Information theory-based analysis of *CYP2C19*, *CYP2D6* and *CYP3A5* splicing mutations

Peter K. Rogan[a], Stan Svojanovsky[a] and J. Steven Leeder[b]

Several mutations are known or suspected to affect mRNA splicing of *CYP2C19*, *CYP2D6* and *CYP3A5* genes; however, little experimental evidence exists to support these conclusions. The present study applies mathematical models that measure changes in information content of splice sites in these genes to demonstrate the relationship between the predicted phenotypes of these variants to the corresponding genotypes. Based on information analysis, the *CYP2C192* variant activates a new cryptic site 40 nucleotides downstream of the natural splice site. *CYP2C19**7* abolishes splicing at the exon 5 donor site. The *CYP2D6**4* allele similarly inactivates splicing at the acceptor site of exon 4 and activates a new cryptic site one nucleotide downstream of the natural acceptor. *CYP2D6**11* inactivates the acceptor site of exon 2. The *CYP3A5**3* allele activates a new cryptic site 236 nucleotides upstream of the exon 4 natural acceptor site. *CYP3A5**5* inactivates the exon 5 donor site and *CYP3A5**6* strengthens a site upstream of the natural donor site, resulting in skipping of exon 7. Other previously described missense and nonsense mutations at terminal codons of exons in these genes affected splicing. *CYP2D6**8* and *CYP2D6**14* both decrease the strength of the exon 3 donor site, producing transcripts lacking this exon. The results of information analysis are consistent with the poor metabolizer phenotypes observed in patients with these mutations, and illustrate the potential value of these mathematical models to quantitatively evaluate the functional consequences of new mutations suspected of altering mRNA splicing. *Pharmacogenetics* 13:207–218 © 2003 Lippincott Williams & Wilkins**

Pharmacogenetics 2003, **13**:207–218

Keywords: drug biotransformation, mRNA splicing, cytochrome P450, single nucleotide polymorphism, information theory, individual information

[a]Laboratory of Human Molecular Genetics and [b]Division of Pediatric Pharmacology and Medical Toxicology, Children's Mercy Hospital and Clinics, Kansas City, Missouri, USA.

Sponsorship: This work was supported by grant ES10855 from the National Institute of Environmental Health Sciences.

Correspondence and requests for reprints to Peter K. Rogan, Children's Mercy Hospital and Clinics, 2401 Gillham Road, Kansas City, MO 64108, USA.
Tel: +1 816 960 4820; fax: +1 816 753 1307; e-mail: progan@cmh.edu

Received 30 October 2002
Accepted 29 January 2003

Introduction

Chemical modification of drugs via biotransformation reactions generally results in termination of biological activity through decreased affinity for receptors or other cellular targets, as well as more rapid elimination from the body. The cytochromes P450 (CYPs) are heme-containing proteins that catalyse the oxidative biotransformation of exogenous compounds, such as drugs and environmental toxicants, as well as the metabolism of many lipophilic endogenous substances (e.g. steroids, fatty acids, fat-soluble vitamins, prostaglandins, leukotrienes and thromboxanes). Over the past 10 years, it has become increasingly apparent that some differences in the interindividual responses to medications, particularly those with relatively narrow therapeutic indices, are a consequence of genetic variations in specific drug biotransformation pathways.

The best-studied genetic variation in drug response is the debrisoquine/sparteine polymorphism attributed to a specific CYP designated as *CYP2D6* [1,2]. The *CYP2D6* gene locus is highly polymorphic with more than 70 allelic variants having been described to date (http://www.imm.ki.se/CYPalleles/cyp2d6.htm). Five to 10% of Caucasians and approximately 1–2% of Asian subjects are classified as poor metabolizers (PMs) of drugs metabolized by *CYP2D6* [3]. Generally, the PM phenotype is conferred by inheritance of two recessive loss-of-function alleles – variant forms of CYP genes that do not give rise to functional protein. Another well-characterized polymorphic CYP activity was originally observed as a marked inability to 4′-hydroxylate the *S*-enantiomer of mephenytoin [4]. *CYP2C19* or 'mephenytoin hydroxylase' deficiency is present in 3–5% of the Caucasian population and in approximately 20% of Asians [5]. Constituents of the CYP3A subfamily metabolize up to 60% of all currently used drugs [6] and collectively form the largest portion of the human liver CYP protein. *CYP3A5* is polymorphically expressed in approximately 33% of Caucasians and 60% of African-Americans [7]. Only people with at least one wild-type *CYP3A5* allele appear to express appreciable amounts of CYP3A5 protein.

Single nucleotide polymorphisms (SNPs) that affect mRNA splicing are important causes of defective CYP alleles. For example, *CYP2D6**4* has an allele frequency of approximately 0.2 in Caucasians and accounts for

0960-314X © 2003 Lippincott Williams & Wilkins

DOI: 10.1097/01.fpc.0000054078.64000.de

approximately 75% of loss-of-function alleles in this population [8–11], while the *CYP2D6*11* allele is less common but is also associated with aberrant mRNA splicing [12]. Similarly, the principal allelic variant of *CYP2C19*, *CYP2C19*2*, is present in approximately 75% of the defective alleles found in Japanese and Caucasian PMs. A cryptic splice site that alters the reading frame is activated by a substitution in exon 5 and results in a premature stop codon approximately 20 amino acids downstream of the mutation [13]. Defective mRNA splicing is also thought to be the functional consequence of the less frequent *CYP2C19*7* allele [14].

The *CYP3A5*3* allele is similarly reported to activate an out-of-frame strong cryptic splice site 236 nucleotides downstream of the exon 4 natural acceptor site. However, the mechanistic basis for the inclusion of exon 4B in splice variant 2 and the concurrent inclusion of exon 5B and exclusion of exon 6 in splice variant 3 reported by Kuehl *et al.* 2001 [7] is less clear.

Previously, the functional consequences of these mutations affecting mRNA splicing have been established by immunochemical analysis of microsomal proteins [13] or through population genotype–phenotype correlation studies of *CYP2C19* [15–17], *CYP2D6* [8–11] and *CYP3A5* [7,18,19]. However, with the exception of *CYP2D6*4* [20] and *CYP3A5*6* [7], confirmation of aberrant splicing has not been conducted at the mRNA level, and as a consequence, aberrant splicing is largely inferred from sequence analysis.

Information theory-based models of donor and acceptor splice sites reveal which nucleotides are permissible at both highly conserved and variable positions of these sites [21,22]. Individual information is related to thermodynamic entropy and therefore to the free energy of binding [21]. Because splice sites are recognized before intron excision [23], the sequence of the splice site dictates the strength of the spliceosome–splice site interaction and thus splice site utilization. The individual information contents (R_i, in bits) of natural and mutant splice sites and coding sequence variants have been compared in genes responsible for a wide variety of genetic disorders [24–31]. In these investigations, the effects of nucleotide substitutions were predicted from changes in R_i values [28] and comparisons between normal and mutant splice site R_i values identified substitutions that impaired splicing from those that did not, distinguished null alleles from those that were partially functional, and detected activated cryptic splice sites.

The aim of this investigation was to utilize information analysis as a quantitative tool to characterize splice site location and strength within *CYP2C19*, *CYP2D6* and *CYP3A5* exon–intron junctions and to predict the functional effects of mutations within these regions of the genes. The results presented indicate that information analysis may be useful in predicting the consequences of novel SNPs in drug metabolizing genes on mRNA splicing.

Methods

Sequences

The structure of the *CYP2C19* gene was deduced from the sequences of two non-overlapping contigs, AL583836.18 and AL133513.12, which themselves were derived from different clones (RP11-466J14 and RP11-400G3, respectively) of the same library. Despite the high degree of similarity of the *CYP2C19* gene with other members of this subfamily (*CYP2C8*, *CYP2C9* and *CYP2C18*), it was apparent that both contigs were derived from the same gene because the original mRNA, M91854, aligned with 99.9% nucleotide identity (exons 1–5 with AL583836.18 and exons 6–9 with AL133513.12). This is consistent with the gene structure reported in the November 2002 human genome draft assembly (chromosome 10, positions 95733563–95868877).

The *CYP2D6* wild-type gene was analysed using GenBank accession M33388 [32]. Accession AC005020 was the reference sequence for the *CYP3A5* gene. The coordinates of the exon–intron junctions in AC005020 are concordant with those given by Kuehl *et al.* [7] and the sequence NG_000004 (http://www.imm.ki.se/CYPalleles/cyp3a5.htm).

Source of mutations

The sequences of variants proposed to affect splicing of the *CYP2C19*, *CYP2D6* and *CYP3A5* genes were obtained from the CYP Allele Nomenclature database (http://www.imm.ki.se/CYPalleles/). In some instances, a single sequence could be used to specify multiple alleles at the same locus. The following list indicates the wild-type and variant sequences analysed: 681G>A for the *CYP2C19*2A* and *CYP2C19*2B* alleles [13,14]; IVS5 + 2T>A for the *CYP2C19*7* allele [14]; 1846G>A for the *CYP2D6*4A-L* allele [8,33] and 883G>C for the *CYP2D6*11* allele [12]. The coordinates for the two *CYP2D6* mutations listed in the CYP Allele Nomenclature database are shifted 88 bp relative to those of the M33388 sequence in which the transcription initiation site is designated +1 (e.g. 1934G>A for *CYP2D6*4* and 971G>C for *CYP2D6*11*). For *CYP3A5*, the *CYP3A5*3* allele corresponds to 6986A>G [7], *CYP3A5*5* to 12952T>C [19] and the *CYP3A5*6* allele to 14690G>A [7].

