Silencer elements as possible inhibitors of pseudoexon splicing

Manuela Sironi1,*, Giorgia Menozzi1, Laura Riva1,2, Rachele Cagliani1, Giacomo P. Comi3, Nereo Bresolin1,3, Roberto Giorda1 and Uberto Pozzoli1

1IRCCS E. Medea, Associazione La Nostra Famiglia, 23842 Bosisio Parini (LC), 2Department of Biomedical Engineering, Polytechnic University, Milan, Italy and 3Centro Dino Ferrari, Dipartimento di Scienze Neurologiche, Università di Milano, IRCCS Ospedale Maggiore Policlinico, 20100 Milan, Italy

Received December 18, 2003; Revised February 11, 2004; Accepted February 24, 2004

ABSTRACT

Human pre-mRNAs contain a definite number of exons and several pseudoexons which are located within intronic regions. We applied a computational approach to address the question of how pseudoexons are neglected in favor of exons and to possibly identify sequence elements preventing pseudoexon splicing. A search for possible splicing silencers was carried out on a pseudoexon selection that resembled exons in terms of splice site strength and exon splicing enhancer (ESE) representation; three motifs were retrieved through hexamer composition comparisons. One of these functions as a powerful silencer in transfection-based splicing assays and matches a previously identified silencer sequence with hnRNP H binding ability. The other two motifs are novel and failed to induce skipping of a constitutive exon, indicating that they might act as weak repressors or in synergy with other unidentified elements. All three motifs are enriched in pseudoexons compared with intronic regions and display higher frequencies in intron-less gene-coding sequences compared with exons. We consider that a subpopulation of pseudoexons might rely on negative regulators for splicing repression; this hypothesis, if experimentally verified, might improve our understanding of exonic splicing regulatory sequences and provide the identification of a novel mutation target for human genetic diseases.

INTRODUCTION

Production of functional mRNAs in eukaryotic organisms is critically dependent upon the accuracy of pre-mRNA splicing, a highly regulated process assuring that intervening sequences are removed and an ordered array of exons is maintained in mature transcripts. Splicing also represents a powerful and versatile mechanism to control gene expression and to provide functional diversification of proteins (1).

Splicing relies on the correct identification of exons that must be exactly recognized within pre-mRNAs despite their being extremely short compared with intronic regions. Our knowledge indicates that the presence of well-defined cis-elements, namely the 5’ and 3’ splice sites and the branch point, is necessary but not sufficient to define intron–exon boundaries (2). It is now established that sequences within exon bodies have a prominent role in promoting exon definition and inclusion in mature transcripts. The best understood exonic elements are represented by the so-called exonic splicing enhancers (ESEs). ESEs represent binding sites for SR proteins, which are thought to have a role in the initial steps of spliceosome assembly (3–5). Sequences that act as exonic splicing silencers (ESSs) have also been described (6–11) but are less well characterized than ESEs. In some instances, ESSs have been shown to bind negative regulators belonging to the heterogeneous nuclear ribonucleoprotein (hnRNP) family (11,12). The function of ESEs and ESSs appears to be especially important for the regulation of alternative splicing events, but these sequences probably also play a relevant role in the definition of constitutive exons. Human introns are typically thousands of bases long, and it has been reported that, in the hprt gene, sequences that match splice site consensuses (pseudosites) are highly abundant in intronic regions and that pseudoexons (i.e. intronic sequences displaying good 3’ and 5’ splice sites) outnumber real exons by an order of magnitude (13). Detailed analysis of one of these pseudoexons indicated that it was affected by multiple splicing defects that prevented its inclusion in the transcript. Nonetheless, other observations (14,15) suggest that a subpopulation of pseudoexons might exist in the human genome requiring only subtle changes to become splicing competent. Indeed, two recent reports (14,15) indicated that single base pair mutations or microdeletions deep within intronic regions could determine novel exon definition without creating novel splice sites, but rather altering pseudoexon sequences.

Here we have applied a biocomputational approach to address the question of why pseudoexons are ignored and to identify putative splicing repressor elements.

