Silencer elements as possible inhibitors of pseudoexon splicing

M. Sironi [(1)], L. Riva[(1, 2)], G. Menozzi[(1)] and U. Pozzoli[(1)],

[(1)] IRCCS E. Medea, Associazione La Nostra Famiglia, 23842 Bosisio Parini (LC) -Italy
upozzoli@bp.lnf.it
[(2)] Department of Biomedical Engineering, Polytechnic University, Milan –Italy

Keywords: splicing, pseudoexon, silencer

Introduction

Production of functional mRNAs in eukaryotic organisms is critically dependent upon the accuracy of pre-mRNA splicing. 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 (1). Sequences within exon bodies have a prominent role in promoting exon definition; the best understood exonic elements are represented by exonic splicing enhancers (ESE) which represent binding sites for SR proteins (2). Sequences that act as exonic splicing silencer (ESS) have also been described but are less well characterized than ESEs. It has been reported that pseudoexons (i.e. intronic sequences displaying good 3' and 5' splice sites) outnumber real exons by an order of magnitude (3). Recent observations (4, 5) suggest that a subpopulation of pseudoexons might exist in the human genome requiring only subtle changes to become splicing competent. Here we have applied a biocomputational approach to address the question of why pseudoexons are ignored and to identify putative splicing repressor elements.

Materials and Methods

Exon and pseudoexon selection criteria

110 human genes were used for exon and pseudoexon selection (elite set); all genes were represented in the human gene mutation database (HGMD) and definitively annotated in RefSeq. For the construction of the real and pseudoexon test sets, 1000 RefSeq reviewed human genes were used. Genomic sequences and intron/exon boundaries were obtained from the UCSC human genome annotation database (release hg13). 5' and 3' splice site consensus values (CVs) were calculated following the matrices as described (6). 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> 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 and the presence of at least one branch point-like sequence (YNYURAY) was considered an absolute requirement. All selected pseudoexons were checked against the UCSC 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.

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. ESE scores were calculated using scoring matrices (7).

Search for potential splicing silencer

Potential splicing silencer motifs have been searched for using a slightly modified version of a previously proposed method (8). Briefly, from the 4096 (46) possible hexamers we selected those having significantly higher frequency in Pex compared to Rex and to Pex flanking sequences; in particular, for each hexamer and each comparison, a statistical significance threshold of 2.5 standard deviations above the mean (corresponding to a $p<0.01$) was applied.

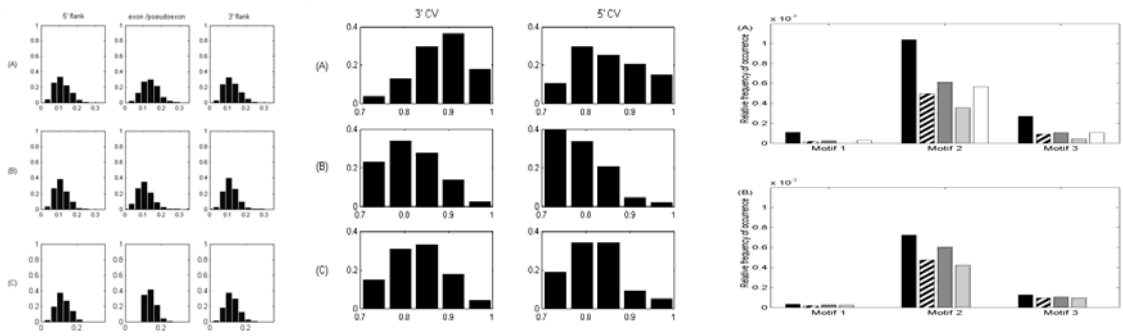
This double comparison was intended to exclude those hexamers whose higher frequency was due to differences in the background composition of exons and introns. Then we clustered the selected hexamers using a previously described (8) measure of dissimilarity with a cut-off of 3.1. Only clusters with more than 3 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.

