Published online September 28, 2005

5574–5582 *Nucleic Acids Research, 2005, Vol. 33, No. 17*
doi:10.1093/nar/gki858

The importance of being divisible by three in alternative splicing

Alon Magen and Gil Ast*

Department of Human Genetics and Molecular Medicine, Sackler Faculty of Medicine, Tel Aviv University, Ramat Aviv 69978, Israel

Received May 22, 2005; Revised August 10, 2005; Accepted September 7, 2005

ABSTRACT

Alternative splicing events that are conserved in orthologous genes in different species are commonly viewed as reliable evidence of authentic, functionally significant alternative splicing events. Several recent bioinformatic analyses have shown that conserved alternative exons possess several features that distinguish them from alternative exons that are species-specific. One of the most striking differences between conserved and species-specific alternative exons is the high percentage of exons that preserve the reading frame (exons whose length is an exact multiple of 3, termed symmetrical exons) among the conserved alternative exons. Here, we examined conserved alternative exons and found several features that differentiate between symmetrical and non-symmetrical alternative exons. We show that symmetrical alternative exons have a strong tendency not to disrupt protein domain structures, whereas the tendency of non-symmetrical alternative exons to overlap with different fractions of protein domains is similar to that of constitutive exons. Additionally, skipping isoforms of non-symmetrical alternative exons are strongly underrepresented, compared with their including isoforms, suggesting that skipping of a large fraction of non-symmetrical alternative exons produces transcripts that are degraded by the nonsense-mediated mRNA decay mechanism. Non-symmetrical alternative exons also show a tendency to reside in the 5′ half of the CDS. These findings suggest that alternative splicing of symmetrical and non-symmetrical exons is governed by different selective pressures and serves different purposes.

INTRODUCTION

Alternative splicing creates multiple mRNA products from a single gene. This mechanism is widespread in higher eukaryotes: bioinformatic analyses indicate that more than half of human genes undergo alternative splicing (1,2). As the estimated number of human proteins (~90 000) far exceeds the number of known protein-coding genes (~26 000), alternative splicing was suggested as the key mechanism for bridging this numerical disparity (3–5). The ability of alternative splicing to greatly increase the protein-coding capability of a single gene is well demonstrated in the Drosophila's Dscam gene, which has the potential capability to generate up to 38 016 distinct protein isoforms by means of combinatorial alternative splicing of multiple 'exon cassettes' (6).

In many cases, alternative splicing creates protein isoforms, either with partially or completely different functionality from a single gene, by selectively inserting or removing protein domains encoded by alternative 'cassette' exons (i.e. alternative exons of the exon-skipping type: they are either completely included or completely removed from the mRNA product). Some examples of domain architecture alternations caused by alternative splicing are the removal of signal peptides and transmembrane regions, which affects protein localization (7); the insertion of different structural elements to ion-channel genes in different nervous system cells (8,9) and the removal of ethylene receptors in peaches, which affects the fruit's ripening time (10).

In addition to its role in increasing proteome diversity, alternative splicing is also known to produce splice forms that are not being translated into proteins, but, rather, play a regulatory role. The best-described example for such 'unproductive' splicing is the generation of splice variants targeted for degradation by the nonsense-mediated mRNA decay mechanism (NMD) (1,11–13). NMD is an RNA surveillance function that recognizes mRNAs containing premature termination codons (PTCs) and targets them for degradation, rather than translation into proteins [reviewed in (14)]. It was even suggested that as much as 35% of human alternative isoforms

*To whom correspondence should be addressed. Tel: +972 3 640 6893; Fax +972 3 640 9900; Email: gilast@post.tau.ac.il

contain PTCs (termed PTC$^+$ mRNAs) (15). This widespread coupling of NMD and alternative splicing suggests that NMD-targeted transcripts play an important regulatory role, probably in the downregulation of gene expression (1,11,15).

However, for the vast majority of alternative splicing events predicted from bioinformatic analyses, which are predominately based on expressed sequence tag (EST) data, the functional significance is unknown (4,16,17). Moreover, as a large fraction of ESTs in dbEST comes from cancer-related tissues, many of these alternative splicing events might represent aberrant splicing or EST artifacts (18–20). In this study, we examined a dataset of alternatively spliced cassette exons that are conserved between human and mouse. The conservation of these alternative exons in both species, which diverged from their common ancestor 75–130 million years ago, suggests that these exon-skipping events represent truly functional splicing events (19,21).

It has been previously reported that alternative cassette exons that are conserved in several species possess some unique features that distinguish them from species-specific exons (i.e. alternative exons that exist in a human gene, but not in its mouse counterpart or *vice versa*). Compared with species-specific exons, conserved exons tend (i) to be shorter (19); (ii) to have a high inclusion level (i.e. the number of ESTs/mRNAs that represent their inclusion is higher than the number of those that skip them) (22); (iii) not to contain repetitive elements (19) and (iv) to be in a length that is divisible by 3 [termed 'symmetrical exons', see also (23)], if they reside within the protein coding region (CDS) (19,24–26). The high percentage of symmetrical exons among conserved alternative exons within the CDS (53–77%, with respect to an expected percentage of 33% without any bias for divisibility by 3) reflects a functional pressure for preserving protein coding, as insertion or skipping of these exons does not alter the reading frame (21,25), whereas alternative splicing of non-symmetrical exons always induces a frameshift (Figure 1). Constitutive exons were also found to have a bias toward being symmetrical, but this bias is much more moderate (~40% are symmetrical) (23,25,27).

In this study, we examined the differences between symmetrical and non-symmetrical conserved alternative exons with respect to several characteristics. We found that they differ (i) in their tendency to overlap with coding sequences for protein domains, (ii) in their inclusion level, (iii) in the coverage of their skipping isoforms in GenBank and (iv) in their preference for specific locations along the CDS. These findings suggest that symmetrical and non-symmetrical conserved alternative exons are under different selective pressures and may play different functional roles, such as NMD activation by skipping of non-symmetrical alternative exons.

MATERIALS AND METHODS

Database compilation

We used two databases in our analysis: one of conserved alternative exons (only cassette exons) and one of conserved constitutive exons. The conserved alternative exons consisted of 823 human exons, 664 mouse exons, 86 rat exons and 22 dog exons, along with their corresponding mRNAs from RefSeq, NCBI's database of curated and non-redundant reference sequences (http://www.ncbi.nlm.nih.gov/RefSeq/) (28). The conserved constitutive exons database consisted of 39 720 human–mouse homologous constitutive exons and their matching 6124 RefSeq mRNAs, already compiled by us (http://www.tau.ac.il/~gilast/sup_mat2.htm) (29).

