Humu-2012-0484

Informatics

Supporting Information for this preprint is available from the *Human Mutation* editorial office upon request (humu@wiley.com)

Prediction of mutant mRNA splice isoforms by information theory-based exon definition

Eliseos J. Mucaki[1], Ben C. Shirley[2], and Peter K. Rogan[1,2§]

Departments of [1]Biochemistry and [2]Computer Science, Western University, London, ON

[§]Corresponding author:

Peter K. Rogan

Dept of Biochemistry

Schulich School of Medicine and Dentistry

Western University

London ON N6A 2C1

519-661-4255

E-mail: progan@uwo.ca

Funding: Canadian Breast Cancer Foundation, Natural Sciences and Engineering Research Council of Canada, Canada Foundation For Innovation, Canada Research Chairs, Compute Canada, Western University, and Cytognomix Inc.

Abstract

Mutations that affect mRNA splicing often produce multiple mRNA isoforms, resulting in complex molecular phenotypes. Definition of an exon and its inclusion in mature mRNA relies on joint recognition of both acceptor and donor splice sites. This study predicts cryptic and exon skipping isoforms in mRNA produced by splicing mutations from the combined information contents (R_i, which measures binding site affinity) and distribution of the splice sites defining these exons. The total information content of an exon ($R_{i,total}$) is the sum of the R_i values of its acceptor and donor splice sites, adjusted for the distance separating these sites, ie. the gap surprisal. Differences between total exon information contents ($\Delta R_{i,total}$) are predictive of the relative abundance of these exons in distinct processed mRNAs. Constraints on splice site and exon selection are used to eliminate non-conforming and poorly expressed isoforms. Molecular phenotypes are computed by the Automated Splice Site and Exon Definition Analysis server (ASSEDA; http://ossify.sg.csd.uwo.ca). Predictions of splicing mutations were highly concordant (85.2%; n=61) with published expression data. *In silico* exon definition analysis will contribute to streamlining assessment of abnormal and normal splice isoforms resulting from mutations.

Keywords

Exon definition, mRNA, cryptic splicing, gap surprisal, information theory

Background

mRNA processing mutations, which are responsible for a wide range of human diseases [Divina et al., 2009], alter the abundance and/or structures of mature transcripts. These mutations often occur proximate to exon/intron boundaries, but are frequently found at other sequence locations within introns or exons. Mutations which abolish or weaken recognition of natural splice acceptor or donor sites often produce transcripts lacking corresponding exons or activate adjacent cryptic splice sites of the same phase. Alternatively, mutations activate cryptic splice sites whose strength exceeds existing natural sites elsewhere in the unspliced transcript. The resultant molecular phenotypes may include isoforms with altered exon length and, in some instances, reduced or leaky expression of normal isoforms. We propose an approach based on information theory to predict the structures and approximate abundance of the output molecules generated directly or indirectly by splicing mutations.

Berget's exon definition model [Berget, 1995] provides a mechanism for recognizing multiple small exons against a background of considerably larger intronic sequences. Accurate exon recognition can be complicated by pseudo-exonic structures present in introns that mimic natural exon structures [Ibrahim et al., 2005]. To discriminate between these structures, accurate spliceosomal recognition relies on relatively high affinities of the recognition sequences in natural exons and the presence of other splicing regulatory elements. Exons and adjacent introns also contain splicing enhancer (ESE, ISE) and silencer (ESS, ISS) sequences close to or overlapping constitutive splice sites, which may assist or suppress exon recognition through interactions with additional proteins [Berget, 1995; Graveley and Maniatis, 1998]. Recognition of an exon may therefore depend to some degree on the combined effects of each of these

proteins [Goren et al., 2010], however the factors that recognize the acceptor and donor splice sites are often sufficient [Hwang and Cohen, 1997].

Information theory can be used to measure the conservation of nucleotide sequences bound by individual proteins or protein complexes. In splicing, information theory-based models of donor and acceptor splice sites reveal which nucleotides are permissible at both highly conserved and variable positions in individual sites [Schneider, 1997; Robberson et al., 1990]. These sequences are recognized prior to intron excision, and this recognition is related to the strength of the spliceosome-splice site interaction [Berget, 1995]. The strengths of spliceosome-splice site interactions are related to the corresponding individual information content, R_i, of the RNA sequence [Rogan et al., 1998]. We propose that an exon can be defined by the cumulative R_i values of each of these distinct binding sites contributing to exon recognition ($R_{i,total}$), based on the fact that information is additive for independent sources of uncertainty [Jaynes 1957].

In contrast with splice sites across an intron, cognate pairs of donor and acceptor splice sites from the same exon tend to be separated by a narrow range of distances in the unspliced transcript (the most common internal exon size is 96 nt). Single exon recognition tends to be constrained by preferred distances between the U2 and U1 splicesomal binding sites across the same exon [Hwang and Cohen, 1997]. We previously presented a model to define exon sequences that incorporates the information contents of both splice sites and preferences for certain exon lengths of all natural exons [Rogan, 2009]. A general approach was used that minimized entropy of a pair of binding sites separated by a variable length interstitial sequence. Given a set of exons flanked on either side by 100 nucleotides (nt) intron sequences, the most accurate model (99% correctly detected exon boundaries) was derived by bootstrapping sets of 4000 sequences with left (acceptor) and right (donor) sites of 31 (9.7 bits) and 15 nts (8.1 bits) in

length. In the present study, we ensure that pairs of splice sites of opposite polarity are derived from the same exon by incorporating the surprisal function ([Tribus, 1961]; also termed self-information by Shannon [Cover and Thomas, 2006]), which corrects for both frequent and uncommon or rare inter-site distances that are unlikely to form an exon. This is based on the observation that long internal exons are recognized inefficiently [Robberson et al., 1990], though they do occur (1115 known internal exons > 1000 nt; [Bolisetty and Beemon, 2012]). The total exon information content ($R_{i,total}$) is significantly reduced by this gap surprisal value, if either the predicted exon length is suboptimal or splice site pairs are derived from different exons, but is nearly unchanged for common exon lengths.

