SusMiRTrain: ab initio SVM classifier for porcine microRNA precursor prediction

Peng-Fang Zhou , Fei Zhang, Zhen-Hua Zhao, Wen-Qian Zhang, Wen-Chao Lin, Yang Zhang, De-Li Zhang §

Investigation Group of Molecular Virology, Immunology, Oncology & Systems Biology, Center for Bioinformatics, and Research Laboratory of Virology, Immunology & Bioinformatics, College of Veterinary Medicine, Northwest A & F University, Yangling 712100, Xi'an City, Shaanxi Province, P.R.China
§ Corresponding author
zhangdeli@tsinghua.org.cn

Abstract: MicroRNA (miRNA), which is short non-coding RNA, plays important roles in almost all biological processes examined. Several classifiers have been applied to predict humans, mice and rats precursor miRNAs (pre-miRNAs), but no classifier is applied to classify porcine pre-miRNAs only based on the porcine pre-miRNAs because of the little known miRNA component in the porcine genome. Here, we developed a novel classifier, called SusMiRTrain, to predict porcine pre-miRNAs. Trained on 60 porcine pre-miRNAs and 65 pseudo porcine hairpins, SusMiRTrain achieved 86.4% (5-fold cross-validation accuracy) and 0.9144 (ROC score). Tested on the remaining 14 porcine pre-miRNAs and 1000 pseudo hairpins, it reported 100% (sensitivity), 87.3% (specificity) and 87.5% (accuracy),respectively. Furthermore, a Java package, called SusMiRPred, was developed to filter out the short sequences which have not the pre-miRNAs structure features and to extract features for porcine pre-miRNAs prediction. [Researcher. 2010;2(2):61-63]. (ISSN: 1553-9865).

Key words: MicroRNA; Swine; SVM

1. Introduction

MicroRNAs (miRNAs) are short RNAs (~20-22nt) that can regulate gene expression by binding to the mRNAs at the post-transcriptional level in eukaryotes (Bartel 2004; Mendell 2008). These short RNAs are generally derived from long, primary transcripts (pri-miRNA) which are processed into fold-back precursor miRNA (pre-miRNAs) with characteristic stem-loop RNA structures (Lee, Kim et al. 2004). The pre-miRNAs are cleaved into ~22 nt duplexes which then unwind, leaving the mature miRNA sequence preferentially incorporated into RNA-induced silencing complex (RISC) to regulate protein-coding gene expression (Khvorova, Reynolds et al. 2003; Schwarz, Hutvagner et al. 2003; Bartel 2004). To date, miRNAs have been shown to play critical roles in almost all biological processes examined, such as control of developmental timing, cell fate specification, limb development, apoptosis, angiogenesis, fat metabolism, insulin secretion, and even cancer (Lee, Feinbaum et al. 1993; Mendell 2008) .

Many miRNA families are conserved among the vertebrate animals. However, many of the new miRNAs recently discovered are not conserved beyond mammals, and ~10% are taxon specific (Berezikov, van Tetering et al. 2006). Comparative approaches suffer lower sensitivity in detecting novel pre-miRNAs without known homology pre-miRNAs (Berezikov, Guryev et al. 2005). But all of the porcine miRNA sequences in the latest miRBase were computationally predicted on the basis of sequence homology to known miRNAs from other species (Wernersson, Schierup et al. 2005; Huang, Zhu et al. 2008; McDaneld, Smith et al. 2009; Reddy, Zheng et al. 2009). The current release of miRBase contains only 72 porcine miRNA sequences while 718 human miRNAs, 595 mouse miRNAs, and 330 rat miRNAs have been identified (Griffiths-Jones, Saini et al. 2008).

Here, we developed a Java package: SusMiRPred, to filter out the short sequences according to the pre-miRNA structure features and to convert the structures to triplet-SVM features (Xue, Li et al. 2005). Then, we trained a SVM model called SusMiRTrain based on porcine pre-miRNA using libSVM package. Trained on 60 pre-miRNAs in miRBase and 65 pseudo pre-miRNAs, the SusMiRTrain achieved 86.4% (5-fold cross-validation accuracy) and 0.9144 (ROC score).

2. Materials and Methods

We downloaded all 77 porcine miRNA sequences from miRBase version 14.0 [13]. Genomic sequences were from NCBI of June 2009 and were downloaded

from the Ensemble FTP site (ftp://ftp.ncbi.nlm.nih.gov/genomes/Sus_scrofa/). The protein coding regions (CDS) sequences were downloaded from the Ensemble FTP site

(ftp://ftp.ncbi.nlm.nih.gov/genomes/Sus_scrofa/RNA).

We used RNAfold [14] version 1.7 with default parameters to predict RNA secondary structures.

Three datasets were built to train SVM and to evaluate the classifier performance. One was training set called it as the "TRAINING-SET", and two were test sets named as the "CDS-SET" and "MIRBASE-SET" according to the ways we collected. The porcine pre-miRNAs whose secondary structures do not contain multiple loops were considered, which gave us 74 pre-miRNAs, covering more than 96% of all the reported porcine pre-miRNAs. We extracted 60 pre-miRNAs from them as one part of "TRAINING-SET" set and the remaining 14 pre-miRNAs composed of the "MIRBASE-SET" set.

We filtered the CDS sequences keeping the length distribution of the extracted segments with that of porcine pre-miRNAs. The criteria for selecting the pseudo-miRNAs from the segments are: minimum of 18 base pairings on the stem of the hairpin structure (included the GU wobble pairs), maximum of -15 kcal/mol free energy of the secondary structure, and no multiple loops. These criteria ensured that the extracted pseudo pre-miRNAs were similar to real pre-miRNAs according to the widely accepted characteristics. As most of reported miRNAs are located in the un-translated regions or intergenic regions, we took the hairpins collected from CDS as examples of pseudo pre-miRNAs. Totally, 8494 pre-miRNA-like hairpins (pseudo pre-miRNAs) were collected in this dataset. We randomly selected 65 pseudo pre-miRNAs from them as the other part of the "TRAINING-SET" set, and 1000 pre-miRNA-like hairpins from remaining pseudo pre-miRNAs were extracted as "CDS-SET" set.

