How prevalent is functional alternative splicing in the human genome?☆

Rotem Sorek[1,2], Ron Shamir[3] and Gil Ast[1]

[1]Department of Human Genetics, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel
[2]Compugen, 72 Pinchas Rosen Street, Tel Aviv 69512, Israel
[3]School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

Comparative analyses of ESTs and cDNAs with genomic DNA predict a high frequency of alternative splicing in human genes. However, there is an ongoing debate as to how many of these predicted splice variants are functional and how many are the result of aberrant splicing (or 'noise'). To address this question, we compared alternatively spliced cassette exons that are conserved between human and mouse with EST-predicted cassette exons that are not conserved in the mouse genome. Presumably, conserved exon-skipping events represent functional alternative splicing. We show that conserved (functional) cassette exons possess unique characteristics in size, repeat content and in their influence on the protein. By contrast, most non-conserved cassette exons do not share these characteristics. We conclude that a significant portion of cassette exons evident in EST databases is not functional, and might result from aberrant rather than regulated splicing.

Numerous studies have shown that alternative splicing is prevalent in mammalian genomes. Using ESTs and cDNAs aligned to the genomic sequence, these studies estimate that between 35% and 59% of all human genes undergo alternative splicing [1,2]. However, it is not clear how many of the splice variants predicted from ESTs are functional and how many represent aberrant splicing ('noise') or EST artefacts (such as genomic contamination) [3–5]. An mRNA variant can be defined as being 'functional' if it is required during the life-cycle of the organism and activated in a regulated manner.

Aberrant alternative splicing

Somatic mutations within splice sites or introns could result in aberrant splicing, leading to non-functional mRNAs; ESTs derived from these mRNAs would be indistinguishable from normal splice variants. Because somatic mutations are prevalent in cancer related tissues, and >50% of the ESTs in dbEST come from cancer, cell-lines or tumor tissues [6], such spurious variants can be ubiquitous in dbEST.

Splicosomal mistakes have also been proposed as a mechanism that can result in non-functional transcripts [3]. In common with any complex biological machine, the splicosome can 'slip' and identify cryptic splice sites in introns as normal splice sites, thus, inserting part of an intron into the mature mRNA. Obviously such mistakes would not represent functional, regulated alternative splicing.

What portion of the observed splice variants represents functional alternative splicing? We employed a comparative genomics approach to address this question, by compiling a dataset of exon-skipping events (cassette exons) that are conserved between human and mouse. The conservation of such events in both human and mouse species, which diverged from their common ancestor 75–110 million years ago, suggests the functional importance of these exons.

Detecting cassette exons conserved between human and mouse

We have recently collected a dataset of 980 EST-predicted human alternatively spliced cassette exons [7]. From these 980 exons, 243 (25%) were also found to be alternatively spliced in mouse ['conserved alternatively spliced exons' (CAS exons)]. The remaining 737 (75%) are 'non-conserved alternatively spliced exons' (non-CAS exons) (Box 1). Low levels of alternative splicing conservation between human and mouse were also observed in other studies [8,9]. The method that was used to locate cassette exons and human–mouse conservation is described in detail in Ref. [7]. Calculated features of the 980 exons used for this study appear as supplementary material online.

Box 1. Finding exon-skipping events that are conserved between humans and mouse

The initial set of exons comprised 980 apparently alternatively spliced human exons, for which a mouse EST spanning the intron that represents the exon-skipping variant was found. Two strategies were used to identify an exon as being conserved in mouse: (i) identification of mouse ESTs that contain the exon and the two flanking exons; and (ii) if the exon was not represented in mouse ESTs, the sequence of the human exon was searched against the intron spanned by the skipping mouse EST on the mouse genome. If a significant conservation (>80%) was found, the alignment spanned the full length of the human exon, and the exon was flanked by the canonical AG acceptor and GT donor sites in the mouse genome, then the exon was declared as conserved. For 243 exons (25% of 980), conserved alternative splicing was detected in mouse. Detailed description of the methods can be found in Ref. [7].

☆ Supplementary data associated with this article can be found at doi: 10.1016/j.tig.2003.12.004

Corresponding author: Gil Ast (gilast@post.tau.ac.il).

Comparing conserved with non-conserved cassette exons

Presumably, orthologous exons that are alternatively spliced both in human and mouse have functional importance. We can therefore regard the group of CAS exons as a representative group of functional, alternatively spliced exons.