Here, we analyze splicing mutations according to changes in total exon information. Since ASSEDA predicts changes in expression relative to wild type levels, it is assumed that the gene is expressed in the tissue being assayed, and that all splicing regulatory factors required for its expression are present in the relevant cell type in which the mutation is analyzed. Multiple splice isoforms may be produced from activated cryptic splice sites of the same polarity as the mutated splice site. The exons with highest information contents have the highest abundance, analogous to previous analyses of individual splicing mutations [Rogan et al., 1998]. Comparison of $R_{i,total}$ values for different exons in normal and mutant sequences are used to estimate their relative inclusion or exclusion in mature mRNA. Information-theory based exon definition models generate testable predictions of splice isoforms and can reveal splice isoforms that have not been previously described.

Materials and Methods

Exon Information Content

We derive the information content of a spliced exon from the cumulative contributions of the nucleic acid binding sites recognized by the spliceosomal machinery and the distribution distances separating binding sites within the same exon. Given a set *S* of *n* different binding sites in an exon, each of which are recognized by *m* different proteins, then S = {x_n, where $1 \leq n \leq m$}. The total information content, I_s, of all sites in S is

$$I_S = \sum_{n=1}^{m} R_i(x_n) \dots\dots\dots bits \ (1)$$

The information content of each site, $R_i(x_n)$ (measured in bits) is derived from a weight matrix (R_{iw}) representing the sequence conservation of each nucleotide in that sequence. The derivation has been presented previously [Schneider, 1997; Rogan et al., 1998].

The information contents of each set of binding sites are modified to account for the probability that these sites occur within the same exon. This requires a gap surprisal term that depends on the transcriptome-wide distribution of the lengths separating them. The gap surprisal is applied to a set of sites within the same exon. Each combination of different binding proteins ($x_1, x_2 \dots$) is described by a distinct distribution. The number of different, unordered pairs of binding sites, given n different sites, correspond to $\binom{n}{2}$, different gap surprisal terms. The gap surprisal for two binding sites (x_p and x_q), separated by L nucleotides $g(L_{pq})$, is

$$g(L_{pq}) = -log_2\left(P(L_{pq})\right) bits \dots\dots\dots (2)$$

where L_{pq} is the distance between x_p and x_q sites. We calculate $P(L_{pq})$ from experimentally validated inter site distances from human genes. Equation (4) signifies that the greater the distance between two sites, the larger the gap surprisal (greater penalty) will be, resulting in a biological reduction of larger than consensus exon length occurrence.

Denoting $G(L_s)$, the total gap surprisal of $\binom{n}{2}$different pairs of sites in set S,

$$G(L_s) = \sum_{1 \le p \le n} \sum_{p < q \le n} g(L_{pq}) \dots\dots\dots (3)$$

The total information content ($R_{i,total}$) is defined by combining Equations (1) and (3),

$$R_{i,total} = \sum_{n=1}^{m} R_i(x_n) + \sum_{1 \le p \le n} \sum_{p < q \le n} g(L_{pq}) \dots\dots\dots (4)$$

To calculate the $R_{i,total}$ of an internal exon, we consider the simplest case with a constitutive set of donor and acceptor splice sites (n=2). We define x_1 as the acceptor and x_2 to be the donor site. x_n has been extended to incorporate other types of binding sites, including splicing regulatory factors, SF2/ASF (*SRSF1*) and SC35 (*SRSF2*), that modify exon recognition. These factors act to enhance splicing when the recognition sites are located within exons (ESE) and repress splicing (ISS) if occurring in the intron adjacent to constitutive splice sites [Lim et al., 2011]. The sign of this term in $R_{i,total}$ is positive if the binding site is exonic and negative if it is intronic. The pairwise distribution of functional binding sites in the transcriptome is required to determine $g(L_{pq})$. For the first and last exons of a gene, $R_{i,total}$ is the sum of the R_i value of the single splice site in that exon adjusted for $g(L)$, where L is exon length, and is based on length distributions for the corresponding terminal exons. The sign of the $g(L_{pq})$ term is negative for exonic locations (ESE) and reversed for intronic sites (ISS). We calculate and compare $R_{i,total}$ values for the strengths of the constitutive splice sites in an exon prior to and after a mutation (details are provided in Supp. Methods). Isoforms with either different donor or acceptor sites may be predicted for each mutation. Because the lengths of these isoforms may vary considerably from each another, analysis of compound mutations at different gene locations has been disabled in molecular phenotypic analysis. The exon definition algorithm requires at least one natural site from an exon to be contained in the predicted isoforms; thus, cryptic or pseudo-

exons activated by intronic mutations are not reported. Nevertheless, the point mutation analysis capability of the ASSA server may detect these sites.

Populating the annotation database

The ASSEDA server is based on human genome reference sequence hg19 (GRCh37), GenBank and RefSeq cDNA accessions (downloaded from genome.ucsc.edu, July 2011), and SNP (dbSNP 135) tables. Genome-wide information weight matrices for automatically curated acceptor (n=108,079) and donor (n=111,772) splice sites (acceptor_genome and donor_genome, respectively; described in [Rogan et al., 2003]), were used in the $R_{i,total}$ calculation. The reference sequence was scanned with these matrices to determine the R_i's of known natural splice sites and used to populate a MySQL database table (ALL_RI, modified from the *all_mRNA.txt* and the *refSeqAli.txt* from the UCSC genome browser).

The frequencies of different exon lengths occurring in the RefSeq database were determined for the gap surprisal calculation. Gap surprisals were normalized, based on highest frequency distance separating splice sites of opposite polarity, which was assigned $G(L_s) = 0$ bits. Separate distributions were compiled, respectively, for first, internal, and last exons, and stored in separate database tables. The start and end positions of first and last exons were relaxed to include any coordinate within a 200 nt window once in order to avoid duplication of exons in the gap surprisal calculation (this accounts for variation in the methods used to generate the cDNAs that are mapped onto the genomic sequence).

Incorporating models of splicing regulatory sequences into $R_{i,total}$

The impact of mutations in ISS or ESE's at SF2/ASF or SC35 binding sites on constitutive splicing can be predicted by selecting the option to incorporate this term into the $R_{i,total}$ computation (on the Advanced Options page). Information weight matrices, $R_i(b,l)$, for

SF2/ASF, SC35, SRp40 (*SRSF5*), and SRp55 (*SRSF6*) were derived from previously published data [Liu et al., 1998; Liu et al., 2000; Smith et al., 2006], and supplemented by experimentally-validated binding sites curated from subsequent publications (sequence logos and weight matrices are available in Supp. Table S1). After scanning the reference genome and locating all predicted binding sites with the SF2/ASF and SC35 R_i(b,l) matrices, their distributions, $g(L_{pq})$ were determined separately for intronic and exonic binding sites in closest proximity to adjacent constitutive splice sites. In computing $R_{i,total}$, the strongest pre-existing splicing regulatory site affected by the mutation (with the highest initial R_i value) is selected by the server, unless the final R_i value of a second site surpasses that of the pre-existing site upon introduction of the mutation (then the second site is reported). The gap surprisal table that is applied is based on which splicing regulatory protein is selected, and the location of the site.

