DATABASES

Automated Splicing Mutation Analysis by Information Theory

Vijay K. Nalla[1,3] and Peter K. Rogan[1,2,3]*

[1]*Laboratory of Human Molecular Genetics, Children's Mercy Hospital and Clinics, University of Missouri-Kansas City, Kansas City, Missouri;* [2]*School of Medicine, University of Missouri-Kansas City, Kansas City, Missouri;* [3]*School of Computing and Engineering, University of Missouri-Kansas City, Kansas City, Missouri*

Communicated by A. Jaime Cutticchia

Information theory–based software tools have been useful in interpreting noncoding sequence variation within functional sequence elements such as splice sites. Individual information analysis detects activated cryptic splice sites and associated splicing regulatory sites and is capable of distinguishing null from partially functional alleles. We present a server (https://splice.cmh.edu) designed to analyze splicing mutations in binding sites in either human genes, genome-mapped mRNAs, user-defined sequences, or dbSNP entries. Standard HUGO-approved gene symbols and HGVS-approved systematic mutation nomenclature (or dbSNP format) are entered via a web portal. After verifying the accuracy of input variant(s), the surrounding interval is retrieved from the human genome or user-supplied reference sequence. The server then computes the information contents (R_i) of all potential constitutive and/or regulatory splice sites in both the reference and variant sequences. Changes in information content are color-coded, tabulated, and visualized as sequence walkers, which display the binding sites with the reference sequence. The software was validated by analyzing ~1,300 mutations from *Human Mutation* as well as eight mapped SNPs from dbSNP designated as splice site variants. All of the splicing mutations and variants affected splice site strength or activated cryptic splice sites. The server also detected several missense mutations that were unexpectedly predicted to have concomitant effects on splicing or appeared to activate cryptic splicing. Hum Mutat 25:334–342, 2005. © **2005 Wiley-Liss, Inc.**

KEY WORDS: information theory; mRNA processing; splicing; haplotypes; SNP; genotype–phenotype; HUGO; HGVS; software; mutation nomenclature

DATABASES:
https://splice.cmh.edu(Splicing Mutation Analysis Server)
http://genome.ucsc.edu/goldenPath (UCSC Genome Browser)
www.hgvs.org/mutnomen/ (HGVS Mutation Nomenclature Guidelines)

INTRODUCTION

Accurate interpretation of mutations that alter noncoding, conserved sequence elements in human genes is important for diagnosis and prognosis of inherited or acquired genetic disorders. Mutations that affect mRNA splicing are common in human diseases [Krawczak and Cooper, 1991]. Strengths of one or more splice sites may be altered and, in some instances, concomitant with changes in coding sequences [Richard and Beckmann, 1995; Rogan et al., 1998].

The effects of such mutations can be predicted in silico by information theory [Rogan and Schneider, 1995; Rogan et al., 1998] and predictions confirmed in vitro by experimental studies [Vockley et al., 2000; Rogan et al., 2003; Lamba et al., 2003; Susani et al., 2004]. Changes in the affinity of a protein or protein complex for its cognate binding site can be estimated from the individual information content of the natural and variant sequences [Schneider, 1997]. The individual information content (R_i) can be evaluated for any variant that occurs within a binding site in the genome or transcriptome, given an adequate model or weight matrix ($R_i(b.l)$) based on a set of functional sites recognized by the same protein(s) [Gadiraju et al., 2003].

Traditionally, individual information analysis required coordinate-based instructions to define mutation coordinates in GenBank (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=nucleotide) entries stored in locally installed sequence libraries. We have previously used the Delila system to extract sequences and analyze mutations (http://delila.ncifcrf.gov/~toms/delilaserver.html) [Schneider et al., 1984; Rogan et al., 1998]. Though this software has always been capable of handling large sequence coordinates present in some human chromosomal sequences, there is a tendency to introduce computational and data entry errors when specifying such mutations.

Received 16 September 2004; accepted revised manuscript 6 December 2004.

*Correspondence to: Peter K. Rogan, Ph.D., Laboratory of Human Molecular Genetics, Children's Mercy Hospital and Clinics, 2401 Gillham Rd., Kansas City, MO 64108. E-mail: progan@cmh.edu

Grant sponsor: National Institute of Environmental Health Sciences (NIEHS); Grant number: ES 10855.

DOI 10.1002/humu.20151
Published online in Wiley InterScience (www.interscience.wiley.com).

The profusion of mutation and SNP catalogs has created a demand for resources to rapidly catalog and evaluate the functional impacts of sequence variants throughout the human and other genomes [Horaitis and Cotton, 2004]. Sequence conservation by information analysis has been a successful approach for recognizing nondeleterious variants [Rogan and Schneider, 1995; Ng and Henikoff, 2003] and for distinguishing of milder from more severe mutations [Rogan et al., 1998; Kodolitsch et al., 1999]. In haplotype analysis, such in silico approaches could assist in identifying one or more variants potentially having an adverse impact on splicing or protein function.

We developed standalone and web-based software to evaluate prospective splicing mutations of any established or unannotated gene, user-defined sequence, or SNP based on standard nomenclature. This obviates the requirement to determine genomic coordinates corresponding to these genes or to learn complex syntaxes in order to introduce mutations into these sequences. Prospective mutations are entered using either common gene and mutation or dbSNP designation, or within custom user-defined sequences or unannotated expressed genomic sequences. After parsing the corresponding sequences, mutations are introduced, and the software dynamically computes and displays the information contents of all relevant splice sites and regulatory sequences, and computes changes in affinity at such sites.