Based upon our previous studies indicating that information present in the terminal codons of internal exons make these positions more susceptible to mutation

[25], we also analysed such substitutions in CYP genes, specifically the donor sites corresponding to *CYP2D6*8* and *CYP2D6*14* alleles. Mutations involving the terminal codons of exons in the *CYP2C19* and *CYP3A5* genes have not been reported.

Information models of splice donor and acceptor sequences

Information analysis of nucleic acid binding sites is based upon information theory weight matrices derived from a comprehensive set of aligned functional sites. The frequencies of nucleotides at each position are used to calculate the individual information weight matrix, $R_i(b,l)$ [21,22], where b represents the particular base at position l. The models are then utilized to evaluate the effect of splicing mutations by comparing the information contents (R_i value) for wild-type and mutant sequences. Functional binding sites have R_i values > 0 bits.

The previous information weight matrices of donor and acceptor splice sites [34] were updated by deriving matrices from a more comprehensive set of sites in the human genome working draft sequence. A detailed description of this procedure and the resultant weight matrices are presented in the Appendix.

Mutation analyses

The locations and strengths of splice sites for the *CYP2C19*, *CYP2D6* and *CYP3A5* genes and corresponding splicing mutations were then evaluated with the updated information matrices. Wild-type and mutant CYP sequences were parsed with the Delila software package [21]*, and the corresponding individual information contents (R_i in bits) of these sites, were evaluated for normal and mutated sites [28].

The consequences of mutating a binding site can be quantified because information content is cumulative over all positions in the splice site [28]. The previously described weight matrices were used to scan genomic sequences for sites with positive R_i values, and the results were displayed visually using the Walker and Lister programs [22]. A sequence window of ±400 nucleotides circumscribing the natural and mutant splice sites was scanned to detect potential cryptic splice sites potentially activated by splicing variants, as the majority of natural or cryptic exons do not exceed this length. Mutational severity was inferred by comparison of R_i values of the mutant sequence with the cognate wild-type splice sites [28]. Derivation of error associated with the individual information values is available online as supplementary material.

Results

Splice junction models

Comprehensive updated models of splice acceptor and donor splice sites were derived from the human genome draft sequence (10/1/2000 Version). To compare our results with the original weight matrix [34], we used the same sequence intervals to define the average information content (i.e. for acceptor sites, from −25 to +2, and for donor sites, from −3 to +6). Sequence logos depicting the average information contents of acceptor and donor sites for the updated models are shown in Fig. 1. Several features of the updated models deserve comment. Although a histogram of individual splice site strengths comprising all sites in the model approximates a Gaussian distribution, it is actually a multinomial distribution because of the large number of sites from which the matrices were derived. The standard deviation of R_i is reduced for both donor and acceptor sites by approximately 1 bit relative to the original models [34] that were based upon 31-fold fewer sequences. As a consequence, the updated splice junction database exhibits greater sequence variability, resulting in increased uncertainty, and therefore a lower average information content (R_{sequence}; decreased from 8.00 to 6.73 bits for donor and from 8.87 to 7.45 bits for acceptor sites) than the previous models [28].

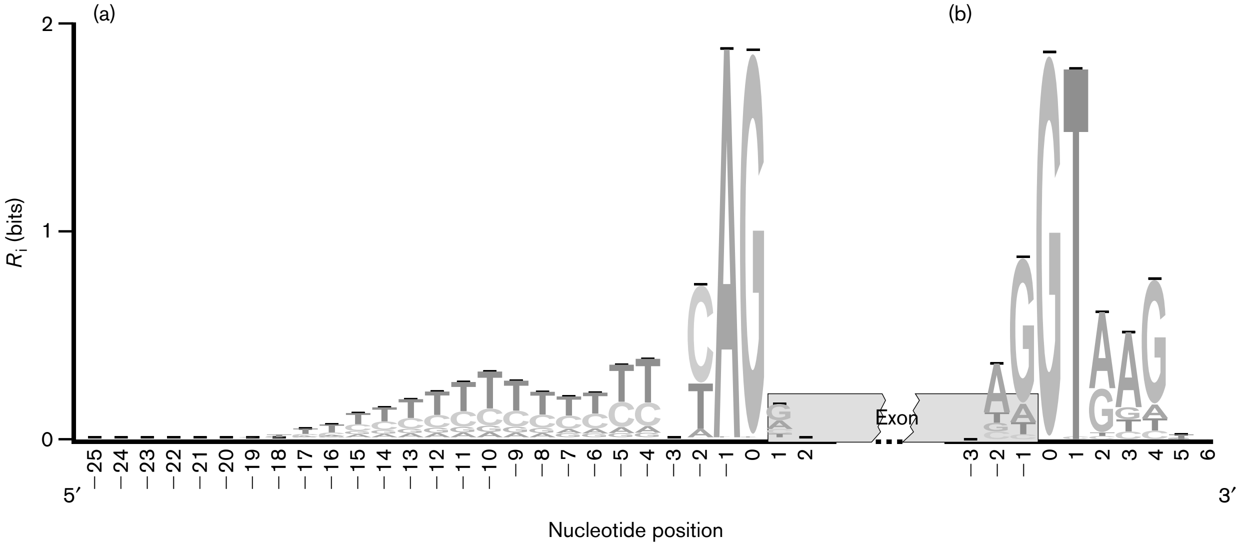
To estimate the minimum information content required for splice site recognition (e.g. $R_{i,\min}$ (previously determined to be approximately 2.4 bits)) [28], we reevaluated a large set of previously analysed, phenotypically correlated splicing mutations [28] with the updated information models. The new models redefined the minimum information content required for splice site recognition from 2.4 bits to 1.6 bits, a decrease of approximately 0.8 bits. This nominal difference does not alter the interpretation of either the present or previous individual information analyses of mRNA splicing.

R_i analyses of natural CYP splice sites

The individual information contents (R_i) of the donor and acceptor sites for the *CYP2C19*, *CYP2D6* and *CYP3A5* genes are presented in Table 1. All of the natural splice junctions in these genes contain canonical AG and GT dinucleotides at the splice junctions of the respective acceptor and donor sites. Nearly all of these sites exceeded the minimum information content required for splice site recognition. However, some splice sites had borderline R_i values, with very weak (*CYP2C19* acceptor site of exon 6, *CYP2D6* exon 6 donor and acceptor sites and acceptor site of exon 5 in the *CYP3A5* gene) or only small, positive R_i values in

*The Delila software for information analysis of protein and nucleic-acid sequences is described at http://www.lecb.ncifcrf.gov/~toms/delila.html. This software may be obtained from the National Institutes of Health (see http://www.lecb.ncifcrf.gov/~toms/walker/iipp.html). Splice sites may also be analysed on-line at http://www.lecb.ncifcrf.gov/~toms/delilaserver.html.

Fig. 1



Sequence logo for human (a) acceptor ($n = 53\ 985$) and (b) donor ($n = 56\ 286$) splice sites based on the (+) strand of the 7 October 2000 genome draft. The logo shows the distribution of information content (in bits) at each position over the region of 28 nucleotides for acceptor (−25, +2) and 10 nucleotides for donor (−3, +6) from the first nucleotide of the splice junction (position 0). The height of each nucleotide represents its frequency and the I-bar at the top of each stack indicates the standard deviation at the corresponding position.

the range from 0.8–1.6 bits. Unexpectedly, two acceptor sites (*CYP2D6*, exon 7 and *CYP3A5*, exon 12) were found to have negative information contents of −0.8 and −2.3 bits, respectively and, theoretically, they should not be recognized. It is notable that both sites contain highly unfavourable nucleotides (mostly purines instead of the preferred pyrimidines) upstream of the AG dinucleotide adjacent to the splice junction.

Because the cytochrome P450 family contains a number of pseudogenes [35], we were concerned about the possibility that published sequences of *CYP2D6* and *CYP3A5* might have been misassembled from closely related, but non-functional sequences. We therefore evaluated the splice site information contents of sequences of additional CYP genes (*CYP1A1*, *CYP1A2*, *CYP2A6*, *CYP2B6*, *CYP2C8*, *CYP2C9*, *CYP3A4*, *CYP3A7* and *CYP3A43*) at natural exon–intron junctions. We were surprised to find that the *CYP2B6*, *CYP2C8*, *CYP3A4*, *CYP3A7* and *CYP3A43* sequences each possess a single acceptor splice junction with a negative R_i value. For *CYP2B6*, the acceptor site of exon 2 had an R_i value < 0, while in the case of *CYP2C8*, a negative R_i value was predicted for the acceptor site of exon 6. All of the acceptor sites for exon 12 of genes in the *CYP3A* subfamily had negative R_i values (ranging from −4.3 bits for CYP3A4 to −2.0 for CYP3A43).