Description of server

The ASSEDA server retains ASSA's capability to analyze changes in individual information content, but also predicts molecular phenotypes based on changes in $R_{i,total}$. ASSEDA and ASSA use the same interface to input sequence variants: HUGO-approved gene symbols, HGVS mutation nomenclature, and dbSNP identifiers, sequence window range around the mutation coordinate, and selected weight matrices as input (Figure 2a; [Nalla and Rogan, 2005]). Mutation syntaxes are then translated into equivalent Delila instructions [Schneider et al., 1984]. The ASSEDA server contains a new option that allows analysis of either splice site information, molecular phenotype based on exon information, or both (for system architecture and program flow diagrams, see Supp. Figures S1 and S2). Upon submission of a mutation, a set of GenBank accession identifiers (ID) corresponding to mRNAs associated with the submitted gene is suggested. These IDs now include mRNAs in the NCBI Reference Gene Sequence

database (http://www.ncbi.nlm.nih.gov/RefSeq/; RefSeq). The IDs are differentiated according to GenBank accessions (in green) and RefSeq ID's (in blue). The longest mRNA accession number is selected by default, and the genomic structure of each RefSeq accession is hyperlinked to the selected ID.

The window range is a primary determinant of the number of potential isoforms reported, since larger windows capture additional potential cryptic splice sites. The feasibility of exon formation is assessed by their $R_{i,total}$ values, and by using rule-based filters to ensure that only likely isoforms are reported. These eliminate cryptic exons with misordered splice sites, overlapping donor and acceptor sites, internal exons less than 30 nt in length [Dominski and Kole, 1991], predicted splice isoforms with <1% of exon inclusion relative to the mutated, natural exon strength ($\Delta R_{i,total}$ between two isoforms < 6.65 bits). The server highlights isoforms with negligible expression when their $R_{i,total}$ values are at least 1 bit below that of the $R_{i,total}$ of the mutated exon. Tabular results can be sorted by column and is paginated, which is particularly helpful for mutations in which numerous cryptic exons are predicted. All rows with potentially expressed isoforms are uncolored, but the wild type exon is indicated in red. Splice isoforms that either cannot be expressed or minor forms (<5% of the major expressed form) that would not be detectable experimentally are, by default, filtered out. Without filtering, rows containing non-functional or minimally expressed predicted isoforms are highlighted in distinct colors: (1) Exons with misordered splice sites (light blue), (2) Potential cryptic exons with lower $R_{i,total}$ values than normal or mutated exon (≤1% predicted expression; pink). (3) Isoforms with both incorrect splice site order and have low $R_{i,total}$ values (green). The minimum reportable $R_{i,total}$ value may also be selected using horizontal sliding scale bar which filters out potential exons below this threshold.

The server draws a set of box glyphs (Figure 3a) depicting a set of exon structures and lengths of potential isoforms that are most likely to form exons. The index of each isoform and its $R_{i,total}$ value are also indicated next to each structure as well as the approximate chromosome coordinates of the normal and cryptic exons.

The server also generates separate custom tracks of each isoform and uploads them to the UCSC genome browser, where they are displayed in the context of the exon containing the mutation as an embedded window within ASSEDA. Each isoform is spectrally color coded based on $R_{i,total}$ content.

Relative abundance of predicted splice isoforms

The server also displays pairwise differences in relative abundance for all predicted isoforms. The relative abundance or fold change in binding affinity of a single binding site is $\leq 2^{\Delta Ri}$, where ΔR_i is the difference between the respective individual information contents of wild type and mutant type of the site [Schneider, 1997]. We extend the idea of relative abundance of single binding site to multiple binding sites by comparing their $R_{i,total}$ values. Suppose n and m are two alternative splice isoforms sharing at least one common splice site and their respective total information contents are $R_{i,total(n)}$ and $R_{i,total(m)}$. If $R_{i,total(n)} > R_{i,total(m)}$, then the relative abundance of n over m will be $\leq 2^{\Delta Ri,total(nm)}$, where $\Delta R_{i,total(nm)} = R_{i,total(n)} - R_{i,total(m)}$. Relative transcript abundance is displayed as a multidimensional graph (with *scatterplot3d*, an R package for visualization of three dimensional multivariate data). The graph shows predicted pairwise differences in exon abundance (Z axis) of the X axis isoform relative to the one on the Y axis, both before (left graph) and after mutation (right graph). The isoform designations correspond to those shown in the other molecular phenotype tabs.

Results

Exon definition by information analysis of functional exons

Gap surprisal values of all exon lengths were determined from their respective frequencies in the exome of all RefSeq genes. The gap surprisal penalty was then normalized so that the most common internal exon length (96 nt; n=172,250) was zero bits, by subtracting a constant value of 6.59 bits (its $\log_2$ frequency). Less frequent exon lengths were scaled to this value by subtracting this constant from their respective gap surprisal values. First and terminal exons are, respectively, missing either a donor or an acceptor splice site, and exhibit a broader range of exon lengths. Separate gap surprisal distributions were computed for these exons. The most frequent first and last exons were, respectively, 158 (n=23,471) and 232 (n=21,261) nt in length, corresponding to gap surprisals of 7.8 and 9.4 bits, respectively. $R_{i,total}$ values were > 0 bits for 98.9% of internal exons, 95.3% of first exons, and 93.1% of last exons (Figure 1). Although inclusion of the gap surprisal term resulted in fewer false positive splice isoforms [Robberson et al., 1990; Dominski and Kole, 1992], a slightly higher proportion of first and last exons had negative $R_{i,total}$ values. Since most of these splice sites in these exons exhibited positive R_i values (72% of first, 87% last exons), the negative $R_{i,total}$ values may be the result of other unknown factors contributing to recognition of these exons not accounted for, or to suboptimal gap surprisal functions.

