A new method for measuring functional similarity of microRNAs

Guangchuang Yu1,a, Chuan-Le Xiao1,a Xiaochen Bo3, Chun-Hua Lu1,3, Yide Qin4, Sheng Zhan1, Qing-Yu He*1.

1Institutes of Life and Health Engineering, Jinan University, 601 Huangpu Ave. W. Guangzhou, 510632, China; 2Beijing Institute of Radiation Medicine, Beijing, 100850, China; 3College of Life Science and Technology, Guangxi University, Nanning, 530004, China; 4Department of Biochemistry and Molecular Biology, Anhui Medical University, Hefei, 230032, China. a These authors contributed equally in this work.

Received: 28 July 2010 Accepted: 18 August 2010 Available Online: 10 September 2010

MicroRNAs (miRNAs) are a group of small RNAs with regulatory roles at post-transcriptional level. Although they have been clustered based on their sequence or structure similarities, there is still no effective method to determine their functional similarities due to the lack of miRNAs functional annotation. To address this critical need, we presented here a novel method for systematic study of functional similarities among human miRNAs by using their target genes GO semantic similarities. The functional similarities were validated by comparing with miRNA expression similarities. To extract the highly significant clusters, we used multi-scale bootstrap re-sampling in clustering miRNAs functional similarities. The clustering of human miRNAs based on target genes molecular function annotation led to 44 significant clusters. The clustering results were coherent with biological knowledge. Our analysis suggests that systematic clustering based on target genes GO semantic similarities can aid to reveal the functional diversity of miRNA families. Additionally, this method can be extended to other species and used to predict novel miRNA functions.

Keywords: MicroRNA; Gene Ontology; Semantic Similarity; Clustering.

1. Introduction

MicroRNAs (miRNAs), approximately 22 nucleotides in length, are non-coding RNAs that play crucial roles in post-transcriptional regulation. MiRNAs are evolutionarily conserved, and generally transcribed by RNA polymerase II. MiRNAs perform their functions by RNA induced silencing complex (RISC), leading their target miRNAs to direct destructive cleavage or translational repression. MiRNAs are considered to represent one of the most important components of the cell. They involve in many critical biological processes, including cell development and differentiation [1,2], proliferation [3], apoptosis [4], development [5], immune system regulation [6,2], cancer progression [7], and virus-host interaction [8] and therefore represent potential targets for therapeutic applications. According to miRBase [9], the number of registered miRNA genes continues to grow rapidly. However, hundreds of recently identified miRNAs have unknown functions due to the lack of experimental strategy for systematic identification of their regulating targets.

In order to better understand miRNAs, it is increasingly necessary to measure their functional similarity and thus to infer novel potential functions for miRNAs. Human miRNAs have been grouped into 46 families on the basis of hairpin sequences conservation by Rfam [10], and 60 families according to pre-miRNAs sequence and secondary structures by using FOLDALIGN [11]. Many human clusters containing miRNAs without sequence homology was found [12] which indicating the current strategies for measuring miRNA similarity have some flaws. Since miRNA-mRNA duplex allows mismatch, and target recognition only matches the 6-8nt long seed region, miRNAs with similar sequences and pre-miRNAs with similar structures may have distinctive functions. Therefore, a new method for measuring miRNAs functional similarity is necessitated.

Gene Ontology (GO) is the de facto standard for annotation of gene products. The relationship of different genes was
organized as a directed acyclic graph (DAG), making it feasible for quantitative semantic comparisons. Measuring gene functional similarity based on GO has been widely used in novel GO annotations [13,14], gene function prediction [15] and similarity measurement [16,17], microarray analysis [18], cluster analysis [19,20], pathway analysis [21,22], and localization prediction [23]. The powerfulness of GO semantic similarity measurement has been verified in terms of the correlations with sequence similarities [24], gene expression profiles [25], and protein-protein interactions [26]. However, it is impractical to measure miRNAs functional similarity directly due to the lack of GO terms annotation of most miRNAs and the lack of functional annotation database [27] since most miRNAs functions remain unknown. Fortunately, most of the genes miRNAs regulated are well annotated, making it possible for inferring functional similarity of miRNAs based on their targets. The functions of miRNAs can be inferred by GO enrichment analysis of their predicted targets were reported [28,29], which also indicating that the functional similarity of miRNAs can be measured by quantitatively calculating the similarity of their target genes.

