microRPM: a system for novel microRNA prediction

Yi-Fan Chiang-Hsieh
College of Biosciences and Biotechnology,
Institute of Tropical Plant Sciences,
National Cheng Kung University, Tainan 70101, Taiwan

Wen-Chi Chang
College of Biosciences and Biotechnology,
Institute of Tropical Plant Sciences,
National Cheng Kung University, Tainan 70101, Taiwan

***Abstract*—MicroRNAs (miRNAs) are endogenous non-coding small RNAs, which play important roles in post-transcriptional regulation of gene expression via mRNA cleavage or translation inhibition. Several machine learning-based approaches have been developed to identify novel miRNAs from next generation sequencing (NGS) data. Usually, precursor/genomic sequences were essentially required as references for most methods. However, the non-availability of genomic sequences is often a limitation in miRNA discovery in non-model organisms. It is thus necessary to develop a systematic approach to determine novel miRNAs without reference sequences. In this study, an effective method was developed to identify miRNAs from non-model plants based only on NGS datasets with several significant structure-related features by using support vector machine (SVM) algorithm. The accuracy of the independent test reached 96.61 % and 93.04 % for dicots (*Arabidopsis*) and monocots (rice), respectively. This novel approach was also complied as a user-friendly program, called microRPM (microRNA Prediction Model), and freely available at http://microRPM.itps.ncku.edu.tw.**

Keywords—microRNA; next-generation sequencing; support vector machine

I. RESULTS

MiRNAs are important gene regulators in different biological pathways and species, and a growing number of studies thus focus on identifying miRNAs from non-model organisms. However, most of the existing miRNA identification tools depend on identifying miRNA precursors, and the need for reference sequences means that they cannot be used with organism that lack whole genome/transcriptome sequences. An effective miRNA prediction model was therefore developed in this study for non-model organisms. Only a small RNA sequencing dataset is required in microRPM for novel miRNA identification. As we known, miR-BAG [1], miRDeep2 [2, 3], and miReader [4] are three microRNA prediction methods that discover microRNA from NGS data currently. Among them, miReader [4] is only one algorithm that does not need sequenced genome to predict microRNAs. However, the program of miR-BAG provided on the web were crashed. Therefore, Table 1 only illustrates the comparisons among miReader, miRDeep and microRPM. Due to the model running time of miReader was time-consuming, and the performance was bad (accuracy=56.95%) when using *Arabidopsis* dataset, we did not test rice on miReader. Significantly, microRPM shown the best performance in all datasets. Especially, the sensitivity, specificity, precision, and accuracy reached 98.15 %, 95.07 %, 95.22 %, and 96.61 % when using *Arabidopsis* dataset based on reference-non-required model of microRPM, respectively (Table 1). The effective model is provided in microRPM and freely available at http://microRPM.itps.ncku.edu.tw. We hope that the microRPM can advance the miRNA researches.

Table 1. Performance comparisons among miReader, miRDeep2, and microRPM.

Model	Test dataset	Organism	Sensitivity (%)	Specificity (%)	Precision (%)	Accuracy (%)	Reference sequence
miReader [4]	GSM850278	*Arabidopsis*	14.12	99.77	98.4	56.95	No
miRdeep2 [3]	GSM850278	*Arabidopsis*	87.67	86.02	86.25	86.85	Yes
	GSM816687	Rice	85.02	79.19	80.34	82.11	Yes
microRPM[a]	GSM850278	*Arabidopsis*	98.15	95.07	95.22	96.61	No
	GSM816687	Rice	94.91	91.17	91.49	93.04	No

[a]Reference-non-required model. Training feature: Triplet element.

ACKNOWLEDGMENT

The authors would like to thank the Ministry of Science and Technology (MOST 105-2311-B-006 -004 –MY3 and MOST 104-2313-B-006-002) and Academia Sinica (Innovative Translational Agricultural Research Grant) of the Republic of China for financially supporting this research.

REFERENCES

[1] A. Jha, R. Chauhan, M. Mehra, H. R. Singh, and R. Shankar, "miR-BAG: bagging based identification of microRNA precursors," *PLoS One,* vol. 7, p. e45782, 2012.

[2] M. R. Friedlander, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel*, et al.*, "Discovering microRNAs from deep sequencing data using miRDeep," *Nat Biotechnol,* vol. 26, pp. 407-15, Apr 2008.

[3] S. D. Mackowiak, "Identification of novel and known miRNAs in deep-sequencing data with miRDeep2," *Curr Protoc Bioinformatics,* vol. Chapter 12, p. Unit 12 10, Dec 2011.

[4] A. Jha and R. Shankar, "miReader: Discovering Novel miRNAs in Species without Sequenced Genome," *PLoS One,* vol. 8, p. e66857, 2013.