Prediction and statistical analysis of alternatively spliced exons

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

[1] European Bioinformatics Institute,
Wellcome Trust Genome Campus,
Hinxton,
Cambridge, CB10 1SD
U.K.

[2] Institute for Biochemistry
University Erlangen-Nurenberg
Fahrstrasse 17
91054 Erlangen
Germany
Corresponding author
Phone: +49 9131 8524622
Fax: + 49 9131 85 24605
e-mail: Stefan@stamms-lab.net

Abbreviations used: EST: expressed sequence tag, SELEX: systematic evolution of ligands by exponential enrichment; SR-protein: serine/arginine-rich protein; mean: average of a distribution; mode: value that occurs most frequent in a distribution; hnRNP: heterogenous ribonuclear protein;

Abstract

The completion of large genomic sequencing projects revealed that metazoan organisms abundantly use alternative splicing. Alternative 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 length, sub optimal 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.

1.0 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). 59% of the 245 genes present on chromosome 22 are alternatively spliced and the 544 genes of chromosome 19 result in 1859 different messages. The comparison of ESTs with the human genome sequence indicates that 47% of human genes might be alternatively 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 discuss the major findings and the computer programs used to generate them.

2.0 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; Burset et al. 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 is binding 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 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 currently used 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 SR family of proteins (Fu 1995; Manley and Tacke 1996; Graveley 2000) and 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 (Figure 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 (i) an actual competition of the factors for an RNA binding site, (ii) through sequestration of the factors by protein:protein interaction and, (iii) by changes 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.

3.0 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 53295 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 in a way that this similar 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 (Thanaraj and Clark 2001) (Burge et al. 1998). 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, twenty 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 diversity, as introns flanked by GT-AC, GT-CG, GT-TG, AT-AG and GA-AG have been detected (Burset et al. 2000; Burset et al. 2001). However, only a total of 39 such introns have been detected and the exact splicing mechanism remains to be determined. The results of these studies of mammalian introns correlate well with the data from the human genome project. It was found that 99.24% of all human introns are flanked by GT-AG, 0.69% by GC-AG, 0.05% by AT-AC and 0.02% by other nucleotides (Burset et al. 2000; Burset et al. 2001). Weight matrices for some of the non-canonical splice sites and different categories of GT-AG introns have been compiled in AltExtron database (Clark and Thanaraj 2002) (Table 2).

3.2 Types of exons

Alternative exons can be subdivided into different groups. Depending on their splicing mode, they can be classified as cassettes, mutually exclusive, retained intron, and alternative 3' and 5' splice sites (Breitbart et al. 1987)(Figure 2). Database analysis from human curated sets revealed that cassette exons are the most common type, representing about half (53%) of alternative exons (Stamm et al. 2000). This is in agreement with an EST based study, where roughly 61% of alternative exons were found to be either cassette or mutually exclusive exons (Mironov et al. 1999). The next most common splicing modes in this sample of data are alternative 3’ splice sites (10%), mutually exclusive cassette exons (9%), alternative 5’ splice sites (7%) and retained introns (6%). Details of the molecular mechanism that regulates alternative splicing of the various subgroups will most likely be different: in alternative 3’ and 5’ splice sites, there is often a competition between different splice sites. Intronic sequences between mutually exclusive exons are often short and there is evidence for unusual branch point locations in some of these system studied (Helfman and W.M. 1989; Southby et al. 1999). Often, alternatively spliced exons show tissue or developmental specific expression. A survey of the methods used to obtain information about the expression patterns shows that most data are obtained by RT-PCR from whole tissue RNA (Stamm et al. 2000). As a result, expression data will be averaged over a vast number of different cell types. When organs are compared, it is striking that most alternative exons are present in multiple tissues, which is in agreement with EST based studies (Modrek et al. 2001). This could indicate that an intrinsic balance of regulatory factors expressed in all tissues (Hanamura et al. 1998) could be responsible for alternative splice site selection of numerous exons. The organs that express alternative exons devoted only to one tissue were predominantly brain, then muscle, blood and liver (Stamm et al. 2000). Since these data are obtained from the published literature, they could also reflect the current focus on research. However, high expression of the genetic information in the nervous system has been found earlier in clonal and kinetic analysis (Chaudhari and Hahn 1983; Milner and Sutcliffe 1983) and in a recent EST based study (Modrek et al. 2001). The analysis of tissue specificity is further complicated by the fact that not all researchers use the same tissue to determine alternative splicing patterns. A systematic and unified approach is urgently needed here.

3.3 Regulatory elements

Since the regulatory elements reside on the pre-mRNA, the comparison of alternative with constitutive exons should reveal overall regulatory features. Several studies therefore compared the nucleotide usage of alternative exons at the 5’ and 3’ splice sites. A major result from this comparison is that the majority of alternative exons contain splice sites that strongly deviate from the consensus sequence. This deviation is more pronounced in exons that are specific for a certain tissue or developmental stage.

5’splice site

The nucleotide composition at the 5' splice site reflects binding to U1 snRNA (Zhuang and Weiner 1986) as well as an interaction with U6 snRNA (Wise 1993). Overall, vertebrate alternative exons deviate most strongly at the +4 and +5 position. In constitutive cassette exons, the +3 position is occupied by adenosine while in the case of alternative exons it is more often occupied by guanosine (Clark and Thanaraj 2002). In exons that are exclusively expressed in neurons (Figure 3) and muscle cells, there is an additional deviation at the -3 position where splice sites from both categories use more adenosine than the consensus (Stamm et al. 1994; Stamm et al. 2000) (Figure 3). The corresponding bases U5 and U6 of U1 snRNA that would bind to the +4 and +3 positions are modified to pseudouracils (Reddy 1989). It is interesting that alternative exons deviate from the consensus at positions of the 5' splice site that are modified in the complementary base position of U1 and U6 snRNA. None of the 16 human and six drosophila U1 snRNA genes show a sequence diversion at the positions binding to the 5'splice site, but the posttranscriptional modifications have not been systematically studied.

3' Splice site

In general, alternative exons have a more purine rich polypyrimidine tract. The reduced pyrimidine content in the polypyrimidine tract of alternative exons will most likely reduce its affinity towards U2AF, which was shown to bind to the consensus TTTYYYYTNTAGG (Wu et al. 1999). When subclasses of exons are analyzed, the purine content is strongest in retained introns. In contrast, polypyrimidine tracts of mutually exclusive exons adhere well to the consensus sequence. Inspection of subgroups of cassette exons with different biological regulation reveals some remarkable features. Exons that are expressed only in brain, neurons or muscle, as well as all exons that are developmentally regulated, deviate more from the consensus than any other classes of exons. This deviation appears not to be random, as it is most pronounced at the -3 and -10 position of developmentally regulated exons, at the +1, -3, -9 and -11 position of brain specific exons, at the +1, -3 and -9 position of neuron specific exons and at the +1, -7 and -12 position of muscle specific exons. In all these classes, the single most divergent nucleotide usage is the presence of an adenosine at the -3 position of the 3' splice site, (Figure 4) (Stamm et al. 1994; Stamm et al. 2000). The mechanistically implication for this usage is not clear, but in the Holliday like structure proposed for the spliceosome (Steitz 1992) the nucleotide at the -3 position would base pair with the G9 of U1 snRNA, which would favor a cytosine at this position. Mutations of this nucleotide at the -3 position influence exons usage (Epstein et al. 1994; Stamm et al. 1999). Interestingly, the SELEX consensus sequence of U2AF (Wu et al. 1999) has a T at this position and does not reflect the mammalian splice site consensus.