Here, we proposed a new method for systematic study of miRNAs functional similarity. The fundamental concept of our method is based on the functions of miRNAs targets GO semantic similarities. We validated our method by comparing it with miRNA expression similarity and showed that clustered miRNAs have functional relatedness through co-function of targeting genes. The similarities obtained by our method are consistent with biological knowledge of miRNA functional relationship.

2. Material and methods

2.1 Human microRNA target prediction

Experimental identification of target miRNAs is difficult, and TarBase [30] currently lists only 1093 verified target miRNAs for 110 human miRNAs. Due to the lack of experimental targets of miRNAs, we used computational method for miRNA genome-wide target prediction in this study, where target genes were predicted by the algorithm of Probability of Interaction by Target Accessibility (PITA) [31]. PITA uses a thermodynamic model for miRNA-mRNA interaction that was scored by an energy score, ∆ΔG, which equals to the difference between the energy expended on opening the target site structure, ∆Gopen, and the energy gained by forming the duplex, ∆Gduplex. ∆ΔG correlates well with the experimentally measured degree of mRNA suppression were reported [31].

Human miRNAs were downloaded from miRBase, version 12 [9] and human genome was downloaded from UCSC, version 18 [32] which corresponds to the human genome build 36.1 assembled by NCBI. We extracted 3′ untranslated region (UTR) sequences in a single FASTA format file. For genes missing 3′-UTR annotation, 800bp downstream annotated end of the coding sequence were used as the putative UTR. Since miRNA-target interaction requires unpairing of bases flanking the targets, we used a flank of 3 upstream and 15 downstream nucleotides when performing prediction. To reduce false positive, the prediction results were narrowed down using the criteria of 7-8 bases seed length, no G:U wobble or loops, no mismatch and conservation score of 0.9 or higher. To assign an overall miRNA-target score, we computed the statistical weight to sum all the ∆ΔG generated by different sites bounded to the same miRNA as defined in formula 1. Finally we screened out the results by ∆ΔG below -10 kcal/mol. ∆ΔG is an energetic score, and the lower its value, the stronger of miRNA-target binding. ∆ΔG < -10 kcal/mol is expected to be functional in endogenous miRNA expression levels.

\[
Target\text{-Score} = -\log(\sum(\exp(-\Delta \Delta G)))
\]

2.2 Human microRNA functional similarities measurement

For measuring functional similarities among miRNAs, we used GO annotation of their target genes to define the similarity index.

Here, we developed an R package called GOSemSim [33], which implemented five methods proposed by Resnik [34], Lin [35], Jiang [36], Schlicker [37], and Wang [38] respectively, and was extended to support 19 species, including Anopheles, Arabidopsis, Bovine, Canine, Chicken, Chimp, Coelicolor, E coli strain K12 and Sakai, Fly, Human, Malaria, Mouse, Pig, Rhesus, Rat, Worm, Xenopus, Yeast, and Zebrafish. The program is freely distributed under GPL2 and can directly be installed from Bioconductor. The manual and source code are available at http://bioconductor.org/packages/release/bioc/html/GOSemSim.html. Since GOSemSim package only supports Entrez Gene identifier for measuring similarities among human genes, we used Bioconductor package bioRaR [39] to query BioMart [40] databases for mapping RefSeq identifiers to Entrez Gene identifiers. Molecular function ontology was used to annotate target genes, and Wang method was used to calculate similarity. Wang method was based on GO graph structure, and outperformed other algorithms based on information content and thus being more consistent with human perspectives [38]. In order to give a single distance between two miRNAs, we combined similarity scores of multiple target genes as defined in formula 2. Similarities between two gene sets associated with two miRNAs form a matrix. The similarity of these two miRNAs is the average of maximum row scores and column scores. We used this strategy and finally obtained pairwise semantic similarities among human miRNAs.