*To whom correspondence should be addressed. Tel: +39 031 877111; Fax: +39 031 877499; Email: msironi@bp.lnf.it

Nucleic Acids Research, Vol. 32 No. 5 © Oxford University Press 2004; all rights reserved
Exon and pseudoexon selection criteria

A total of 110 human genes were used for exon and pseudoexon selection (elite set); genes were selected according to the following criteria: representation in the human gene mutation database (HGMD; http://archive.uwcm.ac.uk/uwcm/hg/hgmd0.html) and definitive annotation in the NCBI Reference Sequence (RefSeq) collection; provisionally annotated genes were not considered. For the construction of the real and pseudoexon test sets, 1000 RefSeq reviewed human genes were used; in particular, genes with evidences of multiple alternative splicing events were discarded. In both cases, intronless HGMD loci were ignored.

Genomic sequences and intron–exon boundaries were obtained from the UCSC human genome annotation database (release hg13) (http://genome.ucsc.edu/cgi-bin/hgGateway).

5′ and 3′ splice site consensus values (CVs) were calculated following the matrices described by Shapiro and Senapathy (16). CVs are plain numbers describing the degree to which individual splice junctions match the consensus (CV varies from 0 to 1).

Real exons were selected according to the following criteria: 50 bp < exon length < 200 bp, presence of both ‘AG’ splice acceptor and ‘GT’ splice donor nucleotides, 3′ CV and 5′ CV greater than 0.7. Exons that undergo alternative splicing events were not included. First and last exons were also excluded.

Pseudoexon selection followed the same criteria used for the selection of real exons; in addition, the presence of at least one branch point-like sequence (YNYURAY) within 60 nt upstream of the 3′ pseudosite was considered an absolute requirement. All selected pseudoexons were checked against the UCSC expressed sequence tag (EST) and mRNA annotation tables, and all aligning sequences were discarded. Finally, we purged pseudoexons that shared a 3′ or 5′ pseudo splice site by choosing the best scoring consensus.

Database creation and sequence analysis

Selected exons (Rex) and pseudoexons (Pex) were collected and organized in a database with the following information: exonic/pseudoexonic sequences, flanking 400 bp sequences (200 bp on each side), consensus values and ESE frequencies. The terms Pex and Rex designate sequences constituting pseudoexon and real exon bodies.

The UCSC annotation tables were used for the identification and categorization of repeated elements. ESE scores were calculated using previously described ESE scoring matrices (17). Intronless gene sequences were derived from the UCSC human genome annotation database by selecting all RefSeq gene collection loci displaying no exon/intron annotations. These sequences were then checked for their translatability so as to purge the set from non-translated mRNAs. Uncertain entries were inspected by hand. Finally, we pruned the data set to remove highly homologous sequences (>70% identity at the nucleotide level).

Search for potential splicing silencer

Potential splicing silencer motifs have been searched for using a slightly modified version of a previously proposed method (18). Briefly, from the 4096 (4⁶) possible hexamers, we selected those having significantly higher frequency in Pex compared with Rex and with Pex flanking sequences; in particular, for each hexamer and each comparison, a statistical significance threshold of 2.5 SDs above the mean (corresponding to P < 0.01) was applied.

This double comparison was intended to exclude those hexamers whose higher frequency was due to differences in the background composition of exons and introns. Then we clustered the selected hexamers using a previously described (18) measure of dissimilarity with a cut-off of 3.1. Only clusters with more than three hexamers were considered. Hexamers from each cluster were aligned using CLUSTALW.

To define consensus matrices, we extracted all occurrences of each aligned hexamer in the Pex set and calculated base frequencies at each position. For each hexamer, flanking nucleotides were also extracted to allow padding of missing edge positions. Frequencies have been normalized by the Pex background base composition and expressed in bits.

Given that, for any consensus threshold greater than 0, more motifs were found in Pex1 than in Rex, the actual threshold was chosen so as to maximize the ratio between the number of motifs found in Rex and in Pex1.

All sequences, data and thresholds are available upon request.

