The Author(s) *BMC Bioinformatics* 2016, **17**(Suppl 19):507
DOI 10.1186/s12859-016-1367-0

BMC Bioinformatics

RESEARCH Open Access

Grouping miRNAs of similar functions via weighted information content of gene ontology

Chaowang Lan[1,2], Qingfeng Chen[1*] and Jinyan Li[2*]

From 15th International Conference On Bioinformatics (INCOB 2016)
Queenstown, Singapore. 21-23 September 2016

Abstract

Background: Regulation mechanisms between miRNAs and genes are complicated. To accomplish a biological function, a miRNA may regulate multiple target genes, and similarly a target gene may be regulated by multiple miRNAs. Wet-lab knowledge of co-regulating miRNAs is limited. This work introduces a computational method to group miRNAs of similar functions to identify co-regulating miRNAsfrom a similarity matrix of miRNAs.

Results: We define a novel information content of gene ontology (GO) to measure similarity between two sets of GO graphs corresponding to the two sets of target genes of two miRNAs. This between-graph similarity is then transferred as a functional similarity between the two miRNAs. Our definition of the information content is based on the size of a GO term's descendants, but adjusted by a weight derived from its depth level and the GO relationships at its path to the root node or to the most informative common ancestor (MICA). Further, a self-tuning technique and the eigenvalues of the normalized Laplacian matrix are applied to determine the optimal parameters for the spectral clustering of the similarity matrix of the miRNAs.

Conclusions: Experimental results demonstrate that our method has better clustering performance than the existing edge-based, node-based or hybrid methods. Our method has also demonstrated a novel usefulness for the function annotation of new miRNAs, as reported in the detailed case studies.

Keywords: Gene ontology, Functions of miRNAs, Information content, GO graphs, Spectral clustering

Background

MiRNA is a small non-coding RNA molecule highly conserved in plants and animals. Many investigations have reported that miRNAs can play important roles in various vital biological processes such as gene expression, cell development, cancer progression, and immune process by binding to the 3′ untranslated regions of their target genes, which can result in the translational repression or rapid degradation of the target transcripts [1]. As miRNA function is usually carried out by groups of miRNAs rather than individually [2], clustering miRNAs for the function annotation of new miRNAs is a problem of wide interests, given that the knowledge of co-regulating miRNAs is limited in wet-labs.

Sequence or structure-based similarity measurements have been previously proposed to cluster miRNAs for similar functions. For example, the Rfam [3] and miRBase [4] databases use sequence similarities to classify the functions of miRNAs. The concern is that some miRNAs having a high sequence similarity may have distinct functions. Also, the structure-function relationships used in the function annotation of miRNAs have been reported to show serious limitations in the case of complex substructures [5].

*Correspondence: qingfeng@gxu.edu.cn; jinyan.li@uts.edu.au
[1] School of Computer, Electronic and Information, and State Key Laboratory for Conservation and Utilization of Subtropical Agro-bioresources, Guangxi University, No.100 Daxue Road, 530004 Nanning, China
[2] Advanced Analytics Institute, Faculty of Engineering and IT, University of Technology Sydney, PO Box 123, Broadway, Sydney, NSW 2007, Australia

Recently, individual target genes of differentially expressed miRNAs have been explored for clustering miRNAs into groups of similar functions. However, a miRNA can regulate multiple target genes. To overcome this limitation, we explore a novel similarity measurement between the two sets of target genes corresponding to two miRNAs. We propose to transfer the function similarity between the two sets of target genes as the function similarity of the two miRNAs.

The function similarity between two sets of target genes has been previously investigated and can be derived from the structure information of gene ontology (GO) trees of these target genes [6]. The hierarchical structure of a GO tree is a directed acyclic graph (DAC), containing structured vocabularies to describe the functions at different levels of the gene products [7]. The nodes of a GO tree are called *terms*. An *edge* in a GO tree represents a relationship between two terms. The two most common relationships between two terms are *is_a* for subclass and *part_of* for component [8]. GO terms, their relationships and the similarity between two GO trees have been considered in many bioinformatics applications by literature such as for pathway analysis [9], gene network analysis [10], and gene expression research [11].

We introduce a novel measurement of information content, a weighted information content of gene ontology, to estimate the similarity between two GO trees. The weighted information content of a term in a GO tree is determined by three factors: the number of descendants of the term, the depth of the path from the term to the root node or to the most informative common ancestor (MICA), and the relationships along the edges in the path. Every term in a GO tree has its unique information content. Based on this definition of information content, the similarity between two GO trees is proposed to be measured by the information contents of all the common terms between the two GO trees, relative to the information contents of all the unique terms. Two GO trees are more similar in function than others if they have more common terms and fewer unique terms. When we are given two sets of GO trees, the similarity between the two sets are derived by computing all the pair-wise similarities of the GO trees from the two sets. This similarity between the two sets of GO trees is then transferred as a similarity measurement between the two miRNAs whose target genes correspond to the two set of GO trees.

In the literature, node-based [12] and edge-based methods [13] have been proposed to measure the similarity between GO trees or subtrees. By their definitions, the nodes in the same hierarchy are assumed to have an equal distance to the root, an idea which was criticized by [14]. Further, the information content of a term in a GO graph is exactly the same as another's, even if the two terms have different depths in the graph [15]—it ignores important properties of edges such as the depth and the topology information of the term in the GO graph. Node-based methods also focus on the most informative common ancestor like our method, but they neglect the whole path structures of GO terms. Moreover, the edge-based methods do not distinguish the weight of terms at different depths of a GO graph. Our weighted information content of gene ontology can overcome these shortcomings.

For enhancing the performance on clustering the miRNAs into subgroups of similar functions, a self-tuning technique is applied to determine the optimal parameter σ for the spectral clustering method [16]. Further, an appropriate cluster number is estimated by the eigenvalues of the normalized Laplacian matrix. Our approach has been used for grouping miRNAs of similar functions associated with diseases stored at several databases. Most of the experimental results showed good accuracy and the annotation results for new miRNAs can be supported by evidence found from the other databases or from recently published literature.

Methods

MiRNAs and their target genes were downloaded from http://mirtarbase.mbc.nctu.edu.tw/php/download.php (the file hsa_MTI.xls). This file is a relational table having 39110 lines and 9 columns: miRTarBase ID, miRNA, Species (miRNA), Target Gene, Target Gene (Entrez ID), Species (Target Gene), Experiments, Support Type, and References (PMID). Each line of this table stores information of one miRNA and the information of one of its target genes. We note that some multiple lines in this table actually refer to the same miRNA—researchers have done many different experiments to confirm the same miRNA's target genes. As a result, there are only 289 distinct miRNAs in this file. We used all of them in this work. The disease information associated with each of the miRNAs was searched at the HMDD database (http://www.cuilab.cn/hmdd). Out of the 289 miRNAs, 24 did not have disease information available.

The GO terms of a target gene were searched at the EMBL-EBI website (http://www.ebi.ac.uk/). The relationships (i.e., is_a and part_of) of these GO terms were derived from the AmiGo database (http://amigo.geneontology.org/amigo). These GO terms and their relationships were integrated and represented by graphs.

