Prediction and Statistical Analysis of Alternatively Spliced Exons

T.A. Thanaraj[1] and S. Stamm[2]

The completion of large genomic sequencing projects revealed that metazoan organisms abundantly use alternative splicing. Alternatively spliced exons can be found in these sequences by sequence comparison of genomic, mRNA and EST sequences. Furthermore, a large number of alternative exons have been described in the literature. Here, we review computer and manually curated databases of alternative exons and discuss the various approaches used to generate them. Sequence analysis shows that alternative exons often have unusual lengths, suboptimal splice sites and characteristic nucleotide patterns. Despite this progress alternative exons cannot be predicted *ab initio* from genomic data, which is due to the degenerate nature of splicing signals.

1 Overview

The first draft of the human genome has demonstrated that an average human gene contains a mean of 8.8 exons with an average size of 145 nt. The mean intron length is 3365 nt and the 5′ and 3′ UTR are 770 and 300 nt, respectively. As a result, a “standard” gene spans about 27 kbp. After pre-mRNA splicing, the mature message consists of 1340 nt coding sequence and 1070 nt untranslated regions and a poly (A) tail (Lander et al. 2001). The vertebrate splicing machinery is not only capable of accurately recognizing the small exons within the larger intron context, but is also able to recognize exons alternatively. In this process, an exon is either incorporated into the mRNA, or is excised as an intron. This process of alternative splicing is abundantly used in higher eukaryotes. In humans, a detailed analysis was performed for chromosome 22 and 19 (Lander et al. 2001). Of the 245 genes present on chromosome 22, 59% are alternatively spliced and the 544 genes of chromosome 19 result in 1859 different messages. The comparison of expressed sequence tags (ESTs) with the human genome sequence indicates that 47% of human genes might be alternatively

[1] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

[2] Institute for Biochemistry, University Erlangen-Nürnberg, Fahrstrasse 17, 91054 Erlangen, Germany; e-mail: stefan@stamms-lab.net

Progress in Molecular and Subcellular Biology, Vol. 31
Philippe Jeanteur (Ed.)

spliced (Modrek et al. 2001). This is in contrast to data obtained from *C. elegans*, where about 22% of the genes are alternatively spliced. Several databases have been generated that contain a wealth of information about sequences that are involved in alternative splicing. Here, we shall discuss the major findings and the computer programs used to generate them.

2 Mechanism of Splicing

2.1 General Splicing Mechanism

Three major *cis*-elements of the pre-mRNA define an exon, the 5′ splice site, the 3′ splice site and the branch point. All these elements are short (7–14 nt) and can only be described by degenerate consensus sequences. A given *cis*-element in a human gene will follow the consensus sequence only to a certain degree (Burset et al. 2000, 2001). This divergence from consensus sequences in higher eukaryotes is in contrast to the situation in yeast, where splice sites follow a strict consensus. To allow for exon recognition, in higher eukaryotes, additional elements known as exonic or intronic enhancers, depending on their location, are present (Cooper and Mattox 1997; Smith and Valcárcel 2000). Through recognition of these regulatory elements, sequence-specific RNA binding proteins regulate spliceosome assembly. The spliceosome is a 60S complex containing small nuclear RNPs (U1, U2, U4, U5 and U6) and over 50 different proteins. In this complex, U1 snRNP binds to the 5′splice site. SF1 and U2 snRNP bind to the 3′ splice site and the branch point. The consensus sequences of the 5′ splice site and the branch point reflect their binding to U1 and U2 snRNA, and the polypyrimidine tract of the 3′ splice site is reminiscent of the systematic evolution of ligands by exponential enrichment (SELEX) sequence of U2AF (TTTYYYYTNTAGG; Wu et al. 1999).

2.2 Exon Definition

Since the splice sites in higher eukaryotes are less conserved than the ones in yeast, the question arises how these exons are recognized. It was proposed that in vertebrates with their larger introns, the splicing machinery searches for a pair of closely spaced 5′ and 3′ splice sites (Berget 1995). The exon is then defined by binding of U1 and U2 snRNAs, as well as associated splicing factors to the exon. After an exon has been defined, neighboring exons must be juxtaposed. Due to the degenerate nature of splice sites in higher eukaryotes, it is difficult to predict exons from genomic DNA sequences and current computer programs cannot accurately predict exons from genomic DNA (Thanaraj

2000). This finding *in silico* contrasts with the high accuracy and fidelity characteristic for splice sites in vivo.

