1974–1983 *Nucleic Acids Research, 2003, Vol. 31, No. 7*
DOI: 10.1093/nar/gkg279

Conserved sequence elements associated with exon skipping

Elana Miriami[1,2,*], Hanah Margalit[2] and Ruth Sperling[1]

[1]Department of Genetics, The Alexander Silberman Institute of Life Sciences and [2]Department of Molecular Genetics and Biotechnology, The Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 91904, Israel

Received October 26, 2002; Revised and Accepted January 28, 2003

ABSTRACT

One of the major forms of alternative splicing, which generates multiple mRNA isoforms differing in the precise combinations of their exon sequences, is exon skipping. While in constitutive splicing all exons are included, in the skipped pattern(s) one or more exons are skipped. The regulation of this process is still not well understood; so far, *cis*-regulatory elements (such as exonic splicing enhancers) were identified in individual cases. We therefore set to investigate the possibility that exon skipping is controlled by sequences in the adjacent introns. We employed a computer analysis on 54 sequences documented as undergoing exon skipping, and identified two motifs both in the upstream and downstream introns of the skipped exons. One motif is highly enriched in pyrimidines (mostly C residues), and the other motif is highly enriched in purines (mostly G residues). The two motifs differ from the known *cis*-elements present at the 5′ and 3′ splice site. Interestingly, the two motifs are complementary, and their relative positional order is conserved in the flanking introns. These suggest that base pairing interactions can underlie a mechanism that involves secondary structure to regulate exon skipping. Remarkably, the two motifs are conserved in mouse orthologous genes that undergo exon skipping.

INTRODUCTION

The reported number of genes in the human genome is about 32 000, exceeding that of *Drosophila melanogaster* (14 000 genes) or *Caenorhabditis elegans* (19 000 genes) (1), but still not large enough to account for the diversity of human proteins. Indeed it was suggested that human complexity is to a large extent generated by alternative splicing, and up to 35–60% of human genes have been estimated to be alternatively spliced (2–5), generating distinct protein isoforms (6,7). Of the alternatively spliced forms, exon skipping is the most common form, where, during splicing, skipping of specific exon(s) occurs in different tissues or at different developmental stages (4,8).

The specific isoforms of the skipped mRNAs, discussed in this study, are involved in a wide range of functions. For example, the Survivin gene, with both skipped (lacking exon 3, survivin-ΔEx3) and unskipped forms, functions as an inhibitor of apoptosis (9). However, the full-length product inhibits apoptosis in the cytoplasm, whereas its variant, survivin-ΔEx3, functions in the nucleus (10). Another example is provided by the Clathrin light chain (CLa), which functions as the main structural component of the coated vesicles involved in intracellular transport. The tissue-specific variant, the unskipped form, is expressed specifically in the brain and reflects the special demands of neurons, such as axonal transport (11). The Clk kinases function in splicing, generating catalytically active (Clk) and inactive (Clk^T) isoforms by exon skipping (12).

Several cancers and inherited diseases in humans are associated with mutations that cause unnatural exon skipping (13; for review see 7). Usually, the mutations affect the splice sites; for example, a point mutation at the 5′ splice site of exon 7 of the Wilm's tumor suppressor gene causes unnatural skipping of exon 7 and generates a truncated protein that is associated with Wilm's tumor (14). Mutations located outside of the traditional splice sites, either internally within the exon or in the flanking intron sequences have also been reported to be associated with exon skipping and diseases. For example, mutations in exon 18 of the breast cancer BRCA1 gene cause inappropriate skipping of the entire constitutive exon (15). In the latter case, it was found that these mutations disrupt *cis*-acting regulatory elements that function as enhancers of splicing (for review, see 16). Serious efforts have been carried out to identify these *cis*-regulatory elements within the exons (exonic splicing enhancers) and the *trans* factors that interact with them, such as SR (17,18) and hnRNP (19) proteins. Less is known, however, about regulatory elements that reside in the introns. One interesting example of intron regulatory element disruption is found in the autosomal dominant form of the Isolated Growth Hormone Deficiency (IGHD-II), which is caused by mutations in intron 3 of the growth hormone gene. These mutations cause the increase in exon skipping of exon 3, accompanied by a change in the 5′ end of exon 3. This leads to a truncation in that exon and prevents secretion of the growth hormone. Interestingly, some of these IGHD-II mutations

*To whom correspondence should be addressed. Tel: +972 2 658 6034; Fax: +972 2 658 6975; Email: elanam@gene.md.huji.ac.il

perturbed an intronic splicing enhancer that is located in the downstream intron (intron 3) (20).

To explore the possibility that exon skipping is regulated by specific sequences in the flanking introns, a database of gene sequences that are known to undergo exon skipping was analyzed. In particular, we focused on events in which only one exon at a time was skipped, without additional changes in the flanking exons. Application of an algorithm that identifies common motifs in unaligned sequences has revealed two motifs, each in one of the introns flanking the skipped exon. One motif is highly enriched in pyrimidines (mostly C residues), and the second motif is highly enriched in purines (mostly G residues). The relative order of the elements in each of the flanking introns is conserved, and they show base pairing potential, suggesting that exon skipping may be regulated by a mechanism that involves base pairing interactions.

MATERIALS AND METHODS

Database of genes that undergo exon skipping

The data were derived from ISIS, the Intron Sequence Information System (http://www.introns.com/) (2). These data were filtered, and various sequences were excluded based on several criteria. (i) Cases where exon skipping was accompanied by changes in the splice sites of the flanking exons, thereby extending or truncating the exons. (ii) Cases in which the data were derived from BAC, PAC or cosmid, and the proteins were described as hypothetical, putative or pseudo. (iii) Cases where two or more exons were skipped in one event. (iv) Cases in which the information in GenBank contradicted the information in ISIS. After this filtering the dataset included 54 genes that exhibited exon-skipping events. These gene transcripts had only one exon skipped, and there were no additional changes in the flanking splice sites.

Interestingly, the genes that have been included show a wide range of tissue and developmental expression patterns, based on the literature and on the information in SpliceNest [http://splicenest.molgen.mpg.de/; (21)] and http://www.ncbi.nlm.nih.gov/SAGE/. Sequences corresponding to each of the skipped exons were used to query the GenBank database for the corresponding genomic sequences, from which the flanking intron sequences could be retrieved. Additionally, when the mRNA sequence information of the skipped exons was available, we used it for further validation of the intron boundaries. Each exon in the database was examined to confirm that it was appropriately flanked by consensus 5′ and 3′ splice sites that followed the GT/AG rule for sequences at the intron boundaries (22). These skipped exons, together with the entire upstream and the entire downstream intron sequences, were used for the analysis.

