Biochemical and Biophysical Research Communications 471 (2016) 368–372

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

Recognition of alternatively spliced cassette exons based on a hybrid model

Xiaokang Zhang, Qinke Peng*, Liang Li, Xintong Li

Systems Engineering Institute, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Article history:
Received 28 January 2016
Accepted 6 February 2016
Available online 8 February 2016

Keywords:
Cassette exons
Constitutive exons
Hybrid model
Gene expression programming
Random forests

ABSTRACT

Alternative splicing (AS) is an important mechanism of gene regulation that contributes to protein diversity. It is of great significance to recognize different kinds of AS accurately so as to understand the mechanism of gene regulation. Many *in silico* methods have been applied to detecting AS with vast features, but the result is far from satisfactory. In this paper, we used the features proven to be useful in recognizing AS in previous literature and proposed a hybrid method combining Gene Expression Programming (GEP) and Random Forests (RF) to classify the constitutive exons and cassette exons which is the most common AS phenomenon. GEP will firstly make prediction to the samples of strong signal, and the other samples of weak signal will be distinguished with a more complex classifier based on RF. The experiment result indicates that this method can highly improve the recognition level in this issue.

1. Introduction

A great difference between prokaryotes and eukaryotes is that the pre-mRNAs of most eukaryotes contain not only expressed sequences (exons), but also intervening sequences (introns). Discovered in the late 1970s, pre-mRNA splicing removes the introns and joins exons together to form mature mRNAs [1,2]. And alternative splicing (AS) contributes to the diversity of expression of genes, which affects nearly 95% of mammalian genes [3,4].

The experiments *in vivo* to distinguish the different kinds of AS are expensive both in time and money. So computational methods in solving biological issues become more and more popular. Cassette exon, or exon skipping, where an exon is entirely included in, or excluded from the mature transcript, is the most prevalent form of AS in humans [5]. So in this paper, we'll focus on distinguishing cassette exons from constitutive exons which are always included in mature transcript.

To apply computational methods, the first task is to extract features from the DNA sequences. As time goes by, more and more features were used to identify alternative splicing. Sorek et al. [6] used seven features which are all statistic features. Based on that feature set, Dror, Sorek and Shamir [7] added the motif features to the feature set, and also used the features of pyrimidines and 5′ UTR, rising the feature dimension up to 228. Sinha et al. [8] added some motif features with biological significance and added the feature number up to 365. A much larger feature set came from Barash et al. [4], since they explored some remote regions next to the exon and made the feature dimension up to more than one thousand. And the feature set was used several times after by the search group [9,10]. However, too many features will increase the work burden and slow down the calculating speed, and what's more, that may not improve the accuracy but worsen it, since useless features will increase noise to the algorithm [7,11]. We collect the previously used features and pay attention to their performances, then choose only the putative effective features to train our prediction models afterward. Bush and Hertel [12] used information gain to describe the “worth” of the features to predict the splicing type of an exon. But information gain can only reflect the feature's independent attribution to the recognition of alternative splicing and ignores the interwork of them. So we utilize the Gene Expression Programming (GEP) algorithm to find out the features' interactive contribution [13].

Previous studies have applied state-of-the-art classifiers such as Support Vector Machine (SVM) [7] and Adaboost [14] in predicting exon skipping, but the result is far from satisfactory. A problem is that they treated all the samples equally without discrimination, but the samples are in fact different according to the signal strength. Busch and Hertel analyzed the inclusion level of cassette exons, and found that the majority of cassette exons are biased towards high (>=80%) or low (<=20%) splice-site exon inclusion

* Corresponding author.
E-mail address: qkpeng@mail.xjtu.edu.cn (Q. Peng).

http://dx.doi.org/10.1016/j.bbrc.2016.02.022
0006-291X/

levels [12]. The cassette exons of high inclusion level are of higher probability to be included into the mature transcript when splicing process occurs. So the constitutive exons can be regarded as a special kind of cassette exons whose inclusion level is 100%. So the higher inclusion level a cassette exon possesses, the more similar it is to a constitutive exon, which makes it more difficult to distinguish the cassette exons from the constitutive exons. To solve this problem we adopt a hybrid model combining two classifiers. Consistent with above, we keep GEP as the first classifier to select features and make prediction for the samples of strong signal at the same time. Then an ensemble learning algorithm, Random Forests (RF) [15], will be adopted to learn the delicate difference between the samples of weak signal to make further prediction.

