Bioinformatics, 34(20), 2018, 3547–3556
doi: 10.1093/bioinformatics/bty343
Advance Access Publication Date: 27 April 2018
Original Paper

Databases and ontologies

MiRGOFS: a GO-based functional similarity measurement for miRNAs, with applications to the prediction of miRNA subcellular localization and miRNA–disease association

Yang Yang[1,2,*], Xiaofeng Fu[1], Wenhao Qu[1], Yiqun Xiao[1] and Hong-Bin Shen[3,4,*]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, [2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai 200240, China, [3]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China and [4]Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on November 5, 2017; revised on March 31, 2018; editorial decision on April 22, 2018; accepted on April 26, 2018

Abstract

Motivation: Benefiting from high-throughput experimental technologies, whole-genome analysis of microRNAs (miRNAs) has been more and more common to uncover important regulatory roles of miRNAs and identify miRNA biomarkers for disease diagnosis. As a complementary information to the high-throughput experimental data, domain knowledge like the Gene Ontology and KEGG pathway is usually used to guide gene function analysis. However, functional annotation for miRNAs is scarce in the public databases. Till now, only a few methods have been proposed for measuring the functional similarity between miRNAs based on public annotation data, and these methods cover a very limited number of miRNAs, which are not applicable to large-scale miRNA analysis.

Results: In this paper, we propose a new method to measure the functional similarity for miRNAs, called miRGOFS, which has two notable features: (i) it adopts a new GO semantic similarity metric which considers both common ancestors and descendants of GO terms; (i) it computes similarity between GO sets in an asymmetric manner, and weights each GO term by its statistical significance. The miRGOFS-based predictor achieves an F_1 of 61.2% on a benchmark dataset of miRNA localization, and AUC values of 87.7 and 81.1% on two benchmark sets of miRNA–disease association, respectively. Compared with the existing functional similarity measurements of miRNAs, miRGOFS has the advantages of higher accuracy and larger coverage of human miRNAs (over 1000 miRNAs).

Availability and implementation: http://www.csbio.sjtu.edu.cn/bioinf/MiRGOFS/

Contact: yangyang@cs.sjtu.edu.cn or hbshen@sjtu.edu.cn

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

1 Introduction

MicroRNA (miRNA), as an important regulatory molecule, plays a crucial role in many fundamental biological processes (He and Hannon, 2004). Especially, numerous studies have demonstrated that miRNAs can be promising biomarkers for various diseases (Esquela-Kerscher and Slack, 2006; Lu *et al.*, 2005). Therefore, identifying the cellular functions of miRNAs and predicting their associations with diseases have been important tasks in the field of bioinformatics, and a lot of computational tools have been developed, e.g. for whole-genome expression analysis (Thomson *et al.*, 2004), target prediction (Agarwal *et al.*, 2015; Fan and Kurgan, 2015; Peterson *et al.*, 2014) and functional enrichment (Bleazard *et al.*, 2015; Gusev *et al.*, 2007; Vlachos *et al.*, 2012).

In gene function analysis, the domain knowledge, like gene ontology (GO) (Ashburner *et al.*, 2000) and KEGG pathway (Kanehisa and Goto, 2000), is often utilized as complementary information to the high-throughput experimental data in various computation tasks, such as gene clustering (Huang and Pan, 2006), disease-related gene identification (Schlicker *et al.*, 2010) and protein–protein-interaction prediction (Mahdavi and Lin, 2007). In order to quantitatively measure the functional similarity between two genes, the annotation terms associated with the genes are extracted and their semantic correlation is computed. However, due to the lack of functional annotation of miRNAs in public database, such analyses for miRNAs are relatively difficult. The measuring of miRNA functional similarity is usually based on the shared targets or associated diseases. For instance, Wang *et al.* (2010) inferred miRNA similarity by their association with diseases and the semantic correlation between diseases (diseases are organized hierarchically via MeSH descriptor), but the number of miRNAs having known disease association is very limited. In order to use more available information source, Yu *et al.* (2011) proposed to convert GO information from target genes into the functional similarity between miRNAs. They provided a similarity matrix for 533 human miRNAs, and performed clustering on the matrix. However, the authors did not show how to apply the similarity scores to miRNA-related prediction tasks and there was no large-scale experiment to quantitatively evaluate the similarity scores. Besides, the coverage of miRNAs is still low, considering the human genome may encode over 1000 miRNAs (Bentwich *et al.*, 2005; Griffiths-Jones *et al.*, 2006).

In this paper, we also propose a GO-based functional similarity for miRNAs, called miRGOFS. Different from Yu *et al.* (2011)'s method, miRGOFS is featured by a new GO semantic metric and a new integration strategy for comparing the sets of GO terms. Using miRGOFS, we compute pairwise similarities for a total of 1100 miRNAs, including all the human miRNAs which have predicted targets in the `microRNA.org` target resource (Betel *et al.*, 2007). This is, as far as we know, the largest similarity matrix of human miRNAs.

To verify the performance of miRGOFS, we apply it to two computational tasks of miRNAs. One is the prediction of miRNA–disease associations, which has been a hot topic in miRNA research (Chen and Zhang, 2013; Chen *et al.*, 2016a,b, 2017a,b,c, 2018; You *et al.*, 2017; Zeng *et al.*, 2016). The other is the identification of miRNA subcellular locations, which has not been well studied in computational biology. Similar to proteins, miRNAs should be located in the right subcellular compartments to play their functions (Lee *et al.*, 2002; Leung and Sharp, 2006). Recent studies have revealed the extremely abundant localization patterns of miRNAs: the mature miRNAs can target to multiple cellular compartments in cytoplasm, such as mitochondria, endoplasmic reticulum, RNA granules, or be secreted out of cells via exosomes; and they can even locate at nucleus and function in epigenetic regulation (Leung, 2015). Moreover, in the pharmaceutical industry, recent miRNA-targeted therapeutics can not only interfere with the miRNA–target interaction, but also modify the subcellular localization of miRNAs, thus changing their roles in disease progression (Abba *et al.*, 2017). The localization information can provide important insights into miRNAs functions. However, to our knowledge, there has been no computational method for miRNA subcellular localization. It could be due to the lack of information source. Most protein subcellular localization methods utilize sequences and annotation data, like gene ontology and function domain. By contrast, miRNAs have very short sequences which have limited discriminative power, and their functional annotation in public databases is also scarce. While the prediction of protein subcellular localization benefits a lot from GO information, whether the GO information of target genes provides a hint for miRNA subcellular localization is unknown. In this study, we assess the performance of GO-based functional similarity in the prediction of miRNA subcellular localization. The proposed GO-based similarity scores can serve as a basic feature in various prediction tasks related to miRNA functions.