In order to evaluate the quality of the identified motifs three datasets were used. (i) As a negative control we used a dataset of human genes with annotated exon/intron boundaries (http://www.fruitfly.org/seq_tools/datasets/Human/coding_data/ Reese 1999). From these data we excluded genes that undergo alternative splicing (2,4,5,8). At the end of these procedures the control dataset contained 259 genes with a total of 1069 introns and 1328 exons of which only 810 exons had flanking introns. In these data no exon skipping was documented. However, given that the EST data is not complete, it is possible that this control dataset includes a few exon-skipping events that have not yet been documented. (ii) As a positive control we used 25 genes derived from altExtron data (8), which were annotated there as undergoing exon skipping and were not included in the training set used to identify the motifs. These genes followed the same criteria used to build the training set. (iii) An additional control set contained introns that interact with the Polypyrimidine Tract-Binding protein (PTB) (23); those interactions were confirmed by biochemical binding assays as well as cross-linking. PTB, also known as hnRNP-I, functions as a negative regulator of alternative splicing by a mechanism that requires its binding to intronic repressor sequences (23; for review see 24). This dataset was checked in order to exclude the possibility that the identified motifs overlapped with the PTB binding motif.

Table 1. Nucleotide composition (in percentage)

	Introns Control (1069)[a]	Upstream[b] (54)[a]	Downstream[c] (54)[a]	Exons Control (1328)[a]	Skipped exons (54)[a]
A	25.12	25.38	25.42	22.34	23.97
G	24.15	22.44	22.49	28.49	27.51
C	23.58	22.93	22.76	29.32	26.61
T	27.15	29.25	29.33	19.85	21.91

[a]Number of sequences.
[b]Introns upstream of the skipped exons.
[c]Introns downstream of the skipped exons.

Identification of common motifs

Two algorithms that search for common motifs in unaligned sequences were employed, MEME (GCG package) (25) and Gibbs sampling [http://argon.cshl.org/ioschikz/gibbsDNA/; (26)]. These algorithms were run separately on each of the sequence groups under investigation, the skipped exons, the entire upstream introns and the entire downstream introns. MEME was run with the default parameters when the length of the searched motif was limited to 20 nucleotides. The identified common motifs were represented as weight matrices and presented in a graphic representation produced by the PICTOGRAM program [http://genes.mit.edu/pictogram.html (27)]. The representation of a motif as a weight matrix allows also for the evaluation of its degree of conservation by computing its information content.

The identified motifs were searched in the different datasets using the program Motifsearch (GCG package).

RESULTS

Features of the skipped exons and the flanking intron sequences

Nucleotide composition. As a first step we determined the nucleotide composition of the skipped exons and of their flanking introns, to find out if they deviate from the control sequences (exons and introns that are not associated with exon skipping). As shown in Table 1, the composition of the skipped exons was not significantly different from that of the control exons. In both skipped and control exons, guanine and cytosine residues were over-represented. However, cytosine

residues were less represented in the skipped exons compared with the control. The introns flanking the skipped exons were not significantly different from the control introns, and adenosine and especially uridine residues were slightly over-represented in flanking introns. Thus, exons that are skipped seem to have nucleotide compositions similar to constitutive exons that are not skipped, and their composition is different from that of introns.

'Quality' of splice sites. It is possible that the boundaries of skipped exons and their flanking introns have 'weak' splice sites that diverge from the consensus splice sites, and therefore they are not selected for splicing. To test this possibility we examined the splice sites of both upstream and downstream introns, as well as of the control introns (22,28). For each of the 5′ splice site signals we generated a weight matrix that represented the frequency of nucleotides in each position. The three weight matrices representing the 5′ splice site signals of the upstream, downstream and control introns, had equivalent values of information content (~8 bits), and they were not significantly different (as revealed by a χ^2 test, with *P*-values ranging between 0.3 and 0.99). A similar analysis was carried out for the 3′ splice sites of the introns and the polypyrimidine tract adjacent to it (including 15 nucleotides upstream from the intron 3′ end). The information content of each of the three matrices was ~8.5 bits and they were not significantly different (with *P*-values ranging between 0.95 and 0.99).

Unlike the 5′ and 3′ splice sites, the position of the branch site cannot be directly determined. Therefore, a region of 50 bases upstream of the 3′ splice site was extracted, and potential branch sites were identified (27). The consensus sequence for the potential branch in our data set is CTNAC. The three potential weight matrices representing the branch signals of the upstream, downstream and control introns, had equivalent values of information content (5.5, 5.2, 5.7 bits, respectively), and they were not significantly different (as revealed by a χ^2 test, with *P*-values ranging from 0.4 to 0.9).

Sequence length. The lengths of the skipped exons ranged from 12 to 236 bases (with an average of 109.1 ± 50.5 bases). The lengths of the flanking introns ranged from 37 to 5114 bases (upstream) and from 40 to 14 884 bases (downstream). These lengths are very similar to the lengths of exons and introns in the control set, and also reflect a large variability in intron lengths. The entire intron sequences were taken for the motif search.

Identification of common motifs

Next, we searched for motifs that may be specific for the exon-skipping process. No assumption was made regarding the location of these motifs, and hence they were searched in the group of exon sequences as well as in both upstream and downstream intron sequences. Two different algorithms that are aimed at detecting common motifs in unaligned sequences were used, MEME (25) and Gibbs sampling (26). We first ran the MEME algorithm and obtained the motifs in a weight matrix representation. This representation allows for the evaluation of the motifs by their information contents in bit units, a measure that reflects the general level of conservation throughout the motif's positions. The motifs can then be ranked by their values of information content, where higher values of information content are associated with motifs that would be more significant and less expected to be found at random. MEME identified common motifs both in the skipped exons and in their flanking introns. However, the information content of the intronic motifs was much higher (~15 bits) than that of the exonic motifs (between 5 and 9 bits for different motifs of the same length), and the intronic motifs were more abundant. It is possible that exonic motifs that play a role in alternative splicing exist, but either they are less well defined and less conserved or they may appear in multiple forms (16). Therefore, we could not identify them in the current analysis that relies on multiple alignment. Hereinafter, we focus on the identified intronic motifs.

The application of the algorithm to the intron sequences has revealed two motifs (Fig. 1). As can be seen in Figure 1A, one motif is highly enriched in pyrimidines (mostly C residues), and is composed of CTCC or CCTCCC repeats. The second motif is highly enriched in purines (mostly G residues), and is composed of AGGG repeats (Fig. 1B). Application of the Gibbs sampling algorithm has produced very similar results. To ascertain that these two motifs are different we aligned their weight matrices by dynamic programming and applied a χ^2 test to compare the nucleotide distributions throughout the aligned positions. The null hypothesis that these two weight matrices are not different was rejected ($P \leqslant 0.005$).

The program Motifsearch was used to locate the identified motifs in the various datasets requiring a significance level of 0.01 (probability to find the motif by chance). In the training data, we found the two motifs in 50 genes (out of 54), where at least one G-rich and one C-rich motif appeared in a complementary manner in the flanking introns. In 33 of these genes (out of 54) the two motifs appeared in both flanking introns of the skipped exons (Fig. 2). Indeed, if these motifs were signals for exon skipping we would not expect to find them frequently in the negative control sequences, but would expect to find them relatively frequently in the flanking introns of skipped exons that were not used for training. A search with Motifsearch program for these motifs in 25 additional sequences that undergo exon skipping identified them in a complementary manner in the flanking introns of 16 out of the 25 skipped exons (64%). In contrast, these motifs were very scarce in the negative control set. Only for 35 out of 810 exons (4.3%) the two motifs appeared in a complementary manner in both flanking introns (at least one motif per flanking intron). In the former analysis the exon-skipped sequences that were used for training and test were derived from two different databases. To assure that our findings are not dependent on the choice of the training set, we repeated the MEME analysis for all 79 sequences and for randomly chosen sets of 54 sequences out of the 79 sequences. As before, the G-rich and C-rich motifs were identified and their weight matrices were not significantly different from the ones defined in the first analysis (with *P*-values ranging between 0.9 and 0.99).