Branch point

The determination of the exact location of the branch point is difficult. In cases where branch points in alternative exons have been studied, the distance between them and the 3' splice site was found to vary significantly. Cases of alternative 3' splice site usage were studied systematically. It was found that 15% of the alternative isoforms show strong poly pyrimidine tract sequences compared to 59% of constitutive exons. In equal fraction (40%) of cases both the normal and alternative isoforms show strong branch point signals (Clark and Thanaraj 2002). This implied that the isoforms from alternative 3' splice site usage probably use the same branch points and it is the variation and the composition in distance that plays a dominant role in determining the competitiveness of the alternative AGs.

3.4 Overall splice site quality

Calculating a score for the splice site can assess the overall quality of a splice site. The score expresses how well the splice site adheres to the consensus sequence. Several methods have been used to calculate splice site scores. (Shapiro and Senapathy 1987; Zhang and Marr 1993; Stamm et al. 1994; Zhang and Marr 1994). In general, the score expresses the coincidence of a splice site with the consensus sequence. The higher the score, the better the splice site follows the consensus. When the distribution of splice site scores of alternative exons is analyzed, several features emerge. Alternative splice sites have a broader distribution and a large proportion of the exons have sub optimal splice site scores (Stamm et al. 2000; Clark and Thanaraj 2002). However, about half of the alternative cassette exons are surrounded by splice sites within the range of constitutive splice sites. The weakest 3' splice sites are found in retained introns and alternative 3' splice sites. In a similar way, the weakest individual 5' splice sites are found in retained introns and in alternative exons with an alternative 5' splice site. According to the exon definition model (Berget 1995), cassette exons were analyzed by adding the 3' and 5' splice site scores of a single exon. When alternative exons from different tissues were compared, exons that were only expressed in a single tissue were characterized by the weakest splice site scores (Figure 5) (Stamm et al. 1994; Stamm et al. 2000).

3.5 Exonic elements

Since a considerate number of alternative exons are flanked by sub optimal splice sites, the question arises, why and how these exons are recognized. Experimental data show that purine-rich and AC rich sequences can act as enhancers. These sequences were either identified by mutagenesis in model substrates, found as consensus sequence that binds to regulatory proteins, or were shown to cause human disease when mutated (Table 1). In order to find common motifs in groups of exons with sub optimal splice sites, these exons were compared using a Gibbs sampler. This algorithm allows the determination of redundant sequence motifs that are least likely to occur by chance in a given nucleotide composition (Neuwald et al. 1995). When exons with sub optimal splice sites were tested, the motif GNCNTCCAA was found to be being highly abundant. Using the same algorithm, subgroups of alternative exons that were only expressed in neurons, brain, blood, liver, muscle, testis and thymus were tested. Only short motifs from 4 to 9 nucleotides could be found (Figure 6). Common to all motifs is a high degree of degeneracy, which has also been found when SELEX procedures were employed *in vivo* or *in vitro* (Lui et al. 1998; Lui et al. 1999; Tacke and Manley 1999) . Furthermore, the occurrence of a given motif was always restricted only to certain exons in a given subclass. When all motifs were considered together, the nucleotide composition was A=22%, G=28%, C=33% and T=17%, indicating that overall these motifs are not purine rich. Another interesting feature of these motifs is that they do not contain many GAR-enhancer motifs that have been described for a variety of systems. Similar, AAC and ACC triplets could not be detected, which marks the absence of A/C rich enhancer (ACE) motifs (Coulter et al. 1997). Together, these data argue that in addition to the previously described GAR and ACE enhancer sequences, other motifs exist in alternatively spliced exons. Often, these motifs are pyrimidine rich. Interestingly, such pyrimidine rich motifs have been previously identified as activating motifs in *in vitro* SELEX procedures (Tian and Kole 1995) . However, the trans-acting factors binding to these motifs remain to be determined.

In the future, comparison of the complete mouse and human genomes will allow the identification of intronic regulatory sequences. The comparison between the genomes of the related nematodes C. elegans and C. briggsae has been performed using a newly developed three pass algorithm (Kent and Zahler 2000). It was found that several alternatively spliced genes, such as the let-2 and bli-4 genes, contain conserved intronic GT-repeat regions that could have regulatory functions (Kent and Zahler 2000). A similar genome wide analysis in vertebrate system is expected to yield intronic motifs, since the comparison of model systems already revealed conserved regions (Zhang et al. 1999). However, due to the larger size of vertebrate introns, these analyses will be more difficult.

3.6 Length of alternative exons

The length of exons is conserved among species. Most internal cassette exons have a length between 50 to 200 nt in humans, drosophila and C.elegans. However, in the length distribution in C.elegans and drosophila, more exons can be found in the classes larger than 250 nt, which results in a larger mean size for internal exons (218 nt for C.elegans, versus 145 nt for humans). The difference in size distribution is even more strikingly, when introns are analyzed. Here, the mean size of introns is 267 nt for C. elegans, 487 nt for drosophila but 3365 nt for humans (Lander et al. 2001).

Experimental evidence suggests that the length of alternative exons (Black 1991) , as well as the length of its flanking introns (Bell et al. 1998) are involved in alternative splicing regulation. A computer study showed that human introns associated with alternative 3’ splice sites seem to be shorter on average, with a median length of 625 nts (Clark and Thanaraj 2002). When all types of vertebrate alternative exons are considered, they have a mean length of 174± 288 nt (mean of the distribution ± standard deviation), which is comparable to the distribution of human cassette exons (137±123).

However, when cassette exons that are expressed in a single tissue, namely brain, neurons or muscle, were analyzed, they were found to be shorter on average (78±58; 101± 125 and 58±55) than constitutive exons (Figure 7). Exons with alternative 5’ and 3’ splice sites still have means of 136±187 and 144±199, respectively, which is comparable to constitutive exons. Inspection of the length distribution shows that the mode, indicating the highest frequency of exons, is smaller in alternative exons than in constitutive ones. This is apparent with alternative 3' and 5' splice sites, for which the mode of the distribution is within 1-25 and therefore significantly smaller than the mode for constitutive exons, which is 100-125. In contrast, exons arising from intron retention are on average larger than constitutive cassette exons (308±320) and smaller than constitutive introns that average around 3365 in humans (Lander et al. 2001). However, the mode of their distribution is in the range of 75-100 nt. Common to all distributions is that alternative exon length does not follow a normal distribution but is skewed toward the smaller size exons, which underlines a bias towards smaller exons in alternative splicing.

