International Journal of Basic and Applied Biology
Print ISSN: 2349 – 5820; Online ISSN: 2349 – 5839; Volume. 2 No. 2, November 2014. pp. 105
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/ijbab.html

Eukaryotic Donor Splice Site Prediction: A Machine Learning Approach

Prabina Kumar Meher[1], A. R. Rao[2] and S. D. Wahi[1]

[1]Division of Statistical Genetics, Indian Agricultural Statistics Research Institute, New Delhi-12, India
[2]Centre for Agricultural Bioinformatics, Indian Agricultural Statistics Research Institute, New Delhi-12, India

Abstract : Identifying the genes accurately is one of the most important and challenging task in bioinformatics and its success depends on the precise identification of splice sites. As AG and GT di-nucleotide represent possible donor and acceptor splice sites, every AG and GT in a DNA sequence is a candidate acceptor and donor splice site and they need to be classified as either a real splice site or a pseudo splice site. Given that AG and GT di-nucleotide occurs very frequently at non-splice-site positions, it is very hard to identify a true donor/acceptor splice site from a false splice site. Various computational methods have been developed for splice site prediction and among them machine learning methods have been more successful. In splice site prediction using machine learning approaches, features vector are generated through different encoding schema. In this investigation, an attempt is made to develop a new sequence encoding approach based on the di-nucleotide association. The encoded sequence data are then used for the prediction of donor splice sites using Artificial Neural Network (ANN), Support Vector Machine (SVM) and Random Forest (RF) methodology, following 10-fold cross validation techniques. Combination of SVM and RF coupled with proposed encoding approach achieved better accuracy as compared to the other combinations in terms of area under Receiving Operating Characteristics (ROC) curve (AUC). The performance of the proposed was also compared with several existing approaches using an independent data set of 50 genes. The proposed approach outperformed other approaches in terms of prediction accuracy.