The C-rich motif differs from the PTB motif

PTB protein is a known splicing regulatory protein that functions as a negative regulator of splicing by a mechanism that requires its binding to intronic sequences (silencer elements). Since the potential binding sites for PTB are TCTT (23), TTCTC (29) and CTCTCT (30), we verified that the identified motifs differ from the known elements to which

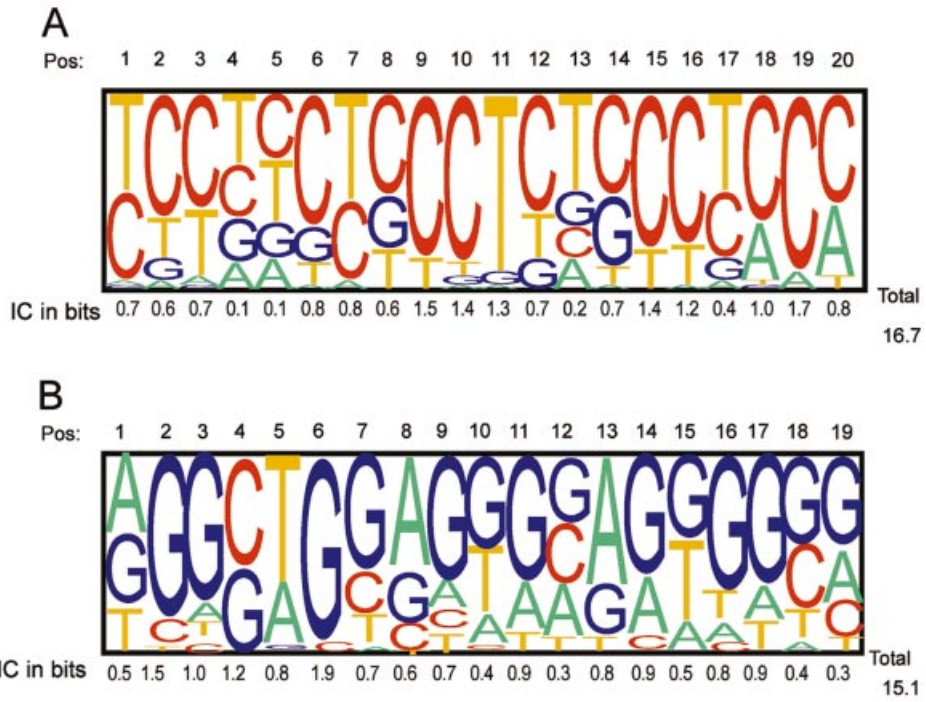

Figure 1. Exon skipping motifs. Two common sequence motifs were identified in introns flanking skipped exons: a C-rich motif (**A**) and a G-rich motif (**B**), represented here graphically. Both upstream and downstream introns show these two motifs. The height of each letter is proportional to the frequency of the corresponding base at the given position, and bases are listed in descending order of frequency from top to the bottom. The information content (IC in bits) relative to the background of the intron base composition (Table 1) is also shown.

PTB binds. We generated the weight matrices for the PTB motifs based on known binding sequences of the PTB protein (listed in Table 2), and compared them with the weight matrices of the newly discovered motifs. The best alignment of the two weight matrices was found by dynamic programming and they were compared by a χ^2 test. The two matrices were found to be significantly different ($P < 0.005$). Still, since our motif is composed of a relatively long C-rich sequence and PTB is known to bind short polypyrimidine-rich tracts, we cannot rule out the possibility that it binds to a sub-site of the identified C-rich motif.

The C-rich motif differs from the known polypyrimidine tract at the 3′ end of introns

Since the 3′ ends of the constitutive introns exhibit a polypyrimidine tract (22) it was important to verify that the C-rich motif does not overlap with this known sequence. Sequences of 50 nucleotides upstream to the 3′ end of introns in the control set were used to define a weight matrix of the known 3′ splice site motif (22). Similarly, an alignment of 3′ ends of the introns upstream and downstream of the skipped exons yielded an alignment consensus, which was not significantly different from the 3′ end of the constitutive introns. Moreover, the weight matrix that was generated by the alignment of the 3′ end sequences was compared with the weight matrix of the C-rich consensus found here, and these were found to be significantly different ($P \leqslant 0.005$).

In addition, while the known signal of the polypyrimidine tract is adjacent to the 3′ end of the intron, the newly identified C-rich motif was found to be distal to the 3′ end in most cases. In about 10% of the cases, the C-rich signal in the upstream intron partially overlapped the 3′ end motif (see also Fig. 2). These exceptions could arise in two possible ways: (i) these introns lack the usual polypyrimidine tract and instead they contain this C-rich motif. (ii) These introns exhibit an unusually long polypyrimidine tract (31), and, in part, it overlaps the C-rich motif reported here.

The purine motif

The second identified motif is enriched in G-nucleotides (Fig. 1B) and is found to be distal to the 5′ splice sites in most cases (Fig. 2). Interestingly, when we ran the MEME program on sets of 54 sequences randomly chosen from the control set, a shorter G-rich motif [composed of (A/T)GGG, or GGG(A/T)] was identified, that was located immediately adjacent to the 5′ splice sites (32), and may correspond to the proposed elements that may play a role in bridging the 5′ and 3′ splice sites, as suggested by Carlo *et al.* (33). A few cases of G-rich elements within the introns of alternatively spliced exons have been reported previously (20,34–36). Extensive mutational analyses

Figure 2. (Next page) The elements' order in the introns flanking the skipped exons is conserved. Exons are represented by boxes, introns (upstream and downstream) by lines, C-rich elements by small red boxes, and G-rich elements by small blue boxes. For each gene (presented by locus number), the locations of the motifs with the highest scores are given in scale. A bar indicates 100 nucleotides. In cases where two motifs received equivalent best scores they were both displayed. Sequences that fit the motif weakly are denoted with the letter w (for weak). In several introns one of the motifs was missing. The intron sequences are of different lengths; however, in all cases the entire flanking intron sequences were taken for the analysis.