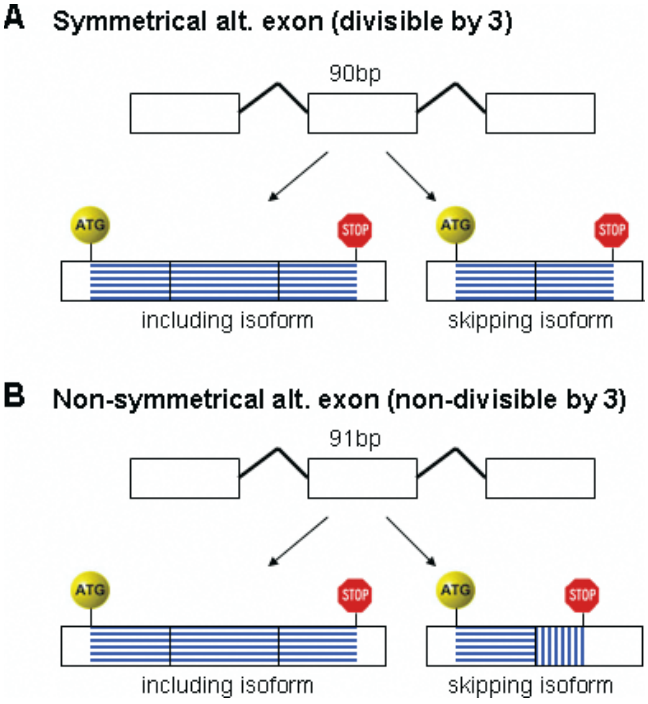

Figure 1. Different impact on the reading frame of symmetrical and non-symmetrical alternatively spliced exons. (**A**) A gene consisting of three exons whose second exon is a symmetrical cassette exon (90 nt). Skipping and inserting of the symmetrical exon does not alter the reading frame: the sequence upstream and downstream to the alternative exon encodes for the same peptide sequence in the including and skipping isoforms. (**B**) A gene consisting of three exons whose second exon is a non-symmetrical cassette exon (91 nt). Skipping of the non-symmetrical exon results in a frameshift downstream to the alternative exon.

Compilation of the alternative exons database was performed in several stages: we started with a dataset of human–mouse homologous alternative cassette exon pairs, which was built by integrating three source databases:

(i) http://www.soe.ucsc.edu/~sugnet/psb2004/altGraphXCon.html (30).
(ii) http://www.tau.ac.il/~gilast/sup_mat2.htm (29).
(iii) http://www.sciencedirect.com/science/MiamiMultiMediaURL/B6TCY-4B9558H-1/B6TCY-4B9558H-1-C/5183/f9bdffb6d98512adebc2e545a0151900/ Application_1.xls (19).

We removed all duplications and remained with 1285 pairs of human–mouse homologous alternative exons. As all three source databases were compiled using EST data, we blasted all exons against RefSeq, NCBI's database of curated and non-redundant reference sequences (http://www.ncbi.nlm.nih.gov/RefSeq/) (28), to find their matching full-length mRNAs for further analysis. We found such RefSeq mRNAs for 823 human exons and 664 mouse exons.

To find rat conserved alternative cassette exons, we searched NCBI's HomoloGene database (www.ncbi.nlm.

nih.gov/entrez/query.fcgi?db=homologene) for rat RefSeq mRNAs that are homologous to the 823 human RefSeq mRNAs in our dataset. For each pair of human and rat homologous mRNAs, we aligned the human alternative exon and the rat mRNA, using clustalW. This yielded 319 rat mRNAs that include a sequence with >80% identity with the human alternative exon incorporated in their human homologous mRNA. We then wanted to verify that these sequences in rat mRNAs are indeed alternative cassette exons. For that, we ran two blast searches against dbEST: (i) of the putative rat alternative exon and 50 nt on both sides (from its flanking exons), representing the exon-including isoform; and (ii) of a concatenation of the flanking 50 nt, representing the exon-skipping isoform. We used only blast results with E-value $< 1 \times 10^{-4}$. If an EST was hit on both blasts, we performed pairwise alignment both of the EST and the sequence representing the exon-including isoform and of the EST and the sequence representing the exon-skipping isoform, using the clustalW algorithm. According to the clustalW score, we determined whether the EST includes the exon or skips it. This left us with 86 rat exons that have EST evidence showing they incorporate an alternative cassette exon (which is conserved in human). We repeated the same procedure with dog exons and found 22 dog exons conserved in human and with evidence for both their skipping and inclusion in dog.

Finding exon-skipping and exon-including isoforms

Detection of exon-skipping isoforms was conducted in the same manner as the search for ESTs that provide evidence for the skipping of rat and dog alternative exons (see above). However, for this analysis, the blast was performed against human and mouse GenBank mRNAs (rather than dbEST), because we searched for full-length mRNAs where the cassette exons are skipped. GenBank mRNAs were downloaded from the UCSC genome browser (http://hgdownload.cse.ucsc.edu/downloads.html).

Inclusion level calculation

We calculated the inclusion level only for exons for which at least one including and one skipping mRNA were found. The inclusion level of an exon was determined as the quotient of the number of mRNAs representing its exon-including isoform divided by the number of all mRNAs that include or skip the exon.

Normalization of exon location within CDS

To normalize the exon location within the CDS, we first calculated N, which is the number of all possible locations for the start point of this exon in the CDS without exceeding the boundaries of the CDS (N = CDS length − exon length + 1). The normalized location was determined as the quotient of the actual location of the exon's start point within the CDS divided by N.

Manual analysis of the effect on the reading frame of a random sample of non-symmetrical and symmetrical alternative exons

We used the mRNA/genomic alignments in the UCSC genome browser to manually compare between exon-skipping and exon-including mRNAs. We first blasted each mRNA against the genome, at http://genome.ucsc.edu/cgi-bin/hgGateway. Then we followed the link to the mRNA information and examined its mRNA/genomic alignment to determine its CDS boundaries and exon–intron junctions. This allowed us to determine the differences between the two isoforms in the CDS level (i.e. if they use different start and stop codons, if a frameshift occurs, etc.). All the CDS annotations we used are based on SWISS-PROT, TrEMBL, mRNA and RefSeq.

Statistical analysis

We used two statistical tests in this work: t-test and Fisher's exact test. t-test could be applied due to large sample size. For all statistical analyses, the level of statistical significance was fixed at 0.01.