The decreased size of alternative exons expressed only in a single tissue could indicate the necessity of a certain amount of RNA binding factors to be assembled on constitutive exons prior to their recognition. This has been observed in constitutive splicing as well: the 42 detected human exons that are smaller than 19 nt contain 72% purines, which are most likely part of exonic GAR-type enhancers (Lander et al. 2001).

3.7 A major function of alternative exons are the introduction of premature stop codons

From sequence comparison, a general function of alternative exons is not readily visible. Only in 55-60% of alternative splicing events, the coding frame is maintained (Clark and Thanaraj 2002). In a sample of 1000 alternative spliced exons compiled from the literature, about 22% of the exons contained a stop codon or introduced a frameshift, resulting in a premature stop codon (Stamm et al. 2000), which is in agreement with EST based studies that found 19% frameshifts (Modrek et al. 2001). Therefore, the introduction of premature stop codons seems to be an important biological feature of alternative exons. These stop codons have been compiled and were found to often disrupt purine-rich or AC rich enhancer sequences (Valentine 1998). Most of these stop codons appear in coding exons and fulfill the requirements for nonsense mediated decay. In conclusion, the introduction of stop codons and the following nonsense mediated decay seem to be an important biological role for alternatively spliced exons.

4.0 Bioinformatics Resources to identify alternative exons

Publicly available databases contain a large amount of transcript sequences, which are present in mainly two forms: partial transcript sequences such as ESTs and full-length mRNA sequences. The EST sequences are available from dbEST or from EMBL/GenBank databases. Since partial and complete sequencing of mRNAs from many different tissues, developmental stages, and pathologies creates these collections they contain diverse transcripts. There is a high amount of redundancy in EST sequences. Related EST sequences that correspond to a unique mRNA transcript are available as clusters and consensus sequences and are compiled in TIGR Gene Indices, NCBI Unigene clusters, and SANBI STACK clusters. TIGR gene indices are created by comparing EST sequences with transcripts and clustering them if there is a minimum of 40 base pair match with at least 95% of similarity and if there is only a maximum unmatched region of 20 base pairs. These clusters are assembled into consensus sequences. The Unigene database clusters EST's in a similar manner but does not create a consensus sequence. In the STACK database the gene sequences are organized according to tissue and each gene is represented with alignments of its expressed fragments; consensus sequences are presented (Table 3).

4.1 Detection of alternative splice events using sequence comparison

To date, data on alternative splicing have derived by either examining annotated gene forms in nucleotide and protein sequence databases (such as EMBL/GenBank/DDBJ, SwissProt) and bibliography databases such as MedLine or through examining alignments of nucleic acid sequences. The major sequences alignments are between a) genomic DNA and ESTs, b) mRNA and ESTs, as well as c) mRNA and mRNA. In these approaches, BLAST is often used to identify regions of identical nucleotide bases denoted as high-scoring segments. A comparison that includes genomic DNA gives the most information, as the gene structure and splice sites can be delineated.

a) When genomic *DNA and ESTs* are aligned, the gaps on the DNA sequence correspond to putative introns. They are further analyzed by comparing their terminal bases with splice site consensus sequences. The satisfying gaps are denoted as introns and those that are not compatible are ignored. The region between two such confirmed introns is denoted as confirmed exon. If these confirmed exons overlap with one another, they are called alternative exons. Specialized versions of BLAST such as Sim4 and SplicedAlignment are used for this computation (Table 3).

b) Gaps that are produced in the alignment of *mRNA with EST* sequences indicate different isoforms. Databases that used this approach present the data in terms of insertions/deletions in the transcripts rather than in terms of exon-intron structures (Brett et al. 2002) and do not discriminate between the different alternative exon types.

c) *mRNA-mRNA* alignments were mainly performed from mouse sequences by the RIKEN research group (Kawai et al. 2001).This program attempts to detect alternative splice events through clustering cDNAs from redundant clone sets. Redundant cDNAs showing more than two portions of alignments longer than 20 bases are clustered. Similar to that of mRNA-EST alignments, the gaps/inserts are deduced as alternative splice events. Recent databases based on the above approaches are listed in Table 2.

4.2 Computer Programs for detecting alternative splicing

Using alignments between different nucleic acids, alternative splicing events can be detected in databases. Different programs are used, depending whether the gene structure is known or needs to be determined.

Finding alternative exons from known gene structures

To find alternative exons from known gene structures, Thanaraj (Thanaraj 1999) and Hide (Hide et al. 2001) first constructed 'theoretical' mRNA parts from the gene and then compared it with the EST database. Exon skipping events that correspond to cassette exons were identified by concatenating a 50-bp tag from the 3' terminus of the proceeding exon with a 50-bp tag from the 5' terminus of each of succeeding exons. This creates a set of all consecutive and nonconsecutive exon-exon junctions. Each of such 100-bp exon-exon constructs is searched for similarity with the collection of EST sequences. A skipping event is reported when an EST aligns with the construct without a gap.

Exon isoform events are created by alternative 5' and 3' splice sites. They are detected by connecting 50-bp exon sequence tags from the 5' and 3' end of a given intron. These exon-exon constructs are searched for similarity against EST databases. While an EST sequence that shows uninterrupted similarity at the junction confirms the normal event, a gap in the alignment at the splice junction indicates either truncation or extension of intron sequences depending on whether the gap occurred in the EST sequence or in the exon-exon construct. The gaps are further mapped to the gene sequence and the end nucleotides are checked for splice site consensus sequences.

To detect retained introns, for every intron in the gene structure, two constructs, the exon-intron and the intron-exon constructs are built by using the 50-bp tags flanking the exon-intron junction or those flanking the intron-exon junctions. These constructs are searched for similarity with the collection of EST sequences. An EST sequence showing similarity to both the constructs belonging to an intron indicates potential candidates for intron retention events.

Finding alternative splice events in genome regions of unknown gene structures

The increasing amount of genomic sequence data makes it necessary to identify (alternative) splice sites in sequences of unknown structure. This is performed by using either consensus EST cluster sequence or self-clustering methods to map and compare individual EST sequences with genomic DNA.

Human ESTs were used for a genome-wide detection of alternative splicing in humans (Modrek et al. 2001). In this work the genome region that aligns with an EST consensus sequence is identified. Every EST from the cluster is then searched for similarity with the identified genome region and the potential splice sites are identified. The putative splice sites are then validated against splice site consensus weight matrices (generally the GT-AG type). Subsequently all the mutually exclusive splice site pairs are identified as alternative splicing events.

Two programs, procrustes-EST and TAP have been used to identify alternative exons with a self-clustering approach.

