Nucleic Acids Research, 2006, Vol. 34, No. 1 ***23–31***
doi:10.1093/nar/gkj408

Alternative splicing regulation at tandem 3′ splice sites

Martin Akerman and Yael Mandel-Gutfreund*

Department of Biology, Technion-Institute of Technology, Haifa 32000, Israel

Received October 17, 2005; Revised and Accepted December 6, 2005

ABSTRACT

Alternative splicing (AS) constitutes a major mechanism creating protein diversity in humans. Previous bioinformatics studies based on expressed sequence tag and mRNA data have identified many AS events that are conserved between humans and mice. Of these events, ∼25% are related to alternative choices of 3′ and 5′ splice sites. Surprisingly, half of all these events involve 3′ splice sites that are exactly 3 nt apart. These tandem 3′ splice sites result from the presence of the NAGNAG motif at the acceptor splice site, recently reported to be widely spread in the human genome. Although the NAGNAG motif is common in human genes, only a small subset of sites with this motif is confirmed to be involved in AS. We examined the NAGNAG motifs and observed specific features such as high sequence conservation of the motif, high conservation of ∼30 bp at the intronic regions flanking the 3′ splice site and overabundance of *cis*-regulatory elements, which are characteristic of alternatively spliced tandem acceptor sites and can distinguish them from the constitutive sites in which the proximal NAG splice site is selected. Our findings imply that AS at tandem splice sites and constitutive splicing of the distal NAG are highly regulated.

INTRODUCTION

In eukaryotic cells, gene expression is controlled at the level of transcription as well as by post-transcriptional events, such as alternative splicing (AS). The intron/exon boundaries in eukaryotic genes are defined by short and degenerate classical splice sites (5′ splice site, 3′ splice site and the branch point), which are involved in recognition and interaction with the basal splicing machinery. AS allows an individual gene to express different combinations of exons, which in turn can encode proteins with diverse and even antagonistic functions (1). In themselves, the splice sites do not contain enough information to explain the complex regulation of AS. However, recently it has become clear that this fine-tuned mechanism is achieved by multiple weak signals across the exons and introns which are recognized by an extensive number of different proteins (2).

Recent bioinformatics studies have shown that exons which undergo AS are distinguishable from constitutively spliced exons by several features. Among these characteristics is the divisibility by three of the exon length, which is likely to ensure the preservation of the reading frame in the mRNA (3). Another feature is the evolutionary conservation of intronic sequences flanking AS exons (3–8). Though not fully understood, the unusually high conservation of the introns surrounding AS exons suggests the presence of *cis*-regulatory elements within these regions, such as intronic enhancers and silencers that control AS by interacting with regulatory proteins (5).

The most common and extensively studied form of AS is exon skipping (ES), which represents ∼40% of all AS events conserved between humans and mice (4). However, AS can also be attained by altering the position of the splice donor or acceptor sites. These processes are known as alternative 3′ splice site (3′AS) and alternative 5′ splice site (5′AS) and contribute together to ∼25% of all AS events conserved between humans and mice (4). Recently, Hiller *et al*. (9) reported a widespread occurrence of a NAGNAG 3′ acceptor splice site motif in the human genome. The NAGNAG motif includes two 3′ splice site motifs in tandem and thus has the potential of producing mRNA isoforms which differ by a 3 nt sequence (NAG). Based on their analysis, Hiller *et al*. (9) suggested that the NAGNAG motif, which can insert or delete a single amino acid in the protein, is present in 30% of human genes and is functional in at least 5% of the genes. In addition it has been observed that alternative spliced isoforms, resulting from the NAGNAG motif are differentially expressed in human and mouse tissues (9,10).

Since the process of AS is highly specific and thus expected to be strictly regulated (11), it is intriguing to postulate that similar to other AS events the AS at the NAGNAG 3′ splice site is controlled by an extensive regulatory mechanism. If AS at the NAGNAG acceptor sites is related to other AS events, one expects to find common features between these sites and other sites in the genome that undergo AS, such as ES. Moreover, it is likely that some NAGNAG acceptor sites are not alternatively spliced and instead, one of the acceptor sites will be constitutively chosen. Previous studies demonstrated that when two AG sites are closely located, the one proximal to the branch point is selected as an acceptor by a scanning

*To whom correspondence should be addressed. Tel: +972 4 8293958; Fax: +972 4 8225153; Email: yaelmg@tx.technion.ac.il

mechanism (12,13). However, it was also observed that when the distance between the two AG sites is short (<6 bp), the distal site might be chosen (14). In these cases, mutations in the proximal AG prevent recognition of the distal AG, while mutations surrounding the proximal AG may shift the selection towards the distal AG (14,15). These findings suggest that the selection of an acceptor site depends on the sequence environment, and can be altered by subtle changes such as point mutations. This is consistent with other studies showing that the pattern of AS can be altered by mutations in exonic splicing enhancer sites (ESEs) and exonic splicing enhancer sites (ESSs) (16).

In this work we studied the AS at the NAGNAG tandem 3′ splice sites. We have observed that NAGNAG sites are the major form of alternative 3′ splice site in the human genome, accounting for half of all alternative 3′ splice sites. Based on a comprehensive analysis of ~6000 NAGNAG sites we propose that, by itself, the NAGNAG motif is not sufficient for AS. Nevertheless, analysis of a subset of the NAGNAG sites confirmed by expressed sequence tag (EST) data to be alternatively spliced shows that they encompass several characteristics of other known AS events. Comparison between the constitutively and alternatively spliced NAGNAG sites revealed that they differ principally by three major properties: (i) the sequence and evolutionary conservation of the NAGNAG motif, (ii) the conservation of intron sequences flanking the NAGNAG site, and (iii) the abundance of known *cis*-regulatory elements in the neighboring regions of the 3′ splice sites.

MATERIALS AND METHODS

Data extraction

We have analyzed a dataset of 460 3′AS and 207 5′AS events, which are EST-confirmed and conserved between human and mouse (4). Among them we extracted 215 events with exact 3 nt 3′AS.

From the dataset of 102 461 constitutively spliced exons (4), 5634 exons with the NAGNAG motif at the 3′ splice site were extracted. Among them 5050 are confirmed by EST to be spliced at the proximal NAG site, 'Proximal' group [E transcript, as defined by Hiller *et al* (9)], while 584 are confirmed by EST to be spliced at the distal NAG site, 'Distal' group [I transcript, as defined by Hiller *et al*. (9)].

