ELSEVIER

The types and prevalence of alternative splice forms

Mihaela Zavolan and Erik van Nimwegen

The finding that eukaryotic gene structures are extremely complex prompted the development of new experimental techniques for the accurate measurement of transcription start site usage and of the expression of alternative splice forms. On the computational side, analyses of large databases of splice variants revealed differences in the length, motif composition and selection pressure between constitutive and alternatively spliced exons. Such features are being incorporated into novel computational tools for gene structure prediction. The result of these investigations is a continuously improving catalogue of alternative splice forms. How the expression of these alternative splice forms is regulated remains one of the major open questions.

Addresses
Division of Bioinformatics, Biozentrum, University of Basel, Klingelberstrasse 50-70, Basel, CH-4056, Switzerland

Corresponding author: Zavolan, Mihaela (mihaela.zavolan@unibas.ch)

Current Opinion in Structural Biology 2006, **16**:362–367

This review comes from a themed issue on
Sequences and topology
Edited by Nick V Grishin and Sarah A Teichmann

Available online 18th May 2006

0959-440X/$ – see front matter

DOI 10.1016/j.sbi.2006.05.002

Introduction

Recent probing of mammalian transcriptomes using high-throughput techniques [1••,2] revealed that much of the diversity of cell types in multicellular eukaryotes is due to the complex patterns of expression of gene structures. Most genes initiate transcription from multiple promoters, show multiple splice variations and have multiple polyadenylation sites. The complexity of the observed splice forms challenges the classical concept of a gene and requires novel ways of describing sets of transcripts that derive from a common genomic region. In addition, the function and tissue-specific expression profiles of most splice variants are unknown.

Here, we will review recent developments in the approaches to identify splice variants, the methods and vocabulary for describing splice variations, the measurement and analysis of their tissue-specific expression and regulation, and their impact on protein structure.

Identification of alternative splice forms

Virtually all recent studies aimed at the identification of splice variants use the genome sequence as a reference onto which various other types of information are projected, such as the genome sequences of other species, transcript sequences (expressed sequence tags [ESTs] and full-length cDNAs), protein domain models, and expression information from tiling and exon–exon junction microarrays.

Probably the most reliable method for identifying splice variations starts with full-length cDNA sequences [1••,3]. These are mapped to their respective genomes and clusters of overlapping transcripts are analyzed to determine alternative splicing events. The great advantage of full-length cDNAs is that, from a complete transcript, one can accurately identify the transcription start site, the polyadenylation signal and the splice sites that were used to generate the mature mRNA. The drawback is that the number of full-length cDNAs currently available is still relatively small, especially compared with EST data.

For this reason, splice variation is more commonly inferred from ESTs, but because ESTs are typically short and only represent fragments of transcripts, one is confronted with the issue of having to reconstruct full-length transcripts computationally. This 'EST assembly problem' can be best approached using so-called 'splice graphs' [4,5]. The nodes of this graph are individual genomic nucleotides, and a directed edge from node x to node y exists if and only if nucleotide y occurs immediately downstream of x in at least one transcript. Sets of consecutive nodes with in- and out-degree 1 correspond to genomic regions that are always included in transcripts as contiguous blocks and are therefore often collapsed into single nodes to simplify the graph. An alignment of overlapping ESTs can be efficiently represented as such a splice graph and each possible full-length transcript consistent with the EST data corresponds to a path through this graph [5]. One may also assign likelihoods to the different possible transcripts [6,7]. In a related approach, Eyras *et al.* [8] constructed a graph whose nodes correspond to ESTs and ESTs that are consistent with a common underlying transcript are connected with edges. This graph is then used to construct a minimal set of transcripts that can account for all of the EST data.

The drawback of the splice graph approach is that it does not capture correlations between splice events in different parts of the transcript. As the number of splice sites inferred from a set of overlapping transcripts grows, the number of possible full-length transcripts grows

combinatorially, yet most of them probably never or rarely occur *in vivo*. In addition, EST coverage is biased toward the 5′ and 3′ ends of transcripts, and this may lead to underestimation of alternative splicing affecting the central regions of transcripts [9].

