GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

High-accuracy splice site prediction based on sequence component and position features

J.L. Li[1,2]*, L.F. Wang[1]*, H.Y. Wang[3], L.Y. Bai[2] and Z.M. Yuan[1,2]

[1]Hunan Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, Hunan Agricultural University, Changsha, China
[2]College of Bio-Safety Science and Technology, Hunan Agricultural University, Changsha, China
[3]Department of Statistics, Kansas State University, Manhattan, KS, USA

*These authors contributed equally to this study.
Corresponding author: Z.M. Yuan
E-mail: zhmyuan@sina.com

Genet. Mol. Res. 11 (3): 3432-3451 (2012)
Received December 15, 2011
Accepted March 30, 2012
Published September 25, 2012
DOI http://dx.doi.org/10.4238/2012.September.25.12

ABSTRACT. Identification of splice sites plays a key role in the annotation of genes. Consequently, improvement of computational prediction of splice sites would be very useful. We examined the effect of the window size and the number and position of the consensus bases with a chi-square test, and then extracted the sequence multi-scale component features and the position and adjacent position relationship features of consensus sites. Then, we constructed a novel classification model using a support vector machine with the previously selected features and applied it to the *Homo sapiens* splice site dataset. This method greatly improved cross-validation accuracies for training sets with true and spurious splice sites of both equal and different proportions. This method was also applied to the NN269 dataset for further evaluation and independent testing. The results were superior to those obtained with previous methods, and demonstrate the stability and superiority of this method for prediction of splice sites.

Key words: Splice site prediction; Multi-scale component features; Position features; Adjacent position relationship features; Support vector machine