Gene ontology (GO)

The GOA database (http://www.geneontology.org/ontology/function.ontology) was downloaded and used to annotate the genes in which the different events have been observed. The set of genes for which annotations were available included 122 genes with the EST-confirmed NAGNAG 3′AS sites, occurring within coding sequences (not including events in untranslated regions), 483 genes including the EST-confirmed ES cases, and 8377 genes in which conserved constitutive splicing events were observed. The two latter groups were used as reference sets. For simplifying the classification and comparison results, for each dataset the overall frequency of the GO terms at the fourth hierarchy of the 'molecular function' sub-ontology was calculated. An in house Perl script was used to automatically extract the GO annotations. GO terms for which the number of counts was less than five were pooled in a group we defined as 'others'.

A χ^2-test for goodness of fit was applied to compare the GO terms between the alternative spliced NAGNAG set of genes and (i) ES gene set and (ii) constitutive spliced gene set. For each GO term in the alternatively spliced NAGNAG group that had more than five counts a *P*-value was calculated separately, comparing the counts of the term in the NAGNAG 3′AS group to the counts in each of the two reference sets. As the large numbers did not allow calculating the exact *P*-value under the hypergeometric distribution, for each GO term a 2×2 χ^2-test was applied in order to approximate the *P*-value (17). To account for multiple testing the Bonferroni correction was applied, thus only *P*-values lower than 0.0038 (0.05/13) were considered statistically significant.

Conservation at the tandem acceptor sites

The human/mouse (mm5) pairwise alignment database was downloaded from the UCSC site (http://hgdownload.cse.ucsc.edu/goldenPath/hg17/vsMm5). Tandem acceptor sites were analyzed only in the following groups in which information was available: 179 EST-confirmed NAGNAG 3′AS sites, 4014 'Proximal' and 406 'Distal'. Graphical representation was conducted with the LOGOS software (18) on the three different sets of aligned mouse homologous acceptors and the percent conservation per-position was calculated. The height of the symbols represents the relative frequency of each nucleotide in a given position.

Intronic conservation

Sequence conservation was calculated in a range between 30 bp upstream the splice acceptor and 30 bp downstream the splice donor, including the splice site in both cases. Human/mouse (mm5) pairwise alignments were analyzed and conservation scores were calculated based on the average number of conserved base pairs in overlapping windows of length 8. Full alignments for introns flanking the following sets of exons were available for 104 EST-confirmed alternatively spliced NAGNAGs, 1834 constitutively spliced exons, 710 skipped exons, 1670 'Proximal' group from which 1249 are human/mouse conserved NAGNAG sites and 421 are not conserved at the NAGNAG site. Of the 188 'Distal' groups 91 are human/mouse conserved at the NAGNAG site and 97 are not. Average values were calculated for 8 bp windows for each group separately.

Word search

The presence of *cis*-regulatory elements in both exon and intron regions close to the NAGNAG acceptor sites, not including the NAGNAG site, was studied using the following procedure. Overlapping words of length 5 and 4 were counted in 100 bp pre-mRNA sequences in the upstream intron and downstream exon flanking the NAGNAG motif. An additional search was applied on a subset of these sequences which is conserved between human and mouse at 30 bp flanking the NAGNAG acceptor, respectively. Since the conserved regions were shorter, to avoid statistical bias which could arise from sparse data, in the latter set we counted only words of length 4. In cases in which the flanking exonic or intronic sequences were shorter than 100 or 30 bp, in the respective datasets,

we analyzed the available sequences. Important to note that in cases were the downstream exon or upstream intron were shorter than 100 bp, the analysis was not extended to the following exon or intron. Conservation values were extracted from human/mouse (mm5) pairwise alignments downloaded from the UCSC website. Words were counted separately in the flanking regions of each of the following groups: (i) 215 EST-confirmed NAGNAG 3′AS sites, (ii) a subset of 78 sequences from set 1 with the strong CAGCAG tandem acceptor, (iii) 5050 'Proximal' NAGNAGs, (iv) 584 'Distal' NAGNAGs, (v) 984 Skipped exons, and (vi) 102 461 constitutively spliced NAGNAG acceptors. All word counts were normalized to the size of the dataset. In addition each of the above datasets was randomly shuffled and words were counted as described above for the shuffled sets. From each group (1–5) the counts in the equivalent shuffled set were subtracted. The CS set was used as the background control group for calculating the $\log_2$ ratio (lr) for each word, as shown in the equation below.

$$\log_2 \frac{(N(i)-N(si))/T(i)}{(N(\mathrm{cs})-N(\mathrm{scs}))/T(\mathrm{cs})} \quad \mathbf{1}$$

where $N(i)$ is the number of counts in set i (1–5), $N(si)$ is the number of counts in set i after random shuffling, $T(i)$ is the total number of words in set I, $N(\mathrm{cs})$ is the number of counts in the CS set and $N(\mathrm{scs})$ is the number of counts in the CS group after random shuffling. $T(\mathrm{cs})$ is the total number of words in the CS set.

To ensure that we do not get significant results from low counts, in each set only words that were found in the top 1% were considered.

Words with lr >1 detected in the exon regions were screened against the matrices in ESEfinder (http://rulai.cshl.edu/tools/ESE/ESEmatrix.html) and the RESCUE-ESE database (http://genes.mit.edu/burgelab/rescue-ese/) to search for known ESE's and ESS's and their putative binding proteins. In addition a manual literature search was applied to examine the possible function of the overabundant words in the intronic regions.

RESULTS

Biased length distribution at alternative exon ends

We have analyzed the length distribution of the alternating exon ends, namely the distance between two alternative splice sites of the same exon, in a dataset of 460 3′AS and 207 5′AS events, which are EST-confirmed and conserved between human and mouse (4). This analysis revealed an extremely high occurrence of alternative 3′ splice sites separated by 3 nt, accounting for >50% (246/462) of all conserved alternative 3′ splice site events (Figure 1 and Supplementary Table 1). This set of 3′AS that differ by exactly 3 nt correspond precisely to the exons with NAGNAG 3′ splice sites which were recently found to be highly abundant in the human genome (9). Similar to alternatively skipped exons which tend to be divisible by three, the difference in length between alternative exons resulting from 3′ or 5′AS events show clear periodicity of three. As in ES the divisibility by three is most probably important for preserving the reading frame of the proteins. However, different than in the 3′AS sites, exact 3 nt indels at 5′ splice sites are very rare, found in only 3 out of 207 conserved events (<1.5%) (Figure 1). The infrequent occurrence of exact 3 nt indels at the 5′ splice site is expected based on the longer consensus of the 5′ splice site GU(G/A)AGU and the strong requirement for a G at position 1 and 5 (19). Nonetheless, the pattern of the 5′ splice site consensus GUXXGU is consistent with the peaks at 4, 6 and 9 continuing with a periodicity of 3 nt, we observed for the 5′AS events (Figure 1).