2 Related work

2.1 Semantic similarity between GO terms

GO describes gene attributes in standardized terms, and organizes them in directed acyclic graphs (DAGs) including nodes (terms) and relationships (edges). How to measure semantic correlations between terms has been a central problem in the information extraction of GO database (Couto *et al.*, 2007; Lord *et al.*, 2002; Xu *et al.*, 2013). In GO DAGs, the similarity of two GO terms relies not only on the distance between them, i.e. length of the path connecting the two nodes, but also on the location/depth of these two terms in the DAG. With the same length of path from each other, a pair of GO terms close to the root of the DAG would be less similar than a pair of leaf nodes, because the lower level terms represent more specific attributes, corresponding to higher information content (IC) (Resnik *et al.*, 1999), as defined in Eq. (1),

$$IC(x) = -\log p(x) = -\log\left(\frac{|\mathcal{G}_x|}{|\mathcal{G}_{root}|}\right), \tag{1}$$

where the probability of GO term x is computed according to a gene–GO mapping file. $\mathcal{G}_x$ is the set of genes that are associated with x, i.e. the genes annotated by x or the descendants of x, and $\mathcal{G}_{root}$ is the set of genes that are associated with the root term, i.e. the total number of genes that the DAG annotates. In GO DAGs, parent nodes denote generalized concepts of their child nodes. Thus, for x, the genes annotated by its descendants also associate with it. This is the so-called 'true path rule'. According to this rule, $p(x)$ can be estimated as,

$$p(x) = \sum_{t\in\mathcal{D}_x} p^*(t) = \sum_{t\in\mathcal{D}_x} \frac{|\mathcal{G}_t^*|}{\sum_{t'\in\mathcal{N}} |\mathcal{G}_{t'}^*|}, \tag{2}$$

where $\mathcal{D}_x$ denotes the descendant set of x, $p^*(x)$ is the probability of x annotating a gene or gene product. $\mathcal{G}_x^*$ is the set of genes that are directly annotated by x ($\mathcal{G}_x^* \subseteq \mathcal{G}_x$), and $\mathcal{N}$ is the full set of nodes in the DAG. For two query nodes, Resnik *et al.* (1999) selected the maximum value among the ICs of lowest common ancestors (LCAs) as the similarity of the query pair. Considering that the lengths of paths from the two query nodes to their LCAs may differ,

Lin (1998) proposed to divide the IC of LCA by the averaged IC of the two query nodes.

Early studies on semantic metric between GO terms are mostly based on information content, where an external data source recording all the gene–GO associations is required. Most of the methods treat GO DAGs as trees or consider only one LCA (with the maximum IC value) (Jiang and Conrath, 1997; Lin, 1998; Resnik *et al.*, 1999). Later, the structure-based methods have emerged (Wang *et al.*, 2007; Wu *et al.*, 2005), which rely on structural information, e.g. edges, of the DAGs. These methods addressed the multi-LCA issue, but very few of them utilized the common descendant information (Yang *et al.*, 2012). In this study, we propose a new IC-based method, which takes both common ancestors and descendants into consideration.

2.2 GO-based similarity for miRNAs

As a gene product is generally annotated by multiple GO terms, the functional similarity between two genes can be inferred by integrating GO term similarities from their GO sets (Pesquita *et al.*, 2008; Teng *et al.*, 2013). Several methods have been proposed, such as MAX, AVG, RCMAX and BMA (Supplementary Equations (S1)) (Yu *et al.*, 2010). The BMA, i.e. best match average, is widely accepted because of its excellent performance (Azuaje *et al.*, 2005; Schlicker *et al.*, 2006).

Since miRNAs mainly play their functions via their target genes, the functional similarity of miRNAs can be estimated by computing the similarity between the two gene sets corresponding to the two miRNAs. Especially, Yu *et al.* (2011) proposed to compute GO-based similarity for miRNAs via two steps: (i) integrating GO semantic similarities into the similarities between two target genes, (ii) integrating target similarities into miRNA similarities, where both steps adopted the BMA rule, and they suggested to use the GO semantic similarity proposed by Wang *et al.* (2007), i.e. G-SESAME.

In this study, we also infer the functional similarity for miRNAs based on the GO annotation of their target genes. However, our method is more straightforward. By combining the GO sets of the target genes into a whole set, each miRNA is associated with a GO set with redundant terms. Then, we compute the similarity for each pair of miRNAs by evaluating the similarity between their corresponding GO sets. Especially, we assign a weight for each GO term based on its statistical significance in the set, and also develop a new integration rule of the GO semantic similarities.

3 Materials and methods

3.1 Datasets

3.1.1 MiRNAs and their target genes

A lot of public databases provide target information for miRNAs, such as TarBase (Vlachos *et al.*, 2015), TargetScan (Lewis *et al.*, 2003), PicTar (Krek *et al.*, 2005), miRanda (John *et al.*, 2004), DIANA-microT-v4 (Reczko *et al.*, 2012) and mirDB (Wong and Wang, 2015). TarBase houses the experimentally validated miRNA-gene interactions, while the others provide computational tools for miRNA target prediction. In order to enable the analysis for large-scale miRNA dataset, we download miRNA target information from two databases, i.e. microRNA.org and mirDB, which adopt miRanda and MirTarget V3 as the prediction tool, respectively. The microRNA.org (released August, 2010) provides computationally predicted targets with good mirSVR scores for both conserved and non-conserved human miRNAs from www.microrna.org, including 1100 miRNAs (here the 'good mirSVR score' means the mirSVR value is less than –0.1); while mirDB (Version 5.0 released August, 2014) covers even more miRNAs (2588 human miRNAs). Note that these two databases have different settings of stringency for the predicted targets. Specifically, the average numbers of targets per miRNA in microRNA.org and mirDB are 717 and 4016, respectively, indicating that microRNA.org has a much looser confidence threshold for the identification of targets.

3.1.2 The construction of benchmark set for miRNA subcellular localization

We extract the subcellular locations of miRNAs from a comprehensive RNA database, RNAlocate (Zhang *et al.*, 2016), which covers localization information of mRNAs, miRNAs, lncRNAs, etc. (http://www.rna-society.org/rnalocate). It houses more than 37 700 manually curated RNA-associated subcellular localization entries with experimental evidence. In this study, we focus on the human miRNAs due to their important roles in the development of complex diseases. The construction of the benchmark dataset consists of three steps:

1. Download all 7449 human miRNA entries with curated subcellular localization from the RNAlocate database, and merge them into 1048 unique miRNAs, as multi-locational miRNAs have multiples records in the database (We check aliases in `miRBase.org`);
2. Remove miRNAs that are not covered in `microRNA.org` and get 813 miRNAs, including 266 mono-locational ones and 547 multi-locational ones;
3. Further remove three locations, i.e. endoplasmic reticulum, extracellular vesicle and nucleolus, because they have too few samples.