IMPLEMENTATION

Parsing Mutations

One or more allelic variants are entered using the approved HUGO nomenclature [den Dunnen and Antonarakis, 2001; den Dunnen and Paalman, 2003] (see www.hgvs.org/mutnomen/ for the nomenclature, and the mutation checklist at www.hgvs.org/mutnomen/checklist.html). Since information analysis depends on the coordinates of wild-type and variant sequences, the server only supports syntax that delimits the nucleotides affected by the mutation. Rearrangements that do not precisely specify the nucleotide boundaries of the mutation are not supported (e.g., EX3_5del, or 77+?_923+?) [den Dunnen and Antonarakis, 2001]. Several unconventional grammars that have been adopted by certain locus specific mutation databases are also acceptable (Table 1).

The gene is first identified either by its HUGO-approved gene name or by its GenBank mRNA accession number. Depending upon the gene, certain sequence variants that alter splice site information content may not be of interest, because the relevant splice form would not normally include the exon defined by this site. Since alternative mRNA processing from the same gene is fairly common [Modrek et al., 2001], the available mRNAs associated with this gene that have been mapped onto the genome are presented, each hyperlinked to their respective genomic locations. These mRNA accessions are retrieved from the Known Genes Cross Reference (kgXref) table from the University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/goldenPath/gbdDescriptions.html#KgXref). Unless the genomic sequence is indicated as a gene fragment, the longest splice form of the mRNA is presented as the default selection; however, any genome-mapped transcript may be selected. The closest natural splice site to the mutation is inferred from the mutation designation or cDNA coordinate.

The server is intended for analyses of one or more allelic variants. Analysis of multiple mutations(s) *in cis*, i.e., haplotypes, is supported by separating each variant with either the " + " or " ; " symbols (either enclosed by a single set of square brackets or without brackets; spaces on either side of the " + " or " ; " symbols are required). Mutations that occur on different chromosomes or are of uncertain phase affect distinct mRNA transcripts and should be analyzed separately; the corresponding nomenclature for these mutations is not supported.

Parsing SNPs

Single- and multiple-nucleotide polymorphisms in dbSNP may be analyzed according to their Reference Sequence (rs) numbers or they may be explicitly defined by their corresponding genome coordinates and variant(s) in HUGO notation. Queries are looked up in the SNPmap and dbSNP MySQL tables (MySQL AB; www.mysql.com) derived from the UCSC Genome Browser website. Since these tables currently do not contain all mapped SNPs, missing variants can be evaluated by specifying genome coordinates and sequence variant(s) or indel (e.g., g.1805377A>G). Since the dbSNP database is derived independently of either the reference genome sequences or the gene annotations, the dbSNP (or the normal sequence for a gene mutation entry) and given strand of the reference sequence can be discordant. In such instances, the server suggests change(s) in the reference and variant sequences based on the possibility that the reference sequence contains the polymorphic variant or that the reference SNP occurs on the antisense strand. Additionally, since many SNPs are located within genes, but the polarity of transcription cannot be inferred from the SNP itself, either one or both strands may be analyzed for changes in information content.

Genome-wide Precomputation of Natural Splice Site R_i Values

The complete human genome reference sequence (April, 2003 assembly; National Center for Biotechnology Information [NCBI] Build 33) was scanned with donor and acceptor weight matrices to determine the information contents of all natural splice sites in expressed genomic sequences. A MySQL database that maps natural sites to genomic and cDNA sequence coordinates is populated from the data obtained from the above scan. This database serves as a basis for parsing submitted mutations. The genomic coordinate of the exon boundary is extracted from a precomputed MySQL table (ALL_RI; modified from the mRNA.txt from the UCSC genome browser) containing: 1) chromosome name; 2) the accession number of the mRNA; 3)

TABLE 1. Supported Nonstandard Mutation Formats

Mutation	HGVS designation	Reference
IVS9+3_+delAAGTATTTACT	c.901+3delAAGTATTTACT	Nicholls et al. [1992]
903_904insT	c.903_904insT	Cardoso et al. [2002]
IVS3-39_40insT	c.434_39insT	Climente et al. [2002]
IVS16+3del6	c.2056+3delAGCAAA	Messiaen et al. [2000]
IVS2nt5g->c	c.640+5G>C	Scriver et al. [2003]

gene orientation; 4) exon number; 5) the respective mRNA and genomic sequence coordinates; and 6) the respective R_i values of each type of splice site.

The mutation type—either intronic or coding sequence variant—is determined from the mutation name and the chromosomal coordinates of the gene. The orientation of the gene is looked up in the ALL_RI table. The genomic location of the mutation is computed from the offset of its coordinate relative to the nearest natural splice site from this database. The splice site polarities and the signs of the offset are reversed for the genomic coordinates of exon blocks from mRNAs that map onto the antisense strand of the reference sequence. Depending on the specific mutation designation and directionality of transcription, the genomic coordinate is determined by addition or subtraction of the relative distance from the exon boundary coordinate. The HUGO mutation syntax is translated into the equivalent Delila instruction [Schneider et al., 1984]. A sequence interval circumscribing the mutation is retrieved from the reference genome and the variant is introduced into the extracted sequence. To analyze sites within wild-type sequences, identical initial and variant nucleotides are specified (e.g., IVS2+1G>G). If a mutation cannot be processed due to misspecification of the reference sequence (due to entry error or polymorphism), the server reports the correct sequence and permits the instruction to be modified and reprocessed.

Noninteractive Mutation Analysis

This mutation and sequence extraction procedure was verified noninteractively with software (interpretmut.pl) designed to analyze nucleotide mutations embedded in text files. Mutation data were compiled from peer-reviewed articles from *Human Mutation* conforming to HUGO-approved format and from various online locus-specific databases. Text files typically contained a single mutation per line; however, multiple mutations on the same line were treated as single haplotypes (assumed to be *in cis*). Duplicate occurrences of mutations in the same published tables were sorted and eliminated. Alternative mRNA GenBank accession numbers were derived from browser tables for genes specified with mRNA accessions that were not originally indexed in these tables. The software also correctly parses syntaxes for some locus-specific databases with unconventional mutation nomenclature (Table 1).

