Observed Extent of Alternative Splicing in a Limited Data Set of Human Genes of EST−Confirmed Splice Sites.

Classification of the data from our data set on EST−confirmed splice sites from human. [See http://www.ebi.ac.uk/~thanaraj/splice.html and "A clean data set of EST−confirmed splice sites from Homo sapiens and..." T.A. Thanaraj (1999) Nucleic Acids Research 27(13), 2627−2637.]

- Of 601 EST−confirmed pairs of donor and acceptor sites (*see http://www.ebi.ac.uk/~thanaraj/splice.out for the data set*), **143 pairs (24%)** showed occurrences of EST−confirmed alternative events (of the types illustrated in the figures below; intron retention was NOT included) at either the 5' and/or 3' splice site.

- Of 40 EST−confirmed pairs of donor and acceptor sites from **UTR, 18 pairs (45%)** showed occurrences of EST−confirmed alternative events at either the 5' and/or 3' splice site. The share of detected alternative splicing that occurs in UTR's is 12.6% (Of the 143 EST−confirmed splice pairs that were affected by alternative splicing, 18 occurrend in UTR's).

- These set of 601 EST−confirmed pairs of donor and acceptor sites were derived from 171 unique gene entries. The 143 pairs of 5' and 3' splice sites, which displayed alternative splicing, were found in 83 of the 171 genes. **Thus alternative splicing could be observed in at least 49% of the cases in human genes.**
A similar estimate is subsequently reported in a recent article by Kan et al (2001) Genome Research 11 (5) 889−900.

The value observed by us can be an underestimate for the following reasons:

(1) Analysed in this study were only those genes that have NOT been annotated in the databases as exhibiting alternative splicing.

(2) In addition to the above number of cases of alternative splicing, intron retention (either the complete intron or a fragment of length > 50 nucleotides downstream of donor or upstream of acceptor site) was observed in 66 cases of splice sites. These cases were not included in our calculations. Some of them might correspond to genuine transcripts and not artifacts.

(3) Only a region of 50 nucleotides on either side of donor or acceptor site was scrutinised for the occurrencs of (a) exon extension/truncation and (b) the

start/end of crypic exons.

(4) Ambiguous and doubtful cases were excluded. Only those splice sites for which both the annotated constituent donor and acceptor sites had proof from a single EST sequence were used in the study.

(5) The splice sites that were analysed in our work were predominantly from the coding regions. Most the of the regulation via alternative splicing occurs in the 5' region of the genes.

Estimate by others

Croft et al., (2000, Nature Genetics, 24[4], 340−341) has observed an estimate of 22%. Gelfand and coworkers advocate a theory that one in every three human genes might give rise to alternatively spliced transcripts (Mironov et al, 1999, Genome Research, 9, 1288−1293). Similar estimates of 30−35% have been reported by others Hanke et al (1999) Trends Genet 15(10) : 389−90. A subsequent report by Kan et al (2001 Genome Research 11 (5) 889−900) presents an estimate as high as our above estimate.

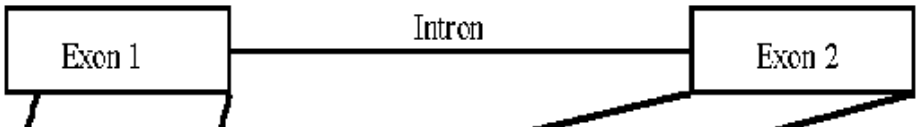
Types of alternative splicing observed

The types of the observed alternative splicings with the EST−confirmed pairs of donors and acceptors are as shown in the following figures.

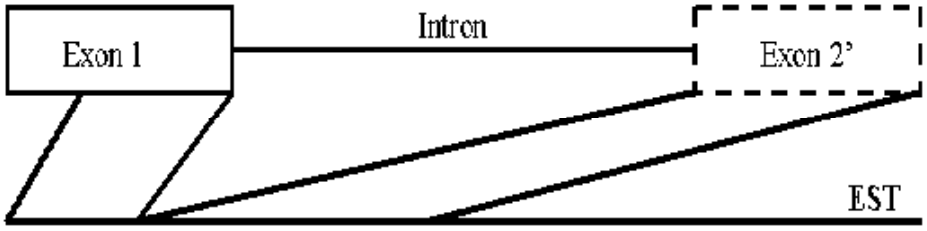
- 1. 5' exon skipping = 23 cases of pairs of donors and acceptors;
- 2. 3' exon skipping = 20 cases;
- 3. exon extension/truncation at 5' end = 29 cases;
- 4. exon extension/truncation at 3' end = 50 cases;
- 5. Cryptic exons = 13 cases;
- 6. Multiple events = 8 cases;

Number of genes (out of the 171 genes forming the data set) showing the different types of alternative splicing.