Non-CAS exons, on the other hand, can also be functional, representing exons created after the divergence of the human and the mouse lineages. However, if these exons, as a group, were indeed functional, we might expect them to have the same general characteristics as CAS exons. Therefore, we compared several features between the two groups of exons predicted by ESTs.

We first sought to understand the influence of the alternatively spliced exons on the proteins in which they are inserted. From the 243 CAS exons identified, 195 had an unambiguous coding-region annotation in the GenBank cDNAs. Of these, 191 (98%) were located within the protein-coding region and four were located within the untranslated region (UTR). This finding is not surprising because UTRs are mostly found in the terminal exons [10], whereas the exons in our study were internal exons. A similar percentage of the non-CAS exons were located in the coding sequence and UTR.

The influence of the CAS exons on the protein coding sequence (CDS) was significantly different from the influence of the non-CAS exons (Figure 1). In 147 of 191 (77%) CAS exons that were located within the protein-coding region, the insertion of the alternatively spliced exon did not alter the reading frame and only eight of these 147 exons (5%) contained an in-frame stop codon. This means that 73% of the conserved alternative exons are 'peptide cassettes' (i.e. they insert a short amino-acid sequence into the translated protein without changing the coding frame) (Figure 1). These results indicate a strong tendency of functional cassette exons to add or remove amino acids within the protein sequence, rather than inflict a dramatic change on it.

By comparison, only 206 of the 510 (40%) non-CAS exons preserved the reading frame; when an alternatively spliced exon was inserted in frame, almost half (47%) of these contained a stop codon. Thus, only 109 of the 510 (21%) non-CAS exons were indeed 'peptide cassette' exons (Figure 1).

We examined in detail the cases in which the alternatively spliced exons caused a frame shift in the CDS. Of the 44 CAS exons that caused a frame shift, 27 (61%) actually made the protein longer by suppressing a nearby stop codon and only four exon insertions (9%) resulted in a protein shorter than 100 amino acids. This indicates that, in a substantial fraction of the conserved alternatively spliced exons that cause a frame shift, the exon insertion changes only the C-terminus of the protein.