Finally, we obtain a benchmark dataset of 813 miRNAs, covering 6 subcellular compartments, as shown in Table 1.

3.1.3 Benchmark datasets of miRNA–disease association

During the last decade, a lot of computational methods for the prediction of miRNA–disease associations have been developed and various public databases and benchmark datasets have merged, such as HMDD (Li *et al.*, 2013) and miR2Disease (Jiang *et al.*, 2009), which house the miRNA–disease associations reported in the existing literatures. In order to compare the proposed functional similarity metric with other metrics, we select two widely used benchmark sets, and name them as Data1 and Data2 as follows.

> **Data1** contains all the records in HMDD database (released on September 2009) created by Wang *et al.* (2010), including 1616 miRNA–disease associations. In order to generate the MISIM (miRNA similarity) scores, Wang *et al.* (2010) merged the records with the same mature miRNAs (such as hsa-mir-376a-1

Table 1. Data distribution of the benchmark set of miRNA subcellular localization

Location	# miRNA
Cytoplasm	206
Microvesicle	329
Mitochondrion	294
Nucleus	319
Circulating	419
Exosome	712
Total label #	2279
Total miRNA#	813

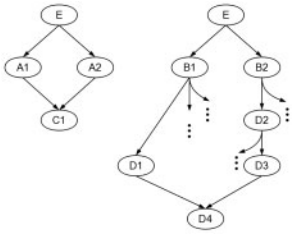

Fig. 1. An example of two pairs of nodes with the same ancestor but different local structure of descendants

and hsa-mir-376a-2), then the set includes 1395 associations, covering 271 miRNAs and 137 diseases. In this set, 267 miRNAs have target information in `microRNA.org`. Thus, the final dataset used in our experiment includes 1388 associations, 267 miRNAs and 137 diseases.

Data2 was initially collected from HMDD and miR2Disease by Jiang *et al.* (2010), and contains 270 experimentally verified microRNA–disease associations. Then, Chen and Zhang (2013) used this dataset to validate their proposed methods, PBSI, MBSI and NetCBI, which also rely on MISIM, and they removed 19 miRNAs that are not covered in MISIM, thus remaining 242 miRNA–disease associations, including 99 miRNAs and 51 diseases. In order to compare our result with these three methods, we adopt the reduced set (242 associations) in our experiments, where all the miRNAs are covered in miRGOFS.

3.2 Methods

3.2.1 GO term similarity

As stated in Section 2.1, for IC-based methods, the closeness of a pair of GO terms is normally measured by the IC of their lowest common ancestors (LCAs). In this study, we take not only LCAs but also HCDs (highest common descendants) into consideration, which were often neglected in previous studies. The importance of descendant information is illustrated in Figure 1. For the two pairs, (A_1, A_2) and (B_1, B_2), the LCAs are their adjacent ancestor, E, but the local structure below the two pairs are very different. Apparently, A_1 and A_2 are more similar than B_1 and B_2.

According to the definition of LCA (Bender *et al.*, 2005), if a node is a common ancestor of x and y, and it is not an ancestor of any other common ancestor of x and y, then the node belongs to the LCA set. The HCD set can be defined analogously. For example, in Supplementary Figure S1, GO: 0006119 is an HCD of GO: 0006091 and GO: 0016310, and GO: 000977 is also an HCD of them although it is lower than GO: 0006119 in the DAG.

In order to utilize HCD information, we propose a new IC-based metric as shown in Eq. (3). It has three terms, denoting the relative distance from LCA to x, y and HCD, respectively,

$$Sim(x,y) = \frac{IC(L_{x,y})}{IC(x)} + \frac{IC(L_{x,y})}{IC(y)} + \frac{IC(L_{x,y})}{IC(H_{x,y})}, \tag{3}$$

where $L_{x,y}$ and $H_{x,y}$ represent the LCA set and HCD set of (x, y), respectively. Considering that GO terms are organized in DAGs instead of trees, i.e. two terms may have more than one common ancestors and descendants, here we consider all LCAs and HCDs. Eq. (3) suggests that high IC(LCA) and low IC(HCD) lead to a high pairwise similarity. This is consistent with the intuition that in order to get a high similarity, the LCAs should be located as low as possible and HCDs should be located as high as possible to be close to the query terms.

In order to use the new metric, we need to estimate the ICs first. Although Section 2.1 gives the definition of IC, it is only applicable to single nodes. The ICs for the sets of LCAs and HCDs should be defined specifically. Instead of using common operations for set computation, e.g. max, min, average, we introduce two special rules, namely intersection ($\cap$) and union ($\cup$), to compute the probabilities for the LCA set and HCD set, i.e.

$$p(L_{x,y}) = p\left(\bigcap_{l \in L_{x,y}} D_l\right) = \sum_{t \in CD_{L_{x,y}}} p^*(t), \tag{4}$$

where D_l is the descendant set of l, and $CD_{L_{x,y}}$ denotes the set of common descendants for the nodes in $L_{x,y}$,

$$p(H_{x,y}) = p\left(\bigcup_{h \in H_{x,y}} D_h\right) = \sum_{t \in D_{H_{x,y}}} p^*(t) = \sum_{h \in H_{x,y}} p(h), \tag{5}$$

where D_h is the descendant set of h. In Eq. (4), the intersection operation takes the common descendants of all LCAs; while in Eq. (5), the union operation merges all the descendants of HCDs. Then the probability of LCA/HCD set is converted to IC values by a negative logarithm transformation.

It is straightforward to justify these two rules. Generally, for a pair of GO terms, the more LCAs/HCDs they share, the more similar they are. According to the intersection rule, more LCAs will lead to a smaller set with higher information content; while by using the union rule, more HCDs lead to a larger set with lower information content. And, high IC of LCAs and low IC of HCDs result in large similarity score, as shown in Eq. (3).

To illustrate these two set operations, we extract a partial GO DAG from GO database as shown in Supplementary Figure S1. Take the pair of GO: 0006091 and GO: 0016310 as an example, they are the LCAs for GO: 0042773 and GO: 0009777. The intersection set of their descendants includes GO: 0006119, GO: 0009777, descendants of GO: 0006119 and descendants of GO: 0009777, while the union set of their descendants consists of all descendants of GO: 0006091 and GO: 0016310. Obviously, the intersection set of LCAs would never be empty, which at least contains the query pair of nodes. However, two GO terms may have no HCD at all. In such case, Eq. (3) will degenerate to the sum of the first two terms.

In addition, Eq. (3) only reflects the relative distances from LCAs to the query nodes and from LCAs to HCDs. We should also consider the depth of the query nodes. Thus the complete equation for measuring GO semantic similarity is defined in Eq. (6),

$$Sim'(x,y) = \left(\frac{IC(L_{x,y})}{IC(x)} + \frac{IC(L_{x,y})}{IC(y)} + \frac{IC(L_{x,y})}{IC(H_{x,y})}\right) \times (IC(x) + IC(y)), \tag{6}$$

where $Sim'(x, y)$ denotes the final similarity score for (x, y).