CYP splicing mutations

Information analyses of the *CYP2C19*, *CYP2D6* and *CYP3A5* splicing mutations are presented in Table 2. The *CYP2C19*2* mutation, 681G>A (G19154A), activates a novel, out-of-frame 7.6 bit cryptic acceptor site 40 nucleotides downstream from the corresponding natural acceptor site, whose strength (10.0 bits) is unaffected (Fig. 2a). Because this site is at least five-fold weaker than the natural site, it is somewhat surprising that this cryptic site is recognized efficiently; however, it is conceivable that other, as yet, unknown exon splicing regulatory elements adjacent to this site could strengthen it futher. The *CYP2C19*7* mutation, IVS5 + 2T>A (T19294A), abolishes splicing (a change in R_i value from 6.6 bits to −11.0 bits, i.e. $R_i < 0$ bits) at the natural donor site of exon 5. The *CYP2D6*4* mutation (IVS3−1G>A; G3465A) abolishes recognition of the exon 4 acceptor (a decrease from 5.8 bits to −5.5 bits), while strengthening a new 1.1 bit cryptic site one nucleotide downstream of the natural acceptor (Fig. 2b). In *CYP2D6*11*, the mutation IVS1-1G>C (G2502C) is predicted to abolish recognition of the exon 2 acceptor and results in skipping of this exon (decrease from 3.6 bits to −9.3 bits; not shown).

The *CYP3A5*3* mutation (6986A>G; A22893G in accession AC005020), located 236 nucleotides upstream

Table 1 **Information contents and coordinates of *CYP2C19*, *CYP2D6* and *CYP3A5* splice junctions**

Gene: Source:	*CYP2C19* AL133513, AL359672*		*CYP2D6* M33388		*CYP3A5* AC005020	
Exon	Acceptor	Donor	Acceptor	Donor	Acceptor	Donor
1		8.1 ± 0.02; 169		9.7 ± 0.01; 1800		3.2 ± 0.03; 15984
2	7.5 ± 0.04; 12352	5.3 ± 0.02; 12516	3.6 ± 0.04; 2502	3.0 ± 0.05; 2675	8.6 ± 0.04; 19600	10.6 ± 0.01; 19695
3	9.6 ± 0.04; 12684	5.8 ± 0.02; 12835	2.0 ± 0.04; 3224	5.8 ± 0.02; 3378	12.7 ± 0.03; 21223	7.5 ± 0.02; 21277
4	4.1 ± 0.04; 17793	9.8 ± 0.01; 17955	5.8 ± 0.04; 3465	5.5 ± 0.02; 3627	4.6 ± 0.04; 23129	7.5 ± 0.02; 23230
5	10.0 ± 0.04; 19115	6.6 ± 0.02; 19293	3.0 ± 0.04; 4059	9.5 ± 0.01; 4237	1.3 ± 0.04; 28743	6.8 ± 0.02; 28858
6	0.7 ± 0.04; 57790	7.3 ± 0.02; 57933	1.6 ± 0.04; 4426	1.2 ± 0.03; 4569	3.4 ± 0.04; 29119	6.2 ± 0.02; 29209
7	6.4 ± 0.04; 80131	11.2 ± 0.01; 80320	−0.8 ± 0.05; 4775	6.7 ± 0.02; 4964	8.7 ± 0.04; 30494	4.5 ± 0.02; 30644
8	8.7 ± 0.04; 87211	6.6 ± 0.02; 87354	5.3 ± 0.04; 5417	3.9 ± 0.03; 5560	12.8 ± 0.03; 31713	6.6 ± 0.02; 31842
9	11.2 ± 0.04; 90027		6.8 ± 0.04; 5657		8.2 ± 0.04; 32926	6.1 ± 0.02; 32994
10					11.9 ± 0.03; 35149	9.6 ± 0.01; 35311
11					14.5 ± 0.03; 43029	3.8 ± 0.02; 43257
12					−2.3 ± 0.05; 45576	7.3 ± 0.03; 45737
13					5.3 ± 0.04; 47408	

*Coordinates given are from the assembled sequence of the *CYP2C19* gene in Build 31 (November, 2002) of the human genome draft.

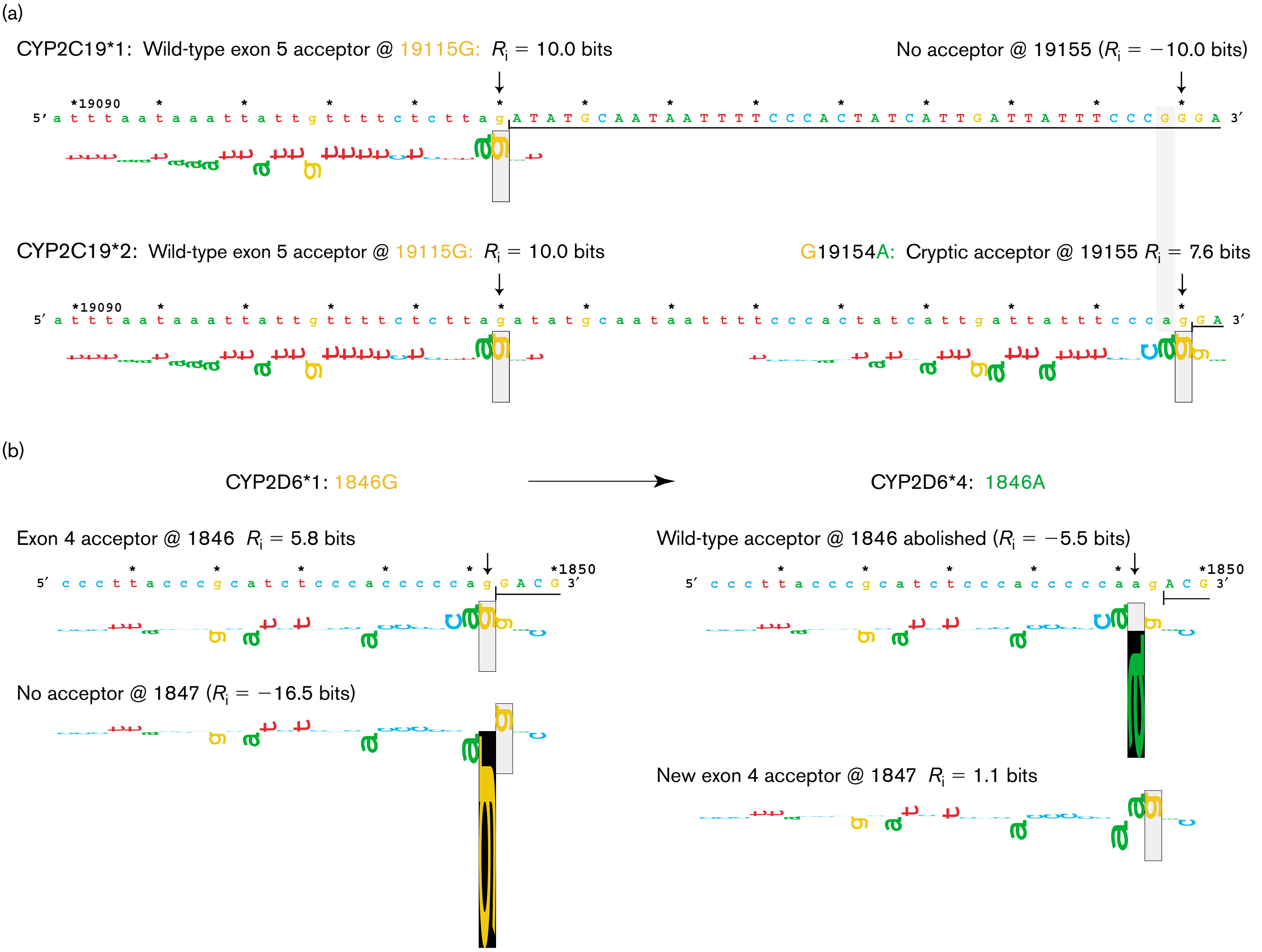
Table 2 **Information analysis of Cytochrome P450 gene splicing mutations**

No.	Allele	Designation	Location	Nucleotide change	Nucleotide position	R_i (wild type) → R_i (mutant)	Interpretation
1	CYP2C19*2	681G>A	Exon 5	G19154A	19115	A 10.0 ± 0.04 → 10.0 ± 0.04	No change (natural site)
					19155	A −10.0 ± 1.0 → 7.6 ± 0.04	Activates new cryptic site
2	CYP2C19*7	IVS5+2T>A	Intron 5	T19294A	19293	D 6.6 ± 0.02 → −11.0 ± 1.0	Inactivates natural site
3	CYP2D6*4	1846 G>A IVS3-1G>A	Intron 3	G3465A	3465	A 5.8 ± 0.04 → −5.5 ± 0.22	Inactivates natural site
					3466	A −16.5 ± 1.0 → 1.1 ± 0.04	Possibly activates new cryptic site
4	CYP2D6*11	883 G>C IVS1-1G>C	Intron 1	G2502C	2502	A 3.6 ± 0.04 → −9.3 ± 0.04	Inactivates natural site
5	CYP2D6*8	1758 G>T	Exon 3	G3377T	3378	D 5.8 ± 0.02 → 2.2 ± 0.03 ■	Mild mutation (natural site)
6	CYP2D6*14	1758 G>A	Exon 3	G3377A	3378	D 5.8 ± 0.02 → 2.8 ± 0.02 ▲	Mild mutation (natural site)
7	CYP3A5*3	6986 A>G	Intron 3	A22893G	22893	A −3.2 ± 0.22 → 8.2 ± 0.04	Activated new cryptic site
8	CYP3A5*5	12952 T>C	Intron 5	T28859C	28858	D 6.8 ± 0.02 → −0.2 ± 0.05	Inactivates natural site
9	CYP3A5*6	14690 G>A	Exon 7	G30597A	30494	A 8.7 ± 0.04 → 8.7 ± 0.04	No change (natural site)
					30622	A 2.9 ± 0.04 → 3.6 ± 0.04	Possibly activates cryptic site
					30644	D 4.5 ± 0.02 → 4.5 ± 0.02	No change (natural site)