Unique distribution of GO terms

It has been shown previously that AS exons do not occur equally within all populations of genes but are rather restricted to certain gene families (20). To examine whether EST-confirmed NAGNAG 3′AS events are also found in specific gene families, we studied the distribution of the GO terms for the genes in which the NAGNAG 3′AS events were found. The distribution of the GO terms was compared with the distribution of the GO terms in two reference sets: (i) genes including ES events (AS) and (ii) genes including constitutive exons (CS), by applying the χ^2-test for goodness of fit. As can be seen in Table 1, among the 13 GO terms for which we found more than five entries in the NAGNAG 3′AS group, only the DNA binding were found to be significantly more frequent (after Bonferroni correction) than in both the AS and the CS group. Interestingly, the RNA-binding term was not significantly different between the NAGNAG 3′AS group and the AS group, while highly significant compared with the CS group. These results are motivating as it is well known that RNA-binding proteins are involved in AS, both as targets and

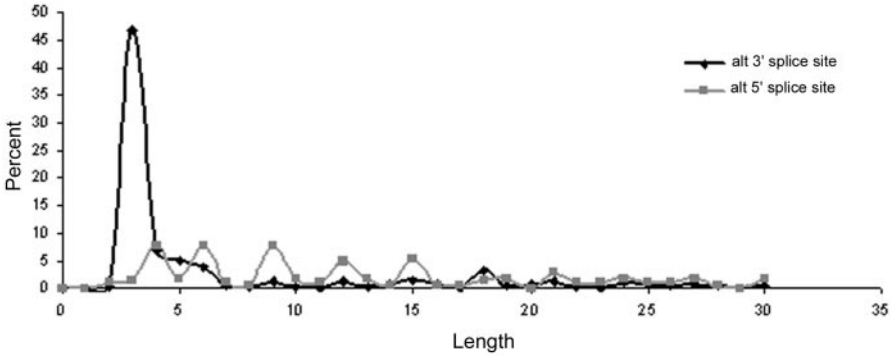

Figure 1. Distribution of indel (insertion/deletion) length in alternative 3′ (black, diamonds) and 5′ (gray, squares) splice sites revealing the high occurrence of exact 3 nt indels at alternative 3′ splice sites.

Table 1. Comparison of the GO distribution between our dataset of genes with NAGNAG 3′AS and alternatively spliced and constitutively spliced genes

Molecular function	χ^2-NAG3′AS/AS	*P*-value	χ^2-NAG3′AS/CS	*P*-value
DNA binding	17.1897	$P \sim 0.000$	62.5636	$P \sim 0.000$
Transcription corepressor activity	7.2061	$0.01 > P > 0.005$	22.1898	$P \sim 0.000$
RNA binding	4.5283	$0.05 > P > 0.025$	46.3431	$P \sim 0.000$
Transcription factor binding	1.3925	$P > 0.1$	7.1942	$0.01 > P > 0.005$
Cytoskeletal protein binding	1.4861	$P > 0.1$	0.0013	$P > 0.1$
Hydrolase activity	0.6673	$P > 0.1$	0.434	$P > 0.1$
Protein kinase activity	0.3231	$P > 0.1$	1.5253	$P > 0.1$
Cation binding	0.2559	$P > 0.1$	0.0525	$P > 0.1$
Purine nucleotide binding	0.2034	$P > 0.1$	0.1869	$P > 0.1$
Ligase activity	0.1995	$P > 0.1$	0.036	$P > 0.1$
Metal ion binding	0.1522	$P > 0.1$	0.0287	$P > 0.1$
Peptidase activity	0.0113	$P > 0.1$	0.7435	$P > 0.1$
Transferase activity	0.0014	$P > 0.1$	0.0004	$P > 0.1$

GO entries for the 'Molecular function' sub-ontology at node 4 were computed. The χ^2-tests were carried out between a set of genes containing EST-confirmed alternatively spliced NAGNAGs against genes with skipped exons (NAGNAG/ES) and against constitutively spliced genes (NAGNAG/CS). *P*-values were calculated applying the Bonferroni correction for multiple testing.

regulators, and in many cases RNA-binding proteins autoregulate their own splicing (21). These results are also in accordance with the finding that proteins belonging to the RNA-recognition group in Pfam are enriched for transcripts with NAGNAG 3′ splice sites (9). Overall, the GO analyses suggest that genes included in the NAGNAG 3′AS group share some functional similarity with genes that undergo AS. Nevertheless, the exceptionally high occurrence of DNA-binding proteins, and other proteins with transcription activity, in the NAGNAG 3′AS group imply that the insertion or deletion of a single amino acid may play a role in fine tuning the binding properties of the different DNA-binding protein isoforms. Direct evidence for an effect of AS of a NAGNAG motif in the PAX-3 paired domain on DNA binding affinity has been shown previously (22).

Nucleotide preferences at the NAGNAG motif

Analysis of the NAGNAG consensus sequence in the 215 EST-confirmed alternatively spliced NAGNAG group (excluding events outside of translated regions) showed that the N at the NAG site proximal to the branch point [E acceptor, as defined by Hiller *et al*.(9)] is preferred to be C>T>A>>G and the N at the NAG site distal to the branch point [I acceptor, as defined by Hiller *et al*. (9)] is C>A>T>>G (Figure 2). This is in agreement with the relative strength of the splicing acceptor sites (23) which are CAG>TAG>AAG. The tri-nucleotide GAG, which very rarely appears as a splice acceptor (23) was found to be almost absent at both sites. Another interesting phenomenon we have found in the NAGNAG motifs in the alternatively spliced group is the unexpected preference for A over T at the N of the proximal NAG site (Figure 2C). This may be due to the fact that in the alternatively spliced group the proximal NAG site is occasionally found within the coding sequence. Thus TAG, which (depending on the reading frame) could insert a termination codon, seems to be avoided at this position of the NAGNAG 3′AS group. Putting together the splice site strength and avoidance of stop codons in the coding sequence, it is understandable that the most common tandem acceptor in the alternative spliced group is CAGCAG (78 cases) followed by TAGCAG (44 cases), CAGAAG (22 cases), AAGCAG (20 cases) and TAGAAG (16 cases).