RESULTS

To analyze conserved alternative splicing events, we compiled a dataset of 823 human alternative cassette exons (only exons for which the mouse homolog is alternatively spliced as well). Of these exons, 91.3% (751/823) are located within the CDS (CDS-internal exons). The rest either reside in the 5′ or 3′ untranslated regions (UTRs) (2.2%—18/823 and 0.4%—3/823, respectively) or contain the start or the stop codon (2.8%—23/823 and 3.4%—28/823, respectively). For each exon, we found a matching mRNA in NCBI's RefSeq database (see Materials and Methods). As a control, we repeated all analyses on a dataset of 39 720 human constitutively spliced exons (only exons for which the mouse homolog is constitutively spliced as well) and their matching 3062 RefSeq mRNAs, which was compiled by us previously (29). Analyses were repeated with a dataset of conserved alternatively spliced and constitutively spliced mouse exons, as well as with a dataset of conserved alternatively spliced exons of rat and dog. The results of these analyses are in correlation with the results obtained from analyzing the human exons dataset (see Supplementary Data).

We first confirmed that the CDS-internal alternative exons in our dataset are biased to be symmetrical. Indeed, 68.6% (515/751) were found to be symmetrical. As expected, the alternative exons that are not CDS-internal exhibit a much milder bias to be symmetrical (41.6%—30/72, are symmetrical), as do the CDS-internal constitutive exons (39.7%—15608/39253).

Percentage of symmetrical alternative exons that overlap with protein domains is low compared with non-symmetrical ones

We compared symmetrical and non-symmetrical alternative exons with respect to their potential to insert or remove protein domains. For that, we searched against Pfam and Smart databases (31,32) for protein domains in the mRNAs of the 751 alternative and 39 253 constitutive exons that are CDS-internal, and compared the percentage of symmetrical and non-symmetrical exons that overlap with domains (i.e. encode for a fragment of domain or possess the full extent of a coding sequence of a domain).

We found that the percentage of symmetrical alternative exons that overlap with protein domains is significantly

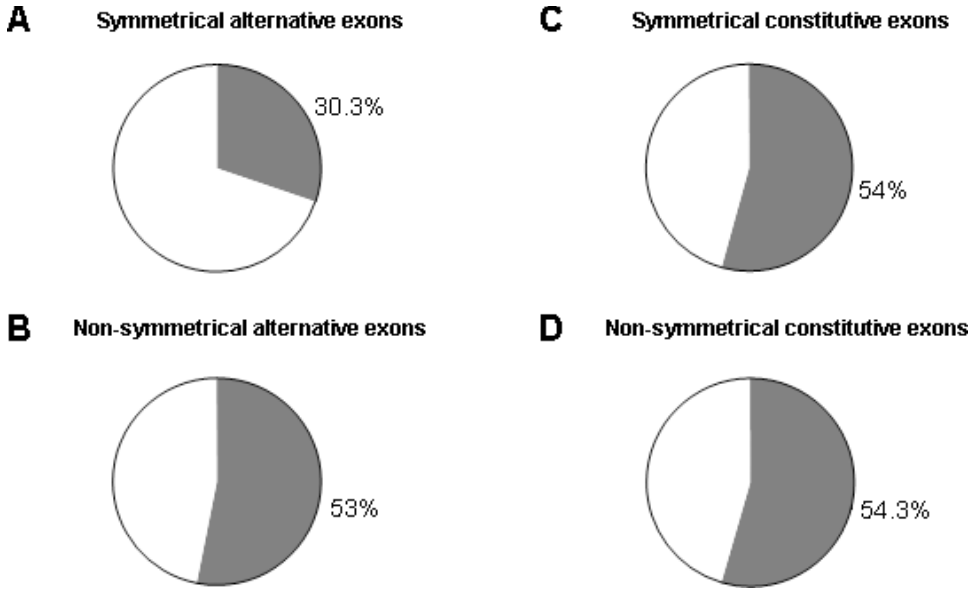

Figure 2. Percentage of exons that overlap with protein domain coding sequence. Gray and white indicate the fraction of exons that encodes protein domain fragments or full domains and those that do not encode for any fragment of a protein domain, respectively. (**A**) and (**B**) are for alternatively spliced exons, (**C**) and (**D**) are for constitutively spliced ones.

lower than that of non-symmetrical alternative exons and all constitutive exons. As shown in Figure 2, 30.3% (156/515) of symmetrical alternative exons overlap with domains, compared with 53% (126/236) of non-symmetrical alternative exons, $P < 9 \times 10^{-10}$ (Fisher's exact test). With respect to constitutive exons, 54% (8433/15 608) of symmetrical exons and 54.3% (12 850/23 645) of non-symmetrical exons overlap with coding sequences for protein domains. The differences between symmetrical and non-symmetrical constitutive exons and between non-symmetrical alternative exons and all constitutive exons are not statistically significant ($P > 0.2$ in all cases).

It has been previously shown that conserved alternative exons tend to be shorter than conserved constitutive exons (33). Therefore, the bias of symmetrical alternative exons against overlapping with protein domains could potentially stem from their relative short length. We found that the average length of symmetrical alternative exons is indeed slightly shorter than that of non-symmetrical ones: 113.3 and 116.9 nt, respectively, but this difference is not statistically significant ($P < 0.6$, *t*-test). Moreover, among constitutive exons, symmetrical exons are also shorter than non-symmetrical ones: 128.4 and 131.2 nt, respectively, and this difference does reach statistical significance ($P < 0.01$). Therefore, the slight difference in the average length of symmetrical and non-symmetrical alternative exons cannot account for the tendency of symmetrical alternative exons not to overlap with protein domains.

Symmetrical alternative exons tend not to overlap with medium-sized fragments of protein domains

To further analyze the tendency of symmetrical alternative exons not to overlap with protein domains, we compared the percentage of symmetrical and non-symmetrical alternative and constitutive exons that encode different fractions of domains, in increments of 10% (Figure 3A). Among alternative exons, we found that symmetrical exons are underrepresented among exons that overlap with domain fragments consisting of >10% and <60% of domains (which we term medium-sized fragments). Among constitutive exons, the differences between symmetrical and non-symmetrical exons reach a statistical significance only among exons that encode complete domains (rather than only a fraction). This finding is in correlation with previous studies, which showed that most exons (constitutive and alternative) that hold complete protein domains are symmetrical (23,27). Symmetrical alternative exons also show a higher tendency to encode complete domains compared with non-symmetrical ones, but this result does not reach statistical significance.