Procrustes-EST (Gelfand et al. 1999) uses EST contigs and assemblies from TIGR Human Gene Index as the transcript information to delineate alternative intron-exon structures. First, the EST contigs that show similarity to a given genome region are selected. Then candidate splice sites are predicted on the genome region and then chains of exons with the highest similarity score with one or more of the EST contigs are identified. Such predicted exon-intron structures are merged into superstructures through the following simple procedure: Each of triples (intron-exon-intron) from these predicted structures is considered against each other - if the right intron of a triple coincided with the left intron of a second triple, the triples are merged into a superstructure. This procedure is repeated until no triples could be added to the constructed superstructure. All such possible superstructures are constructed. A superstructure that intersected neither the annotated gene structure nor any other superstructure in a common exon is omitted as an orphan. The remaining superstructures represent the different alternative gene structures for the genome region.

TAP (Transcript Assembly Program) (Kan et al. 2001) is built to delineate gene structures using EST sequences that align in a DNA region on the genome. Firstly, exon segments are identified through DNA-EST alignments. The gaps between exons are accepted as introns (splice junction pairs), if they have the consensus GT-AG sequences or have more than two EST sequences supporting such gaps. Such observed splice junction pairs along the length of the gene are sorted out to assemble the joint gene structures; the algorithm uses the EST-encoded connectivity and redundancy information to sort out the complex alternative splicing patterns. The connectivity information from EST sequences is decoded into four types and scored appropriately. The program scores conflicting, contiguous, transitive and gapped connections and builds a matrix to record the connectivity relationship between pairs of introns. The gene structure is assembled by tracing a path through the filled connectivity matrix from 5’ to 3’ direction in a manner that at each elongation step from an intron i to j, the connection with the maximum matrix score is chosen. Thus, at each step contiguous connection takes precedence over transitive connection, which in turn takes precedence over gapped connections. Mutually exclusive connections are thus not included in the same path. Alternative paths are generated by branching the path when multiple downstream splice pairs have the same connectivity score.

4.3 Limitations, problems and accuracy of predicting alternative splicing events

The *expression of alternative exons* in a given tissue, cell type or developmental stage is not addressed by most computer generated databases. Attempts have been made to decipher the tissue specificity of alternative poly-A site usage (Beaudoing and Gautheret 2001). One of the major problems to delineate tissue specificity of alternative splicing events is the lack of a proper classification systems for EST libraries. Currently the anatomical site, pathology conditions, and developmental stage of EST libraries is ill defined. However, there are attempts for improving the situation, as NCBI has recently released its own classifiers.

Repeats, paralogous genes, and duplicated or multiple copies of the genes cause false positive and negative exon predictions. Repeats in gene sequences as well as in inter-genic regions are common in mammalian genomes (Lander et al. 2001). Furthermore, genomic regions and genes can be duplicated. Therefore, it is often difficult to unambiguously assign a transcript or EST cluster to a gene. As a result highly similar gene products could be interpreted as the result of alternative splicing, but are generated by paralogous genes. Repeats are masked before gene-transcript alignments are carried out. Masking is done by checking every gene for the occurrence of known repeats, which requires an *a priori* knowledge of repeat sequences. Duplicated genes and redundant gene entries in a start-up data set are identified by checking for similarity among the entries of start-up gene set and retaining one representative entry. However, it is impossible to test whether a matching EST sequence is derived from the representative gene entry or from one of the removed gene entries. One possible solution is to identify repeats in the alignment data by observing where multiple matches to EST sequences occur. When two or more genes show matches to the same region of a transcript, these matches are considered to be repeat matches and are discarded. Similarly, a transcript is also removed if a region of it shows matches to more than one region on the gene (Clark and Thanaraj 2002). This method has the distinct advantage that no *a priori* knowledge of the repetitive elements is required.

Quality of alignments. Generally, stringent criteria such as that a high scoring segment has at least 95% of identity and the E value is $< 10^{-15}$ are required to be imposed. It is our experience that these E values often remove any high scoring segment of length < 30-35 bases, which is characteristic for exons generated by alternative 5' and 3' splice site usage. Performing local alignments with slightly relaxed criteria between the EST gap regions and the region on the gene between the partial structure could solve this problem (Clark and Thanaraj 2002).

The Exact assignment of splice sites is often difficult, but is necessary for the *ab initio* gene prediction programs and to delineate gene structures from the gene-EST alignments. In gene-EST alignments, there can be more than one candidate splice site pair. Since this does not change the coding sequence, computer programs often ignore it. However, for studies related to splicing, it is essential to exactly pinpoint the splice sites and programs such as SPC (SpliceProximalCheck) have been developed for it (Thanaraj and Robinson 2000).

Most of the prediction programs rely on EST data, which have several intrinsic problems. Programs that assemble gene structures by matching and overlapping EST sequences do not discriminate whether an EST is derived from the same or from different clones. Alternative splicing events, such as mutually exclusive exons can therefore not be identified. Alternative variants are often rare (Hide et al. 2001; Kan et al. 2001; Clark and Thanaraj 2002) and many programs that require at least two or more EST sequences before confirming an event will ignore them.

Since ESTs are derived by sequencing the ends of a cDNAs, the EST-coverage is non-uniform along the length of the gene. The internal regions of the gene have low coverage; in a similar manner, the 3' ends of the genes can be over represented. These bias the EST data sets to events in the 5' and 3' UTR and under represents internal exons. This bias is evident when studies that include genomic data are compared with pure EST based studies. A genome wide survey of alternative splicing finds 74% of alternative events in the coding region (Modrek et al. 2001), whereas datasets derived purely from ESTs find only 20% such events in the coding region (Mironov et al. 1999).

The success rate of detecting alternative splice events in silico has been tested by experimental validation of a limited set of the predicted alternative events (Brett et al. 2002). RT-PCR and subsequent sequencing of a limited set of the predicted alternative events indicated a success rate of 92% and a false negative rate of 2.5%.

4.4 Ad initio prediction of alternative exons

Currently the efforts that derive data on alternative splice events do not 'predict' the events but instead locate them by positioning EST sequences on either the genes or mRNAs. The current gene prediction programs do not predict alternative gene structures. Though some of the available programs such as GenScan and HMMgene can predict sub optimal genes (Table 3), it has not been assessed whether such predicted sub optimal genes correspond to alternative gene structures. Predicting alternative exons from genomic data is still at an early stage. At the best, we are now able to detect alternative spliced exons based on similarity to

EST sequences. Since alternative splicing is signaled by weak signals as marked by poor consensus sequences, poor correlation among the different structural elements, and use of minor intron/exon types, prediction of them is difficult. It is also the case that alternative splicing often results in altered reading frames in downstream coding exons or to premature stop codons. Thus while selection of optimal exons is the key in the current gene prediction programs, the emphasis on alternative gene structure prediction is the choice of sub-optimal exons. Further, the EST-based methodologies locate alternative exons in a non-contextual manner along the native gene structure and do not illustrate which combinations of these exons exist in a full-length isoform transcripts. Thus, efforts need to be concentrated on dynamic programming procedures that can put together sub-optimal and optimal exons in a proper context.