Consistent with the results of Hiller *et al*. (9), in the constitutively spliced exons which are conserved between human and mouse we found many exons with the NAGNAG motif at the 3′ splice site (5634). These may be a mixture of constitutively spliced NAGNAG acceptors in which the same 3′ splice site is always selected, and putative alternatively spliced NAGNAG sites which are not supported by EST data. Of them 5050 are cases in which the EST data confirms splicing at the NAG site proximal to the branch point (Figure 2A); this group was defined the 'Proximal' group (Supplementary Table 2). The other 584 were cases in which the EST data confirms splicing at the NAG site distal from the branch point (Figure 2A), which we defined as the 'Distal' group (Supplementary Table 3). In an attempt to understand whether AS at the NAGNAG acceptor sites is regulated, we analyzed each of these groups (alternative spliced, 'Proximal', 'Distal') separately. In comparison with the 'N preferences' found at the confirmed alternative spliced NAGNAG sites, examination of the motif in the 'Proximal' group revealed a striking difference between the two half sites, G being the most frequent nucleotide in the distal NAG site (Figure 2C and Supplementary Table 8). However, the nucleotide preferences at the proximal NAG site were very similar in both groups. The preferences observed at the 'Proximal' NAGNAG are consistent with the idea that the proximal NAG site is the actual splice acceptor while the distal NAG site may be in most of the cases part of the coding sequence. This is confirmed by the high occurrence of GAG at the distal site, which is practically not observed as a splice site. Moreover, the high occurrence of GAG at the distal site is also in accordance with the known preference for G at the first position of mammalian exons (19). On the contrary, at the 'Distal' group we found that the N at the proximal NAG site is G>C>A>T and the N at the distal NAG site is C>>T>A>>G (Figure 2B and C and Supplementary Table 8). Interestingly, at the 'Distal' group there is a strong bias for C at the distal NAG site. This could be a result of the general tendency of the splicing machinery to choose a proximal splice acceptor when two potential acceptor sites are found in close proximity (12,13,24,25). Hence in order to force the splicing machinery to the distal NAG splice site it is expected to find a strong acceptor at that site.

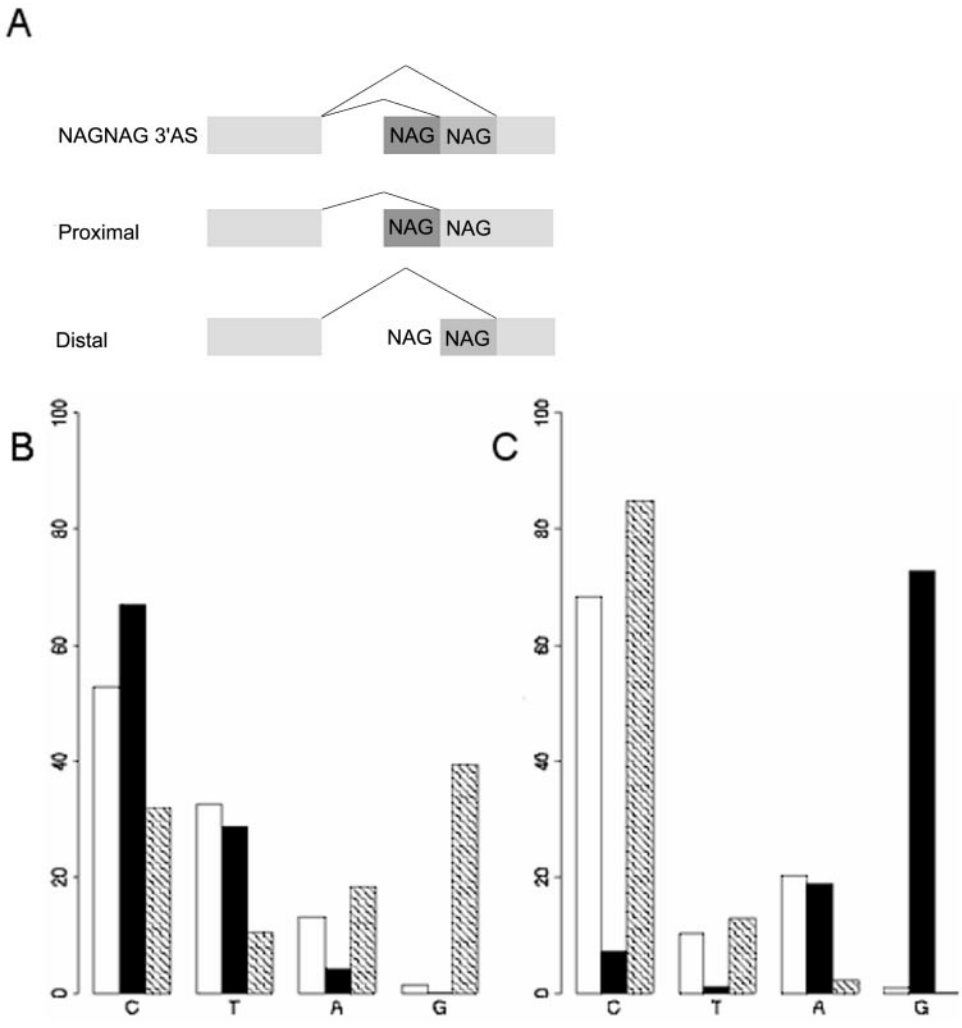

Figure 2. (**A**) Schematic diagram of the three groups of NAGNAG splice sites: NAGNAG 3′AS, proximal and distal. (**B**) Observed nucleotide frequency for the N of the proximal NAG site and (**C**) the N of the distal NAG site in the three different groups: NAGNAG 3′AS (white), 'Proximal' (black) and 'Distal' (striped lines).

Sequence conservation at the 3′ splice sites and flanking regions

Applying comparative analysis to the NAGNAG sequence itself between the human and mouse genomes, we found that both AGs at the tandem acceptors are 100% conserved in all of the 215 examined cases of EST-confirmed NAGNAG 3′AS sites (Figure 3A), while in the 'Proximal' and 'Distal' NAGNAG groups only the AG that is expected to serve as a splice acceptor was found to be 100% conserved (Figure 3B and C). These results reinforced that in the alternatively spliced group both NAG sites are true splice acceptors, while in the 'Proximal' and 'Distal' groups only one of the NAG splice sites seems to be a real acceptor.

