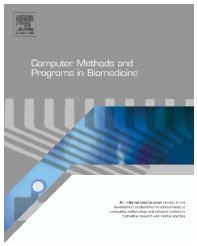
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 128 (2016) 1–11

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



"iSS-Hyb-mRMR": Identification of splicing sites using hybrid space of pseudo trinucleotide and pseudo tetranucleotide composition

*Muhammad Iqbal, Maqsood Hayat**

Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan

ARTICLE INFO

Article history:
Received 24 August 2015
Accepted 16 February 2016

Keywords:
Splicing sites
PseTNC
PseTetraNC
KNN
mRMR

ABSTRACT

Background and objectives: Gene splicing is a vital source of protein diversity. Perfectly eradication of introns and joining exons is the prominent task in eukaryotic gene expression, as exons are usually interrupted by introns. Identification of splicing sites through experimental techniques is complicated and time-consuming task. With the avalanche of genome sequences generated in the post genomic age, it remains a complicated and challenging task to develop an automatic, robust and reliable computational method for fast and effective identification of splicing sites.

Methods: In this study, a hybrid model "iSS-Hyb-mRMR" is proposed for quickly and accurately identification of splicing sites. Two sample representation methods namely; pseudo trinucleotide composition (PseTNC) and pseudo tetranucleotide composition (PseTetraNC) were used to extract numerical descriptors from DNA sequences. Hybrid model was developed by concatenating PseTNC and PseTetraNC. In order to select high discriminative features, minimum redundancy maximum relevance algorithm was applied on the hybrid feature space. The performance of these feature representation methods was tested using various classification algorithms including K-nearest neighbor, probabilistic neural network, general regression neural network, and fitting network. Jackknife test was used for evaluation of its performance on two benchmark datasets S_1 and S_2, respectively.

Results: The predictor, proposed in the current study achieved an accuracy of 93.26%, sensitivity of 88.77%, and specificity of 97.78% for S_1, and the accuracy of 94.12%, sensitivity of 87.14%, and specificity of 98.64% for S_2, respectively.

Conclusion: It is observed, that the performance of proposed model is higher than the existing methods in the literature so for; and will be fruitful in the mechanism of RNA splicing, and other research academia.

* *Corresponding author*. Tel.: +92 937542194.
E-mail addresses: m.hayat@awkum.edu.pk, Maqsood.hayat@gmail.com (M. Hayat).

http://dx.doi.org/10.1016/j.cmpb.2016.02.006

1. Introduction

Gene splicing plays prominent role in protein diversity and thus enable a single gene to increase its coding capability. The precursor messenger RNA (pre-mRNA) transcribed from one gene can lead to different mature mRNA molecules during a typical gene splicing event, which causes to generate multiple functional proteins. In eukaryotes gene, splicing takes place prior to mRNA translation by the differential inclusion or exclusion of regions called exons and introns of pre-mRNA. Exons that code for proteins are interrupted by non-coding regions called introns in eukaryotic genomes. There is a line between introns and exons called splice site (Fig. 1). Sides of introns have splice sites, the former is called the 5′ splice site or donor site and the latter is called the 3′ splice site or acceptor site. The vast amounts of donor and accepter sites form a pattern which is recognized by the presence of GT and AG, respectively. Spliceosome, which is comprises of 300 proteins and five small nuclear RNAs (snRNAs U1, U2, U4, U5, and U6) that is responsible for identification of donor and acceptor sites in genome sequence [1]. When splice sites become identified, spliceosome bind to both 3′ and 5′ ends of the introns and cause the intron to form a loop. With the help of two sequential transesterification reactions the given intron is eradicated from the genome sequence as shown in Fig. 1, while the remaining two exons are linked together [2,3]. Eliminating non-coding regions (introns) from (pre-mRNA) and fusing the required consecutive coding regions (exons) to form a mature messenger RNA (mRNA) is a prominent and notable step in gene expression. Therefore, to better understand the splicing mechanism; it is essential to identify the splicing sites in genome accurately.