Interpretation of splicing mutations by exon definition analysis

The Automated Splice Site Analysis server (https://splice.uwo.ca), which computes information changes at individual splice sites, has been upgraded to additionally analyze changes in total exon information to the Automated Splicing and Exon Definition Analysis server (ASSEDA; http://ossify.sg.csd.uwo.ca). Mutations are input through a web interface [Nalla and

Rogan, 2005] with a predefined sequence window of sufficient length to encompass the spectrum of potential splice sites activated by a mutation. The $R_{i,total}$ values of prospective exons within this window are computed with the Molecular Phenotype option on the main page of the server.

The server accepts and reports genomic DNA coordinate notations (as well as the IVS notations) supported by the most recently proposed Locus Reference Genomic (LRG) variant format [Dalgleish et al., 2010] which is consistent with HGVS recommendations. A typical molecular phenotypic prediction is indicated in Figure 2 (*BRCA1* IVS20+1G>A or HGVS designation chr17: g.41209068C>T; Supp. Table S2, Mutation #4). The tabular results indicate genomic coordinates of donor and acceptor sites, their relative distance from the closest natural site, and the change in R_i for these sites. Each row indicates $R_{i,total}$ both before and after mutation for a different set of exon boundaries corresponding to a distinct predicted isoform. Predicted isoforms are sorted according to these values, whose fold differences in binding affinity are $\leq 2^{\Delta R_{i,total}}$ [Schneider, 1997].

Initially, 20 potential isoforms are found for this mutation, of which those with the highest $R_{i,total}$ values and the affected natural exon are indicated (Figure 2b). Based on the mechanism of exon recognition and the $\Delta R_{i,total}$ values, only a subset of these indexed isoforms is likely to be expressed. Splice site polarity is specified such that a functional acceptor splice site cannot occur downstream of a natural donor splice site to define an exon, and vice versa [Berget, 1995]. The server eliminates exons with misordered splice sites, removing many false positive splice isoforms which do not conform to the natural mRNA splicing mechanisms. Pairs of splice donor and acceptor sites that either overlap each other are also not considered as potential exons [Nalla and Rogan, 2005; Robberson et al., 1990]. Predicted low abundance natural and cryptic isoforms with undetectable expression (Figures 2b and 2c) are also filtered out.

The structures and lengths of each potential isoform (natural, cryptic, skipped) are also displayed in a separate tab (Figure 3a). The central exon affected by the mutation is drawn to scale, however flanking intron sequences are condensed for presentation. In the example above, the exon 20 donor site in chr17: g.41209068C>T ($R_{i,total}$ 11.9 -> -6.6 bits) is inactivated and an corresponding isoform with exon skipping is shown. The relative abundance (Z axis) of different pairs of indexed isoforms (X and Y) before (Figure 3b) and after (Figure 3c) mutation also predicts a number of cryptic isoforms. Isoform 1 uses a pre-existing donor 87 nt downstream that is at least 13,307 (i.e. $\leq 2^{13.7\ bits}$) fold more abundant than the mutated exon, but would not normally be detected because it is 32 fold ($\leq 2^{5.0}$) less abundant than the normal exon. mRNA analyses have shown that this mutation results in both cryptic and skipped splice forms [Sanz et al., 2010], however isoform 4 which contains 133 of intronic sequence (Figures 2c and 3a), was not detected.

The molecular phenotype analysis panel also displays wildtype and predicted mutant isoforms as individual BEDGRAPH custom tracks on the UCSC genome browser. The combination of these predictions with other browser tracks (ie. sequenced mRNAs, ESTs, and known SNPs within a gene) can distinguish mutant from naturally occurring alternative splice forms or previously described variants that may be associated with a suspected mutation.

Validation

To assess whether the proposed model of exon definition produced results consistent with observed mutant spliced products, we evaluated a series of reported splicing mutations for which end-point (Supp. Table S2) and quantitative (Supp. Table S3) expression studies had been performed. Mutant isoforms and relative abundance were predicted for splicing mutations in the nephropathiccystinosis gene product (*CTNS*), cardiac myosin binding protein C (*MYBPC3*),

fumarylacetoacetate hydrolase (*FAH*), the Ellis van Creveld syndrome gene product (*EVC*), thalassemia (*HBB*), coagulation factor XII (*F12*), adenosine deaminase (*ADA*), cyclin-dependent kinase inhibitor 2A (*CDKN2A*), iduronate 2-sulfatase (*IDS*), phenylalanine hydroxylase (*PAH*), paired-like homeodomain 2 (*PITX2*), phosphomannomutase 2 (*PMM2*) and early onset breast cancer (*BRCA1* and *BRCA2*).

Detailed analysis of one of these mutations illustrates the importance of the gap surprisal function and of post-hoc filtering out misordered and weak exons. A splicing mutation, *CDKN2A*: IVS2+1G>T (g.21970900G>T; Supp. Table S2, #21) abolishes a natural donor site, and numerous potential cryptic donor sites are unmasked (n=61 potential exons). After filtering non-conforming or negligibly expressed isoforms, 8 prospective exons of sizes 133, 234, 166, 680, 435, 973, 742, 515 and 308 nt remain. Application of the gap surprisal term reduces the number of exon combinations. Exons of length 308, 435, 515, 680, 742 and 973 nt are much less common, and the gap surprisal penalties are correspondingly larger (5.9, 7.1, 8.3, 8.6, 8.6, 9.9 bits respectively), significantly lower their $R_{i,total}$ values. The three highest predicted 133, 166 and 234 contain correctly ordered splice sites (with penalties of 0.8, 1.4 and 3.0 bits respectively), two of which have previously been reported [Rutter et al., 2003]. Similarly, *PMM2:* IVS3-1G>C (Supp. Table S2 #41) mutation predicts 263 possible exon structures, 35 isoforms when the gap surprisal penalty is applied, and only 4 after filtering. Clearly, one of the strengths of the present approach is that it effectively eliminates improbable or poorly expressed isoforms.

The accuracy of information-based exon definition prediction was compared with expression data for these mutations (Supp. Table S2). There is generally good concordance with documented splice isoforms. All mutations either weakened or inactivated natural splice sites or

activated cryptic splice sites as expected. Among the mutations tested, 36 of 41 mutations were completely consistent with published results, including cryptic isoforms. In some instances, a reported cryptic exon was predicted, but was not determined to be the most abundant splice isoform.

