

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 Xu¹, Yu Xue³, Yunxia Cao⁴, Qian Zhai⁵, Yong Zhai⁵, Mingqing Xu⁶, Howard J. Cooke^{1,7} and Qinghua Shi^{1,*}

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

