BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 10 2011, pages 1436–1437
doi:10.1093/bioinformatics/btr148

Sequence analysis

Advance Access publication March 23, 2011

Prediction of novel pre-microRNAs with high accuracy through boosting and SVM

Yuanwei Zhang[1,†], Yifan Yang[2,†], Huan Zhang[1,†], Xiaohua Jiang[1,†], Bo Xu[1], Yu Xue[3], Yunxia Cao[4], Qian Zhai[5], Yong Zhai[5], Mingqing Xu[6], Howard J. Cooke[1,7] and Qinghua Shi[1,*]

[1]Department of Life science, Hefei National Laboratory for Physical Sciences, Microscale and School of Life Sciences, [2]Department of Statistics and Finance, University of Science and Technology of China, Hefei 230027, [3]Department of Systems Biology, Huazhong University of Science and Technology, Wuhan 430074, [4]Reproductive Medicine Center, The First Affiliated Hospital, Anhui Medical University, Hefei, [5]Anhui Research Institute for Family Planning, Hefei, Anhui 230031, [6]Department of Obstetrics and Gynecology, The People Hospital, Luan, China and [7]MRC Human Genetics Unit and Institute of Genetics and Molecular Medicine, Western General Hospital, Edinburgh, UK

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: High-throughput deep-sequencing technology has generated an unprecedented number of expressed short sequence reads, presenting not only an opportunity but also a challenge for prediction of novel microRNAs. To verify the existence of candidate microRNAs, we have to show that these short sequences can be processed from candidate pre-microRNAs. However, it is laborious and time consuming to verify these using existing experimental techniques. Therefore, here, we describe a new method, miRD, which is constructed using two feature selection strategies based on support vector machines (SVMs) and boosting method. It is a high-efficiency tool for novel pre-microRNA prediction with accuracy up to 94.0% among different species.

Availability: miRD is implemented in PHP/PERL+MySQL+R and can be freely accessed at http://mcg.ustc.edu.cn/rpg/mird/mird.php.

Contact: qshi@ustc.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 23, 2010; revised on February 22, 2011; accepted on March 9, 2011

1 INTRODUCTION

MicroRNAs, short RNAs (~20–25 nt) that perform their functions by guiding mRNA transcriptional degradation or translational suppression (Carthew and Sontheimer, 2009; Wu *et al.*, 2010), have various functions in organ development. For example, they mediate switching of chromatin remodeling complexes in neural development and participate in transcriptional circuits that control skeletal muscle gene expression and embryonic development (Chen *et al.*, 2006; Yoo *et al.*, 2009). Increasingly, evidence demonstrates that they can also function either as tumor suppressors or oncogenes (Bonci *et al.*, 2008; He *et al.*, 2005). Although more microRNA functions are being discovered, there are still many novel microRNAs whose functions remain to be elucidated.

To predict novel pre-microRNAs in specific animals and plants, comparative genomic-based methods have been developed, including MiRscan, MIRcheck, miRAlign and MIRFINDER (Huang *et al.*, 2007; Laufs *et al.*, 2004; Lim *et al.*, 2003; Wang *et al.*, 2005). Although these tools are capable of identifying phylogenetically conserved stem–loop precursor RNAs, they do not work well when applied to genomes that lack close homologs. Recently, several machine learning-based algorithms have been introduced to predict microRNAs (Hsieh *et al.*, 2010; Jiang *et al.*, 2007; Xu *et al.*, 2008). In addition, some modified no-learning methods, based on simple and widely accepted principles, have been used, where pre-microRNAs are detected by manually choosing the optimal filter (Quail *et al.*, 2008). Although these methods have simple structures and flexibility, their performance can still be improved by combination with machine-learning methods.

In this study, we developed a novel machine-learning tool, named miRD (microRNA Detection) for accurate and efficient detection of novel pre-microRNAs. There are two sets of features and each was used to build a support vector machines (SVMs) model. (Vapnik, 2000). A boosting method was then applied to combine the two independent SVM models (Freund and Schapire, 1996). We tested the performance of miRD on a small RNA deep-sequencing dataset of human fetal ovary. Altogether, 92 novel candidate pre-microRNAs were predicted by miRD and were sorted in descending order of the predicted probability (Supplementary Table S8). To confirm the expression of the predicted pre-microRNA, the top 16 candidates were selected for further experimental validation. Surprisingly, all these selected pre-microRNA from human fetal ovary were verified by real-time PCR (Supplementary Fig. S5). miRD was more efficient than any published algorithm (tripleSVM, MIReNA), with its *AC* and *MCC* reaching 94.0% and 0.872, respectively (Supplementary Table S6).

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first four authors should be regarded as joint First Authors.

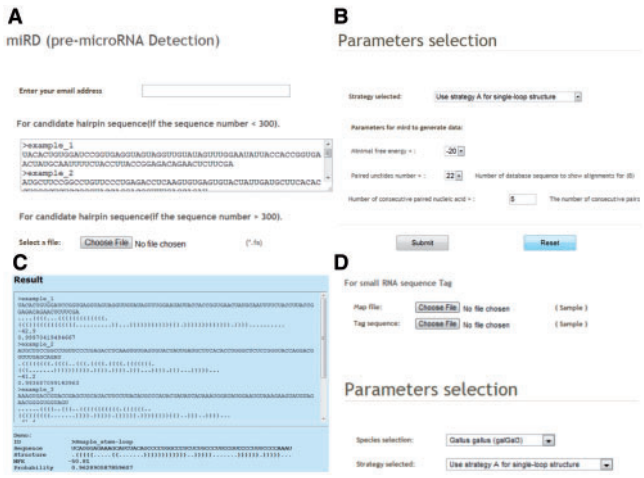

Fig. 1. Application of miRD web server. (**A**) Input page of miRD web server. (**B**) Parameters selection for miRD prediction. (**C**) Prediction result of miRD. (**D**) The page for uploading small RNA deep-sequencing file.

