Available online at www.sciencedirect.com
ScienceDirect
Genomics 89 (2007) 471–479
GENOMICS
www.elsevier.com/locate/ygeno

# Applying genetic programming to the prediction of alternative mRNA splice variants

Ivana Vukusic [a], Sushma Nagaraja Grellscheid [b], Thomas Wiehe [a,*]

[a] Institut für Genetik, Universität zu Köln, Zülpicher Strasse 47, 50674 Köln, Germany
[b] Institute of Human Genetics, International Centre for Life, University of Newcastle upon Tyne, Newcastle upon Tyne NE1 7RU, UK

Received 5 August 2006; accepted 2 January 2007
Available online 5 February 2007

**Abstract**

Genetic programming (GP) can be used to classify a given gene sequence as either constitutively or alternatively spliced. We describe the principles of GP and apply it to a well-defined data set of alternatively spliced genes. A feature matrix of sequence properties, such as nucleotide composition or exon length, was passed to the GP system "Discipulus." To test its performance we concentrated on cassette exons (SCE) and retained introns (SIR). We analyzed 27,519 constitutively spliced and 9641 cassette exons including their neighboring introns; in addition we analyzed 33,316 constitutively spliced introns compared to 2712 retained introns. We find that the classifier yields highly accurate predictions on the SIR data with a sensitivity of 92.1% and a specificity of 79.2%. Prediction accuracies on the SCE data are lower, 47.3% (sensitivity) and 70.9% (specificity), indicating that alternative splicing of introns can be better captured by sequence properties than that of exons.


*Keywords:* Alternative splicing; Cassette exon; Intron retention; Genetic programming; Feature matrix; Splice signals

Alternative pre-mRNA splicing is a major source of transcriptome and proteome diversity. In human, aberrant splicing is an important cause of genetic diseases and cancer [1–5]. Until a few years ago it was believed that almost 95% of all genes undergo constitutive splicing, in which introns and exons are uniquely defined objects (Fig. 1a). It is now widely accepted that alternative splicing is the rule rather than the exception and that perhaps more than 75% of all human genes are alternatively spliced [6–10]. The various forms of alternative splicing are illustrated in Figs. 1b–1f, of which the cassette exon splicing is the most frequent type of alternative splicing [11].

Whether an exon or an intron will be included or excluded in the transcripts of a gene of a certain cell type is influenced by the information contained in the sequence of the exon and the flanking intronic region. This includes sequences that indicate exon–intron boundaries, binding sites for essential splicing factors, and binding sites for splicing enhancer and splicing silencer sequences. Often the sequences are very degenerate and bear little similarity to a consensus sequence. This makes bioinformatic analysis of splicing very challenging. In addition, it is commonly accepted that no single factor determines whether an exon will be spliced into a transcript. Instead, it is perhaps a combined effect of various factors including *cis*-acting sequences and *trans*-acting splicing factors.

Early approaches for large-scale detection of alternative splicing were based on observed transcripts. The search for instances of alternative splicing was performed by the alignment of expressed sequence tags (ESTs) to the genome and to other ESTs or cDNAs [11]. Other studies have relied on specifically generated microarrays for the detection of alternative splicing [9,12]. However, since these methods produce only a snapshot of the tissue that is sampled at a certain time and under certain conditions, many alternative events may still remain undiscovered. Therefore innovative, non-EST-based approaches are required to detect these events and to complete the knowledge about the transcriptome.

Recent studies have focused on comparative genomics, since functional parts of the DNA tend to be conserved between species [13–15]. Sorek et al. described a non-EST-based method that uses characteristic features of alternative exons to

* Corresponding author. Fax: +49 221 470 1630.
*E-mail address:* twiehe@uni-koeln.de (T. Wiehe).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.01.001

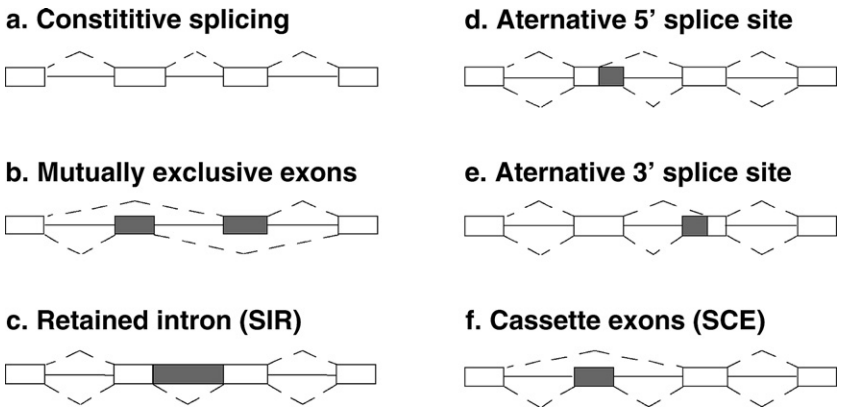


Fig. 1. Schematic representations of patterns of alternative splicing. Constitutive exons are shown in white and alternatively spliced exons in gray. SIR: simple intron retention; SCE: simple cassette exon.

