Discovery Function Labels Based On Numerical Features Of MicroRNA Found In Plant

Authors and Affiliations

Rongsheng Zhu^{1#*}; Jinglu Liu^{1#}; Dawei Xin^{2#}; Zhanguo Zhang¹; Zhenbang Hu²; Yang Li¹; Qingshan Chen^{2*}

¹ College of Science, Northeast Agricultural University, Harbin 150030, China; ⁴ College of Agriculture, Northeast Agricultural University, Harbin 150030, China

Equal contribution

* Corresponding author, Phone: 086-0451-55191945, Fax: 086-0451-55190960, e-mail: rshzhu@126.com

Abstract: MicroRNAs(miRNAs),are a class of small endogenous non-coding RNAs that play Important post-transcriptional regulation role by degrading targeted mRNAs or repressing mRNA translation. We screened 84 miRNAs belonging to 21 conserved family from 4014 miRNAs in miRBase database which distributed in 47 plant species. Of the predicted 274 target genes,42 GO terms were found in the Gene Ontology. With the 135 numerical features which extracted by perl program, the difference significantly result of ANOVA ($P<0.001$) and multiple comparison show that the function labels G...,GC content and Helix in biological process, cellular component and biological function, respectively. Our result have suggested a potential connection between numerical features of miRNAs and the function of target genes of miRNAs.

Keywords: MicroRNA; Numerical features; ANOVA; Function labels

1. Introduction

MicroRNAs(miRNAs),are a class of small endogenous non-coding RNAs that play Important post-transcriptional regulation role and found in plant and animal. Plant miRNAs, with a mature length of approximately 22nt, recognized nearly perfect complementary binding to 3'-UTRs of target mRNA[1-3]. Recent studies have suggested that miRNA have essential function including developmental timing, cellular differentiation, proliferation, apoptosis, cell identity and fate and response to environmental stress[1, 4-8]. Since the first miRNA lin-7 and lin-4 was found in *C. elegans* have been more than two decade[9],the previous“Junk DNA”is considered to play an important role in the biological processes.

With the development of computational identification[10, 11] , majority of the new function of predicted miRNAs and their numerical features have been found. For example, previous studies have shown that unlike transfer RNAs and ribosomal RNAs, the minimum free energy(MFE) of majority miRNAs are lower than shuffled sequences[12]. The adjusted minimum free energy (AMFE)[13], and the minimum free energy index(MFEI)[14] also indicate that the high tendency of stable secondary structure of miRNAs. Although the small mature miRNAs segment with high degree of similarity across species,however, it is difficult for experimental verified and high cost[15]. Computational analysis of numerical features of miRNAs is a new research method that may bring new discoveries for the miRNAs research with a new vision.

A.thaliana, a model crop, as a major material in biological research which has a small genome,

low repetitive sequences and easily converted. Previous studies suggest that miR156 could inhibit transcripts of zygotic SPL as to prevent premature accumulation in embryonic maturation phase in *A.thaliana* [16]. The miR395 family which was predicted to code for ATP sulphurylases by target mRNAs[1, 17]. The miR162 and miR168 regulate genes required for miRNA biogenesis or activity [18, 19].All of these revealed that functional diversification of conserved miRNA which also highlighted the importance of research in numerical features. Because the numerical features reflect the structure and function significantly[20]. Whether these miRNAs with different function related to the differences of numerical feature? In this paper we aim to try to find the numerical features to explain the diversity of functions.

In this study, we used ANOVA and Multiple comparison to evaluate significant differences between numerical features from 21 conserved families and GO terms of predicted targets . The conclusions as follow:first, We got 274 predicted target genes and found that they enriched in 42 GO terms which 23 in biological process, 11 in cellular component and 8 in molecular function. Second, as the result of ANOVA, there are significant difference between numerical features and the GO terms, respectively. Third, we got function labels based on numerical features of miRNAs in plant. The secondary structure feature G... , GC content and Helix show significant difference in biological process, cellular component and biological function, respectively. Our conclusions reveal the relationship between numerical features of miRNAs and its functions.

