disease terms. As a result, 1395 miRNA–disease associations, including 271 miRNAs and 137 diseases, were utilized in our study (Supplementary Material 2).

2.3 Method for measuring miRNA functional similarity

A critical step in our method was the measurement of miRNA functional similarity, the basis of miRNA functional network construction. In this study, we presented a method, MISIM, to measuring the functional similarity of miRNAs. MISIM adapted the method for measuring the functional similarity of protein-coding genes based on GO terms (Du *et al.*, 2009; Wang *et al.*, 2007). MISIM contains four main procedures that can be measured based on the functional similarity of two miRNAs, for example, MA and MB. First, the diseases associated with these two miRNAs were identified, denoted as DA and DB. Next, the semantic values of diseases were calculated based on the DAG of corresponding diseases. Third, the semantic similarity for any pair of diseases between DA and DB was calculated based on disease semantic value. Finally, the functional similarity of MA and MB was calculated based on the semantic similarity of DA and DB. Details of the procedures are given in the following sections.

2.3.1 Semantic value of a disease All the denominations of diseases used in this study were in accordance with the MeSH database (http://www.ncbi.nlm.nih.gov/). The MeSH database provided a strict system for disease classification and could be helpful for studying the relationship of diseases. It can be described as a DAG, in which the nodes represent diseases while the links represent relationship between nodes. There is only one type of relationship, defined as 'is-a', to connect a child node to a parent node. Each disease has one or more addresses in the DAG, referred herein as codes, to numerically define its location in the MeSH graph. The codes of a child node are defined as the codes of its parent nodes appended by the child's addresses. For instance, the entry on breast neoplasms has two possible addresses or codes: C04.588.180 and C17.800.090.500. Their corresponding parent nodes are C04.588 neoplasms by site and C17.800.090 breast diseases (Fig. 2). A disease A can be represented as a graph, $\mathrm{DAG}_A = (A, T_A, E_A)$, where T_A is the set of all ancestor nodes of A including node A itself and E_A is the set of corresponding links. We define the contribution of a disease t in

DAG_A to the semantics of disease A as the D value of disease t related to disease A, $D_A(t)$, which can be calculated by

$$\begin{cases} D_A(A)=1 \\ D_A(t)=\max\{\Delta * D_A(t') | t' \in \text{children of } t\} & \text{if } t \neq A \end{cases} \quad (1)$$

where Δ is the semantic contribution factor for edges (E_A) linking disease t with its child disease t'. In the DAG of disease A, disease A is the most specific disease and therefore we define its contribution to its own semantic value as one. Those ancestor nodes which are located farther from node A are more general denominations. For example as shown in Figure 2, C17 'Skin and Connective Tissue Diseases' is more general than C17.800 'Skin Diseases' and the later is more general than C17.800.090 'Breast Diseases'. We here presume that such farther ancestor nodes contribute less to the specific semantic value of node A. Therefore, Δ should be chosen between 0 and 1 to reduce the contributions of ancestor nodes that are far from A. To find a suitable value for Δ, we tuned the parameter of Δ using different values, such as 0.3, 0.4, 0.5, 0.6 and 0.7. We found that MISIM similarity shows better correlation with expression similarity when $\Delta = 0.5$ than other values. Therefore, we set Δ as 0.5 in this study. Based on Equation (1), we then define the semantic value of disease A, DV(A) as

$$\text{DV}(A)=\sum_{t \in T_A} D_A(t). \quad (2)$$

As an example, the DV value of 'breast neoplasms' is 1.0 (breast neoplasms) + 0.5 (breast diseases) + 0.5 (neoplasms by site) + 0.5 × 0.5 (neoplasms) + 0.5 × 0.5 (skin diseases) + 0.5 × 0.5 × 0.5 (skin and connective tissue diseases) = 2.6250.