Definition 1. ***GO graph of a gene****. Given a gene g, its GO terms and the relationships of these GO terms are represented by a DAC (direct acyclic) graph* $G(g) = (Term^g, Edge^g)$*, where* $Term^g$ *represents the set of nodes each labeled with a GO term, and* $Edge^g$ *represents the set of edges each labeled with a relationship (is_a or part_of)* *between a pair of terms of g. Such a graph is also called a GO graph or GO tree of g.*

Definition 2. ***Root node****. The root node of a GO graph is the term node which has an in-degree only. A GO graph has one and only one root node.*

Definition 3. ***Leaf nodes****. A leaf node of a GO graph is a term node which has an out-degree only. The GO graph of a gene may have multiple leaf nodes.*

Figure 1a, b and Fig. 2 show three examples of GO graphs, where root nodes, leaf nodes, and the relationships of some pairs of terms are explained.

Definition 4. ***Term graph****. A term graph is a special form of GO graph. Given a GO graph, if it has only one leaf term A, such a GO graph is called A's term graph, denoted by* $TG_A = (A, Term_A, Edge_A)$*, where* $Term_A$ *and* $Edge_A$ *represent the set of GO terms and the set of edges of the GO graph, respectively.*

Given a GO graph $G = (Term, Edge)$, for every term $t \in Term$, we can construct one term graph $TG_{(t,G)} = \left(t, Term_{(t,G)}, Edge_{(t,G)}\right)$, where $Term_{(t,G)}$ is the set of terms in the path from t to the root node of G, and $Edge_{(t,G)})$ is the set of edges in the path from t to the root node of G.

In particular, a leaf node l_node of GO graph G can form a leaf term graph $TG_{(l\_node,G)} = (l\_node, Term_{(l\_node,G)}, Edge_{(l\_node,G)})$. Leaf term graphs of a GO graph are used later to define the similarity between two GO graphs. The subscript G is sometimes omitted when it is understood. Figure 1a and b are actually the two leaf term graphs of Fig. 2.

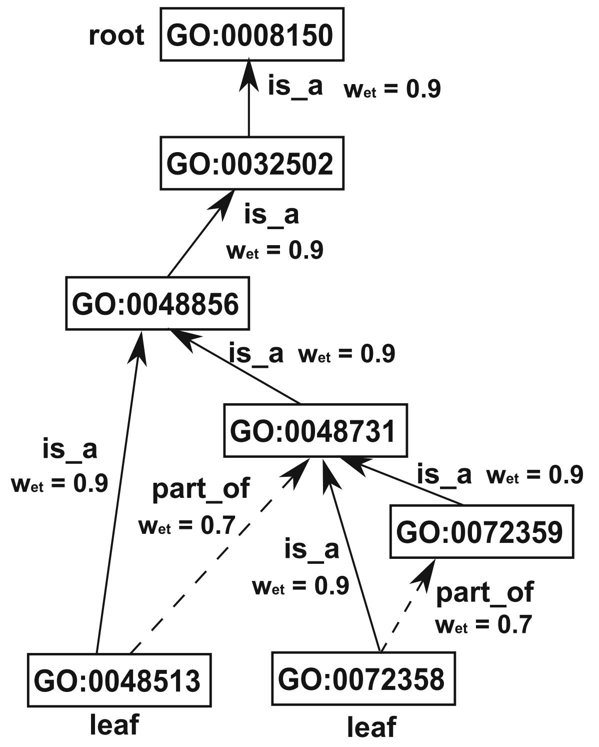

Fig. 2 A GO tree integrating graphs of GO:0048513 and GO:0072358

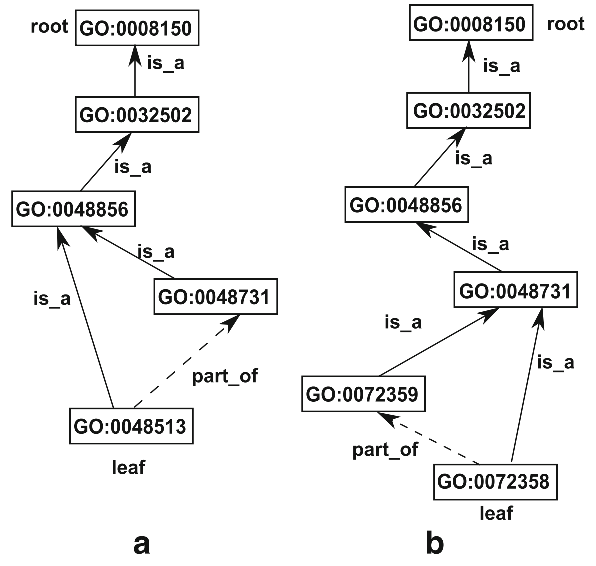

Fig. 1 **a** The GO graph of GO:0048513 and **b** The GO graph of GO:0072358. Each includes one leaf term only

Definition 5. ***Depth and level of a node****. The depth of a term node t in a GO graph is the number of edges in the longest path from t to the root node of the graph. For example, the depth of 0048513 is 4 as shown in Fig.* 1*a. If the depth of a term is d, the term is also said to be at level d.*

Given two term graphs $TG_A = (A, Term_A, Edge_A)$ and $TG_B = (B, Term_B, Edge_B)$. There may exist many common terms (at least the root node) between $Term_A$ and $Term_B$. For example, term 0048856 is a common term between $TG_{0048513}$ (Fig. 1a) and $TG_{0072358}$ (Fig. 1b). For all other terms in $Term_A$ or $Term_B$, they are called *uncommon* or unique terms.

Clustering miRNAs for similar functions

Suppose we are given h number of miRNAs, the first process of our method is to construct a $h \times h$ similarity matrix of these miRNAs. For every pair of miRNAs in the matrix, their similarity is transferred from the similarity between their two sets of target genes. As every gene can be represented by a GO tree, the similarity between the two sets of target genes can be determined by computing the similarity between the two sets of GO trees. With this $h \times h$ similarity matrix as input, we use a spectral clustering method to group miRNAs of similar functions. We present details for these steps:

1. Compute the weighted information content of every term in a GO graph to determine the similarity between two GO trees;
2. Compute the similarity between two sets of GO trees to determine the similarity between two miRNAs;
3. Construct a similarity matrix of the h miRNAs, and subgroup them for a similar function in each group using the similarity matrix as input.

The framework of our method is showed in Fig. 3.