Wild-type and mutant type sequences are scanned with the splice donor and acceptor and regulatory splice site individual information weight matrices [$R_i(b.l)$]. The predicted changes in binding sites are presented as hypertext markup language (HTML) tables categorized by type of site and change in information content and as sequence walker visualization maps [Rogan et al., 1998]. The initial results web page provides links to each of these tables (data not shown), which are based on the type of change in R_i values that occurs (decreased, increased, or unchanged R_i values), as well as a comprehensive table containing all of these categories of sites. The table rows indicate: 1) the genomic location of each site; 2) the relative distance of the site to the nearest splice junction; 3) the genomic coordinate of the nearest splice junction; 4) the magnitude of the change in R_i value; 5) the corresponding minimum fold and percentage change in predicted binding affinity (based on Rogan et al. [1998]); 6) the Z scores of the natural sites; 7) the Z scores of the variant sites; 8) the change in score [ΔZ], which is a measure of the deviation of the mean information content of all natural sites, $R_{sequence}$ (or the absolute site strength); and 9) the fold changes in binding affinity for R_i contents exceeding zero bits, determined from the differences between the respective R_i values of the site in the wild-type and variant sequences. The original text file is also converted to HTML and mutations are hyperlinked to their respective tables and sequence walker visualization maps (see below). Hyperlinks to original citations and results of these analyses can be found at www.sce.umkc.edu/~roganp/diseasemutinfo/1.html.

All potential splice and regulatory sites (and sites whose information contents change due to mutation) are also displayed as a graphical visualization map of sequence walkers [Schneider, 1997] (as Adobe PDF files; www.adobe.com) depicting each site for the both the natural and variant sequence windows (Fig. 2A). The locations of the natural and variant nucleotides are respectively marked with the tail and heads of an arrow.

Web Server Application and Interface

A web server was developed to compute and display changes in individual information contents of binding sites resulting from specified nucleotide sequence alterations (https://splice.cmh.edu). Registration is required to access the server. Guest registration is usually sufficient for small scale analyses ($\leq$20 runs); however, more extensive access requires full registration via a National Cancer Institute website (see www.lecb.ncifcrf.gov/~toms/contacts.html), which is free for noncommercial application.

The Automated Splice Site Analysis secure web server has a front-end interface containing a form for specifying the sequence to be analyzed, mutation(s), sequence window length, and information weight matrices (Fig. 1). The interface links to an online User Guide and other local and external reference materials. A backend server executes Perl programs and scripts that dynamically process multiple individual information analysis requests from different users.

After checking mutation syntax, the server verifies the gene name and mRNA GenBank accession number for inclusion in the kgXref MySQL table. The variant is then parsed to ensure that the intron number and mRNA coordinates are consistent with the gene structure, and the requested sequence change(s) is analyzed to determine the validity of the nucleotide at that genomic location. Mutations or variants are then processed to form instructions, which are then interpreted by the Delila program [Schneider et al., 1984].

The reference sequence is retrieved from the chromosome sequence library containing the gene corresponding to the requested mRNA and the variant(s) are introduced into a second copy of the sequence. Both sequences are scanned with each $R_i(b,l)$ matrix using default or custom-defined individual information thresholds. Results are retained for 1 hr, after which a scheduling program deletes the related files. This approach permits multiple users to access the server or multiple mutations to be analyzed by a single user simultaneously.

Unannotated, user-supplied sequences are evaluated by comparing the given or reference sequence and a modified version of this sequence created from one or more defined mutations. Both sequences are scanned with the $R_i(b,l)$ matrices over a specified length window circumscribing the coordinate(s) of the variant(s). Since a mutation may affect the R_i value of any potential binding site, the default minimum length sequence window spans an interval upstream and downstream equal to the length of the longest scanned $R_i(b,l)$ matrix. If the calculated window range extends beyond the boundaries of a user-defined sequence, the default range is truncated at the terminus of the sequence. With few exceptions, individual splicing mutations

FIGURE 1. **Entry of sequence variants in established genes into the Automated Splice Site Analysis server. A: Partial screenshot showing the analysis of donor and SC35 serine-rich, arginine-rich (SR) protein binding sites for the IVS2+1G>A splicing mutation in the human phenylalanine hydroxylase (PAH) gene. Access to alternative portals for user-defined sequence entry, human mRNA sequences defined by corresponding GenBank accession number, and SNP analysis are available from this page. Links are provided to an online User Guide and other web-based references useful for formulating and interpreting mutations. B: Genomic sequences corresponding to mRNAs found at the PAH locus in the UCSC Genome Browser. The default selection corresponds to the genome sequence of the longest mRNA in the list.**

have relatively short-range effects, which dictate the length of the sequence window interval that may be scanned (currently ≤ 1 kb). This reduces overhead from scanning long sequences and reduces the size of the sequence walker visualization map that depicts the binding sites. This restriction on sequence window length is relaxed for haplotypes (up to 100 kb in length), in which both the complete interval spanned by the variants and the adjacent flanking windows may be analyzed.

Potential binding sites exceeding threshold information contents ($R_{i,min}$) within the defined sequence window are reported above the default threshold set at the theoretical limit (zero bits) [Schneider, 1997], however this value can be explicitly defined by the user. Sites with information contents below $R_{i,min}$ are generally not reported, except for those affected by variants that change corresponding R_i values. These exceptions include weak or nonsites ($R_i < R_{i,min}$) that are strengthened or activated by the mutation ($\geq R_{i,min}$) and valid sites ($R_i \geq R_{i,min}$) that are abolished or weakened below the threshold value ($R_i < R_{i,min}$).