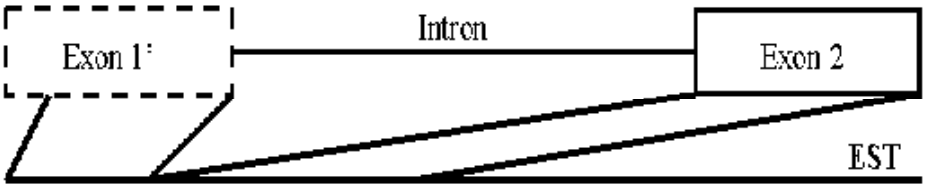
- 1. Exon skipping = 35 genes ===> (20% of 171 genes)
- 2. Exon extension/truncation = 56 genes ===> (33% of 171 genes);
- 3. Cryptic exons = 12 genes ===> 7%.



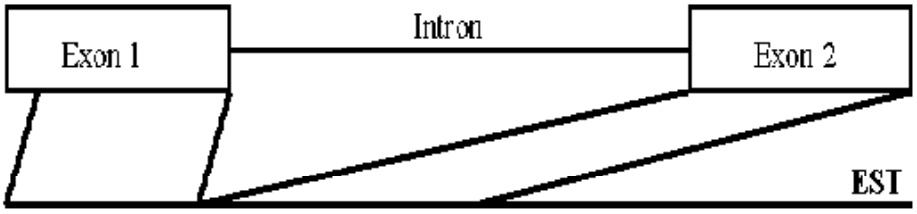
splice site.
601 entries.



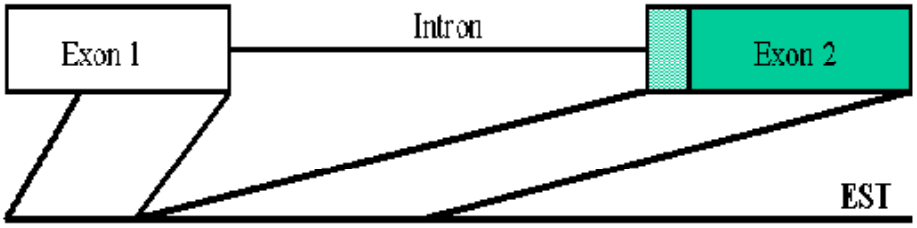
3' Exon Skipping:
Exon 2' was different from Exon 2.
Observed in 20 entries.



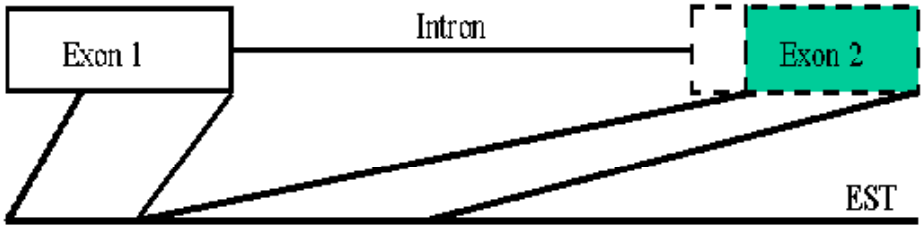
5' Exon Skipping:
Exon 1' was different from Exon 1.
Observed in 23 entries.



As per annotation.
EST-confirmed splice site.
601 entries

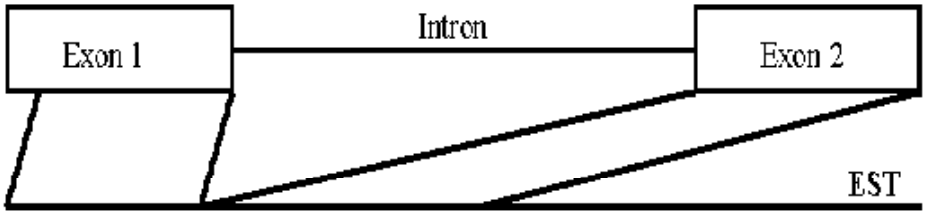


3' Exon Extension:
Exon 2 was in an extended form.

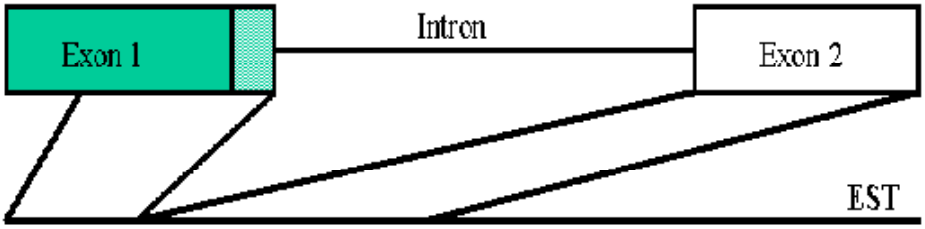


3' Exon Truncation:
Exon 2 was in a truncated form.

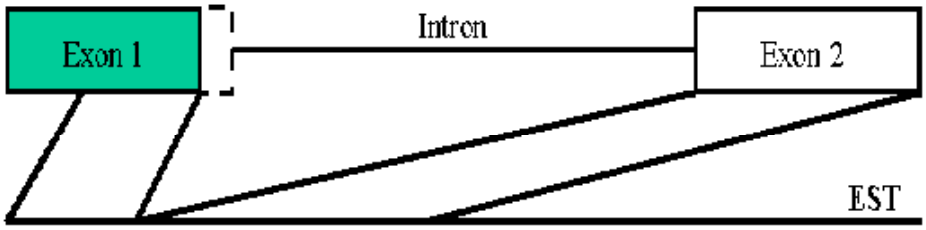
The above two observed in 50 cases.