Our analysis is summarized in Figure 3B–E. It shows that only symmetrical alternative exons tend not to encode medium-sized fragments of domains (Figure 3B compared with Figure 3C–E). Therefore, the low percentage of alternative symmetrical exons that encode protein domain fragments or complete domains stems from the tendency of these exons not to encode medium-sized fragments of domains. The pattern of encoding different fractions of domains is similar between non-symmetrical alternative exons and all constitutive exons (symmetrical or non-symmetrical).

Skipping isoforms of non-symmetrical alternative exons have a low representation in GenBank

Inclusion or skipping of a non-symmetrical alternative exon induces a frameshift that may result in the introduction of a PTC that induces degradation of the mRNA by the NMD mechanism [see examples in (34,35)]. Alternatively, an alternative exon (symmetrical or non-symmetrical) can harbor a PTC [see examples in (36,37)]. As only 3.4% of the alternative exons in our dataset encode for a stop codon, we searched for evidence of PTC generation only as result of frameshifts

A

		0%	<10%	<20%	<30%	<40%	<60%	<80%	<70%	<30%	<80%	<100%	100%
Alternative exons	Symmetrical (616 exons)	69.7%	11.8%	7.6%	4.3%	1.7%	1.2%	0.2%	0.6%	0.2%	0.0%	0.4%	2.3%
	Non-sym. (238 exons)	47.0%	11.0%	14.4%	9.7%	5.5%	4.2%	5.5%	1.3%	0.0%	1.4%	0.0%	0.0%
Constitutive exons	Symmetrical (16803 exons)	46.0%	14.0%	16.1%	9.0%	4.5%	3.0%	2.2%	1.2%	0.6%	0.5%	0.3%	2.6%
	Non-sym. (23846 exons)	45.7%	14.8%	17.1%	8.6%	4.6%	3.4%	2.2%	1.3%	0.8%	0.5%	0.4%	0.6%

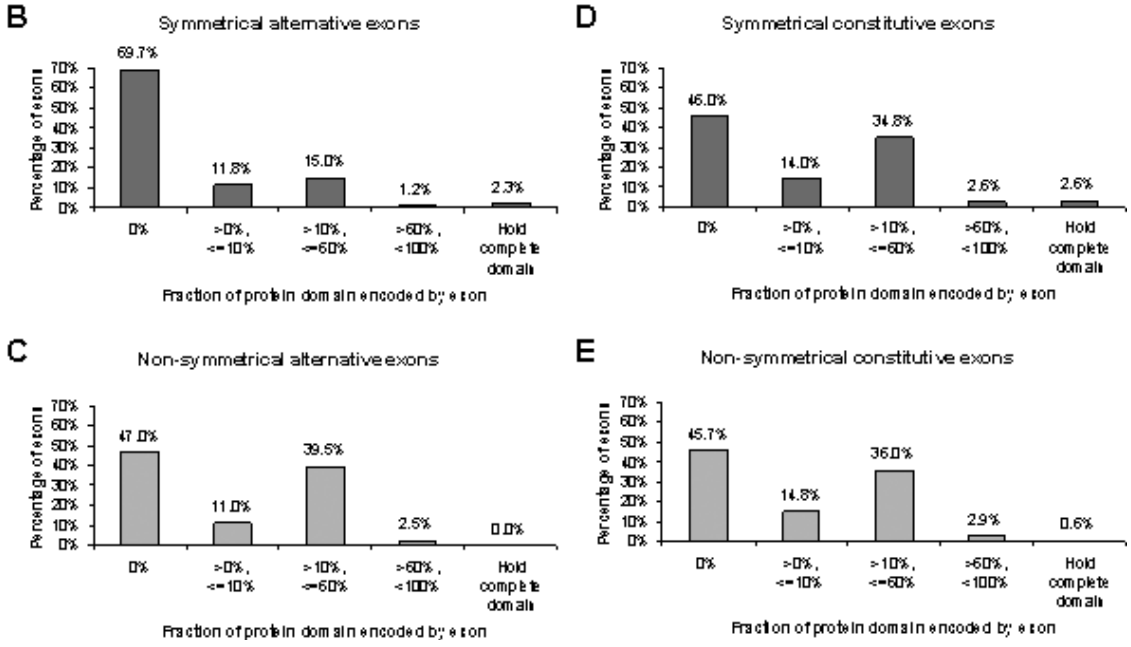

Figure 3. Percentage of symmetrical and non-symmetrical exons that encode different fractions of protein domains. (**A**) The percentage of symmetrical and non-symmetrical alternative and constitutive exons that encode for different fragments of protein domains, starting with exons not encoding any domain fragment (0%), then encoding different sizes of fragments (in increments of 10%) up to encoding complete domains (the bar marked 100%). Statistically significant differences between symmetrical and non-symmetrical alternative exons and between symmetrical and non-symmetrical constitutive exons are highlighted in gray. (**B**–**E**) A histogram representation of the results shown in (A), indicating the percentage of exons that do not encode protein domains (0%), encodes more than 0% and up to 10% of protein domains (>0%, ≤10%), encodes medium-sized fragments (>10%, ≤60%), encodes fragments that consist of >60% and <100% of protein domains or holds complete domains. B and C show the distribution of alternative exons, D and E show that of constitutive exons.

induced by alternative splicing of non-symmetrical exons. For that, we used BLAST to search against GenBank mRNAs for the exon-including and exon-skipping isoforms of the alternative exons that are CDS-internal (see Materials and Methods). We used GenBank mRNAs for this analysis because PTC^+ mRNAs were found to have a low representation in RefSeq, compared with GenBank (15).

Interestingly, we detected an exon-skipping mRNA for only 42.7% (101/236) of the non-symmetrical exons, compared with 92.2% (475/515) of the symmetrical exons (exons for which no skipping mRNA was detected are known to be alternative based on EST-data solely). Then, for all alternative exons for which skipping mRNAs were found, we calculated the inclusion level, which is the fraction of total transcripts of the gene that includes the exon (see Materials and Methods). As already mentioned, conserved alternative exons usually exhibit a high inclusion level (22). However, we found that the average inclusion level of non-symmetrical exons is significantly higher than that of symmetrical exons: 0.74 and 0.59, respectively, $P < 1 \times 10^{-31}$ (*t*-test). The low percentage of non-symmetrical exons with mRNA evidence for their skipping isoform, along with the high inclusion level of these exons, shows that the exon-skipping isoforms of non-symmetrical alternative exons have a very low coverage among GenBank mRNAs. This may indicate that skipping of a large fraction of conserved non-symmetrical alternative exons leads to NMD activation.