Mutation entry. The server offers other portals for mutation entry that obviate the requirement to specify variants in HUGO gene nomenclature. The genomic locations corresponding to unannotated mRNAs mapped onto the genome sequence have also been indexed in the server's MySQL database, and potential mutations in these presumptive genes can be predicted with the server. In instances in which transcripts have been mapped by sequence comparison to multiple genomic locations but the genomic sequence has not been assigned a HUGO-approved gene name, any single-transcription templates may be selected for analysis. Each of the possible gene locations is hyperlinked to the corresponding UCSC genome browser coordinates, since these sequences could represent either unrecognized members of a multigene family or pseudogenes (which may contain nonfunctional splice sites with $R_i < R_{i,min}$).

The server can be used to analyze variants in user-defined DNA sequences that are not found within the reference set of human genes, mRNAs, or dbSNP accession numbers. Because such sequences are not annotated and can be derived from any source, only unambiguous mutations that are specified by their sequence coordinate and nucleotide change(s) can be analyzed (e.g., 454C>G). Raw sequences entered on the web form can be numbered using a reformatting button, which facilitates determining the coordinate(s) of the sequence variant(s) on either strand. For user-defined sequences, however, only the syntax of the coordinate and nucleotide change are verified.

Selection of precomputed information weight matrices. The $R_i(b.l)$ matrices for all validated donor and acceptor splice sites were updated from the set of intron–exon junctions annotated in the April 2003 genome reference sequence (with $R_i > 0$ bits) [Rogan et al., 2003]. Donor and acceptor binding site models were also derived from the February 2002 mouse genome reference sequence. The corresponding murine and human splice site matrices are nearly identical, but there are subtle, statistically significant differences between them, especially at the positions with lower overall conservation (P. Rogan, unpublished

observations). Mouse mutations can be evaluated by user-defined sequence entry.

Pathological mRNA splicing abnormalities can be affected by mutations in splicing regulatory elements that produce either exon skipping or inclusion [Dietz et al., 1993; Shiga et al., 1997]. These elements are recognized by accessory serine-arginine rich proteins, which are known to promote early recognition of both donor and acceptor sites by enhancing formation of prespliceosomal complexes with U1 small nuclear ribonucleoproteins [Staknis and Reed, 1994]. To evaluate the consequences of potential mutations affecting these sites, we have computed information theory–based weight matrices for sequences recognized by the splicing regulatory factors, SRp40, ASF/SF2 [Liu et al., 1998], and SC35 [Liu et al., 2000]. Genomic sequence may be scanned either with the default human donor and acceptor or any of these regulatory protein information weight matrices selected from the list of available $R_i(b,l)$ matrices.

Modifications of the analysis and display. Upon specification of the variant and selection of $R_i(b,l)$ matrices and a genomic sequence, the following ("Advanced options") can be used to customize the analysis and resulting output: 1) The threshold $R_{i,min}$ value may be defined. Recent comprehensive models of splice sites have revised estimates of $R_{i,min}$ to 1.6 bits, based on reanalysis of minimally functional sites [Rogan et al., 1998, 2003]. High threshold values can be used to eliminate display of weak sites; low thresholds detect weak sites or unbound sequences ($R_i < 0$). 2) Some electronic databases have adopted a convention in which the first nucleotide of a full-length GenBank cDNA accession is stipulated as position 1. Since this offset affects parsing of mutations relative to the cDNA sequence, the first position may be defined relative to either the beginning of the open reading frame or to the transcription start site. 3) Amino acid translation of the three forward frames can also be displayed on the sequence walker visualization map in addition to potential splice sites or splicing regulatory sequences. This option can reveal silent or missense mutations with concomitant effects at the RNA level. 4) Results can be restricted to tabular output by eliminating Walker visualization maps. Server performance is improved, particularly for haplotypes of widely separated mutations.

Displaying results. The server uses the same backend engine to generate results as the noninteractive text parser; however, the web version has additional capabilities. The table (Fig. 2A) and sequence walker visualization map (Fig. 2B) indicate the information contents of all predicted sites above the $R_{i,min}$ value (which can be defined by the user) in the normal and mutated sequences. Both the initial and final information contents at any natural, preexisting, or novel cryptic sites affected by sequence changes are reported, including values of altered sites that fall below the $R_{i,min}$ threshold.

The table cells indicating the locations and information contents of predicted binding sites are color-coded (using cascading style sheets) by direction and type of change in R_i value. Mutations that inactivate or create leaky, cryptic sites and display preexisting cryptic sites whose R_i values are unchanged are separately coded to facilitate interpretation. A legend indicating the codes and their significance is presented at the end of each analysis.

Information contents of all natural splice sites in the gene containing the variant are retrieved from the ALL_RI MySQL table and presented as a popup window that is hyperlinked to the results table. The internal cells in the main table on each of the results pages are linked to the cell of the natural splice site closest to the variant. The data in this table is used to compare the R_i values of natural sites with the corresponding mutated or cryptic sites. The popup table is not produced for user-defined sequences, since the locations of exon/intron boundaries are not available for these sequences. In SNP analyses, these cells are linked to the corresponding entry in the UCSC Genome Browser.

Exon boundaries are annotated on the sequence walker visualization maps as bracketed dotted lines below the sequence, which are based on chromosomal coordinates read from the ALL_RI MySQL table.