As per annotation. EST-confirmed splice site. 601 entries.

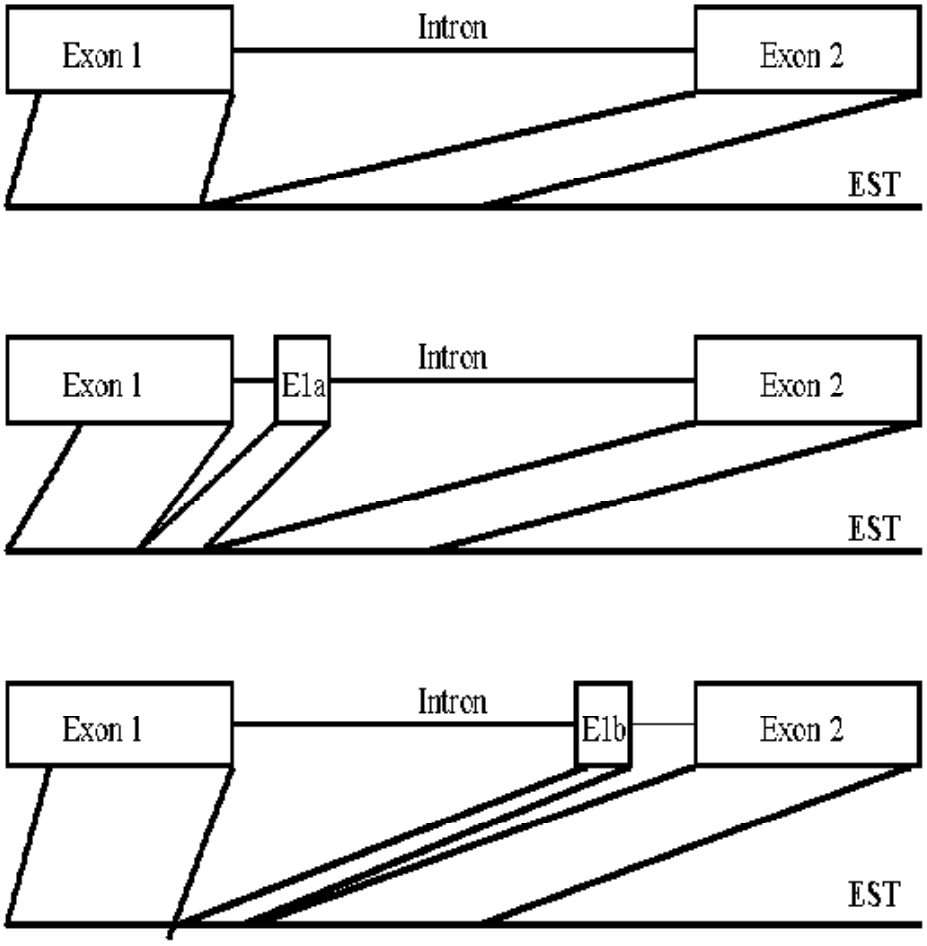


5' Exon Extension:
Exon 1 was in an extended form.



5' Exon Truncation:
Exon 1 was in a truncated form.

The above two observed in 29 entries.



In 8 cases, multiple events were observed.

Note on the derivation of the genes used in this study (See my NAR paper cited at the top of this page for details). The salient points are as below::
(1) Initially selected a set of 4300 entries (not results of HTG) from EMBL data base were selected.
(2) Stringent cleaning procedures were applied.
(3) Redundant sequences were removed.
(4) These steps reduced the 4300 genes to 310 genes.
(4) Entries having unusual splice sites (as determined by decision trees) were

removed.
(5) This yielded 219 entries.
(6) Of these 219, only for a set of 171 genes EST could confirm at least one pair of donor and acceptor site.
(7) These 171 genes were used in the study.

Note on the derivation of the value for the extent of alternative splicing in human genes

I started with a data set of 219 genes out of which only 171 genes had at least one of its introns (the pair of donor and its partner acceptor site) confirmed by EST. Such EST−confirmed introns were 601. Of these introns, I checked how many exhibited alternative splice through other EST sequences. They were 143 in number. Now I went back and checked in how many genes these 143 occurred. They were 83. So the fraction of the number of genes that showed up Alternative Splicing is 83/171 == 49%. So what I have done is : I have sort of normalised the normal form of gene as well as the isoform gene for EST coverage.

I have taken care of, in the above work, two important bottlenecks in the use of ESTs to work out the estimates of alternative splicing. The two bottle necks are (i) the coverage of genes by EST is only around 35−40%; (ii) there is a bias in that the 3' regions on the gene tend to be represented more by EST than the 5' regions.

The analysis presented in this web page is being written up as a short note for a journal.