distinguish between constitutive and cassette exons [16]. In addition to the length of an exon and avoidance of reading frame disruption, an important feature employed by these authors was a high sequence conservation of alternative exons and their flanking intronic regions in human–mouse orthologs [17]. The prediction accuracy could be raised by including additional features (e.g., different trimer counts and the composition of the splice sites) and by using a machine learning approach based on support vector machines (SVMs) [18]. In 2005 Rätsch and colleagues designed an SVM kernel with position-specific motifs to classify alternative exons in *Caenorhabditis elegans*. This approach does not require any information of the conservation level [19]. Yeo et al. [20] have developed a statistical machine learning algorithm, named ACEScan, that is based on regularized least-squares classification. ACEScan distinguishes exons with evolutionarily conserved alternative splicing from constitutively spliced or lineage-specific-spliced exons [21]. This approach uses features similar to the ones employed by Sorek et al., for instance, conservation level, splice site scores, exon and intron lengths, and oligonucleotide composition. Ohler et al. [22] have developed an algorithm that uses a pair hidden Markov model on orthologous human–mouse introns. This approach is applied to detect alternative exons that were completely missed in current gene annotations. A method proposed by Hiller et al. [23] does not depend on the existence of orthologous sequences. They use information from protein domain families (Pfam) to predict exon skipping and intron retention events. In this study, we have used genetic programming (GP), a machine learning approach, to generate classifiers of cassette exons and retained introns.

## Results and discussion

### *Sequence features*

Exon length is known to be one distinguishing feature for alternatively and constitutively spliced exons: alternative exons are usually shorter [8]. Fig. 2 shows the length distributions from our data set of cassette and constitutively spliced exons. The average length of simple cassette exons (SCE) is 139 bp. This value is 8% smaller than the average length of constitutively spliced exons (151 bp). The maximal length of a constitutively spliced exon is 7572 bp; in contrast the largest SCE has a length of 3726 bp. Both length distributions are qualitatively very similar. However, the SCE length distribution is shifted to smaller values. This difference is statistically significant (two-tailed $t$ test, $p = 0.0001$). A much larger difference was observed in the data set of constitutively spliced and simple retained introns (SIRs) (Fig. 2). The average length of introns of the constitutive data set is 6367 bp; 68% of the introns are longer than 1 kb. In contrast, the average length of retained introns is

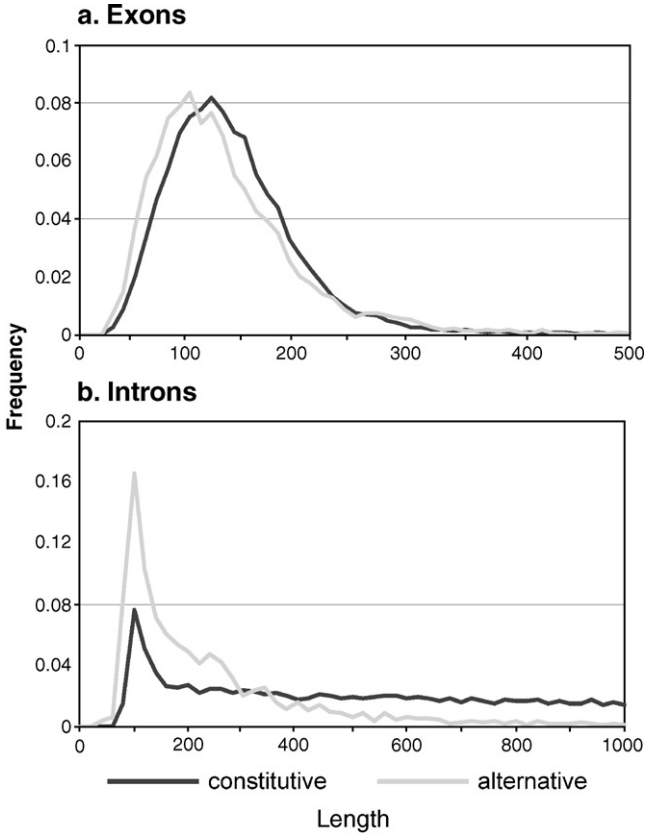


Fig. 2. (a) Length distribution of cassette and constitutively spliced exons. (b) Length distribution of retained and constitutively spliced introns. Note that the length of constitutive introns has an extreme heavy-tailed distribution.