Biochemical experimental approaches provide little details about identifying splicing sites with certain limitations, thus to rely only on these techniques is not appropriate, because

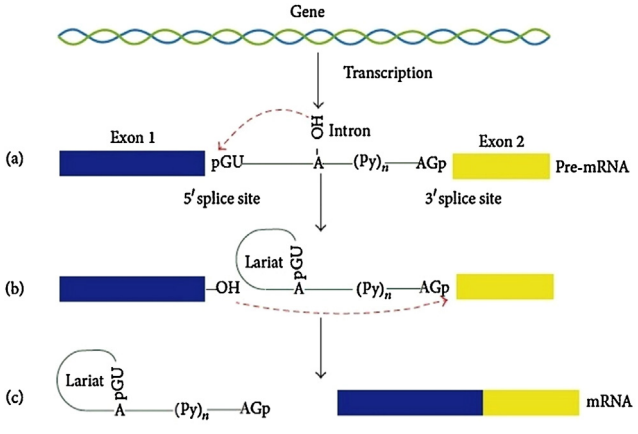


Fig. 1 – A schematic drawing to show the pathways of RNA splicing. (a)The 2′OH of the branchpoint nucleotide within the intron (solidline) carries out a nucleophilic attack at the first nucleotide of the intron at the 5′ splice site (GU) forming the lariat intermediate; (b) the 3′OH of the released 5′ exon then performs a nucleophilic attack at the last nucleotide of the intron at the 3′ splice site (AG); (c) joining the exons and releasing the intron lariat.

these are time-consuming and expensive operations. In addition, these are not mostly applicable. Hence with increasing the density of logic, it is a great challenge, and extremely desirable task to develop computational methods for precise, consistent, robust and automated system for timely identification of splicing sites. A series of methods have been proposed to identify splicing sites consequently, considerable results have been achieved, but still it contains large vacuum for further improvements in term of prediction performance. After the comprehensive review [4] and also a series of latest publications [5–11] revealed that, to develop a really effective statistical predicator for biological system, we need to pass from the following steps: (i) in order to train and test the predictor, we need to construct or select a valid benchmark dataset; (ii) for correct reflection of biological sample in their intrinsic correlation with the target to be predicted, we have to formulate the sample with an effective mathematical expression; (iii) to operate the predication, a powerful algorithm is needed; (iv) also to evaluate the anticipated accuracy of the predictor objectively, properly cross validation tests is needed to be performed.

In view of the importance of splicing sites for genome analysis, the present study was initiated to develop a computational method for predicting splice sites. In the present work, a hybrid model "iSS-Hyb-mRMR" is proposed, which used pseudo trinucleotide composition and pseudo tetranucleotide composition strategies to extract numerical descriptors. To eradicate the irrelevant and redundant features from feature space, minimum redundancy and maximum relevance (mRMR) was applied. Classification algorithms including K-nearest neighbors (KNN), probabilistic neural network (PNN), generalized regression neural network (GRNN) and fitting network (FitNet) were utilized in order to select the best one among these. Jackknife test was applied to assess the performance of the classification algorithms using two datasets S_1 and S_2 for donor sites and acceptor sites, respectively.

The rest of the paper is organized as; Section 2 describes materials and methods, Section 3 describes evaluation criteria for performance measurement, Section 4 describes result and discussions and finally conclusion has been drawn in Section 5.

2. Materials and methods

2.1. Dataset

In order to develop a statistical predictor, it is preliminary to establish a reliable and stringent benchmark dataset for training and testing the predictor. However, in case of erroneous and redundant benchmark dataset, consequently, the outcomes of predictor must be unreliable and inconsistent. In order to remove the redundancy and reduce the similarity from the dataset usually CDHIT is applied. In addition, as pointed out in a comprehensive review [12], for examining the performance of a prediction method there is no need to split a benchmark dataset into a training and testing dataset. Because, the performance of predictor is evaluated by leave one out cross validation or subsampling tests, actually, the predicted outcomes are the