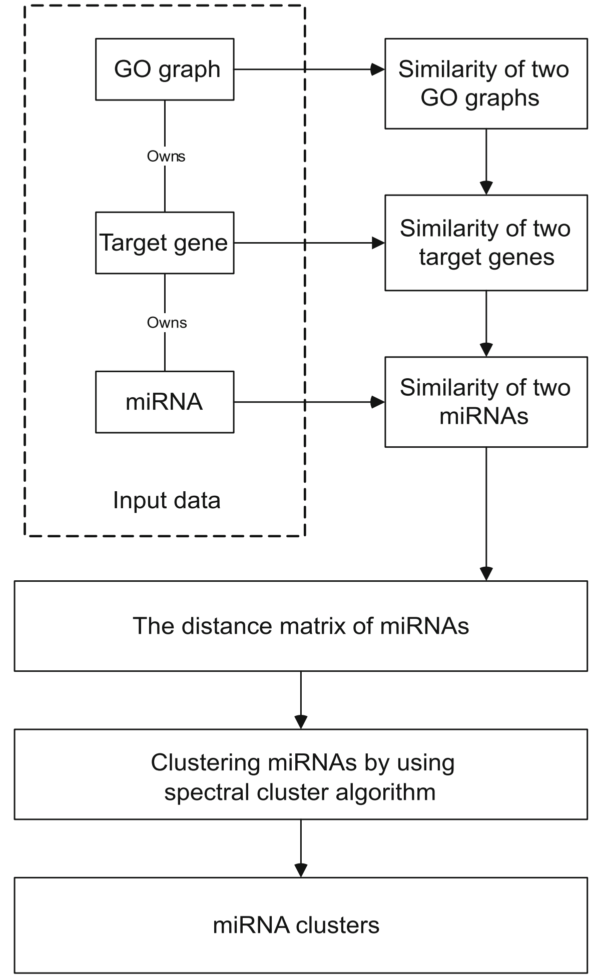
GO graph
Owns
Target gene
Owns
miRNA
Input data
Similarity of two GO graphs
Similarity of two target genes
Similarity of two miRNAs
The distance matrix of miRNAs
Clustering miRNAs by using spectral cluster algorithm
miRNA clusters

Fig. 3 Our method for clustering miRNAs. The raw input to the method consists of data of miRNAs, the target genes of the miRNAs, and the GO graphs of the miRNAs' target genes. Firstly, the similarity of two GO graphs is computed according to the weighted information content. Secondly, the similarity of two target genes of two miRNAs is calculated by using the similarity of their GO graphs. Thirdly, the similarity between two miRNAs is computed by their similarity of their target genes. Then the miRNA distance/similarity matrix is constructed by these similarity scores. Finally, the spectral clustering method is applied to cluster the miRNAs

Compute a weighted information content of a term in a GO graph

The traditionally defined information contents of two terms in a GO graph can be exactly the same even if the two terms have different depths in the graph [17]. We propose a new measurement for the information content to deal with this issue. It is a descendant-based information content, adjusted by a weight proportional to the depth and the relationships of the nodes in the path of the term to the root node.

For a GO graph $G = (Term, Edge)$, the information content (IC) of a term $t \in Term$ is computed by

$$IC(t) = -\log \frac{1 + \|descends(t)\|}{\|Term\|} \tag{1}$$

where $\|descends(t)\|$ is the number of t's descendants in G, and $\|Term\|$ is the number of terms of G. This equation implies that a parent node's IC is always smaller than its child node (i.e., a GO term closer to the root node has a smaller IC value); and that two different leaf terms have the same IC (because they do not have any descendants). For example, term 0048731 in Fig. 2 has 3 descendants (0072359, 0072358, and 0048513). Its IC is $-\log(\frac{1+3}{7}) = 0.243$. The IC values of the other terms in Fig. 2 are listed in Table 1.

Definition 6. ***The most informative common ancestor.*** *Given a term graph, if a leaf term can reach to a node by walking through a direct line, this node is called an ancestor term of the leaf term. A common ancestor term is such an ancestor term that two input leave terms can both reach. The most informative common ancestor (MICA) is the common ancestor term that has the maximum IC value of two term graphs.*

These information contents are then adjusted by a weight of the path of the term to the root node or to the MICA node. The weight is named *edge weight* which is determined by two factors: the relationships of the edges in the path and the distance of the path. Let $TG_A = \{A, Term_A, Edge_A\}$ and $TG_B = \{B, Term_B, Edge_B\}$ be two term graphs, G be the merged graph, and *mica* be the two term graphs' MICA. For a term t in the graph G, its distance weight $\omega_{edge}(t, G)$ is defined as

$$\omega_{edge}(t, G) = \frac{2}{\pi} * \arctan \frac{1}{\omega_{depth}(t, G)} \tag{2}$$

where $\omega_{depth}(t, G)$ is 1 if t is the root node. If t is *mica*'s ancestor or *mica*, $\omega_{depth}(t, G)$ is the product of all the relationships in the longest path from t to the root node of TG_A or TG_B, otherwise it is the product of all the relationships in the longest path from t to the *mica*.

Table 1 The *IC*, length weights, edge weight, and weighted information content of the terms in Fig. 2

Term	GO:0008150	GO:0032502	GO:0048856	GO:0048731	GO:0072359	GO:0048513	GO:0072358
IC	0	0.067	0.146	0.243	0.544	0.845	0.845
$\omega_{depth}(t, G)$	1	0.9	0.81	0.729	0.9	0.9	0.63
$\omega_{edge}(t, G)$	0	0.533	0.567	0.599	0.533	0.533	0.642
$\omega IC(t, G)$	0	0.189	0.288	0.382	0.538	0.671	0.737

The *is_a* relationship is more important than the *part_of* relationship. Thus, we set *is_a* as $W_{edge} = 0.9$ and *part_of* is set as $W_{edge} = 0.7$. We note that the edge weight of a term increases when the term is farther to the root node or to the MICA. The arctan transformation is to standardize the reciprocal of two length weights as they can be very large.

For example, ω_{depth} of term 0008150 in Fig. 2 is 1, since it is the root node of the GO tree. The ω_{depth} value of its child 0032502 is 0.9, as the relationship between these two terms is *is_a*. The ω_{depth} value of $GO : 0072359$ is 0.9, because this term is MICA's descendant term and the relationship between MICA and this term is *is_a*. The other terms' ω_{depth} values are listed in Table 1. We note that if a term has multiple longest paths to the root node or MICA, we choose the one which provides the biggest edge weight for the term. The edge weights of the terms in Fig. 2 are also listed in Table 1 (see the second-last row).

By Eq. 2, if an ancestor term of the MICA is near to the root, this term contributes less similarity to the two term's trees as it is more general. For a descendant term of the MICA, which is near to MICA, contributes less dissimilarity. Unlike traditional edge-based methods [18] which set all the edges as the same weight, our method considers both the distance of the terms to the root or MICA node and the difference between *is_a* and *part_of* to measure the distance weight of a term.

We combine the initial information content (i.e., Eqn. 1) of a term t in a merged GO graph G and its edge weight (i.e., Eqn. 2) to derive a weighted information content for the term. It is denoted by $\omega IC(t, G)$, defined as

$$\omega IC(t, G) = \sqrt{IC(t) * \omega_{edge}(t, G)} \tag{3}$$

The weighted information contents of all the terms in Fig. 2 are shown in the last row of Table 1.

By this definition, only the root node has a weighted information content of 0. It is understandable because a root node does not contribute to the weight—it has no parent node and it is the ancestor of all other terms. As some terms (e.g., the leaf nodes of a graph) having the same IC can occur at different levels of the graph, the IC value alone cannot reflect the different importance of these terms. This is the main reason why edge weights are used to resolve this issue.

Determine the similarity of two genes based on weighted information content

As the GO graph of a gene may contain multiple leaf term graphs, we first define the similarity between two term graphs, and then define the similarity between two GO graphs.

Given two term graphs $TG_A = (A, Term_A, Edge_A)$ and $TG_B = (B, Term_B, Edge_B)$, the similarity of these two graphs is measured through the weighted information contents of their common terms as well as their uncommon terms. We use Fig. 4 to illustrate this definition. The common terms between the two leaf term graphs $TG_{0048513}$ and $TG_{0072358}$ are shown in the dashed square box. The terms outside the square box are the uncommon terms of these two leaf term graphs. The MICA of these two term graphs is $GO : 0048731$. The ancestry terms of the MICA are all in the square box, and all the descendant terms of MICA are outside the box.