Plasmid construction

Exons 23–25 of the NF1 gene were PCR amplified from human genomic DNA using restriction site-tagged primers; in particular, the forward and reverse primers carried XmaI (Ex23 ForXma: tagcggggctcaggtagttaaagtc) and SalI (Ex 25 RevSal: CTTCTGTCACGGCTGCAGAGGGAGACCAGC) cleavage sites, respectively. The PCR was carried out with JumpStart RED Accu Taq DNA polymerase (Sigma). The PCR fragment was purified using Microcon spin columns (Millipore), digested with XmaI and SalI and directionally ligated into pDisplay (Invitrogen). Colonies were selected on LB kanamycin plates and positive clones were fully sequenced. The construct was named pDisplay NF23–25.

The three identified motifs were inserted in the middle of NF1 exon 24 using a PCR-based approach. Two unrelated sequences (a heptamer and a hexamer) were also inserted as a negative control for length variation on splicing efficiency. A forward motif-tagged primer and reverse oligonucleotide were designed so as to anneal 20 bp downstream (reverse) and 20 bp upstream (forward) of residue 34 in exon 24. The two oligonucleotides were used for PCR amplification using pDisplay NF23–25 as a template and Pfu polymerase (Promega). Thirty-five reaction cycles were carried out with annealing and extension temperatures of 59 and 72°C, respectively. The PCR was then purified using Microcon spin columns and digested with DpnI to eliminate the parental plasmid. The digestion was followed by phosphorylation of the PCR-amplified plasmid and by intramolecular ligation using the Rapid DNA Ligation Kit (Roche); 7 μl of the ligation reaction were used to transform chemically competent TOP10 Escherichia coli cells (Invitrogen). Transformants were selected on LB kanamycin plates and motif insertion was verified through direct sequencing of plasmid DNA.
All sequencing was performed using the BigDye™ Terminator Cycle Sequencing (PE Applied Biosystems) and sequences were run on an ABI PRISM 3100 Genetic Analyzer.

For each construct, at least two independent clones were used in transfection/RT–PCR experiments.

**In vitro splicing assays**

Cos-7 cells were grown in Dulbecco’s modified Eagle’s medium (DMEM) supplemented with 10% fetal calf serum (FCS). Transfections were performed using Lipofectamine™ 2000 (Invitrogen) in OPTI-MEM 1 medium (Gibco) following the manufacturer’s recommendations. At 36 h after transfection, cells were washed with ice-cold phosphate-buffered saline (PBS) and total RNA was extracted using EuroZol (Celbio) in accordance with the manufacturer’s specifications. cDNA synthesis was performed using random hexamers and the first strand cDNA Synthesis Kit (Roche).

PCR amplifications were carried out with AmpliTaq Gold (PE Applied Biosystems) using primers annealing to pDisplay-transcribed regions up- and downstream of NF1 exons. In particular, primer sequences were as follows: HA1-For, CCATAATGATGGTCAGATTG, ABI-Tet labeled; PDGFR-Rev, CAGGAGTGTGCAAGCCAGAT; 25 cycles were carried out with annealing and extension temperatures of 55 and 72°C, respectively. Amplified products were run on an ABI PRISM 310 Genetic Analyzer and visualized using the GeneScan program.

**RESULTS**

**Pseudoxon retrieval and analysis**

A total of 110 human genes were selected for pseudoxon search. In particular, we selected HGMD genes that had been fully annotated and included in RefSeq; this requirement originated from the need to analyze only extensively studied genes whose alternative splicing events have been described in detail. As a matter of fact, the human transcriptome has not been saturated and we aimed at minimizing the possibility of including in the pseudoxon data set sequences that might turn out to be spliced in only a few cell types or physiological states.

Exon and pseudoxon selection followed the criteria described in Materials and Methods: a total of 770 real exons (Rex) and 8128 pseudoxons (Pex) were obtained. In particular, a total of about 7.8 million intronic nucleotides were searched for pseudoxons, resulting in an approximate frequency of one pseudoxon per kb. Since repeated elements have been reported to contain pseudosites, we analysed 5′ and 3′ pseudoxon splice sites for the presence of repeats; 17% of 5′ and 15% of 3′ pseudosites were found to be accounted for by repeated elements. Moreover, in ~19% of cases, the same repeat carried both the 3′ and 5′ splice sites, indicating that the whole pseudoxon was constituted of a transposable element. The categorization of pseudosite-associated repeats is reported in Table 1. LINE elements (of both the L1 and L2 families) accounted for ~41% of pseudosite-containing repeats, while Alu sequences represented the second more abundant repeat family with 24% of pseudosites.

