Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites*

R. Michael Stephens†‡ and Thomas Dana Schneider§¶

version = 3.51 of splice.tex 1996 Feb 20

Key words: splice, spliceosome, information theory, evolution, human.

Manuscript: HIP 86/91

Subject Category: *Proteins, Nucleic acids and other biologically important macromolecules: in vivo modification and processing.*

Running title: Spliceosome Evolution

*This paper was published in (Stephens & Schneider, 1992) and is available on the world wide web at
http://www-lmmb.ncifcrf.gov/~toms/paper/splice
or
ftp://ftp.ncifcrf.gov/pub/delila/splice.ps

†National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Mathematical Biology, P. O. Box B, Frederick, MD 21702-1201 and Linganore High School, 12013 Old Annapolis Rd., Frederick, MD 21701.

‡Current address: Massachusetts Institute of Technology, EC Box R, 3 Ames St., Cambridge, MA 02139, Internet address: stephens@athena.mit.edu.

§National Cancer Institute, Frederick Cancer Research and Development Center, Laboratory of Mathematical Biology, P. O. Box B, Frederick, MD 21702-1201. Internet address: toms@ncifcrf.gov.

¶corresponding author

An information analysis of the $5'$ (donor) and $3'$ (acceptor) sequences spanning the ends of nearly 1800 human introns has provided evidence for structural features of splice sites that bear upon spliceosome evolution and function:

(1) 82% of the sequence information (*i.e.* sequence conservation) at donor junctions and 97% of the sequence information at acceptor junctions is confined to the introns, allowing codon choices throughout exons to be largely unrestricted. The distribution of information at intron-exon junctions is also described in detail and compared with footprints.

(2) Acceptor sites are found to possess enough information to be located in the transcribed portion of the human genome, whereas donor sites possess about one bit less than the information needed to locate them independently. This difference suggests that acceptor sites are located first in humans and, having been located, reduce by a factor of two the number of alternative sites available as donors. Direct experimental evidence exists to support this conclusion.

(3) The sequences of donor and acceptor splice sites exhibit a striking similarity. This suggests that the two junctions derive from a common ancestor and that during evolution the information of both sites shifted onto the intron. If so, the protein and RNA components that are found in contemporary spliceosomes, and which are responsible for recognizing donor and acceptor sequences, should also be related. This conclusion is supported by the common structures found in different parts of the spliceosome.

1. Introduction

In eukaryotic cells, nuclear RNA is usually spliced prior to translation (for review see (Green, 1986; Sharp, 1987; Green, 1991)). Removing introns is the function of the spliceosome, which is made up of small nuclear ribonucleoprotein particles (snRNPs) (Brody & Abelson, 1985; Frendewey & Keller, 1985; Grabowski *et al.*, 1985; Reed *et al.*, 1988; Steitz *et al.*, 1988). Because reliable splicing is necessary for cell survival, there must be a precise way for the spliceosome to identify RNA splice sites (Aebi & Weissmann, 1987).

These patterns are defined by nucleotides at the ends of the introns and are probably not affected by folded RNA structures, since introns can have large interior deletions without affecting the splicing mechanism (Breathnach & Chambon, 1981). A model of splice site identification utilizing just the GT and AG dinucleotides is not acceptable because bases other than these dinucleotides affect splicing (Aebi & Weissmann, 1987; Aebi *et al.*, 1987). Even consensus sequences are not sufficient to characterize splice patterns (Breathnach & Chambon, 1981; Green, 1986; Padgett *et al.*, 1986; Aebi & Weissmann, 1987; Aebi *et al.*, 1987; Mount, 1982; Nakata *et al.*, 1985), because this qualitative (and arbitrary) simplification of the base frequencies destroys subtle details.