The similarity of the two term graphs TG_A and TG_B, denoted by $sim(TG_A, TG_B)$, is defined as

$$sim(TG_A, TG_B) = \frac{\sum\limits_{t \in common} \omega IC(t, G)}{\sum\limits_{t \in common} \omega IC(t, G) + \sum\limits_{t \in uncommon} \omega IC(t, G)} \tag{4}$$

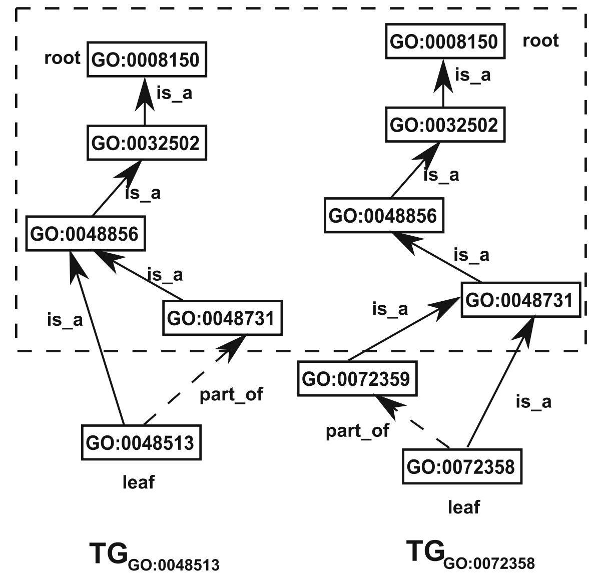

Fig. 4 A schematic illustration of the similarity between two graphs

where, *common* is the set of common terms between TG_A and TG_B, *uncommon* is the set of their uncommon terms, and G is the merged graph of TG_A and TG_B.

The similarity $sim(TG_A, TG_B)$ ranges its values between 0 and 1. When the MICA of the two term graphs is the root node, the similarity between these two graphs is 0. If the two term graphs are the same, their similarity is 1.

As mentioned, the GO graph of a gene may contain multiple leaf nodes which correspond to multiple term graphs. Use G_1 to denote the GO graph of gene g_1 and G_2 as the GO graph of gene g_2. The similarity between G_1 and G_2 is measured by averaging the similarities of every leaf term graph of one GO graph (G_1 or G_2) with the other GO graph (G_2 or G_1). Assume G_1 has n_1 number of leaf terms $LeafTerms_1 = \left\{l\_node_1^1, l\_node_1^2, \ldots, l\_node_1^{n_1}\right\}$, and their leaf term graphs are denoted by $TG(LeafTerms_1) = \left\{TG_{(l\_node_1^1,G_1)}, TG_{(l\_node_1^2,G_1)}, \ldots, TG_{\left(l\_node_1^{n_1},G_1\right)}\right\}$. Also assume G_2 has n_2 number of leaf terms $LeafTerms_2 = \left\{l\_node_2^1, l\_node_2^2, \cdots, l\_node_2^{n_2}\right\}$, and their leaf term graphs are denoted by $TG(LeafTerms_2) = \left\{TG_{(l\_node_2^1,G_2)}, TG_{(l\_node_2^2,G_2)}, \cdots, TG_{\left(l\_node_2^{n_2},G_2\right)}\right\}$.

The similarity between G_1 and G_2, denoted by $sim(G_1, G_2)$, is given by

$$sim(G_1, G_2) = \frac{\sum\limits_{tg \in TG(LeafTerms_1)} sim(tg, G_2) + \sum\limits_{tg \in TG(LeafTerms_2)} sim(tg, G_1)}{n_1 + n_2} \tag{5}$$

where, $sim(tg, G_2) = \max\limits_{1 \le i \le n_2} sim(tg, TG_{(l\_node_2^i,G_2)})$; and $sim(tg, G_1)$ is similarly defined. We note that the maximal similarity of leaf-leaf term graph pairs is applied to measure the similarity between one leaf term graph and one GO graph.

Clustering miRNAs for similar functions based on their target genes' similarity/distance matrix

A miRNA usually has several target genes. In this work, the similarity between two miRNAs is measured by the similarity between the two sets of their target genes. We first introduce the similarity between a set of genes and a gene. Given a set of genes $GS = \{g_1, g_2, \ldots, g_m\}$ and a gene g', the similarity between GS and g' is given by

$$sim(GS, g') = \max_{1 \le i \le m} sim(G(g_i), G(g')) \tag{6}$$

where $G(g_i)$ is the GO graph of g_i, and $G(g')$ is the GO graph of g'.

An alternative method for measuring the similarity between a gene set and a gene is to take the average of the individual GO terms' similarities. However, the average of the individual GO terms' similarities can underestimate the true similarity between a gene set and a gene [15], as we use the similarity between a gene set and a gene to compute the similarity between two gene sets. This underestimated value will lower down the similarity between two gene sets.

Suppose we are given two miRNAs denoted by R_1 and R_2. Assume R_1 has s number of target genes $GS_1 = \{g_1^1, g_1^2, \ldots, g_1^s\}$ and R_2 has k number of target genes $GS_2 = \{g_2^1, g_2^2, \ldots, g_2^k\}$. The similarity of these two miRNAs R_1 and R_2 is defined as

$$sim(R_1, R_2) = \frac{\sum\limits_{1 \le i \le k} sim(GS_1, g_2^i) + \sum\limits_{1 \le j \le s} sim(GS_2, g_1^j)}{s + k} \tag{7}$$

The distance *dsim*, or dissimilarity, between two miRNAs R_1 and R_2 is computed by

$$dsim(R_1, R_2) = 1 - sim(R_1, R_2) \tag{8}$$

The dissimilarity between two miRNAs can be viewed as their distance, and thus it can be applied for clustering a group of miRNAs.

For a number h of miRNAs $R_1, R_2, \ldots, R_h$, a spectral clustering method [19] is applied to the dissimilarity matrix of these miRNAs to detect subsets of miRNAs which each have a similar function. The spectral clustering method is described as follows:

- For a set of data points $X = \{x_1, x_2, \ldots, x_n\}$, construct a complete graph *SPG* in which the data point of X is the node of *SPG*. The weight $\omega_{(x_i,x_j)}$ of each edge that connects with nodes x_i and x_j, is defined as:

$$\omega_{(x_i,x_j)} = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \tag{9}$$

  Let *Wam* denotes denote the weighted adjacency matrix of the graph *SPG*.
- Calculate the normalized Laplacian L from *Wam* and compute the first k eigenvectors of L. The k is the number of clusters. Then, the k eigenvectors can be used to construct a $n * k$ matrix U.
- The matrix U can be seen as a set of n data points under k features. Apply the k-means clustering algorithm to divide these data points.

For the h number of miRNAs $R_1, R_2, \ldots, R_h$, the weighted adjacency matrix *Wam* for the spectral clustering is determined by

$$\omega_{(R_i,R_j)} = e^{-\frac{dsim(R_i,R_j)^2}{2\sigma^2}} \tag{10}$$

where, $1 \le i, j \le h$.