Batch analysis of mutations is available for National Institutes of Health (NIH)-registered users. The mutation list is uploaded as a text document, verified, and processed to eliminate duplicate entries. Results are concatenated on a single HTML page of tables and visualization maps grouped by mutation. The uploaded mutation list containing is modified to provide links to these results. Mutations that are unable to be parsed are listed on a separate page with an explanation of the error.

Validation of the Server

Mutations. The server was initially validated by analyzing ~1,300 mutations from *Human Mutation*, including several multivariant haplotypes. Variants were parsed directly from published text or interactively. We confirmed that all of the previously recognized splicing mutations affected splice site strength or activated cryptic splice sites. The analysis also revealed several unexpected and potentially significant predicted effects on splicing (Table 2), which were not evident from their descriptions in the original sources. The information analysis indicated eight missense mutations that appear to have concomitant effects on adjacent splice donor and acceptor sites, four partially functional splice sites previously thought to be null alleles, and six mutations with previously unrecognized effects on cryptic splicing. In addition, reduction in the strengths of potential SR protein binding sites for ASF/SF2, SC35, and SRp40 that may have an impact on splicing were also detected; however, these cannot be interpreted unequivocally (results not shown).

SNPs. The capabilities of the server for SNP analysis were tested with relevant splicing-related SNPs. A search of Entrez SNP http://ncbi.nlm.nih.gov/entrez/query/fcgi?CMD=search&DB=snp for all genotyped human entries known to occur within splice sites of known genes identified 22 hits (rsID numbers 20381, 25404, 32506, 140601, 363808, 1197060, 1419973, 1805377, 2066504, 2234733, 2241524, 2243187, 2307356, 3093513, 3093513, 4150000, 5744954, 5745908, 9332736, 11509437, 11575789, and 12722699). Eight of these SNPs were listed in the UCSC SNPmap MySQL table (Table 3). As expected, all of these variants were predicted to affect mRNA splicing. The substitution in dbSNP represented the mutant nucleotide for all of the variants except for rs5744954, in which the reference sequence appeared to harbor the mutant allele.

Surprisingly, an apparently deleterious SNP within the exon 7 acceptor site of the *XRCC4* gene (MIM# 194363), rs1805377 (chromosome 5: g.82687655A>G), has been shown to be a common allele in multiple genotyping surveys (dbSNP: ss2667795, ss4963080, ss6903985, ss11763271, ss17845748, and ss23674295), and its frequency varies widely among different ethnic populations. The A allele activates 11.5- and 11.3-bit splice sites at positions 82687655 and 8268761, whereas the G allele inactivates the 11.5-bit site, resulting in exclusive use of the in-frame 11.3-bit site. These predictions are supported by GenBank accessions containing each of these *XRCC4* splice forms

A)

Gene:	PAH
mRNA Accession:	U49897
Mutation:	IVS2+1G>A
Genomic Designation:	chr 12: c103239515t

Sequence Walker

IVS2+1G>A on - strand

Type of change	Genomic Coordinate	Position Relative to Natural Site	Closest Natural Site	Initial (Ri)	Final (Ri)	Δ Ri	Fold change	% Binding (Final/Initial)	Initial (Z)	Final (z)	Δ Z
No change	103239551	-36	103239515	1.8	1.8	0.0	1.0	100	-1.9	-1.9	0.0
No change	103239547	-32	103239515	1.5	1.5	0.0	1.0	100	-2.0	-2.0	0.0
No change	103239533	-18	103239515	1.1	1.1	0.0	1.0	100	-2.1	-2.1	0.0
No change	103239498	17	103239515	1.6	1.6	0.0	1.0	100	-1.9	-1.9	0.0
DECREASED	103239515	0	103239515	7.2	-5.6	-12.8	-144.4	NA	-0.3	-4.1	-3.9

Legend: Site Abolished; Leaky Site; Cryptic Site created; Strengthened Pre-existing site; Pre-existing Site; Weakened Pre-existing Site

B)

