Poster M-22

Choosing the End: Regulation of Alternative 3' Splice Sites

Authors:
Yael Mandel-Gutfreund (*Faculty of Biology, Technion-Israel Institute of Technology*)
Martin Akerman (*Faculty of Biology, Technion-Israel Institute of Technology*)

Short Abstract: In this study we apply a Support Vector Machine to identify alternative 3' splice site events. We show that we can distinguish between alternative and constitutive event both for tandem acceptor motifs and for splice sites which are distant apart. Finally, we suggest a possible mechanism of splice site selection.

Long Abstract:
Alternative splicing (AS) constitutes a major mechanism creating protein diversity in humans. This can result from skipping entire exons or by altering the selection of the splice sites that define the exon borders. Alternative 3' Splice Sites (A3SS) represents ~18.4% of all AS events conserved between human and mouse. Half of these events involve the NAGNAG motif at the 3′ splice sites. Though the NAGNAG motif is found frequently in 3' splice sites, only ~4% of these tandem acceptors are confirmed by EST to be alternatively spliced in human and mouse, while in 86% of the cases the proximal splice site is constitutively selected (P) and in 10% the distal splice site is chosen (D). We have previously shown that it is possible to distinguish between sequences that undergo alternative versus constitutive splicing (CS) at the NAGNAG motifs without relying on EST data. Among the features which are characteristic of the AS NAGNAG sites are: high sequence conservation of the motif, high conservation of ∼30 bp at the intronic regions flanking the 3′ splice site and overabundance of cis-regulatory elements in the flanking intronic and exonic regions. In this study we compute these and others characteristics for a variety of training sets of alternatively and/or constitutively spliced NAGNAG motifs and use a Support Vector Machine (SVM) in order to automatically classify them. We show that we can discriminate AS events from both P and D events with relative high accuracy. Interestingly, a high performance is attained when separating P from D events. The most discriminative parameters between the groups are splice site composition, namely the sequence of the NAGNAG motif, intronic conservation near the splice site and the strength and relative position of the polypyrimidine tract. Based on the assumption that tandem acceptors are a subset of all Alternative 3' Splice Site events, we expand our analysis to further cases in which the acceptors sites are spaced from each other at distances varying from 4 to 100 nucleotides. Using similar parameters the SVM succeeds to discriminate between A3SS events and a control set of Constitutive Splicing events including an AG site in varying distances. In addition, we observe that when using a control set of sequences in which the AG which does not serve as a splice site is not evolutionarily conserved, the SVM performance is considerably higher than when the conservation of the AG dinucleotide is not accounted for. The later result could suggest that 1- there is contamination of the CS ESTs within the AS events, 2- In some cases an AS-like regulatory process in required in order to avoid the selection of a proximal splice site, which is usually chosen as the default.
In order to examine the second assumption we have carried out additional tests comparing

subsets of CS sequences to the AS datasets. In a subset of sequences in which the distance between the splice site and the nearest AG site is 4-12 nt, we find that when the AG dinucleotide is evolutionarily conserved, SVM performance considerably drops compared to cases in which the AG dinucleotide is not conserved. We interpret that only in the former group the sequence environment resembles that of Alternative 3' Splice Site events. However, when the AG dinucleotide is located far from the splice site (30-100 nt), the sequence environment at the CS events does not display the characteristic features of AS neither when the AG site is conserved nor when it is not. Overall, our results imply that regulation is necessary in order to avoid the selection of undesired AG sites. Interestingly, we do not observe this regulation when the AG site is not conserved, perhaps because these sites are still evolving. Moreover such regulation seems not to be required when an AG is placed far from the real splice site, presumably upstream the branch point, since this AG will be sequestrated in the lariat during the second step of transesterification. In Conclusion our findings indicate that both the selection between two alternative splice sites and also the recognition of a constitutive AG site from a range of possible splice sites are regulated processes.