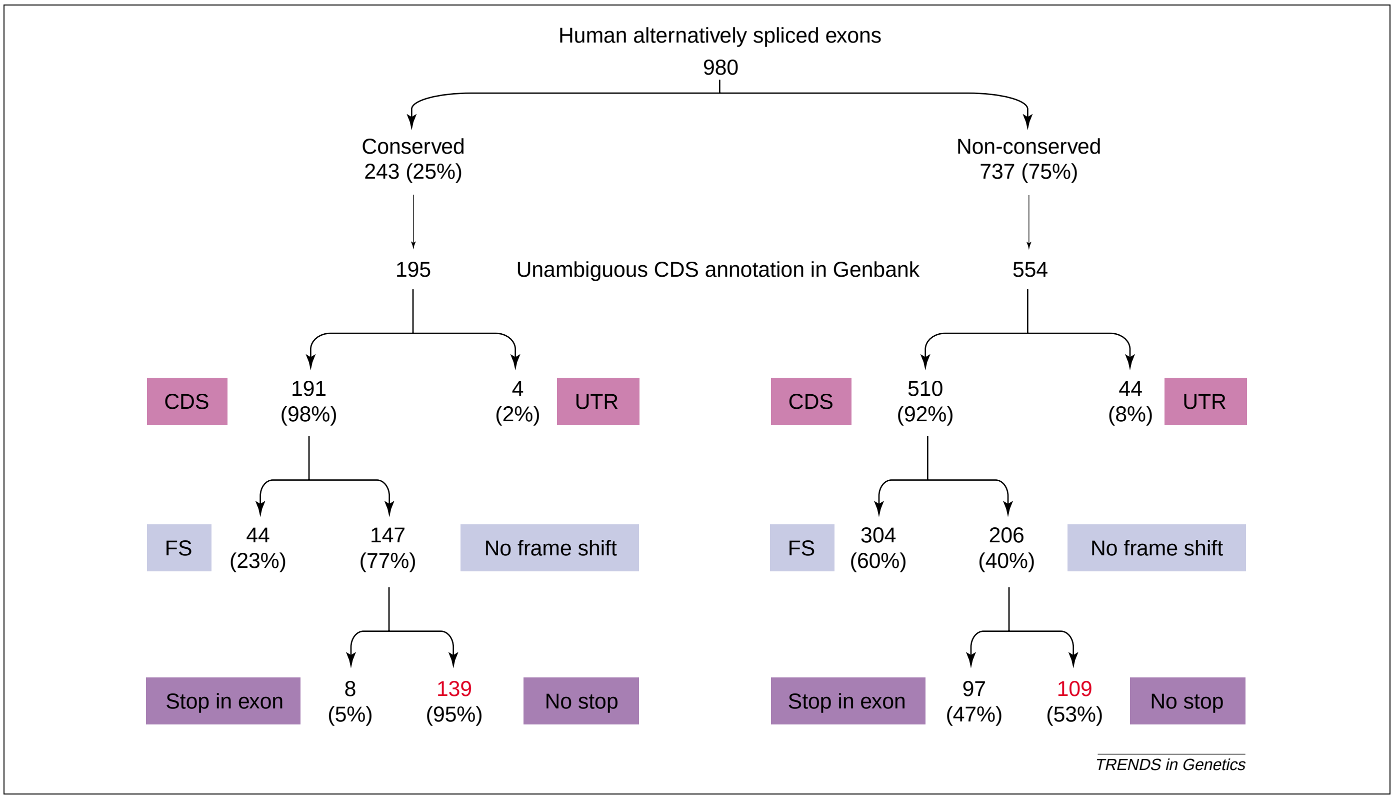

Figure 1. The influence of alternatively spliced exons on the protein-coding sequence. Our dataset of human alternatively spliced exons contained 980 exons. These were divided into two subsets: (i) exons whose existence was also confirmed by mouse ESTs or mouse genomic sequence (243 exons, termed 'conserved alternatively spliced exons'); and (ii) exons for which no such evidence was detected (737 exons, termed 'non-conserved alternatively spliced exons'). The influence of these exons on the protein-coding sequence is indicated by the tree-like diagram. Those exons with 'unambiguous CDS annotation' [i.e. all annotated GenBank cDNAs in which the exon appeared had the same annotation (either UTR or CDS)] were analyzed further. Peptide cassettes were found in 73% (139 of 191) and 21% (109 of 510) of the conserved alternatively spliced and non-conserved alternatively spliced exons, respectively (shown in red). Definitions: frame shift, the length of the exon is not a multiple of three; stop in exon, the exon contains an in-frame stop codon; peptide cassette, contains exons that neither cause a frame shift nor contain a stop codon. Abbreviations: CDS, coding sequence; FS, frame shift; UTR, untranslated region.

By contrast, of the 304 non-CAS exons that caused a frame shift, only 25 (8%) resulted in a longer protein, whereas 91 (30%) resulted in a protein shorter than 100 amino acids. This implies that insertion of non-CAS exons into the mature mRNA frequently has a major effect on the protein sequence.

Frequencies of ESTs, repeats and size as measures of function

It was recently suggested that a higher number of ESTs/mRNAs supporting a splice variant correlates with its functionality [11]. Our results agree with this observation: although CAS exons were on average supported by nine sequences (median 3), the average EST and mRNA support for non-CAS exons was 2.2 sequences (median 1). However, sequence support by itself is not sufficient for detecting functional alternative splicing: in our study, 30% of the CAS exons were supported by a single human expressed sequence.

We have previously shown that point mutations within silent intronic *Alu* elements can result in the creation of new alternatively spliced exons [12]. Such *Alu*-derived exons represent at least 5% of the alternatively spliced cassette exons found in dbEST [13]; however, it is unclear whether these *Alu* exons are functional [14]. Other repetitive elements, such as the RTE-1 retrotransposon in cattle, have also shown the ability to be exonized (i.e. become exons via a splicing-mediated process) [15].

To check how many alternatively spliced exons are the result of such exonization, we performed a BLAST search of the exons against a database of mammalian repeats. Only one of the 243 CAS exons had a significant ‘hit’ (e-value $<10^{-10}$) to a mammalian interspersed repeat (MIR). By contrast, 191 of the 737 non-CAS exons (26%) had significant ‘hits’ to a repeat. In 147 (77%) of these cases, the repeat was an *Alu* retrotransposon, which is unique to primates. The repeat content within exons, therefore, is another feature in which CAS exons differ from non-CAS exons.

Conserved and non-conserved cassette exons also differ in their exon length distribution. The average length of a CAS exon was 87 bases (median 76). By contrast, the average length of non-CAS exons was 116 (median 104). The difference between these two distributions is statistically significant ($P < 10^{-6}$). (For comparison, the average length of constitutively spliced exons is 129 bases [16].) Thus, non-CAS exons are significantly longer than CAS exons.