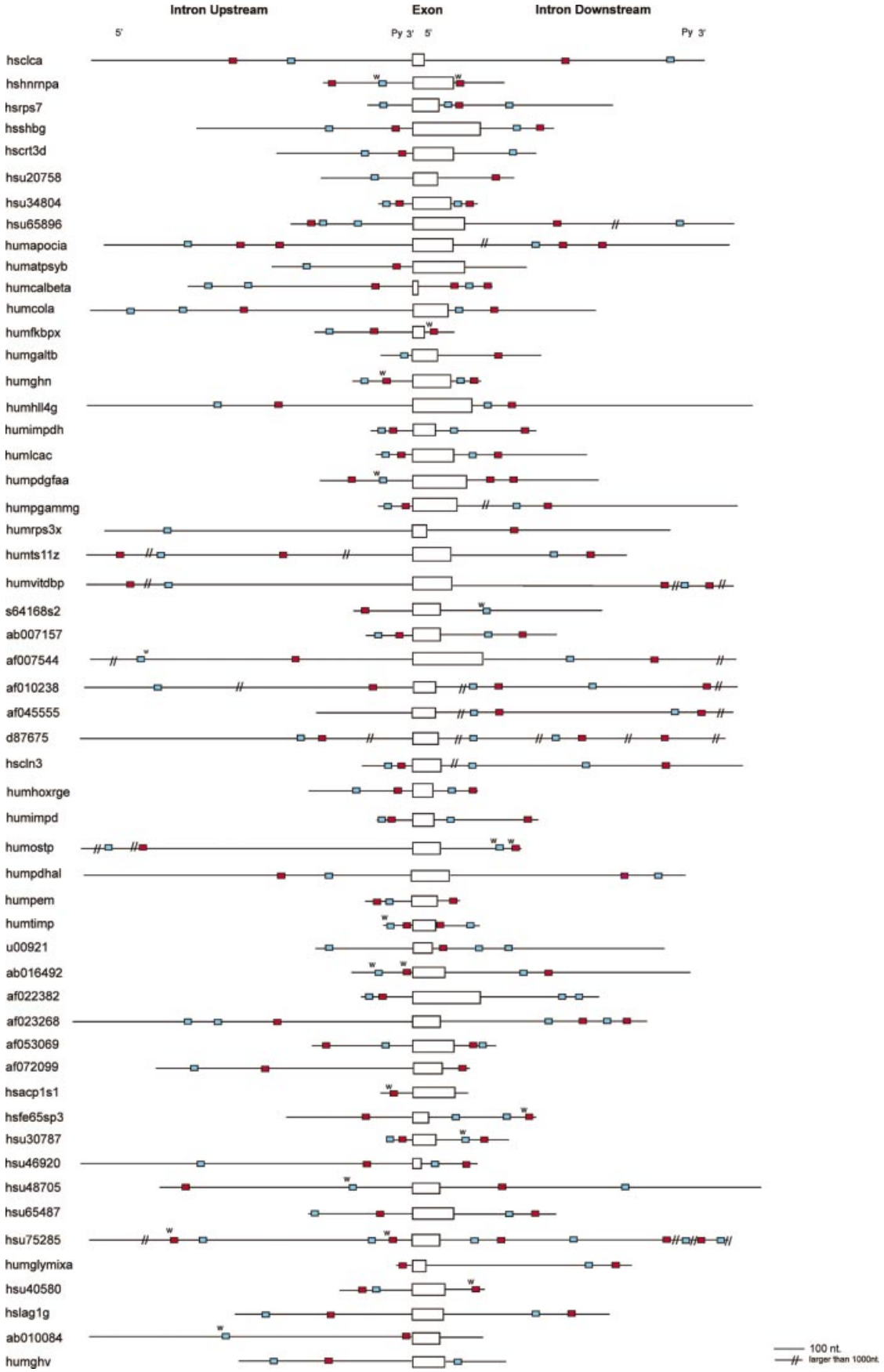
Intron Upstream
Exon
Intron Downstream
5'
Py 3'
5'
Py 3'
hsclca
hshnrnpa
hsrps7
hsshbg
hscrt3d
hsu20758
hsu34804
hsu65896
humapocia
humatpsyb
humcalbeta
humcola
humfkbpx
humgaltb
humghn
humhll4g
humimpdh
humlcac
humpdgfaa
humpgammg
humrps3x
humts11z
humvitdbp
s64168s2
ab007157
af007544
af010238
af045555
d87675
hscln3
humhoxrge
humimpd
humostp
humpdhal
humpem
humtimp
u00921
ab016492
af022382
af023268
af053069
af072099
hsacp1s1
hsfe65sp3
hsu30787
hsu46920
hsu48705
hsu65487
hsu75285
humglymixa
hsu40580
hslag1g
ab010084
humghv
100 nt.
larger than 1000nt.

Table 2. PTB dataset: sequences bound by PTB

Gene	Accession no.	Intron no.	Intron size	Reference
Alpha-tropomyosin	M16432	2	217	(23,55)
	M16432	3	885	(23,55)
Fibroblast growth factor 2 (FGF-R2)	Af169399	Downstream IIIb	1217	(56)
Rat-FGF-R2	Af456422	Upstream IIIb	1059	(57)
	Af456422	Downstream IIIb	1200	(57)
FGF-R1		Upstream α exon	[40]	(58, fig. 3; 59, fig. 1)
		Downstream α exon	[62]	(58, fig. 4)
Calcitonin/CGRP	X15943	Four (partial)	375	(60, fig. 4; 61)
Gamma aminobutyricacid-A (GABA-A)γ2		Upstream γ2 exon	[60]	(39;62, fig. 10)
Caspase-2		Nine (partial)	[50]	(29, fig. 2)
c-src tyrosine kinase	X74765	Upstream and down N1 (entire intron 3)	332	(30,44,63)
Actinin (SM)		Upstream SM	379	(23, figs 5 and 9; 64)

The intron size column gives the sizes of entire introns; otherwise the sizes of the binding elements in brackets are given (length taken for the analysis). In the cases that lack the accession numbers, the sequence regions proposed to interact with PTB are taken from the figures in the references cited.

have shown these elements to function as splicing enhancers (20,34–36). Only one of these genes, the gene encoding growth hormone, was shown to be associated with exon skipping (20). The other enhancers affected differential 5′ selection, such as in the case of the gene encoding thyroid hormone (34), and affected efficiency of splicing (35,36). In these introns, we could identify by local alignment a common core motif of the sequence (A/G)GGG that is shared also by the G element found here. However, the extended consensus of those introns was significantly different from the G-rich element found here ($P \leqslant 0.005$).

Potential RNA structure

Interestingly, the order of the C-rich and G-rich motifs in corresponding flanking introns was conserved. Thus, when the upstream intron has a G-rich motif followed by a C-rich motif, the same order, G-rich followed by C-rich, was conserved in the downstream intron, and vice versa (Fig. 2). However, the distance of the elements from the 5′ or the 3′ splice sites is different in each case. Also, the distances between the motifs within each intron do not seem to be conserved; however, a distance of 40–60 nucleotides between the motifs was the most frequent.

Since one of the motifs is enriched in C and T residues and the other motif is enriched in G and A residues, these two elements can base pair to create a stem–loop structure. Indeed, folding these two sequences by the Mfold program (GCG package) revealed a predicted stem–loop structure in the region surrounding the motifs with free energy of –18.1 kcal/mol (Fig. 3). In Figure 3, a model for the possible role of these elements in exon skipping is illustrated for a gene having three exons and two introns (Fig. 3A). Potential base pairing between the elements located in the flanking introns would bring the 5′ splice site of intron 1 into proximity with the 3′ splice site of intron 3, increasing the frequency of exon 2 skipping (Fig. 3B). Potential base pairing interactions between the elements within the same intron would separate the distal pair of splice sites, decreasing the frequency of exon 2 skipping. These latter interactions could increase the frequency of constitutive splicing by bringing the 5′ splice site of intron 1 into proximity with the 3′ splice site of intron 1, and the 5′ splice site of intron 2 into proximity with the 3′ splice site of intron 2 (Fig. 3C).