The source code of spectral clustering is available at website http://sourceforge.net/projects/spectralcluster/?source=typ_redirect. Our source code of computing the weighted information contents can be downloaded from http://bioinformatics.gxu.edu.cn/bio/data/CWLan/spectralcode.tar.gz. Our results on clustering are available at http://bioinformatics.gxu.edu.cn/bio/data/CWLan/spectralresult.tar.gz.

There are two vital parameters in the spectral clustering method. The first one is σ in Eq. 10 and the other is the number of clusters. These two parameters have heavy influence to the clustering result. Traditional methods usually use several different choices of σ to test and choose the best σ by comparing the results. However, such approaches are time consuming. The selection of a good cluster number has been a challenging issue. In general, the cluster number relies on the user's experience. In this paper, a self-tuning method is applied to decide an optimal value of σ and we also employ the eigenvalues of the normalized Laplacian matrix to determine an optimal number for the clusters.

Self-tuning for the selection of σ. Equation 10 uses the square of σ. The concern is that σ will be the same even though for computing two different data points. The self-tuning method employs two different σ values to calculate the weight of an edge. For the set of miRNAs $R = \{R_1, R_2, \ldots, R_h\}$, the weight of its adjacency matrix by our self-tuning method is:

$$\omega_{self}(R_i, R_j) = e^{-\frac{dsim(R_i,R_j)^2}{2\sigma_i * \sigma_j}} \quad i,j = 1,2\ldots,h \tag{11}$$

where σ_i is the average distance of R_i to all other miRNAs, given by

$$\sigma_i = \frac{\sum_{j=1}^{h} dsim(R_i, R_j)}{h-1} \tag{12}$$

Select the optimal cluster number. An optimal number of clusters of the miRNAs is determined from the trend of the eigenvalues of the normalized Laplacian matrix L. Suppose these eigenvalues are $Eig = \{\lambda_1, \lambda_2, \ldots, \lambda_h\}$ sorted in a descending order. If the eigenvalue λ_{k+1} $(1 \leq k < h)$ is very small and the trend of the subsequent eigenvalues goes stable, then the number of clusters can be set as k. If the differences between two consecutive eigenvalues are very small, we said that the trend of the consecutive eigenvalues goes stable. Figure 5 presents the first 50 largest eigenvalues of the normalized Laplacian matrix L of the miRNA data set used in the second section of "Data Sets and Definitions Related to GO Graphs". Therefore, the cluster number 13 is selected.

Results

Our method was applied to the data set of 289 Human miRNAs downloaded from http://mirtarbase.mbc.nctu.edu.tw/ to cluster their function groups. (The details of the data set have been described in the second Section.) 265 of these miRNAs are associated with a disease. The disease information of the remaining 24 miRNAs are not available from the database at the time of this work. Instead their functions were predicted by our clustering method. We report four parts of computational results in this section. The first part shows the importance of our edge weights to the information contents of term nodes in GO graphs. The second part selects a good edge relationship weight and discusses the effect of the edge relationship weight on the cluster number. The third part compares our method with three existing methods to understand our superior clustering performance. The forth part reports the function annotation results for new miRNAs by our clustering method.

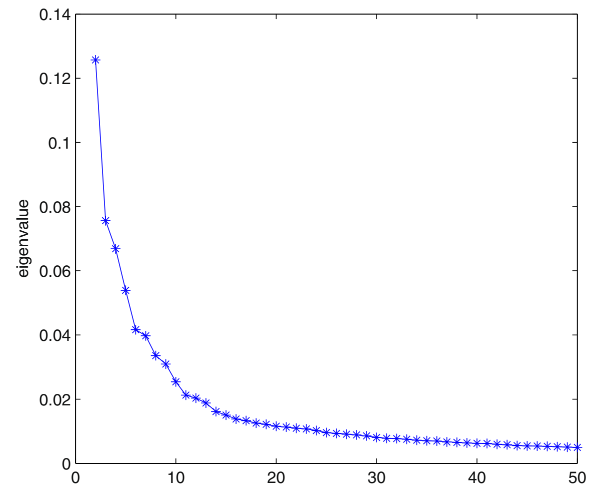

Fig. 5 The first 50 largest eigenvalues

The effect of edge weights on the information contents of term nodes

The results in this section explain why we introduce the edge weight of a term to adjust the information content of the term (using our Eqn. 1). Figure 6 presents the numbers of leaf terms of the GO trees when the term level varies. The majority of these leaf terms are at level 5. By the traditional definition of information content, all these leaf terms have the same *IC* value, although they are at different levels of the trees. This is why we use an edge weight to adjust the information content of a leaf term and make it proportional to the distance of the path from the leaf node to the MICA. Namely, a leaf term having a far distance to the MICA should contain more information than a leaf term closer to the MICA.

Figure 7 shows the numbers of different *IC* values for the terms at the same levels of the GO trees, where the *IC* values are computed according to our definition of information content. For example at level 4 of these GO trees, there are many terms having different *IC* values.

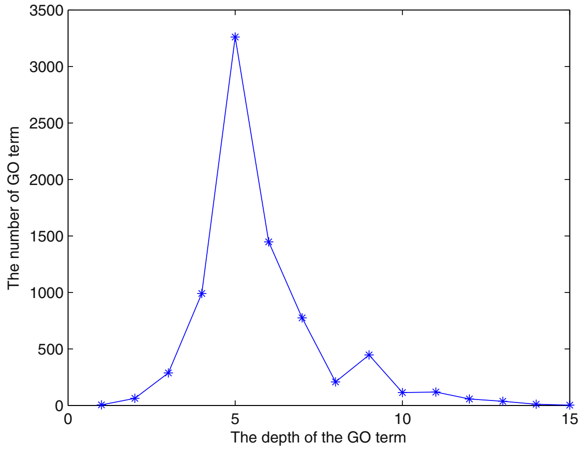

Fig. 6 The numbers of leaf terms at different levels of the GO trees

These terms should have different *IC* values, as they contain different numbers of descendants. The traditional edge-based method [18] assigns the same weights to the terms at the same level. The combination of *IC* value and the edge weight by our Eqn. 3 overcomes this weak point of the node-base method [12] and the edge-based method [18].

Effect of edge relationship weights on the number of clusters

We have proposed to use edge relationships in GO trees to define an edge weight. As this work focuses on the prediction and annotation of miRNA functions, we use the molecular functions of GO terms which only have relationship *is_a* between GO terms. We tested and compared the effectiveness of 9 different weights of the *is_a* relationship from 0.1 to 0.9 with step increase of 0.1 on the function prediction performance for the 265 miRNAs.

