Neurocomputing 179 (2016) 283–289

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Plant miRNA function prediction based on functional similarity network and transductive multi-label classification algorithm

Jun Meng [a], Guan-Li Shi [a], Yu-Shi Luan [b,*]

[a] School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China
[b] School of Life Science and Biotechnology, Dalian University of Technology, Dalian 116023, China

ARTICLE INFO

Article history:
Received 25 May 2015
Received in revised form
1 December 2015
Accepted 1 December 2015
Communicated by: L. Kurgan.
Available online 22 December 2015

Keywords:
MiRNA functional similarity
TRAM
Protein–protein interaction network
Prediction

ABSTRACT

Plant miRNAs play critical roles in the response to abiotic and biotic stress. The advancement in the number of plant miRNA functions lags far behind that of plant miRNAs. In this paper, a method to predict the functions of plant miRNAs is proposed. The functional similarity between each pair of miRNAs is inferred based on a weighted protein–protein interaction network (WPPIN) and graph-theoretic properties. A miRNA functional similarity network (MFSN) is constructed by a simple but robust rank-based approach. Transductive multi-label classification (TRAM) is applied to the MFSN. The experimental results demonstrate that our prediction approach obtains high effectiveness in *Arabidopsis thaliana*. It can also be applied to other plant species when protein–protein interaction networks of various organisms are available.

1. Introduction

MicroRNAs (miRNAs) are endogenous, small, non-coding RNAs, approximately 21–25 nucleotides in length [1]. In the growth of an organism, these molecules are widely involved in the post-transcriptional regulation of genes via the translational repression or degradation of target messenger RNAs [2]. As major regulators of gene expression, miRNAs play pivotal roles in many biological processes, such as growth, developmental timing, nutrient homeostasis, and biotic and abiotic stress responses. Currently, the prediction of unknown miRNAs is the focus of miRNA research. We have proposed a useful tool to predict plant pre-miRNAs and their miRNAs in genome-scale sequences [3]. After the accurate identification and prediction of miRNAs are completed, the main objective is to determine their functions. Currently, the research on miRNA function is concentrated on the experimental determination phase. The function of miRNA is inferred by the up-regulating or down-regulating expression of miRNA and its target genes in a direct or indirect method. Zhu et al. [4] found that the accumulation of miR172 during development would lead to missing organs. The phenotype is consistent with that of the mutant miR172 target gene Apetala2. The miRNA-related function can be determined by analyzing the function of its target genes or promoters [5,6]. Many efforts have been made to identify miRNA target genes. For example, a recent study identified 119 targets in *Solanum lycopersicum,* and 106 of them appeared to be new [7]. In other methods, several miRNA target prediction tools were integrated to identify miRNA target genes in plants and animals [8,9].

Although the experimental analysis to determine miRNA function is increasingly accurate and reliable, the process is complicated and time-consuming, and limited functions are available. To overcome these shortcomings, a calculation method used to predict miRNA functions is proposed. A method to predict human disease-associated miRNAs based on a miRNA functional similarity network and a related network algorithm was proposed [10]. In this method, the semantic similarity measure between two groups of miRNA-related diseases is used to evaluate the functional similarity of two miRNAs, and a miRNA function network is built. However, there is no any database similar to the human disease in plants. Therefore, this method cannot be applied to the prediction of plant miRNA function. In recent years, with the rapid development of high-throughput sequencing technologies, plant protein–protein interaction (PPI) data have increasingly been published. These data have been widely used in protein function prediction and other research fields [11–13] because miRNA function is reflected by regulating its target genes via degradation or repressing the transcription of target genes that also affect the synthesis of proteins directly. Therefore, it can be implied that the functions among miRNAs, target genes and proteins are closely linked. By recognizing the target gene function to determine miRNA function, the current method also confirms this hypothesis [5,6]. According

* Corresponding author.
E-mail address: luanyush@dlut.edu.cn (Y.-S. Luan).

http://dx.doi.org/10.1016/j.neucom.2015.12.011
0925-2312/

to the above assumption, a plant miRNA function prediction method is proposed in this paper. A miRNA functional similarity network is constructed based on an integrated analysis of multiple PPI data and a rank-based similarity network construction method. Then, miRNA function is predicted using a multi-label classification algorithm on the miRNA functional similarity network. Furthermore, a comparative analysis showed that our method was more effective and reliable than a widely used computational method, miRFunSim [14], in *Arabidopsis thaliana* and a human dataset.

2. MiRNA functional similarity calculation

In this paper, the calculation of miRNA functional similarity is based on weighted protein–protein interaction network (WPPIN) and graph-theoretic properties. First, a WPPIN is constructed by combining a protein–protein interaction network (PPIN) with gene ontology (GO)-term semantic similarity weights [15]. Gene selection integrated with GO terms is used to analyze gene-expression data [16,17]. Chen and Wang [18] conducted a theoretical analysis and gave an experimental demonstration of integrating gene-expression data with biological knowledge, such as GO terms, which improved prediction accuracy and interpretability over gene-based prediction models. GO is composed of three categories: biological process (BP), molecular function (MF) and cellular component (CC) [19]. We separately calculate the weight of the PPIN in these three categories to obtain their WPPINs. Second, the miRNAs target genes are predicted with two widely used tools (psRNATarget [20] and Targetfinder [21]). Third, the functional similarity between target genes is calculated based on the WPPIN and an improved weighted breadth-first-search (BFS) algorithm. It is calculated by the best average accumulated weight method, which is represented as follows:

$$F_{i,j} = \max\left(\prod_{e \in shortestpath(gene_i, gene_j)} weight(e)\right). \quad (1)$$

where $shortestpath(gene_i, gene_j)$ includes all edges of the shortest path between $gene_i$ and $gene_j$ in the WPPIN. $Weight(e)$ represents the edge weight of the WPPIN. Function $\max(x)$ indicates that $F_{i,j}$ is the maximum accumulated weight in all paths if there is more than one shortest path between $gene_i$ and $gene_j$.