Results and Discussion

A total of 770 real exons (Rex) and 8128 pseudoexons (Pex) were obtained.5' and 3' CV distributions for Rex and Pex (Fig.1 A and B)indicate that, in general, real exons (median 3'and 5' CVs = 0.8871 and 0.8434, respectively)

display stronger splice sites than pseudoexons (median 3'and 5' CVs = 0.8078 and 0.7760, respectively). Moreover, our data indicate (fig 2 A and B)that the great majority of pseudoexons are depleted of exonic splicing enhancers (ESEs), displaying significantly lower frequencies than the intron average and raising the possibility that some selective processes might act against ESE-enriched pseudoexons.



We thus set out to identify a pseudoexon subset that resembled real exons in both splice-site strength and ESE representation. The following selection criteria were applied to the pseudoexon data set:1) the presence of a 5'splice site (8 bp were considered, as it is for CV calculation) exactly matching at least one 5' splice site among the real exon dataset; 2) the presence of a polypyrimidine tract (-21 to -6 bp from 3'ss) containing at least 11 C/T (11 pyrimidines represents the average C/T content of 3' real site polypyrimidine tracts); 3) an ESE frequency higher than the 25th percentile ESE frequency value in real exon ESE distribution. This pseudoexon subset (Pex1) resembles real exons in terms of both CVs distribution (Fig. 1C) and of ESE frequency (Fig. 2C).
We used Pex1 elements to search for putative silencer elements and retrieved 3 motifs (figure 3). For each of them a consensus matrix was calculated after normalization by Pex background base composition. Higher scoring consensuses for the three motifs are as follows: **CUAGAGGU, U/GGU/AGGGG, UCUCCCAA.**

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 3 A. As it is evident from the histogram, for each motif the frequency is higher in whole introns than in exons but, in turn, in intronic regions, all motives appear to be enriched within Pex1 and depleted in non-Pex1 pseudoexons. In order to verify these result, 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). The relative frequencies of the three identified motifs in Pex1T, non-Pex1T pseudoexons, RexT and introns and are plotted in figure 4B and indicate that, motif frequency is higher in Pex1T as respect to both intronic regions, RexT and PexT, the latter again displaying fewer motifs than introns.

Literature search revealed that motif 2 exactly matches a previously identified silencer element with hnRNP H binding ability (9), while motifs 1 and 3 are novel. In order to verify the effect on splicing regulation of the three identified motifs, we constructed an *NF1* minigene plasmid (exons 23-25). 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. Exon 24 splicing behavior was then evaluated after transient transfections in Cos-7 cells performing RT-PCRs on total RNA: it 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.
Overall, our analyses of splicing features indicate that the majority of pseudoexons display lower splice-site CVs and fewer ESEs compared to the real exon average; these features most probably account for the splicing incompetence of a great number of Pex. Nonetheless, we consider that our data indicate the existence of a sub-population of pseudoexons that might rely on negative regulators for splicing repression.

References

(1) Senapathy, P., Shapiro, M.B., Harris, N.L. (1990). *Methods Enzymol*., **183,** 252-78
(2) Manley, J.L., Tacke, R. (1996). *Genes Dev*., **10**, 1569-79.
(3) Sun, H., Chasin, L.A. (2000). *Mol. Cell. Biol,*. **20,** 6414-25.
(4) Ishii, S., Nakao, S., Minamikawa-Tachino, R., *et al.* (2002) *Am. J. Hum. Genet.,* **70,** 994-1002.
(5) Pagani, F., Buratti, E., Stuani, C., et al.. (2002). *Nat. Genet.,* **30,** 426-9.
(6) Shapiro, M.B. and Senapathy, P. (1987). *Nucleic Acids Res.,* **15,** 7155-7174
(7) Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., Krainer, A.R. (2003). *Nucleic Acid Res.,* **13,** 3568-3571
(8) Fairbrother, W.G., Yeh, R.F., Sharp, P.A., Burge, C.B. (2002). *Science,* **297,** 1007-13
(9) Chen, C.D., Kobayashi, R., Helfman, D.M. (1999). *Genes Dev*., **13**, 593-606.