Chapter 7
MicroRNA Precursor Prediction
Using SVM with RNA Pairing Continuity Feature

Huan Yang  
Jilin University, China

Yan Wang  
Jilin University, China

Trupti Joshi  
University of Missouri, USA

Dong Xu  
University of Missouri, USA

Shoupeng Yu  
Changchun Institute of Technology, China

Yanchun Liang  
Jilin University, China

ABSTRACT

MicroRNAs (miRNAs) are endogenous single-stranded non-coding RNAs of ~22 nucleotides in length and they act as post-transcriptional regulators in bacteria, animals and plants. Almost all current methods for computational prediction of miRNAs use hairpin structure and minimum of free energy as characteristics to identify putative pre-miRNAs from a pool of candidates. We discovered a new effective feature named “basic-n-units” (BNU) to distinguish pre-miRNAs from pseudo ones. This feature describes pairing continuity of RNA secondary structure. Simulation results show that a classification method, called Triplet-SVM-classifier, achieved an accuracy of 97.24% when this BNU feature was used. This is a 3% increase caused solely by adding this new feature. We anticipate that this BNU feature may increase the accuracy for most classification methods.

DOI: 10.4018/978-1-60960-064-8.ch007

Copyright © 2011, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
INTRODUCTION

A large number of microRNAs (miRNAs) have been identified through a variety of methods. But the number of miRNAs discovered is far less than that in the forecast level of theory. It is difficult to identify true pre-miRNA from their sequences. Pre-miRNA prediction methods include those that depend on sequence alignment of homologous fragments as well as those that use scoring functions based on sequences and structural characteristics. Machine learning prediction methods, which are widely applied in bioinformatics research, are becoming a common tool for pre-miRNA predictions. However, the accuracies of existing methods show that there is still room for improvement in distinguishing true pre-miRNAs from other hairpin sequences with similar stem-loops (pseudo pre-miRNAs). Because most of these algorithms rely on the characteristics of stem-loop hairpin structures, in which a large amount of pseudo pre-miRNAs can be found, it becomes increasingly important to identify more unique and essential features that define true pre-miRNAs. In our study, we discovered a new effective structural feature of pre-miRNAs whose inclusion in classification studies has improved the prediction accuracy.

BACKGROUND

MicroRNAs (miRNAs), one of the non-coding RNA families, are a class of endogenous, single-stranded, small (19-27nt) nucleic acids that have an extremely conserved structure. In the nucleus, a gene encoding miRNA is first transcribed into a pri-microRNA, which is cut into pre-miRNA with a hairpin (stem-loop) structure (Bartel, 2004) of about 70nt in length using an enzyme called “Drosha RNase” (Lee et al., 2003). The mature miRNAs are derived from cleavage of pre-miRNA by the Dicer enzyme out of the nucleus. miRNAs are essential in animal and plant development (Bartel, 2004), stress response in plants (Bari et al., 2006), and various diseases including cancers (Blenkiron et al., 2007). They also play key roles in the regulation and control of a variety of the metabolic processes of different organisms (Hua & Xiao, 2005).

Since 2003, thousands of miRNAs have been experimentally identified. At the same time, a variety of prediction methods have been developed too. For examples, miRScan relies on the observation that the known miRNAs are derived from phylogenetically conserved stem-loop precursor RNAs (Lim et al., 2003). Xue et al. (2005) proposed an SVM-based method for classification of real and pseudo pre-miRNAs based on local contiguous structure-sequence composition feature. Nam et al. (2005; 2006) constructed hidden Markov models (HMM) to search for distant homologs of miRNA families. Yousef et al. (2006) used a Naïve Bayes classifier along with the integration of data from multiple species to predict miRNA genes. Jiang et al. (2007) tried to construct a classifier, called MiPred, between per-miRNA and pseudo miRNA using Random forest and P-value. As suggested by Helvik et al. (2007), the miRNA gene prediction methods can also be improved by reliable predictions of Drosha-processing sites. Aiming at identifying miRNAs from genomes with a few known miRNA, Xu et al. (2008) proposed and developed a novel miRNA prediction method, miRank, with a novel ranking algorithm based on random walks. RNAmicro (Hertel & Stadler, 2006) was designed to classify the surveys of large-scale comparative genomics for predicting putative RNAs. Recently, Ahmed et al. (2009) demonstrated that guide and passenger strands of miRNA precursors can be distinguished using nucleotide sequence and secondary structures.

However, new problems also occurred: the features extracted using the existing forecasting methods are limited, and the range of species that can be predicted is small. Some methods can only predict one type of pre-miRNAs from either animal, plant or bacteria; some can predict
Related Content

Dis2PPI: A Workflow Designed to Integrate Proteomic and Genetic Disease Data
[www.igi-global.com/article/dis2ppi-workflow-designed-integrate-proteomic/77811?camid=4v1a](www.igi-global.com/article/dis2ppi-workflow-designed-integrate-proteomic/77811?camid=4v1a)

Social Network Sites and Their Role in the Sharing of Health Information
[www.igi-global.com/chapter/social-network-sites-their-role/64075?camid=4v1a](www.igi-global.com/chapter/social-network-sites-their-role/64075?camid=4v1a)

Spatial Uncertainty Analysis in Ecological Biology
[www.igi-global.com/article/spatial-uncertainty-analysis-ecological-biology/78389?camid=4v1a](www.igi-global.com/article/spatial-uncertainty-analysis-ecological-biology/78389?camid=4v1a)

Unsupervised Feature Selection
[www.igi-global.com/chapter/unsupervised-feature-selection/53905?camid=4v1a](www.igi-global.com/chapter/unsupervised-feature-selection/53905?camid=4v1a)