2. Materials and methods

2.1 Data prepare

2.1.1 MiRNA genes and predicted target

We selected 4014 miRNA genes from miRBase Sequence Database , release 18 (<http://www.mirbase.org/>)[21]which covered 47 species in viridiplantae. We screened 84 miRNA genes belong to 21 conserved families in plant .All 274 predicted targets genes in *A. thaliana*, obtained from PMRD (<http://bioinformatics.cau.edu.cn/PMRD/>)[22]. Detailed information is listed in supplemental **Table 1**.

2.2 Data analysis

2.2.1 Numerical features of miRNA

We designed a perl program for assessing 135 numerical features from the selected miRNA genes. These features were roughly divided into these classes which contain frequency characteristics of nucleotides, frequency characteristics of the secondary structure matching state, energy features and entropy features. details see our previous study by Zhu et al [23].The novel numerical features of external loop, hairpin loop and multi-loop are added in this study. (see **Table 2**)

2.2.2 Go annotation

In order to detect functional categories enriched in genes targeted by selected conserved miRNA family, we analyzed the annotation of transcripts with putative target sites in The Gene Ontology(<http://www.geneontology.org/> , AmiGO version 1.8)[24]. Matlab was used to visualize the enriched the GO terms and miRNAs.(see **Fig. 1**)

2.2.3 Basic statistical methods

We used ANOVA ($P<0.001$)to evaluate significant differences in biological process , molecular

function and cellular component, respectively. Multiple comparison was used to further analysis the significant result of ANOVA.

3. Result

3.1 Screening conserved miRNA families

We got 21 conserved miRNA families from 4014 plant miRNAs in miRBase database which covered 47 plant species. The miR156 is the most conserved miRNA families which covered 34 plant species, and the relatively conserved miRNA families miR162-1/394 covered 14 plant species. Most of the conserved families are multigene families in which each miRNA product highly similar mature miRNAs. In general, more conserved miRNAs have a tendency to express at a higher levels than do relatively conserved miRNAs[25]. In the case of miR166 binding the mRNAs which encoded class III HD-ZIP transcription factors, the complementary sites are almost identical in seed plants, lycopods, and mosses, hornworts and liverworts[26, 27]. It suggested that the conserved miRNA families tend to regulate the same function in different species. (seetable.1)

3.2 Functional diversification of predicted target genes of miRNAs

In this study, we focused on the relationship between GO terms and conserved miRNAs. We mapped the predicted target genes of conserved miRNAs to the Gene Ontology (Fig.1). we find that the predicted miRNA target genes from conserved families have a significant enrichment. The result showed that there were more than 75 miRNAs enriched in 9 GO terms for each. The GO term of 'Metabolic process' which enriched the most miRNAs. On the contrary, there are only 4 miRNAs enrich in 'Pigmentation', 5 in 'Electron carrier activity', 'Enzyme regulator activity' and 'Receptor activity'. Only 2 miRNAs enriched in 'Structural molecule'. These results suggested that conserved miRNA exercise with a variety of regulatory functions in the process of life activities. As shown in Fig.1. We also detected that the members of conserved miRNA families of miR160, miR164, miR165, miR167-1, miR169-1 which have the same enrichment, respectively. There are only 9 and 8 GO terms in miR390 and miR394. This result indicated that conserved mature products of miRNAs of plant have a tendency to function consistency. Interestingly, we find that different member of conserved miRNAs families have different function category, such as miR-159 (319), miR-171-1, miR-395 and miR-397.

Fig. 1 Cluster GO terms and conserved miRNAs. The x-axis represents 84 conserved miRNAs, the y-axis represents the 42 GO terms in which the three GO classification. The number of biological processes, cellular component and molecular function is 23, 11 and 8, respectively. The white box represents present and black box represents absence. The analysis results are summarized in **table**

3.3 Significant difference among GO terms based on numerical features of miRNAs

In order to investigate whether there is a significant difference between GO terms and 135 numerical features, we used ANOVA to analyze the differences between them (Fig.2). The ANOVA ($P < 0.001$) results showed significant differences in these three GO categories. As shown in Fig. 2A, the numerical features A%, G..., stack, (G+C)%, TE (information entropy of triple nucleotides) in biological process showed significant difference with their sequence number of 1, 93, 123, 128, 134. It is noteworthy that the five of numerical features in cellular component GC

content, GCG, UGG, G., (G+C)/(A+U) belong to the sequence number 14, 42, 78, 102, 129 show in Fig.2B, which are all closely related to nucleotide frequent, secondary structure and information entropy feature. We also found some strong difference in Fig.2C, the numerical feature of Bulge loop, Helix, GA, AGC and SSE (Secondary structure paired state information entropy) may be important in the molecular function.

