The Logic and Formulation of Exon Definition for Splice and Splicing Regulatory Sites with Negative Information Content

Update on:

Mucaki EJ, Shirley BC, Rogan PK. Prediction of mutant mRNA splice isoforms by information theory-based exon definition. Hum Mutat. 2013 Apr;34(4):557-65.

and the "Automated Splice Site and Exon Definition Analysis server (ASSEDA)"

In Mucaki et al. (2013), we described a method of predicting the overall strength of an exon by calculating its total information content ($R_{i,\text{total}}$) from the sum of the R_i values of its donor and acceptor splice sites, adjusted for their gap surprisal (the self-information of the distance between the two sites). Differences between total information contents of an exon ($\Delta R_{i,\text{total}}$) are predictive of the relative abundance of these exons in distinct processed mRNAs.

Splice sites altered by mutations that prevent stable interaction with splicesomes are said to be abolished. Information theory predicts abolition of binding below their minimum binding affinity, $R_{i,\text{minimum}}$, which is empirically derived. This value is slightly above zero bits, the theoretical minimum for binding at equilibrium ($\Delta G = 0$; Schneider 1997). Sites with $R_i < 0$ are unbound, forming stable interactions would be endergonic ($\Delta G > 0$). This raises the question, when predicting the change in exon strength ($\Delta R_{i,\text{total}}$) due to a mutation that inactivates binding, whether mutant sites with varying degrees of negative information content are energetically distinguishable from one another.

The computation of $R_{i,\text{total}}$ sums the R_i values of component binding sites, irrespective of their initial or final strengths. Thus, a mutated site with $R_i << 0$ would result in greater $\Delta R_{i,\text{total}}$ compared to a

site with R_i ~ 0. To assess whether the degree of unfavorable binding should be applied to the exon definition calculation, or if values below 0 bits should be computed similarly to a binding site at equilibrium (Ri ~ 0), we reevaluated experimentally validated natural and regulatory splicing mutations in our paper with both approaches. In Table 1 of our study, $R_{i,total}$ was calculated for 10 variants from Supplementary Table 2, both including and excluding the negative impact (ie. R_i <0 vs R_i = 0) of inactivated splice sites. Mutation #2 [*ADA*:g.43249658G>A] abolishes a natural donor site, from 8.8 to -9.9 bits. When applying the full decrease in strength ($\Delta R_{i,total}$: -18.7 bits), the natural exon decreases from 21.0 to 2.3 bits. When it is not applied, the change is significantly smaller (21.0 -> 12.2 bits; $\Delta R_{i,total}$ -8.8 bits). When a weak natural site is abolished, the difference in $\Delta R_{i,total}$ can be quite small (9; -14.8 vs -3.1 bits). In one case (#38), the reduction in $\Delta R_{i,total}$ leads to a partially discordant prediction where the abolished natural exon is weaker than the experimentally confirmed activated cryptic exon. This mutation was concordant when including the negative bit value of the mutated natural site.

The impact of mutations in splicing regulatory (SR) factors can also be predicted on ASSEDA, where the R_i of the SR binding site is added to the $R_{i,total}$, as well as a secondary gap surprisal value for the particular SR protein. These sites can also be abolished. But when a SR binding site is no longer active, should the SR gap surprisal still be applied, or is the SR gap surprisal no longer applicable? To investigate, we test mutations from Mucaki et al (2013; Supplementary Table 4) which abolish the splicing enhancer SF2/ASF with and without the SR gap surprisal when R_i of the SR site is < 0 bits. The removal of the gap surprisal for mutation 2 of Suppl. Table 4 leads to a discordant prediction, where the ΔR_i is less than the SR gap surprisal at that distance and therefore the $\Delta R_{i,total}$ is positive. As experimental evidence shows an increase in skipping, it is a discordant prediction. Therefore, we still apply the gap surprisal on both initial and final $R_{i,total}$ when the SR protein of interest is abolished as the site is naturally present and therefore expected for binding. Conversely, when we apply the gap surprisal to the initial $R_{i,total}$ for a splicing factor that is being created, we are essentially applying a penalty for a

site that does not normally exist. Therefore, we no longer apply the SR gap surprisal value to the initial Ri,total in these cases.