Non-symmetrical alternative exons tend to reside near the ends of the CDS, with a higher preference for 5′ locations

As NMD activation depends upon the location of the termination codon relative to the last exon–exon junction, we wanted to examine whether non-symmetrical alternative exons tend to reside in specific portions of the CDS. For that, we found the location of all CDS-internal exons in their including mRNA by normalizing the CDS length between 0 and 1 (see Materials and Methods), and compared, in increments of 0.1, the percentage of symmetrical and non-symmetrical alternative and constitutive exons at different normalized locations (Figure 4). We found that non-symmetrical exons exhibit a tendency to reside in the 5′ half of the CDS, with average location of 0.39, median 0.31 (Figure 4B). There is also a slight rise in the percentage of non-symmetrical alternative exons toward the 3′ end of the CDS. Symmetrical alternative exons and

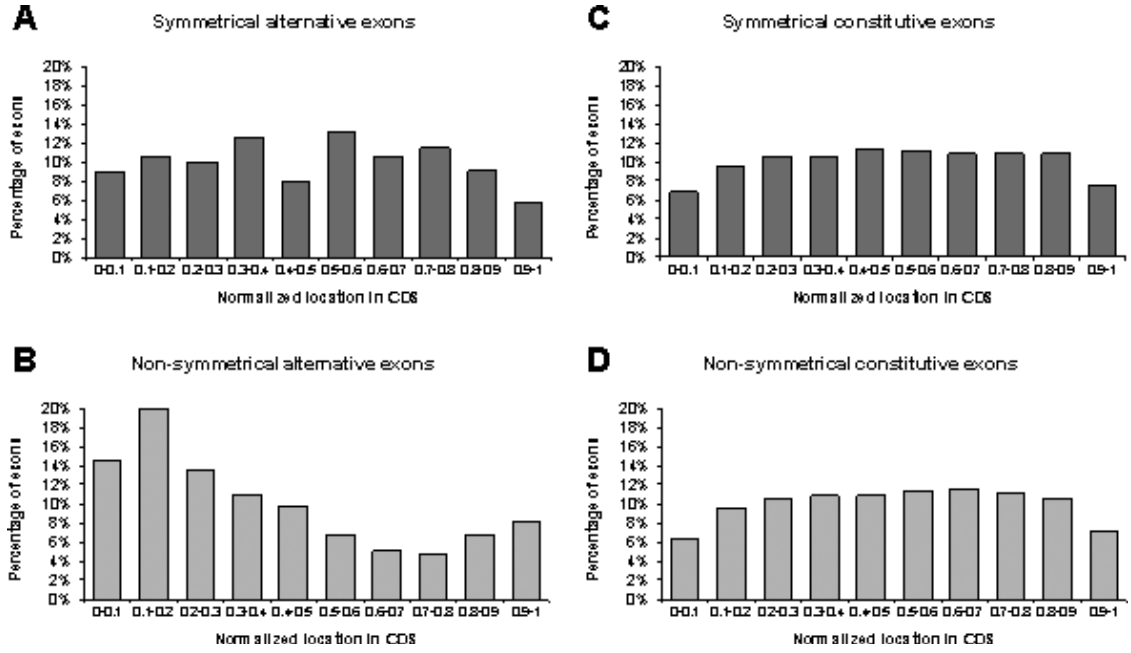

Figure 4. Percentage of symmetrical and non-symmetrical exons along the CDS. (**A**–**D**) Percentage of exons in different locations along the CDS, in increments of 0.1 (exon locations in CDS are normalized between 0 and 1, see Supplementary Data). A and B show the distribution of alternative exons, C and D show that of constitutive exons.

all constitutive exons do not show any bias for specific locations within the CDS: average location of 0.49 and 0.51 for symmetrical alternative and constitutive exons, respectively (Figures 4A, C and D). The difference between the average normalized location of symmetrical and non-symmetrical alternative exons is statistically significant: $P < 2.7 \times 10^{-0.5}$ (*t*-test). There is also a slight drop in the percentage of symmetrical alternative exons and all constitutive exons near the ends of the CDS, which may result from a tendency of the start and stop codons not to lie near exon–exon junctions (as it is in exons with average normalized location which is between 0 and 0.1 or 0.9 and 1). We also checked whether the 43% of non-symmetrical alternative exons, which have mRNA evidence for their skipping, differ from the rest of the non-symmetrical alternative exons with respect to a preference to reside in a specific location along the CDS, but found no such difference (data not shown).

Analysis of the effect on the reading frame of alternative splicing events of non-symmetrical exons

Skipping or inserting of cassette exons can be accompanied by other alternative splicing events (splicing of other exons, use of alternative initiation or termination, etc.). Careful analysis of all the differences between the exon-including isoforms and the exon-skipping ones may require a very sophisticated computational algorithm. Therefore, we decided to perform such an analysis manually on a random sample of alternative exons. For that, we randomly sampled a group of 30 symmetrical alternative exons and 30 non-symmetrical ones of the 101 and 475 symmetrical and non-symmetrical exons, respectively, that are CDS-internal and have mRNA evidence in GenBank for both their inclusion and skipping. For each of the 60 sampled exons, we used the mRNA/genomic alignments in the UCSC genome browser (see Materials and Methods) to compare one of their exon-skipping isoforms with one of their exon-including isoforms, with regard to preservation of the reading frame between the two isoforms and generation of a PTC. Stop codons were considered as PTCs if they were located more than 55 nt upstream from the 3′-most exon–exon junction (38).

For the 30 non-symmetrical exons, we detected three categories of differences between the including and skipping isoforms (Figure 5). For ∼53% (16/30) of these exons (category A), the reading frame downstream of the alternative exon is different in the skipping and including isoforms. Among these 16 exons, the skipping isoform contains a PTC in four cases and the including isoform contains a PTC in two cases. In category B, which contains ∼37% (11/30) of the exons, an alternative start codon ensures the use of the same reading frame downstream from the alternative exon (but introduces a different reading frame upstream to the alternative exons). Among these 11 exons, the skipping isoform is PTC^+ in one case. Category C contains three exons for which the including and skipping isoforms have a different reading frame (both upstream and downstream from the alternative exon). Among these three exons, the skipping isoform contains a PTC in one case, and, in another case, the including isoform is PTC^+. To summarize, in 20% (6/30) of pairs the skipping isoform was PTC^+; in 10% (3/30), the including isoform was PTC^+; and, in 70%, none of the isoforms harbored a PTC. For comparison, none of the skipping or including isoforms of the 30 symmetrical alternative exons meets the definition of a PTC^+ mRNA.

