Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

Effective sample selection for classification of pre-miRNAs

K. Han

School of Computer and Information Engineering,
Harbin University of Commerce, Harbin, Heilongjiang, China

Corresponding author: K. Han
E-mail: hanke@hrbcu.edu.cn

Genet. Mol. Res. 10 (1): 506-518 (2011)
Received September 8, 2010
Accepted October 26, 2010
Published March 29, 2011
DOI 10.4238/vol10-1gmr1054

ABSTRACT. To solve the class imbalance problem in the classification of pre-miRNAs with the *ab initio* method, we developed a novel sample selection method according to the characteristics of pre-miRNAs. Real/pseudo pre-miRNAs are clustered based on their stem similarity and their distribution in high dimensional sample space, respectively. The training samples are selected according to the sample density of each cluster. Experimental results are validated by the cross-validation and other testing datasets composed of human real/pseudo pre-miRNAs. When compared with the previous method, *microPred*, our classifier *miRNAPred* is nearly 12% more accurate. The selected training samples also could be used to train other SVM classifiers, such as *triplet-SVM*, *MiPred*, *miPred*, and *microPred*, to improve their classification performance. The sample selection algorithm is useful for constructing a more efficient classifier for the classification of real pre-miRNAs and pseudo hairpin sequences.

Key words: Sample selection; Class imbalance; Pre-miRNA; Information gain; Conservation