Information analysis correctly predicted several types of splicing abnormalities in different genes. There were 31 mutations which resulted in formation of one or more cryptic exons (Supp. Table S2). Exons using these cryptic splice sites were predicted for 28 of the 31 mutations, 20 of which had the highest $R_{i,total}$ values. The other 8 mutations were ranked these cryptic splicing isoforms among the highest 6 in abundance, save one (Supp. Table S2 #10). Complete intron retention was reported for one mutation (#40), while 9 mutations were found to result in exon skipping only (#1, 7, 8, 11, 14, 23, 26, 37 and 41). Previously, we have shown that large changes in ΔR_i can result in exon skipping as well as leaky splicing [Rogan et al., 1998]. All of these mutations decreased $R_{i,total}$ of the natural exon, although in one case, the extent was marginally below significance (#14; 0.8 bits). Exon skipping was reported for mutations #7, 8, 23 and 24 rather than reduced levels of exon inclusion suggested by the exon definition analysis. These mutations reduced the predicted exon abundance by 9 to 23 fold relative to the normally spliced product. This level of expression is close to the detection limit of a minor cryptic splice isoform for most analytic methods [Rogan et al., 1998], and may explain why only exon skipping was documented for these mutations [Macias-Vidal et al., 2009; Tompson et al., 2007; Claes et al., 2002; Claes et al., 2003]. Additionally, the discrepancy could simply be due to the limitations of the *in vitro* analyses used.

Exon definition analysis of the remaining mutations showed partial discordance to published mRNA evidence. In 3 cases, the reported cryptic site used had an $R_i < 0$ bits (#10, 15,

32). Mutation #27, $R_{i,total}$ of the natural and the proven activated cryptic site does not quite reach the threshold for a functional site defined by information theory. In the final case (#22), the creation of a cryptic donor is predicted (2.7 bits), but the resultant 425 nt exon is not observed ($R_{i,total} < 0$).

In simple molecular recognition systems, information theory based methods predict binding site affinity changes, as R_i is a direct related to binding affinity. However, while splicesomal binding affinity is crucial to constitutive splice site recognition [Berget, 1995], other factors, such as interactions with regulatory factors can influence splicing outcomes [De Conti et al., 2012]. The present model attempts to account for the effects of these factors on exon inclusion by incorporating the contributions of multiple binding sites. Predicted isoforms were compared to detected mRNA species for 8 splicing mutations that have been assessed with quantitative methods (predominantly quantitative RT-PCR; Supp. Table S3). ASSEDA correctly predicted a decrease in wildtype splicing in most instances. Cryptic exons detected (Supp. Table S3: #2, 4, 5, 8) were also correctly predicted, and were ranked highest based on $R_{i,total}$ values (using a window size 200 nt) in all but one case (in which the cryptic isoform ranked second; Supp. Table S3: #8). Decreases in the predicted strength of the natural exon was accurately predicted for 6 of 8 mutations (Supp. Table S3: #1-4, 7, 8). The mutation c.653A>G (Supp. Table S3: #5) was found to abolish normal splicing of exon 6 of *PCCB*, although the exon was predicted to have some residual splicing (82.3% decrease in binding efficiency). Conversely, g.6622214G>C (Supp. Table S3: #6) in *NSUN2* was predicted to abolish normal splicing (99.97% decrease in binding efficiency), but the a low level of the splice form was still detected (5% of control expression). When both cryptic isoforms and exon skipping occur (Supp. Table S3: #2 and 8), the ratio between the two splice forms does not seem to relate directly to predicted

strength changes. In both cases, splicing of the normal exon is abolished and a cryptic isoform (where $R_{i,total}$ < initial $R_{i,total}$) appears, but they differ in the ratio of exon skipping to cryptic exon expression. This may be associated with the instability of mRNAs with frameshifted cryptic isoforms (Supp. Table S3: #4).

Impact of ESE/ISS Elements

Elements recognized by splicing regulatory proteins, SF2/ASF, SC35, SRp40, SRp55, and hnRNP-H (*HNRNPH1*), can now be analyzed with ASSEDA, however these matrices are based on many fewer sites (usually <50), and the R_i values may not be as accurate as constitutive splice sites, especially at the low end of the distribution. The server computes R_i values of any of these individual sites and can incorporate mutations at either SF2/ASF or SC35 sites into the $R_{i,total}$ computation. Since a mutation can affect multiple predicted sites, the site with the highest R_i value altered by the mutation is analyzed, unless a second cryptic site is strengthened resulting in final R_i is exceeding that of the original binding site.

A second gap surprisal function, based on the distances between known natural constitutuve sites and the closest predicted splicing regulatory site of the same type, was also applied in the $R_{i,total}$ calculation. Exonic (ESE) and intron (ISS) have independent gap surprisal distributions (Supp. Figure S4). The ubiquity of these splicing regulatory sequences suggested that their predicted distributions would be biased towards shorter inter-site distances, however there were distinct preferences for certain distances. 17.2% of all exonic SF2/ASF sites were separated by 4nt from a natural splice site (n=562,786; comparatively, all other distances between 0-10nt range from 1.5-4.4% in frequency). The most common intronic SF2/ASF sites were 1, 3 and 5 nt from the natural site (9.3%, 7.1% and 10.5% respectively; n= 562,788). The most common SC35 site inter-site exonic distances were 0, 4 and 7nt (9.5%, 6.5%, 6.6%

respectively) and intronic distances were spaced 1 and 2 nt from the splice site (9.9% and 9.5%). In all cases, frequency decreased with increased inter-site distance. The distribution of predicted SRp40 distances showed no distance bias; there was a gradual inverse relationship between frequency and distance from the natural site (maximum frequency was < 0.1 % of the sites).