We also examined the location in CDS of the alternative non-symmetrical exons in categories A and B, for which neither inclusion nor skipping introduces a PTC. For the 10 exons in category A, where both isoforms share the same reading frame upstream from the alternative exon, the average normalized location is 0.68 (median 0.76). For the 10 exons in

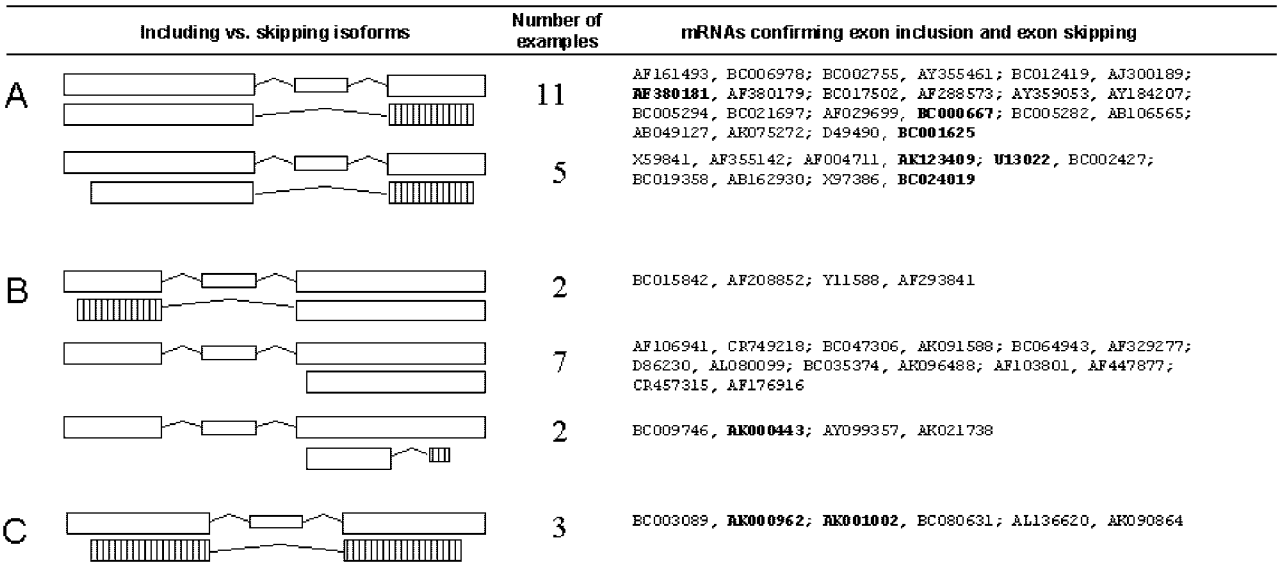

Figure 5. Differences between the including and skipping isoforms of 30 random non-symmetrical alternative exons. Six different types of pairs of inclusion/skipping isoforms were grouped into three categories (A–C). In each pair, the upper and lower isoforms represent the including and skipping mRNA, respectively. The accession numbers of all inclusion/skipping pairs that represent this category are indicated on the right (including isoform, skipping isoform). Accession numbers of PTC^+ mRNAs are in bold. Open boxes indicate the same reading frame for both isoforms, and boxes with vertical lines indicate different reading frames. The boxes represent only the CDS portion of the mRNA. In category (**A**), skipping of the alternative exon induces a frameshift downstream from the exon. In 11 cases, the including and skipping isoforms have the same start codon (ATG). In five cases, the two isoforms use different ATGs, but the reading frame upstream from the alternative exon remains the same. In category (**B**), use of an alternative ATG ensures that both isoforms share the same reading frame downstream from the alternative exon. In two cases, both ATGs are upstream from the exon, and, thus, the two isoforms have a different reading frame upstream from the exon. In seven cases, the skipping isoform uses an alternative in-frame ATG downstream from the alternative exon. In the other two cases, an alternative in-frame ATG upstream from the exon is used, but an alternative splicing event downstream from the exon introduces a different stop codon at the skipping isoform. In category (**C**), both isoforms differ in their start and stop codons and do not share the same reading frame throughout the CDS.

Table 1. Differences between symmetrical and non-symmetrical alternative exons and constitutive ones, with respect to several characteristics

Characteristic	Alternative exons		Constitutive exons	
	Symmetrical	Non-Symmetrical	Symmetrical	Non-Symmetrical
Percentage of total CDS-internal exons	68.6%	31.4%	40%	60%
Average exon size (nt)	113.3	116.9[a]	128.4	131.2
Percentage of exons that overlap with protein-domains				
Total[b]	30.3%	53%	54%	54.3%[a]
Medium-sized[c]	15%	39.4%	34.8%	36%[a]
Full domain[d]	2.3%	0%[a]	2.6%	0.06%
Percentage of exons with mRNA evidence of skipping	92.2%	42.7%	N/A	N/A
Average inclusion level	0.59	0.74	N/A	N/A
Average location in RefSeq mRNAs	0.49	0.39	0.51	0.51[a]

[a]Difference did not reach statistical significance.
[b]Percentage of exons overlapping with protein-domains fragments or complete protein-domains.
[c]Percentage of exons overlapping with medium-sized fragments (10–60%) of protein-domains.
[d]Percentage of exons that possess a complete protein-domain.

category B, where both isoforms share the same reading frame downstream from the alternative exon, the average location is 0.21 (median 0.19). These findings show that the location of these alternative non-symmetrical exons ensures that most of the reading frame remains unchanged in both the skipping and including isoforms. For comparison, the average normalized location of the 30 symmetrical exons is 0.48 (median 0.49), indicating no preference of these exons for specific locations within the CDS.

DISCUSSION

In this report, we found that conserved symmetrical and non-symmetrical alternative exons differ with respect to several characteristics. We showed that the tendency of symmetrical alternative exons to overlap with protein domains is significantly lower, compared with non-symmetrical alternative exons and all constitutive exons. This low tendency results from the bias of symmetrical alternative exons not to overlap with medium-sized fragments (10–60%) of protein domains. We also showed that non-symmetrical alternative exons tend to reside in the 5′ half of the CDS and have a higher inclusion level, compared with symmetrical alternative exons. Our findings are summarized in Table 1. Supplementary Table 1 shows that alternative cassette exons in mouse, rat and dog that are conserved in human possess similar characteristics.