5.0 Overview of available databases and programs

During the last years, several databases of alternatively spliced exons were generated (Table 2). Their data sets have derived from one of two main approaches: i) examining annotated forms in databases such as EMBL/GenBank/DDBJ, SWISS-PROT and MedLine – these include ASDB, Alternative-Exon Database, and AsMamDB, or ii) examining alignments of EST/cDNA sequences with genomic DNA sequences – these include the Intronerator and AltExtron. These individual databases differ in the sizes of the generated data sets, and in the detail of the methodologies employed and hence in the quality. The first database generated was a compilation of neuron-specific alternative exons that was based on sequences published in peer reviewed journals (Stamm et al. 1994). A more general alternative splicing database (ASBD) was developed later. In the first version of ASDB protein sequences generated by alternative splicing were identified by cluster analysis of Swiss-Prot entries (Gelfand et al. 1999). Later this database was extended with a DNA division that contains complete genes for which alternative splicing has been mentioned in the Genbank annotations (Dralyuk et al. 2000). AsMamDB was developed to systematic study alternative spliced genes in mammalian systems and contains alternative exons based on human, mouse and rat Genbank entries (Ji et al. 2001). Another specialized database is the intronerator that compiles cDNA alignments and a catalog of alternatively spliced genes from C. elegans (Kent and Zahler 2000). A compilation of human genes that use GC-AG introns shows that one in every twenty alternative introns is a GC-AG intron (Thanaraj and Clark 2001). Finally, starting from an early compilation (Mount 1982), several databases of splice sites are available. SpliceDB is a database of mammalian splice sites (Burset et al. 2001) and mutations in splice sites that are implicated in human diseases have been compiled (Nakai and Sakamoto 1994). Splice sites used by GC-AG introns have been compiled on the AltExtron web data (Thanaraj and Clark 2001). Intronic sequences were compiled in the intron and sequence information database (ISIS) (Croft et al. 1999). Finally, RNA recognition motif binding proteins (Birney et al. 1993) and all drosophila splicing regulatory proteins (Mount and Salz 2000) have been compiled. The various computer programs and specialized databases that were used to predict and analyze alternative exons are compiled in Table 3.

Acknowledgments

This work was supported by the EBI (TAT) and the Deutsche Forschungsgemeinschaft (Sta399/7-1) and SFB473 C8 to SS.

Figure and Table legends

Figure 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). Upper-case 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 U170K (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 multi-protein: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

Figure 2: Types of alternative exons

Type of alternative splicing events: Alternative exons are shown as boxes with different shading. Flanking constitutive exons are shown as white boxes. The open arrow indicates the position of the alternative 3' splice site analyzed; a closed arrow indicates the position of the 5' splice sites analyzed.

Figure 3: Nucleotide usage at the 5' splice site of neuron-specific exons.

The percent nucleotide usage of constitutively (left) and neuron-specific alternatively spliced exons (right) is pair wise compared. The U1 sequence and the vertebrate consensus splice site sequence are shown at the bottom. The different nucleotides are indicated by different patterns, as indicated in the figure.

Figure 4: Nucleotide usage at the 3' splice site of neuron specific exons

The percent nucleotide usage of constitutively (left) and neuron-specific alternatively spliced exons (right) is pair wise compared. The vertebrate consensus sequence is indicated at the bottom, exon sequences are in capital letters. Note the use of A in alternatively spliced exons at the -3 position. The different nucleotides are indicated using the same patterns as in Figure 1.

Figure 5: Distribution of splice site scores in constitutive and alternative exons

Top: schematic diagram that illustrates the location of the splice sites

A. Splice site scores of constitutive exons. The combined scores for an exon are plotted on the x-axis, the number of exons on the y-axis. The ◆ sign indicates the mean of the distribution in human cassette exons. An exon surrounded by a "perfect" splice sites would have a score of 27.8. The mean of the distribution is 16.9 for constitutively spliced exons.

B. Splice site scores of human cassette exons that are expressed in multiple tissues (more than two tissues, but not detectable in all tissues analyzed). The mean of the distribution is 12.9. The box indicates exons surrounded with sub optimal splice sites that were analyzed by sequence comparison to find common motifs.

Figure 6: Nucleotide motifs found in exons specific for certain tissues.

The motifs were identified using a Gibbs algorithm on the subset of weak exons present in multiple tissues and on exons that are specifically expressed in one tissue.

Figure 7: Length distribution of alternative exons

(A) Histogram showing the length distribution of human constitutive cassette exons. The ◆ sign indicates the mean of the distribution in human constitutive cassette exons. For comparison the length distribution of brain cassette exons are shown. The y-axis of each histogram represents the numbers of exons; the x-axis represents the nucleotide length.

Table 1: Compilation of RNA elements that influence splice site selection

The first column shows the gene and exon that contains the RNA element, which is characterized in the next columns according to its type (ESE: exonic sequence element, ISE: intronic sequence element) and sequence. Trans-acting factors are indicated in bold under the sequence if they were identified experimentally. Most RNA elements will work in combination with additional RNA regulatory sequences that are not shown. Meth. Indicated the experimental method used: 1. deletion analysis, 2: in vivo splicing assay, 3: in vitro splicing assay, 4: gel mobility shifts; 5: autogenesis; 6: in vitro binding; 7: UV-crosslink; 8: competition experiment; 9: SELEX; 10: immunoprecipitation; 11: spliceosomal complex formation; 12: nuclease protection

Table 2: Overview of existing databases that compile alternative exons, regulatory proteins and diseases

Table 3: Tools to identify predict and analyze alternative splicing events.

6.0 Literature cited:

Amendt, B. A., Z. H. Si, et al. (1995). "Presence of exon splicing silencers within human immunodeficiency virus type 1 tat exon 2 and tat-rev exon 3: evidence for inhibition mediated by cellular factors." Mol Cell Biol **15**(11): 6480.

Ashiya, M. and P. J. Grabowski (1997). "A neuron-specific splicing switch mediated by an array of pre-mRNA repressor sites: evidence of a regulatory role for the polypyrimidine tract binding protein and a brain-specific PTB counterpart." Rna **3**(9): 996-1015.

Ayane, M., U. Preuss, et al. (1991). "A differentially expressed murine RNA encoding a protein with similarities to two types of nucleic acid binding motifs." Nucl. Acids Res. **19**: 1273-1278.

Beaudoing, E. and D. Gautheret (2001). "Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data." Genome Res **11**(9): 1520-6.

Bell, M. Y., A. E. Cowper, et al. (1998). "Influence of intron lenght on alternative splicing of CD44." Mol. Cell. Biol. **18**: 5930-5941.

Berget, S. M. (1995). "Exon Recognition in vertebrate splicing." J. Biol. Chem. **270**: 2411-2414.