One reason for the specificity observed in vertebrate cells are additional regulatory elements known as silencers or enhancers. Based on their location they can be intronic or exonic. These sequence elements are again characterized by loose consensus sequences. They can be subdivided into purine-rich (GAR-type) and AC-rich (ACE-type) enhancers (Cooper and Mattox 1997). Enhancers bind to proteins that are able to recruit spliceosomal components, which results in the recognition of splice sites that are located near an enhancer (Hertel and Maniatis 1998). Since enhancers are often exonic, their loose consensus sequences, e.g., their degeneracy, is most likely necessary to allow for the amino acid usage needed in a given protein.

Proteins binding to sequence elements on the pre-mRNA can be subdivided into two major groups: members of the serine/arginine-rich (SR) family of proteins (Fu 1995; Manley and Tacke 1996; Graveley 2000) and heterogenous ribonuclear proteins (hnRNPs) (Weighardt et al. 1996). The binding of individual proteins to enhancer sequences is intrinsically weak and not highly specific. However, in most cases studied several such sequence elements are present. Furthermore, the proteins binding to *cis*-elements often bind to other RNA binding proteins. As a result, a protein : RNA complex is formed and the exon is recognized with high specificity. The composition of the protein : RNA complex is dependent on the concentration of various regulatory proteins, their phosphorylation status and the sequences of the regulatory elements on the pre-mRNA (Fig. 1).

2.3 Splice-Site Recognition Is Influenced by the Relative Concentration of Regulatory Proteins

The regulation of alternative splicing is still under intense investigation. The relative concentration of splicing-associated proteins can regulate alternative splice site selection (Hastings and Krainer 2001). Experiments both in vivo and in vitro show that the relative concentration of SR proteins and hnRNPs can dictate splice site selection (Mayeda and Krainer 1992; Mayeda et al. 1993; Caceres et al. 1994). Furthermore, the expression levels of various SR proteins (Ayane et al. 1991; Zahler et al. 1993; Screaton et al. 1995) and hnRNPs (Kamma et al. 1995) vary amongst tissues and could therefore account for differences in splice site selection. Several examples of antagonistic splicing factors have been described (Mayeda et al. 1993; Caceres et al. 1994; Gallego et al. 1997; Jumaa and Nielsen 1997; Polydorides et al. 2000). Here, one factor promotes inclusion of an exon and the other factor promotes its skipping. In most of these cases, it remains to be determined whether this antagonistic effect is achieved by (1) an actual competition of the factors for an RNA binding site, (2) through sequestration of the factors by protein:protein interaction and (3) by changes