A - acceptor, D - donor site, ■ fold change ⩾ 12x, ▲ fold change ⩾ 8x.

from the exon 4 acceptor activates a new strong cryptic site (increase from −3.2 to 8.2 bits) at this position, while the natural acceptor site remains unchanged at 4.6 bits. This change results in the inclusion of a non-canonical cryptic exon 131 nucleotides in length, termed exon 3B by Kuehl *et al.* [7], that is bounded at the 3′ end by a preexisting cryptic 4.4 bit donor site at position 23026 (Fig. 3). The *CYP3A5*3* allele is also associated with transcripts containing intron-derived cryptic exons denoted 4B and 5B (Fig. 4a) (SV3 mRNA in Kuehl *et al.* [7]). Exon 4B appears to be activated by substitution of the wild-type adenosine with guanosine at position 24035. This substitution creates a 3.5 bit donor site (whereas the wild-type donor site is 0.1 bits) which, in combination with a preexisting 5.4 bit acceptor at 23930, results in the inclusion of this cryptic exon in the mature transcript (Fig. 4b). Inclusion of both exons 3B and 4B in the same transcript would result from inheritance of a haplotype consisting of both 22893G and 24035G variants. By contrast to exons 3B and 4B, the cryptic splice sites in exon 5B are not activated by a specific change in the nucleotide sequence in IVS4. Inclusion of cryptic exon 5B was also shown to occur concomitant with skipping of wild-type exon 6. Information analysis indicates that the exon 5B donor site at position 29097 (3.8 bits) overlaps the

Fig. 2


(a)
CYP2C19*1: Wild-type exon 5 acceptor @ 19115G: R_i = 10.0 bits
No acceptor @ 19155 (R_i = −10.0 bits)
*19090
5′ a t t t a a t a a a t t a t t g t t t t c t c t t a g A T A T G C A A T A A T T T T C C C A C T A T C A T T G A T T A T T T C C C G G G A 3′
CYP2C19*2: Wild-type exon 5 acceptor @ 19115G: R_i = 10.0 bits
G19154A: Cryptic acceptor @ 19155 R_i = 7.6 bits
*19090
5′ a t t t a a t a a a t t a t t g t t t t c t c t t a g a t a t g c a a t a a t t t t c c c a c t a t c a t t g a t t a t t t c c c a g G A 3′
(b)
CYP2D6*1: 1846G
CYP2D6*4: 1846A
Exon 4 acceptor @ 1846 R_i = 5.8 bits
*1850
5′ c c c t t a c c c g c a t c t c c c a c c c c c a g G A C G 3′
No acceptor @ 1847 (R_i = −16.5 bits)
Wild-type acceptor @ 1846 abolished (R_i = −5.5 bits)
*1850
5′ c c c t t a c c c g c a t c t c c c a c c c c c a a g A C G 3′
New exon 4 acceptor @ 1847 R_i = 1.1 bits

Information changes resulting from mutations in CYP genes. Sequence walkers [22], illustrate the information contents of corresponding natural and mutant splice sites. The height of a letter is the contribution of that particular base to the total conservation of the site. The upper bound of the vertical rectangle at the splice junction is at +2 bits and the lower bound is at −5 bits. Letters that are point downwards represent negative contributions of these bases to the overall information content of that site. The coding region is indicated by upper case letters. (a) *CYP2C19*2*: the walker for the wild-type exon 5 acceptor at nucleotide 19155 is presented for both the *CYP2C19*1* and *CYP2C19*2* alleles. The shaded rectangle indicates the position of the splice site created by the 19154G>A in the *CYP2C19*2* coding region variant. (b) *CYP2D6*4*: the wild-type exon 4 acceptor (1846G, left side of panel) is abolished in the presence of the 1846A mutation characteristic of the *CYP2D64* allele concurrent with activation of a new exon 4 acceptor at nucleotide 1847 (right side of panel). Accession numbers for the sequences utilized are given in Table 1.

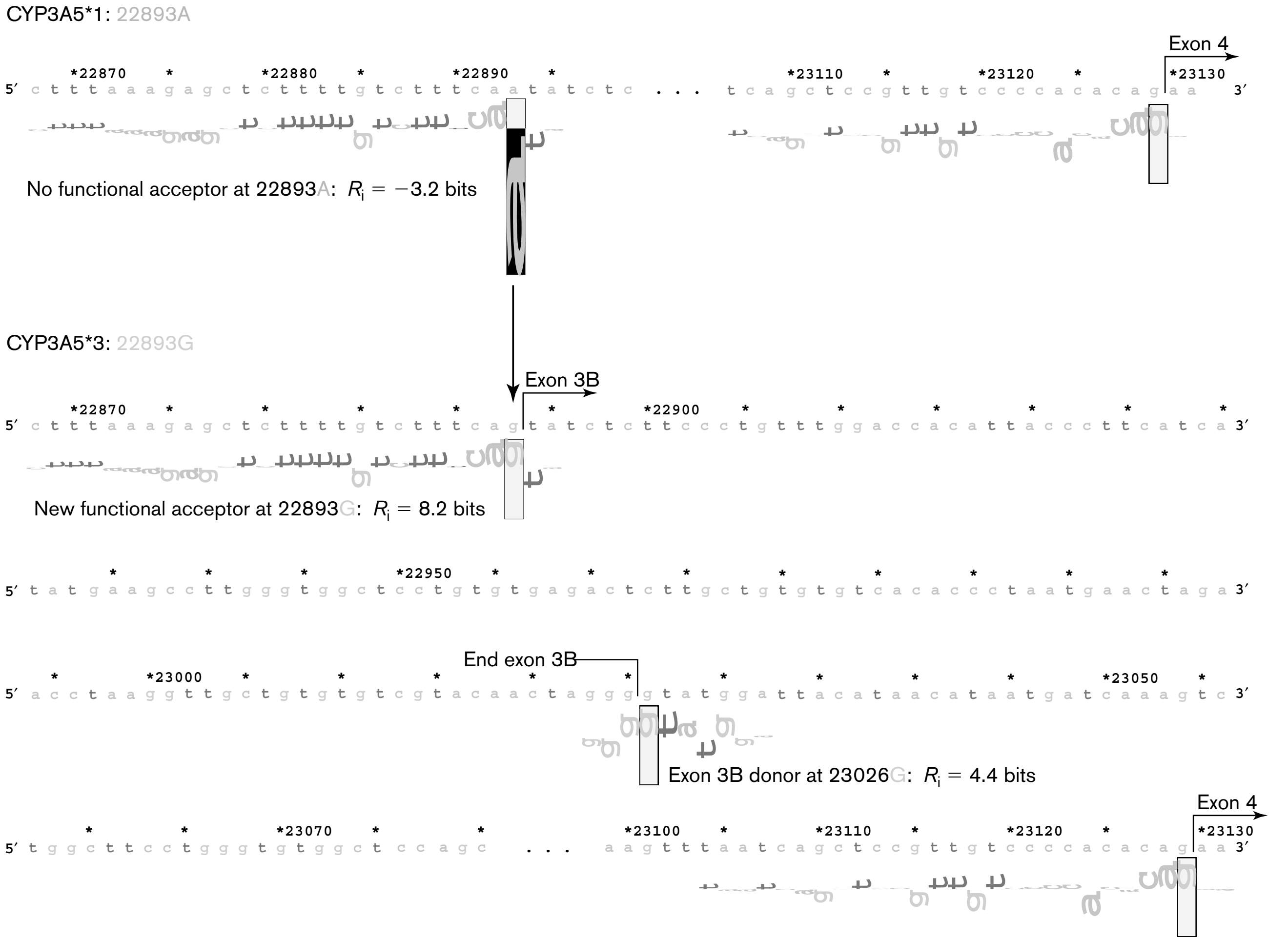
downstream acceptor of this adjacent exon at position 29119 (3.4 bits; Fig. 4c). Activation of the exon 5B donor site impairs recognition of the downstream acceptor site, thus preventing incorporation of exon 6.