3.2.2 Similarity between miRNAs

In order to compute functional similarities between miRNAs, we need to find the target genes of miRNAs, and then map these target genes into sets of GO terms. We directly merge all the GO terms of target genes, and get a redundant GO set for each miRNA, i.e. some GO terms may have multiple copies in the set. The common integration strategies described in Section 2.2 can certainly be applied to such redundant set, by treating each copy of the GO terms as an individual member of the set. However, this simple treatment may lose important statistical information. In order to reflect the statistical significance of GO terms in the set, we define a weight for each

GO term based on the cumulative hypergeometric distribution. Specifically, suppose that there are N genes annotated by some GO in the database, and M genes are annotated by the term x, while the miRNA has n target genes, and k of them are annotated by x, i.e. the term x occurs k times in the GO set corresponding to the miRNA, then the weight of x, w_x, is defined as,

$$w_x = -\log p(X \geq k) = -\log\left(1 - \sum_{i=0}^{k-1} \frac{C_M^i \times C_{N-M}^{n-i}}{C_N^n}\right), \quad (7)$$

where C_N^n is the number of combinations of N items taken n at a time.

Besides, in order to predict function-related properties for a query gene, the biggest issue is to find the closest genes to it. Thus, we only need to know the functional similarity scores between the query gene and other known genes. Therefore, instead of generating a symmetric similarity matrix, we develop a method based on Euclidean distance to calculate the similarity scores between a query miRNA and other miRNAs. By incorporating the aforedescribed weights into the Euclidean distance equation, the similarity between the query miRNA m_q and a training miRNA m_t is defined as,

$$Sim_{m_q,m_t} = \sqrt{\sum_{i=1}^{n} \left(Sim(a_i, \mathcal{B}) \times w_{a_i}\right)^2}, \quad (8)$$

where n is the number of non-redundant GO terms of m_q, $\mathcal{B}$ is the GO set of m_t. We call the new integration rule WED, i.e. weighted Euclidean distance.

The miRGOFS method is implemented in C# with the task parallel library (TPL), which enables multiple threads to run in parallel. And the source code is available at https://github.com/yangy09/MiRGOFS.

3.2.3 Inference rule

For a query miRNA, its association with the labels, like diseases or subcellular locations, can be inferred based on its similarities with the miRNAs in the training set. Here, we adopt Eq. (9) (Zhou *et al.*, 2017) to infer the correlation between the query miRNA q and the label d,

$$Cor_{q,d} = \frac{\sum_{j \in \mathcal{I}_{N_d}} sim_{q,j} + \frac{num_d}{num}}{\sum_{i \in \mathcal{I}_N} sim_{q,i} + 1}, \quad (9)$$

where $\mathcal{I}_N$ is the index set of all the nearest neighbors of the query miRNA, $\mathcal{I}_{N_d}$ is the index set of the nearest neighbors which are associated with d, num_d is the number of miRNAs associated with d, and num is the total number of miRNAs in the training set. This is a modified version of commonly used correlation model, e.g. the microRNA-based similarity inference (MBSI) model (Chen and Zhang, 2013). We use only the training samples within the neighborhood of the query miRNA, and add a Bayesian prior, which is equal to the proportion of miRNAs associated with d in the whole training set, because there may be no neighbor associated with d.

Given the correlation values, we use them to draw ROC curve for performance evaluation in the prediction of miRNA–disease association, which has a large number of labels (diseases); while for miRNA subcellular localization, we treat it as a multi-label classification problem, and convert the correlation values into feature vectors as suggested in Zhou *et al.* (2017). Suppose there are a total of k locations, we generate a k-Dim vector for each miRNA, where the elements of the vector are the correlation values to the k locations respectively. Since the three GO categories, BP, MF and CC, yield different similarities for miRNAs, which may be complementary with each other, we generate a k-Dim vector for each GO category and combine them as a $(3 \times k)$-Dim vector.

Table 2. The Pearson's correlation coefficients between expert's scores and computed scores[a]

Method	ZZL	Resnik	Lin	COMBINE	Wang	New Metric
PCC	0.7144	0.8241	0.8496	0.8638	0.8257	**0.8763**

Note: The bold value is the maximum value of the row.

[a]PCCs of Resnik's, Lin's, ZZL's and COMBINE methods were extracted from Li *et al.* (2006). PCC of Wang's method was computed in our experiments.

4 Experimental results

4.1 Performance of the new GO semantic metric

Since miRGOFS is featured by a new IC-based GO semantic metric, we first evaluate the performance of the GO metric. Here, we use a golden-standard dataset including 25 pairs of GO terms (Supplementary Table S1). Their similarities were given by domain experts (Li *et al.*, 2006), i.e. averaged scores from ten biologists. Note that the GO database updates constantly and some GO terms in this set have been obsolete. For a fair comparison, we run our method on the GO database and gene–GO mapping files released on 2005 as used in Li *et al.* (2006), and also implement Wang's method (Wang *et al.*, 2007) on the same version. The similarity scores are shown in Supplementary Table S1. The experts' scores range from 0 (not similar) to 10 (synonymous). We compute Pearson's correlation coefficients (PCCs) between the experts' scores and the computed scores from the new metric and five other methods, as shown in Table 2. Our method has the highest PCC among the six methods, indicating that its scores are most consistent with the experts' ratings. Compared with other IC-based methods, e.g. Resnik's and Lin's, the new method has a significant improvement, which demonstrates the effectiveness of the structural information from descendants.

4.2 MiRNA subcellular localization

4.2.1 Experimental procedure and settings

The experimental procedure consists of the following steps:

1. Correlation scores calculation: using the miRNA pairwise similarity and the labels from the training samples according to Eq. (9);
2. Feature encoding: each miRNA is represented by an 18-D feature vector combining the correlation scores obtained from BP, CC and MF DAGs;
3. 10-fold cross-validation: the classifiers are support vector machines (SVMs) with RBF kernel.

Specifically, in order to search the parameters (C and γ) for SVMs, we perform a nested 5-fold cross-validation. We also repeat the 10-fold cross-validation for 10 times, and get the averaged accuracies of the 100 tests.