**Comparing real exons with pseudoxons**

As reported above, pseudoxons and real exons were selected only if they presented both 5′ and 3′CVs higher than 0.7. Nonetheless, we wished to determine 5′ and 3′ CV distributions for Rex and Pex; the results are plotted as histograms in Figure 1A and B and indicate that, in general, real exons (median 3′ and 5′ CVs = 0.8871 and 0.8434, respectively) display stronger splice sites than pseudoxons (median 3′ and 5′ CVs = 0.8078 and 0.7760, respectively).

ESE frequencies were also analyzed in exons and pseudoxons (Fig. 2) as well as in the respective 5′- and 3′-flanking sequences (200 bp up- and downstream). Comparison of real exons with pseudoxons indicated that, as shown in Figure 2, the former presented a significantly higher ESE frequency compared with the latter (medians = 0.1329 and 0.1042, respectively; Wilcoxon rank sum test, P < 0.001) while no significant difference was evidenced when real exon 5′- and 3′-flanking sequences (median ESE frequencies = 0.1150 and 0.1100, respectively) were compared with the respective pseudoxon flanks (medians = 0.1100 and 0.1050). Moreover, ESE frequencies were found to be significantly higher in real exons compared with their 5′- and 3′-flanking sequences (Wilcoxon signed rank test of equality of medians, P < 0.0001). Conversely, when pseudoxons were considered, the opposite situation was observed, with ESEs being significantly less represented within pseudoxons compared with flanking sequences (Wilcoxon signed rank test of equality of medians, P < 0.0001); the same result was obtained (P < 0.0001) when ESE frequencies were calculated for a Pex subgroup purged of repeat-containing pseudoxons (data not shown). Thus we compared ESE frequency in

### Table 1. Analysis of repeat-containing pseudosites

<table>
<thead>
<tr>
<th>Class</th>
<th>Family</th>
<th>Pseudosite number</th>
<th>3′</th>
<th>5′</th>
<th>Both</th>
<th>Total pseudosites n</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ALU</td>
<td></td>
<td>358</td>
<td>393</td>
<td>248</td>
<td>999</td>
<td>23.98%</td>
<td></td>
</tr>
<tr>
<td>MIR</td>
<td></td>
<td>156</td>
<td>213</td>
<td>83</td>
<td>452</td>
<td>10.85%</td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td></td>
<td>273</td>
<td>263</td>
<td>765</td>
<td>1301</td>
<td>31.23%</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td></td>
<td>133</td>
<td>121</td>
<td>132</td>
<td>386</td>
<td>9.27%</td>
<td></td>
</tr>
<tr>
<td>L3/CRI</td>
<td></td>
<td>19</td>
<td>10</td>
<td>12</td>
<td>41</td>
<td>0.98%</td>
<td></td>
</tr>
<tr>
<td>LTR elements</td>
<td></td>
<td>119</td>
<td>150</td>
<td>215</td>
<td>484</td>
<td>11.62%</td>
<td></td>
</tr>
<tr>
<td>MER</td>
<td></td>
<td>111</td>
<td>112</td>
<td>113</td>
<td>336</td>
<td>8.07%</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td>43</td>
<td>57</td>
<td>3</td>
<td>103</td>
<td>2.47%</td>
<td></td>
</tr>
<tr>
<td>Low complexity/simple</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>19</td>
<td>30</td>
<td>13</td>
<td>62</td>
<td>1.49%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0.05%</td>
<td></td>
</tr>
</tbody>
</table>
pseudoexons with the average ESE frequency in intronic regions. For this purpose, the longest intron of each gene was searched for ESEs and the median ESE frequency was calculated. The Wilcoxon rank sum test was then applied to compare median ESE frequencies in intronic regions and in Pex: ESEs were found to be significantly less represented within pseudoexons than they are in intronic regions (Pex median = 0.1042; intron median = 0.1273; Wilcoxon rank sum test, \( P < 0.0001 \)).