The *CYP3A5*5* mutation, 12952T>C (T28859C in accession AC005020), located immediately adjacent to the natural donor site, most likely abolishes splicing (decrease from 6.8 bits to −0.2 bits). By contrast, the *CYP3A5*6* mutation, 14685G>A (G30597A), is predicted to have only a minor effect on splicing, strengthening a cryptic acceptor site 22 nucleotides upstream of the exon 7 donor site from 2.9 to 3.6 bits (as well as a weak cryptic donor from 0.7 to 1.2 bits that is 44 nucleotides upstream of this natural donor site). It is not obvious why the *CYP3A5*6* allele induces skipping of exon 7; however, the mutation appears to be mild. In *CYP3A5*1/CYP3A5*6* heterozygotes (Fig. 1b of Kuehl *et al.* [7]), the wild-type transcript is considerably more abundant than that synthesized from the mutant allele, which may be consistent with leaky splicing of exon 7 in *CYP3A5*6* rather than complete exon skipping.

R_i analysis of missense mutations affecting splice sites

We previously reported that mutations with $R_i < 2.4$ bits inactivate splice sites, often producing severe

Fig. 3


CYP3A5*1: 22893A
Exon 4
5′ ctttaaagagctcttttgtctttcaatatctc . . . tcagctccgttgtccccacacagaa 3′
No functional acceptor at 22893A: $R_i = -3.2$ bits
CYP3A5*3: 22893G
Exon 3B
5′ ctttaaagagctcttttgtctttcagtatctcttccctgtttggaccacattacccttcatca 3′
New functional acceptor at 22893G: $R_i = 8.2$ bits
5′ tatgaagccttgggtggctcctgtgtgagactcttgctgtgtgtcacaccctaatgaactaga 3′
End exon 3B
5′ acctaaggttgctgtgtgtcgtacaactagggtatggattacataacataatgatcaaagtc 3′
Exon 3B donor at 23026G: $R_i = 4.4$ bits
Exon 4
5′ tggcttcctgggtgtggctccagc . . . aagtttaatcagctccgttgtccccacacagaa 3′

Sequence logos for the *CYP3A5*3* allele. Sequence walkers and logos are as described in Fig. 2. Mutation 22893A>G creates a new functional acceptor at position 22893 which, together with an existing donor site at 23026G, gives rise to an inserted exon designated exon 3B by Kuehl *et al.* [7]. Sequence numbering according to accession number AC005020.

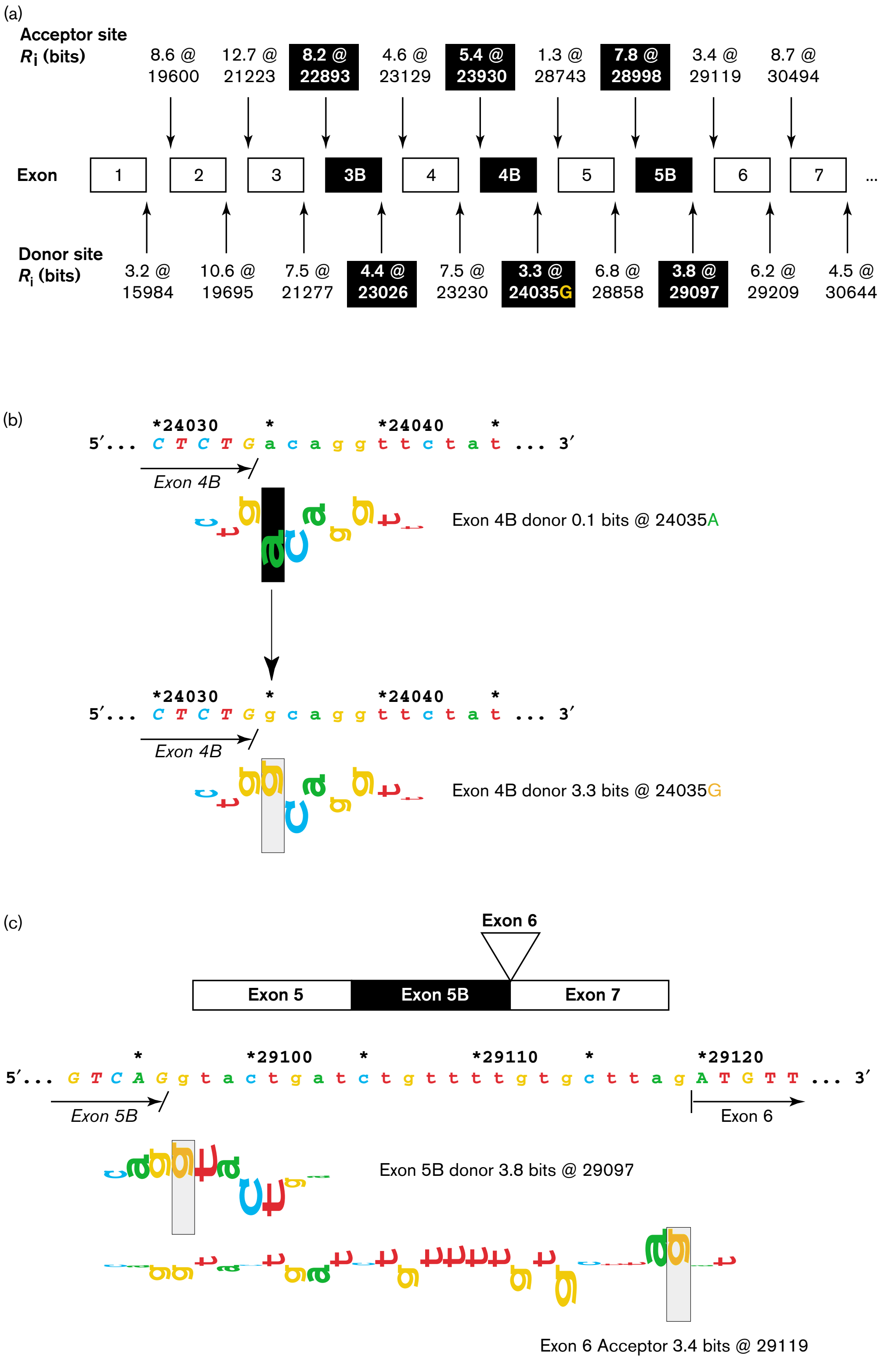
phenotypes while, in contrast, nucleotide substitutions with R_i values exceeding this threshold may reduce but do not necessarily abolish splicing, often producing milder phenotypes [29]. Assuming the mutated splice site is the limiting factor in defining the exon, the residual amount of correctly spliced (functional) mRNA can be estimated from the minimum-fold change in binding affinity ($\geqslant 2^{\Delta R_i}$), where ΔR_i represents the difference between R_i values of the wild and mutated variant. Results are usually expressed as the maximum percentage of normal mRNA product synthesized.

*CYP2D6*8* (1758G>T) and *CYP2D6*14* (1758G>A) both significantly reduce the strength of the adjacent exon 3 donor site. In the case of the 1758G>T transversion characteristic of *CYP2D6*8*, the strength of the natural exon 3 donor site is decreased from an initial value of 5.8 bits to 2.2 bits, corresponding to an at least 12-fold reduction in the strength of this site. For *CYP2D6*14*, the missense 1758G>A mutation has a similar effect (a decrease from 5.8 to 2.8 bits), but reduces splice site use at least eight-fold. Thus, exon 3 skipping would be predicted to be the predominant consequence of these mutations, and full-length transcripts containing these nonsense and missense mutations would be expected to comprise a minor proportion of the corresponding mRNA.

Discussion

We have analysed mutations and other sequence variants in several members of the cytochrome P450 gene family using information theory-based models of donor and acceptor splice sites. To derive these models, we extracted unique acceptor and donor splice junctions from genes on the given strand of the October, 2000 version of the human genome draft sequence. While the overall information weight matrices are consistent with previous studies [34], the current models exhibit a

Fig. 4



(a) Information content (in bits) for donor and acceptor sites of exons 1–7 and splice variants (exons 3B, 4B and 5B) of *CYP3A5*. The donor splice site for exon 4B has an R_i value of 0.1 bits in the presence of the wild-type nucleotide 24035A but is increased to 3.5 bits in the presence of the variant 24035G. (b) Sequence walkers for the exon 4B donor sites –24035A and 24035G. (c) The sequence walker for the exon 5B donor site (3.8 bits at 29097) overlaps with the exon 6 acceptor (3.4 bits at 29119). Overlapping sites could impair splicing and lead to skipping of exon 6. Sequence numbering according to accession number AC005020.

tighter distribution of R_i values and reduced standard deviations (approximately 1 bit for donor and the acceptor) around the average information content, $R_{sequence}$, which presumably reflects the significant increase in the numbers of splice sites. Standard deviations for individual information values are small, resulting in narrow confidence intervals around R_i values (all < 0.05 bits) and making all differences in information contents of mutant and cognate wild-type splice sites statistically significant.

Reevaluation of previously studied mutations with the updated splice site models for donor/acceptor splice sites agreed with our previous findings [28] and redefined the minimum value required for the splice site recognition to approximately 1.6 bits. Interestingly, this value is closer to the theoretical threshold of zero bits. Thus, the addition of new sites to the model has more clearly delineated the cut-off between true binding sites and sequences unlikely to be bound by the splicesome.