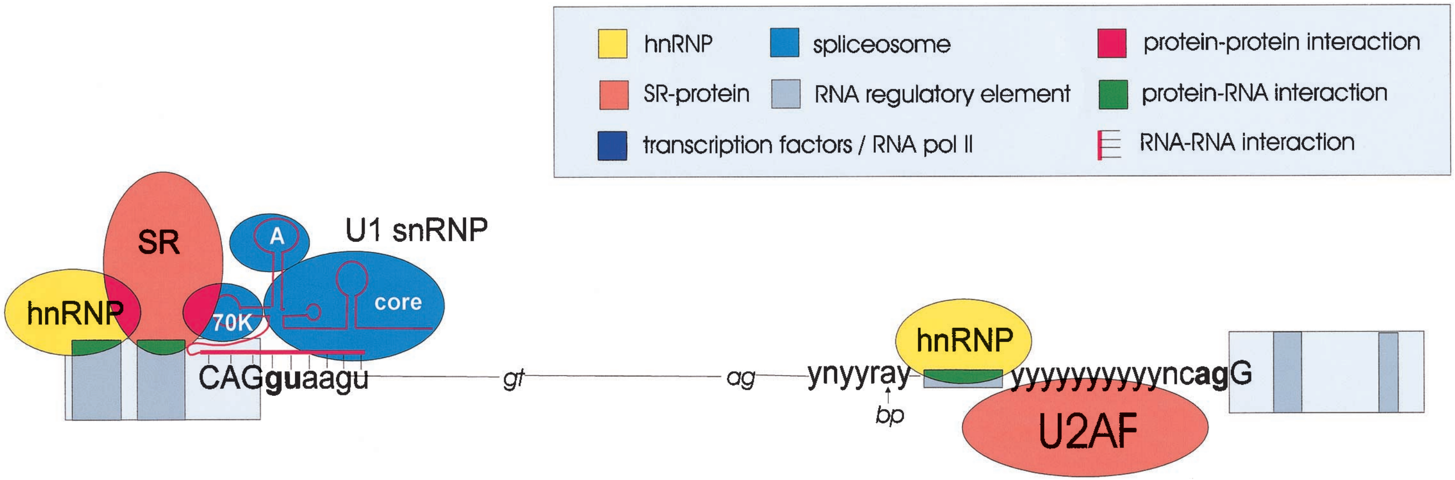

Fig. 1. Elements involved in alternative splicing of pre-mRNA. Exons are indicated as *boxes*, introns as *thin lines*. Splicing regulator elements (enhancers or silencers) are shown as *gray boxes* in exons or as *thin boxes* in introns. The 5′ splice site (CAGguaagu) and 3′ splice site $(y)_{10}$ncagG, as well as the branch point (ynyyray), are indicated (y = c or u, n = a, g, c or u). *Uppercase letters* refer to nucleotides that remain in the mature mRNA. Two major groups of proteins, hnRNPs (*yellow*) and SR or SR related proteins (*orange*), bind to splicing regulator elements; the protein: RNA interaction is shown in *green*. This protein complex assembling around an exon enhancer stabilizes binding of the U1 snRNP close to the 5′ splice site, for example, due to protein:protein interaction between an SR protein and the RS domain of U1 70K (shown in *red*). This allows hybridization (*thick red line with stripes*) of the U1 snRNA (*red*) with the 5′ splice site. The formation of the multiprotein:RNA complex allows discrimination between proper splice sites (*bold letters*) and cryptic splice sites (small *gt ag*) that are frequent in pre-mRNA sequences. Factors at the 3′ splice site include U2AF, which recognizes pyrimidine-rich regions of the 3′ splice sites, and is antagonized by binding of several hnRNPs (e.g., hnRNP I) to elements of the 3′ splice site. *Orange*: SR and SR related proteins; *yellow*: hnRNPs; *green*: protein:RNA interaction; *red*: protein:protein interaction; *thick red line with stripes*: RNA:RNA interaction

in the composition of protein complexes recognizing the splicing enhancer. In addition, cell-type-specific splicing factors have been detected. In *Drosophila*, for example, the expression of the SR protein transformer is female-specific (Boggs et al. 1987) and determines the sex by directing alternative splicing decisions. Other tissue-specific factors include the male germline-specific transformer-2 variant in *D. melanogaster* (Mattox et al. 1990) and *D. virilis* (Chandler et al. 1997), an isoform of its mammalian homologue htra2-beta3 that is expressed only in some tissues (Nayler et al. 1998), the neuron-specific factor NOVA-1 (Jensen et al. 2000) as well as testis and brain-enriched factor rSLM-2 (Stoss et al. 2001) and NSSR (Komatsu et al. 1999). For most of these factors, the tissue-specific target genes remain to be determined. However, a combination of knockout experiments and biochemical analysis allowed the identification of doublesex, fruitless, and transformer-2 as a target of the transformer-2/transformer complex in *Drosophila* (Hoshijima et al. 1991; Mattox and Baker 1991; Heinrichs et al. 1998) and glycine receptor alpha2 and GABA$_A$ pre-mRNA as a target for NOVA-1 (Jensen et al. 2000). Although this analysis is currently limited, it is likely that a given splicing factor will influence several pre-mRNAs. SR-proteins (Fu 1995; Graveley 2000) from all species and splicing regulatory proteins for *Drosophila* (Mount and Salz 2000) have been compiled.

3
Properties of Alternative Spliced Exons

3.1
Several Types of Splicing

The analysis of human intron sequences demonstrates the existence of several intron types. Out of 53,295 human confirmed exons, 98.12% use the canonical GT and AG dinucleotides at the 5′ and 3′ site respectively. Another 0.76% of introns contain GC-AG dinucleotides at this position (Lander et al. 2001); which is comparable to the estimated 1.1% obtained by modeling ESTs on the human genome (Clark and Thanaraj 2002).These introns are processed similarly to the GT-AG introns. They have been compiled in a database and it was found that one in every twenty alternative introns is a GC-AG intron. About 60% of all GC-AG introns are alternatively spliced (Burge et al. 1998; Thanaraj and Clark 2001). Finally, a different class of introns exists that are flanked by AT-AC dinucleotides at the 5′ and 3′ position. These introns are processed by a variant U12 splicing system (Burge et al. 1998). Further, Clark and Thanaraj (2002) observed that 0.4% of the observed introns can be of the type U12-spliceosome GT-AG. Levine and Durbin could identify 404 EST-confirmed U12-type introns in the whole genome, 20 of which had termini dinucleotides different from GT-AG and AT-AC (Levine and Durbin 2001). A systematic survey of mammalian splice sites revealed even more intron sequence diver-