Selection of a pseudoexon subset and search for putative splicing silencers

Our comparison of real exons with pseudoexons revealed that the latter display, on average, weaker splice site consensuses and lower ESE frequencies. We speculated that Pex displaying stronger splice signals might be enriched in splicing silencer elements (and maybe rely on them for splicing repression) compared with the whole pseudoexon population. Thus we set out to identify a pseudoexon subset resembling real exons in both splice site strength and ESE representation. The following selection criteria were applied to the pseudoexon data set: (i) the presence of a 5’ splice site (8 bp were considered, the same as for CV calculation) exactly matching at least one 5’ splice site among the real exon data set; (ii) the presence of a polypyrimidine tract (–21 to –6 bp from the 3’ splice site) containing at least 11 C/T (11 pyrimidines represent the average C/T content of 3’ real site polypyrimidine tracts); and (iii) an ESE frequency higher than the 25th percentile ESE frequency value in real exon ESE distribution. Moreover, we purged pseudoexons that contained repeated elements. The application of the above-mentioned selections resulted in a new pseudoexon data set, we named Pex1, composed of 454 elements. This pseudoexon subset more closely resembles real exons in terms of both distribution of CVs (Fig. 1C) and ESE frequency (Fig. 2C). Indeed, ESE frequencies were significantly higher in Pex1 compared with both 5’- and 3’-flanking sequences (medians for Pex1, 5’ and 3’ flanks = 0.1344, 0.1200 and 0.1150, respectively; Wilcoxon signed rank test of equality of medians, \( P < 0.0001 \)). Moreover, no significant differences were observed when Rex and Pex1 ESE frequencies were compared.

We used Pex1 elements to search for hexamers that were significantly over-represented in pseudoexons compared with real exons but that also presented significantly higher frequencies in pseudoexon bodies compared with 200 bp up- and downstream flanking sequences. This latter criterion allowed for correction of background nucleotide frequencies, in that exons and introns are known to display a different base composition (19,20).

The search, followed by hexamer clustering, allowed the definition of three motifs (Fig. 3). For each of them, a consensus matrix was calculated after normalization by Pex background base composition.

Motif distribution and possible role as splicing silencers

The relative frequencies of the identified motifs in Pex1, non-Pex1 pseudoexons (Pex-Pex1 set), real exons, introns and intronless gene coding sequences are plotted in Figure 4A. As is evident from the histogram, for each motif the frequency is higher in whole introns (the longest intron for each gene was analyzed, as above) than in exons but, in turn, in intronic regions all motifs appear to be enriched within Pex1 and depleted in non-Pex1 pseudoexons. It is worth noting that this finding cannot be expected to originate from the method through which these motifs were selected since only 400 bp of flanking sequence per pseudoexon were used for hexamer composition comparison; these sequences account for only a small percentage of intron sequences. Remarkably, we found that these motifs display increased frequency in intronless gene-coding sequences compared with Rex.

In order to verify these results, we generated a new data set (test set) by selecting 1000 RefSeq reviewed genes and by extracting real exons (RexT, 4598 elements), pseudoexons (PexT, 39514 elements) and Pex1 pseudoexons (Pex1T, 3314 elements) as described above. Again, for each gene, the longest intron was extracted. The relative frequencies of the three identified motifs in Pex1T, non-Pex1T pseudoexons, RexT and introns are plotted in Figure 4B and indicate that motif frequency is higher in Pex1T compared with intronic regions, RexT and PexT, the latter again displaying fewer motifs than introns.