DISCUSSION

In the current study we have searched for signals for exon skipping in a database of 54 skipped exons and their flanking introns, extracted from the Intron Sequence Information System (ISIS) (2). We have not addressed the issue of several human genetic diseases that are associated with unnaturally occurring exon skipping. In those cases, mutations within the exons caused abnormal skipping of the exons that harbored those mutations, and the skipping events were attributed to nuclear scanning mechanism or disruption of exonic splicing enhancers (for reviews see 13,15,16,37).

We have focused on genes that normally show exon skipping events, excluding from the data genes with multiple skipping events or events that were followed by additional truncation or extension of the 3′ and/or 5′ ends. Two elements were found to be conserved both in the upstream and downstream flanking introns of the skipped exons: a G-rich element and a C-rich element that demonstrate relative positional order conservation in the two introns. The G-rich and C-rich elements identified in this study are present within most of the flanking introns of the skipped exons in the training set. It is proposed that the C-rich and G-rich elements are candidates to be involved in exon skipping. However, additional elements might be involved in exon skipping in those few introns that lack the C-rich and G-rich elements.

G-rich motifs have been reported previously in association with splicing enhancers. For example, in the case of the growth hormone, exon 3 is skipped, and intron 3 harbors a splicing enhancer (CGGGGATGGGGG) (20). The G repeats are required for the enhancer function, since deletion of the runs of Gs caused a significant increase in exon 3 skipping (from 6 to 60% compared with wild-type) (20). The enhancer sequence localized in the growth hormone (20) coincides with the G-rich element identified here (see Fig. 2, gene humghn). In this gene we have identified the G-rich and C-rich elements in both flanking introns of exon 3 (Fig. 2). The increase in exon skipping by deletion of the G-rich element in intron 3 (20), might be explained by shifting the equilibrium from intra-intronic to inter-intronic base pairing of the elements. Thus, the study of McCarthy and Phillips (20), with regard to exon skipping of the growth hormone transcript, provides experimental support to our suggestion that the elements identified in this study may be involved in exon skipping.

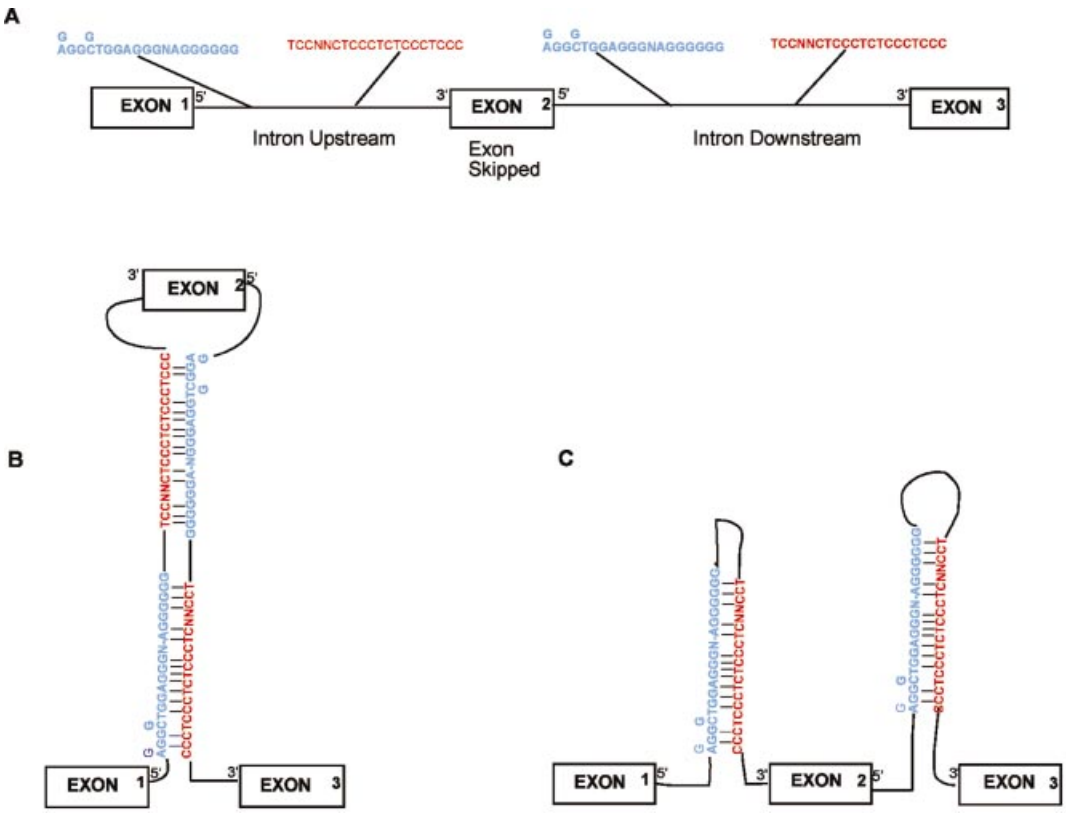

Figure 3. A proposed model for the role of the G-rich and the C-rich elements in exon skipping. A schematic gene with three exons (boxes) and the flanking introns (lines) having two motifs in each flanking intron is displayed. For the motif consensus sequences, only the most frequent nucleotide at each position is illustrated (**A**). Potential base pairing interactions between the elements of the flanking introns would bring the 5′ splice site of intron 1 into proximity with the 3′ splice site of intron 2, increasing the frequency of exon 2 skipping (**B**). Potential base pairing interactions between the elements within the same intron would separate the distal pairs of splice sites, decreasing the frequency of exon 2 skipping. These interactions would bring the 5′ splice site of the upstream intron in proximity to its 3′ splice site, and the 5′ splice site of the downstream intron into proximity with its 3′ splice site, increasing the frequency of constitutive splicing (**C**). Note that the conserved order of the motifs in the flanking introns is important for the potential base pairing interaction that can facilitate the formation of the skipped form.

Other purine-rich elements that affected splice site selection were identified within the intron of the thyroid hormone receptor gene (34), within an α-globin intron (35) and within a β-tropomyosin (36). Therefore, it remains to be seen if the G-rich motif identified in this study is unique to this specific group of genes that show exon skipping, or whether the G motif is also associated with other forms of alternative splicing events.

Regarding the *trans* factor/s that may interact with the purine-rich motif identified in this study, hnRNP A1 protein has been proposed to interact across exons as in the case of hnRNP A1 pre-mRNA (38). When we searched for the consensus sequence for hnRNP A1 binding site, TAG(A/G)G(T/A) (38), we identified it only within two of the motifs found in the 108 introns in our dataset, suggesting that it is unlikely that hnRNP A1 would bind to the purine motif identified in this study.