Fig. 2 The analysis indicated the relationship between numerical features of miRNAs and GO terms at GO categories level. A, B and C show the result of ANOVA based on three GO classification and 135 miRNA numerical features. We used the F value to evaluate the difference. Original data have been standardized in table2

Fig. 3 The significant difference results of ANOVA ($P < 0.001$) in different GO categories

3.4 Function labels based on numerical features of miRNA

Whether there is a class of numerical features which could distinguish between different biological function of the target gene in these significant difference numerical features? With this question, we extended a further analysis, based on the result of ANOVA, Multiple Comparisons used to the results of significant difference selected from ANOVA ($P < 0.001$). We find some function labels which could distinguish the function at numerical features level. The result indicated that the result of Multiple Comparisons in three GO categories all have a significant different numerical feature.

The probability density curve and the case of structure of miRNAs (Fig.4 A, D) show that miR398a/172e could distinguish the GO term 'establishment of localization' and 'metabolic process' at the numerical feature G... in biological process. Fig.4 B, E show that miR398b/160a could distinguish the GO term 'cell' and 'membrane' at the the numerical feature of GC content. Fig.4C.F show that miR162a/164a could distinguish the GO term 'catalytic activity' and 'nucleic acid binding transcription factor activity'. The secondary structure feature G..., Helix and GC content with significant result indicated that these three features can be used as the specific label on the miRNA level to distinguish the biological functions.

Fig. 4 The probability density curve and the case of structure of miRNAs. A, B and C represents the distribution of numerical of different GO terms. D. The case of G... numerical feature that the yellow arrow indicates the characteristic sites. E. The case of GC content numerical feature that the blue arrow indicates the characteristic sites. F. The case of helix numerical feature that the green arrow indicates the characteristic sites. All the mature of miRNAs showed in red. The VARNA(<http://varna.lri.fr/index.php?lang=en&page=home&css=varna>) was used to visualize the structure of miRNAs[28].

4. Discussion

Our analysis showed that the highly conserved miRNAs families have a tendency to the regulation of target genes consistently (Fig.1) The different members of conserved miRNAs families tend to enrich in the same terms. Owing to the highly conserved mature of miRNAs and the extensive complementarity with target genes, indicating to regulate the same target genes and to exercise the same function. The result also show that the same miRNA enrichment in multiple functional classification, suggesting the functional diversity of miRNA regulation. Previous studies have

shown that microRNA families were conserved between plant species and target sites were also conserved in target mRNAs[25, 29-31]. We used the measure of numerical features to evaluate conserved miRNA sequences. Because of the numerical features of miRNA reflect the structure and function.

In order to evaluate the numerical features of conserved miRNAs whether there were a Intrinsically linked to the biological activities. Analyzes using ANOVA ($P < 0.001$) at GO terms level(Fig.2) showed that there are some significant difference numerical features. The result indicated that these shared significant difference numerical features is likely related to the biological activities of miRNA target genes(Fig.3). However, those differences were not significant one is likely to be a stable part of supporting biological activities. The multiple comparisons result suggested that the numerical feature of miRNA could distinguish the GO terms. The numerical feature G... wsa the function lable of 'establishment of localization' and 'metabolic process' in biological process. Previous studies have shown that miRNA398 likely regulates plant responses and the plant stress regulatory network ,like oxidative stress, water deficit, salt stress, abscisic acid stress and so on [32]. This feature likely to be tend to affect the secondary structure of the pre-miRNA, often associated with an internal loop structure.

