Using Information Content and Base Frequencies to Distinguish Mutations from Genetic Polymorphisms in Splice Junction Recognition Sites

Peter K. Rogan *and Thomas D. Schneider†‡§

version = 1.38 of colonsplice.tex 2000 Nov 8
Research Article, *Human Mutation* vol. 6, No. 1, pages 74-76, 1995
short title: Identifying Splice Site Polymorphisms

Predicting the effects of nucleotide substitutions in human splice sites has been based on analysis of consensus sequences. We used a graphic representation of sequence conservation and base frequency, the sequence logo, to demonstrate that a change in a splice acceptor of *hMSH2* (a gene associated with familial nonpolyposis colon cancer) probably does not reduce splicing efficiency. This confirms a population genetic study that suggested that this substitution is a genetic polymorphism. The information-theory based sequence logo is quantitative and more sensitive than the corresponding splice acceptor consensus sequence for detection of true mutations. Information analysis may potentially be used to distinguish polymorphisms from mutations in other types of transcriptional, translational or protein coding motifs.

KEY WORDS: Information theory, Human splice sites, DNA sequencing, mutation, polymorphism

Nucleotide substitutions in a human splice donor or acceptor recognition sequence often disrupt processing of the normal transcript. Although altered mRNA splicing must ultimately

*Department of Pediatrics, Division of Genetics, Milton S. Hershey Medical Center, P. O. Box 850, Pennsylvania State University, Hershey, Pennsylvania 17033. (717) 531-8414, fax: (717) 531-8985, email: rogan@ncifcrf.gov

†To whom reprint requests/correspondence should be addressed.

‡Laboratory of Mathematical Biology, National Cancer Institute, Frederick Cancer Research and Development Center, P. O. Box B, Building 469, room 144, Frederick, Maryland 21702-1201. (301) 846-5581 (-5532 for messages), fax: (301) 846-5598, email: toms@ncifcrf.gov

be confirmed experimentally, tools capable of predicting the effects of these substitutions may be useful in recognizing those which are most likely deleterious. Mutations in human splice sites have been conventionally identified by comparing the sequence of the putative mutation with the consensus sequence [Mount, 1982; Nakai and Sakamoto, 1994]. This type of comparison can produce misleading results since the consensus sequence contains only the most representative nucleotides at each position. In some instances, it may not be possible to distinguish between deleterious mutations and silent genetic polymorphisms. This paper describes a more robust alternative.

Information analysis of normal splice junctions reveals partially conserved nucleotide sequences that are not always reflected in the corresponding consensus sequence [Stephens and Schneider, 1992]. Information content may be represented by a sequence logo (Fig. 1), which ⇐Fig 1 depicts the relative contribution of each position of the splice site and the relative frequencies of each nucleotide at every position [Schneider and Stephens, 1990]. The logo illustrates the full range of normal variants in the splice junction. To determine whether a nucleotide substitution in a splice site represents a polymorphism or a mutation, the individual information content of the site is compared with the overall distribution of individual information in a set of ~1800 human splice sites.

As an example of this method, we have analyzed the T→C transition found at position -5 of the intervening sequence of the *hMSH2* gene from multiple, independent sporadic colon carcinomas and patients with Lynch syndrome [Fishel et al., 1993]. Other mutations in the coding domain of this gene cause hereditary nonpolyposis colon cancer by disrupting the repair of somatic lesions that accumulate in genomic DNA [Leach et al., 1993]. Although the substitution at -5 was proposed to cause aberrant splicing of *hMSH2* mRNA [Fishel et al., 1993], our analysis suggests that it is probably not deleterious to maturation of the *hMSH2* message. First, upon inspection of the sequence logo, there is a nearly equal probability of observing C or T at position -5 in this set of splice acceptor sequences (Fig. 1; this corresponds to position -6 in [Fishel et al., 1993]). Second, cytosine at this position does not impede the normal splicing of 691 of 1712 acceptor sites derived from numerous human genes [Stephens and Schneider, 1992]. Third, we find that the common allele contains 6.5 bits of information, and the substitution weakens it to 6.3 bits. The average of the distribution of sites is 9.3 bits, and the distribution has a standard deviation of 4.6 bits. Nonfunctional sites are predicted to be below zero on this scale [Schneider, 1994]. Indeed, 2 of 20 unrelated normal individuals displayed this variant, consistent with the suggestion that this change represents a polymorphism [Leach et al., 1993].