In numbers, we found that a total (Rex plus RexT) of 297 real exons (5.79%) harbor at least one motif (14 display more than one) while, globally, 398 Pex1 (10.56%) contain one of these motifs. In all cases, motif 2 greatly outnumbers the
We hypothesized that the identified consensuses might act as splicing silencers and, consequently, we speculated that real exons harboring one of them might need stronger splicing signals to be efficiently included in mature transcripts. Thus we analyzed CVs and ESE frequencies in Rex and RexT harboring at least one putative silencer compared with those carrying none. While no difference was evidenced when 5′ and 3′ CV distributions were calculated, ESE analysis indicated that real exons carrying putative silencers displayed a significantly higher ESE frequency compared with those presenting none (mean ESE frequency: 0.1498 and 0.1339, respectively; Wilcoxon rank sum test, \(P < 0.0001\)); histograms are shown in Figure 5.

**Splicing assays**

In order to verify the effect on splicing regulation of the three identified motifs, we constructed an NF1 minigene plasmid. NF1 exons 23–25 were PCR amplified and cloned in an expression vector (pDisplay); the best scoring sequences for the three motifs were then inserted in the middle of NF1 exon 24 so as not to disrupt any ESE. NF1 exon 24 displays relatively low splice site consensus values (0.80 and 0.79 for CV3 and CV5, respectively) and an ESE frequency of 0.119.

In order to verify that length variations had no impact on splicing efficiency, an unrelated heptamer (sequence: TGACTAT) and a hexamer (sequence: TCATGT) were also inserted in exon 24, giving rise to two control vectors. Exon 24 splicing behavior was then evaluated after transient transfections in Cos-7 cells by performing RT–PCRs on total RNA; GeneScan analysis of the PCR products is shown in Figure 6. Exon 24 was efficiently spliced in all cases, except for the construct containing motif 2, where the more prominent signal was accounted for by the exon skipped transcript. Transfection with the two control vectors indicated that size variation had no appreciable influence on splicing.

**DISCUSSION**

One of the major concerns of molecular genetics in the post-genomic era has been the identification of cis- and trans-acting elements regulating constitutive and alternative splicing processes in the human genome. Despite these efforts, we still lack some information concerning the mechanisms that drive exon definition and invariably determine intron removal. It has long been recognized that exonic regions differ statistically from intronic sequences in terms of base composition and hexamer frequency (19,20); nonetheless, these differences must be converted into discrete biological signals.

![Figure 2. ESE frequency distribution. ESE frequencies were calculated for exon or pseudoexon bodies (central panels) and for the respective 3′- (left) and 5′- (right) flanking sequences (200 bp each side). (A) Rex; (B) Pex; (C) Pex1.](image-url)
in order to understand how they can be exploited by the splicing machinery to allow molecular recognition. In recent years, the availability of a great wealth of sequence data has allowed the application of *in silico* strategies to identify novel splicing signals (18,21–23). Here we have applied computational methods to address the question of how pseudoexons are neglected in favor of real exons and possibly to identify sequence elements preventing pseudoexons from being spliced. It had been previously reported that in the human HPRT gene, pseudoexons outnumber real exons by an order of magnitude and the analysis of an arbitrarily selected pseudoexon had indicated that it was affected by multiple splicing defects (13). Our search for pseudoexons in 110 human genes shows that the average occurrence of these sequences is about one per kb and confirms the estimation of a real to pseudoexon ratio of about 1:10. Analysis of CV distributions as well as ESE frequencies indicated that, most probably, the great majority of pseudoexons are affected by multiple splicing defects. Indeed, despite our selection criteria (both 3' and 5' CV higher than 0.7), Pex and Rex CV distributions were very different, with the former displaying higher frequencies for lower values. The same applied to ESE representation with Rex showing significantly higher ESE frequencies compared both with their flanking intronic sequences and with Pex. Interestingly, Pex were found to display significantly lower ESE frequencies than their flanks; the difference was not accounted for by a bias in Pex selection due to the presence of repeated elements since purging for repeat-containing Pex did not modify the result. Moreover, analysis of ESE frequency in whole introns indicated that Pex are depleted of ESEs, displaying significantly lower frequencies than the intron average. Although Pex selection procedures can hardly be considered to introduce a bias on ESE frequency (given that inclusion criteria reside on branchpoint and splice site strength without any concern for Pex body sequence), the existence of a non-random association of DNA sequences might account for this result. Indeed, this same consideration might also explain the lower motif frequency in non-Pex1 pseudoexons compared with the intron average (Fig. 4A and B). One hypothesis to explain these findings is that they might be due to the presence of ancestral repeated sequences that are no longer recognizable as a result of base substitution accumulation but still retain some of their original features accounting for non-random associations.