It was surprising to find two acceptor sites with negative information contents in exon 7 of *CYP2D6* ($R_i = -0.8$ bits) and exon 12 of *CYP3A5* ($R_i = -2.3$ bits). In addition, each of the three genes contained sites (*CYP2C19* exon 6 acceptor, *CYP2D6* exon 6 donor and *CYP3A5* exon 5 acceptor) with R_i values less than the minimum value of a functional splice site.

The *CYP* superfamily is the first one in which multiple functional splice sites with negative information contents have been described. A single splice site with negative information content (−0.1 bits) has been reported at the exon 2 acceptor of the *XPC* gene [36]. However, a small positive R_i value (+0.1 bits) was found by reevaluating this sequence with the splice acceptor weight matrix derived in the present study. Nevertheless, sites with R_i values less than the minimum functional R_i (1.6 bits) are extremely rare, comprising only 0.0008% of all validated donors ($n = 218$) and none of the validated acceptors in the present splice site models.

A functional binding site can exhibit negative information if the site is being recognized by a factor(s) different from the one that defines the information matrix. In this instance, these acceptor sites in the *CYP* family do not appear to be substrates for the U12 splicesome, which contains a completely different set of conserved nucleotides at intron–exon junctions [37]. It also seems unlikely sequencing errors could account for all of the sites with negative R_i values, and the conservation of negative information at the same exon junction in *CYP3A* orthologs does not appear to be coincidental.

We hypothesize that a novel splicing regulatory mechanism(s) compensates for weakness of these sites and promotes their recognition (e.g. through the presence of an uncharacterized splice enhancer sequence or by structural, rather than sequence-specific, recognition of the splice junction). The exon 5 acceptor in CYP3A5 may be an example of weak splice sites (or sites with negative R_i values) that is particularly susceptible to alternative splicing and skipping of the exons. However, if negative information content sites were recognized by a tissue-specific splicing factor, post-transcriptional regulatory control could be exerted by limiting translation of functional gene products to cells expressing the factor. In other tissues, the exon containing such sites would likely be skipped. The resultant mRNA would most likely not encode a functional protein, and may be degraded prior to translation via nonsense-mediated decay.

Mutations occurring at splice sites in the *CYP2C19*, *CYP2D6* and *CYP3A5* genes associated with a poor metabolizer phenotype were analysed by comparison of the wild-type and mutant genomic sequences. As expected, the *CYP2C19*2*, *CYP2C19*7*, *CYP2D6*4*, *CYP2D6*11*, *CYP3A5*3*, *CYP3A5*5* and *CYP3A5*6* mutations inactivate or reduce recognition of the cognate wild-type sites or activate cryptic sites, thus producing aberrantly processed mRNAs encoding truncated, non-functional proteins.

In *CYP2C19*2*, a 7.6 bit cryptic acceptor is activated 40 nucleotides downstream from the 5′ end of exon 5, although this site is weaker than the natural site ($R_i = 10.0$ bits). Interestingly, the wild-type sequences of other members of this gene subfamily, *CYP2C8* and *CYP2C18*, also contain potential acceptor splice sites at the corresponding position (8.1 bits for *CYP2C8*; 6.3 bits for *CYP2C18*); however, these sites are rarely used [13]. Evaluating splice junctions of the *CYP2C8* and *CYP2C18* genes, we determined that strong exon 5 acceptor sites were present for both genes, having R_i values of 14.8 and 12.3 bits, respectively. However, the differences in R_i values for the natural exon 5 and the cryptic acceptor sites for the *CYP2C8* ($\Delta R_i = 14.8-8.1 = 6.7$ bits; $2^{\Delta R_i} \geqslant 104$-fold difference in binding strength) and *CYP2C18* genes ($\Delta R_i = 6.0$ bits; $\geqslant 64$-fold difference) were considerably greater than the 2.4-bit difference observed for the corresponding sites in *CYP2C19*2*. Based on an analysis of expressed sequence tags encoded by *CYP2C* subfamily members, splicing occurs predominantly at the natural splice site (as expected); however, use of the cryptic site within exon 5 has been documented in at least one instance (in a hepatocellular carcinoma; dbEST accession AV688442) of *CYP2C18*.

Information theory-based analysis provides compelling explanations for the various splice variants described for

CYP3A5, including the insertion of exon 3B into the three CYP3A5*3 splice variants (SV1, SV2 and SV3), the insertion of exon 4B in splice variant 2 (AF355801) and the insertion of exon 5B/deletion of exon 6 in splice variant 3 (AF355802) reported by Kuehl *et al.* [7]. We previously showed that a cryptic exon in the *CFTR* gene activated by a downstream splicing mutation is bounded by splice donor and acceptor sites with R_i values exceeding the minimum information content [28] (Table 2, mutation #3). Exons in alternatively spliced transcripts that are skipped exhibit generally weaker acceptor sites than downstream acceptors that are recognized and spliced, and the respective differences in information contents are related to the prevalence of exon skipping [38]. By analogy, the R_i values of acceptor sites of non-canonical exons in *CYP3A5*3* that are ectopically spliced into the wild-type transcript and generate the SV2, SV3 and SV4 splice forms of Kuehl *et al.* [7] exceed those of the adjacent downstream acceptors of the wild-type exons. The results of the current and previous study are consistent, in that both suggest that acceptor splice site strength significantly influences the dynamics of exon inclusion and, as a consequence, the stochiometries of alternative splice forms.

Our analyses suggest that exons 3B and 4B are incorporated in the *CYP3A5*3* transcript by independent cryptic splicing mutations, respectively, in IVS3 and IVS4. An analysis of expressed sequence tags (http://www.ncbi.nlm.nih.gov/dbEST/) indicates that each of these exons are incorporated in some CYP3A5 transcripts to the exclusion of the other. Four tags (GenBank accessions BG565695, AV704349, AV660478 and AI310154) contain exon 4B and wild-type exons only, whereas one transcript (X90579) contains exons 3B and 5B and wild-type exons, but does not include exon 4B. Thus, inclusion of exons 3B and 4B does not appear to be a concerted event. Independent segregation of 22893G and 24035G variants could explain the existence of transcripts containing either, but not both of these exons; however, proof that the *CYP3A5*3* allele is itself comprised of multiple alleles will require examination of splicing patterns for each of the possible haplotypes.

The concomitant inclusion of exon 5B and exclusion of exon 6 from the SV3 isoform of the *CYP3A5*3* allele can also be explained by information analysis. Overlapping natural and cryptic acceptor sites in the IVD gene with similar information contents can interfere with splice site recognition [31]. In that study, we found exonic mutations (149G>C and 148 C>T) that strengthened the cryptic site and prevented either site from being recognized appropriately. This led to loss of definition of the exon, and skipping of this exon during splicing. In *CYP3A5*3*, overlap of the exon 5B donor and exon 6 acceptor sites appears to prevent the simultaneous recognition of both sites. Despite the fact that both sites are of similar strength, the exon 5B donor is preferentially recognized, possibly because the splicesome is already bound to the exon 5B acceptor.

Even though the *CYP3A5*6* allele interferes with definition of exon 7, the mutation strengthens the adjacent cryptic sites by only small amounts (1.6- and 1.4-fold, respectively, for the acceptor and donor sites). The mutation would not be expected to activate these cryptic sites because both variant sites still contain less information than either the natural donor or acceptor sites, whose R_i values are unaltered by the mutation. Disruption of interactions between SR proteins and an exonic splicing enhancer remains a plausible explanation for exon skipping. However, information analysis of exon 7 with preliminary models of SR protein binding sites (i.e. SC35 and ASF/SF2) failed to reveal any change in the strengths or distribution of these sites as a result of the *CYP3A5*6* mutation (P. Rogan, unpublished data).

Mutations involving the terminal codon of exon 3, *CYP2D6*14* and *CYP2D6*8* are predicted to be leaky and to be consistent with decreased (but not totally abolished) mRNA production. The reduced activity of the *CYP2D6*14* (1758G>A) gene product has been attributed to a glycine to arginine change at amino acid position 169 [39]. However, information analysis of the nonsense mutation, *CYP2D6*8*, and the missense change, *CYP2D6*14*, at this position suggests that the splice site use is decreased $\geqslant$ 12- and $\geqslant$ eight-fold, respectively, and mRNA would be expected to be approximately 8% of the normal levels of transcript. Interestingly, this transcript remains in frame, and in theory, could be translated into a protein 50 amino acids shorter than the full-length enzyme with a novel codon (i.e. an amino acid substitution, Gly, at the position of the junction between exons 2 and 4). Alignment of the *CYP2D6* amino acid sequence with that of CYP102 [40] reveals that the deleted amino acids correspond to the C and D helices of CYP102 [41]. In addition to the major changes in structure that would be expected to dramatically alter catalytic activity, loss of the conserved W-X-X-X-R motif in particular would preclude the normal interaction between tryptophan and arginine of this motif and the proprionate side chain of the prosthetic heme [41]. Thus, the poor metabolizer phenotype of these mutations may be predominantly the result of the decreased enzymatic activity of the internally deleted *CYP2D6* protein, an eight- to 12-fold reduction of intact mRNA or a combination of both effects.

