Computational Biology and Chemistry 68 (2017) 219–223

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Prediction and feature analysis of intron retention events in plant genome

Ying Cui[a,c], Chao Zhang[a,*], Meng Cai[b,c]

[a] School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China
[b] School of Economics and Management, Xidian University, Xi'an 710071, China
[c] Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA

ARTICLE INFO

Article history:
Received 30 December 2016
Received in revised form 7 February 2017
Accepted 11 April 2017
Available online 13 April 2017

Keywords:
Alternative splicing
Intron retention
Feature selection
Random forest
Plant

ABSTRACT

Alternative splicing (AS) is a major contributor to increase the potential informational content of eukaryotic genomes by creating multiple mRNA species and proteins from a single gene. In plants, up to 60% genes are alternatively spliced and the most common type of AS is intron retention (IR). Genomic analyses of IR have illuminated its crucial role in shaping the evolution of genomes, in the control of developmental processes, and in the dynamic regulation of the transcriptome to influence phenotype. To explore the relationship between the sequence feature and the formation mechanism of IR, we statistically analyzed the retained introns and proposed an improved random forest-based hybrid method to predict intron retention events in plant genome. The results indicate that IR has significant relationship with individual introns which have weaker 5' splice sites, lower GC content and less termination codon occurrence. By the method we proposed, 93.48% retained introns can be correctly distinguished from constitutive introns. Strikingly, our study will facilitate a better understanding of underlying mechanisms involved in intron retention.

1. Introduction

In eukaryotic genes, pre-mRNA undergoes 5' end capping, splicing and 3' end polyadenylation during the transcription process. The essential step called splicing in gene expression refers to the removal of introns and ligation of exons by a large RNA-protein complex known as spliceosome, consisting of five small nuclear RNAs (snRNAs) and over 200 proteins with different functions (Wahl et al., 2009). While many splice sites are selected in the same way (Constitutive Splicing, CS), others are chosen by various splicing levels, resulting in alternative splicing (AS). AS is a process which produces multiple distinct mRNA isoforms from a single gene locus through variable selection of splice sites during pre-mRNA splicing. It plays a paramount regulatory role in proteome diversity and biological complexity (Petrillo et al., 2014; Raczynska et al., 2013). It is estimated that more than 95% of intron-containing genes undergo AS in humans (Pan et al., 2008), and roughly 15% of genetic diseases are connected with AS (Marquez et al., 2012). Recent research claims that more than 60% of intron-containing genes exit AS in plants, notably higher than previous studies reported (Syed et al., 2012).

In plants, AS events predominantly involve intron retention (Wong et al., 2016), the process by which an intron remains unspliced and consequently is transcribed into pre-mRNA. The frequency of IR is as high as 64% of all AS types (exon skipping, intron retention and the alternative 5′/3′ splice-site selection) in model plant *Arabidopsis thaliana* (Kalyna et al., 2012). The original view that intron retention is a consequence of mis-splicing and is non-functional (Roy and Irimia, 2008; Jaillon et al., 2008), was compromised by a string of recent discoveries (Buckley et al., 2011; Rocchi et al., 2012; Boothby et al., 2013; Palazzo et al., 2013; Wong et al., 2013; Remy et al., 2014). Existing evidence declares that IR has important biological consequences in plant development, stress responses, and tissue specific physiology (Mandadi and Scholthof, 2015; Wong et al., 2016). Although the importance of IR in plants has been undoubtedly confirmed, the mechanism contributing to the high incidence remains a puzzle. So in this paper, we pour attention into computationally analyzing the sequence feature of IR and distinguishing retained introns (RIs) from constitutive introns (CIs) which are invariably spliced during RNA transcription.

Machine learning, as an effective instrument widely used in the domain of bioinformatics, was also employed in the prediction of

* Corresponding author at: School of Mechano-Electronic Engineering, Xidian University, No. 2 South Taibai Road, Xi'an 710071, China.
E-mail address: zhangchaomee@163.com (C. Zhang).

http://dx.doi.org/10.1016/j.compbiolchem.2017.04.004
1476-9271/

IR both in human and plant genome (Hiller et al., 2005; Xia et al., 2006; Mao et al., 2014), such as Support Vector Machine and Random Forest. However, the results of existing methods are far from satisfactory. Besides, due to the reliance on annotation information, some existing methods have limitation in application.

In this work, we comparatively analyzed the sequence feature of retained and constitutive introns in model plant *Arabidopsis thaliana* in order to explore the reason why intron retention occurs in plant and what kind of introns tend to have higher possibility to escape from splicing. The results of sequence analysis provide clues that introns with weaker 5' splice sites, lower GC content, and less termination codons are more inclined to be retained during the process of splicing. Furthermore, we proposed a hybrid method based on improved random forest to predict the retained introns in *Arabidopsis thaliana*. The simulation results show that our approach achieves significant improvement of the prediction of IR success over the original RF method.