The numerical feature GC content, as a function lable, could distinguish the 'cell' and 'membrane' in cellular component within miR398b/160a. The miR160a participated in seed germination of *Lotus japonicas*[33]. The case of miR162a/164a showed the numerical feature Helix could distinguish 'catalytic activity' and 'nucleic acid binding transcription factor activity' in molecular function. The evidence indicated that miR162a were experimentally verified to be the targets of nodal modulator 1-like protein gene[34]. The CUC1 and CUC2 transcripts are targeted by miR164 family, encoded by MIR164a, b and c[35]. GC content and Helix are the affect factors about the secondary structure of pre-miRNA. The more GC content and Helix in single pre-miRNA the higher energies they have. Mature miRNAs are the products of inverted repeat transcripts that are precisely cleaved by RNase III enzyme(s) in the Dicer and/or Drosha protein families[36] may need more greater energies. The previous evidences have shown that unlike other non-coding RNAs, the microRNA precursors have lower MFE (minimum free energy) than random sequences[12].

In addition, the numerical feature could not accurately describe all function of miRNAs in palnt. There are many computational predicted miRNAs need to experiment to verify and some function were unknown, because of the low quality species genome. With the development of more numerical characteristics and the next-generation sequencing technology, the miRNA biogenesis its regulatory network will be more clear in the near future.

Acknowledgments

References

1. Bartel, D. P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function, *Cell*. **116**, 281-97.
2. Singh, S. K., Pal Bhadra, M., Girschick, H. J. & Bhadra, U. (2008) MicroRNAs--micro in size but macro in function, *FEBS J*. **275**, 4929-44.
3. Bartel, D. P. (2009) MicroRNAs: target recognition and regulatory functions, *Cell*. **136**, 215-33.
4. Ambros, V. (2004) The functions of animal microRNAs, *Nature*. **431**, 350-5.
5. Croce, C. M. & Calin, G. A. (2005) miRNAs, cancer, and stem cell division, *Cell*. **122**, 6-7.
6. Meltzer, P. S. (2005) Cancer genomics: small RNAs with big impacts, *Nature*. **435**, 745-6.

7. Wienholds, E. & Plasterk, R. H. (2005) MicroRNA function in animal development, *FEBS Lett.* **579**, 5911-22.
8. Zhang, B., Pan, X., Cobb, G. P. & Anderson, T. A. (2006) Plant microRNA: a small regulatory molecule with big impact, *Developmental biology.* **289**, 3-16.
9. Lee, R. C., Feinbaum, R. L. & Ambros, V. (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*, *Cell.* **75**, 843-54.
10. Frazier, T. P., Xie, F., Freistaedter, A., Burklew, C. E. & Zhang, B. (2010) Identification and characterization of microRNAs and their target genes in tobacco (*Nicotiana tabacum*), *Planta.* **232**, 1289-308.
11. Xie, F., Frazier, T. P. & Zhang, B. (2010) Identification and characterization of microRNAs and their targets in the bioenergy plant switchgrass (*Panicum virgatum*), *Planta.* **232**, 417-34.
12. Bonnet, E., Wuyts, J., Rouze, P. & Van de Peer, Y. (2004) Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences, *Bioinformatics.* **20**, 2911-2917.
13. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic Acids Res.* **31**, 3406-3415.
14. Freyhult, E., Gardner, P. P. & Moulton, V. (2005) A comparison of RNA folding measures, *BMC bioinformatics.* **6**, 241.
15. Guerra-Assuncao, J. A. & Enright, A. J. (2012) Large-scale analysis of microRNA evolution, *BMC Genomics.* **13**, 218.
16. Nodine, M. D. & Bartel, D. P. (2010) MicroRNAs prevent precocious gene expression and enable pattern formation during plant embryogenesis, *Genes & development.* **24**, 2678-92.
17. Jones-Rhoades, M. W. & Bartel, D. P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA, *Molecular cell.* **14**, 787-99.
18. Xie, Z., Kasschau, K. D. & Carrington, J. C. (2003) Negative feedback regulation of Dicer-Like1 in Arabidopsis by microRNA-guided mRNA degradation, *Curr Biol.* **13**, 784-9.
19. Vaucheret, H., Vazquez, F., Crete, P. & Bartel, D. P. (2004) The action of ARGONAUTE1 in the miRNA pathway and its regulation by the miRNA pathway are crucial for plant development, *Genes & development.* **18**, 1187-97.
20. Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P. & Anderson, T. A. (2006) Conservation and divergence of plant microRNA genes, *Plant J.* **46**, 243-59.
21. Griffiths-Jones, S., Saini, H. K., van Dongen, S. & Enright, A. J. (2008) miRBase: tools for microRNA genomics, *Nucleic Acids Res.* **36**, D154-8.
22. Zhang, Z., Yu, J., Li, D., Zhang, Z., Liu, F., Zhou, X., Wang, T., Ling, Y. & Su, Z. (2010) PMRD: plant microRNA database, *Nucleic Acids Res.* **38**, D806-13.
23. Zhu, R., Li, X. & Chen, Q. (2011) Discovering numerical laws of plant microRNA by evolution, *Biochem Biophys Res Commun.* **415**, 313-8.
24. Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Ami, G. O. H. & Web Presence Working, G. (2009) AmiGO: online access to ontology and annotation data, *Bioinformatics.* **25**, 288-9.
25. Axtell, M. J. & Bowman, J. L. (2008) Evolution of plant microRNAs and their targets, *Trends in plant science.* **13**, 343-9.
26. Floyd, S. K. & Bowman, J. L. (2004) Gene regulation: ancient microRNA target sequences in plants, *Nature.* **428**, 485-6.