We have shown that individual information analysis can be used to interpret the majority of known genomic mutations responsible for abnormal and non-canonical splicing of cytochrome P450 transcripts. This approach may also find application in evaluating genetic variants identified by DNA sequencing of other drug metabolizing genes. Given the inherently high density of polymorphism at these loci, information analysis may be particularly valuable in efficiently and economically prioritizing potentially deleterious mutations for subsequent functional studies.

Acknowledgements

We thank Tom Schneider for valuable comments on the manuscript and Joseph Lucke for assistance with statistical analysis of individual information.

References

1 Mahgoub A, Idle JR, Dring LG, Lancaster R, Smith RL. Polymorphic hydroxylation of debrisoquine in man. *Lancet* 1977; **1**:584–586.

2 Eichelbaum M, Spannbrucker N, Steincke B, Dengler HJ. Defective N-oxidation of sparteine in man: a new pharmacogenetic defect. *Eur J Clin Pharmacol* 1979; **16**:183–187.

3 Bertilsson L, Lou Y-Q, Du Y-L, Liu Y, Kuang T-Y, Liao X-M, *et al*. Pronounced differences between native Chinese and Swedish populations in the polymorphic hydroxylations of debrisoquin and S-mephenytoin. *Clin Pharmacol Ther* 1992; **51**:388–397.

4 Küpfer A, Preisig R. Pharmacogenetics of mephenytoin: a new drug hydroxylation polymorphism in man. *Eur J Clin Pharmacol* 1984; **26**:753–759.

5 Goldstein JA, de Morais SMF. Biochemistry and molecular biology of the human *CYP2C* subfamily. *Pharmacogenetics* 1994; **4**:285–299.

6 Evans WE, Relling MV. Pharmacogenomics: translating functional genomics into rational therapeutics. *Science* 1999; **286**:487–491.

7 Kuehl P, Zhang J, Lin Y, Lamba J, Assem M, Schuetz J, *et al*. Sequence diversity in *CYP3A* promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. *Nature Genet* 2001; **27**:383–391.

8 Marez D, Legrand M, Sabbagh N, Lo Guidice J-M, Spire C, Lafitte J-J, *et al*. Polymorphism of the cytochrome P450 *CYP2D6* gene in a European population: characterization of 48 mutations and 53 alleles, their frequencies and evolution. *Pharmacogenetics* 1997; **7**:193–202.

9 Sachse C, Brockmöller J, Bauer S, Roots I. Cytochrome P450 2D6 variants in a Caucasian population: allele frequencies and phenotypic consequences. *Am J Hum Genet* 1997; **60**:284–295.

10 Griese E-U, Zanger U, Brudermanns U, Gaedigk A, Mikus G, Mörike K, *et al*. Assessment of the predictive power of genotypes for the in-vivo catalytic function of CYP2D6 in a German population. *Pharmacogenetics* 1998; **8**:15–26.

11 Gaedigk A, Gotschall RR, Forbes NS, Simon SD, Kearns GL, Leeder JS. Optimization of cytochrome P450 2D6 (CYP2D6) phenotype assignment using a genotyping algorithm based on allele frequency data. *Pharmacogenetics* 1999; **9**:669–682.

12 Marez D, Sabbagh N, Legrand M, Lo Guidice JM, Boone P, Broly F. A novel *CYP2D6* allele with an abolished splice recognition site associated with the poor metabolizer phenotype. *Pharmacogenetics* 1995; **5**:305–311.

13 de Morais SMF, Wilkinson GR, Blaisdell J, Nakamura K, Meyer UA, Goldstein JA. The major genetic defect responsible for the polymorphism of *S*-mephenytoin metabolism in humans. *J Biol Chem* 1994; **269**: 15419–15422.

14 Ibeanu GC, Blaisdell J, Ferguson RJ, Ghanayem BI, Brøsen K, Benhamou S, *et al*. A novel transversion in the intron 5 donor splice junction of *CYP2C19* and a sequence polymorphism in exon 3 contribute to the poor metabolizer phenotype for the anticonvulsant drug *S*-mephenytoin. *J Pharmacol Exp Ther* 1999; **290**:635–640.

15 Brøsen K, de Morais SMF, Meyer UA, Goldstein JA. A multifamily study on the relationship between *CYP2C19* genotype and *S*-mephenytoin oxidation phenotype. *Pharmacogenetics* 1995; **5**:312–317.

16 Bathum L, Skjelbo E, Mutagingwa TK, Madsen H, Hørder M, Brøsen K. Phenotypes and genotypes for CYP2D6 and CYP2C19 in a black Tanzanian population. *Br J Clin Pharmacol* 1999; **48**:395–401.

17 Sviri S, Shpizen S, Leitersdorf E, Levy M, Caraco Y. Phenotypic–genotypic analysis of CYP2C19 in the Jewish Israeli population. *Clin Pharmacol Ther* 1999; **65**:275–282.

18 Gellner K, Eiselt R, Hustert E, Arnold H, Koch I, Haberl M, *et al*. Genomic organization of the human *CYP3A* locus: identification of a new inducible *CYP3A* gene. *Pharmacogenetics* 2001; **11**:111–121.

19 Chou F-C, Tzeng S-J, Huang J-D. Genetic polymorphism of cytochrome P450 3A5 in Chinese. *Drug Metab Disp* 2001; **29**:1205–1209.

20 Gonzalez FJ, Skoda RC, Kimura S, Umeno M, Zanger UM, Nebert DW, *et al*. Characterization of the common genetic defect in humans deficient in debrisoquine metabolism. *Nature* 1988; **331**:442–446.

21 Schneider TD. Information content of individual genetic sequences. *J Theoret Biol* 1997; **189**:427–441.

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

23 Berget SM. Exon recognition in vertebrate splicing. *J Biol Chem* 1995; **270**:2411–2414.

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

25 Allikmets R, Wasserman WW, Hutchinson A, Smallwood P, Nathans J, Rogan RK, *et al*. Organization of the *ABCR* gene: analysis of promoter and splice junction sequences. *Gene* 1998; **215**:111–122.

26 Kannabiran C, Rogan PK, Basti S, Rao GN, Kaiser-Kupfer M, Hejtmancik JF. Autosomal dominant zonular cataract with sutural opacities is associated with a splice mutation in the A3/A1-crystallin gene. *Mol Vision* 1998; **4**:21–26.

27 O'Neill JP, Rogan PK, Cariello N, Nicklas JA. Mutations that alter RNA splicing of the human *HPRT* gene: a review of the spectrum. *Mutation Res* 1998; **411**:179–214.

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

29 von Kodolitsch Y, Pyeritz RE, Rogan PK. Splice-site mutations in atherosclerosis candidate genes. Relating individual information to phenotype. *Circulation* 1999; **100**:693–699.

30 Svojanovsky SR, Schneider TD, Rogan PK. Redundant designations of BRCA1 intron 11 splicing mutation. *Hum Mut* 2000; **16**:264.

31 Vockley J, Rogan PK, Anderson BD, Willard J, Seelan RS, Smith DI, *et al*. An unusually high frequency of abnormal splicing of IVD RNA in isovaleric acidemia, including exon skipping caused by missense mutations in the IVD gene. *Am J Hum Genet* 2000; **66**:356–367.

32 Kimura S, Umeno M, Skoda RC, Meyer UA, Gonzalez FJ. The human debrisoquine 4-hydroxylase (*CYP2D*) locus: sequence and identification of the polymorphic *CYP2D6* gene, a related gene, and a pseudogene. *Am J Hum Genet* 1989; **45**:889–904.

33 Hanioka N, Kimura S, Meyer UA, Gonzalez FJ. The human *CYP2D* locus associated with a common genetic defect in drug oxidation: A G_{1934}>A base change in intron 3 of a mutant *CYP2D6* allele results in an aberrant 3′ splice recognition site. *Am J Hum Genet* 1990; **47**:994–1001.

34 Stephens RM, Schneider TD. Features of splicesome evolution and function inferred from an analysis of information at human splice sites. *J Mol Biol* 1992; **228**:1124–1136.

35 Heim MH, Meyer UA. Evolution of highly polymorphic human cytochrome P450 gene cluster: CYP2D6. *Genomics* 1992; **14**:49–58.

36 Khan SG, Muniz-Medina V, Shahlavi T, Baker CC, Inui H, Ueda T, *et al*. The human *XPC* DNA repair gene: arrangement, splice site information content and influence of a single nucleotide polymorphism in a splice acceptor site on alternative splicing and function. *Nucl Acids Res* 2002; **30**:3624–3631.

37 Hall SL, Padgett RA. Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites. *J Mol Biol* 1994; **239**: 357–365.

38 Thompson TE, Rogan PK, Risinger JI, Taylor JA. Splice variants but not mutations of DNA polymerase b are common in bladder cancer. *Cancer Res* 2002; **62**:3251–3256.

39 Wang S-L, Lai M-D, Huang J-D. G169R mutation diminishes the metabolic activity of CYP2D6 in Chinese. *Drug Metab Disp* 1999; **27**: 385–388.

40 Lewis DFW, Eddershaw PJ, Goldfarb PS, Tarbit MH. Molecular modelling of cytochrome P4502D6 (CYP2D6) based on an alignment with CYP102: structural studies on specific CYP2D6 substrate metabolism. *Xenobiotica* 1997; **27**:319–340.