Birney, E., S. Kumar, et al. (1993). "Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors." Nucleic Acids Res **21**(25): 5803-16.

Black, D. L. (1991). "Does steric interference between splice sites block the splicing of a short c-src neuron-specific exon in non neuronal cells?" Genes & Dev. **5**: 389-402.

Boggs, R. T., P. Gregor, et al. (1987). "Regulation of sexual differentiation in D. melanogaster via alternative splicing of RNA from the transformer gene." Cell **50**(5): 739-47.

Breitbart, R. E., A. Andreadis, et al. (1987). "Alternative splicing: a ubiquitous mechanism for the generation of multiple protein isoforms from single genes." Annu Rev Biochem **56**: 467-95.

Brett, D., H. Pospisil, et al. (2002). "Alternative splicing and genome complexity." Nat Genet **30**(1): 29-30.

Bruzik, J. P. and T. Maniatis (1995). "Enhancer-dependent interaction between 5' and 3' splice sites in trans." Proc Natl Acad Sci U S A **92**(15): 7056-9.

Burge, C. B., R. A. Padgett, et al. (1998). "Evolutionary fates and origins of U12-type introns." Mol Cell **2**(6): 773-85.

Burset, M., I. A. Seledtsov, et al. (2000). "Analysis of canonical and non-canonical splice sites in mammalian genomes." Nucleic Acids Res **28**(21): 4364-75.

Burset, M., I. A. Seledtsov, et al. (2001). "SpliceDB: database of canonical and non-canonical mammalian splice sites." Nucleic Acids Res **29**(1): 255-9.

Caceres, J. F., S. Stamm, et al. (1994). "Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors." Science **265**(5179): 1706-9.

Caputi, M., G. Casari, et al. (1994). "A novel bipartite splicing enhancer modulates the differential processing of the human fibronectin EDA exon." Nucleic Acids Res **22**(6): 1018-22.

Carlo, T., D. A. Sterner, et al. (1996). "An intron splicing enhancer containing a G-rich repeat facilitates inclusion of a vertebrate micro-exon." Rna **2**(4): 342-53.

Chandler, D., M. E. McGuffin, et al. (1997). "Evolutionary conservation of regulatory strategies for the sex determination factor transformer-2." Mol. Cell. Biol.: 2908-2919.

Chaudhari, N. and W. E. Hahn (1983). "Genetic expression in the developing brain." Science **220**: 924-928.

Clark, F. and T. A. Thanaraj (2002). "Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from humans." Hum. Mol. Gent. **11**: 1-14.

Cooper, T. A. (1998). "Muscle-specific splicing of a heterologous exon mediated by a single muscle-specific splicing enhancer from the cardiac troponin T gene." Mol Cell Biol **18**(8): 4519-25.

Cooper, T. A. and W. Mattox (1997). "The Regulation of Splice-Site Selection, and Its Role in Human Disease." Am.J.Hum.Genet. **61**: 259-266.

Coulter, L. R., M. A. Landree, et al. (1997). "Identification of a new class of exonic splicing enhancers by in vivo selection [published erratum appears in Mol Cell Biol 1997 Jun;17(6):3468]." Mol Cell Biol **17**(4): 2143-50.

Croft, L., S. Schandorff, et al. (1999). "ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome." Nat. Genet. **24**: 340-341.

Del Gatto, F. and R. Breathnach (1995). "Exon and intron sequences, respectively, repress and activate splicing of a fibroblast growth factor receptor 2 alternative exon." Mol Cell Biol **15**(9): 4825-34.

Dralyuk, I., M. Brudno, et al. (2000). "ASDB: database of alternatively spliced genes." Nucleic Acids Res **28**(1): 296-7.

Elrick, L. L., M. B. Humphrey, et al. (1998). "A short sequence within two purine-rich enhancers determines 5' splice site specificity." Mol Cell Biol **18**(1): 343-52.

Epstein, J. A., T. Glaser, et al. (1994). "Two independent and interactive DNA-binding subdomains of the Pax6 paired domain are regulated by alternative splicing." Genes Dev **8**(17): 2022-34.

Fogel, B. L. and M. T. McNally (2000). "A cellular protein, hnRNP H, binds to the negative regulator of splicing element from Rous sarcoma virus." J Biol Chem **275**(41): 32371-8.

Fu, X.-D. (1995). "The superfamily of arginine/serine-rich splicing factors." RNA **1**: 663-680.

Gallego, M. E., R. Gattoni, et al. (1997). "The SR splicing factors ASF/SF2 and SC35 have antagonistic effects on intronic enhancer-dependent splicing of the beta-tropomyosin alternative exon 6A." Embo J **16**(7): 1772-84.

Gelfand, M. S., I. Dubchak, et al. (1999). "ASDB: database of alternatively spliced genes." Nucleic Acids Res **27**(1): 301-2.

Graham, I. R., M. Hamshere, et al. (1992). "Alternative splicing of a human alpha-tropomyosin muscle-specific exon: identification of determining sequences." Mol Cell Biol **12**(9): 3872-82.

Graveley, B. R. (2000). "Sorting out the complexity of SR protein functions." RNA **6**: 1197-1211.

Hanamura, A., J. F. Caceres, et al. (1998). "Regulated tissue-specific expression of antagonistic pre-mRNA splicing factors." RNA **4**: 430-444.

Hastings, M. L. and A. R. Krainer (2001). "Pre-mRNA splicing in the new millennium." Curr Opin Cell Biol **13**(3): 302-9.

Heinrichs, V., L. C. Ryner, et al. (1998). "Regulation of sex-specific selection of fruitless 5' splice sites by transformer and transformer-2." Mol. Cell. Biol. **18**: 450-458.

Helfman, D. M. and R. W.M. (1989). "Branch point selection in alternative splicing of tropomyosin pre-mRNAs." Nucl. Acids Res. **17**: 5633-5640.

Hertel, K. J. and T. Maniatis (1998). "The function of multisite splicing enhancers." Mol. Cell **1**: 449-455.

Hide, W. A., V. N. Babenko, et al. (2001). "The contribution of exon-skipping events on chromosome 22 to protein coding diversity." Genome Res **11**(11): 1848-53.

Hoshijima, K., K. Inoue, et al. (1991). "Control of doublesex alternative splicing by transformer and transformer-2 in Drosophila." Science **1991**(252): 833-836.

Huh, G. S. and R. O. Hynes (1994). "Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element." Genes Dev **8**(13): 1561-74.

Humphrey, M. B., J. Bryan, et al. (1995). "A 32-nucleotide exon-splicing enhancer regulates usage of competing 5' splice sites in a differential internal exon." Mol Cell Biol **15**(8): 3979-88.

Jensen, K. B., B. K. Drege, et al. (2000). "Nova-1 regulates neuron-specific alternative splicing and is essential for neuronal viability." Neuron **25**: 359-371.