As for the C-rich element, its comparison with the PTB binding motif has shown that these are different motifs. However, it is still possible that PTB can also bind to the C-element that is discovered here. It may bind either to polypyrimidine-rich sub-sequences of the motif or even to other sub-sequences, as PTB was also found to bind to unusual binding sites, such as the CUG repeats (39). Additionally, some of the exons silenced by PTB are flanked by PTB binding sites on both adjacent introns, thus it has been suggested that PTB proteins can interact across the exon, due to their ability to multimerize (24). Besides PTB, other hnRNP proteins, such as hnRNP K and hnRNP L were found to be interacting with intronic polypyrimidine enhancer sites, as in the chicken β-tropomyosin gene (40). The hnRNP K and α-CP-2KL proteins are part of polyC binding proteins (PCBPs) (41). Less is known about activities of PCBPs proteins in alternative splicing, and whether they can stabilize secondary structures that are involved in this process or/and whether they can bring the elements into juxtaposition. However, given that hnRNP K is an RNA binding protein that can dimerize and oligomerize with multiple proteins (41,42), it is possible that it can interact across the exon via the C-rich element found here.

Recently, an intron regulatory element, TGCATG, was detected at a relatively high frequency within downstream introns for both brain specific and muscle specific alternatively spliced exons (not necessarily skipped forms) (43). This motif has been implicated previously in alternative splicing of several genes including c-src (44), 4.1 (31), myosin heavy chain B (45), fibronectin (EIIIB exon) (46) and other genes (43). However, when we searched for this element within our intron dataset, this element did not appear as frequent as in the data of Brudno *et al*. (43). In our data, this element appeared in

5 of the 54 downstream and 8 of the 54 upstream introns (two genes had this element in both upstream and downstream introns). Hence, it seems that this element may be required for tissue-specific regulation, but it is not required for the exon skipping events, and additional intron regulatory elements should be recruited.

While it is possible that the G-rich and C-rich elements by themselves affect exon skipping, more appealing is the conjecture that they regulate exon skipping by forming a secondary structure, due to their base complementarities (Fig. 3). This postulation is based on the finding that once two elements are present in both flanking introns, the order of the C-rich and G-rich motifs within the flanking introns is conserved. Otherwise, if the relative positions of the elements within the flanking introns were not conserved, the potential duplex structures are less likely to be formed. Thus, we speculate that base pairing interactions between the different *cis*-acting elements identified here may shift the 'exon skipping decision' from 'on' to 'off', simply by switching from inter-intronic base pairing between the elements located in the flanking introns, to base pairing within the intron, respectively (Fig. 3).

This supports a possible mechanism for exon skipping in which *cis*-elements in both upstream and downstream introns that are involved in the formation of a secondary structure mask the splice sites of the alternative exons (Fig. 3B). At this stage this mechanism is suggestive, however it can gain support from previous observations of inter- and intra-intronic base pairing that were shown to affect alternative splicing, as described below.

In several cases of alternative splicing, masking of the splice sites was obtained by base pairing interactions between the exon and its adjacent intron (47–52). Early studies of the mutually exclusive exons 6A and 6B of β-tropomyosin, predicted a large secondary structure to be involved in regulation of alternative splicing. However, only a limited secondary structure (of one stem, including exon 6B and the 3′ splice site of the upstream intron) has been shown to play a role in selecting the alternative exon (48,49). Furthermore, local stable base paired structures that mask the splice sites reduce the usage of the nearby donor splice site (50,51), or recognize efficiently a weak splice signal at the acceptor site (52).

An example for intra-intronic base pairing that affects alternative splicing is found in fibroblast growth factor receptor-2 (FGFR-2) gene, where two complementary elements within an intron enhanced its splicing (53). In the latter case, a short base paired region could occur, but due to the large distance between the elements, proteins were proposed to assist the interaction (53).

Interestingly, RNA structure similar to that postulated in Figure 3C was observed in yeast YL8A gene transcript (54). In this gene, two distinct pairs of complementary sequences act within each intron to prevent exon skipping and ensure inclusion of internal exons. Destroying either intron self-complementarities allows exon skipping to occur, and restoring the complementarities using compensatory mutations rescues exon inclusion. When new complementarities were introduced to the introns, forming a structure similar to that presented in Figure 3B, exon skipping was increased (54). Thus, suggesting that complementary sequences are positioned to function as internal identity elements that bring into juxtaposition only the appropriate 5′ splice sites.

Our results show that in 50 genes, at least one complementary element is present in each flanking intron. In 33 of these genes (out of 54 in the training set) the two motifs appeared in both flanking introns of the skipped exons (Fig. 2). It seems that for an exon to loop out, only one complementary element at each side of the skipped exon is sufficient. In that case, however, potential intra-intronic base paring is less likely to occur, thus reducing constitutive splicing. When the two elements are present in both flanking introns, equilibrium between inter-intronic and intra-intronic base pairing is expected. Proteins that interact with the base paired elements may affect this equilibrium. It is possible that proteins and/or U snRNAs are involved in this switch through binding to one of the stem–loop structures and shifting the equilibrium towards it. Such binding may assist in bringing the elements closer together, or stabilizing the potential secondary structure, if it is formed.

It remains to be seen whether the motifs found here are conserved throughout evolution. As a first step towards this end, we examined the conservation of the motifs in mouse orthologs of the 50 human genes in our data in which the conserved G-rich and C-rich elements were found in both introns flanking the skipped exons (Fig. 2). This analysis was done in three steps. (i) We first searched for the genomic sequences of the mouse orthologs. (ii) Next we searched either for documentation on exon skipping or for EST data that support exon skipping events that correspond to the exon skipping events in the orthologous human genes. (iii) Finally, the flanking introns of the mouse and human corresponding skipped exons were compared in order to examine the conservation of the identified motifs. In total, eight orthologous mouse genes were identified with either documentation or EST data supporting the skipping of similar corresponding exons. Strikingly, in all of these cases the presence, location and the sequences of the G-rich and C-rich elements in the upstream and downstream introns flanking the skipped exons were similar in mouse and men. These examples include the following genes (Fig. 2): survivin, an inhibitor of apoptosis (human gene HSU75285/mouse gene AF115517); amyloid precursor protein, APP (human gene D87675/ mouse contig NW_000108); phospho-glycerate mutase, PGAM2 (human gene HUMPGAMMG/mouse gene AF317587); LST1 (human gene U00921/mouse gene AF109719); uroporphyrinogen decarboxylase, URO-D (human gene HSU30787/mouse contig NW_000211); CLN3 (human gene HSCLN3/ mouse contig NW_000332); galactose-1-phosphate uridyl transferase, GALT (human gene HUMGALTB/mouse contig NW_000206); ribosomal RNA upstream binding transcription factor, UBTF) (human gene HSU65487/mouse contig NW_000040). This list might expand when further exon skipping is identified in homologous mouse genes. Notably, intronic sequences in homologous introns of different species are usually not conserved, except for elements that play a role in splicing or other pre-mRNA processing events. The G-rich and C-rich elements are among the very few intronic sequences conserved between the mouse and human homologous genes that show exon skipping. This conservation further supports their putative regulatory role in exon skipping.

It should be also pointed out that our proposal that the G-rich and C-rich motifs are associated with exon skipping, provides tools to recognize potential alternatively spliced introns by exon skipping in men and other species, a phenomenon that should be further analyzed.

