

Exonic splicing motif discovery using positional bias

Brad Friedman^{1,2}, Michael B. Stadler¹, Christopher B. Burge¹

Keywords: RNA splicing, regulatory motifs, positional bias, k -mer statistics

Introduction

A great majority of human proteins are non-contiguously encoded in the genome. Excision of the long non-coding interruptions (introns) from transcribed RNA and subsequent splicing together of the much shorter remaining coding portions (exons) is essential for the production of these proteins. In the cell, the pattern of splicing is largely controlled by a set of RNA sequence motifs found at or near the splice sites. Although the motifs that mark the exact splice sites are well characterized, similar sequences also appear within introns or exons and therefore these motifs do not completely specify the boundaries of introns. Identifying a complete set of splicing elements including both exonic splicing enhancers (ESE) and silencers (ESS) and intronic splicing enhancers and silencers (ISE and ISS) is an area of active research.

The problem of creating a library of splicing motifs has been attacked from many angles. Experimentally, motifs have been amplified by artificial evolution using known splicing factors, ([3], [1]), and through high-throughput library screens ([4]). Computationally, several studies have effectively predicted splicing motifs. These studies calculated the prevalence of particular oligomers in some positive set relative to some negative set. For instance, studies examined oligomers' prevalence in exons with strong splice sites versus weak splice sites [2], and in non-coding exons versus intronic regions flanked by splice sites [5].

Our current effort seeks to identify exonic splicing motifs by taking advantage of the positional bias of previously identified motifs near splice sites: sequences that promote splicing are expected to be enriched near constitutive splice sites, and sequences that silence splicing may be depleted near constitutive splice sites. This analysis differs from previous studies in that we analyze the complete positional distribution of each hexamer and also completely control for patterns related to the protein coding function of exons. We are therefore able to use all known coding exons in the human genome in our analysis, rather than restricting to non-coding exons, for example. This gives statistical power to detect weak biases.

Methods

We think of synonymous oligonucleotides as competing for particular spots in an mRNA. For example, any of the hexamers CAACAA, CAGCAA, CAACAG or CAGCAG read in phase 0 would code for consecutive glutamine amino acids, but for each instance of diglutamine in the

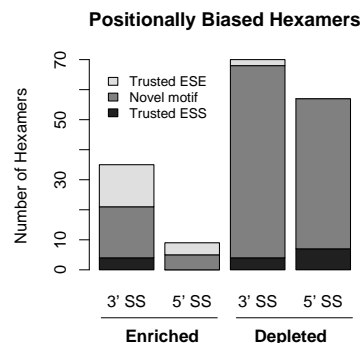


Figure 1: Comparison of positionally biased hexamers and previously reported splicing motifs. Trusted splicing motifs are from [4], [2] and [5]

¹Department of Biology and ²Department of Mathematics, MIT, 77 Massachusetts Ave, Cambridge, MA, 02139. E-mails: baf@mit.edu, stadler@mit.edu, and cburge@mit.edu.

human genome, only one of the four is the winner. Our null model is that although certain hexamers may be preferred overall, these preferences are constant for diglutamine-coding positions independent of position relative to splice sites.

Exonic splicing motifs should deviate from the null model in the sense that enhancers should win the competition and silencers should lose more often near splice sites than in the middle of exons. By scanning the genome for winning and losing positions, we can determine for each hexamer whether it exhibits a significant bias towards or away from splice sites. Using the median of the winning positions we are able to compute the extent and significance of such biases without binning or monte carlo simulations. Indeed, the cumulative distribution function for the median of the winning sites is given by

$$\begin{aligned} P(\text{median} \leq m) &= P(\text{at least } \frac{w}{2} \text{ winning sites at distance } \leq m \text{ from the splice site}) \\ &= \sum_{r=\lceil w/2 \rceil}^w P(\text{exactly } r \text{ winning sites at distance } \leq m \text{ from the splice site}) \\ &= \sum_{r=\lceil w/2 \rceil}^w \frac{\binom{M}{r} \binom{N-M}{w-r}}{\binom{N}{w}}, \end{aligned}$$

where w is the total number of winning sites, M is the total number of sites (winning *or* losing) at distance $\leq m$ from the splice site, and N is the total number of sites.

Results

Nine hexamers were found to be significantly enriched (putative ESE) near the 3' splice site, and 35 near the 5' splice site, while 57 were depleted (putative ESS) near the 3' splice site and 70 near the 5' splice site. The depleted sets tended to contain more previously reported exonic splicing silencers and the enriched sets tended to contain more previously reported exonic splicing enhancers (Figure 1).

Further analysis of these results and experimental validation of these predictions are currently underway and will be presented in greater detail at the meeting.

References

1. L.R. Coulter, M.A. Landree, and T.A. Cooper. Identification of a new class of exonic splicing enhancers by in vivo selection. *Mol Cell Biol*, 17(4):2143–50, 1997.
2. W.G. Fairbrother, R.F. Yeh, P.A. Sharp, and C.B. Burge. Predictive identification of exonic splicing enhancers in human genes. *Science*, 297(5583):1007–13, 2002.
3. R. Tacke and J.L. Manley. The human splicing factors ASF/SF2 and SC35 possess distinct, functionally significant RNA binding specificities. *EMBO J*, 14(14):3540–51, 1995.
4. Z. Wang, M.E. Rolish, G. Yeo, V. Tung, M. Mawson, and C.B. Burge. Systematic identification and analysis of exonic splicing silencers. *Cell*, 119(6):831–45, 2004.
5. X.H. Zhang and L.A. Chasin. Computational definition of sequence motifs governing constitutive exon splicing. *Genes Dev*, 18(11):1241–50, 2004.