2.3.2 Semantic similarity of two diseases We presented the measurement of the semantic similarity of two diseases by considering their relative positions in the MeSH disease DAG. We assumed that diseases that share larger part of their DAGs tend to have higher semantic similarity. The semantic similarity of two diseases is defined as

$$S(A,B)=\frac{\sum_{t \in T_A \cap T_B}(D_A(t)+D_B(t))}{\text{DV}(A)+\text{DV}(B)} \quad (3)$$

where $D_A(t)$ is the semantic value of disease t related to disease A and $D_B(t)$ is the semantic value of disease t related to disease B. Formula (3) calculates the semantic similarity of two diseases based on both the addresses of these diseases in DAG graphs and their semantic relations with their ancestor diseases.

2.3.3 miRNA MISIM functional similarity To accurately measure the functional similarity between two miRNAs, we need also consider the contributions from similar diseases that are associated with these two genes, respectively. Therefore, we need to first define semantic similarity between one disease and one group of disease. Here we let 'dt' represent one disease and let 'DT' represent one disease group. We then define the similarity of dt and DT, $S(dt, DT)$, as the maximum similarity between one disease and a disease group, e.g. $DT = \{dt_1, dt_2, \ldots dt_k\}$. It is calculated as follows:

$$S(dt,DT)=\max_{1 \le i \le k}(S(dt,dt_i)). \quad (4)$$

To better describe the method measuring miRNA functional similarity, here we take hsa-mir-103 and hsa-mir-151 as an example. Assuming DT_1 represents the related diseases (a group of diseases) of hsa-mir-103 and DT_2 represents the related diseases (another group of diseases) of hsa-mir-151. DT_1 contains m diseases, and DT_2 contains n diseases. The functional similarity of two miRNAs need consider all diseases DT_1 in and DT_2. We therefore define the functional similarity of two miRNAs as

$$\text{MISIM}(\text{M1},\text{M2})=\frac{\sum_{1 \le i \le m} S(dt_{1i},DT_2)+\sum_{1 \le j \le n} S(dt_{2j},DT_1)}{m+n}. \quad (5)$$

As a result, the functional similarity of hsa-mir-103 and hsa-mir-151 is calculated to be 0.80.

2.3.4 Construction of miRNA functional network It is increasingly important to investigate the biology problems at the systems level (Horvath *et al.*, 2006; Sharan *et al.*, 2007). In various biological networks, gene functional network construction and analysis is considered one of the most popular topics, and some important findings have been obtained in this respect. In the same vein, constructing a reliable miRNA functional network is increasingly necessary for better understanding of miRNAs. As described above, MISIM is a reliable measurement of miRNA functional similarity, the most critical problem in miRNA functional network construction; in doing so, constructing a miRNA functional network can be easily performed. For a list of interesting miRNAs, we first calculated their pairwise MISIM functional similarity coefficients. We then set up a MISIM threshold, for example, 0.7, to determine whether two miRNAs have a link. miRNA pairs with MISIM coefficient greater than or equal to the threshold will be connected by a direct link; otherwise, they are not connected directly. Finally, a miRNA functional network can be constructed by this approach.

3 RESULTS

3.1 MISIM functional similarity of miRNAs

We applied our method on the MISIM to miRNAs recorded in HMDD, and then calculated the functional similarity of all miRNA pairs. As a result, we obtained the pairwise MISIM functional similarity of 271 miRNAs. We further evaluated the accuracy of our method by investigating the relationship of the calculated functional similarity with miRNA expression similarity, family and cluster, manually annotated functional relationship for host genes of intronic miRNAs, and miRNA targets.

3.2 miRNA MISIM functional similarity is correlated with expression similarity

miRNAs with similar functions tend to be involved in similar biological processes and interact with similar cellular components. Hence, it is possible that miRNAs with similar functions tend to have similar expression profiles. To validate this, we explored the relationship of miRNA functional similarity calculated by MISIM to expression similarity. Like protein-coding genes (Horvath *et al.*, 2006), in this study, we used absolute Pearson's correlation coefficients as the measure for expression similarity of miRNAs. We obtained the miRNA expression data from 40 normal tissues from Liang's study (Liang *et al.*, 2007). We then calculated miRNA expression similarity using Liang's miRNA expression data (Liang *et al.*, 2007), followed by a correlation analysis for functional similarity and expression similarity. As a result, miRNA functional similarity showed positive correlation with expression similarity ($R = 0.05$, $P = 2.70 \times 10^{-12}$, Pearson's correlation). We further grouped miRNA pairs into different groups according to functional similarity by a step of 0.1 and calculated the average functional similarity and expression similarity of each group. Clearly, miRNA functional similarity is positively correlated with expression similarity ($R = 0.8685$, $P = 5 \times 10^{-4}$; Fig. 3A). Results indicate that functional similarity inferred by our method is correlated with expression similarity, which is well known to be associated with functional similarity.

3.3 miRNAs in the same family or cluster show high MISIM functional similarity

A family of miRNAs incorporates similar mature miRNA sequences and complete identical seed regions, which are widely accepted as

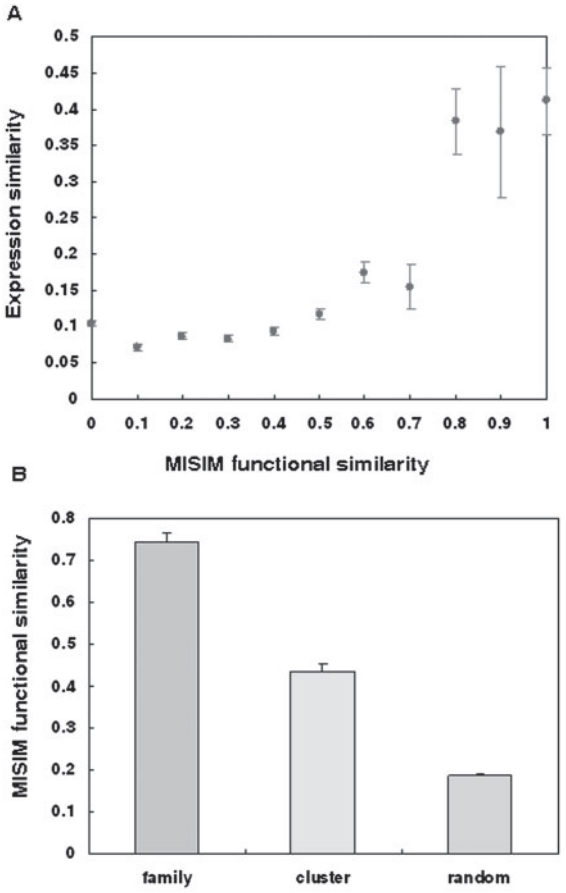

Fig. 3. (**A**) The relationship between miRNA MISIM functional similarity and their expression similarity. (**B**) A comparison of MISIM functional similarity of miRNAs in the same family, miRNAs in the same cluster and miRNAs of random pair.

the 'key' regions for miRNA target reorganization (Bartel, 2009). Therefore, miRNAs of the same family tend to show high functional similarity. We first downloaded miRNA family data from miRBase (Griffiths-Jones, 2004). To evaluate the reliability of functional similarity in our method calculations, we compared the functional similarity of miRNAs in the same family to miRNAs that are not from the same family. As a result, the functional similarity of miRNAs in the same family was significantly greater compared with other miRNAs ($P = 2.7 \times 10^{-52}$, *t*-test; Fig. 3B).

It has been reported that some miRNAs tend to organize into very compact clusters in the genome (Baskerville and Bartel, 2005). A cluster of miRNAs is usually transcribed and expressed synchronously and functions coordinately (Baskerville and Bartel, 2005). Therefore, miRNAs in the same cluster are expected to have higher functional similarity. In order to confirm this, we first downloaded the miRNA genome coordinate data from miRBase (Griffiths-Jones, 2004) and then identified miRNA clusters by setting the distance cutoff between miRNAs as 50 kb base on the miRNA genome coordinate data, as previously suggested (Baskerville and Bartel, 2005). We next calculated the functional similarity of miRNAs in the same clusters and compared it with that of random miRNA pairs that belong to neither the same family nor the same

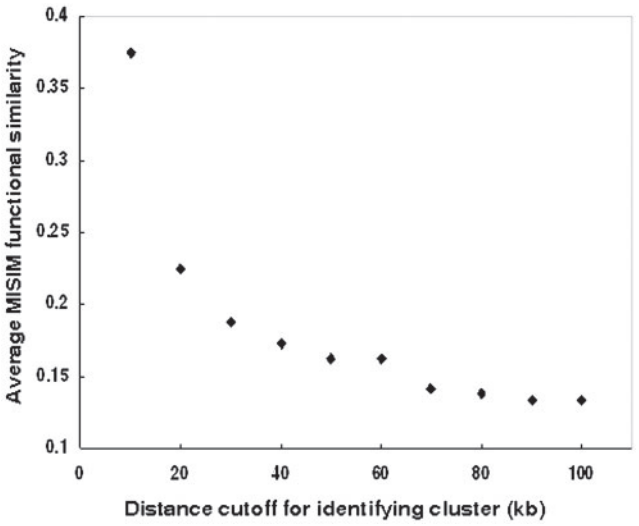

Fig. 4. The effects of distance cutoff for identifying miRNA clusters on the MISIM functional similarity of miRNAs in the same clusters.

cluster. As expected, clustered miRNAs showed higher MISIM functional similarity than random miRNA pairs ($P = 6.5 \times 10^{-34}$, *t*-test; Fig. 3B).

The reason we selected 50 kb as a distance cutoff to identify miRNA clusters is that Baskerville *et al.* observed that the correlation of expression profiles decreases sharply for miRNAs whose distances are more than 50 kb, while miRNAs whose distances are within 50 kb usually show high expression similarity (Baskerville and Bartel, 2005). Specifically, 50 kb is suggested based on observations from miRNA expression. However, the relationship of distance cutoff and miRNA functions remains unclear. Therefore, it would be interesting to investigate the MISIM functional similarity of miRNAs in the same clusters identified using different distance cutoffs. We first identified miRNA clusters using different distance cutoffs from 100 kb to 10 kb by a step of 10 kb. Next, we calculated the MISIM functional similarity of miRNAs in the same clusters. As a result, from 30 to 100 kb, the MISIM similarity in a cluster changed minimally, whereas for clusters identified using distance cutoff <30 kb, significantly higher MISIM functional similarity was observed (Fig. 4). This suggests that a distance cutoff of 20 kb is more reliable for inferring functional relationships of miRNAs based on miRNA clusters.

3.4 Novel highly functionally similar miRNAs

As we have described above, expression similarity and sequence similarity of miRNAs can only interpret part of the functional similarity. Some functionally similar miRNA pairs neither have high expression similarity nor belong to the same family or cluster. Their functional similarities have not been supported by existing knowledge or data. Although we cannot provide direct evidence for the high functional similarity of this part of miRNAs, the data for the relationship of intronic miRNAs and their host genes may present some clues. It is well known that some miRNAs are located within the intron regions of protein-coding genes, which are named as their host genes. It has been confirmed that many intronic miRNAs show high expression similarity with their host