2. Material and methods

2.1. Materials

The human cassette exons and constitutive exons are extracted from the HEXEvent database [16]. HEXEvent is a free database that provides a list of human internal exons and reports all their known splice events based on EST information from the UCSC Genome Browser [17]. We downloaded the first human chromosome and aligned the exon positions against the Human Reference Sequence (hg19, GRCh37.73) [17]. In order to analyze the context information of the exons, we also extended the length of exons using their flanking nucleotides to the upstream −70 bp and the downstream +70 bp. And only the splice sites obeying the GT-AG rule are kept. So finally a total of 1416 cassette exons were collected. And a total of 1416 constitutive exons were randomly selected.

2.2. Feature definition

The features used to identify cassette exons can be divided into several types, such as conservation based features, statistic features, RNA secondary structures, splicing regulator elements, and so on.

And among all those features, conservation based features are proved to be the most important [8,12]. Therefore we adopt the conservation based features as an important part of our used features.

Among the conservation based features, the Position Weight Matrix (PWM) is a generally used model to measure the conservative property of a sequence, and it's very popular because of its simplicity [18–20]. From a given set of N aligned sequence, the PWM was computed as:

$$P_{ib} = \frac{1}{N}\sum_{k=1}^{N} I(x_{ik}) \tag{1}$$

$$I(x_{ik}) = \begin{cases} 1, & if\ \ x_{ik} = b \\ 0, & otherwise \end{cases} \tag{2}$$

where $b \in \{A, C, G, T\};\ i = 1, 2, \cdots, n;\ k = 1, 2, \cdots, N$.

And the PWM scoring function (SF) is represented as:

$$SF = \sum_{i=1}^{n} ln\frac{P_{ib}}{P_{0b}} \tag{3}$$

This represents the strength of the splice site. In acceptor site sequence, the interested sequence is from −23 to 2 (1 corresponds to the start of an exon) and −3 to 6 (1 corresponds to the start of an intron) for donor site sequence.

But a disadvantage of PWM is that all positions are assumed to be independent of each other, and this assumption is unilateral in molecular sequence motifs [21]. Taking this into consideration, an Inhomogeneous first-order Markov Model (I1MM) which takes second-order nearest-neighbor constraints into consideration solves this problem [22]. An I1MM can be generated as follows:

$$P_{1MM}(X) = P(X_1)\prod_{i=2}^{n} P(X_i|X_{i-1}) \tag{4}$$

However, I1MM ignores dependencies between nonadjacent positions, so there comes the Maximum Entropy Model (MEM) [22]. An MEM consists of two distributions, namely the signal ($P^+(X)$) and the decoy probability distribution ($P^-(X)$). Given a new sequence, the Maximum Entropy Score (MES) is calculated as:

$$L(X = x) = \frac{P^+(X = x)}{P^-(X = x)} \tag{5}$$

MES denotes the like-hood of splicing at different splice donor and acceptor sequence, which is calculated using a Perl script provided by the Burge laboratory.

Besides the conservation based features, GC content is also found to play an important role in the recognition of cassette exons from constitutive exons [12,23,24]. So we take the GC contents of last 70 bp in upstream intron, the exon, and first 70 bp in downstream intron, into consideration (Fig. 1). We also use a 10 bp sliding window to count the GC contents in the adjacent 70 bp of the upstream and downstream introns (Fig. 1).

And the finally selected feature set is listed in Table 1.

2.3. Algorithm

On the aspect of exons, there are exons of high inclusion level and low inclusion level [12]. When we reflect that onto the data, that becomes the different strengths of signal. Apparently, the positive samples (cassette exons) of high-level splicing signal can be recognized more easily from the constitutive exons than the samples of low-level splicing signal. We design a hybrid model combining Gene Expression Programming and Random Forests, which classifies the samples of high-level splicing signal with the first classifier that is simple, and uses a much more complex classifier to classify the samples of low-level splicing signal in the second classifier. Fig. 2 illustrates the process. That idea partly comes from the Three-way Decisions [26]. The main idea of this strategy is to divide the entirety into three independent parts, and to adopt different methods to deal with them separately.

One advantage of the algorithm GEP is that it can find out the relationship among the variables, which tells us the important variables, or rather the principal features, at the same time. As we will see in the *Results* part, using the selected features by GEP, the second classifier will get a better performance, which verifies the statement before that too many noisy and useless features can weaken the performance of the classifier.

Throughout this paper we will use the notation $x_j^{(i)}$ with $i : 1 \cdots m$ and $j : 1 \cdots n$ to refer to the value of feature j in i^{th} training sample, where m is the number of samples and n is the number of features. In the first classifier, we will use GEP to learn the function F between the features as the independent variables $x^{(\bullet)} \in \mathbb{R}^n$ and the labels as the dependent variables $y \in \{1, -1\}$:

$$y = F\left(x^{(\bullet)}\right) \tag{6}$$

Say that after training, we get the GEP expression f. We set the threshold as $t \in [0, 1]$, then we define: