Gene 424 (2008) 115–120

Contents lists available at ScienceDirect

Gene

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

Using estimative reaction free energy to predict splice sites and their flanking competitors

Hong-Ying Jin, Liao-Fu Luo *, Li-Rong Zhang

Laboratory of Theoretical Biophysics, School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China

ARTICLE INFO

Article history:
Received 29 April 2008
Received in revised form 3 July 2008
Accepted 31 July 2008
Available online 7 August 2008

Received by M. Di Giulio

Keywords:
Maximum information principle
Alternative splicing
Competitors of splice site
Splice site prediction

ABSTRACT

A crucial part in the gene structure prediction is to identify the accurate splice sites, not only constitutive but also alternative ones. Here, we use the maximum information principle (MIP) to analyze the conservative segments around splice sites. According to the MIP, a reaction free energy (RFE) expression is deduced, which can be employed to estimate the free energy change during splicing reaction involving a donor or acceptor site. The expression contains not only the background probability factors, but also all kinds of dependencies among both adjacent and non-adjacent bases. We apply the RFE expression to recognize splice sites and their flanking competitors in human genes, the results show high sensitivity and specificity, so the RFE expression accords well with the splicing reaction process. Moreover, the RFE expression is better than previous methods for predicting competitors of splice sites, and it outperforms the reaction free energy subtraction (RFES), that implies RFE competition between a given splice site and its flanking competitor may not be an only primary factor for alternative splice site selection. The work is helpful to not only the understanding of splicing reaction from its relation to MIP, but also the research on computational recognition of splicing sites and alternative splice events.

1. Introduction

Most of the eukaryotic protein coding genes consist of introns and exons, and the borders between introns and exons are called splice sites (SSs). Approximately 99% of the 5′ SSs (donor sites) contain consensus dinucleotides GT and the 3′ SSs (acceptor sites) contain consensus AG (Burset et al., 2000). After precursor messenger RNA (pre-mRNA) of a gene has been transcribed, the introns must be removed from the precursor to give an mRNA that consists only of the series of exons, such is called RNA splicing. In the process of splicing, the pre-mRNAs of many genes follow patterns of alternative splicing, in which a single gene gives rise to more than one mRNA sequence. The mechanisms leading to alternative splicing include the blocking of splicing factor binding sites (such as the polypyrimidine tract), affinity increase of splicing factors by splice enhancers, the inhibition by pre-mRNA secondary structures, etc. (Smith and Valcárcel, 2000; Graveley, 2001; Modrek et al., 2001; Wang and Marin, 2006). Recent studies have indicated that Up to 74% of all human genes are alternatively spliced (Modrek and Lee, 2002; Johnson et al., 2003). Alternative splicing is a key mechanism for enriching proteomic diversity and regulating tissue-specific developmental processes by producing several transcripts from a single gene (Lopez, 1998; Kazan, 2003).

The prediction of the complete gene structure is an important problem in genome annotation, it depends on the precise identification of the SSs, not only constitutive but also alternative ones (Florea et al., 2005). However, it is difficult to detect alternative splicing events (Thanaraj and Stamm, 2003; Florea et al., 2005; Wang and Marin, 2006). An important type of alternative splicing event is alternative 5′ SS or alternative 3′ SS (i.e. exon extension/truncation) event, which takes place in competing with each other between adjacent SSs (Xia et al., 2006). If we can successfully predict flanking competitors of given splice sites, we will be able to find such events. To solve the problem, a method was proposed recently (Xia et al., 2006), which selects more than 300 parameters, including nucleotide composition in SS flanking region, U1 snRNA binding energy to 5′ SS, features of polypyrimidine tract of 3′ SS, and distance and frame-preservation between a given SS and its competitor; and then combines them in SVM to predict flanking competitors (namely, to recognize competitive splice site pairs and non-competitive splice site pairs in this method). Another method employs only one parameter (Yang and Li, 2008), the position weight matrix (PWM) scoring function subtraction between a given splice site and its competitor to make prediction and obtains the basically same accuracy as the method of Xia et al. (2006). However, it lacks any explanation on the scoring function subtraction (SFS); and furthermore, lacks any

Abbreviations: MIP, maximum information principle; RFE, reaction free energy; RFES, reaction free energy subtraction; SS, splice site; pre-mRNA, precursor messenger RNA; PWM, position weight matrix; SFS, scoring function subtraction; IC, information content; S_n, sensitivity; S_p, specificity; ROC, receiver operating characteristics; MEM, maximum entropy model; MDD, maximum dependence decomposition model; MM1, first-order Markov model; WMM, weight matrix model.

* Corresponding author. School of Physical Science and Technology, Inner Mongolia University, Hohhot 010021, China. Tel.: +86 471 4992676; fax: +86 471 4951761.
E-mail addresses: liumy312@sohu.com, lolfcm@mail.imu.edu.cn (L.-F. Luo).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.07.038

consideration for the background influence in genome and the dependency among bases around SSs.

In this paper, we apply MIP to splicing process and deduce a RFE expression, which can be employed to estimate the free energy change during splicing reaction involving a donor or acceptor site. Here, the RFE includes not only binding free energy, but also free energy change during the subsequent reaction. In a previous work, a method was proposed for the estimation of binding free energy of transcription factor to DNA (Berg and von Hippel, 1987, 1988). The method assumes that all of the binding sites of a given transcription factor are equally likely to occur (Berg and von Hippel, 1987; Kinney et al., 2007). The assumption is too simplified; moreover, the method has used some approximations and not considered the complicated dependence structures among bases. Our RFE expression will exceed these limitations. Based on the consideration that the RFE of a true donor or acceptor site during splicing reaction is less than that of a pseudo donor or acceptor site, we shall use the RFE expression to identify SSs and their flanking competitors. This will give deeper insight into the research on SS selection.

2. Materials and methods

2.1. Datasets

Human splice site data are downloaded from the latest AltSplice database (Human release 3) of EBI, which is available at http://www.ebi.ac.uk/asd/altsplice/index.html (Thanaraj et al., 2004; Stamm et al., 2006). Then, from the alternative 5′ and 3′ splicing events that are annotated as 'INTRON ISOFORM (II-5P)' and 'INTRON ISOFORM (II-3P)' types respectively in the database, 6706 Alternative 5′ SSs and 8334 Alternative 3′ SSs are extracted, together with their competitive SSs. These alternative 5′ and 3′ SSs have 7797 and 9472 competitors respectively. From this database, 7382 constitutive 5′ SSs and 7695 constitutive 3′ SSs, which exist in all splicing patterns and whose flanking exons and introns do not show any alternative splicing events are also extracted. All these selected SSs obey GT–AG rule.

Statistical analysis shows that the >90% distances between two 5′ (3′) SSs in an alternative 5′ (3′) splicing event are smaller than 200 (150) nucleotides (nt) (Xia et al., 2006), so we extract all non-splice-site GT (AG) dinucleotides located within 200 nt of the given 5′ (3′) SSs as the negative sets.

2.2. Introduction to the MIP

The MIP is a fundamental principle in non-equilibrium statistical theory (Jaynes, 1957; Haken, 1988; Luo, 2004; Lezon et al., 2006). The principle indicates: information content (IC) of any a non-equilibrium system tends towards a maximum under a set of constraint conditions. According to the MIP, if we select appropriate constraints, and apply the Lagrange multiplier method to evaluate the constrained maximum of the IC of a non-equilibrium system, we can solve any distribution problem of the system in principle. Nucleic acid sequence is a typical non-equilibrium system, where bases often undergo the mutation stochastically under the inherent perturbation in the micro-environment. Thereby the MIP as a guiding principle in the genetic language research and bioinformatics of nucleic acid sequence was proposed by Luo et al. (Luo and Bai, 1995; Luo, 2004).

2.3. Using the MIP to deduce RFE expression

We extract large numbers of donor (acceptor) conservative sites (including surrounding regions) from protein coding genes in the human genome. Supposing that the length of a conservative site segment is l bases, we randomly select k ($k \le l$) bases from l base sites and combine them together to form a class of k-mer selective mode. For given l and k there are C_l^k classes, and for each class there are 4^k kinds of k-mers. The IC of donor (acceptor) site is defined as follows:

$$I = -\sum_s \sum_i p_{si} \log_2 \frac{p_{si}}{p_{0i}} \quad i = b_1 b_2 \ldots b_k \tag{1}$$

where I means the IC, s is the index of a class (range 1 to C_l^k) of C_l^k k-mer classes, i is the index of a k-mer (range 1 to 4^k) in a given class of C_l^k k-mer classes, p_{si} is the probability of k-mer i in class s, and p_{0i} denotes background probability of k-mer i in the genome.

The value of I has a range of $C_l^k \cdot \min\left(-\log_2 \frac{1}{p_{0i}}\right) \le I \le 0$. If the 4^k k-mers in any of C_l^k classes are distributed according to the background probabilities, I is maximized to 0. If donor (acceptor) sites are so conservative that they are composed of only one kind of segment, I tends to be minimized. In fact, k-mers in donor (acceptor) sites are not distributed according to the background, so the IC does not take the maximum. The reason is that they are constrained by their biological function. If a site segment has not any biological function, the neutral mutation will drive it towards the k-mer distribution of background. The integrative effect of neutral mutation and functional constraint is: the IC of the site segment tends towards a constrained maximum under some functional constraints. This is the generalization of MIP to the case of given background probabilities.

The k-mers in donor (acceptor) sites have not been distributed according to the background in evolution because of the functional constraints, and one major constraint is such k-mers should be advantageous to the process of splicing reaction. That is to say, the RFE for a real donor (acceptor) splicing reaction is less than that for a hypothetic one between a pseudo donor (acceptor) site and spliceosome. Thus, the true donor (acceptor) sites have enough competitive advantage; they can specifically react with spliceosome. In view of this, for donor (acceptor) sites, we are able to deduce an RFE expression according to the MIP by use of following constraints.

A. The constraint for the normalization of probability:

$$\sum_s \sum_i p_{si} = C_l^k \tag{2}$$

B. The constraint for the RFE:

We know that the splicing mechanisms of most protein coding genes are basically the same. The splicing reaction is performed by the spliceosome, comprising more than 150 proteins and five complexes of RNA and proteins called snRNPs (Kol et al., 2005). Splicing includes a series of biochemical processes, and it is so complex that the direct measurement of the RFE is difficult. But regardless of the complexity of the splicing reaction, we can make following hypothesis.

For a k-mer in a given class of donor (acceptor) sites, the RFE between the k-mer and spliceosome is determined by its segment type. The total RFE of a donor (acceptor) site is the sum of the RFEs over the C_l^k k-mers divided by C_{l-1}^{k-1}, the repeated count number per position, since when we add all C_l^k RFEs together, each of l donor (acceptor) positions has been repeatedly counted C_{l-1}^{k-1} times. As far as a large number of donor (acceptor) sites in the genome are concerned, their average RFE tends towards a stable value.

This assumption is the modified additivity assumption (Berg and von Hippel, 1987; Benos et al., 2002; Lässig, 2007). Previous additivity assumption focuses on the binding free energy of transcription factor to DNA; here we use reaction free energy instead of binding free energy. The reason is that the evolution of splice sites is not only related to the binding energy, but also to the free energy change during the subsequent reaction. For example, some studies indicated that in splicing of some genes, if U1 snRNA binds too tightly to 5′ SS, splicing will be impaired through delayed release of U1 snRNA, which prevents the formation of the spliceosome's catalytic core (Staley and Guthrie, 1999; Lund and Kjems, 2002).