genes (Baskerville and Bartel, 2005; Gennarino *et al.*, 2009; Wang *et al.*, 2009). This information had been confirmed to be helpful in the prediction of miRNA targets (Gennarino *et al.*, 2009). Although there is no experimental evidence showing that intronic miRNAs and host genes share common molecular functions, we believe that intronic miRNAs and host genes should have higher functional similarity than random miRNA and protein-coding gene pairs. Taking off from this point, we made the following presumption: if the host genes of two intronic miRNAs are functionally related, intronic miRNAs tend to have more functional common parts. We downloaded miRNA genome coordinate data from miRBase (Griffiths-Jones, 2004) and downloaded protein-coding gene genome coordinate data from the University of California at Santa Cruz (Karolchik *et al.*, 2004). We identified intronic miRNAs and their host genes by mapping miRNAs that are within introns of protein-coding genes. We further manually annotated the function of these host genes based on the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) and determined whether the function of two host genes is highly similar or not. As a result, we identified 55 pairs of intronic miRNAs that have high MISIM coefficients, after which the function of their host genes were manually annotated. Surprisingly, 27 (49%) pairs of host genes appeared to be functionally connected with each other (Supplementary Material 3). For example, mir-103 and mir-151 have a high MISIM coefficient of 0.9. The host gene of mir-103, PANK3, has functions like ATP binding, nucleotide binding, pantothenate kinase activity and transferase activity. The host gene of mir-151, PTK2, has functions like ATP binding, nucleotide binding, protein binding, signal transducer activity and transferase activity. This indicates that the host genes of the two miRNAs are functionally related, which further suggests that miRNAs with high MISIM coefficient are indeed functionally related. Interestingly, there are still over 100 pairs with high MISIM coefficient that cannot be supported by current knowledge (Supplementary Material 4). Considering the high reliability of our method, which was confirmed by the above analysis, novel and highly functionally similar miRNAs are deemed helpful in understanding miRNA function and disease, as well as in exploring the novel mechanism connecting miRNAs in the function.

3.5 MISIM functional similarity of two miRNAs is correlated with the fraction of their common targets

miRNAs exert their functions by regulating target genes. It is expected that miRNAs that have higher fraction of common targets will have higher functional similarity. Therefore, it is reasonable to validate our method by comparing the MISIM functional similarity of two miRNAs with the fraction of their common targets. To confirm this, we first performed the analysis based on experimentally supported miRNA targets from TarBase (Papadopoulos *et al.*, 2009). As a result, only a limited number of experimentally supported targets are available. The common targets of different miRNAs are rare. Therefore, it is not feasible to perform correlation analysis for MISIM functional similarity and targets. We next repeated the above analysis based on targets predicted by TargetScan (Lewis *et al.*, 2005). As expected, miRNAs with higher MISIM functional similarity indeed have higher fraction of common targets ($R = 0.80$, $P = 0.006$, Pearson's correlation; Fig. 5). This result further supported that MISIM is valid to measure miRNA functional similarity. Furthermore, similar with Figure 3A, for miRNA pairs with MISIM similarity higher than 0.7, the fraction of their common targets of their total targets increases dramatically, suggesting that miRNA pairs with MISIM similarity higher than 0.7 are highly reliable to be really functionally related.

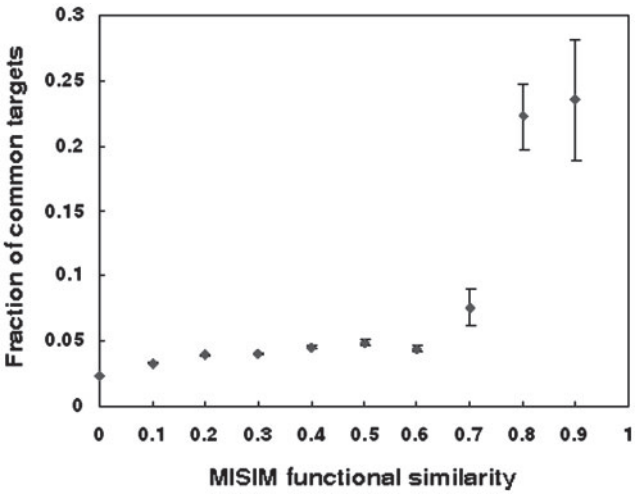

Fig. 5. The relationship between miRNA MISIM functional similarity and the fraction of their common targets. The *y*-axis indicates the fraction of common targets shared by two miRNAs.