41 Hasemann CA, Kurumbail RG, Boddupalli SS, Peterson JA, Deisenhofer J. Structure and function of cytochromes P450: a comparative analysis of three crystal structures. *Structure* 1995; **3**:41–62.

Appendix

Development and validation of information theory-based matrices of donor and acceptor splice sites using the human genome working draft sequence.

A non-redundant set of accurately localized splice sites was derived from all donor and acceptor splice sites from the given (or +) strand of known genes of the 10/7/00 Human (http://genomearchive.cse.ucsc.edu/goldenPath/07oct2000/database) Genome Working Draft sequence. The initial database comprised 153 231 donor and an equal number of acceptor sites comprising, in some instances, different GenBank accessions corresponding to different cDNAs aligned with the identical gene sequence. Splice sites of the same type (donor or acceptor) less than 3 nucleotides apart were considered to be duplicate (or multiplicate) sites. 86 068 acceptor and 86 221 donor segments remained in the set, once these duplicate sites were removed.

Based on our previous studies [28,34], a sequence window 28 nucleotides in length spanning positions −25 to +2 relative to the exon–intron splice junction at position 0 was used to define the acceptor information weight matrix while the donor matrix was computed over a 10 nucleotide sequence window from position −3 to +6. The presumptive coordinates of exon junctions from mRNAs mapped onto the genome draft sequence were parsed from the all_mrna.txt annotation table at http://genome-archive.cse.ucsc.edu/goldenPath/07oct2000/database. Accurately localized exon junctions were identified by iteratively recomputing the weight matrix, and selecting all sites with positive R_i values. According to information theory, binding of sites with negative R_i values is energetically unfavourable ($\Delta G \geqslant 0$). After eight cycles of removing sites with $R_i < 0$, the final models were based upon 53 985 unique acceptor and 56 286 donor sites (Appendix: Tables 1 and 2). Sites with negative R_i values are the result either of misalignment of the mRNA and genomic sequences or are recognized by proteins having a different binding pattern. The coordinates of intron–exon junctions in the genome draft sequence may be inaccurate due to the rapid alignment procedures (comparing cDNA and genomic sequences with algorithms such as BLAT or BLAST) used to determine the locations of these sites. However, atypical functional splice sites were eliminated by this procedure include those recognized by the U12 splicesome, which comprise a minority of sites in the genome.

Table 1 **Weight matrix $R_i(b,l)$ for acceptor sites**

R_i (a,l)	R_i (c,l)	R_i (g,l)	R_i (t,l)	l
−0.107075	−0.117162	−0.736935	0.383654	−25
−0.112706	−0.125239	−0.763776	0.405411	−24
−0.143967	−0.098777	−0.812644	0.429678	−23
−0.181144	−0.074375	−0.816704	0.438961	−22
−0.233510	−0.069523	−0.850323	0.482276	−21
−0.333400	−0.035904	−0.859398	0.520905	−20
−0.439090	−0.049756	−0.852315	0.582583	−19
−0.600740	−0.007482	−0.897213	0.645200	−18
−0.770779	0.036919	−0.937582	0.696879	−17
−0.909978	0.037505	−0.958881	0.750295	−16
−1.060285	0.064146	−1.018417	0.796151	−15
−1.187987	0.122464	−1.107940	0.818572	−14
−1.316805	0.089079	−1.187987	0.888378	−13
−1.466006	0.074762	−1.249169	0.940551	−12
−1.607146	0.088512	−1.401490	0.989186	−11
−1.725607	−0.023814	−1.465174	1.075143	−10
−1.697994	0.093035	−1.360973	0.993476	−9
−1.509629	0.152930	−1.274678	0.909795	−8
−1.372359	0.294579	−1.391451	0.816522	−7
−1.300152	0.337939	−1.621278	0.815610	−6
−1.685352	0.388875	−2.213358	0.940499	−5
−1.671860	0.176031	−2.182965	1.064386	−4
−0.169585	0.082080	−0.426492	0.088701	−3
−2.264999	1.280768	−7.436372	0.152659	−2
1.904015	−15.720324	−15.720324	−6.311565	−1
−9.419094	−11.004056	1.903478	−6.376783	+0
−0.055063	−0.949161	0.899288	−1.288094	+1
−0.098454	−0.508344	−0.455097	0.473630	+2

7.448748 bits = mean (R_{sequence}); 22.933991 bits = R_i of consensus sequence. −63.849965 bits = R_i of anticonsensus sequence. −25.268638 bits = average R_i for random sequence. 53985 = number of sequences.

Table 2 **Weight matrix $R_i(b,l)$ for donor sites**

R_i (a,l)	R_i (c,l)	R_i (g,l)	R_i (t,l)	l
0.283450	0.396834	−0.568674	−1.220276	−3
1.229685	−1.397305	−1.276078	−1.001612	−2
−1.456455	−3.334596	1.551773	−2.047699	−1
−15.780540	−15.780540	1.862810	−15.780540	0
−15.780540	−5.149494	−15.780540	1.851591	+1
1.123510	−3.416837	0.340151	−3.338363	+2
1.349145	−1.915567	−1.266851	−1.290145	+3
−1.679871	−2.346400	1.511565	−1.850917	+4
−0.617326	−0.913107	−0.517867	0.812155	+5
0.111954	−0.523082	0.098607	−0.340368	+6

6.729847 bits = mean (R_{sequence}). 11.801021 bits = R_i of consensus sequence. −46.628249 bits = R_i of anticonsensus sequence. −25.962089 bits = average R_i for random sequence. 56286 = number of sequences.

We validated these donor and acceptor information models with the original model which effectively comprised a ‘normalized’ set of representative splice sites [34]. The Euclidean distances between these matrices are 31.4 bits for donors and 34.2 bits for acceptors, with the differences concentrated at the positions with the highest information contents. Evaluation of the present information model were evaluated with the original weight matrices [34] showed only a single donor site (0.002%) and 45 acceptor sites (0.08%) matrix with negative R_i values. Similarly, only 14 of 1744 acceptors and four of 1799 donor sites used to build the original models had negative R_i values when analysed with the present $R_i(b,l)$ matrices. Careful inspection of these sequences indicated that 45/46 of the sites with negative R_i values were the result of misalignments between mRNA and genomic sequences in the draft human genome sequence.

Online supplemental material

Estimate of the error associated with the individual information statistic R_i.

To calculate the error associated with the individual information statistic, R_i, we assume that the distribution of observed nucleotide frequencies $f(b,l)$, $b = 1, \ldots, 4$ and $l = 1, \ldots, L$, follow a product-multinomial distribution. In particular, each vector of base proportions $(f_{1l}, f_{2l}, f_{3l}, f_{4l})^t$ at given location l is assumed to arise from a multinomial distribution with parameters $(\pi_{1l}, \pi_{2l}, \pi_{3l}, \pi_{4l})^t$ and η_l. The multinomial parameters are approximated by the observed proportions for distributions with > 5000 sites. The multinomial distributions are assumed to be independent across locations.

To estimate the error across the entire binding site, we determined the information statistic for a given base at a specific location and summed over all positions in the site. The sample information statistic [1] for a given base b at location l is:

$$R_{i\,bl} = 2 - (-\log_2 f_{bl}).$$

This statistic is asymptotically normal [2] with mean:

$$E(R_{i\,bl}) \approx 2 + \log_2\pi_{bl} - [(1 - \pi_{bl})/(2\eta_l\pi_{bl})]$$

and variance:

$$V(R_{i\,bl}) \approx (1 - \pi_{bl})/(\eta_l\pi_{bl})$$

The asymptotic normality holds in this instance because the sample size is large ($>$ 5000 sites [3]). The estimated variance is obtained by substituting the f (frequency) for the multinomial parameter π.

Given a sequence of bases b_l across all locations [$\sum$ from $l = 1, 2, \ldots L$], the sample information statistic is:

$$R_{i\,b+} = \sum R_{i\,bl} = 2L + \sum\log_2 f_{bl}$$

This statistic is also asymptotically normal with mean:

$$E(R_{i\,b+}) = 2L + \sum\{\log_2\pi_{bl} - [(1 - \pi_{bl})/(2\eta_{bl}\pi_{bl})]\}$$

and variance:

$$V(R_{i\,b+}) = \sum(1 - \pi_{bl})/(\eta_l\pi_{bl})$$

Similarly, the estimated variance is obtained by substituting the sample of π with f. For nucleotides that are not represented at a particular position in the matrix (e.g. zero frequency), Staden's substitution was used, where $f(b,l) = 0$ is replaced with $f(b,l) = 1/(n + t)$, where $n =$ number of sequences used to create the weight matrix and $t = 2$. Because the population of sites included in the database considerably exceeds 5000, the assumption of normality is valid, and the original multinomial parameters approximate the observed proportions (frequencies).

References

1 Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *J Mol Biol* 1986; **188**:415–431.

2 Basharin GP. On a statistical estimate for the entropy of a sequence of independent random variables. *Prob Appl* 1959; **4**:333–336.

3 Miller GA. Note on the bias of information estimates. In: H. Quastler (editor). Information Theory in Psychology. Glencoe Illinois: Free press; 1955. pp. 95–100.

0960-314X © 2003 Lippincott Williams & Wilkins