Recently it has been observed that skipped exons are characterized by high sequence conservation at the flanking intronic sequences (3–8). The high conservation of intronic regions around alternative exons presumably indicates the presence of regulatory elements, responsible for regulating the AS of these exons. Thus, if NAGNAG 3′AS events are true cases of AS we expect to detect a high conservation at the 5′ flanking intron adjacent to the NAGNAG site. Figure 4A shows the average of human–mouse sequence conservation for windows of length 8 surrounding each position of the 30 bp flanking the 5′ introns. As illustrated, the conservation of the introns adjacent to the NAGNAG 3′AS sites is significantly higher than in constitutively spliced exons, albeit lower than the average conservation of introns flanking alternatively skipped exons. Interestingly, also the length of the conserved intronic regions around NAGNAG 3′AS sites is shorter (∼30bp) than the conserved region flanking skipped exons (100 bp) observed in previous works (4,5). As can be noticed from the figure, at the 3′ intron far from the NAGNAG 3′ splice sites the conservation level is similar to constitutively spliced exons. This latter result is reasonable since regulation of AS at the NAGNAG acceptor site might not involve the downstream 3′ intron, whereas ES most likely will involve both 5′ and 3′ introns. We also examined human–mouse conservation of the introns at the 'Proximal' and 'Distal' NAGNAG groups, and separated each into two sequence sets, those whose AG at the NAGNAG sites are conserved between human and mouse and those whose AGs are not conserved (Figure 4B). Interestingly,

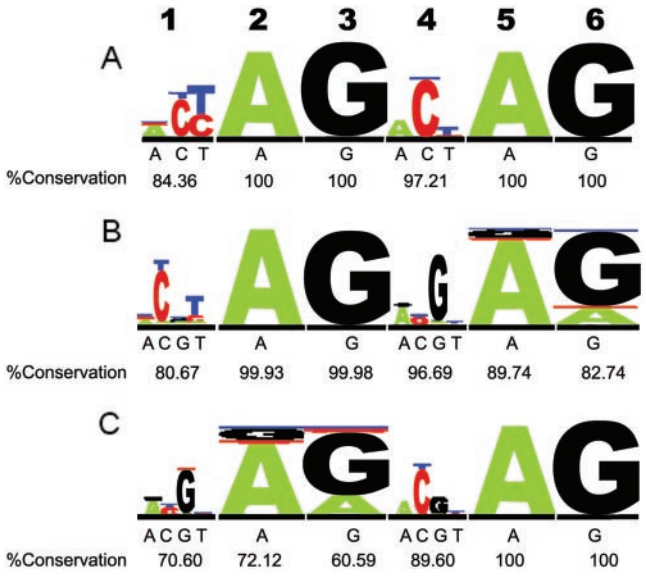

Figure 3. Human/mouse evolutionary conservation of NAGNAG acceptor sites. Percent of conservation was calculated for human/mouse pairwise alignments (hg17/mm5) derived from the UCSC site http://hgdownload.cse.ucsc.edu/downloads.html#human in three different datasets. (**A**) EST-confirmed NAGNAG 3′AS sites, (**B**) 'Proximal' group and (**C**) 'Distal' group. The numbers 1–6 indicate the nucleotide position at the NAGNAG splice site. The letters below each position indicate the base found in human. The heights of the letters represent the relative frequency of each nucleotide in a given position. The values below indicate the overall percent nucleotide conservation at that position. For positions (2, 5 and 3, 6), which are occupied in human by a single type of nucleotide, the percent conservation represents the conservation of the nucleotides (A or G) shown above. Graphic representations were carried out with the program Logos http://weblogo.berkeley.edu/logo.cgi.

in the 'Proximal' group we observed low intronic conservation in both sub-groups while the subset of the 'Distal' group that are conserved at the NAGNAG motif have an intronic conservation level that resembles that of the NAGNAG 3′AS sites. This suggests the existence of a very strict regulation upon the selection of the distal NAG splice site even if it is constitutively spliced. Nonetheless, this could also indicate that the subsets of the 'Distal' group are true NAGNAG 3′AS sites, though not confirmed by EST data. Since 'Distal' acceptors are rarely chosen by the splicing machinery, it is likely that the proximal NAG has to be weaker for the splicing machinery to identify the distal NAG instead (as seen in Figure 2B and C). In the case of NAGNAG 3′AS sites, it may bias the choice towards the distal NAG site, allowing the production of major and minor splice variants.

Overabundance of *cis*-acting elements flanking alternatively spliced 3′ splice sites

To further study the regulation at alternatively spliced NAGNAG sites, we searched for presence of *cis*-acting elements in both the exon and intron regions close to the NAGNAG acceptor sites in pre-mRNA sequences. For that we first counted overlapping words of length 4 and 5 and analyzed the words which are overrepresented in the flanking exons and introns in the alternative spliced NAGNAG group relatively to the equivalent regions in the constitutively spliced group. In addition we performed the search on the following datasets: 'Proximal' NAGNAGs, 'Distal' NAGNAGs and Skipped exons (Supplementary Tables 4–7). As a control we also counted words in the same set of introns and exons flanking the NAGNAG sites after random shuffling the sequences. The word counts in the shuffled sequences were subtracted from the word counts in the equivalent sets. This was done to ensure that high abundance of the word is not related to the base content of that region. As it is expected to find ESE's within both AS and CS exons (26) words which were found 2-fold higher than in the constitutive set (lr > 1), after subtraction of words from the shuffled sets (Materials and Methods), were considered overabundant. We performed the search in two fashions: a standard word count at both the 100 bp exonic and intronic flanking sequences. In addition we applied a more stringent search, counting only words that are conserved between human and mouse at the 30 bp exonic and intronic flanking sequences. The reason for selecting a shorter region for the stringent test was the limited amount of data of high quality alignments of intronic regions. To be consistent, also the dataset of conserved exonic sequences was limited to 30 bp. Nevertheless, to assure that we do not get significant results from very low counts in the conserved set, in this limited set we counted only words of length 4. Also important to note is that the tandem acceptor sequences (NAGNAG) were excluded in all cases.