2. Materials and methods

2.1. Datasets

Data from Mao et al. (2014) was used to test the proposed method in this work. In this dataset, introns of model plant *Arabidopsis thaliana* were extracted from TAIR10 genome sequence (ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes/). After flitting the introns too short, 2520 retained introns (RIs), 110254 constitutive introns (CIs) and their flanking regions were obtained.

2.2. Improved random forest

Random Forest (RF) algorithm, which is an ensemble machine-learning method developed by Breiman (2001), has been widely applied to classification problems in bioinformatics area (Pan et al., 2014; Hu et al., 2016). RF is consisted of many base tree-structured classifiers such as CART (Classification And Regression Tree) and is proven to be robust to noise, no over-fitting and computationally feasible. By applying CART as a base classifier, RF collects the outputs of all decision trees to vote for the final result, then a sample is classified according the voting result. However, the majority voting rule in traditional RF algorithm renders the minority categories more likely to be misclassified. In response to this limitation of traditional RF, an improved RF function called cforest available in R-package party is employed in this paper (Hothorn et al., 2006, 2014; Strobl et al., 2008). In contrast to standard RF based on CART with unfair splitting criterion, cforest is established by unbiased base classification trees based on a conditional inference framework.

In cforest, the association between predictor variable and the response is evaluated by conditional inference trees. The procedure of cforest is: firstly, the algorithm tests the global null hypothesis of independence between any of the input variables and the response, it doesn't stop until this hypothesis is accepted. Otherwise select the input variable with strongest association to the response. Second, the selected predictor variable splits into two disjoint sets. Then, repeat the first and second steps.

For a forest framework, the most crucial point is the splitting objective function. In traditional RF, information gain and the Gini impurity are the most common measures which show a bias towards relevant predictor variables. Therefore, Strobl et al. (2008) proposed a conditional permutation importance scheme in cforest framwork. The new permutation importance scheme is based on a partition of the entire feature space that is determined directly by the fitted forest model. Accordingly, it is practicable for demonstrating the influence of a variable and computing its permutation importance conditional on correlated covariates of any type.

2.3. Feature analysis

In order to find out the closely related features for the prediction of intron retention and to further explore the factors that might be the contributors of intron retention events, we present a sequence feature analysis of retained introns (RIs) and constitutive introns (CIs) on the data sets described in section 2.1. We statistically calculated the features including length, nucleotide composition (mononucleotide, dinucleotide and trinucleotide), GC content, termination codon frequency and splice site strength.

To calculate the splice site strength of RIs and CIs, a supervised learning method, MaxEnt model (Yeo and Burge, 2004) was employed due to its excellent performance on splicing signal modeling. The basic idea of MaxEnt is to maximum the entropy of a distribution subject to certain constraints derived from training process. Comparing to Weight Matrix Model, MaxEnt is not bound by the limitation that hypothesizing independence between different positions. Additionally, rather than First-order Markov Model, MaxEnt takes connections between nonadjacent positions into consideration. A MaxEnt model consists of two distributions, scilicet signal model $(P^{+}(x))$ and the decoy probability distribution $(P^{-}(x))$.

Let p denotes the unknown probability distribution. For a specific DNA sequence x, $p(x)$ represents its probability in this distribution. let $\hat{p}$ be our approximation of p, the entropy of $\hat{p}$ is defined as,

$$H(\hat{p}) = -\sum \hat{p}(x)\log_2(\hat{p}(x)). \quad (1)$$

For a given sequence x, the signal strength of it could be scored by the following expression,

$$L(X = x) = \frac{P^{+}(X = x)}{P^{-}(X = x)}. \quad (2)$$

Here, $P^{+}(X = x)$ and $P^{-}(X = x)$ indicate the probability of x from the distributions

of signals (+) and decoys (−), respectively.

The results of feature analysis are shown in Table 1 and Fig. 1. Here, p value in Table 1 is computed by t-test to measure the statistically significant differences between two groups of data. A p value <0.05 is generally considered significant difference and a p value <0.01 usually means highly significant. After comparative analysis, it is found that, comparing to CIs, RIs tend to be shorter, have lower GC content, less termination codons and weaker 5' and 3' splice strength.

2.4. Feature selection

A total of 88 features are generated in section 2.3 including length, nucleotide composition (mononucleotide, dinucleotide and trinucleotide), GC content, termination codon frequency and

Table 1
Results of sequence feature analysis.

	Mean value of RIs	Mean value of CIs	p value
length	144.7	158.7	0.0002571
GC content	0.02942757	0.03599057	<2.2e-16
termination codon	0.05765421	0.05946783	0.005972
5' splice strength	2.988	5.986	<2.2e-16
3' splice strength	−15.41154	−14.70420	0.0007036