This change is unlikely to affect the recognition of other nucleotides in the same acceptor site, as mutational analysis of the polypyrimidine tract in which it resides suggests that these nucleotides are independently recognized by the spliceosome [Stephens and Schneider, 1992; Roscigno et al., 1993]. We have found 196 normal human sites with the same or lower information content as the *hMSH2* acceptor containing this substitution. 51 of these contain cytosine at position -5. Either the true mutation lies elsewhere, in this or another gene [Leach et al., 1993; Bronner et al., 1994; Papadopoulos et al., 1994], or the change indicates that this base is involved in a genetic control mechanism other than mRNA splicing [Amrein et al., 1994].

To summarize, inference of genetic mutations in splice junction recognition sites based on consensus sequences may be inaccurate, whereas information analysis of sequence variants can distinguish between polymorphic nucleotides and mutant sites. True mutations are expected to reside in positions in which the sequence conservation in bits significantly exceeds the background variation [Stephens and Schneider, 1992] and where the base frequency decreases significantly. However, the identification of a mutation by information analysis does not always imply that the substitution will have a phenotype. For example, incomplete penetrance may affect the reliability of molecular diagnosis based on information analysis.

A similar approach could be applied to the analysis of other conserved transcriptional and translational signals or protein motifs in human sequences.[1]

ACKNOWLEDGMENTS

We thank Denise Rubens and Paul N. Hengen for comments on the manuscript. P.K.R. is supported by the Leukemia Research Foundation, the American Cancer Society, the March of Dimes Birth Defects Foundation, the Four Diamonds Pediatric Cancer Research Foundation, the David S. and Amy S. Goldberg Memorial Fund for Pediatric Research, and PHS R29-HD 29098-01.

[1]The computer programs for creating sequence logos and the data for human splice junctions are available across the internet by anonymous file transfer protocol (ftp) to ftp.ncifcrf.gov in the directory pub/delila. See the README and splice.info.Z files. The files are also available on the World Wide Web at http://www-lmmb.ncifcrf.gov/~toms/, email: toms@ncifcrf.gov.