To analyze the sequence conservation at splice sites, we chose to use Shannon's information measure (Shannon, 1948; Shannon & Weaver, 1949; Pierce, 1980) as applied previously (Schneider *et al.*, 1986; Schneider, 1988; Schneider & Stormo, 1989; Eiglmeier *et al.*, 1989; Fields, 1990; Penotti, 1990; Penotti, 1991; Herman & Schneider, 1992). (See also (Berg & von Hippel, 1987; Berg & von Hippel, 1988*a*; Berg, 1988; Berg & von Hippel, 1988*b*; Stormo, 1990).) The information measure has three advantages over other quantitative measures of binding site patterns (Stormo *et al.*, 1982*b*; Beacham *et al.*, 1984; Goodrich *et al.*, 1990; Mars & Beaud, 1987). First, since it is the *only* possible consistent additive measure of sequence conservation (Shannon, 1948; Shannon & Weaver, 1949) it allows one to make quantitative statements about the relative importance of various portions of a binding site. Second, unlike significance measures such as the standard deviation or χ^2, the information converges to a steady value as the number of sequences increases. This allows one to compare different binding sites (*e.g.* donors to acceptors) and to resolve finer details by increasing the number of sequences. Finally, by adding up the information from many individual positions, the total information of the pattern at the binding sites can be determined. This total may then be compared to the amount of information required to locate the sites in the transcribed or physically available portion of the genome. If the two numbers differ significantly, we are alerted to the existence of previously unnoticed phenomena (Schneider *et al.*, 1986; Schneider & Stormo, 1989).

It is not our intention to create a method for locating splice sites. This worthy and difficult task has been attempted by many others (Stormo *et al.*, 1982*a*; Quinqueton & Moreau, 1985; Nakata *et al.*, 1985; Hopfield & Tank, 1986; Ohshima & Gotoh, 1987; Fichant & Gautier, 1987; Kudo *et al.*, 1987; Gelfand, 1990; Brunak *et al.*, 1991). Identifying sites by using a black box search system such as a neural network (Hopfield & Tank, 1986; Stormo *et al.*, 1982*a*; Nakata *et al.*, 1985; O'Neill, 1991) does not tell us how the spliceosome finds sites, nor does it teach us how the sites are constructed. Here, we are mainly concerned with these two problems.

Many groups use coding sequences and sequence composition of long stretches to help locate coding regions and splice junctions (Fichant & Gautier, 1987; Gelfand, 1990; Brunak *et al.*, 1991). These techniques use information which is probably not available to the spliceosome as it searches, and so we call them "impure". A "pure" technique would be one which attempts to match the natural mechanism. It seems unlikely that the spliceosome is capable of looking for long open reading frames or responding to the general composition of the pre-mRNA. Such a mechanism has no precedent for any DNA or RNA binding protein we are aware of. Furthermore, there is no precedent for memory in nucleic-acid recognizers, so these molecules almost certainly cannot keep track of the reading frame composition beyond the small region they are in physical contact with. Those regions can be determined by footprint experiments and sequence logos (Schneider & Stephens, 1990). Search techniques which use open reading frames or other global clues are therefore almost certainly using information which is not available to the spliceosome. The models generated by these methods, useful as they may be for site identification, do not increase our understanding of the spliceosome recognition mechanism, and in the long run they must be rejected.

We chose to use human sequences because they form a large, reliable data set (Watson, 1990; Cantor, 1990). By using human sequences exclusively we avoided the possibility of subtle species differences (Csank *et al.*, 1990). We did assume, however, that the mechanism of splicing is the same in all tissues.

2. A Measure of Sequence Pattern Conservation

Shannon's measure of information indicates how much choice is involved in a particular selection from among two or more alternatives. Our measurements are given in bits, where one bit is the amount of information necessary to select one of two possible states (*e.g.* a yes-no question). In general, to select one possibility from M possibilities requires $\log_2 M$ bits. To specify one of the four nucleotide bases requires two yes-no choices and so can be done with two bits.

Shannon's theory further states that the information is related to the *decrease* in the number of possibilities, not the original number of possibilities (Shannon, 1948; Shannon & Weaver, 1949; Pierce, 1980; Schneider *et al.*, 1986; Brillouin, 1951; Tribus & McIrvine, 1971; Rothstein, 1951). Thus, if there are four possibilities and we only answer one yes-no question, then two possibilities remain, so the information gained is $\log_2 4 - \log_2 2 = 1$ bit. Although this distinction may appear trivial, erroneous applications of information theory continue to appear in the literature. For example, Brunak's information curves are upside down (Brunak *et al.*, 1991; Lukashin *et al.*, 1992), showing more information *outside* the binding sites than inside them. Our measure shows the site itself carrying the information needed for binding.