Which exons are functional?

We detected a conserved mouse exon for 25% of the human cassette exons in our set. These CAS exons most probably have functional importance. In principle, some of the 75% human non-CAS exons could also be functional; however, these might be expected to have similar characteristics to those of the CAS exons. This is not the case. We have shown that the group of non-CAS exons significantly differs from the group of CAS exons in many important parameters (discussed previously) (these differences are summarized in Table 1). This suggests that many of the apparently non-conserved splice variants in the human genome are non-functional. It is noteworthy that this claim is based on the assumption that functional non-CAS exons ought to have properties similar to those of CAS exons – this has yet to be verified.

To further test this assumption, we examined non-CAS exons that are supported by multiple ESTs. Of the non-CAS exons, 21% have no frame shift and no stop codon. However, in the subset of non-CAS exons that are supported by five sequences or more and are found in the CDS (37 exons), 54% have no frame shift and no stop codon. Therefore, some of the non-CAS exons indeed represent new functional exons that are specific to the human lineage. In addition, this supports our claim that functional non-CAS exons have properties similar to those of CAS exons.

Non-CAS exons might have an important evolutionary role even if they are not functional. These rarely expressed splice isoforms ‘suggest’ new variants while keeping the genomic repertoire intact. Such new variants can have a positive effect on the organism and become fixed during evolution. Indeed there are several examples, such as the *Alu*-derived exon in the ADARB1 gene [17], in which repetitive elements were found to be recruited from an intron during evolution and fixed as new alternatively

Table 1. Features differentiating between conserved alternatively spliced exons and non-conserved alternatively spliced exons

Features	Conserved alternatively spliced exons	Non-conserved alternatively spliced exons[a]	*P* value[b]
Average size	87	116	$P < 10^{-6}$
Percentage of exons that are a multiple of three	77% (147/191)	40% (206/510)	$P < 10^{-5}$
Percentage of exons that are ‘peptide cassettes’ [c]	73% (139/191)	21% (109/510)	$P < 10^{-15}$
Percentage of exon insertions that result in a longer protein (from a total of exons that are not a multiple of three)	61% (27/44)	8% (25/304)	$P < 10^{-9}$
Percentage of exon insertions that result in a protein <100 amino acids (from a total of exons that are not a multiple of three)	9% (4/44)	30% (91/304)	$P < 0.02$
Average supporting expressed sequences	9	2.2	$P < 10^{-6}$
Percentage of exons that contain repetitive elements[d]	<0.5% (1/243)	26% (191/737)	$P < 10^{-20}$

[a] Non-conserved alternatively spliced exons are human exons that were found to be alternatively spliced in human ESTs but were not found in the mouse genome.

[b] The *P* value was calculated using Fisher’s exact test, except for the ‘average size’ and ‘average support’, for which *P* values were calculated using student’s *t* test.

[c] A ‘peptide cassette’ exon is defined such as it neither causes a frame shift nor contains a stop codon, so that the effect of its insertion or deletion on the translated protein is a local insertion or deletion of a peptide.

[d] Exons were aligned to a database of repetitive elements, and ‘hits’ with e-value $<10^{-10}$ were considered positive.

Box 2. Methods

Detailed methods for compilation of exon sets are found in Refs [7] and [13]. Fisher's exact statistical test was used for calculating *P* values of the parameters differentiating between the two exons sets, except for the 'average size' and 'average support', for which *P* value was calculated using student's *t* test. Human–mouse exon skipping data from ASAP were downloaded from http://www.bioinformatics.ucla.edu/ASAP/data/Comparative_Genomics/Hs_Mm_exon_skip_status_table. Lengths of these exons were extracted from: http://www.bioinformatics.ucla.edu/ASAP/data/Alt_Splice/Human/January_2002/ exon_obs_table.

spliced exons. Such a recruitment of intronic sequences as new alternatively spliced exons was previously proposed to be major evolutionary driving force towards the emergence of new protein sequences in eukaryotes [8,9,13,15,18].

Comparison with other results

We examined the results of Modrek and Lee [9], who compared exon-skipping events between human and mouse using ASAP, an EST-based alternative splicing database. 127 ASAP exons were identified in which both variants (exon inclusion and exon skipping) were observed in human and in mouse (equivalent to our CAS exons); in 78 (61%) the length of the exon was a multiple of three (i.e. did not cause a frame shift). By contrast, of the 427 ASAP human exons that were predicted to be skipped in human but were not conserved in mouse (non-CAS exons), only 164 (38%) had an exon length that was a multiple of three. These numbers are similar to the numbers calculated from our set of exons, showing that our results are not database dependent (Box 2).