The results of the word count are shown in Table 2. Only overrepresented words which were ranked in the top 1% in the NAGNAG 3′AS set are shown. In the table it is also specified whether the words were found generally in the HAGHAG (H = A or C or T) or in the subset of only strong CAGCAG acceptors. Interestingly, the overabundant words were not detected in either the 'Proximal' NAGNAGs or the 'Distal' NAGNAGs. Moreover, all of the overrepresented words, excluding TTTTA, were not overabundant in the equivalent regions surrounding skipped exons. As can be easily noticed, most of the overabundant words that were found in the exon regions, both in the larger dataset flanking acceptors with the HAGHAG consensus site and the subset including only the strong CAGCAG acceptor site, contain the core CAGC. Two additional words which do not include the core CAGC were found only in the larger set flanking the HAGHAG consensus. Overall we found more overabundant words in the limited set with the strong CAGCAG site. This may be due to a couple of reasons: (i) CAGCAG tandem acceptors, having two equally strong acceptors may require more extensive regulation via *cis*-regulatory elements, (ii) the CAGCAG could represent a distinct sub-group in which regulatory elements are easier to be detected. This is supported by a recent study by Wang *et al*. (26) that there is no simple correlation between the weakness of the splice site and the high occurrence of ESE's surrounding it. To ensure that the overabundance of CAGC does not result from a GC-content bias in our dataset, in addition to subtracting the words from the shuffled sequences we also analyzed the GC content in our sequences compared with a set of 107 000 constitutively spliced exons and found a similar GC content (50%) in both sets (data not shown). In the intronic regions we found two types of overrepresented words: T rich (TATTT, TTTTA) and G rich (GTGGG). Interestingly, the T rich sequences, which could be related to the polypyrimidine track, were highly frequent both in the NAGNAG

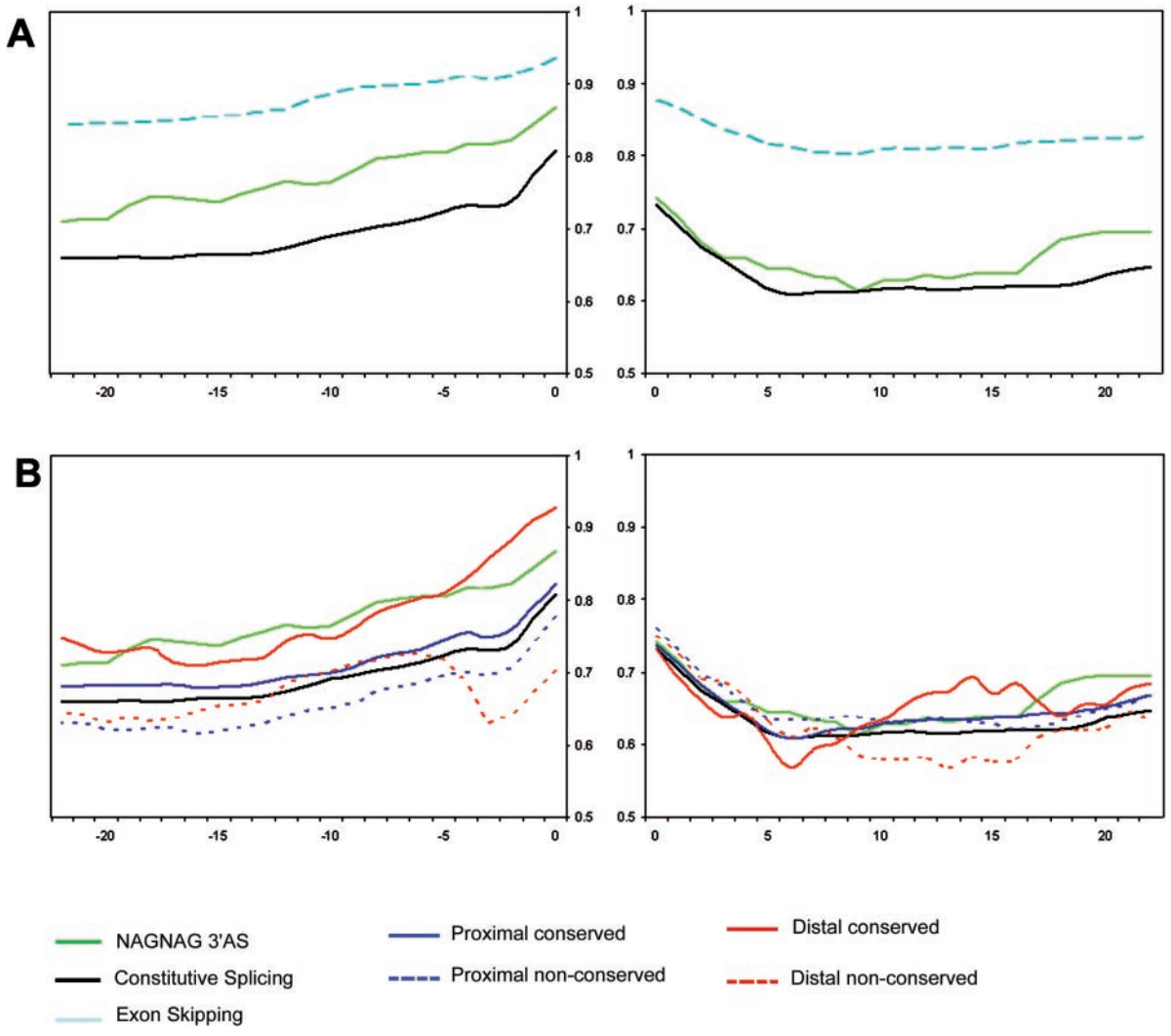

Figure 4. Human/mouse evolutionary conservation at the intronic flanking regions of exons with a NAGNAG sequence at the acceptor site. (**A**) Average conservation scores for 8 nt overlapping windows spanning 30 nt upstream the splicing acceptor and downstream the splicing donor at NAGNAG 3′AS sites (green), Skipped Exons (turquoise) and Constitutive spliced exons (black). (**B**) Conservation scores for conserved 'Proximal' and 'Distal' groups (plain, blue and red, respectively) and non-conserved 'Proximal' and 'Distal' groups (dashed, blue and red, respectively). Numbers in the *X*-axis represent the position of the first nucleotide of the window relative to the splice site. Values for each point i represent the average conservation of positions $i - (i + 7)$.

3′AS and the constitutive set. However, since they were also equally frequent in the shuffled sets causing the background to be close to zero (Equation 1), they were mistakably overestimated and detected in our analysis as highly overabundant.