Please note that the values found in Table 2's "Gap Surprisal Included when SR is Abolished" columns are slightly different from those reported in Mucaki et al., 2013 (Supplementary Table 4). We've updated the gap surprisal distributions for the following factors: SF2/ASF, SC35 and SRp40. We re-scanned the genome with an updated version of these models, which slightly changed the distributions for SF2/ASF and SC35. SRp40 changed significantly, and now better resembles the other gap surprisal functions. The updated graphs can now be found here: http://splice.uwo.ca/gapsurprisals.html. While this should not significantly affect $\Delta R_{i,total}$, it may affect the initial and final $R_{i,total}$ values.

Table 1: Exon definition of cryptic splicing mutations : Impact of mutations with negative Ri values on natural exons

TableID[1]	Gene	Mutation	Logic: Negative Ri value Considered			Logic: Negative Ri value set to zero bits (Not Considered)			Final $R_{i,total}$ (cryptic exons)	Concordant (C) / Discordant (D)
			$R_{i,total}$ Initial	$R_{i,total}$ Final	$\Delta R_{i,total}$	$R_{i,total}$ Initial	$R_{i,total}$ Final	$\Delta R_{i,total}$		
2	ADA	c.975+1G>A	21.0	2.3	-18.7	21.0	12.2	-8.8	14.9	C/C
3	BRCA1	c.212+1G>A	15.2	-3.4	-18.6	15.2	9.3	-5.9	14.1	C/C
4	BRCA1	c.5340+1G>A	11.9	-6.8	-18.7	11.9	5.3	-6.6	6.9	C/C
9	BRCA1	c.213-2A>G	8.4	-6.4	-14.8	8.4	5.3	-3.1	13.8	C/C
15	BRCA2	c.8633-2A>G	18.6	3.8	-14.8	18.6	8.3	-10.3	15.0	C/C
18	BRCA2	c.8395A>G	-0.6	18	18.6	9.6	18	8.4	13.5	C/C
21	CDKN2A	c.457+1G>A	13.0	-5.7	-18.7	13.0	3.3	-9.7	9.1	C/C
35	IDS	c.880-2A>G	18.8	4.1	-14.7	18.8	6.9	-11.9	10.1	C/C
37	MYBPC3	c.772+1G>A	8.4	-10.3	-18.7	8.4	1.4	-7.0	3.6	C/C
38	MYBPC3	g.47361343A>G	9.6	-5.1	-14.7	9.6	2.8	-6.8	2.6	**C/D**
[1] IDs from Supplemental Table 2 of Mucaki et al., 2013										

Table 2: Exon Definition of Mutations of SR Sites with/without Impact of Gap Surprisal for Abolished Sites

TableID[1]	Gene	Mutation	Logic: Gap Surprisal Included when SR is Abolished			Logic: Gap Surprisal Not Included when SR is Abolished			SR Protein	Concordant (C) / Discordant (D)
			$R_{i,total}$ Initial	$R_{i,total}$ Final	$\Delta R_{i,total}$	$R_{i,total}$ Initial	$R_{i,total}$ Final	$\Delta R_{i,total}$		
1	SMN1/2	g.70247773C>T	16.4	12.1	-4.3	17.8	13.7	-4.1	SF2/ASF	C/C
2	PAH	g.103237478T>C	9.9	5.6	-4.3	9.9	11.0	1.1	SF2/ASF	**C/D**
4	ACAT1	g.108014720C>T	12.9	7.1	-5.8	12.3	9.1	-3.2	SF2/ASF	C/C
6	BEST1	g.61724926T>C	19.5	15.2	-4.3	19.5	17.7	-1.8	SF2/ASF	C/C
[1] IDs from Supplemental Table 4 of Mucaki et al., 2013										