Trained on "TRAINING-SET" set using libSVM package with triplet features(Xue, Li et al. 2005), a training model called "SusMiRTrain" was achieved for porcine pre-miRNA prediction. Then, we used "CDS-SET" and "MIRBASE-SET" set to evaluate the SusMiRTrain performance.

3. Results and Discussion

.

Trained on the "TRAINING-SET" set, SusMiRTrain achieved 86.4% (5-fold cross-validation accuracy) and 0.9144 (ROC score) (Filgure 1). Tested on the "CDS-SET" set and "MIBASE-SET" set, it reported 100% (sensitivity), 87.3% (specificity) and 87.5% (accuracy), respectively. The good performance of SusMiRTrain showed that it was available for the prediction of porcine pre-miRNAs. Since the little number of porcine pre-miRNAs in miRBase, researchers predicted porcine pre-miRNAs on basis of porcine pre-miRNAs and other species pre-miRNAs, which also was on the basis of sequence homology to known miRNAs from other species (Xue, Li et al. 2005). Here, we proposed an ab initio method for porcine pre-miRNAs prediction without known homology pre-miRNAs.

There are four layers in the SusMiRPred. First, SusMiRPred filtered short sequences with mfe >= -15kcal/mol and stem < 18. Second, it filtered the multiloop sequences. Then, the sequences with the length of inter < 9nt; the length of bulge < 6nt; the numbers of inter and bulge < 10; the numbers of inter < 8 and the numbers of bulge < 6 was remained. At last, triplet features were extracted from the remaining sequences to predict pre-miRNAs.

We scanned for the genome sequences using a 100-nt query window with 10-nt increments at a time. Sequences with potential hairpin-like structures were extracted as candidate miRNA precursors. A GC content requirement of 30% to 75% for the 100-nt query sequences was applied. Additionally, low-complexity sequences, such as those with dinucleotides repeated >= 4 times (for example, ATATATAT), trinucleotides repeated >=3 times (for example, ATGATGATG), or any single nucleotide repeated >6 times (for example, AAAAAAA), were removed using a repeat filter. Such sequences have been observed little in known miRNAs. The resulting candidate miRNA precursors were analyzed with the program RNAfold for secondary structure prediction. Then, we used the SusMiRPred to filter the pre-miRNAs structures. Using the SVM classifier of SusMiRTrain to predict the candidate hairpins, we got the predicted candidate results of the porcine pre-miRNAs. 9250 porcine pre-miRNAs candidates were found by scanning the whole porcine genome using SusMiRPred and SusMiRTrain (Figure 2).

4. Conclusions

All of porcine microRNAs were computationally predicted on the basis of sequence homology to known miRNAs from other species. There were no ab initio approaches to predict the porcine pre-microRNAs. In this article, we proposed a Java package: SusMiRPred, and a training model: SusMiRTrain to predict the species specific pre-microRNAs. Only trained on porcine pre-microRNAs, SusMiRTrain achieved the accuracy about 87.5% for distinguishing real vs. pseudo porcine pre-miRNAs. The good performance showed that SusMiRTrain was available for porcine pre-miRNAs prediction.

Acknowledgement:

This item is supported by the NorthwestA & F University Foundation for Attracting Foreign Personnel (No. 01140406) from the“985”Project for Higher Educationand and the National Natural Science Foundation of China (No.30270342).

References

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell 116(2): 281-97.

Berezikov, E., V. Guryev, et al. (2005). "Phylogenetic shadowing and computational identification of human microRNA genes." Cell 120(1): 21-4.

Berezikov, E., G. van Tetering, et al. (2006). "Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis." Genome Res 16(10): 1289-98.

Griffiths-Jones, S., H. K. Saini, et al. (2008). "miRBase: tools for microRNA genomics." Nucleic Acids Res 36(Database issue): D154-8.

Huang, T. H., M. J. Zhu, et al. (2008). "Discovery of porcine microRNAs and profiling from skeletal muscle tissues during development." PLoS One 3(9): e3225.

Khvorova, A., A. Reynolds, et al. (2003). "Functional siRNAs and miRNAs exhibit strand bias." Cell 115(2): 209-16.

Lee, R. C., R. L. Feinbaum, et al. (1993). "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." Cell 75(5): 843-54.

Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." EMBO J 23(20): 4051-60.

McDaneld, T. G., T. P. Smith, et al. (2009). "MicroRNA transcriptome profiles during swine skeletal muscle development." BMC Genomics 10: 77.

Mendell, J. T. (2008). "miRiad roles for the miR-17-92 cluster in development and disease." Cell 133(2): 217-22.

Reddy, A. M., Y. Zheng, et al. (2009). "Cloning, characterization and expression analysis of porcine microRNAs." BMC Genomics 10: 65.

Schwarz, D. S., G. Hutvagner, et al. (2003). "Asymmetry in the assembly of the RNAi enzyme complex." Cell 115(2): 199-208.

Wernersson, R., M. H. Schierup, et al. (2005). "Pigs in sequence space: a 0.66X coverage pig genome survey based on shotgun sequencing." BMC Genomics 6(1): 70.

Xue, C., F. Li, et al. (2005). "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine." BMC Bioinformatics 6: 310.

Figure Legends:

Figure 1. ROC score

Figure 2. Flowchart of the porcine pre-miRNA prediction procedure

Date:

02/05/2010