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

Mining SNPs from EST sequences using filters and ensemble classifiers

J. Wang[1]*, Q. Zou[2]* and M.Z. Guo[1]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang, P.R. China
[2]School of Information Science and Technology, Xiamen University, Xiamen, Fujian, P.R. China

*These authors contributed equally to this study.
Corresponding author: M.Z. Guo
E-mail: maozuguo@hit.edu.cn

Genet. Mol. Res. 9 (2): 820-834 (2010)
Received January 11, 2010
Accepted February 8, 2010
Published May 4, 2010
DOI 10.4238/vol9-2gmr765

ABSTRACT. Abundant single nucleotide polymorphisms (SNPs) provide the most complete information for genome-wide association studies. However, due to the bottleneck of manual discovery of putative SNPs and the inaccessibility of the original sequencing reads, it is essential to develop a more efficient and accurate computational method for automated SNP detection. We propose a novel computational method to rapidly find true SNPs in public-available EST (expressed sequence tag) databases; this method is implemented as SNPDigger. EST sequences are clustered and aligned. SNP candidates are then obtained according to a measure of redundant frequency. Several new informative biological features, such as the structural neighbor profiles and the physical position of the SNP, were extracted from EST sequences, and the effectiveness of these features was demonstrated. An ensemble classifier, which employs a carefully selected feature set, was included for the imbalanced training data. The sensitivity and specificity of our method both exceeded 80% for human genetic data in the cross validation. Our method enables detection of SNPs from the user's own EST dataset and can be used on species for which there is no genome data. Our tests showed that this method can effectively guide

SNP discovery in ESTs and will be useful to avoid and save the cost of biological analyses.

Key words: Single nucleotide polymorphisms; Expressed sequence tag; Filter; Ensemble classifier; SNPDigger