In order to correctly apply information theory, we must construct a consistent physical model of how the spliceosome finds the intron ends. A molecule which recognizes patterns on nucleic-acids is called a "recognizer". Recognizers have two distinct thermodynamic states. In their *before* state, they are non-specifically bound to the nucleic-acid, presumably by electrostatic attraction (Weber & Steitz, 1984), while in their *after* state they are specifically bound to their binding sites. (We are not concerned here with how the recognizer locates the nucleic-acid.) *Before* binding, any of the four bases can be presented to the surface of the recognizer. Shannon's terminology to describe this situation is to say that there are 2 bits of "uncertainty" as to which base is present. (N.B. Because it is inconsistent with the *before-after* model, we no longer use the genomic uncertainty here (Schneider *et al.*, 1986).) In contrast to the high uncertainty in the *before* state, the uncertainty is lower in the *after* state. Some positions, having only one base, will have no uncertainty remaining. Other positions, having two possible bases, have 1 bit of uncertainty remaining. Positions outside the binding site still have 2 bits of uncertainty remaining. The information content is the difference in uncertainty between the initial, unbound *before* state and the final, specifically bound *after* state (Schneider, 1991*a*; Schneider, 1991*b*). It can be measured from a set of aligned sequences.

When the frequencies of bases in the binding site are not 0%, 50%, or 100%, a more sophisticated method must be used to calculate the uncertainty. The uncertainty at a position can be found from the probabilities of the bases at that position by using the following equation:

$$H(l) = -\sum_{b \in \{a,c,g,t\}} f(b,l) \log_2 f(b,l) \qquad \text{(bits per position)}, \tag{1}$$

where $H(l)$ is the uncertainty at position l and $f(b,l)$ is the probability of base b at position l. In this study, the observed base frequencies were used for $f(b,l)$ since the true base probabilities of the population are unknown. With this substitution, $H(l)$ is biased when there are few sequences, so a small sample-size correction was added (Schneider *et al.*, 1986). The correction was only $e(n) = 0.0012$ bits because $n \approx 1800$ sites.

The uncertainty is one of four bases (2 bits) before recognizer binding and measured as $H(l) + e(n)$ after binding, so the information at each position in a set of sequences is the difference:

$$R_{sequence}(l) = 2 - \Big(H(l) + e(n)\Big) \qquad \text{(bits per position)}. \tag{2}$$

This measure of sequence pattern shows the relative importance (conservation) at each position. Furthermore, the total conservation of pattern at a binding site can be found by summing up the conservation at each individual position:

$$R_{sequence} = \sum_{l} R_{sequence}(l) \qquad \text{(bits per site)}. \tag{3}$$

(This assumes positions are independent, support for which is given in Methods and (Schneider, 1991*a*).)

The same *before-after* model is used to determine $R_{frequency}$, the amount of information needed to locate the binding sites. In the *before* state, the uncertainty in physical position of a recognizer is $H_{before} = \log_2 G$, where G is the number of places on the genetic material to which the recognizer can be bound non-specifically. In the *after* state, the uncertainty in physical position is lower, depending on the number of binding sites (γ): $H_{after} = \log_2 \gamma$. The decrease in uncertainty is:

$$R_{frequency} = H_{before} - H_{after} = -\log_2(\gamma/G), \tag{4}$$

where γ/G is the frequency of binding sites.

When $R_{sequence}$ is compared to $R_{frequency}$ for several procaryotic genetic control systems, the two numbers are reasonably close to each other (Schneider *et al.*, 1986; Schneider, 1988) meaning that the observed information at a binding site ($R_{sequence}$) is approximately equal to the information needed to locate the sites in the genome ($R_{frequency}$). Since $R_{frequency}$ is determined from the size of the genome and number of sites, it is a function of physiology and history. In contrast, $R_{sequence}$ is determined by patterns in the genome, and, in principle could be any value. To account for the convergence of the two values, we imagine that $R_{sequence}$ evolves toward $R_{frequency}$ (Schneider, 1988). We take the idea that $R_{sequence}$ should be close to $R_{frequency}$ as a "working hypothesis".