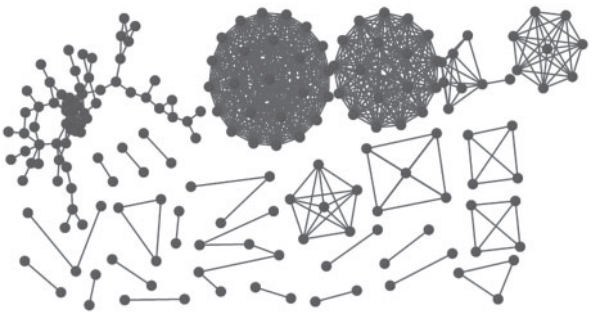

Fig. 6. miRNA functional network constructed based on miRNA MISIM functional similarity. Each node represents one miRNA and the edges linking any two nodes (miRNAs) indicate that the functional similarity of the two miRNAs is equal to or greater than the similarity cutoff (here the cutoff is 0.7). The network is visualized by Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

3.6 A miRNA functional network

By applying this procedure to inputted miRNAs, the original miRNA pairs are converted into a network based on their MISIM coefficients (Fig. 6; Supplementary Material 5). Similar to most of the reported biological networks, the degree of this miRNA functional network (MISIM threshold = 0.7) also shows a scale-free distribution (Supplementary Fig. 1), which means that most of the miRNAs only have a few functionally similar miRNAs, but there are indeed some miRNAs that have a numerous miRNAs that are functionally similar. We next identified network components using Pajek (Supplementary Material 6), a free network analysis

tool (http://vlado.fmf.uni-lj.si/pub/networks/pajek/). Among all the miRNAs, 31.7% (86) of them do not have any link with other miRNAs. The giant network component contains 23.6% (64) of the total miRNAs. To investigate the correlation between link numbers a network has and the functional similarity cutoff used to construct a miRNA functional network. We constructed networks using various similarity cutoffs. As shown in Supplementary Figure 2, the number of links dramatically decreases when the cutoff increases from low value to high value. When the cutoff is equal to or bigger than 0.7, the link numbers remain relatively stable. Considering the results of expression correlation and common target correlation with functional similarity, it seems that the cutoff 0.7 or 0.8 is suitable for the construction of miRNA functional networks.

Considering the relationship of the degree to the gene in a functional network (Horvath *et al.*, 2006), it is convenient to determine the important genes from a group using a gene functional network. This method was carried out by Horvath *et al.* to identify ASPM as a candidate glioblastoma molecular target (Horvath *et al.*, 2006). Furthermore, some general methods of biological networks can also be used to infer useful information from this network. For example, the method presented by Clauset *et al.* can be used to predict potential novel links in the network (Clauset *et al.*, 2008). We applied this method to predict novel links among nodes. The probability of each predicted link was calculated. To confirm the validation of the predicted links, we performed analysis of correlation between the probability of predicted links and the MISIM functional similarity of miRNAs connected by the corresponding links. The result showed a significantly positive correlation between these two variables ($R = 0.34$, $P = 1.0 \times 10^{-6}$, Spearman's correlation test). This indicated that the predicted novel links with high probability tend to be real links. Furthermore, novel miRNA–disease associations can be predicted through miRNA pairs with high MISIM similarity. For example, we predicted novel miRNA–disease associations for miRNAs with MISIM similarity between 0.7 and 0.9. As a result, 10 of the novel predicted miRNA–disease associations were supported by newly published literature. These associations are mir-18a and mir-19a versus neuroblastoma, mir-34 and mir-200a versus pancreatic neoplasms, let-7c versus carcinoma, hepatocellular, mir-200b versus adenocarcinoma, mir-200a versus uterine cervical neoplasms, mir-18b versus breast neoplasms, mir-133a versus myocardial infarction and mir-199a versus stomach neoplasms. Overall, the methods presented in this study can conveniently discover potentially important miRNAs in a biological experiment.