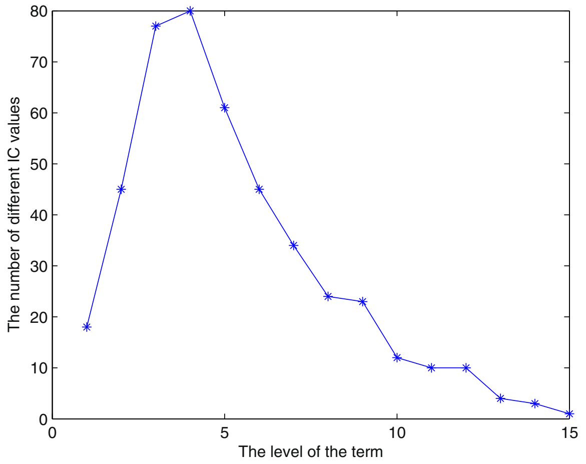

Fig. 7 The numbers of different *IC* values at the same levels of the GO trees

An accuracy rate is used to measure the quality of the clustering results. It is defined as the proportion of miRNAs in a cluster which are associated with the same disease:

$$accr(disease, C) = \frac{nm(disease)}{\|C\|} \tag{13}$$

where $nm(disease)$ is the number of miRNAs associated with the *disease*, and $\|C\|$ is the total number of miRNA in the cluster *C*. Usually, a cluster of miRNAs formed by computational methods can have diverse proportions of miRNAs each sharing a different disease. We used the accuracy of the prevailing disease to represent the accuracy rate of the cluster. A high accuracy of a cluster means that many miRNAs associated with the same disease are clustered into the same group, implying the weight of the *is_a* relationship is properly assigned for the function prediction of new miRNAs.

The breast cancer, stomach cancer, and hepatocellular carcinoma were three diseases which are most prevailing in three clusters for all of the situations of the relationship weight from 0.1 to 0.9. The detailed accuracy rates are presented in Table 2. We found that 0.8 was a good relationship weight.

Figure 8 shows that the eigenvalues from the 10th to the 20th become very stable (i.e., the difference between two consecutive eigenvalues becomes close to 0) under all situations of the relationship weight from 0.1 to 0.9. As discussed above, cluster number 13 was chosen to group miRNAs of similar functions. It can be seen that the effect of the edge relationship weights on the cluster number is very small.

Clustering performance comparison with existing methods

We compared our method with three literature methods to understand the grouping performance for miRNAs of similar functions. The three literature methods are a node-based approach by Lin [12], edge-based approach by Viktor [20], and hybrid approach by Wang [21].

The performance by each clustering method is reported in Table 3.

For the Breast Neoplasms Cluster, all the four methods have very close and competitive accuracy. For the Hepatocellular Carcinoma Cluster, Lin's method has the largest number of miRNAs and the highest accuracy. Our method has the second largest number of miRNAs on the hepatocellular carcinoma cluster and the second highest accuracy. For the Stomach Neoplasms cluster, our method yields the largest number of miRNAs and

Table 2 Accuracy rates of three different clusters by setting 9 different edge relationship weights

Relationship Weight group	0.1 group	0.2 group	0.3 group	0.4 group	0.5 group
Breast Neoplasms Cluster	35/43 = 0.814	35/42 = 0.833	35/42 = 0.833	37/44 = 0.841	30/35 = 0.857
Hepatocellular Carcinoma Cluster	14/25 = 0.56	10/19 = 0.526	16/27 = 0.593	14/24 = 0.583	12/23 = 0.522
Stomach Neoplasms Cluster	15/27 = 0.55	13/24 = 0.542	13/26 = 0.500	15/25 = 0.600	14/26 = 0.538
Relationship Weight group	0.6 group	0.7 group	0.8 group	0.9 group	
Breast Neoplasms Cluster	33/40 = 0.825	36/43 = 0.837	36/43 = 0.837	36/43 = 0.837	
Hepatocellular Carcinoma Cluster	12/27 = 0.444	17/23 = 0.739	18/24 = 0.750	11/26 = 0.423	
Stomach Neoplasms Cluster	11/24 = 0.458	17/28 = 0.607	17/27 = 0.630	14/26 = 0.538	

the highest accuracy rate. Overall, our method generates the best accuracy for the union of the three clusters, and has the largest coverage of the miRNAs (the total number of miRNAs in the clusters). Wang's method has the same coverage of 94 miRNAs as ours, but its accuracy is about 30% lower. Lin's method has a similar overall accuracy as ours, but its coverage is about 20% smaller.

Co-regulating miRNAs and function annotations for new miRNAs

As suggested, miRNAs clustered into the same group should have similar functions. Some of our experiments have verified this point. For example, the pair of miRNA-519d and miRNA-216a in the Hepatocellular Carcinoma cluster have a similar function. In fact miRNA-519d [22] and miRNA-216a [23] had been both found to up-regulate PTEN in hepatocellular carcinoma cells. Another example is from the Breast Cancer cluster about the pair of miRNA-205 and miRNA-145. miRNA-205 is involved in the regulation of breast cancer [24], while miRNA-145 also plays a vital role in regulating breast cancer [25]. In the Stomach Cancer cluster, it can be confirmed that miRNA-150 is related to stomach caner [26] and miRNA-106a is also related to this cancer [27]. Many previous studies have indicated that multiple miRNAs can work together to effect cancer formation [28]. Our method to identify these miRNA clusters can assist in investigating this mechanism [29].

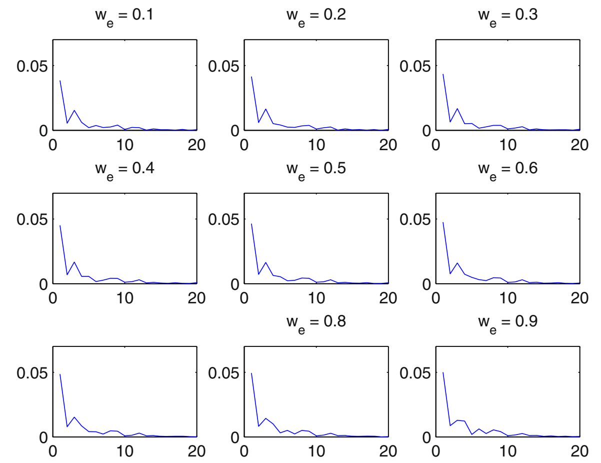

Fig. 8 The consecutive differences of the first 20 largest eigenvalues (i.e., the differences between the kth and $(k+1)$th eigenvalue) under the setting of different edge relationship weights

The functions/disease information of some miRNAs (24) of our 289-miRNA data set are still un-annotated in the HMDD database. However these un-annotated miRNAs can be clustered into some groups by our method, and their functions can be annotated according to the prevailing functions of the groups:

- 5 of the 24 un-annotated miRNAs are grouped into the breast cancer cluster (miRNA-129, miRNA-135a, miRNA-196a, miRNA-5787, and miRNA-9),
- 4 are grouped in the stomach cancer cluster (miRNA-103a, miRNA-181a, miRNA-19b, and miRNA-519a),
- 2 are in the Hepatocellular Carcinoma cluster (miRNA-515 and miRNA-639),
- 8 are classified into the Ovarian cancer cluster (miRNA-512, miRNA-518a, miRNA-521, miRNA-644a, miRNA-876, miRNA-886, miRNA-892b, miRNA-153),
- 2 are clustered in the Prostatic cancer cluster (let-7f and miRNA-219a), and
- 3 are in the Colorectal cancer cluster (miRNA-30c, miRNA-181b, and miRNA-513a).