Ji, H., Q. Zhou, et al. (2001). "AsMamDB: an alternative splice database of mammals." Nucleic Acids Research **29**: 260-263.

Jumaa, H. and P. J. Nielsen (1997). "The splicing factor SRp20 modifies splicing of its own mRNA and ASF/SF2 antagonizes this regulation." EMBO J. **16**: 5077-5085.

Kamma, H., D. S. Portman, et al. (1995). "Cell type specific expression of hnRNP proteins." Exp. Cell Res. **221**: 187-196.

Kan, Z., E. C. Rouchka, et al. (2001). "Gene structure prediction and alternative splicing analysis using genomically aligned ESTs." Genome Res **11**(5): 889-900.

Kawai, J., A. Shinagawa, et al. (2001). "Functional annotation of a full-length mouse cDNA collection." Nature **409**(6821): 685-90.

Kawamoto, S. (1996). "Neuron-specific alternative splicing of nonmuscle myosin II heavy chain-B pre-mRNA requires a cis-acting intron sequence." J Biol Chem **271**(30): 17613-6.

Kent, W. J. and A. M. Zahler (2000). "Conservation, regulation, synteny, and introns in a large-scale C. briggsae-C. elegans genomic alignment." Genome Res **10**(8): 1115-25.

Kent, W. J. and A. M. Zahler (2000). "The intronerator: exploring introns and alternative splicing in Caenorhabditis elegans." Nucleic Acids Res **28**(1): 91-3.

Komatsu, M., E. Kominami, et al. (1999). "Cloning and characterization of two neural-salient serine/arginine-rich (NSSR) proteins involved in the regulation of alternative splicing in neurons." Genes to Cells **4**: 593-606.

König, H., H. Ponta, et al. (1998). "Coupling of signal transduction to alternative pre-mRNA splicing by a composite splice regulator." EMBO J. **10**: 2904-2913.

Kosaki, A., J. Nelson, et al. (1998). "Identification of Intron and Exon Sequences involved in Alternative Slicing of Insulin Receptor Pre-mRNA." J. Biol. Chem. **273**: 10331-10337.

Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." Nature **409**(6822): 860-921.

Levine, A. and R. Durbin (2001). "A computational scan for U12-dependent introns in the human genome sequence." Nucleic Acids Res **29**(19): 4006-13.

Libri, D., M. Goux-Pelletan, et al. (1990). "Exon as well as intron sequences are cis-regulating elements for the mutually exclusive alternative splicing of the beta tropomyosin gene." Mol Cell Biol **10**(10): 5036-46.

Libri, D., A. Piseri, et al. (1991). "Tissue-specific splicing in vivo of the beta-tropomyosin gene: dependence on an RNA secondary structure." Science **252**(5014): 1842-5.

Lou, H. and R. F. Gagel (1999). "Mechanism of tissue-specific alternative RNA processing of the calcitonin CGRP gene." Front Horm Res **25**: 18-33.

Lou, H., R. F. Gagel, et al. (1996). "An intron enhancer recognized by splicing factors activates polyadenylation." Genes Dev **10**(2): 208-19.

Lui, H.-X., S. L. Chew, et al. (1999). "Exonic splicing enhancer motif recognized by human SC35 under splicing conditions." Mol. Cell. Biol. **20**: 1063-1071.

Lui, H.-X., M. Zhang, et al. (1998). "Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins." Genes and Dev. **12**: 1998-2012.

Lynch, K. W. and T. Maniatis (1995). "Synergistic interactions between two distinct elements of a regulated splicing enhancer." Genes Dev **9**(3): 284-93.

Manley, J. L. and R. Tacke (1996). "SR proteins and splicing control." Genes and Dev. **10**: 1569-1579.

Mattox, W. and B. S. Baker (1991). "Autoregulation of the splicing of transcripts from the transformer-2 gene of Drosophila." Genes and Dev. **5**: 786-796.

Mattox, W., M. J. Palmer, et al. (1990). "Alternative splicing of the sex determination gene transformer-2 is sex-specific in the germ line but not in the soma." Genes and Dev **4**: 789-805.

Mayeda, A., D. M. Helfman, et al. (1993). "Modulation of exon skipping and inclusion by heterogeneous nuclear ribonucleoprotein A1 and pre-mRNA splicing factor SF2/ASF [published erratum appears in Mol Cell Biol 1993 Jul;13(7):4458]." Mol Cell Biol **13**(5): 2993-3001.

Mayeda, A. and A. R. Krainer (1992). "Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2." Cell **68**(2): 365-75.

McCarthy, E. M. and J. A. Phillips, 3rd (1998). "Characterization of an intron splice enhancer that regulates alternative splicing of human GH pre-mRNA." Hum Mol Genet **7**(9): 1491-6.

McNally, L. M. and M. T. McNally (1996). "SR protein splicing factors interact with the Rous sarcoma virus negative regulator of splicing element." J Virol **70**(2): 1163-72.

McNally, L. M. and M. T. McNally (1998). "An RNA splicing enhancer-like sequence is a component of a splicing inhibitor element from Rous sarcoma virus." Mol Cell Biol **18**(6): 3103-11.

Milner, R. J. and J. G. Sutcliffe (1983). "Gene expression in rat brain." Nucl. Acids Res. **11**: 5497-5520.

Min, H., C. W. Turck, et al. (1997). "A new regulatory protein, KSRP, mediates exon inclusion through an intronic splicing enhancer." Gen. and Dev. **11**: 1023-1036.

Mironov, A. A., J. W. Fickett, et al. (1999). "Frequent alternative splicing of human genes." Genome Res. **9**: 1288-1293.

Modafferi, E. F. and D. L. Black (1997). "A complex intronic splicing enhancer from the c-src pre-mRNA activates inclusion of a heterologous exon." Mol Cell Biol **17**(11): 6537-45.

Modrek, B., A. Resch, et al. (2001). "Genome-wide detection of alternative splicing in expressed sequences of human genes." Nucleic Acids Res **29**(13): 2850-9.

Mount, S. M. (1982). "A catalogue of splice junction sequences." Nucleic Acids Res **10**(2): 459-72.

Mount, S. M. and H. K. Salz (2000). "Pre-messenger RNA processing factors in the Drosophila genome." J Cell Biol **150**(2): F37-44.

Nagel, R. J., A. M. Lancaster, et al. (1998). "Specific binding of an exonic splicing enhancer by the pre-mRNA splicing factor SRp55." Rna **4**(1): 11-23.

Nakai, K. and H. Sakamoto (1994). "Construction of a novel database containing aberrant splicing mutations of mammalian genes." Gene **141**: 171-177.

Nayler, O., C. Cap, et al. (1998). "Human Transformer-2-beta Gene (SFRS10): Complete nucleotide sequence, chromosomal localization, and generation of a tissue-specific isoform." Genomics **53**: 191-202.

Neuwald, A. F., J. S. Liu, et al. (1995). "Gibbs motif sampling: Detection of bacterial outer membrane protein repeats." Prot. Sci. **4**: 1618-1632.