Anyway, taken together, analyses of splicing features indicate that the majority of pseudoexons display lower splice site CVs and fewer ESEs compared with the real exon average; these features most probably account for the splicing incompetence of a great number of Pex. Nonetheless, some recent observations (14,15) have shown that deep intronic mutations causing subtle changes in pseudoexon sequences can activate their inclusion in mature transcripts and cause a genetic disease. Remarkably, Pagani *et al.* (15) were able to demonstrate that a pathological 4 bp deletion in the ATM gene intron 20 disrupted a novel silencer element that prevented pseudoexon inclusion. This observation raises the interesting possibility that at least a proportion of pseudoexons might be prevented from splicing by *cis*-acting negative regulatory elements rather than by multiple splicing defects. Indeed, it has been reported (24) that splicing inhibitory sequences are rather common in the human genome, indicating that splice silencing might be one of the mechanisms that repress pseudoexon inclusion. In order to test this possibility, we selected a Pex population that resembled real exons in terms of both splice site strength and ESE representation, and we reasoned that these pseudoexons might rely on negative regulators for splicing repression. Three motifs were found to be significantly over-represented in Pex1 compared with Pex1 flanks and Rex. The ability of motif 2 to act as an exonic splicing silencer has already been demonstrated in alternatively spliced exons (11), lending support to the hypothesis whereby pseudoexons displaying good splicing signals are subjected to negative regulators. The sequence UGUGGG in the rat β-tropomyosin gene is responsible for exon 7 skipping in non-muscle cells (11); moreover, a related hexamer (UUGGGU) functions as an ESS in HIV-1 tat gene exon 2.
Chen et al. (11) were able to demonstrate that the UUGGG motif in β-tropomyosin exon 7 exerts its negative effect by binding hnRNP H. The same conclusion was reached by Pagani et al. (26) who demonstrated that a C to G substitution within CFTR exon 9 creates a binding site for hnRNP H (sequence: GUUGGGGG) that closely matches the consensus for motif 2 and that is responsible for diminished exon inclusion. Interestingly, as discussed below, we were able to demonstrate that motif 2 can also function as a splicing silencer when inserted in a constitutive exon.

The third motif we retrieved (maximum consensus: UCUCCCAA) also displays some similarity to a previously identified ESS: Zheng et al. (27) indicated that the C-rich core sequence GGCUCCCC functions as a splicing silencer in bovine papillomavirus type 1. This sequence shares the central C-rich region with motif 3.

Figure 4. Relative motif frequency in the initial set (A) and in the test set (B). The relative frequency of each motif was calculated in Pex1 and Pex1T (black), in non-Pex1/non-Pex1T pseudoexons (striped), whole intron sequences (dark gray), Rex and RexT (light gray) and intronless gene-coding sequences (white). Overall absolute frequencies for the three motifs in real exons (Rex + RexT) and in elite pseudoexons (Pex1 + Pex1T) were as follows: motif 1, 13 and 19; motif 2, 243 and 315; motif 3, 55 and 64.

Figure 5. Comparison of ESE frequencies between motif-containing and motif-lacking real exons (Rex + RexT). ESE frequencies were calculated and expressed as histograms for motif-lacking (upper panel) and motif-containing (lower panel) real exons.