3.7 Comparisons with other semantic similarity methods

In this study, we used the algorithm presented by Wang *et al.* (2007) to measure miRNA functional similarity based on the DAG structures of miRNA-associated diseases. This algorithm is originally presented for GO similarity measuring based on GO. Because many methods for GO similarity measuring based on GO have been presented (Pesquita *et al.*, 2009), here for comparisons, we further implemented other two methods, the method of most informative common ancestor (MICA) and the method of improved shortest path (see the review paper by Pesquita *et al.*, 2009). As a result, the method of improved shortest path did not achieve a good result (data not shown). MICA obtained a significant result but is a little bit worse than the current method we used (i.e. for correlation analysis of miRNA functional similarity with expression similarity, $R = 0.77$, $P = 0.005$). In addition, we provided a choice for MICA method in our web server.

4 DISCUSSION

In summary, we presented a method for measuring miRNA functional similarity and construction of miRNA functional networks. Results show that our method is reliable and can be used to infer potential function and/or associated diseases for miRNAs. Moreover, a web-accessible program that implements the method of MISIM is also available at http://cmbi.bjmu.edu.cn/misim.

As described above, since MISIM has calculated miRNA functional similarity based on miRNA–disease association data and the disease DAG, MISIM may generate bias in some cases, especially when little disease association data is available for a miRNA. Therefore, in the future, MISIM will improve greatly when more miRNA–disease association data and more accurate disease relationship are available. We believe that with the rapid increase of miRNA–disease association data (Fig. 1), MISIM will play more important roles in the analysis of miRNAs.

ACKNOWLEDGEMENTS

We thank the three reviewers for their valuable comments and suggestions.

Funding: Natural Science Foundation of China (Grant no. 30900829).

Conflict of Interest: none declared.

REFERENCES

Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.

Baskerville,S. and Bartel,D.P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, **11**, 241–247.

Berezikov,E. *et al.* (2006) Approaches to microRNA discovery. *Nat. Genet.*, **38**(Suppl. 38), S2–S7.

Clauset,A. *et al.* (2008) Hierarchical structure and the prediction of missing links in networks. *Nature*, **453**, 98–101.

Du,Z. *et al.* (2009) G-SESAME: web tools for GO-term-based gene similarity analysis and knowledge discovery. *Nucleic Acids Res.*, **37**, W345–W349.

Esquela-Kerscher,A. and Slack,F.J. (2006) Oncomirs - microRNAs with a role in cancer. *Nat. Rev. Cancer*, **6**, 259–269.

Gennarino,V.A. *et al.* (2009) MicroRNA target prediction by expression analysis of host genes. *Genome Res.*, **19**, 481–490.

Goh,K.I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res.*, **32**, D109–D111.

Horvath,S. *et al.* (2006) Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. *Proc. Natl Acad. Sci. USA*, **103**, 17402–17407.

Karolchik,D. *et al.* (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.

Latronico,M.V. *et al.* (2007) Emerging role of microRNAs in cardiovascular biology. *Circ. Res.*, **101**, 1225–1236.

Lee,S.G. *et al.* (2004) A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, **20**, 381–388.

Lewis,B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Liang,Y. *et al.* (2007) Characterization of microRNA expression profiles in normal human tissues. *BMC Genomics*, **8**, 166.

Lin,J. *et al.* (2007) A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res.*, **17**, 1304–1318.

Lord,P.W. *et al.* (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, **19**, 1275–1283.

Lu,M. *et al.* (2008) An analysis of human microRNA and disease associations. *PLoS One*, **3**, e3420.

Papadopoulos,G.L. *et al.* (2009) The database of experimentally supported targets: a functional update of TarBase. *Nucleic Acids Res.*, **37**, D155–D158.

Pesquita,C. *et al.* (2008) Metrics for GO based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9**(Suppl. 5), S4.

Pesquita,C. *et al.* (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, e1000443.

Sharan,R. *et al.* (2007) Network-based prediction of protein function. *Mol. Syst. Biol.*, **3**, 88.

Wang,D. *et al.* (2009) Cepred: predicting the co-expression patterns of the human intronic microRNAs with their host genes. *PLoS One*, **4**, e4421.

Wang,J.Z. *et al.* (2007) A new method to measure the semantic similarity of GO terms. *Bioinformatics*, **23**, 1274–1281.