Our results suggest that 73% of functional cassette exons neither change the coding frame nor introduce a premature stop codon. However, Zavolan *et al.* reported that only 178 of 423 (42%) of the alternative splicing detected in mouse full-length mRNAs preserved the reading frame [19]. Another study that detected alternative splicing using ESTs and cDNAs also reported that only 40% of the deletions or insertions of a new sequence in the middle of the protein were in frame [20]. This contradiction probably stems from the fact that these two studies used alternative splicing predicted from alignments of expressed sequences (ESTs and cDNAs), which contained both conserved and non-conserved splice variants. Indeed, combining the numbers from both our sets (CAS and non-CAS exons) results in 248 out of 701 (35%) of the exons preserving the reading frame, similar to the results in these studies. [19,20] In a study of a sample of 1000 alternatively spliced exons compiled from the literature, ~78% were identified as 'peptide cassette' exons (neither contained a stop codon nor introduced a frame-shift) [21] – similar to the 73% we detected in CAS exons. These data strongly supports our results because experimentally confirmed splice variants are more likely to be functional.

We have shown that a comparative genomics approach can be useful for assessing the functionality of splice variants. In the future, the parameters defining functional cassette exons could be used for *de novo* identification of alternative splicing in organisms for which no EST data exist.

Acknowledgements

We thank Dan Graur for inspiration; Han Xie, Amit Novik, Dvir Dahary and Ami Haviv for insightful discussions; Zurit Levine for technical help; and Eli Eisenberg for critical reading of the manuscript.

References

1 Mironov, A.A. *et al.* (1999) Frequent alternative splicing of human genes. *Genome Res.* 9, 1288–1293

2 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

3 Graveley, B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.* 17, 100–107

4 Modrek, B. and Lee, C. (2002) A genomic view of alternative splicing. *Nat. Genet.* 30, 13–19

5 Maniatis, T. and Tasic, B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature* 418, 236–243

6 Baranova, A.V. *et al.* (2001) *In silico* screening for tumour-specific expressed sequences in human genome. *FEBS Lett.* 508, 143–148

7 Sorek, R. and Ast, G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.* 13, 1631–1637

8 Nurtdinov, R.N. *et al.* (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.* 12, 1313–1320

9 Modrek, B. and Lee, C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* 34, 177–180

10 Deutsch, M. and Long, M. (1999) Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.* 27, 3219–3228

11 Kan, Z. *et al.* (2002) Selecting for functional alternative splices in ESTs. *Genome Res.* 12, 1837–1845

12 Lev-Maor, G. *et al.* (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science* 300, 1288–1291

13 Sorek, R. *et al.* (2002) Alu-containing exons are alternatively spliced. *Genome Res.* 12, 1060–1067

14 Pavlicek, A. *et al.* (2002) Transposable elements encoding functional proteins: pitfalls in unprocessed genomic data? *FEBS Lett.* 523, 252–253

15 Makalowski, W. (2003) Genomics. Not junk after all. *Science* 300, 1246–1247

16 Zavolan, M. *et al.* (2003) Impact of alternative initiation, splicing, and termination on the diversity of the mRNA transcripts encoded by the mouse transcriptome. *Genome Res.* 13, 1290–1300

17 Lai, F. *et al.* (1997) Editing of glutamate receptor B subunit ion channel RNAs by four alternatively spliced DRADA2 double-stranded RNA adenosine deaminases. *Mol. Cell. Biol.* 17, 2413–2424

18 Kondrashov, F.A. and Koonin, E.V. (2003) Evolution of alternative splicing: deletions, insertions and origin of functional parts of proteins from intron sequences. *Trends Genet.* 19, 115–119

19 Zavolan, M. *et al.* (2002) Splice variation in mouse full-length cDNAs identified by mapping to the mouse genome. *Genome Res.* 12, 1377–1385

20 Modrek, B. *et al.* (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29, 2850–2859

21 Thanaraj, T.A. and Stamm, S. (2003) Prediction and statistical analysis of alternatively spliced exons. *Prog. Mol. Subcell. Biol.* 31, 1–31

0168-9525/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2003.12.004