4.2.2 Comparison of different strategies to generate miRNA functional similarity

To our knowledge, there has been no predictor specialized for miRNA localization. Moreover, the other methods for measuring miRNA similarity cover a small proportion of the miRNAs in the benchmark set, e.g. over 50% miRNAs are not covered in Yu's method. Therefore, here we compare 3 methods implemented by

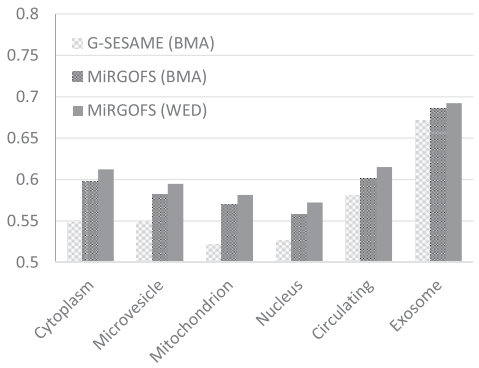

Fig. 2. F_1 scores of different types of miRNA functional similarity on 6 locations

ourselves, which adopt different GO similarity metric or combination rules. Details are given below.

- G-SESAME (BMA): the G-SESAME method for calculating GO semantic similarity (Wang *et al.*, 2007) with the best match average rule to combine the GO similarities.
- MiRGOFS (BMA): the proposed new GO semantic similarity metric with the best match average rule to combine the GO similarities.
- MiRGOFS (WED): the proposed new GO semantic similarity metric with the weighted Euclidean distance rule to combine the GO similarities.

All of these three methods utilize the same target genes and GO annotation information, where the target genes are downloaded from `microrna.org` and `mirdb.org`, and the GO annotation for the target genes are from the NCBI gene2go database (released July, 2017) downloaded from ftp.ncbi.nlm.nih.gov.

Figure 2 shows the F_1 scores of the three methods on each of the 6 subcellular locations. As can be seen, miRGOFS(WED) performs the best and miRGOFS(BMA) ranks the second on all 6 locations. Especially, for cytoplasm and mitochondrion, which have the fewest training samples, miRGOFS(WED) improves the F_1 over 5% compared with G-SESAME(BMA). Apparently, using the same combination rule, the proposed GO semantic is superior to the G-SESAME method with an increase of 1.5–5% on F_1; while using the same GO semantic similarity, the proposed weighted Euclidean distance (WED), is better than the best match average (BMA), and generally increases F_1 by over 1%.

As an overall evaluation, Table 3 shows the averaged *ACC*, F_1 and *AUC* for 10 times of 10-fold cross-validation (Results for each 10-fold cross-validation are shown in Supplementary Fig. S2), where the *ACC* and F_1 are customized for multi-label classification (Briesemeister *et al.*, 2010), as the prediction task is a typical multi-label classification problem (see Supplementary Equations S4–S8). We can observe that:

i. Both the two miRGOFS methods have obvious advantages over G-SESAME(BMA), based on either target source. Driven by the new GO metric, miRGOFS(BMA) increases *AUC* by 3.4% compared against G-SESAME(BMA) using the targets from microRNA.org;
ii. Compared to the commonly used BMA integration rule, the WED strategy can further improve the performance. Generally, the *ACC* and *AUC* increase by around 1%, while the increase for F_1 is not significant;
iii. The mirDB target source leads to better performance for all the methods compared to microRNA.org/miRanda. Especially, the graph-based method benefits a lot from the high-quality targets.

In order to further investigate the significance of the achieved predictive performance, we design two randomized predictors, using different label shuffling strategies. Let L be an $(m \times 6)$-D label matrix, where m is the number of training miRNAs and 6 is the total number of different subcellular compartments. $L_{i,j}$ is a boolean value, indicating whether or not the ith miRNA has the jth label. In the first randomized predictor, we shuffle the order of rows, i.e. the correspondence between miRNAs and their label sets is shuffled; while in the second randomized predictor, we shuffle the elements respectively for each column. Apparently, in either case, the numbers of labels and label distributions are the same as the original dataset. With the same feature vectors generated by miRGOFS (WED), the accuracies obtained by randomized predictors are 6–12% lower than the predictor using right labels. The first randomized predictor is slightly better than the second one, because the original combinations of labels remain unchanged in the first one but are totally changed in the second one. The great gap between original predictor and randomized predictors demonstrates the strong association between GO information and miRNA subcellular localization. Note that the performance of randomized predictors is actually not too bad, because many miRNAs have overlapped labels, and there are hidden correlations between the subcellular compartments.

These experimental results demonstrate the potential of predicting miRNA subcellular locations by using GO information from target genes, and the effectiveness of miRGOFS-driven features in designing computational predictors.

4.3 Prediction of miRNA–disease associations

For the past decade, a lot of computational methods for identifying miRNA–disease association have been proposed, we compare the performance of miRGOFS against 9 state-of-the-art methods using two datasets, Data1 and Data2, as described in Section 3.1.3. Most of the current predictors for miRNA–disease association rely on both miRNA–miRNA functional similarity and disease-disease semantic similarity, while we only use the miRNA–miRNA functional similarity and treat each disease as a separate label, as we mainly focus on developing new functional similarity metric for miRNAs. For Yu's method, we download its similarity matrix (Yu *et al.*, 2011) and use the same inference procedure as in miRGOFS. There are three major differences between these two methods:

1. Yu *et al.* adopted the algorithm of PITA (Probability of Interaction by Target Accessibility) (Kertesz *et al.*, 2007) to predict target genes, while miRGOFS uses the miRanda algorithm from the microRNA.org resource (Betel *et al.*, 2007; John *et al.*, 2004);
2. Yu *et al.* adopted BMA twice for integrating GO similarity scores, while miRGOFS uses the combined GO set and WED method described in Section 3.2.2;
3. Yu *et al.* used the GO similarity metric proposed by Wang *et al.* (2007), while miRGOFS used the new IC-based metric.

Note that Yu's similarity matrix misses 37 and 11 miRNAs in Data1 and Data2, respectively, thus here we only compare with Yu's method on Data2. The AUC values of ROC curves are used as the evaluation criterion, as shown in Table 4.

Table 4 shows not only the AUC values, but also the prior knowledge that is used to infer the functional similarities of miRNAs. Most of the methods utilize MISIM and obtain fairly well prediction results. MISIM was generated by Wang *et al.* (2010) based on HMDD V1, which covers 1395 miRNA–disease associations, i.e.

Table 3. Comparison of different types of miRNA functional similarity[a,b]

Metric	Target source	Method comparison			Shuffled label	
		G-SESAME(BMA)	miRGOFS(BMA)	miRGOFS(WED)	miRGOFS_Rand1	miRGOFS_Rand2
ACC	miRanda	0.442±0.029 •	0.461±0.032•	**0.471±0.028**	0.415±0.034•	0.399±0.027•
	mirDB	0.464±0.030 •	0.472±0.037•	**0.481±0.036**	0.403±0.027•	0.399±0.027•
F_1	miRanda	0.568±0.035 •	0.593±0.035°	**0.597±0.033**	0.536±0.036•	0.529±0.027•
	mirDB	0.612±0.033 °	0.612±0.040°	0.612±0.041	0.526±0.028•	0.532±0.027•
AUC	miRanda	0.557±0.035 •	0.591±0.038°	**0.602±0.039**	0.496±0.024•	0.496±0.032•
	mirDB	0.600±0.035 •	0.605±0.039•	**0.622±0.035**	0.501±0.032•	0.503±0.032•

Note: The significance of bold is the maximum value of each row.