2 METHODS

Mature microRNAs can be processed from pre-microRNAs with two different types of secondary structure: single-stem pre-microRNAs, which typically have one paired stem and one symmetrical loop, and multi-stem pre-microRNAs, which have several symmetrical or asymmetrical loops (Supplementary Fig. S1). We collected 14 197 hairpin precursor microRNAs of 133 species and 8494 pseudo pre-microRNA-like hairpins to construct the training dataset. Different biological features were considered for single- and multiple-stem pre-microRNAs. Strategy A detects the multi-stem pre-microRNAs using a novel method. Since the standard stem–loop structure contains a stable paired stem and a symmetrical loop, the differences between the standard stem–loop and a candidate hairpin structure can be extracted as a measure of the possibility that a candidate hairpin is a real pre-microRNA (Supplementary Fig. S2). Strategy B characterizes the single-stem type, and 59 features were selected based on sequence and structure composition (Supplementary Table S1). The kernel function of the SVM applied in this model was chosen by the try-and-test strategy, and the radial basis kernel was finally selected. A detailed description of the method is provided in the Supplementary Material. The performance of miRD and its comparison with other tools are shown in Supplementary Table S6.

3 APPLICATION

miRD has two applications in pre-microRNA prediction: (i) giving the probability of a candidate pre-microRNA to be a real one; and (ii) extracting the probable pre-microRNAs from deep-sequencing data. To demonstrate how to use the miRD web server, we submitted a sample data in FASTA format as an example (Fig. 1A). The default parameters for feature selecting strategy, minimal free energy of the secondary structure and paired nucleic acid number were selected (Fig. 1B). As shown in Figure 1C, the prediction results contain the probabilities of submitted candidate sequence to be real pre-microRNA, information such as the secondary structure and the minimum free energy. Users can also submit files containing short-read sequences or information of read locations on the genome (Fig. 1D). The candidate pre-miRNA sequences will be extracted from the submitted data. miRD will then analyze these candidate pre-miRNA sequences and return results as shown in Figure 1C.

ACKNOWLEDGEMENTS

We would like to thank Heli Hou and Zexian Liu for critical reading of the manuscript.

Funding: National Basic Research Program (2007CB947401, 2011CB944501) of China (973); Knowledge Innovation Program of the Chinese Academy of Sciences (KSCX2-EW-R-07).

Conflict of Interest: none declared.

REFERENCES

Bonci,D. *et al.* (2008) The miR-15a-miR-16-1 cluster controls prostate cancer by targeting multiple oncogenic activities. *Nat. Med.*, **14**, 1271–1277.

Carthew,R.W. and Sontheimer,E.J. (2009) Origins and mechanisms of miRNAs and siRNAs. *Cell*, **136**, 642–655.

Chen,J.F. *et al.* (2006) The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat. Genet.*, **38**, 228–233.

Freund,Y. and Schapire,R. (1996) Experiments with a new boosting algorithm. *Machine Learning-International workshop then confernce*. Citeseer, pp. 148–156.

He,L. *et al.* (2005) A microRNA polycistron as a potential human oncogene. *Nature*, **435**, 828–833.

Hsieh,C.H. *et al.* (2010) Predicting microRNA precursors with a generalized Gaussian components based density estimation algorithm. *BMC Bioinformatics*, **11** (Suppl. 1), S52.

Huang,T.H. *et al.* (2007) MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans. *BMC Bioinformatics*, **8**, 341.

Jiang,P. *et al.* (2007) MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. *Nucleic Acids Res.*, **35**, W339–W344.

Laufs,P. *et al.* (2004) MicroRNA regulation of the CUC genes is required for boundary size control in Arabidopsis meristems. *Development*, **131**, 4311–4322.

Lim,L.P. *et al.* (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev.*, **17**, 991–1008.

Quail,M.A. *et al.* (2008) A large genome center's improvements to the Illumina sequencing system. *Nat. Methods*, **5**, 1005–1010.

Vapnik,V. (2000) *The Nature of Statistical Learning Theory*. Springer, Verlag New York, Inc.

Wang,J.W. *et al.* (2005) Control of root cap formation by MicroRNA-targeted auxin response factors in Arabidopsis. *Plant Cell*, **17**, 2204–2216.

Wu,H. *et al.* (2010) Genome-wide analysis reveals methyl-CpG-binding protein 2-dependent regulation of microRNAs in a mouse model of Rett syndrome. *Proc. Natl Acad. Sci. USA*, **107**, 18161–18166.

Xu,Y. *et al.* (2008) MicroRNA prediction with a novel ranking algorithm based on random walks. *Bioinformatics*, **24**, i50–i58.

Yoo,A.S. *et al.* (2009) MicroRNA-mediated switching of chromatin-remodelling complexes in neural development. *Nature*, **460**, 642–646.