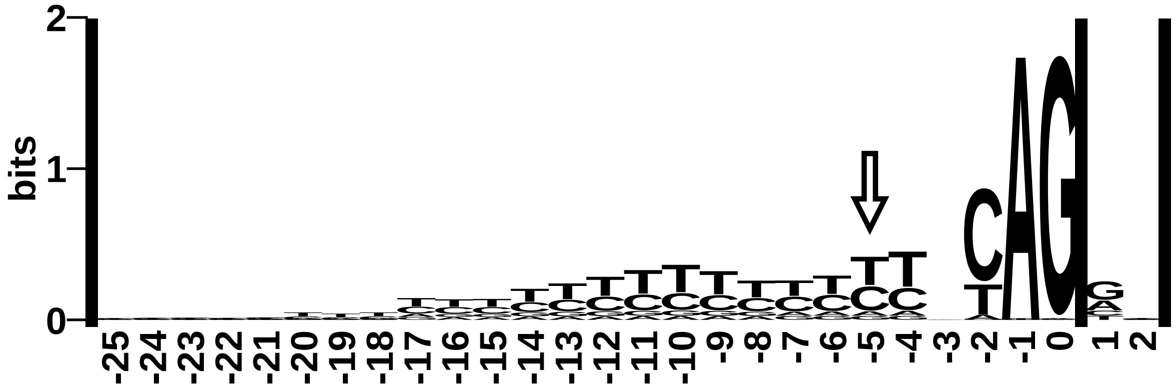

Figure 1: Location of the *hMSH2* polymorphism in the sequence logo of the human splice acceptor site.
This sequence logo was created from 1744 wild-type acceptor sites. The height of each nucleotide is proportional to its frequency at that position, while the height of each entire stack of nucleotides corresponds to the information measure (in bits) or, equivalently, the sequence conservation at that position. When sequence conservation is measured in bits, the relative heights of the stacks can be compared to one another and the total sequence conservation in a region can be found by adding the heights of the stacks together [Shannon and Weaver, 1949; Sloane and Wyner, 1993; Pierce, 1980]. Coordinates in the splice site are defined along the abscissa. RNA strand cleavage during splicing occurs at the vertical line between positions 0 and 1. All positions except -3 in this logo are significantly above background ($p < 8 \times 10^{-8}$). The arrow shows the position of the T$\rightarrow$C substitution in the *hMSH2* gene.

REFERENCES

Amrein H, Hedley ML, Maniatis T (1994): The role of specific protein-RNA and protein-protein interactions in positive and negative control of pre-mRNA splicing by *Transformer 2*. Cell 76:735–746.

Bronner CE, Baker SM, Morrison PT, Warren G, Smith LG, Lescoe MK, Kane M, Earabino C, Lipford J, Lindblom A, Tannergòrd P, Bollag RJ, Godwin AR, Ward DC, Nordenskjold M, Fishel R, Kolodner R, Liskay RM (1994): Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. Nature 362:258–261.

Fishel R, Lescoe MK, Rao MRS, Copeland NG, Jenkins NA, Garber J, Kane M, Kolodner R (1993): The human mutator gene homolog *MSH2* and its association with hereditary non-polyposis colon cancer. Cell 75:1027–1038.

Leach FS, Nicolaides NC, Papadopoulos N, Liu B, Jen J, Parsons R, Peltomäki P, Sistonen P, Aaltonen LA, Nyström-Lahti M, Guan XY, Zhang J, Meltzer PS, Yu JW, Kao FT, Chen DJ, Cerosaletti KM, Fournier REK, Todd S, Lewis T, Leach RJ, Naylor SL, Weissenbach J, Mecklin JP, Järvinen H, Petersen GM, Hamilton SR, Green J, Jass J, Watson P, Lynch HT, Trent JM, de la Chapelle A, Kinzler KW, Vogelstein B (1993): Mutations of a *mutS* homolog in hereditary nonpolyposis colorectal cancer. Cell 75:1215–1225.

Mount SM (1982): A catalogue of splice junction sequences. Nucleic Acids Res. 10:459–472.

Nakai K, Sakamoto H (1994): Construction of a novel database containing aberrant splicing mutations of mammalian genes. Gene 141:171–177.

Papadopoulos N, Nicolaides NC, Wei YF, Ruben SM, Carter KC, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM, Adams MD, Venter JC, Hamilton SR, Petersen GM, Watson P, Lynch HT, Peltomäki P, Mecklin JP, de la Chapelle A, Kinzler KW, Vogelstein B (1994): Mutation of a *mutL* homolog in hereditary colon cancer. Science 263:1625–1629.

Pierce JR (1980):. An Introduction to Information Theory: Symbols, Signals and Noise. second edition, Dover Publications, Inc., New York.

Roscigno RF, Weiner M, Garcia-Blanco MA (1993): A mutational analysis of the polypyrimidine tract of introns: effects of sequence differences in pyrimidine tracts on splicing. J. Biol. Chem. 268:11222–11229.

Schneider TD, Stephens RM (1990): Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18:6097–6100. http://www.lecb.ncifcrf.gov/~toms/paper/logopaper/.

Schneider TD (1994): Sequence logos, machine/channel capacity, Maxwell's demon, and molecular computers: a review of the theory of molecular machines. Nanotechnology 5:1–18. http://www.lecb.ncifcrf.gov/~toms/paper/nano2/.

Shannon CE, Weaver W (1949):. The Mathematical Theory of Communication. University of Illinois Press, Urbana.

Sloane NJA, Wyner AD (1993):. Claude Elwood Shannon: Collected Papers. IEEE Press, Piscataway, NJ.

Stephens RM, Schneider TD (1992): Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. J. Mol. Biol. 228:1124–1136. http://www.lecb.ncifcrf.gov/~toms/paper/splice/.