Confirmation of a theory by comparison of predicted results to observed data is necessary to establish that the theory is reasonable, but new knowledge comes from careful investigation of exceptions. Thus when $R_{sequence}$ is similar to $R_{frequency}$, the working hypothesis is confirmed and, as far as this analysis is concerned, there is not much more to do except perhaps gather more data for higher resolution studies. However, if $R_{sequence}$ and $R_{frequency}$ differ significantly, then the presence of undiscovered biological phenomena is suggested (Schneider *et al.*, 1986; Herman & Schneider, 1992). For example, an enormous discrepancy was found in the case of bacteriophage T7 promoters, where $R_{sequence}$ is 35.4 bits per site but $R_{frequency}$ is only 16.5 bits per site, giving a ratio of $R_{sequence}$ to $R_{frequency}$ of 2.1. One explanation for this result is that two proteins bind at these promoters and that they do not share the same information. If so, the polymerase alone should only use half of the conserved 35 bits. This hypothesis led to experimental work which showed that T7 polymerase does indeed only use 18 ± 2 bits (Schneider & Stormo, 1989). The putative second factor has not yet been identified. In this paper we describe a significant discrepancy between $R_{sequence}$ and $R_{frequency}$ for human donor splice sites.

3. Materials and Methods

(a) Collecting Splice Sites

Sequences were extracted from GenBank 62 (December 1989) using the IDEAS package (Kanehisa, 1988). Duplicate sites were removed from the database and the remaining sites were analyzed by using programs of the Delila system (Schneider *et al.*, 1982; Schneider *et al.*, 1984; Schneider & Stephens, 1990). All programs and data described here are available by anonymous ftp from ncifcrf.gov in directory pub/delila (or by sending electronic mail to toms@ncifcrf.gov). Source code is in Pascal, but many programs have been automatically translated into C. A VAX 11/785 was used to obtain GenBank files. Programs were run on Sun 3/50 and Sun 4/260 workstations, except for Index, Indana, and Cluster which were run on a Cray X-MP for the acceptor splice site analysis.

Dbfilter 1.06 was used to remove entries containing non-*Homo sapiens* and artificial genes. A Delila library (Schneider *et al.*, 1982; Schneider *et al.*, 1984) was then created using Dbbk version 3.11, with corresponding catalog generated by Catal 9.20. The Delila instructions used were computer generated by Dbinst 2.33. These instructions were used with Delila 1.77 to create a Delila book consisting of the sequence regions immediately around splice junctions. (A "book" is a computer database containing a set of sequence fragments.)

Unfortunately, GenBank is full of redundant sequences. To identify duplicate splice junctions, all 50bp long oligomers in the initial Delila book were alphabetized using the program Index 9.17. Indana 5.12 then generated a sublist of sequences sharing repeats of at least 15bp. This length oligomer was chosen since it would appear roughly once in $4^{15} \approx 1 \times 10^9$ bases, which is close to the size of the human genome. Cluster 5.01 reassembled the alphabetized oligomers so that the repeats could be easily identified and removed from the Delila instruction set.

We found many nearly identical sequences from the major histocompatibility complexes (MHCs) and the immunoglobulin complexes (Igs). Since the immunoglobulin supergene family complex represents 0.1% to 0.5% of the human genome (Watson *et al.*, 1987), and there are about 1800 splice junctions in our database, a representative sample would have only about 2 to 9 immunoglobulin genes. Likewise, the MHC complex consists of 2 to 4 thousand kilobases of DNA (Watson *et al.*, 1987), so 2 to 4 MHC genes should be included in the final set. There will be a bias in the final data set no matter what is done. Of several options, we chose to remove all Igs and MHCs in the database except for one sequence from each gene library or class. This produces the most variation and so may tend to reduce the amount of information measured in the sites. Because identical sequences have been removed, however, this method has the great advantage that it should avoid producing spurious patterns outside the binding region. More importantly, there are genes that are not represented *anyway*, since not all genes and supergene families have been sequenced. Because there are many sequences not represented in the calculations, we introduce no further significant error by dropping genes from these two families.