We have found evidence to support our annotation for some of these miRNAs, for example, miRNA-9 which is asigned into the breast cancer cluster. In fact, recent research shows that miRNA-9 is a potential biomarker for breast cancer [30]. The miRNA-129 is also predicted as a regulator in breast cancer by our method. A recent study can support this prediction: miRNA-129 is down-regulated in breast cancer and has effect on breast cancer migration and motility [31]. It has also been claimed that miRNA-135a is very critical in regulating breast cancer — miRNA-135a

Table 3 Accuracy rates on the three clusters by different methods

Method	Our method	Lin's method	Viktor's method	Wang's method
Breast Neoplasms Cluster	36/43 = 0.837	24/29 = 0.829	26/31 = 0.839	36/43 = 0.837
Hepatocellular Carcinoma Cluster	18/24 = 0.750	20/24 = 0.833	15/22 = 0.682	13/21 = 0.619
Stomach Neoplasms Cluster	17/27 = 0.630	12/23 = 0.522	15/29 = 0.517	14/30 = 0.467

can bind to gene ESRR1 which is related with the breast cancer [32].

For the un-annotated miRNAs in the Stomach cancer cluster, it has been found that miRNA-181a is up-regulated in stomach cancer and has effects on cell proliferation in stomach cancer [33]. Literature work also supports that miRNA-19b and miRNA-519a are associated with stomach cancer [34, 35]

In the ovarian cancer cluster, two studies have shown that miRNA-521 and miRNA-153 are indeed associated with the ovarian cancer [36, 37]. In the Colorectal Cancer cluster, three un-annotated miRNAs miRNA-30c, miRNA-181b, and miRNA-513a can be verified that they are related with this cancer [38–40].

Discussion and conclusion

A variety of methods have been developed to study the functional roles of miRNAs by dividing them into functional groups. For example, Kaczkowski applies the miRNAs' sequence and their secondary structure to cluster miRNAs [41]. However, the miRNAs with a high similarity in sequence/structure cannot guarantee similar functions. Thus, the target genes of miRNAs have been taken as an alternative information source to investigate miRNAs functions.

One of the most prevalent comparative methods for the similarity of target genes is GO graph. The approaches can be classified into two categories: (1) those node-based methods and edge-based methods using GO terms, and (2) pairwise methods and groupwise methods using gene products. Typical node-based methods include Resnik's [42], Lin's [12], and Jiang and Conrath's algorithm [43]. This kind of method applies the *IC* for measuring the similarity of two GO graphs.

The Resnik's method uses only the MICA to measure the similarity between two terms. However, this kind of method neglects the dissimilarity of two terms. Other node-based methods consider both the *IC* value of terms as well as the MICA of two GO graphs, such as Lin's method and Jiang and Conrath's method. Although node-based methods are useful in measuring similarity of terms, the original *IC* value relies on a specific corpus and the structure of the GO graph is largely ignored.

The edge-based methods utilize the length between root nodes and terms. The edge-based method applies the length between root node to the MICA and the distances between the MICA and the leaf terms. The edge method reflects the structure of the GO graph. It assumes all edges have equal weight. However, edges in GO graphs can describe two different relationships (is_a and part_of), which should be assigned with different weights. In addition, the edge-based methods view the weight of all GO terms as the same. However, it is reasonable that a term should have lower weight if it is closer to the root node of the GO graph.

Both edge-based methods and node-based methods have their own advantages. Thus, some methods combine the weight of the term and the distance between two terms to measure the similarity of two GO graphs. This kind of method is called hybrid methods. For example, Sevilla applies the *edge* and the *IC* to measure the similarity of two nodes [44]. While this kind of the method always ignores the relationship of the edge. Wang's method [21] is a typical hybrid method that takes the relationship of the edge into consideration. However, if two term pairs have the same structure, they will have the same similarity value.

This work has introduced a new GO-based method to cluster miRNAs for similar functions. A weighted information content is proposed to measure the importance of a term in a GO graph. Its key idea is to integrate the descendant-based information content, the depth of the term, and the relationships of the edges in the path from the term to the root node. Our weighted information content can overcome some limitations of the conventional node-based and edge-based approaches. The similarity between two GO graphs is based on the weighted information contents of the common terms relative to the information contents of the uncommon terms. These similarities are transferred to estimate the similarities of miRNAs. A spectral clustering method has been applied to the similarity/distance matrix of a set of 289 miRNAs for function grouping. Compared with three state-of-the-art clustering methods, our method show better performance in accuracy to measure the similarity/distance between miRNAs. Our method is also useful for the discovery of co-regulating miRNAs and the function annotation of new miRNAs.

Declarations

This article has been published as part of BMC Bioinformatics Volume 17 Supplement 19, 2016. 15th International Conference On Bioinformatics

(INCOB 2016): bioinformatics. The full contents of the supplement are available online https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-17-supplement-19.

Funding
Publication of this article was funded by a National Natural Science Foundation of China project 61363025, two key projects of Natural Science Foundation of Guangxi 2012GXNSFCB053006 and 2013GXNSFDA019029, Australia Research Council Discovery Project ARC130102124, and China Scholarship Council.

Availability of data and materials
Our source code can be downloaded from http://bioinformatics.gxu.edu.cn/bio/data/CWLan/spectralcode.tar.gz. Our results on clustering are available at http://bioinformatics.gxu.edu.cn/bio/data/CWLan/spectralresult.tar.gz.

Authors' contributions
QC and CL conceived the ideas, designed the methods and performed the experiments. JL and QC supervised the study. QC, CL and JL participated in the analysis. QC drafted the manuscript. CL and JL read and revised the manuscript. All authors approved the final version.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Published: 22 December 2016