REFERENCES

1. Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nature Genet.*, **30**, 13–19.
2. Croft,L., Schandorff,S., Clark,F., Burrage,K., Arctander,P. and Mattick,J.S. (2000) ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nature Genet.*, **24**, 340–341.
3. Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.
4. Modrek,B., Resch,A., Grasso,C. and Lee,C. (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.*, **29**, 2850–2859.
5. Kan,Z. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **5**, 889–900.
6. Breitbart,R.E., Andreadis,A. and Nadal-Ginard,B. (1987) Alternative splicing: a ubiquitous mechanism for generation of multiple protein isoforms from single genes. *Annu. Rev. Biochem.*, **56**, 467–495.
7. Grabowski,P.J. and Black,D.L. (2001) Alternative RNA splicing in the nervous system. *Progress Neurobiol.*, **65**, 289–308.
8. Clark,F. and Thanaraj,T.A. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451–464.
9. Mahotka,C., Wenzel,M., Springer,E., Gabbert,H.E. and Gerharz,C.D. (1999) Survivin-ΔEx3 and Survivin-2B: two novel splice variants of the apoptosis inhibitor survivin with different antiapoptotic properties. *Cancer Res.*, **59**, 6097–6102.
10. Rodriguez,J.A., Span,S.W., Ferreira,C.G.M., Kruyt,F.A.E. and Giaccone,G. (2002) CRM1-mediated nuclear export determines the cytoplasmic localization of antiapoptotic protein survivin. *Exp. Biol. Med.*, **275**, 44–53.
11. Ponnambalam,S., Jackson,A.P., LeBeau,M.M., Pravtcheva,D., Ruddle,F.H., Alibert,C. and Parham,P. (1994) Chromosomal location and some structural features of human clathrin light-chain genes (CLTA and CLTB). *Genomics*, **24**, 440–444.
12. Duncan,P.I., Stojdl,D.F., Marius,R.M., Scheit,K.H. and Bell,J.C. (1998) The Clk2 and Clk3 dual-specificity protein kinases regulate the intranuclear distribution of SR proteins and influence pre-mRNA splicing. *Exp. Cell Res.*, **241**, 300–308.
13. Cooper,T.A. and Mattox,W. (1997) The regulation of splice site selection and its role in human disease. *Am. J. Hum. Genet.*, **61**, 259–266.
14. Sakamoto,J., Takata,A., Fukuzawa,R., Kikuchi,H., Sugiyama,M., Kanamori,Y., Hashizume,K. and Hata,J.I. (2001) A novel WT1 gene mutation associated with wilms' tumor and congenital male genitourinary malformation. *Pediatr. Res.*, **50**, 337–344.
15. Liu,H.-X., Cartegni,L., Zhang,M.Q. and Krainer,A.R. (2001) A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes. *Nature Genet.*, **27**, 55–58.
16. Cartegni,L., Chew,S.L. and Krainer,A.R. (2002) Listening to silence and understanding nonsense: exonic mutations that effect splicing. *Nature Rev. Genet.*, **3**, 285–298.
17. Fu,X.D. (1995) The superfamily of arginine/serine-rich splicing factors. *RNA*, **1**, 663–680.
18. Liu,H.-X., Chew,S.L., Cartegni,L., Zhang,M.Q. and Krainer,A.R. (2000) Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. *Mol. Cell. Biol.*, **20**, 1063–1071.
19. Dreyfuss,G., Matunis,M.J., Pinol-Roma,S. and Burd,C.G. (1993) hnRNP proteins and the biogenesis of mRNA. *Annu. Rev. Biochem.*, **62**, 289–321.
20. McCarthy,E.M.S. and Phillips,J.A.,III (1998) Characterization of an intron splice enhancer that regulates alternative splicing of human GH pre-mRNA. *Hum. Mol. Genet.*, **7**, 1491–1496.
21. Coward,E., Haas,S.A. and Vingron,M. (2002) SpliceNest: visualizing gene structure and alternative splicing based on EST clusters. *Trends Genet.*, **18**, 53–55.
22. Burge,C.B., Tuschl,T. and Sharp,P.A. (1999) Splicing of precursors to mRNAs by the spliceosomes. In Gesteland,R.F., Cech,T.R. and Atkins,J.F. (eds), *The RNA World*, 2nd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp. 525–560.
23. Wollerton,M.C., Gooding,C., Robinson,F., Brown,E.C., Jackson,R.J. and Smith,C.W. J. (2001) Differential alternative splicing activity of isoforms of polypyrimidine tract binding protein (PTB). *RNA*, **7**, 819–832.
24. Wagner,E.J. and Garcia-Blanco,M.A. (2001) Polypyrimidine tract binding protein antagonizes exon definition. *Mol. Cell. Biol.*, **21**, 3281–3288.
25. Bailey,T.L. and Elkan,C. (1994) *Fitting a Mixture Model by Expectation Maximization to Discover Motifs in Biopolymers. International Conference on Intelligent Systems for Molecular Biology*. AAAI Press, Menlo Park, CA, pp. 28–36.
26. Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.F. and Wootton,J.C. (1993) Detecting subtle sequence signals: Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
27. Lim,L.P. and Burge,C.B. (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA*, **98**, 11193–11198.
28. Shapiro,M.B. and Senapathy,P. (1987) RNA splice junctions of different classes of eukaryotes: sequences statistics and functional implications in gene expression. *Nucleic Acids Res.*, **15**, 7155–7174.
29. Cote,J., Dupuis,S. and Wu,J.Y. (2001) Polypyrimidine track-binding protein binding downstream of caspase-2 alternative exon 9 represses its inclusion. *J. Biol. Chem.*, **276**, 8535–8543.
30. Chou,M.Y., Underwood,J.G., Nikollc,J., Luu,M.H. and Black,D.L. (2000) Multisite RNA binding and release of Polypyrimidine tract binding protein during the regulation of c-src neural-specific splicing. *Mol. Cell.*, **5**, 949–57.
31. Deguillien,M., Huang,S.-C., Morinière,M., Dreumont,N., Benz,E.J.,Jr and Baklouti,F. (2001) Multiple *cis* elements regulate an alternative splicing event at 4.1R pre-mRNA during erythroid differentiation. *Blood*, **98**, 3809–3816.
32. Nussinov,R. (1988) Conserved quartets near 5′ intron junctions in primate nuclear pre-mRNA. *J. Theor. Biol.*, **133**, 73–84.
33. Carlo,T., Sierra,R. and Berget,S.M. (2000) A 5′ splice sites-proximal enhancer binds SF1 and activates exon bridging of a microexon. *Mol. Cell. Biol.*, **20**, 3988–3995.
34. Hastings,M.L., Wilson,C.M. and Munroe,S.H. (2001) A purine-rich intronic elements enhance alternative splicing of the thyroid hormone receptor mRNA. *RNA*, **7**, 859–874.
35. McCullough,A.J. and Berget,S.M. (2000) An intronic splicing enhancer binds U1 snRNPs to enhance splicing and select 5′ splice sites. *Mol. Cell. Biol.*, **20**, 9225–9235.
36. Sirand-Pugnet,P., Durosay,P., Brody,E. and Marie,J. (1995) An intronic (A/G)GGG repeat enhances the splicing of an alternative intron of the chicken beta-tropomyosin pre-mRNA. *Nucleic Acids Res.*, **23**, 3501–3507.
37. Mendell,J. and Dietz,H.C. (2001) When the massage goes awry: disease-producing mutations that influence mRNA content and performance. *Cell*, **107**, 411–414.
38. Blanchette,M. and Chabot,B. (1999) Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization. *EMBO J.*, **18**, 1939–1952.
39. Zhang,L., Liu,W. and Grabowski,P.J. (1999) Coordinate repression of trio of neuron-specific splicing events by the splicing regulator PTB. *RNA*, **5**, 117–130.
40. Expert-Bezançon,A., Le Caer,J.P. and Marie,J. (2002) Heterogeneous nuclear ribonucleoprotein (hnRNP) K Is a component of an intronic splicing enhancer complex that activates the splicing of the alternative exon 6A from chicken β-tropomyosin pre-mRNA. *J. Biol. Chem.*, **277**, 16614–16623.
41. Makeyev,A. and Liebhaber,S.A. (2002) The poly(C)-binding proteins: a multiplicity of functions and a search for mechanisms. *RNA*, **8**, 265–278.
42. Shnyreva,M., Schullery,D.S., Suzuki,H., Higaki,Y. and Bomsztyk,K. (2000) Interaction of two multifunctional proteins heterogeneous nuclear ribonucleoprotein K and Y-box-binding protein. *J. Biol. Chem.*, **275**, 15498–15503.
43. Brudno,M., Gelfand,M.S., Spengler,S., Zorn,M., Dubchak,I. and Conboy,J.G. (2001) Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. *Nucleic Acids Res.*, **29**, 2338–2348.