The Ig, MHC and other duplicates identified by Cluster were manually removed from the Delila instructions generated by Dbinst. This revised instruction list was sent back through Delila, and the resulting book was once again analyzed with Index, Indana and Cluster.

GenBank contains errors in addition to the duplicate sequences (Brunak *et al.*, 1990). We located some of these as "exceptional" sequences in highly conserved positions, and although these were eliminated, finding other mistakes such as these would require checking every GenBank site against its corresponding paper or an analysis by neural net (Brunak *et al.*, 1990). Because of this, there may still be mistakes in the final database.

The book generated from the cleaned Delila instructions was checked for proper alignment with Alist 4.63 and was then run through Encode 1.26, which converted the sequences into strings of numbers (Schneider *et al.*, 1984). This was in turn run through Rseq 5.31, which took the encoded data and calculated $R_{sequence}$ (Schneider *et al.*, 1986). Rseq also produced a data matrix for Rsgra 4.99, which in turn generated the information curve in the device independent graphics language PostScript® (Adobe Systems Incorporated, 1985*a*; Adobe Systems Incorporated, 1985*b*). Dalvec 2.14 and MakeLogo 7.53 produced the sequence logos (Schneider & Stephens, 1990), also in PostScript®.

(b) Statistical Tests

The information measure given in equation (3) is made on the assumption that the positions of the binding site are independent of one another. To test this assumption, we used the method devised by Olsen in which the χ^2 statistical test is used to detect correlations between positions around a binding site (Olsen, 1983). To express these correlations as information measures, we first calculated the uncertainty

$$H(x,y) = \sum_{b_x} \sum_{b_y} f(b_x, b_y) \log_2 f(b_x, b_y) \tag{5}$$

for every pair of positions (x, y) using the 16 dinucleotide frequencies $f(b_x, b_y)$, where $b_x, b_y \in \{a, c, g, t\}$. The information needed to define those frequencies is

$$R(x, y) = 4 - H(x, y). \tag{6}$$

We then subtracted from this the information at each position, $R(x)$ and $R(y)$, to calculate information in the correlation,

$$R_{cor} = R(x, y) - R(x) - R(y). \tag{7}$$

We analyzed every pair of positions from -50 to +50 around both splice junctions using a program called Diana (version 1.66). Aside from a general correlation between every base and its immediate neighbor, we only found a weak correlation between the ends of the donor junction (positions -1 to +5, -2 to +4 and -2 to +5). These "mutual informations" were about 0.07, 0.05 and 0.04 bits respectively for the three correlations. Other positions show correlations around 0.01 to 0.03 bits, and the sampling correction is 0.01 bits. Although the three correlations appear to be significant, the relationship is not obvious from inspection of the frequency matrixes. The result suggests, however, that the ends of the donor splice junction pattern are functionally related to each other. What this may mean mechanistically is unknown. We did not search for higher order relationships as this would require at least 4 times as much data to retain the same relative error.

To determine the possible variation of $R_{sequence}$, we used the Rsim 1.92 program (Appendix 1) to create a set of frequency matrixes ($f(b, l)_{true}$) that could have produced the observed $R_{sequence}$ value. These matrixes were then used to determine the possible range of $R_{sequence}$ values. This method gives a better upper bound on the standard deviation of $R_{sequence}$ than the earlier method (Schneider *et al.*, 1986).

4. Results

(a) The Sequence Patterns at Splice Junctions

The information curves and sequence logos (Schneider & Stephens, 1990) in Fig. 1 show the location of information at both splice sites, with the amount of information at a position being represented by the height of the information curve or logo at that position. The tallest information spike in each case is within a base of the intron/exon junction, giving evidence ⇐Fig 1 that this pattern is indeed associated with spliceosome recognition. (It is not proof of association since patterns found in the region of a binding site are sometimes unrelated to the known function (Schneider & Stormo, 1989).) Subtle features of the splice sites, such as the gentle sloping of the pyrimidine (C/T) stretch at the acceptor site, can be seen in the logos.