Recently, a number of computational approaches have been proposed for predicting splice variation directly from genome sequence data. In Seneff *et al.* [10], known human gene structures are used to predict orthologous gene structures in mouse. A number of groups [11–14] have developed machine learning approaches to distinguish cassette exons (included in some but not all transcripts) from constitutive exons (always included) based on sequence and conservation statistics. Cassette exons have a wider distribution of lengths, with an enrichment of both short [12,15,16] and long [16] exons. Their sequence composition differs from that of constitutive exons [13,16], in particular in that they are depleted of known splice enhancer elements [16]. Cassette exons are also flanked by 'weaker' splice sites [16,17], but the intronic regions flanking them are more strongly conserved than the intronic regions flanking constitutive exons [18,19]. A large proportion of cassette exons are not conserved between human and mouse [20•], but those that are tend to preserve the reading frame [20•].

Hiller *et al.* [21] computationally predict alternative splice events by mapping PFAM hidden Markov models of protein domains [22] to pre-mRNAs. For each transcript, they consider the transcripts that would be obtained by retaining any of the introns or skipping any of the exons in the pre-mRNA. A prediction of splice variation is made whenever such splice events would significantly improve the match of the transcript to one of the PFAM domains [22]. It appears that the false-positive rate of this method is low, as there is EST evidence for close to 80% of the splice variants predicted in this way [22]. However, by design, the method detects only splice variations that add or improve protein domains in the transcript.

All approaches mentioned above are computational, and operate on the transcript and protein data available in public databases. Another category of studies aims at *de novo* experimental discovery of splice variants. The most widely used approach for this purpose relies on oligonucleotide microarrays and has been called ''design to annotate'' by Srinivasan *et al.* [23•]. Some studies use genome tiling arrays that enable identification of exon–intron boundaries [24] and thereby of expressed exons. Other studies use microarrays that cover a large number of possible exon–exon boundaries, many of which are not yet known to be generated *in vivo*, to discover novel splice events [9]. A unique advantage of the microarray technology is that it enables quantitative measurement of the expression of individual exon forms or even transcript forms. However, a lot of effort has to be invested in designing the microarray to be able to accurately recover information about alternative splicing. Such microarray design issues have been discussed at length by the groups involved in the development of this technology [9,23•,24,25•]; Wang *et al.* [26] designed an algorithm to deconvolve the data about exon and exon–exon junction expression into concentrations of individual splice variants.

Although detailed information about the expression of individual splice forms across tissues is yet to make its way into the currently available genome browsers, there are a number of sequence-based databases of alternative splice forms available online ([4,27–30]; SPAED: Splice Analysis of EST Data http://www.spaed.unibas.ch).

Gene structure vocabulary

Almost all of the terminology that is used by researchers today to describe splice variations derives from experimental studies in which a reference transcript form is compared with typically one or a few variant forms. In this case, variants can usually be described in terms of relatively simple splice variations that we would like to call 'classical' [31]. These include alternative acceptor splice sites, alternative donor splice sites, intron inclusion (also called intron retention), and 'cassette' or 'alternative' exons that are included in some transcripts and excluded in others. Cassette exons that are not included in the reference form, but become included in transcripts either upon sequence mutations or upon specific regulatory events are often referred to as 'cryptic exons'. *Vice versa*, if the reference form includes the exon and the variants do not, the exon is often said to be 'skipped'.

Once large sets of cDNAs and ESTs were aligned to their corresponding genomes, it became clear that the above vocabulary does not suffice to describe the complex variations that are evident in the data. To illustrate these complexities, Figure 1 shows an alignment to the mouse genome of a set of overlapping full-length cDNAs from the FANTOM3 data set [1••]. First, such alignments challenge the simple concept of each transcript deriving from 'a gene': the transcripts annotated with coding regions (CDS) in Figure 1 correspond, in fact, to three different known 'genes' (see legend). Additionally, a number of transcripts (c–i) have been isolated in a study that showed that transcription of this genomic region generates polycistronic transcripts [32]. Finally, a non-coding transcript (k) was isolated during the FANTOM3 project that covers the 3′ end of the hyaluronidase 1 gene and whose function is not known. This example makes it clear that a more general formal terminology is needed to describe gene structures. The solution chosen by the FANTOM consortium was to describe the gene structures evident in their data by a hierarchy of transcript clusters [1••]. At the lowest level of resolution, transcripts with overlapping pre-mRNAs are clustered into