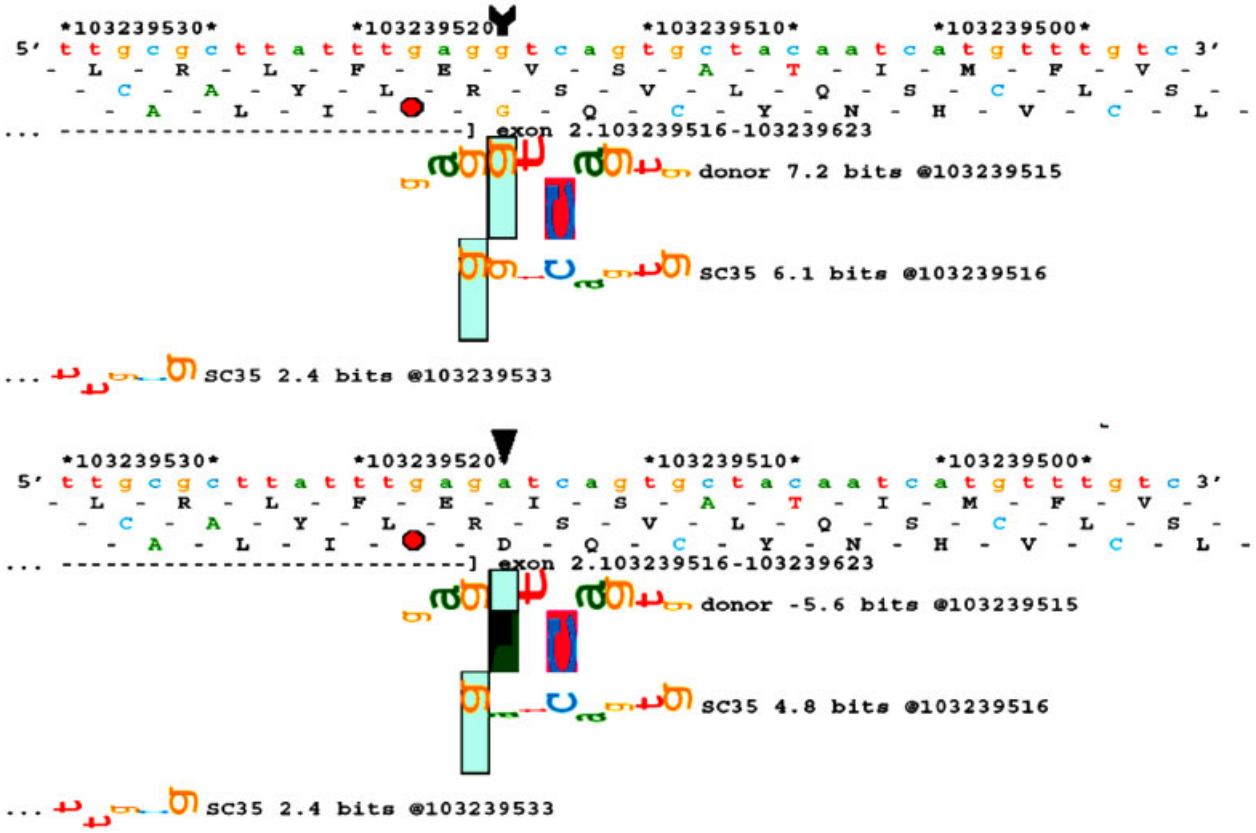

FIGURE 2. **Results generated by the server for the mutation shown in Figure 1. Initially, links are provided to sequence walker visualization map and HTML tabular analysis for the donor and SC35 $R_i(b,l)$ matrices (not shown). A:** Table produced that shows all predicted donor sites and color-coded changes in information content (R_i) over an 108-nucleotide window circumscribing the mutation. **B:** Sequence walker visualization map of all donor and SC35 binding sites within this sequence window. The tail of the arrow indicates the reference sequence and the head points to the mutation. The three forward reading frame translations are shown below the sequence. Bracket and dotted lines delineate the normal and mutated exon 2. The individual information analysis displays sequence walkers for a predicted splice donor site at position 103239515 and an SC35 site at 103239516. The mutation decreases the strength of the SC35 site to 4.8 bits and inactivates the donor site (–5.6 bits).

TABLE 2. Splicing Alterations Predicted by Information Analysis That Were Not Evident in Published Studies

Gene MIM#	Mutation	Published interpretation	Predicted impact	Reference
DHCR7: 602868	321G>C	Missense mutation	Leaky mutation reduces R_i value of natural site from 6.5 to 2.5 bits. Pre-existing 4.1 bit cryptic site is 16 bases away.	Witsch-Baumgartner et al. [2001]
SPG4: 604277	1005–1G>T	Creates cryptic splice site	Predicted exon skipping	Proukakis et al. [2003]
	1728+2T>T	Skipping of exon 16	Creates a cryptic site 4 bases away	
AGL: 232400	IVS21+1G>A	Exon skipping	Possible use of preexisting cryptic site 29 bases away	Lucchiari et al. [2002]
FBNI: 134797	331T>C 344C>G	Missense mutations	Decreases R_i value of natural site	Robinson et al. [2002]
CHM: 300390	1380–6T>G	Skipping of exon 11	Residual splicing predicted at the natural site and activation of pre-existing cryptic site 4 nos away	McTaggart et al. [2002]
	1379+2_+3ins GGT	Aberrant splicing	Strong cryptic site created at the insertion site; also residual splicing predicted at the natural site.	
	1380–1G>A	Skipping of exon 11	Leaky mutation predicted	
NFI: 162200	5546G>A 6858G>C	Missense mutations	Affects R_i values of natural sites	Messiaen et al. [2000]
	IVS39–12T>A	Splice site inactivation 3′	Leaky mutation predicted	
CTNS: 606272	668G>T	Missense mutation	Possibly use of preexisting cryptic site of 16 bases downstream	Kalatzis et al. [2002]
	1020G>A	Out of frame deletion	Leaky mutation predicted	
MSH2: 120435	IVS5+3A>T	Inactivates splice site	Predicted to produce residual wild-type mRNA	Taylor et al. [2003]
ARX: 300382	c.196+2T>C	Skipping of exon1	Leaky mutation predicted	Kato et al. [2004]
	c.1119+1G>C	Skipping of exon 3	Possible use of potential sites 18 or 48 bases downstream	
	c.135–1G>C	Skipping of exon 3	Possible use of cryptic site created 4 bases downstream or pre-existing site 19 bases downstream	
PHYH: 602026	c.497–2A>G	Skipping of exon 6	Leaky mutation, as natural site is not abolished, or possible use of a cryptic site created 1 base upstream	Jansen et al. [2004]
	c.679–1G>T	Skipping of exon 6	Possible use of a novel cryptic site 5 bases downstream	

TABLE 3. Information Analysis of Single Nucleotide Polymorphisms in dbSNP in Established Splice Sites

dbSNP rsID	Gene; MIM#	Predicted impact
1805377[a]	*XRCC4:* 194363	Abolishes acceptor; activates preexisting cryptic site (in frame)
2243187	*ILI9:* 605687	Abolishes acceptor; may activate cryptic site (in frame)
4150000	*EXOI:* 606063	Abolishes acceptor site
2307356	*ORC2L:* 601182	Weakens donor site
1419973	*LRRFIP2:* not assigned	Weakens donor site; may activate cryptic site (in frame)
2234733	*GSTA2:* 138360	Weakens donor site.
5744954[b]	*POLE:* 174762	Strengthens acceptor site, inactivates cryptic site
32506	LOC345645/XM_293923[c]	Weakens acceptor site; may activate cryptic site (in frame)

[a]Both alleles are common.
[b]SNP variant is the common allele.
[c]Gene name and MIM number have not been assigned: mRNA encodes predicted protease ATPase 1 (26S subunit).

