Intron and exon lengths influence on splicing

G. Menozzi[(1)], L. Riva[(1, 2)], M. Sironi [(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: introns length, exon length, consensus values, splicing enhancers

Introduction

Splice site consensus values (CVs) are usually calculated using previously described matrices [1] which are obtained through the analysis of a relatively small splice site number (1500) from different organisms. Now, genome annotation becoming complete, a much more accurate definition is possible. Furthermore, recent studies [2, 3, 4] indicate that consensus value itself is not sufficient to define splice site strength and other parameters must be considered to improve splice site definition. To investigate how intron and exon lengths might be exploited by the splicing machinery to ensure proper splicing control and regulation a human intron database has been developed and analyzed.

Methods and Results

mRNA sequences of reviewed RefSeq genes along with the corresponding genomic sequences have been extracted from the UCSC database [5] release 13; exon/intron boundaries have been checked and corrected by means of Shapiro and Senapathy consensus values.

A total of 50787 different introns were collected from 4652 reviewed genes; for each intron our database contains length, phase, position, junctions sequences and indication of possible alternative splicing events. Gene reference and flanking exons details are also reported.

Our analysis was performed on a data set containing each 5'intron-exon-3'intron group (IEI). The 95.8% of the initial set was accounted for by internal, constitutive exons with AG-GT boundaries. Analyses were restricted to this sub-group.

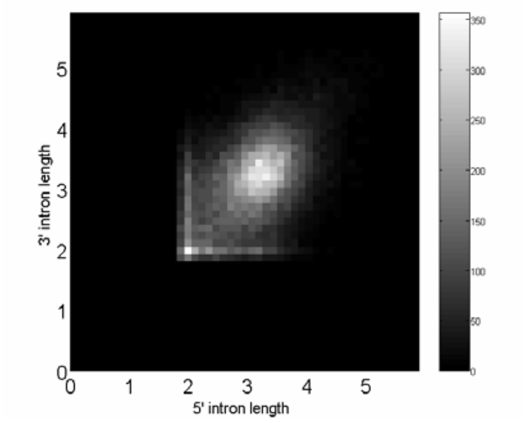

Fig 1. Ditribution of 3' and 5' intron log10 length.

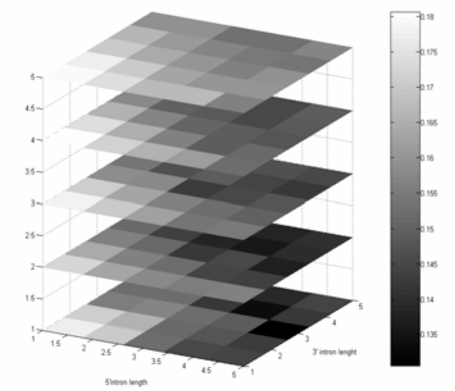

Fig.2, ESE frequency in IEI classes .

Distribution of flanking intron lengths (Fig.1) shows that IEIs tend to cluster in two distinct regions: a major one (IEIR1) with 5' and 3' introns relatively long (2000 bp) and comparable in length, the minor one (IEIR2) with one or both short introns (100 bp).

To efficiently analyze length effects on splicing parameters we divided the IEI set in 5 percentile intervals (20,40,60,80,100) along each dimension (5'intron, exon and 3'intron) obtaining 125 classes whose IEI number ranges from 101 to 957. Class averages were calculated for the following parameters: 3' ss consensus value (CV3), 5'ss consensus value (CV5), putative exonic splicing enhancer (ESE, [6]). We also calculated class averages for consensus values of polypyrimidine tract (CV3py: from -14 to -5) and around the 3'ss (CV3ss: from -4 to +1). Results (Fig.2 and 3) show significant differences (Wilcoxon rank sum test, $p<0.01$) in splicing parameters between IEIR1 and IEIR2 regions. CV3 and CV5 are higher in IEIR1 than in IEIR2. CV3 differences are sharper when CV3py and CV3ss are analyzed separately. ESE are more represented in IEIR2 exons than in IEIR1 ones. Differences between IEIR1 and IEIR2 increase with exon length. These results are consistent with the hypothesis (as reviewed in [2]) whereby exons are the basic units of recognition when flanked by long introns (IEIR1) while introns are instead identified when they are short and flanked by relatively long exons.

Our data also indicate that different frequency matrices should be used in CV calculations so as to account for differences in exon/intron length distribution. We consider that the information reported herein might be useful to implement more efficient exon recognition algorithms.