In our dataset, 68.6% of human CDS-internal alternative exons that are conserved in mouse are symmetrical. Other

studies reported varying percentages of symmetrical exons among conserved CDS-internal alternative exons, ranging from 53 to 77% (19,25,26). However, as conserved CDS-internal alternative exons show a strong bias to be symmetrical in all studies, these discrepancies probably result from differences in dataset compilation and do not reflect meaningful contradictions between the different works.

It has been previously reported that the borders of alternative exons tend to correlate with protein domain borders, probably to eliminate potentially deleterious splice variants that contain broken domains (39). The observed bias of symmetrical alternative exons not to overlap with medium-sized fragments of protein domains shows that this tendency is much stronger among symmetrical alternative exons. This bias may reflect a negative selection on symmetrical exons that produce transcripts with broken domains. Non-symmetrical alternative exons show the same pattern of overlapping with different fractions of domains as do constitutive exons, which may indicate that non-symmetrical alternative exons are indifferent to this selection pressure.

As inclusion or skipping of symmetrical exons does not alter the mRNA's reading frame, it would be reasonable to hypothesize that, compared with non-symmetrical ones, these exons would be more prone to play the role of domain-holders (i.e. alternative exons that encode large fragments of domains and thus their skipping eliminates the domain functionality). We did find that among exons that encode >60% of protein domains, symmetrical alternative exons have a slightly higher representation, but this result did not reach statistical significance. Skipping of exons that encode >60% of protein domains is very likely to eliminate the functionality of the domains with which these exons overlap. Therefore, such exons probably act as domain-holders. However, our findings suggest that only a small fraction (<4%) of alternative exons are domain-holders. This indicates that selective removal or insertion of protein domains (at least of the type represented in Pfam and SMART databases) into different splice forms is not a major role of alternative splicing. We also found that the ratio of symmetrical and non-symmetrical alternative exons that encode such large fragments of protein domains is similar to that of symmetrical and non-symmetrical constitutive exons. This may indicate that the tendency of alternative exons to be domain-holders is not higher than expected by chance.

We also found that mRNAs representing the skipping of non-symmetrical alternative exons are strongly underrepresented in GenBank, compared with those of symmetrical alternative exons. It has already been suggested that mRNAs subjected to NMD have a low coverage in databases because of their rapid degradation (15). Therefore, it is possible that a large fraction of the exon-skipping isoforms of non-symmetrical exons underwent NMD degradation, and, thus, these isoforms are not detected. The finding that non-symmetrical alternative exons tend to reside in the 5′ half of the CDS (average normalized location of half of them is lower than 0.31) can also be related to NMD activation. Reasonably, skipping of non-symmetrical exons from more upstream locations, not accompanied with use of an alternative in-frame start codon, has a higher potential to introduce a stop codon in the frameshift induced downstream from the skipped exon. Also, PTCs closer to the 5′ end may trigger more robust NMD, as in the case of the TCR-β gene (40). Therefore, it is possible that skipping events of non-symmetrical exons are preserved in evolution for NMD activation, which may be more effective if they are skipped from locations within the 5′ half of the CDS.

However, there is also an increase in the representation of non-symmetrical alternative exons near the 3′ end of the CDS. This increase was observed for both non-symmetrical exons with mRNA evidence for their skipping isoforms and those without such evidence. A manual inspection of the effects on the reading frame of 30 random splicing events of non-symmetrical alternative exons suggests that non-symmetrical alternative exons that do not generate PTC^+ mRNAs tend to reside near the ends of the CDS, and this allows the including and skipping isoforms to use the same reading frame throughout most of the CDS. In half of such cases in our sample, the exons resided in proximity to the 5′ end of the CDS and the including and skipping isoforms used a different ATG. This allowed the preservation of the reading frame downstream from the exon. Elimination of an otherwise potent PTC by means of alternative initiation has already been reported in the case of the β-globin and TPI genes (41,42). In the other cases, the exons were found to reside in proximity to the 3′ end of the CDS, and, as a result, the frameshift they induced was short and did not harbor a PTC. This also allowed the preservation of most of the reading frames in both the including and skipping isoforms.

The low coverage of the skipping isoforms of non-symmetrical alternative exons, along with the fact that, in 2/3 of alternative splicing events that generated PTCs in our random sample, the skipping isoform was the PTC^+, suggests that the most prevalent mechanism for the generation of a PTC is skipping of a non-symmetrical exon. As only ~3% of alternative exons in our dataset encode for a stop codon, it is likely that creation of a stop codon by means of insertion of a cassette exon that harbors a PTC is a much less-used mechanism for NMD activation. Moreover, as our findings suggest that the including isoform of many non-symmetrical alternative exons is being translated into protein and that the skipping isoforms undergoes NMD degradation, it is likely that the including isoform of these exons represents the ancestral state of these exons. This supports a hypothesis we recently suggested that constitutively spliced exons can switch to alternative splicing in the course of evolution (21). The average high inclusion level of all conserved alternative exons and the tendency of these exons not to contain repetitive elements further support this idea. According to this model, the selection against transition of non-symmetrical constitutive exons to alternative splicing is weaker near the CDS ends, because skipping these exons either preserves most of the reading frames in the skipping isoform (as described above) or can be used for the generation of NMD-targeted transcripts that may play a regulatory role. Transition of symmetrical exons from constitutive to alternative splicing is probably under negative selection if these exons encode for medium-sized fragments of protein domains.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank David Haussler, Chuck W Sugnet, Ido Carmel, Rotem Sorek and Amir Goren for sharing datasets, and Rotem Sorek for helpful discussion. This work was supported by a grant from the Israel Science Foundation (1449/04 and 40/05) and, in part, by a grant from the German Israeli Project Cooperation Program, German Israeli R&D, and the Chief Scientist of the Israel Health Ministry to G.A. Funding to pay the Open Access publication charges for this article was provided by German Israeli Research and Development.

Conflict of interest statement. None declared.