Then, a functional similarity matrix between target gene sets can be obtained. Finally, the functional similarity of each pair of miRNAs is calculated based on a functional similarity matrix by an improved best-match average method (BMA) [22–24], which is defined as follows:

$$FS_{TarSet_{A,B}} = \frac{\sum_{i=1}^{m-m'} \max_{1 \le j \le n} (F_{i,j}) + \sum_{j=1}^{n-n'} \max_{1 \le i \le m} (F_{i,j})}{(m-m')+(n-n')}. \quad (2)$$

where m and n are the numbers of target genes of $miRNA_A$ and $miRNA_B$, respectively, and m' and n' are the numbers of target genes that are not included in the WPPIN. The calculation of miRNA functional similarity is shown in Fig. 1.

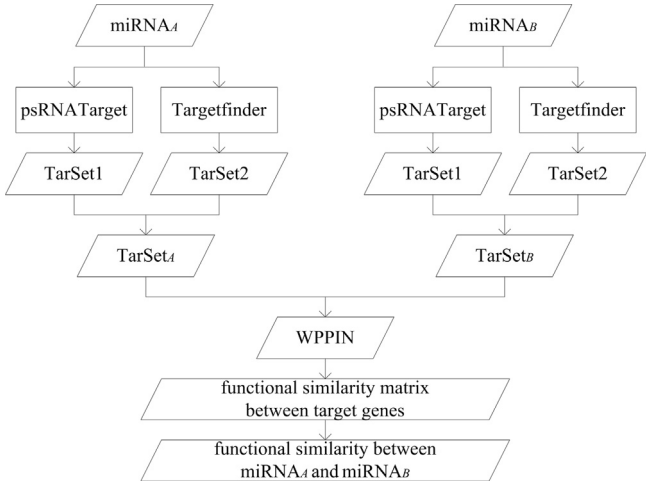

Fig. 1. Functional similarity calculation between miRNAs.

3. Construction of miRNA functional similarity network

3.1. Network construction based on clustering coefficient

The clustering coefficient is a measure of degree in graph theory. It represents the degree to which nodes in a graph tend to cluster together. In the network, the clustering coefficient of node i is defined as $C_i = 2n_i/k_i(k_i-1)$, where n_i is the number of edges between the k_i (>1) first neighbors of node i; if $k_i = 1$, then $C_i = 0$. The clustering coefficient of the network is defined as the average clustering coefficient of all nodes: $C = \sum_{i=1}^{N} C_i/N$, where N is the number of nodes in the network. If $N=0$, we define $C=0$.

The construction of a miRNA function similarity network based on a clustering coefficient can be thought of as a process whereby the edges are removed from the initial graph by increasing the threshold gradually. According to the definitions in Section 2, the functional similarity value between two miRNAs is obtained, and there is an edge between these two miRNAs. The similarity threshold is set a range from 0 to 1 with step 0.01. For each threshold t, the edge between two miRNAs whose functional similarity value is not less than t is reserved and used to construct a miRNA functional similarity network. According to the different value of t, different functional similarity networks and the corresponding clustering coefficient $c(t)$ are obtained. In system biology, a true network should be scale-free and highly modular. As a result, the clustering coefficient $c(t)$ of the biology network should be higher than that of the random network $c_r(t)$; the difference is $\Delta c(t) = c(t) - c_r(t)$. Therefore, the aim of the method is to find the maximum t where $\Delta c(t)$ is increased when the threshold ranges from 0 to t. The critical threshold is the first t value that satisfies $\Delta c(t+0.01) - \Delta c(t) < 0$. The final similarity network is constructed based on the value of t.

The calculation of the clustering coefficient in a random network is different from that of the real biology network. Because there are many random networks and their shapes vary, the clustering coefficient cannot be determined by a network or networks. Therefore, we can use a statistical method to calculate the clustering coefficient of a random network [25]. If the degree of a node is k, and N is the total number of all nodes, the value of the clustering coefficient in a random network is $c_r(t) = (\overline{k^2} - \overline{k})^2/\overline{k}^3$, where $\overline{k} = \sum_{i=1}^{N} k_i/N$ and $\overline{k^2} = \sum_{i=1}^{N} k_i^2/N$.

The network is constructed based on threshold t. The network can be defined as $G(V, E, W, T)$, where $V = \{miR_1, miR_2, ..., miR_N\}$ denotes the node sets, T is the threshold, and $E = \{e_{ij} = \prec miR_i, miR_j \succ | FS_{miR}(miR_i, miR_j) \ge T\}$ represents the reserved edges whose functional similarity values are not less than T. $W = FS_{miR}(miR_i, miR_j)$ represents the edge weight between two miRNAs.

3.2. Rank-based network construction

Clustering coefficient-based network construction is a type of values-based approach. This method is significantly limited by the same threshold for all nodes of the network. Compared to the