(e.g., U40622 vs. AF424542, BC016314, AB017445, or BC010655). The pair of codons excluded from transcripts produced by the G allele occur within the evolutionarily-conserved, DNA binding domain of *XRCC4*. Interestingly, however, the 11.5-bit site is absent in both mouse and rat, because of the absence of a guanine in orthologous sequences corresponding to position 8268755. Considering the essential role that *XRCC4* has in nonhomologous chromosome end joining [Critchlow and Jackson, 1998], it is tempting to speculate that this genotype may significantly contribute to both intra- and interspecies differences in the ability to repair of double strand breaks in the genome that arise from environmental exposures.

DISCUSSION

The system that we have described assists in the interpretation of noncoding sequence variation in functional elements within human genes (or user-defined sequences) by specifying mutations

in commonly accepted mutation formats, user-defined sequences, or to dbSNP references. Information theory–based mutation analysis relies on robust computational models of functional binding sites, rather than allelic frequency or comparative genomic analyses. Analysis is performed dynamically by evaluating the effects, if any, on the information contents of splice sites and accessory splicing factor recognition sites. Other software to process downloaded mutation data in HUGO and nonstandard formats can be used to analyze mutations for genetic disorders from a wide variety of peer-reviewed sources. The server confirmed splicing mutations in a set of unselected, published mutations and also predicted previously unrecognized effects on mRNA splicing in a number of instances.

Sequences entered by the users can be derived from any source, including other species. Analysis of murine sequences can be performed using $R_i(b,l)$ matrices of mouse donor and acceptor splicing models. These and the human splice site models may also be used to perform cross-species analyses for prediction of splicing patterns in heterologous expression or transgenic systems.

Sequence retrieval with this system depends on the accuracy of UCSC genome annotation tables of exon coordinates, gene names, mRNA accession numbers, and localized SNPs. Incorrect exon assignment, typically as a result of incomplete mRNAs mapped onto the April 2003 reference sequence, may result in missing exons. Also, we have noted inaccuracies in exon boundary coordinates, owing to imprecise alignments of cDNA with genomic sequences, although this is uncommon. Polymorphic genomic reference sequences that do not match the specified input nucleotide are not parsed; however, as in mutations, the SNP instruction can be modified and reprocessed. We anticipate that updates of the human genome reference sequence MySQL tables will reduce or eliminate these sources of annotation error.

The system has been designed to facilitate addition of other genome sequences and information weight matrices to analyze other types of binding sites. It will eventually incorporate multipartite information theory–based binding site analysis [Shultzaberger et al., 2001; Bi and Rogan, 2004], a prerequisite for the development of comprehensive mRNA splicing models that predict the structures of normal and abnormal transcripts.

NOTE ADDED IN PROOF

Since May 2004, >700 analyses have been performed by 181 registrants.

ACKNOWLEDGMENTS

We thank Todd Fiedler for Information Services support, Chengpeng Bi for advice, and Julianne Collins, Thomas Schneider, and Paul Rothberg for beta testing.

REFERENCES

Bi C, Rogan PK. 2004. Bipartite pattern discovery by entropy minimization-based multiple local alignment. Nucleic Acids Res 32:4979–4991.

Cardoso C, Leventer RJ, Dowling JJ, Ward HL, Chung J, Petras KS, Roseberry JA, Weiss AM, Das S, Martin CL, Pilz DT, Dobyns WB, Ledbetter DH. 2002. Clinical and molecular basis of classical lissencephaly: mutations in the LIS1 gene (PAFAH1B1). Hum Mutat 19:4–15.

Climente C, Rubio V. 2002. H Intragenic polymorphisms and haplotype analysis in the ornithine transcarbamylase (OTC) gene and their relevance for tracking the inheritance of OTC deficiency. Hum Mutat 20:406–407.

Critchlow SE, Jackson SP. 1998. DNA end-joining: from yeast to man. Trends Biochem Sci 23:394–398.

den Dunnen JT, Antonarakis SE. 2001. Nomenclature for the description of sequence variations. Hum Genet 109:121–124.

den Dunnen JT, Paalman MA. 2003. Standardizing mutation nomenclature: why bother? Hum Mutat 22:181–182.

Dietz HC, Valle D, Francomano CA, Kendzior RJ, Pyeritz RE, Cutting GR. 1993. The skipping of constitutive exons in vivo induced by nonsense mutations. Science 259:680–683.

Gadiraju S, Vyhlidal CA, Leeder JS, Rogan PK. 2003. Genome-wide prediction, display, and refinement of binding sites with information theory based models. BMC Bioinformatics 4:38.

Jansen GA, Waterham HR, Wanders RJ. 2004. Molecular basis of Refsum disease: sequence variations in phytanoyl-CoA hydroxylase (PHYH) and the PTS2 receptor (PEX7). Hum Mutat 23:209–218.

Horaitis O, Cotton RG. 2004. The challenge of documenting mutation across the genome: the human genome variation society approach. Hum Mutat 23:447–452.