REFERENCES

1. Lareau,L.F., Green,R.E., Bhatnagar,R.S. and Brenner,S.E. (2004) The evolving roles of alternative splicing. *Curr. Opin. Struct. Biol.*, **14**, 273–282.
2. Lee,C. and Wang,Q. (2005) Bioinformatics analysis of alternative splicing. *Brief Bioinformatics*, **6**, 23–33.
3. Woodley,L. and Valcarcel,J. (2002) Regulation of alternative pre-mRNA splicing. *Brief Funct. Genomic. Proteomic*, **1**, 266–277.
4. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
5. Leipzig,J., Pevzner,P. and Heber,S. (2004) The alternative splicing gallery (ASG): bridging the gap between genome and transcriptome. *Nucleic Acids Res.*, **32**, 3977–3983.
6. Schmucker,D., Clemens,J.C., Shu,H., Worby,C.A., Xiao,J., Muda,M., Dixon,J.E. and Zipursky,S.L. (2000) Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell*, **101**, 671–684.
7. Cline,M.S., Shigeta,R., Wheeler,R.L., Siani-Rose,M.A., Kulp,D. and Loraine,A.E. (2004) The effects of alternative splicing on transmembrane proteins in the mouse genome. *Pac. Symp. Biocomput.*, 17–28.
8. Grabowski,P.J. and Black,D.L. (2001) Alternative RNA splicing in the nervous system. *Prog. Neurobiol.*, **65**, 289–308.
9. Copley,R.R. (2004) Evolutionary convergence of alternative splicing in ion channels. *Trends Genet*, **20**, 171–176.
10. Bassett,C.L., Artlip,T.S. and Callahan,A.M. (2002) Characterization of the peach homologue of the ethylene receptor, PpETR1, reveals some unusual features regarding transcript processing. *Planta*, **215**, 679–688.
11. Hillman,R.T., Green,R.E. and Brenner,S.E. (2004) An unappreciated role for RNA surveillance. *Genome Biol.*, **5**, R8.
12. Holbrook,J.A., Neu-Yilik,G., Hentze,M.W. and Kulozik,A.E. (2004) Nonsense-mediated decay approaches the clinic. *Nature Genet.*, **36**, 801–808.
13. Frischmeyer,P.A. and Dietz,H.C. (1999) Nonsense-mediated mRNA decay in health and disease. *Hum. Mol. Genet.*, **8**, 1893–1900.
14. Maquat,L.E. (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nature Rev. Mol. Cell Biol.*, **5**, 89–99.
15. Lewis,B.P., Green,R.E. and Brenner,S.E. (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl Acad. Sci. USA*, **100**, 189–192.
16. Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.
17. Maniatis,T. and Tasic,B. (2002) Alternative pre-mRNA splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.
18. Baranova,A.V., Lobashev,A.V., Ivanov,D.V., Krukovskaya,L.L., Yankovsky,N.K. and Kozlov,A.P. (2001) *In silico* screening for tumour-specific expressed sequences in human genome. *FEBS Lett.*, **508**, 143–148.
19. Sorek,R., Shamir,R. and Ast,G. (2004) How prevalent is functional alternative splicing in the human genome? *Trends Genet.*, **20**, 68–71.
20. Xu,Q. and Lee,C. (2003) Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Res.*, **31**, 5635–5643.
21. Ast,G. (2004) How did alternative splicing evolve? *Nature Rev. Genet.*, **5**, 773–782.
22. 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. *Nature Genet.*, **34**, 177–180.
23. Kaessmann,H., Zollner,S., Nekrutenko,A. and Li,W.H. (2002) Signatures of domain shuffling in the human genome. *Genome Res.*, **12**, 1642–1650.
24. Philipps,D.L., Park,J.W. and Graveley,B.R. (2004) A computational and experimental approach toward a priori identification of alternatively spliced exons. *RNA*, **10**, 1838–1844.
25. Resch,A., Xing,Y., Alekseyenko,A., Modrek,B. and Lee,C. (2004) Evidence for a subpopulation of conserved alternative splicing events under selection pressure for protein reading frame preservation. *Nucleic Acids Res.*, **32**, 1261–1269.
26. Baek,D. and Green,P. (2005) Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing. *Proc. Natl Acad. Sci. USA*, **102**, 12813–12818.
27. Liu,M. and Grigoriev,A. (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes—evidence of exon shuffling? *Trends Genet.*, **20**, 399–403.
28. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **33**, D501–D504.
29. Carmel,I., Tal,S., Vig,I. and Ast,G. (2004) Comparative analysis detects dependencies among the 5′ splice-site positions. *RNA*, **10**, 828–840.
30. Sugnet,C.W., Kent,W.J., Ares,M.,Jr and Haussler,D. (2004) Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pac. Symp. Biocomput.*, 66–77.
31. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L. (2000) The Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
32. Letunic,I., Copley,R.R., Schmidt,S., Ciccarelli,F.D., Doerks,T., Schultz,J., Ponting,C.P. and Bork,P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142–D144.
33. Sorek,R., Shemesh,R., Cohen,Y., Basechess,O., Ast,G. and Shamir,R. (2004) A non-EST-based method for exon-skipping prediction. *Genome Res.*, **14**, 1617–1623.
34. Jones,R.B., Wang,F., Luo,Y., Yu,C., Jin,C., Suzuki,T., Kan,M. and McKeehan,W.L. (2001) The nonsense-mediated decay pathway and mutually exclusive expression of alternatively spliced FGFR2IIIb and -IIIc mRNAs. *J. Biol. Chem.*, **276**, 4158–4167.
35. Wollerton,M.C., Gooding,C., Wagner,E.J., Garcia-Blanco,M.A. and Smith,C.W. (2004) Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay. *Mol. Cell*, **13**, 91–100.
36. Winter,J., Lehmann,T., Krauss,S., Trockenbacher,A., Kijas,Z., Foerster,J., Suckow,V., Yaspo,M.L., Kulozik,A., Kalscheuer,V. *et al.* (2004) Regulation of the MID1 protein function is fine-tuned by a complex pattern of alternative splicing. *Hum. Genet.*, **114**, 541–552.
37. Mitrovich,Q.M. and Anderson,P. (2000) Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in *C.elegans*. *Genes Dev.*, **14**, 2173–2184.
38. Nagy,E. and Maquat,L.E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.*, **23**, 198–199.
39. Kriventseva,E.V., Koch,I., Apweiler,R., Vingron,M., Bork,P., Gelfand,M.S. and Sunyaev,S. (2003) Increase of functional diversity by alternative splicing. *Trends Genet.*, **19**, 124–128.
40. Wang,J., Gudikote,J.P., Olivas,O.R. and Wilkinson,M.F. (2002) Boundary-independent polar nonsense-mediated decay. *EMBO Rep.*, **3**, 274–279.
41. Zhang,J. and Maquat,L.E. (1997) Evidence that translation reinitiation abrogates nonsense-mediated mRNA decay in mammalian cells. *EMBO J.*, **16**, 826–833.
42. Inacio,A., Silva,A.L., Pinto,J., Ji,X., Morgado,A., Almeida,F., Faustino,P., Lavinha,J., Liebhaber,S.A. and Romao,L. (2004) Nonsense mutations in close proximity to the initiation codon fail to trigger full nonsense-mediated mRNA decay. *J. Biol. Chem.*, **279**, 32170–32180.