To assess the effect of including SC35 and SF2/ASF sites in the exon definition model, we evaluated 12 reported mutations/variants in either SF2/ASF or SC35 sites that were reported to affect splicing at adjacent splice sites (Supp. Table S4). Eight of 12 predictions of ASSEDA were concordant with the published results (Supp. Table 4 mutations #1-4, 6, 9 and 11 are predicted to weaken splicing and lead to exon skipping; #10 strengthens an intronic SF2/ASF site and activates a cryptic donor). A single nucleotide difference between *SMN1* and *SMN2* (c.840C>T) is known to alter an SF2/ASF exonic site, resulting in skipping of exon 7 in *SMN2* [Cartegni and Krainer 2002]. The SF2/ASF variant in *SMN2* reduces $\Delta R_{i,total}$ of exon 7 in SMN2 by 5.7 bits relative in *SMN1*, corresponding to a 52 fold difference in exon recognition, consistent with skipping of this exon in *SMN2* (Supp. Table S4: #1).

Observed alterations of splicing regulatory sites were not predicted in 3 cases (Supp. Table S4: #5, 7 and 12). For two other mutations, SF2/ASF or SC35 sites were affected, (Supp. Table S4: #3 and 8) but interpretation was confounded by concomitant changes in the opposite direction to SRp55 splicing regulatory sequences (the effect of mutation #3 is predicted with ASSEDA's SRp55 model).

Discussion

We have designed and implemented a novel approach to predict the molecular phenotype of a splicing mutation, producing a probable set of splicing isoforms expressed in mutation

carriers. The system is based on information theory-based methods that accurately quantify binding site affinity [Schneider, 1997; Rogan et al., 1998]. Non-expressed or very low expression exons are filtered out by correcting for suboptimal exon lengths and eliminating incorrectly ordered splice sites. The use of gap surprisal to correct for distances between required sequence features has been previously validated for other types of binding sites [Shultzaberger et al., 2001]. Exon information ($R_{i,total}$) is computed by the ASSEDA server. The backend databases consist of current human genome sequences (hg19/GRCh37) and gene annotations of exon coordinates, gene names and mRNA accession numbers, including NCBI Refseq entries (and dbSNP 135). Mutation entry currently supports c. and g. notation, as well as the deprecated IVS-based mutation description.

We implement a simple model for exon definition based on constitutive splice sites, although the theory for extensible framework for incorporation of multiple splice site recognition sequences is derived. Exon definition-based predictions were compared to known splicing mutations with published mRNA studies, and these predictions were found to be highly concordant (Supp. Table S2). These mutations were sourced from our previous publications so that information theory based modelling of individual splice sites could be compared with exon definition [Rogan et al., 1998; Mucaki et al., 2011].

The exon definition models imply that rare exons (regardless of length) will have large gap surprisal penalties. This is supported by the fact that, for exons beyond a few hundred nucleotides, the penalty function is increases with length until it asymptotes at exon lengths present once in the genome. The significant gap surprisal penalties for long exons raise the question as to how well the model performs at the extreme lengths to correctly distinguish

natural from decoy exons. The model fails if the contributions of the gap surprisal term exceed the R_i values of both natural splice sites. In fact, this is generally not the case.

To assess the ability of the server to predict naturally occurring large exons, 8 large internal exons in genes *BRCA1-ex11*, *BRCA2-ex11*, *TTN-ex253*, *JARID2-ex7*, *KLHL31-ex2*, *C6orf142-ex4* (*MLIP*), *VCAN*-ex8 and *C17orf53-ex3* were evaluated using ASSEDA (Supp. Table S5). Despite the large (> 10 bit) gap surprisal penalties, the $R_{i,total}$ values for each of these exon was still exceeded 0 bits. This can be attributed to their strong donor and acceptor sites, which appear to be essential for large exon recognition ([Bolisetty and Beemon, 2012]; the exception being the donor site of *BRCA1* exon 11 [2.9 bits]). These predicted shorter splice forms are present in *BRCA1* mRNA, however they do not encode full length protein. For example, the highest ranked prospective isoform for *BRCA1*-ex11 was a 118 nt long alternate splice form (NM_007298.3). These large exons were not ranked first, as the $R_{i,total}$ of smaller exons (< 250 nt) tended to have higher overall $R_{i,totals}$ (lower gap surprisal penalty). Larger exons tend to have a higher ratio of enhancers to repressors compared to smaller exons [Bolisetty and Beemon, 2012]. This suggests that gap surprisal function will need to be refined, or contributions of other splicing regulatory proteins will need to be incorporated into $R_{i,total}$ in order to correct the ranking of splice isoforms from long exons.

Although the model we have implemented does predict the preponderance of mutant splice isoforms, it has obvious limitations. For example, misordered splice sites excluded in this model can sometimes be activated through the formation of novel cryptic exons when proximate, pre-existing sites of opposite polarity occur within adjacent introns. The server does not currently handle these situations, which are thought to be uncommon and because of the requirement to include at least one functional site of opposite polarity in the mutated exon.

By default, the current models do not take into account other types of splicing–related binding sites that influence splice site selection, including SR proteins that contribute to exon recognition. This is expected to result in discordant interpretation, (ie. Supp. Table S2 #27), or inaccuracies in predicting the abundance predicted splice isoforms, since $R_{i,total}$ values will be underestimated. Changes in strengths of SF2/ASF and SC35 binding sites can now be included in the $R_{i,total}$ calculation, which can improve its accuracy. However, results might be skewed when the altered splicing regulatory sites do not contribute either positively or negatively to normal splicing. There are many more predicted SF2/ASF and SC35 sites in an exon than there is evidence for each having a specific role in exon definition. Furthermore, only a single splicing regulatory site is currently used in the $R_{i,total}$ calculation, and this assumption may have to be revisited. For example, there are two variants in Supp. Table S4 (#3, 8) which alter either SF2/ASF or SC35 binding sites, but simultaneously change the predicted R_i value of a proximate SRp55 site. This highlights the difficulty in predicting the splicing profile upon mutation of multiple ESE/ISS elements, as the relative contributions of different regulatory elements which support the inclusion of an exon cannot be discerned.

In some cases, the published alternate splice form is detected, but is not the strongest (highest $R_{i,total}$) cryptic exon predicted (Supp. Table S2, #2, 10, 13, 21, 25, 27, 34 and 38). Many of these published mutation phenotypes are based on gel-based RT-PCR analysis, where some cryptic splice isoforms may not have been detected. The gap surprisal term may also be inaccurate, as there are a higher percentage of false negative natural exons with $R_{i,total} < 0$ bits. We suggest using the top 5 splice forms, including cryptic isoforms with the highest $R_{i,total}$ values to perform experimental validation of these predictions.