Kalatzis V, Cohen-Solal L, Cordier B, Frishberg Y, Kemper M, Nuutinen EM, Legrand E, Cochat P, Antignac C. 2002. Identification of 14 novel CTNS mutations and characterization of seven splice site mutations associated with cystinosis. Hum Mutat 20:439–446

Kato M, Das S, Petras K, Kitamura K, Morohashi K-I, Abuelo DN, Barr M, Bonneau D, Brady AF, Carpenter NJ, Cipero KL, Frisone F, Fukuda T, Guerrini R, Iida E, Itoh M, Lewanda AF, Nanba Y, Oka A, Proud VK, Saugier-Veber P, Schelley SL, Selicorni A, Shaner R, Silengo M, Stewart F, Sugiyama N, Toyama J, Toutain A, Vargas AL, Yanazawa M, Zackai EH, Dobyns WB. 2004. Mutations of ARX are associated with striking pleiotropy and consistent genotype–phenotype correlation. Hum Mutat 23:147–159.

Kodolitsch YV, Pyeritz R, Rogan PK. 1999. Splice-site mutations in atherosclerosis candidate genes: relating individual information to phenotype. Circulation 100:693–699.

Krawczak M, Cooper DM. 1991. Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. Hum Genet 86:425–441.

Lamba V, Lamba J, Yasuda K, Strom S, Davila J, Hancock ML, Fackenthal JD, Rogan PK, Ring B, Wrighton SA, Schuetz EG. 2003. Hepatic CYP2B6 expression: gender and ethnic differences and relationship to CYP2B6 genotype and CAR (constitutive androstane receptor) expression. J Pharmacol Exp Ther 307:906–922.

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

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

Lucchiari S, Donati MA, Parini R, Melis D, Gatti R, Bresolin N, Scarlato G, Comi GP. 2002. Molecular characterisation of GSD III subjects and identification of six novel mutations in AGL. Hum Mutat 20:480.

McTaggart KE, Tran M, Mah DY, Lai SW, Nesslinger NJ, MacDonald IM. 2002. Mutational analysis of patients with the diagnosis of choroideremia. Hum Mutat 20:189–196.

Messiaen LM, Callens T, Mortier G, Beysen D, Vandenbroucke I, Van Roy N, Speleman F, Paepe AD. 2000. Exhaustive mutation

analysis of the NF1 gene allows identification of 95% of mutations and reveals a high frequency of unusual splicing defects. Hum Mutat 15:541–555.

Modrek B, Resch A, Grasso C, Lee C. 2001. Genome-wide detection of alternative splicing in expressed sequences of human genes. Nucleic Acids Res 29:2850–2859.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.

Nicholls AC, Oliver J, Renouf DV, Heath DA, Pope FM. 1992. The molecular defect in a family with mild atypical osteogenesis imperfecta and extreme joint hypermobility: exon skipping caused by an 11-bp deletion from an intron in one COL1A2 allele. Hum Genet 88:627–633.

Proukakis C, Auer-Grumbach M, Wagner K, Wilkinson PA, Reid E, Patton MA, Warner TT, Crosby AH. 2003. Screening of patients with hereditary spastic paraplegia reveals seven novel mutations in the SPG4 (Spastin) gene. Hum Mutat 21:170.

Richard I, Beckmann JS. 1995. How neutral are synonymous codon mutations? Nat Genet 10:259.

Robinson PN, Booms P, Katzke S, Ladewig M, Neumann L, Palz M, Pregla R, Tiecke F, Rosenberg T. 2002. Mutations of FBN1 and genotype–phenotype correlations in Marfan syndrome and related fibrillinopathies. Hum Mutat 20:153–161.

Rogan PK, Schneider TD. 1995. Using information content and base frequencies to distinguish mutations from genetic polymorphisms in splice junction recognition sites. Hum Mutat 6:74–76.

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

Rogan PK, Svojanovsky S, Leeder JS. 2003. Information theory-based analysis of CYP2C19, CYP2D6 and CYP3A5 splicing mutations. Pharmacogenetics 13:207–218.

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

Schneider TD. 1997. Sequence walkers: a graphical method to display how binding proteins interact with DNA or RNA sequences. Nucleic Acids Res 25:4408–4415.

Scriver CR, Hurtubise M, Konecki D, Phommarinh M, Prevost L, Erlandsen H, Stevens R, Waters PJ, Ryan S, McDonald D, Sarkissian C. 2003. PAH.db 2003: what a locus–specific knowledge base can do. Hum Mutat 21:333–344.

Shiga N, Takeshima Y, Sakamoto H, Inoue K, Yokota Y, Yokoyama M, Matsuo M. 1997. J Clin Invest 100:2204–2210.

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

Staknis D, Reed R. 1994. Proteins promote the first specific recognition of pre-mRNA and are present together with the U1 small nuclear ribonucleoprotein particle in a general splicing enhancer complex. Mol Cell Biol 14:7670–7682.

Susani L, Pangrazio A, Sobacchi C, Taranta A, Mortier G, Savarirayan R, Villa A, Orchard P, Vezzoni P, Albertini A, Frattini A, Pagani F. 2004. TCIRG1-dependent recessive osteoporosis: mutation analysis, functional identification of the splicing defects, and in vitro rescue by U1 snRNA. Hum Mutat 24:225–235.

Taylor CF, Charlton RS, Burn J, Sheridan E, Taylor GR. 2003. Genomic deletions in MSH2 or MLH1 are a frequent cause of hereditary non-polyposis colorectal cancer: identification of novel and recurrent deletions by MLPA. Hum Mutat 22:428–433.

Vockley J, Rogan PK, Anderson BD, Willard J, Seelan RS, Smith DI, Liu W. 2000. Exon skipping in IVD RNA processing in isovaleric acidemia caused by point mutations in the coding region of the IVD gene. Am J Hum Genet 66: 356–367.

Witsch-Baumgartner M, Loffler J, Utermann G. 2001. Mutations in the human DHCR7 gene. Hum Mutat 17(3):172–182.