[a]•/° indicates the performance difference between the method and miRGOFS (WED) is/is not statistically significant according to pairwise *t*-test at 95% significance level.

[b]miRGOFS_Rand1 and miRGOFS_Rand2 denote the first randomized predictor (shuffling rows in the label matrix) and the second randomized predictor (shuffling elements of each column in the label matrix) using the feature vectors generated by miRGOFS(WED), respectively.

Table 4. Comparison of disease-related miRNA prediction methods[a,b]

Data	Method	Prior Knowledge	AUC
Data1	RWRMDA	MISIM	0.862
	RLSMDA	MISIM, MeSH[c]	0.845
	HD[d]	MISIM	0.778
	miRGOFS	Target, GO	**0.877**
Data2	Yu's	Target, GO	0.762
	Jiang's	Target	0.758
	NetCBI	MISIM, MimMiner[e]	0.807
	MBSI	MISIM	0.748
	PBSI	MimMiner[e]	0.542
	MiRGOFS	Target, GO	**0.811**

Note: The two bold numbers are the maximum values of the last column for Data1 and Data2, respectively.

[a]AUCs of RWRMDA and HD are from Chen *et al.* (2012), AUC of RLSMDA is from RLSMDA (Chen and Yan, 2015), AUC of Jiang's method is from Jiang *et al.* (2010) and AUCs of NetCBI, MBSI and PBSI are from Chen and Zhang (2013).

[b]All of the AUC values are evaluated via the leave-one-out cross-validation, where Jiang's method and PBSI take each association as the test sample while other methods take a miRNA as the test sample.

[c]The relationship between diseases are from MeSH database.

[d]HD: a hypergeometric distribution-based method (Jiang *et al.*, 2010), and Chen *et al.* (2012) implemented this method by using MISIM in their paper.

[e]The disease phenotype similarity scores are from MimMiner (Van Driel *et al.*, 2006).

the Data1. It contains pairwise functional similarities for 271 miRNAs according to the miRNA–disease associations and semantic correlation of diseases, and has been used as a standard similarity metric and basic information source in many miRNA–disease association studies. By contrast, Jiang's, Yu's and miRGOFS use prior knowledge which is completely independent with the miRNA–disease relationship. Jiang *et al.* (2010) constructed a functionally related miRNA network based on the predicted targets by PITA (Kertesz *et al.*, 2007) and TargetScan (Lewis *et al.*, 2003), where a pair of miRNAs are connected if they share a significant number of targets. Yu's method and miRGOFS use the target information and target's GO annotation, which are easy to access. Interestingly, the GO-based methods achieve competitive and even better results than the MISIM-based methods. Especially, on Data1, miRGOFS's AUC value is 1.5% higher than the best AUC of other methods; while on Data2, miRGOFS is slightly better than NetCBI, which uses both miRNA–miRNA functional similarity network and disease-disease semantic similarity network. PBSI only uses the semantic similarity scores between disease phenotypes, and gets the worst accuracy.

The reasons for the advantage of miRGOFS over MISIM-based methods are manifold. First, the GO annotation of target genes can well represent miRNAs' functional features, and serve as a good knowledge source for predicting miRNA–disease associations. Second, although MISIM is built on a known miRNA–disease network, this network contains a lot of missing and error edges, which may influence the calculation of miRNA similarity. Third, our inference rule also contributes to the performance enhancement. Especially, in order to assess the impact of neighbors and the prior term in Eq. (9), we record the AUC values obtained with 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50 neighbors, respectively, and compare the three methods introduced in Section 4.2.2, i.e. G-SESAME (BMA), MiRGOFS (BMA), MiRGOFS (WED). The results (Supplementary Fig. S3) show that MiRGOFS (WED) performs the best no matter how many nearest neighbors are considered. Furthermore, the smoothing technique plays a crucial role when the number of neighbors is small, and its contribution to AUC value becomes stable as the number of neighbors increases. Generally, the smoothing technique brings 1–3% improvement on the two datasets. The results are as expected, i.e. the adjustment by using Bayesian prior is much more necessary when the sample set is sparse. As for the prediction of miRNA–disease association, the label set is large compared with the sample size, thus the data is very sparse, especially in a small neighborhood. By contrast, in the prediction of protein subcellular localization, we only have 6 labels and the data distribution is relatively balanced, thus the effect of smoothing technique is trivial.

4.4 Comparison with other GO-based miRNA similarities

The comparison with other GO-based miRNA similarities could be either conducted on the existing miRNA similarity matrices, or the generated miRNA similarity by integrating GO similarities provided by other studies. However, there are two major obstacles for the comparison:

1. The existing miRNA similarity matrices generally have a small coverage of miRNAs. For instance, Yu *et al.* (2011)'s study covers 533 human miRNAs, and Lan *et al.* (2016)'s study covers 289 miRNAs. Thus, as for the subcellular localization dataset, none of the released miRNA similarity scores could be compared due to the large missing ratio.
2. The current tools for computing GO semantic similarity [e.g. the GOSemSim package (Yu *et al.*, 2010)], as designed for handling

Table 5. Comparison of different types of miRNA functional similarity[a]

Method	AUC		CPU time/pair (seconds)[b]
	Data I	Data II	
Resnik	0.860	0.805	7.3
Lin	0.859	0.800	7.4
Jiang	0.860	0.807	7.3
Rel	0.858	**0.812**	8.9
Wang	0.868	0.804	47.4
miRGOFS	**0.877**	0.811	**0.005**

Note: The significance of bold is the maximum value of each column.

[a]For all the methods, the targets are from mirDB, and the GO semantic similarities are computed based on MF terms.

[b]The average CPU time (seconds) for computing the similarity for a pair of miRNAs.

single GO pairs or two small GO sets, work inefficiently for computing the pairwise similarities between miRNAs which often correspond to very large GO sets.

Here we compare miRGOFS with the miRNA similarities yielded by the R package GOSemSim using various IC-based and graph-based algorithms for the prediction of miRNA–disease association (this task has small datasets, thus allows the comparison with other tools). The results are shown in Table 5. For Data I, miRGOFS has the highest AUC, and the graph-based method (Wang's) is better than the IC-based methods in GOSemSim; while for Data II, all the methods obtain very close performance. The most significant advantage of the new method is the computation efficiency. For each pair of miRNAs, miRGOFS is thousands of times faster than the methods implemented in GOSemSim.