Examination of the overrepresented words showed that the exonic words, sharing the consensus CAGC core, are related to several ESEs sites such as SRp55, SRp40, SC35 and SF2/ASF binding sites. Comparison of these words to the candidate binding sites from the RESCUE-ESE database (27) and the matrices from ESEfinder (28) confirmed that GCAGC; ACAGC, including the core CAGC correspond to the consensus SRp55 binding site; CAGCC, which also includes the core CAGC matches the SF2/AF consensus binding site; and some of these words are also predicted as weak SC35 and SRp40 binding sites. The two words AGGAA and GATGA which were found only in the larger set with the HAGHAG consensus were predicted by the RESCUE-ESE program as ESE's, though their binding protein is yet unknown. We also examined the intronic words and compared them to known consensus sites (29). We found that GTGGG corresponds to the hnRNP H splicing silencer, while TATTT and TTTTA are not annotated as intronic regulatory sequences.

DISCUSSION

NAGNAG 3′AS sites are the major form of alternative 3′ splice sites, accounting for 50% of the cases. In our study we uncovered 215 EST-confirmed events within protein coding regions in the human genome, although it is possible that additional cases remain to be detected. Our data strongly imply that the NAGNAG 3′AS sites undergo AS in a regulated manner. Also, we suggest that the presence of the NAGNAG sequence at the 3′ splice acceptor site is not sufficient to produce two splice variants, as many NAGNAG acceptors with a single EST-confirmed isoform lack regulatory features.

A recent study on aberrant 3′ splice sites, revealed that point mutations occurring in YAG acceptors generally shift the

Table 2. Overabundant words surrounding NAGNAG 3′AS

Word	Location	Tandem 3′ splice site	Occurrence	lr	Predicted binding protein
CAGC	Exon	CAGCAG	19	1.7139	SF2/AF;SRp55;SC35-weak;SRp40-weak
GAGC	Exon	CAGCAG	19	2.6846	SF2/AF-weak
CAGCC	Exon	CAGCAG	25	1.8051	SF2/AF;SRp40-weak
GCAGC	Exon	HAGHAG	70	1.2121	SRp55
GCAGC	Exon	CAGCAG	35	2.0901	SRp55
ACAGC	Exon	CAGCAG	21	2.4655	SRp55;SRp40-weak
AGCAG	Exon	CAGCAG	25	1.3757	SRp55-weak;SRp40-weak
CAGCA	Exon	CAGCAG	29	1.6073	SC35-weak;SRp40-weak
CCAGC	Exon	CAGCAG	25	1.7847	SC35-weak;SRp40-weak
CCCAG	Exon	CAGCAG	23	1.2639	SC35-weak
AGGAA	Exon	HAGHAG	56	1.1704	NN
GATGA	Exon	HAGHAG	55	1.1263	NN
CAGCT	Exon	CAGCAG	21	1.28	NF
TATTT	Intron	HAGHAG	76	7.9365	NF
TTTTA	Intron	HAGHAG	94	4.7052	NF
GTGGG	Intron	CAGCAG	23	1.8705	hnRNP H

Four and five-letter words were counted in 100 nt exonic and intronic regions surrounding the NAGNAG acceptor sites while four-letter words only were counted at human–mouse conserved regions within 30 nt exclusively. The table shows the sequence of the overabundant words, the location (exon or intron), the sequence of acceptor site (CAGCAG or HAGHAG, H = A,C,T), the occurrence of the word within the dataset, the lr relatively to a control set of constitutively spliced exons and the proteins predicted to bind the candidate *cis*-regulatory sequences. NN denote cases in which we found hits in RESCUE-ESE but their binding proteins are not known, NF indicate words that had no hits either in ESEfinder or in RESCUE-ESE.

splice site to cryptic AGs that are found within <21 nt surrounding the authentic acceptor, and these are generally CAG cryptic sites (30). This strengthens the idea that in order to obtain AS at the 3′ splice site, it is not enough to find two adjacent NAG sites. Therefore, a cryptic site has the potential to become functional only under certain circumstances (e.g. mutations in the original splice site). In addition, the overabundance of CAG cryptic sites that replace the original mutated YAG sites is generally in agreement with our results, suggesting that when the splicing machinery does not identify the authentic splice site it prefers to substitute it with a strong CAG site. This is exemplified in the group of the 'Distal' NAGNAG acceptors in which C is usually preferred at the distal NAG site. Moreover, in the 'Distal' NAGNAG group we also found evidence for the involvement of *cis* factors, as indicated by the relative high sequence conservation at the intronic regions, ~30 bp immediately flanking these sites.

Overall we found that alternatively spliced NAGNAG acceptors are distinguishable from constitutively spliced NAGNAG sites based on the following parameters: (i) absence of the nucleotide G at the N position in the NAGNAG motif overall; (ii) high evolutionary conservation at both AG sites; (iii) high conservation at the introns flanking the NAGNAG acceptors. These values are close to the conservation values found recently around skipped exons (5) but are restricted to 30 bp upstream the splice site; (iv) overabundance of *cis*-acting elements surrounding the alternatively spliced NAGNAGs. In addition, it is possible that part of the exons with NAGNAG sequences at the 3′ acceptor site that are not confirmed by EST data to have alternative spliced isoforms, are true NAGNAG 3′AS sites. We suggest that these cases can be tested individually by studying the surrounding features such as intronic conservation, the presence of splicing regulatory protein binding sites and the relative strength of the NAG site.

It is generally assumed that EST data are reliable only when an AS event is evolutionarily conserved, owing to the high rate of splicing aberrations in cancer cell lines that serve for the construction of EST libraries (4). Accordingly, we were able to confirm AS only for 215 events which account for ~1% of human genes, although the range of NAGNAG 3′AS sites may be larger than this. In their study Hiller *et al*. (9) proposed that at least 5% of NAGNAG sites are functional. Additional cases to the ones we have observed may come from different sources. Among these could be AS events that are not conserved between human and mouse. In a recent study (31) it has been suggested that species-specific splicing events, i.e. events that are not conserved between human and mouse, could play critical roles in many process in the cell. Moreover, additional cases could come from AS events in which one of the isoforms is relatively infrequent and thus possibly undetectable in EST experiments. These latter cases may relate to the previously mentioned subset of 'Distal' NAGNAGs. In the future it would be important to experimentally validate NAGNAG 3′AS sites at both the mRNA and the protein level and to focus attention on the effect of the insertion and deletions of exact 3 nt on the protein structure and function.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENTS

We would like to thank Manny Ares, Hanah Margalit, Dan Cassel and Ora Schueler-Furman for valuable comments on the manuscript. Thanks to Omer Haber for computer assistance. This work was supported by the Eunice Geller Cancer Research Fund. Funding to pay the Open Access publication charges for this article was provided by the Israel Science Foundation Grant number 923/05.