Figure 6. Transfection-based splicing assays and RT–PCR. Motifs 1, 2 and 3 were inserted in NF1 exon 24 and splicing efficiency was evaluated after transient transfection and RT–PCR; amplified products (shaded peaks) were run on a Genetic Analyzer. White peaks represent molecular weight markers. CTR, control vector (no insert); motif 1, 2 and 3, vectors carrying each of these motifs within exon 24, 7mer and 6mer, control vectors containing either an unrelated heptamer or hexamer (control of size variation on splicing efficiency).
Remarkably, the novel exon that is included in the α-galactosidase mRNA as a consequence of pseudoexon mutation (14) also carries a motif 3 consensus (sequence UCUCCCCA); the GLA gene was not in our 110 gene database, but application of the Pex1 selection criteria to this novel exon revealed that all requirements were fulfilled. The pathological substitution in GLA intron 4 does not alter the putative silencer sequence but rather is believed to induce exon inclusion through the creation of a novel exonic enhancer. This observation is consistent with our hypothesis whereby Pex1 pseudoexons have the potential of being spliced and thus carry silencer elements; in such a semi-permissive context, mutations that alter the silencer/enhancer balance might determine some degree of exon inclusion.

Further support for the notion whereby the three retrieved elements may function as splicing suppressors in pre-mRNAs came from the analysis of their distribution (Fig. 4). All three motifs appear to be enriched in Pex1 compared with intronic regions and Pex. Most importantly, each motif displays higher frequencies in intronless gene-coding sequences compared with Rex. Intronless genes have been shown to contain sequence elements capable of splicing inhibition when inserted in a heterologous context (28); these sequences, besides functioning in mRNA nuclear export, probably play a relevant role in the suppression of aberrant internal splicing in long coding sequences of intron-lacking genes (28). Also, potential splicing-inhibitory sequences are expected to be selected against in intron-containing genes but to be neutral to intronless ones.

Despite these considerations, our transfection-based assays failed to demonstrate any splicing inhibitory role for motifs 1 and 3. While insertion of motif 2 in NF1 exon 24 determined exon skipping in a consistent proportion of transcripts, no effect was evidenced when the other two motifs were inserted. Nonetheless, it should be considered that NF1 exon 24 is a constitutive exon (notwithstanding its relatively low CVs and ESE frequencies) and, consequently, motifs 1 and 3 might exert silencing effects which are too weak to induce any variation in its splicing efficiency. Indeed, all three motifs are present in a fraction of real exons (5.79%), indicating that, in many circumstances, their presumptive silencing activity is not sufficient to induce exon skipping. In line with this view, we found a significantly higher ESE representation in real exons harboring at least one putative silencer compared with those displaying none, once more suggesting that multiple weak positive and negative elements might have a role in the final decision of whether an exon is spliced and in splicing efficiencies. As a further confirmation of the above-mentioned statement, it is worth noting that motif 2, which acts as a powerful splicing silencer in our assays, is by far the most abundant identified motif in both real and pseudoexons.

Up to now, a single case of pseudoexonic ESS mutation has been clearly identified as pathological by resulting in pseudoexon inclusion; nonetheless, these events might be underestimated in the statistics of disease-causing mutations since their identification depends on transcript analysis (which is only feasible in a minority of cases) and because intronic regions are usually not subjected to mutation screening. The experimental testing of the motifs described here as well as the description of new cases of pseudoexon inclusion might provide the identification of a novel mutation target for human genetic diseases.

Recent data (29) suggest that novel exon creations, despite their potential role in human pathogenesis, might be regarded as a means to obtain an increased rate of change in proteomic diversity. This is probably accomplished through a multistep process where the novel exon is initially included in a minority of transcripts (due to its weak splicing signals) and is then allowed to evolve without compromising the original gene function. Detailed analyses have shown that ~4% of human genes contain transposable elements that have been exonized (30); in particular, it has been reported that exons containing Alu sequences are mostly alternatively spliced, suggesting that constitutive (or high level) Alu exon inclusion might be deleterious and, consequently, be selected against (31). The same possibly holds true for novel exons that do not contain repetitive elements whose origin is, consequently, more difficult to establish. In this regard, the availability of a great number of sequences requiring only subtle changes to become at least partially competent for splicing might be regarded as a potential benefit. Mutations determining novel exon creation might be expected to be pathogenic or not depending on the relative splicing efficiency of the novel exon that might ultimately be the result of the mutation as well as of the pre-existing balance of positive and negative splicing regulatory elements within the pseudoexon.

ACKNOWLEDGEMENTS

We are grateful to Dr M.T. Bassi for useful discussion about the paper.

REFERENCES