5 Discussion

5.1 The data source of GO

In this paper, we infer functional similarity between miRNAs based on GO features. The gene ontology database consists of three separate hierarchies, BP, MF and CC, which yield different functional similarities for each pair of miRNAs. Especially, the Cellular Component (CC) terms directly represent the subcellular localization of genes, thus they contribute most in the protein subcellular localization. For miRNA subcellular localization, in previous experiments, we report the results using the combined feature vectors, i.e. the 18-D feature vectors. We also assess the prediction performance using the 6-D feature vectors generated from BP, CC and MF, respectively (shown in Supplementary Table S2). We find that combining the features extracted from the three DAGs successfully improve the accuracies. BP and MF have close performance, while CC has no superiority over other two categories in the prediction of miRNA subcellular localization. It may be due to two reasons: (i) the CC terms represent the cellular localization of target genes rather than that of miRNAs; (ii) the size of CC DAG is smaller and the CC annotation is relatively sparse compared with BP and MF. Thus, it is not capable to provide sufficient information in the prediction. In the prediction of miRNA–disease associations, since the computed miRNA–disease correlation scores directly yield the AUC value, the similarities obtained by three DAGs are not combined. BP and MF also have close performance, and CC performs the worst.

Furthermore, we investigate the enriched GO terms in the associated GO sets of miRNAs. As mentioned in Section 2.2, for each miRNA, we combine the GO sets of all its target genes and get a redundant GO set associated with miRNA. Apparently, the combined GO set has a large size. On the one hand, as the GO database keeps expanding, both the coverage of annotated genes and the numbers of GO terms increase rapidly; on the other hand, each miRNA has a lot of target genes as we use the computationally predicted targets. Specifically, in this study, the gene ontology database (go-basic.obo, release 08/20/2016) contains 45 217 GO terms, the miRNAs have over 2000 target genes and over 7000 non-redundant GO terms on average, which leads to a high computational cost in the computation of miRNA pair-wise similarities. Actually, many of the computed targets are not real targets of the miRNAs and most of the GO terms in the collection may be useless. Therefore, we only use significant GO terms (P-value < 0.05) in the experiments, but we also compare against the performance of using all GO terms. Take the MF DAG as an example, the enriched GO terms account for ~30% on average of all the GO terms (456/1545), while their accuracies are very close. Therefore, the screening of GO terms improves the computation efficiency significantly.

5.2 The data source of targets

In Section 3.1.1, we introduce the two target sources used in this study, i.e. microRNA.org and mirDB. Different settings on the stringency lead to different target genes. A loose setting may result in too many false positive targets, while a stringent criterion may lead to the lack of GO annotations and low coverage of miRNAs. MicroRNA.org set a quite loose threshold of the evidence value for target prediction, thus all the 1100 miRNAs have corresponding GO terms. Despite the large number of target genes and numerous GO terms, we only focus on the statistically significant GO terms, so the false targets have a small impact on the calculation of similarity. By contrast, mirDB adopts a relatively stringent setting for predicted targets. According to the experimental results (Table 3), mirDB leads to an enhanced performance. However, the much reduced number of target genes may result in sparse GO annotations for miRNAs. For instance, in the similarity matrix generated by using mirDB, hsa-miR-126-3p and hsa-miR-1307-3p have no GO term from the MF category, and hsa-miR-1469 has no GO term from the BP category.

6 Conclusion

Domain knowledge-based functional similarities of miRNAs can help improve the analysis quality of miRNA high-throughput expression data, and predict unknown functional properties of miRNAs as well as miRNA–disease associations. However, due to the lack of functional annotation of miRNAs in public databases, it is not straightforward to utilize current knowledge sources, like gene ontology, to infer miRNA functional similarity. In this paper, we propose a new method, called miRGOFS to convert GO information annotated for the target genes of miRNAs, to infer miRNA functional similarity. The new method consists of two major components, the calculation of GO semantic similarity and the integration of GO set similarity. Each of them is implemented with a newly proposed algorithm, namely the IC-based semantic metric and the weighted Euclidean distance based integration rule (WED). The advantage of the new GO metric lies in a comprehensive utilization of all common ancestors and descendants of GO terms, while the WED algorithm assigns weights to GO terms and computes the similarity between GO sets in an asymmetric manner, which allows us to focus on the similarities between the query miRNA and other known miRNAs. Supported by these two new algorithms, miRGOFS has

shown promising performance for the prediction of miRNA subcellular localization and miRNA–disease associations. Featured by sufficient coverage of human miRNAs and ease of getting the supporting knowledge source (target genes can be computationally identified and their GO terms are available on GO database), miRGOFS would have wide applicability in miRNA functional analysis.

Funding

This work was supported by the National Natural Science Foundation of China (No. 61725302, 61671288, 91530321, 61603161), the Science and Technology Commission of Shanghai Municipality (No. 16ZR1448700, 16JC1404300, 17JC1403500), and the Fundamental Research Funds for the Central Universities.

Conflict of Interest: none declared.

References

Abba,M.L. *et al*. (2017) MicroRNAs as novel targets and tools in cancer therapy. *Cancer Lett*., **387**, 84–94.

Agarwal,V. *et al*. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, **4:e05005**, 1–38.

Ashburner,M. *et al*. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet*., **25**, 25–29.

Azuaje,F. *et al*. (2005). Ontology-driven similarity approaches to supporting gene functional assessment. In: *Proceedings of the ISMB 2005 SIG meeting on Bio-ontologies*, pp. 9–10.

Bender,M.A. *et al*. (2005) Lowest common ancestors in trees and directed acyclic graphs. *J. Algorithms*, **57**, 75–94.

Bentwich,I. *et al*. (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet*., **37**, 766–770.

Betel,D. *et al*. (2007) The microrna.org resource: targets and expression. *Nucleic Acids Res*., **36**, D149–D153.

Bleazard,T. *et al*. (2015) Bias in microRNA functional enrichment analysis. *Bioinformatics*, **31**, 1592–1598.

Briesemeister,S. *et al*. (2010) Yloc: an interpretable web server for predicting subcellular localization. *Nucleic Acids Res*., **38**, W497–W502.

Chen,H. and Zhang,Z. (2013) Similarity-based methods for potential human microRNA–disease association prediction. *BMC Med. Genomics*, **6**, 1.

Chen,X. and Yan,G. (2015) Semi-supervised learning for potential human microRNA–disease associations inference. *Sci. Rep*., **4**, 5501–5501.

Chen,X. *et al*. (2012) RWRMDA: predicting novel human microRNA–disease associations. *Mol. BioSystems*, **8**, 2792–2798.

Chen,X. *et al*. (2016a) Hgimda: heterogeneous graph inference for miRNA–disease association prediction. *Oncotarget*, **7**, 65257–65269.