Conflict of interest statement. None declared.

REFERENCES

1. Mercatante,D. and Kole,R. (2000) Modification of alternative splicing pathways as a potential approach to chemotherapy. *Pharmacol. Ther*., **85**, 237–243.

2. Zheng,Z.M. (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J. Biomed. Sci.*, **11**, 278–294.
3. 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.
4. 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.
5. Sorek,R. and Ast,G. (2003) Intronic sequences flanking alternatively spliced exons are conserved between human and mouse. *Genome Res.*, **13**, 1631–1637.
6. Rahman,L., Bliskovski,V., Kaye,F.J. and Zajac-Kaye,M. (2004) Evolutionary conservation of a 2-kb intronic sequence flanking a tissue-specific alternative exon in the PTBP2 gene. *Genomics*, **83**, 76–84.
7. Kaufmann,D., Kenner,O., Nurnberg,P., Vogel,W. and Bartelt,B. (2004) In NF1, CFTR, PER3, CARS and SYT7, alternatively included exons show higher conservation of surrounding intron sequences than constitutive exons. *Eur. J. Hum. Genet.*, **12**, 139–149.
8. Dror,G., Sorek,R. and Shamir,R. (2005) Accurate identification of alternatively spliced exons using support vector machine. *Bioinformatics*, **21**, 897–901.
9. Hiller,M., Huse,K., Szafranski,K., Jahn,N., Hampe,J., Schreiber,S., Backofen,R. and Platzer,M. (2004) Widespread occurrence of alternative splicing at NAGNAG acceptors contributes to proteome plasticity. *Nature Genet.*, **36**, 1255–1257.
10. Tadokoro,K., Yamazaki-Inoue,M., Tachibana,M., Fujishiro,M., Nagao,K., Toyoda,M., Ozaki,M., Ono,M., Miki,N., Miyashita,T. *et al.* (2005) Frequent occurrence of protein isoforms with or without a single amino acid residue by subtle alternative splicing: the case of Gln in DRPLA affects subcellular localization of the products. *J. Hum. Genet.*, **50**, 382–394.
11. Xing,Y. and Lee,C.J. (2005) Protein modularity of alternatively spliced exons is associated with tissue-specific regulation of alternative splicing. *PLoS Genet.*, **1**, e34.
12. Smith,C.W., Chu,T.T. and Nadal-Ginard,B. (1993) Scanning and competition between AGs are involved in 3′ splice site selection in mammalian introns. *Mol. Cell. Biol.*, **13**, 4939–4952.
13. Smith,C.W., Porro,E.B., Patton,J.G. and Nadal-Ginard,B. (1989) Scanning from an independently specified branch point defines the 3′ splice site of mammalian introns. *Nature*, **342**, 243–247.
14. Chua,K. and Reed,R. (2001) An upstream AG determines whether a downstream AG is selected during catalytic step II of splicing. *Mol. Cell. Biol.*, **21**, 1509–1514.
15. Lev-Maor,G., Sorek,R., Shomron,N. and Ast,G. (2003) The birth of an alternatively spliced exon: 3′ splice-site selection in Alu exons. *Science*, **300**, 1288–1291.
16. Tran,Q. and Roesser,J.R. (2003) SRp55 is a regulator of calcitonin/CGRP alternative RNA splicing. *Biochemistry*, **42**, 951–957.
17. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
18. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
19. Zhang,M.Q. (1998) Statistical features of human exons and their flanking regions. *Hum. Mol. Genet.*, **7**, 919–932.
20. Liu,S. and Altman,R.B. (2003) Large scale study of protein domain distribution in the context of alternative splicing. *Nucleic Acids Res.*, **31**, 4828–4835.
21. Bell,L.R., Horabin,J.I., Schedl,P. and Cline,T.W. (1991) Positive autoregulation of sex-lethal by alternative splicing maintains the female determined state in *Drosophila*. *Cell*, **65**, 229–239.
22. Vogan,K.J., Underhill,D.A. and Gros,P. (1996) An alternative splicing event in the Pax-3 paired domain identifies the linker region as a key determinant of paired domain DNA-binding activity. *Mol. Cell. Biol.*, **16**, 6677–6686.
23. Aebi,M., Hornig,H., Padgett,R.A., Reiser,J. and Weissmann,C. (1986) Sequence requirements for splicing of higher eukaryotic nuclear pre-mRNA. *Cell*, **47**, 555–565.
24. Dix,I., Russell,C.S., O'Keefe,R.T., Newman,A.J. and Beggs,J.D. (1998) Protein–RNA interactions in the U5 snRNP of *Saccharomyces cerevisiae*. *RNA*, **4**, 1675–1686.
25. Liu,Z.R., Laggerbauer,B., Luhrmann,R. and Smith,C.W. (1997) Crosslinking of the U5 snRNP-specific 116-kDa protein to RNA hairpins that block step 2 of splicing. *RNA*, **3**, 1207–1219.
26. Wang,J., Smith,P.J., Krainer,A.R. and Zhang,M.Q. (2005) Distribution of SR protein exonic splicing enhancer motifs in human protein-coding genes. *Nucleic Acids Res.*, **33**, 5053–5062.
27. Fairbrother,W.G., Yeo,G.W., Yeh,R., Goldstein,P., Mawson,M., Sharp,P.A. and Burge,C.B. (2004) RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons. *Nucleic Acids Res.*, **32**, W187–W190.
28. Cartegni,L., Wang,J., Zhu,Z., Zhang,M.Q. and Krainer,A.R. (2003) ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Res.*, **31**, 3568–3571.
29. Ladd,A.N. and Cooper,T.A. (2002) Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biol.*, **3**, reviews0008.
30. Chamary,J.V. and Hurst,L.D. (2005) Biased codon usage near intron–exon junctions: selection on splicing enhancers, splice-site recognition or something else? *Trends Genet.*, **21**, 256–259.
31. Pan,Q., Bakowski,M.A., Morris,Q., Zhang,W., Frey,B.J., Hughes,T.R. and Blencowe,B.J. (2005) Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.*, **21**, 73–77.