27. Floyd, S. K., Zalewski, C. S. & Bowman, J. L. (2006) Evolution of class III homeodomain-leucine zipper genes in streptophytes, *Genetics*. **173**, 373-88.
28. Darty, K., Denise, A. & Ponty, Y. (2009) VARNA: Interactive drawing and editing of the RNA secondary structure, *Bioinformatics*. **25**, 1974-5.
29. Addo-Quaye, C., Snyder, J. A., Park, Y. B., Li, Y. F., Sunkar, R. & Axtell, M. J. (2009) Sliced microRNA targets and precise loop-first processing of MIR319 hairpins revealed by analysis of the *Physcomitrella patens* degradome, *Rna*. **15**, 2112-21.
30. Li, Y. F., Zheng, Y., Addo-Quaye, C., Zhang, L., Saini, A., Jagadeeswaran, G., Axtell, M. J., Zhang, W. & Sunkar, R. (2010) Transcriptome-wide identification of microRNA targets in rice, *Plant J*. **62**, 742-59.
31. Pantaleo, V., Szittyá, G., Moxon, S., Miozzi, L., Moulton, V., Dalmay, T. & Burgyan, J. (2010) Identification of grapevine microRNAs and their targets using high-throughput sequencing and degradome analysis, *Plant J*. **62**, 960-76.
32. Zhu, C., Ding, Y. & Liu, H. (2011) MiR398 and plant stress responses, *Physiologia plantarum*. **143**, 1-9.
33. Hu, J., Zhang, H. & Ding, Y. (2013) Identification of conserved microRNAs and their targets in the model legume *Lotus japonicus*, *Journal of biotechnology*.
34. Wan, L. C., Zhang, H., Lu, S., Zhang, L., Qiu, Z., Zhao, Y., Zeng, Q. Y. & Lin, J. (2012) Transcriptome-wide identification and characterization of miRNAs from *Pinus densata*, *BMC Genomics*. **13**, 132.
35. Nikovics, K., Blein, T., Peaucelle, A., Ishida, T., Morin, H., Aida, M. & Laufs, P. (2006) The balance between the MIR164A and CUC2 genes controls leaf margin serration in *Arabidopsis*, *The Plant cell*. **18**, 2929-45.
36. Axtell, M. J., Westholm, J. O. & Lai, E. C. (2011) Vive la difference: biogenesis and evolution of microRNAs in plants and animals, *Genome biology*. **12**, 221.

The numerical feature distribution of G... in biological processes

bioRxiv preprint doi: <https://doi.org/10.1101/770875>; this version posted September 16, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The numerical feature distribution of GC content in molecular composition

The numerical feature distribution of helix in biological function