The gap surprisal contributions for all exon types (first, last and internal exons) are more variable for larger exons (Supp. Figure S3a, c and d). Although stochasticity is discernible, some of this variability can be attributed to genomic selection for these exon lengths. Certain exon lengths may be overrepresented due to paralogous gene duplications which inflate their frequency relative to other exons of similar length; this can significantly lower gap surprisal values. Additionally, exons that retain an integral number of codons are more frequent than similarly-sized exons that change frame (+1 or -1 length), consistent with previous studies of cassette exons in alternatively spliced genes [Clark and Thanaraj 2002; Stamm et al., 2006]. This triplet periodicity is well defined for exons ranging in size from 9 nt to 200 nt (Supp. Figure S3b), but nevertheless is still apparent in exons beyond 500 nt in length. For these reasons, we avoided fitting gap surprisal values to an algebraic function or smoothing of these distributions.

The development of exon definition-based mutation analysis was motivated by the desire to generate predictions that could be directly compared with laboratory expression data. In some instances, these predictions have included strong cryptic exons that have not been previously detected, possibly because the laboratory studies did not directly anticipate the corresponding splice isoforms. The level of concordance we report for previously validated splicing mutations justifies a prospective study of natural and mutant isoforms predicted by the server, in which all predicted cryptic splice isoforms are tested, and if possible, quantified. It should be feasible to implement algorithms to automate design of isoform specific sequence primers for quantitive expression analysis. This feature will close the circle between bioinformatic methods that predict potential splicing mutations in large scale genomic DNA sequence studies and validation with mRNA obtained from the same individuals.

Acknowledgments

This work was sponsored by The Natural Sciences and Engineering Research Council of Canada (NSERC Discovery Grant 371758-2009). We acknowledge all other laboratory members for valuable comments and contributions. We recognize David Wiseman (Department of Computer Science, Western University) for his assistance with systems administration, server setup and maintenance.

Authors' Contributions

PKR derived and developed the methods, and the gap surprisal distributions. EJM and BCS implemented these methods, which involved modifying previous software, creation of new modules, and updating databases. EJM analyzed experimental data for validation. EJM and PKR wrote the manuscript, which has been approved by all of the authors. Conflict of Interest: PKR is the inventor of US Patent 5,867,402 and founder of Cytognomix, which is developing software based on this technology for complete genome or exome splicing mutation analysis.

References

Berget SM. 1995. Exon recognition in vertebrate splicing. J Biol Chem. 270:2411-2414.

Bolisetty MT, Beemon KL. 2012. Splicing of internal large exons is defined by novel cis-acting sequence elements. Nucleic Acids Res. 40(18):9244-54.

Cartegni L., Krainer A.R. 2002. Disruption of an SF2/ASF-dependent exonic splicing enhancer in *SMN2* causes spinal muscular atrophy in the absence of *SMN1*. Nat. Genet. 30:377-384.

Claes K, Vandesompele J, Poppe B, Dahan K, Coene I, De Paepe A, Messiaen L. 2002. Pathological splice mutations outside the invariant AG/GT splice sites of BRCA1 exon 5 increase alternative transcript levels in the 5' end of the BRCA1 gene. Oncogene. 21:4171-4175.

Claes K, Poppe B, Machackova E, Coene I, Foretova L, De Paepe A, and Messiaen L. 2003. Differentiating pathogenic mutations from polymorphic alterations in the splice sites of BRCA1 and BRCA2. Genes Chromosomes Cancer. 37:314-320.

Clark F, Thanaraj TA. 2002. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum Mol Genet. 11: 451-464.

Clavero S, Pérez B, Rincón A, Ugarte M, Desviat LR. 2004. Qualitative and quantitative analysis of the effect of splicing mutations in propionic acidemia underlying non-severe phenotypes. Hum Genet. 115(3):239-47.

Cook KB, Kazan H, Zuberi K, Morris Q, and Hughes TR. 2011. RBPDB: a database of RNA-binding specificities. Nucleic Acids Res. 39:D301-8.

Cover TM, Thomas JA. 2006. Elements of information theory. Wiley-Interscience, Hoboken, NJ: p.748.

Dalgleish R, Flicek P, Cunningham F, Astashyn A, Tully RE, Proctor G, Chen Y, McLaren WM, Larsson P, Vaughan BW, Beroud C, Dobson G et al. 2010. Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med. 2:24.

De Conti L, Baralle M, Buratti E. 2012. Exon and intron definition in pre-mRNA splicing. Wiley Interdiscip Rev RNA. doi: 10.1002/wrna.1140.

Divina P, Kvitkovicova A, Buratti E, Vorechovsky I. 2009. Ab initio prediction of mutation-induced cryptic splice-site activation and exon skipping. Eur J Hum Genet. 17:759-765.

Dominski Z, Kole R. 1991. Selection of splice sites in pre-mRNAs with short internal exons. Mol Cell Biol. 11(12):6075-83.

Dominski Z, Kole R. 1992. Cooperation of pre-mRNA sequence elements in splice site selection. Mol Cell Biol. 12:2108-2114.

Goina E, Skoko N, Pagani F. 2008. Binding of DAZAP1 and hnRNPA1/A2 to an exonic splicing silencer in a natural BRCA1 exon 18 mutant. Mol Cell Biol. 28(11):3850-60.

Graveley BR, Maniatis T. 1998. Arginine/serine-rich domains of SR proteins can function as activators of pre-mRNA splicing. Mol Cell. 1:765-771.

Goren A, Kim E, Amit M, Vaknin K, Kfir N, Ram O, Ast G. 2010. Overlapping splicing regulatory motifs--combinatorial effects on splicing. Nucleic Acids Res. 38:3318-3327.

Hwang DY, Cohen JB. 1997. U1 small nuclear RNA-promoted exon selection requires a minimal distance between the position of U1 binding and the 3' splice site across the exon. Mol Cell Biol. 17:7099-7107.

Ibrahim EC, Schaal TD, Hertel KJ, Reed R, Maniatis T. 2005. Serine/arginine-rich protein-dependent suppression of exon skipping by exonic splicing enhancers. Proc Natl Acad Sci U S A. 102:5002-5007.

Jaynes E. Information Theory and Statistical Mechanics. Phys. Rev. 106, 620–630 (1957).

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mRNA processing defects in human genes. Proc Natl Acad Sci U S A. 108(27):11093-8.