\[
Sim = \frac{\sum_{1 \leq i \leq m} \max(\text{Score}_{i,j}) + \sum_{1 \leq j \leq n} \max(\text{Score}_{i,j})}{m+n}
\]
2.3 Hierarchical clustering uncertainty assessment

The resulting similarity scores were then clustered by R package Pvclust [41]. When performing clustering, pvclust used multi-scale bootstrap re-sampling to estimate the uncertainty of cluster analysis which has been popular in phylogenetic analysis. The agglomerative method, average linkage, was used, and 10,000 bootstrap replications were run, with relative sample size from 0.5 to 1.4, incrementing in steps of 0.1, for testing p-values. For a cluster with approximately unbiased (AU) p-value > 0.95, the hypothesis with “the cluster does not exist” is rejected with significance level 0.05. Roughly speaking, these clusters not only “seem to exist” attributed to sampling error, but be stably observed when we increase the number of observation [41]. All clusters in this study were extracted with p-value > 0.95.

3. Results and Discussion

We used PITA algorithm to predict human miRNA targets, followed by a strict criteria to control the FDR, and then measured miRNA functional similarity by GOSemSim package. As a result, we obtained the pairwise functional similarity of 533 miRNAs which was provided as Supplemental File 1. The pairwise functional similarity of human miRNA repertoire was shown in Figure 1. The full size of Figure 1 was provided as Supplemental File 2.

Figure 1. Human miRNAs functional similarities.

MiRNAs with high functional similarity may tend to have similar expression profiles. Taking miRNA functional similarity in pairs against expression similarity should show a positive relationship. For further evaluating the quality of our result, we study the relationship of miRNAs functional similarity and expression similarity. We used Pearson’s correlation coefficients to calculate miRNA expression similarity by using miRNA expression profiles of 345 miRNAs in 40 normal human tissues obtained from Liang’s investigation [42]. We classified miRNA pairs into separate groups according to functional similarity values by a step of 0.1, and calculated the average of functional similarity and expression similarity of each group, and then measured the correlation of functional similarity with expression similarity (see Supplemental File 4). As expected, functional similarity obtained by our method showed positive correlation with expression similarity (r = 0.6055), in which the functions of miRNAs can partially be explained by their expression level (r² = 0.3666).

The pairwise miRNAs functional similarity matrix was then clustered by Pvclust package. We obtained 44 Pvclust clusters with AU p-value > 0.95, containing 401 miRNAs. The result of cluster analysis with high p-value highlighted was provided as Supplemental File 3.

Many miRNAs were reported to be associated with diseases. It has been reported that miRNAs implicated in similar disease often have similar functions [27]. In our results, miRNAs associated to similar diseases were more likely to be grouped in the same cluster. For instance, in cluster 23, hasa-miR-215 [43], hasa-miR-221 [44], hasa-miR-194 [45], hasa-miR-193b [46], and hasa-miR-429 [47] were all reported to be associated with adenocarcinoma. In cluster 44, hasa-miR-453 and hasa-miR-219 [48], hasa-miR-93 [49] were reported to relate with breast cancer; while in cluster 29, hasa-miR-30a, hasa-miR-150, hasa-miR-223, and hasa-miR-600 were implicated to relate with Lupus Vulgaris [50].