Pagani, F., E. Buratti, et al. (2000). "Splicing factors induce cystic fibrosis transmembrane regulator exon 9 skipping through a nonevolutionary conserved intronic element." J Biol Chem **275**(28): 21041-7.

Polydorides, A. D., H. J. Okano, et al. (2000). "A brain-enriched polypyrimidine tract-binding protein antagonizes the ability of Nova to regulate neuron-specific alternative splicing." Proc. Natl. Acad. Sci. USA **97**: 6350-6355.

Reddy, R. (1989). "Compilation of small nuclear RNA sequences." Meth. Enzymol. **180**: 521-532.

Screaton, G. R., J. F. Caceres, et al. (1995). "Identification and characterization of three members of the human SR family of pre-mRNA splicing factors." Embo J **14**(17): 4336-49.

Shapiro, M. B. and P. Senapathy (1987). "RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression." Nucleic Acids Res **15**(17): 7155-74.

Shiga, N., Y. Takeshima, et al. (1997). "Disruption of the splicing enhancer sequence within exon 27 of the dystrophin gene by a nonsense mutation induces partial skipping of the exon and is responsible for Becker muscular dystrophy." J Clin Invest **100**(9): 2204-10.

Si, Z. H., D. Rauch, et al. (1998). "The exon splicing silencer in human immunodeficiency virus type 1 Tat exon 3 is bipartite and acts early in spliceosome assembly." Mol Cell Biol **18**(9): 5404-13.

Sirand-Pugnet, P., P. Durosay, et al. (1995). "An intronic (A/U)GGG repeat enhances the splicing of an alternative intron of the chicken beta-tropomyosin pre-mRNA." Nucl.Acids Res. **23**: 3501-3507.

Smith, C. W. J. and J. Valcárcel (2000). "Alternative pre-mRNA splicing: the logic of combinatorial control." TIBS **25**: 381-388.

Southby, J., C. Gooding, et al. (1999). "Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of alpha-actinin mutally exclusive exons." Mol Cell Biol **19**(4): 2699-711.

Staffa, A., N. H. Acheson, et al. (1997). "novel exonic elements that modulate splicing of the human fibronectin EDA exon." J. Biol. Chem. **272**: 33394-33401.

Stamm, S., D. Casper, et al. (1999). "Regulation of the neuron-specific exon of clathrin light chain B." Mol. Brain Res. **64**: 108-118.

Stamm, S., M. Q. Zhang, et al. (1994). "A sequence compilation and comparison of exons that are alternatively spliced in neurons." Nucleic Acids Res **22**(9): 1515-26.

Stamm, S., J. Zhu, et al. (2000). "An alternative-exon database and its statistical analysis." DNA Cell Biol **19**(12): 739-56.

Steitz, J. A. (1992). "Splicing takes a holliday." Science **257**: 888-889.

Stoss, O., M. Olbrich, et al. (2001). "The STAR/GSG family protein rSLM-2 regulates the selection of alternative splice sites." J Biol Chem **276**(12): 8665-73.

Tacke, R., Y. Chen, et al. (1997). "Sequence-specific RNA binding by an SR protein requires RS domain phosphorylation: creation of an SRp40-specific splicing enhancer." Proc Natl Acad Sci U S A **94**(4): 1148-53.

Tacke, R. and J. L. Manley (1999). "Determinants of SR protein specificity." Curr. Opi. Cell Biol. **11**: 358-362.

Tanaka, K., A. Watakabe, et al. (1994). "Polypurine Sequences within a Downstream Exon Function as a Splicing Enhancer." Mol. Cell. Biology **14**: 1347-1354.

Thanaraj, T. A. (1999). "A clean data set of EST-confirmed splice sites from Homo sapiens and standards for clean-up procedures." Nucleic Acids Res **27**(13): 2627-37.

Thanaraj, T. A. (2000). "Positional characterisation of false positives from computational prediction of human splice sites." Nucl. Acids Res. **28**: 744-754.

Thanaraj, T. A. and F. Clark (2001). "Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions." Nucleic Acids Res **29**(12): 2581-93.

Thanaraj, T. A. and A. Robinson (2000). "Prediction of exact boundaries of exons." Briefings Bioinform. **1**: 343-356.

Tian, H. and R. Kole (1995). "Selection of noevel exon recognition elements from a pool of random sequences." Mol. Cell. Biol. **15**: 6291-6298.

Tian, M. and T. Maniatis (1993). "A splicing enhancer complex controls alternative splicing of doublesex pre-mRNA." Cell **74**(1): 105-14.

Tsukahara, T., C. Casciato, et al. (1994). "Alternative splicing of beta-tropomyosin pre-mRNA: multiple cis-elements can contribute to the use of the 5'- and 3'-splice sites of the nonmuscle/smooth muscle exon 6." Nucleic Acids Res **22**(12): 2318-25.

Valentine, C. R. (1998). "The association of nonsense codons with exon skipping." Mutat Res **411**(2): 87-117.

Weighardt, F., G. Biamonti, et al. (1996). "The role of heterogeneous nuclear ribonucleoproteins (hnRNP) in RNA metabolism." BioEssays **18**: 747-756.

Wentz, M. P., B. E. Moore, et al. (1997). "A naturally arising mutation of a potential silencer of exon splicing in human immunodeficiency virus type 1 induces dominant aberrant splicing and arrests virus production." J Virol **71**(11): 8542-51.

Wise, J. A. (1993). "Guides to the heart of the spliceosome." Science **262**: 1978-1979.

Wu, S., C. M. Romfo, et al. (1999). "Functional recognition of the 3' splice site AG by the splicing factor U2AF." Nature **402**: 832-835.

Zahler, A. M., K. M. Neugebauer, et al. (1993). "Distinct functions of SR proteins in alternative pre-mRNA splicing." Science **260**(5105): 219-22.

Zandberg, H., T. C. Moen, et al. (1995). "Cooperation of 5' and 3' processing sites as well as intron and exon sequences in calcitonin exon recognition." Nucleic Acids Res **23**(2): 248-55.

Zhang, L., W. Liu, et al. (1999). "Coordinate repression of a trio of neuron-specific splicing events by the splicing regulator PTB." Rna **5**(1): 117-30.

Zhang, M. Q. and T. G. Marr (1993). "A weight array method for splicing signal analysis." Comput Appl Biosci **9**(5): 499-509.

Zhang, M. Q. and T. G. Marr (1994). "Fission yeast gene structure and recognition." Nucleic Acids Res **22**(9): 1750-9.

Zheng, Z. M., P. J. He, et al. (1999). "Function of a bovine papillomavirus type 1 exonic splicing suppressor requires a suboptimal upstream 3' splice site." J Virol **73**(1): 29-36.

Zhuang, Y. and A. M. Weiner (1986). "A compensatory base change in U1 snRNA suppresses a 5' splice site mutation." Cell: 827-835.