Liu HX, Zhang M, Krainer AR. 1998. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. 12:1998-2012.

Liu HX, Chew SL, Cartegni L, Zhang MQ, Krainer AR. 2000. Exonic splicing enhancer motif recognized by human SC35 under splicing conditions. Mol. Cell. Biol. 20:1063-1071.

Macias-Vidal J, Rodes M, Hernandez-Perez JM, Vilaseca MA, Coll MJ. 2009. Analysis of the CTNS gene in 32 cystinosis patients from Spain. Clin Genet. 76:486-489.

Mucaki EJ, Ainsworth P, Rogan PK. 2011. Comprehensive prediction of mRNA splicing effects of BRCA1 and BRCA2 variants. Hum Mutat. 32:735-42.

Nalla VK, Rogan PK. 2005. Automated splicing mutation analysis by information theory. Hum Mutat. 25:334-342.

Robberson BL, Cote GJ, and Berget SM. 1990. Exon definition may facilitate splice site selection in RNAs with multiple exons. Mol Cell Biol. 10:84-94.

Rogan PK, Faux BM, Schneider TD. 1998. Information analysis of human splice site mutations. Hum Mutat. 12:153-171.

Rogan PK, Svojanovsky SR, Leeder JS. 2003. Information theory-based analysis of CYP219, CYP2D6 and CYP3A5 splicing mutations. Pharmacogenetics. 13:207-18.

Rogan PK. 2009. Ab Initio Exon Definition Using an Information Theory-based Approach. Biochemistry Publications. Paper 10. http://ir.lib.uwo.ca/biochempub/10.

Rutter JL, Goldstein AM, Davila MR, Tucker MA, Struewing JP. 2003. CDKN2A point mutations D153spl(c.457G>T) and IVS2+1G>T result in aberrant splice products affecting both p16INK4a and p14ARF. Oncogene. 22:4444-8.

Sanz DJ, Acedo A, Infante M, Duran M, Perez-Cabornero L, Esteban-Cardenosa E, Lastra E, Pagani F, Miner C, Velasco EA. 2010. A high proportion of DNA variants of BRCA1 and BRCA2 is associated with aberrant splicing in breast/ovarian cancer patients. Clin Cancer Res. 16:1957-67.

Schneider TD, Stormo GD, Yarus MA, Gold L. 1984. Delila system tools. Nucleic Acids Res. 12:129-140.

Schneider TD. 1997. Information content of individual genetic sequences. J Theor Biol. 189:427-441.

Shultzaberger RK, Bucheimer RE, Rudd KE, Schneider TD. 2001. Anatomy of Escherichia coli ribosome binding sites. J Mol Biol. 313:215-228.

Smith PJ, Zhang C, Wang J, Chew SL, Zhang MQ, Krainer AR. 2006. An increased specificity score matrix for the prediction of SF2/ASF-specific exonic splicing enhancers. Hum Mol Genet. 15(16):2490-508.

Stamm S, Riethoven JJ, Le Texier V, Gopalakrishnan C, Kumanduri V, Tang Y, Barbosa-Morais NL, Thanaraj TA. 2006. ASD: a bioinformatics resource on alternative splicing. Nucl Acids Res. 34(suppl 1):D46-55.

Tompson SW, Ruiz-Perez VL, Blair HJ, Barton S, Navarro V, Robson JL, Wright MJ, Goodship JA. 2007. Sequencing EVC and EVC2 identifies mutations in two-thirds of Ellis-van Creveld syndrome patients. Hum Genet. 120:663-670.

Tribus M. 1961. Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications. Van Nostrand, Princeton, NJ: p. 649.

Figure Legends

Figure 1 - Distribution of the $R_{i,total}$ of Annotated Exons

Distribution of the $R_{i,total}$ of Annotated Exons. Histogram of Ri,total values for exons in the RefSeq database are illustrated for first (a), last (b), and internal exons (c). Nearly all internal exons exhibit total information contents exceeding zero bits (98.9%). The gap surprisal functions for first and last exons are not optimized for single splice site exons (4.7% and 7.0%, respectively, have $R_{i,total}$ values below zero bits). The majority of false negative internal exons contain one or both splice sites that are either weak or are not recognized by either the U1- or U2 splicesomes.

Figure 2 - Server input and results for *BRCA1* mutation, chr17:g.41209068G>A

A) User input. The window size of 200 nt increases the number of potential cryptic isoforms reported beyond the default length; B) Resulting table after applying splicing mechanism and exon abundance filters (isoforms 5-14 are not presented due to space limitations). The column headings show key binding site locations, initial and final values and changes in R_i, as well as changes in $R_{i,total}$. The natural or mutated exon is listed in table row 17 (WT in legend below). Cells 1 and 4 (PI) indicate predicted cryptic isoforms with $R_{i,total}$ values comparable or exceeding the strength of the natural exon ($R_{i,total}$ final). Splice isoforms with $R_{i,total} \leq 1$ bit (> 2 fold lower abundance; NE in legend) of the mutated natural exon are minimally expressed and filtered out. Rows 2 and 3 indicate predicted exons with misordered splice sites (NC), and rows 15 and 16 show exons which also would be minimally expressed (NC-NE); C) Only 3 of 35 potential isoforms are reported for the input mutation after filtering on these criteria.

Figure 3 – Structure and Relative Abundance of Predicted Isoforms

Isoforms are depicted graphically according to their exon structures, relative abundance, and custom browser tracks in separate tabs. Isoform numbers in Figure 3 refer to designations in Figure 2c. Panels: (A) The scale above shows the genome coordinates of each of the isoforms. All prospective isoforms (sorted by $R_{i,total}$) are scaled according to their genomic coordinates (above glyphs). The exon skipping splice form is displayed for mutations where resulting $R_{i,total} < 0$ bits; (B and C) Plots indicating predicted pairwise (x,y axes) relative minimum fold differences in abundance (z axis) of each isoform both before and after changes in $R_{i,total}$ due to the mutation. Results are depicted for *BRCA1*, chr17:g.41209068G>A. Panel B shows that the natural wildtype exon (isoform 17) has the highest level of expression. After the mutation (Panel C), isoform 1, which activates a downstream cryptic splice site, is expected to be the dominant splice form. Note that the scale of the Z-axis will change between the panels, depending on the range of $\Delta R_{i,total}$ values resulting from the mutation.