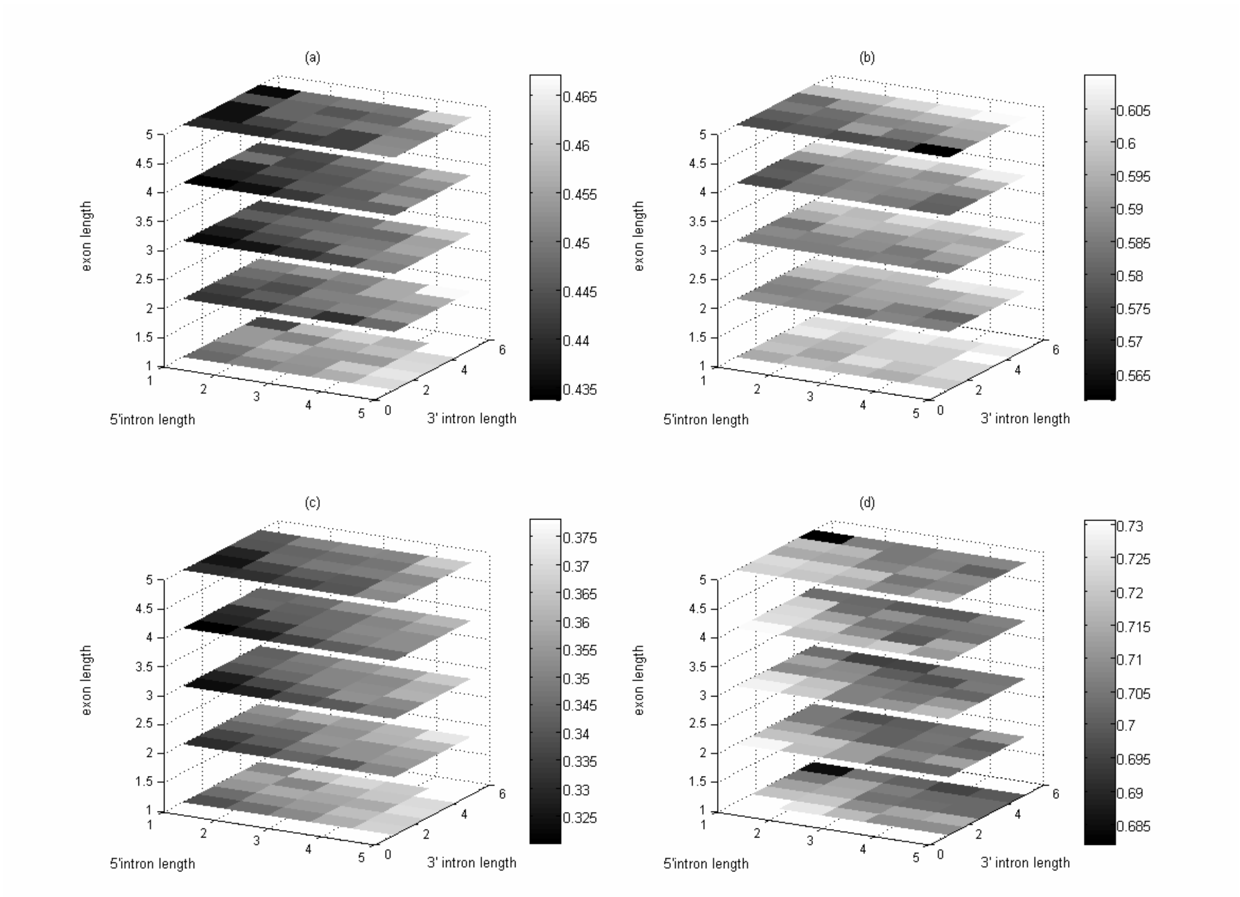

Fig. 3 CV3 (a), CV5 (b), CV3py (c) and CV3ss (d) in IEI classes

References

[1] MB Shapiro and P Senapathy, RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. Nucleic Acids Research, 15(17):7155-74, 1987.

[2] L Cartegni, SL Chew and AR Krainer, Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet, 3(4):285-98, 2002.

[3] SM Mount, C Burks, G Hertz, GD Stormo, O White, C Fields, Splicing signals in Drosophila: intron size, information content, and consensus sequences. Nucleic Acids Research, 20(16):4255-62, 1992.

[4] Z Dominski, R Kole, Cooperation of pre-mRNA sequence elements in splice site selection. Mol Cell Biol., 12(5):2108-14, 1992.

[5] http://genome.ucsc.edu

[6] L Cartegni, J Wang, Z Zhu, MQ Zhang, AR Krainer, ESEfinder: a web resource to identify exonic splicing enhancers, Nucleic Acid Research, 2003, Vol. 31, No.13 3568-3571