MiRNAs conserved in evolution may regulate the cardinal biological process cooperatively. We identified that these miRNAs are tend to cluster together. For example, in cluster 24, hasa-miR-300, hasa-miR-495, hasa-miR-154, and hasa-miR-496 were reported to be conserved in genomes of human, chimp, mouse, rat, dog, and cow [51]. In cluster 33, hasa-miR-410, hasa-miR-377, hasa-miR-668, and hasa-miR-381 are conserved in many mammalian genomes and believed to act cooperatively [51]. Transposable elements (TEs) contribute to the evolution dynamics of miRNAs. We found that hasa-miR-325 and hasa-miR-545 derived from TE L2 [52] were grouped to cluster 30.

Many mammalian viruses have been shown to modulate the expression of host cellular miRNAs [53]. MiRNA expression profiles altered by viruses form a novel regulatory layer, and these miRNAs can be grouped to partially reveal the cross-talk between host and virus. In our clustering result, hasa-miR-181a and hasa-miR-15a that were altered in stable hepatitis B virus expressing cell line [54] were identified in cluster 33. Hsa-miR-24 and hasa-miR-638 that were found to have expression changes during in vitro acute hepatitis C virus infection [55] were grouped in cluster 25.

The clustering results obtained by our method were consistent with many other investigations, suggesting that our method is reliable to calculate functional similarities and sensible to cluster miRNAs. The clustering results are useful to reveal functional diversity of miRNA families.
As described previously, the target genes of miRNA were predicted by PITA algorithm, and our current results were consistent with biological data. Therefore, we can predict novel miRNA functions by miRNA pairs with high similarities or by the GO enrichment analysis of the corresponding miRNA cluster. For instance, we predicted novel miRNA functions with similarities above 0.8. Many of them were supported by newly published literature. For example, hsa-miR-107 and hsa-miR-103 regulate lipid metabolism [56], hsa-miR-449 and hsa-miR-203 regulate pRb-E2F1 activity [57, 58], hsa-miR-200b and hsa-miR-429 regulate EGFR-driven invasion [59], etc. We also used the cluster information to globally assign the predicted functions to novel miRNAs. GO enrichment analysis were performed in all miRNA clusters, and all the enriched GO terms represent each miRNA cluster with their corresponding p-values and other information were provided in Supplemental File 5. Here, we took cluster 23 as an example. After the hypergeometric test, we select the over-represented GO terms of target miRNAs. GO enrichment analysis were performed in all miRNA clusters, and all the enriched GO terms represent each miRNA cluster with their corresponding p-values and other information were provided in Supplemental File 5. Here, we took cluster 23 as an example. After the hypergeometric test, we select the over-represented GO terms of target miRNAs. GO enrichment analysis were performed in all miRNA clusters, and all the enriched GO terms represent each miRNA cluster with their corresponding p-values and other information were provided in Supplemental File 5. Here, we took cluster 23 as an example. After the hypergeometric test, we select the over-represented GO terms of target miRNAs.

4. Concluding remarks

In summary, we proposed a novel method for inferring the functional similarities of human miRNAs by integrating the information provided by miRNA target prediction algorithm PITA with Gene Ontology annotation data. The significant miRNA families were also analyzed. This method can be extended to measure miRNA functional similarities of other species. The current method relies on the prediction of target sites that may contain false positives as well as false negatives and thus may bring bias to some extent. However, target prediction algorithms are necessary when predicting functions of newly identify miRNA. In the future, a more robust and reliable miRNA clustering may be obtained when comprehensive experimental miRNA-targets data are available. We believe that with the rapid increase of experimental miRNA-mRNA deposited in TarBase and the improvement of target prediction algorithms, our method will provide high quality miRNA similarity measurement with high sensitivity and specificity.

5. Supplementary material

Supplementary material regarding this manuscript is online available in the web page of JIOMICS.

http://www.jiomics.com/index.php/jio/rt/suppFiles/21/0

Acknowledgements

This work was partially supported by the 2007 Chang-Jiang Scholars Program, Jinan University "211" Projects (Biotechnology & Bioengineering Medicine and Biomaterial & Tissue Engineering).

References