Chen,X. *et al*. (2016b) Wbsmda: within and between score for miRNA–disease association prediction. *Sci. Rep*., **6**, 21106.

Chen,X. *et al*. (2017a) HAMDA: hybrid approach for miRNA–disease association prediction. *J. Biomed. Inf*., **76**, 50–58.

Chen,X. *et al*. (2017b) MicroRNAs and complex diseases: from experimental results to computational models. *Brief. Bioinf*., doi: 10.1093/bib/bbx130.

Chen,X. *et al*. (2017c) Rknnmda: ranking-based knn for miRNA–disease association prediction. *RNA Biol*., **14**, 952–962.

Chen,X. *et al*. (2018) DRMDA: deep representations-based miRNA–disease association prediction. *J. Cell. Mol. Med*., **22**, 472–485.

Couto,F.M. *et al*. (2007) Measuring semantic similarity between gene ontology terms. *Data Knowl. Eng*., **61**, 137–152.

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

Fan,X. and Kurgan,L. (2015) Comprehensive overview and assessment of computational prediction of microRNA targets in animals. *Brief. Bioinf*., **16**, 780–794.

Griffiths-Jones,S. *et al*. (2006) mirbase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*., **34**, D140–D144.

Gusev,Y. *et al*. (2007) Computational analysis of biological functions and pathways collectively targeted by co-expressed microRNAs in cancer. *BMC Bioinformatics*, **8**, S16–S17.

He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet*., **5**, 522–531.

Huang,D. and Pan,W. (2006) Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, **22**, 1259–1268.

Jiang,J.J. and Conrath,D.W. (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference Research on Computational Linguistics (ROCLING X)*, 1997, Taiwan, pp. 1–15.

Jiang,Q. *et al*. (2009) mir2disease: a manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res*., **37**, D98–D104.

Jiang,Q. *et al*. (2010) Prioritization of disease microRNAs through a human phenome-micrornaome network. *BMC Syst. Biol*., **4**, S2.

John,B. *et al*. (2004) Human microRNA targets. *PLoS Biol*., **2**, e363.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*., **28**, 27–30.

Kertesz,M. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nat. Genet*., **39**, 1278–1284.

Krek,A. *et al*. (2005) Combinatorial microRNA target predictions. *Nat. Genet*., **37**, 495–500.

Lan,C. *et al*. (2016) Grouping miRNAs of similar functions via weighted information content of gene ontology. *BMC Bioinformatics*, **17**, 507.

Lee,Y. *et al*. (2002) MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*., **21**, 4663–4670.

Leung,A. and Sharp,P. (2006) Function and localization of microRNAs in mammalian cells. In: *Cold Spring Harbor Symposia on Quantitative Biology*, **vol. 71**. Cold Spring Harbor Laboratory Press, pp. 29–38.

Leung,A.K.L. (2015) The whereabouts of microRNA actions: cytoplasm and beyond. *Trends Cell Biol*., **25**, 601–610.

Lewis,B.P. *et al*. (2003) Prediction of mammalian microRNA targets. *Cell*, **115**, 787–798.

Li,R. *et al*. (2006) A measure of semantic similarity between gene ontology terms based on semantic pathway covering. *Progress Nat. Sci*., **16**, 721–726.

Li,Y. *et al*. (2013) Hmdd v2. 0: a database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res*., **42**, D1070–D1074.

Lin,D. (1998) An information-theoretic definition of similarity. *ICML*, **98**, 296–304.

Lord,P. *et al*. (2002) Semantic similarity measures as tools for exploring the gene ontology. *Pacific symposium on biocomputing*, **2002**, 601–612.

Lu,J. *et al*. (2005) MicroRNA expression profiles classify human cancers. *Nature*, **435**, 834–838.

Mahdavi,M.A. and Lin,Y.-H. (2007) False positive reduction in protein–protein interaction predictions using gene ontology annotations. *BMC Bioinformatics*, **8**, 262.

Pesquita,C. *et al*. (2008) Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, **9**, S4–16.

Peterson,S.M. *et al*. (2014) Common features of microRNA target prediction tools. *Front. Genet*., **5**, 23.

Reczko,M. *et al*. (2012) Functional microRNA targets in protein coding sequences. *Bioinformatics*, **28**, 771–776.

Resnik,P. *et al*. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res. (JAIR)*, **11**, 95–130.

Schlicker,A. *et al*. (2006) A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, **7**, 302.

Schlicker,A. *et al*. (2010) Improving disease gene prioritization using the semantic similarity of gene ontology terms. *Bioinformatics*, **26**, i561–i567.

Teng,Z. *et al*. (2013) Measuring gene functional similarity based on group-wise comparison of go terms. *Bioinformatics*, **29**, 1424–1432.

Thomson,J.M. *et al*. (2004) A custom microarray platform for analysis of microRNA gene expression. *Nat. Methods*, **1**, 47–53.

Van Driel,M.A. *et al*. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet*., **14**, 535–542.

Vlachos,I.S. *et al*. (2015) Diana-tarbase v7. 0: indexing more than half a million experimentally supported miRNA: mRNA interactions. *Nucleic Acids Res*., **43**, D153–D159.

Vlachos,I.S. *et al*. (2012) Diana mirpath v.2.0: investigating the combinatorial effect of microRNAs in pathways. *Nucleic Acids Res*., **40**, W498–W504.

Wang,D. *et al*. (2010) Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics*, **26**, 1644–1650.

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

Wong,N. and Wang,X. (2015) mirdb: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res*., **43**, D146–D152.

Wu,H. *et al*. (2005) Prediction of functional modules based on comparative genome analysis and gene ontology application. *Nucleic Acids Res*., **33**, 2822–2837.

Xu,Y. *et al*. (2013) A novel insight into gene ontology semantic similarity. *Genomics*, **101**, 368–375.

Yang,H. *et al*. (2012) Improving go semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. *Bioinformatics*, **28**, 1383–1389.

You,Z.-H. *et al*. (2017) Pbmda: a novel and effective path-based computational model for miRNA–disease association prediction. *PLoS Comput. Biol*., **13**, e1005455.

Yu,G. *et al*. (2010) Gosemsim: an r package for measuring semantic similarity among go terms and gene products. *Bioinformatics*, **26**, 976–978.

Yu,G. *et al*. (2011) A new method for measuring functional similarity of microRNAs. *J. Integr. Omics*, **1**, 49–54.

Zeng,X. *et al*. (2016) Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinf*., **17**, 193–203.

Zhang,T. *et al*. (2016) RNAlocate: a resource for RNA subcellular localizations. *Nucleic Acids Res*., **45**(D1), D135–D138.

Zhou,H. *et al*. (2017) Hum-mploc 3.0: prediction enhancement of human protein subcellular localization through modeling the hidden correlations of gene ontology and functional domain features. *Bioinformatics*, **33**, 843–853.