44. Chan,R.C. and Black,D.L. (1997) The polypyrimidine tract binding protein binds upstream of neural cell-specific c-src exon N1 to repress the splicing of the intron downstream. *Mol. Cell. Biol.*, **17**, 4667–4676.
45. Guo,N. and Kawamoto,S. (2000) An intronic downstream enhancer promotes 3′ splice site usage of neural cell-specific exon. *J. Biol. Chem.*, **275**, 33641–33649.
46. Lim,L.P. and Sharp,P.A. (1998) Alternative splicing of the fibronectin EIIIB exon depends on specific TGCATG repeats. *Mol. Cell. Biol.*, **18**, 3900–3906.
47. Estes,P.A., Cooke,N.E. and Liebhaber,S.A. (1992) A native RNA secondary structure controls alternative splice-site selection and generates two human growth hormone isoforms. *J. Biol. Chem.*, **267**, 14902–14908.
48. Libri,D., Piseri,A. and Fiszman,M.Y. (1991) Tissue-specific splicing *in vivo* of beta-tropomyosin gene: dependence on an RNA secondary structure. *Science*, **252**, 1842–1845.
49. Clouetd'Orval,B., d'Aubenton-Carafa,Y., Brody,J.M. and Brody,E. (1991) Determination of an RNA structure involved in splicing inhibition of a muscle-specific exon. *J. Mol. Biol.*, **221**, 837–856.
50. Varani,L., Hasegawa,M., Spillantini,M.G., Smith,M.J., Murrell,J.R., Ghetti,B., Klug,A., Goedert,M. and Varani,G. (1999) Structure of tau exon 10 splicing regulatory element RNA and destabilization by mutations of frontotemporal dementia and parkinsonism linked to chromosome 17. *Proc. Natl Acad. Sci. USA*, **96**, 8229–8234.
51. Blanchette,M. and Chabot,B. (1997) A highly stable duplex structure sequesters the 5′ splice site region of hnRNP A1 alternative exon 7B. *RNA*, **3**, 405–419.
52. Coleman,T.P. and Roesser,J.R. (1998) RNA secondary structure: an important *cis*-element in rat calcitonin/CGRP pre-messenger RNA splicing. *Biochemistry*, **37**, 15941–15950.
53. DelGatto,F., Plet,A., Gesnel,M.-C., Fort,C. and Breathnach,R. (1997) Multiple interdependent sequence elements control splicing of a Fibroblast growth factor receptor 2 alternative exon. *Mol. Cell. Biol.*, **17**, 5106–5116.
54. Howe,K.J. and Ares,M.,Jr (1997) Intron self-complementarity enforces exon inclusion in a yeast pre-mRNA. *Proc. Natl Acad. Sci. USA*, **94**, 12467–12472.
55. Gooding,C., Roberts,G.C. and Smith,C.W. J. (1998) Role of an inhibitory pyrimidine element and general pyrimidine tract binding protein in repression of a regulated alpha-tropomyosin exon. *RNA*, **4**, 85–100.
56. LeGuiner,C., Plet,A., Gallana,D., Gesnel,M.C., Del Gatto-Konczak,F. and Breathnach,R. (2001) Polypyrimidine tract-binding protein represses splicing of a fibroblast growth factor receptor-2 gene alternative exon through exon sequences. *J. Biol. Chem.*, **276**, 43677–43687.
57. Carstens,R.P., Wagner,E.P. and Garcia-Blanco,M.A. (2000) An intronic splicing silencer causes skipping of IIIb exon of fibroblast growth factor receptor 2 through involvement of polypyrimidine tract binding protein. *Mol. Cell. Biol.*, **20**, 7388–7400.
58. Jin,W., Huang,E.S.-C., Bi,W. and Cote,G.J. (1999) Redundant intronic repressors function to inhibit fibroblast growth factor receptor-1α-exon recognition in glioblastoma cells. *J. Biol. Chem.*, **274**, 28035–28041.
59. Jin,W., McCutcheon,I.E., Fuller,G.N., Huang,E.S. and Cote,G.J. (2000) Fibroblast growth factor receptor-1 α-exon exclusion and polypyrimidine tract-binding protein in glioblastoma multiforme tumors. *Cancer Res.*, **60**, 1221–1224.
60. Lou,H., Yang,Y., Cote,G.J., Berget,S.M. and Gagel,R.F. (1995) An intron enhancer containing a 5′ splice site sequence in the human calcitonin/calcitonin gene-related peptide gene. *Mol. Cell. Biol.*, **15**, 7135–7142.
61. Lou,H., Helfman,D.M., Gagel,R.F. and Berget,S.M. (1999) Polypyrimidine tract-binding protein positively regulates inclusion of alternative 3′-terminal exon. *Mol. Cell. Biol.*, **19**, 78–85.
62. Ashiya,M. and Grabowski,P.J. (1997) A neuron-specific splicing switch mediated by an array of pre-mRNA repressor sites: evidence of regulatory role for the polypyrimidine tract binding protein and a brain-specific PTB counterpart. *RNA*, **3**, 996–1015.
63. Markovtsov,V., Nikollc,J.M., Goldman,J.A., Turck,C.W., Chou,M.Y. and Black,D.L. (2000) Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein. *Genes Dev.*, **10**, 208–219.
64. Southby,J., Gooding,C. and Smith,C.W. (1999) Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of α-actinin mutally exclusive exons. *Mol. Cell. Biol.*, **19**, 2699–2711.