References

1. Mazière P, Enright AJ. Prediction of microrna targets. Drug Discov Today. 2007;12(11–12):452–8.
2. Yu J, Wang F, Yang GH, Wang FL, Ma YN, Du ZW, Zhang JW. Human microrna clusters: Genomic organization and expression profile in leukemia cell lines. Biochem Biophys Res Commun. 2006;349(1):59–68.
3. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A. Rfam 11.0: 10 years of rna families. Nucleic Acids Res. 2013;41(D1):226–32.
4. Kozomara A, Griffiths-Jones S. mirbase: annotating high confidence micrornas using deep sequencing data. Nucleic Acids Res. 2014;42(D1): 68–73.
5. Wang D, Wang J, Lu M, Song F, Cui Q. Inferring the human microrna functional similarity and functional network based on microrna-associated diseases. Bioinformatics. 2010;26(13):1644–50.
6. Fontana P, Cestaro A, Velasco R, Formentin E, Toppo S. Rapid annotation of anonymous sequences from genome projects using semantic similarities and a weighting scheme in gene ontology. Plos One. 2009;4(2):4619.
7. Roubelakis MG, Zotos P. Human microrna target analysis and gene ontology clustering by gomir, a novel stand-alone application. BMC Bioinformatics. 2009;10(Suppl 6):S20.
8. Schlicker A, Domingues F, Rahnenfuhrer J, Lengauer T. A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics. 2006;7(1):302.
9. Wu X, Hasan MA, Chen JY. Pathway and network analysis in proteomics. J Theor Biol. 2014;362(0):44–52.
10. Bag S, Ramaiah S, Anbarasu A. fabp4 is central to eight obesity associated genes: A functional gene network-based polymorphic study. J Theor Biol. 2015;364(0):344–54.
11. der Nest MAV, Olson Å, Karlsson M, Lind M, Dalman K, Brandström-Durling M, Elfstrand M, Wingfield BD, Stenlid J. Gene expression associated with intersterility in heterobasidion. Fungal Genet Biol. 2014;73(0):104–19.
12. Lin D. An information-theoretic definition of similarity. In: Proceedings of the 15th International Conference on Machine Learning (ICML 1998), vol. 98. 1998. p. 296–304.
13. Yu H, Gao L, Tu K, Guo Z. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. Gene. 2005;352(0):75–81.
14. Wu X, Pang E, Lin K, Pei ZM. Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge- and ic-based hybrid method. Plos One. 2013;8(5):66745.
15. Bandyopadhyay S, Mallick K. A new path based hybrid measure for gene ontology similarity. IEEE/ACM Trans Comput Biol Bioinformatics. 2014;11(1):116–27.
16. Zelnik-manor L, Perona P. Self-Tuning Spectral Clustering. In: Neural Information Processing Systems. Cambridge; 2004.
17. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009;5(7):1000443.
18. Yu H, Gao L, Tu K, Guo Z. Broadly predicting specific gene functions with expression similarity and taxonomy similarity. Gene. 2005;352(0):75–81.
19. Theodoridis S, Koutroumbas K. Chapter 15 - clustering algorithms iv In: Theodoridis S, Koutroumbas K, editors. Pattern Recognition (Fourth Edition). Boston: Academic Press; 2009. p. 765–862.
20. Pekar V, Staab S. Taxonomy learning: factoring the structure of a taxonomy into a semantic classification decision. In: Proceedings of the 19th International Conference on Computational linguistics-Volume 1. Stroudsburg: Association for Computational Linguistics; 2002. p. 1–7.
21. Wang ZJ, Du Z, Payattakool R, Yu PS, Chen CF. A new Method To Measure The Semantic Similarity Of go terms. Bioinformatics. 2007;23(10): 1274–81.
22. Fornari F, Milazzo M, Chieco P, Negrini M, Marasco E, Capranico G, Mantovani V, Marinello J, Sabbioni S, Callegari E, et al. In hepatocellular carcinoma mir-519d is up-regulated by p53 and dna hypomethylation and targets cdkn1a/p21, pten, akt3 and timp2. J Pathol. 2012;227(3): 275–85.
23. Xia H, Ooi LLPJ, Hui KM. Microrna-216a/217-induced epithelial-mesenchymal transition targets pten and smad7 to promote drug resistance and recurrence of liver cancer. Hepatology. 2013;58(2):629–41.
24. Wu H, Zhu S, Mo YY. Suppression of cell growth and invasion by mir-205 in breast cancer. Cell Res. 2009;19(4):439–48.
25. Sachdeva M, Zhu S, Wu F, Wu H, Walia V, Kumar S, Elble R, Watabe K, Mo YY. p53 represses c-myc through induction of the tumor suppressor mir-145. Proc Nat Acad Sci. 2009;106(9):3207–12.
26. Wu Q, Jin H, Yang Z, Luo G, Lu Y, Li K, Ren G, Su T, Pan Y, Feng B, et al. Mir-150 promotes gastric cancer proliferation by negatively regulating the pro-apoptotic gene egr2. Biochem Biophys Res Commun. 2010;392(3):340–5.
27. Xiao B, Guo J, Miao Y, Jiang Z, Huan R, Zhang Y, Li D, Zhong J. Detection of mir-106a in gastric carcinoma and its clinical significance. Clinica Chimica Acta. 2009;400(1):97–102.
28. Voorhoeve PM. Micrornas: Oncogenes, tumor suppressors or master regulators of cancer heterogeneity? Biochimica et Biophysica Acta (BBA) - Reviews on Cancer. 2010;1805(1):72–86.
29. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, Cui Q. An Analysis of Human MicroRNA and Disease Associations. Plos One. 2008;3(10):e3420.
30. Zhou X, Marian C, Makambi KH, Kosti O, Kallakury BV, Loffredo CA, Zheng YL. Microrna-9 as potential biomarker for breast cancer local recurrence and tumor estrogen receptor status. PLoS One. 2012;7(6): 39011.
31. Wang Q, Tang J, Zhou C, Zhao Q. The down-regulation of mir-129 in breast cancer and its effect on breast cancer migration and motility. Sheng li xue bao:Acta physiologica Sinica. 2012;64(4):403–11.
32. Dai X, Chen A, Bai Z. Integrative investigation on breast cancer in er, pr and her2-defined subgroups using mrna and mirna expression profiling. Sci Rep. 2014;4:6566.
33. Chen G, Shen ZL, Wang L, Lv CY, Huang XE, Zhou RP. Hsa-mir-181a-5p expression and effects on cell proliferation in gastric cancer. Asian Pac J Cancer Prev. 2013;14(6):3871–5.
34. Ooi CH, Oh HK, Wang HZ, Tan ALK, Wu J, Lee M, Rha SY, Chung HC, Virshup DM, Tan P. A densely interconnected genome-wide network of micrornas and oncogenic pathways revealed using gene expression signatures. PLoS Genet. 2011;7(12):1002415.
35. Zhang J, Song Y, Zhang C, Zhi X, Fu H, Ma Y, Chen Y, Pan F, Wang K, Ni J, et al. Circulating mir-16-5p and mir-19b-3p as two novel potential biomarkers to indicate progression of gastric cancer. Theranostics. 2015;5(7):733.
36. Delfino KR, Rodriguez-Zas SL. Transcription factor-microrna-target gene networks associated with ovarian cancer survival and recurrence. PLoS One. 2013;8(3):58608.

37. Kim TH, Kim YK, Kwon Y, Heo JH, Kang H, Kim G, An HJ. Deregulation of mir-519a, 153, and 485-5p and its clinicopathological relevance in ovarian epithelial tumours. Histopathology. 2010;57(5):734–43.
38. Kara M, Yumrutas O, Ozcan O, Celik OI, Bozgeyik E, Bozgeyik I, Tasdemir S. Differential expressions of cancer-associated genes and their regulatory mirnas in colorectal carcinoma. Gene. 2015;567(1):81–6.
39. Bovell LC, Shanmugam C, Putcha B-DK, Katkoori VR, Zhang B, Bae S, Singh KP, Grizzle WE, Manne U. The prognostic value of micrornas varies with patient race/ethnicity and stage of colorectal cancer. Clin Cancer Res. 2013;19(14):3955–65.
40. Mosakhani N, Sarhadi VK, Borze I, Karjalainen-Lindsberg ML, Sundström J, Ristamäki R, Österlund P, Knuutila S. Microrna profiling differentiates colorectal cancer according to kras status. Genes Chromosomes Cancer. 2012;51(1):1–9.
41. Kaczkowski B, Torarinsson E, Reiche K, Havgaard JH, Stadler PF, Gorodkin J. Structural profiles of human mirna families from pairwise clustering. Bioinformatics. 2009;25(3):291–4.
42. Resnik P. Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95). San Francisco; 1995. p. 448–53.
43. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. In: Proc. of the Int'l. Conf. on Research in Computational Linguistics. Taiwan; 1997. p. 19–33.
44. Sevilla JL, Podhorski VS, Guruceaga E, Mato JM, MartinezCru LA. Correlation between gene expression and go semantic similarity. IEEE/ACM Trans Comput Biol Bioinformatics. 2005;2(4):330–8.