Fig. 1 also shows that the location of the pattern as indicated by the donor information curve does not precisely match those bases protected by the spliceosome in a T_1 fingerprint experiment by Mount *et. al.* (Mount *et al.*, 1983) (in positions +7 to +12), nor does it match the RNAase-A data in (Krämer, 1987) (in positions −17 to −4 and +7 to +11). We must point out, however, that there is a difference between a base being *protected* and a base being *specifically bound*. A base can be protected if it is located in a groove on the recognizer, but this does not necessarily imply that specific binding also occurs. Conversely, a base can be externally bound to a recognizer and be exposed. Protection is not necessarily a measure of exactly where the contacts are, whereas the information curve can give a much closer approximation of the actual binding pattern locations. Still, hydroxyl radical protection has been shown from −2 or −3 through +7 or +8, which matches the conserved region closely (R. A. Padgett, personal communication). The $3'$ edge of *E. coli* ribosome binding sites is at base +13 according to the information curve (Schneider *et al.*, 1986), which is remarkably close to the point (+15) where reverse transcriptase is blocked by a bound ribosome (Hartz *et al.*, 1988). This kind of "toeprint" experiment would also be interesting to try on spliceosomes.

We measured the background noise around the donor splice site at 0.012 ± 0.007 bits between positions -50 and +50 excluding the site locations from -3 to +6. From this approximately normal distribution, we calculated the statistical significance of position +6, which possesses only ~ 0.04 bits of information, and found that it is over 5 standard deviations from the mean of the background noise. Therefore, position +6 is highly significant ($p < 4.0 \times 10^{-5}$), even though it represents only $\sim$0.03 bits of information above the background noise. Likewise, position -3 is even more significant (18.75 standard deviations, $p < 1 \times 10^{-12}$). From the resolution of the current graph, we conclude that the overall donor splice site runs at least from position -3 to position +6.

Background noise around the acceptor site was also measured from −50 to +50 excluding the site positions -25 through +2 and was found to be

0.008 ± 0.0055 bits. Relative to the background, positions -25 and +2 are also highly significant ($p < 1.5 \times 10^{-8}$ and $p < 8.2 \times 10^{-8}$, respectively). We therefore used these positions in our definition of the acceptor splice site. We note that the ~ 0.04 bit position -25 coincides with the end of the hydroxyl radical protection data of Wang and Padgett (Wang & Padgett, 1989) (Fig. 1), although as discussed above, protection and binding are not necessarily correlated. Indeed, the frequencies of bases upstream of -25 suggest a continuation of the pyrimidine tract, but this is within the background noise of the information curve, and a larger data set would be required to confirm it.

To avoid misinterpreting these results, we must distinguish between the height of a logo at a particular position (information content) and its statistical significance. The two are *not* correlated since a "highly conserved" position would be insignificant if it had been derived from only two random sequences. In contrast, positions in our data set with 0.04 bits are significantly differentiated from the background noise because they are derived from $\sim$1800 sequences.

We also note that the zero information point at position -3 corresponds to an unprotected base in the footprint data. This suggests that the nucleotide at position -3 is completely exposed in solution, having neither specific nor non-specific contacts. Perhaps it is a divider between two separate recognition regions, the pyrimidine tract and the AG spike. Indeed, several different snRNAs and proteins are involved in binding the acceptor end of the intron (Guthrie & Patterson, 1988; Zieve & Sauterer, 1990; Lührmann, 1988; García-Blanco *et al.*, 1989).

The area under the donor information curve from -3 to +6 is $R_{sequence} = 7.92 \pm 0.09$ bits/site, while $R_{sequence} = 9.35 \pm 0.12$ bits/site from -25 to +2 for the acceptor. The standard deviations were calculated as described in Appendix 1. These results are basically the same as those of Penotti (Penotti, 1991), who used less than half as many sequences, and did not report the standard deviations.

(b) Information Needed to Locate Splice Junctions

We estimated $R_{frequency}$ (the information needed to locate a set of sites) as 9.66 bits per site by adding the average length of introns (592 bp, $n = 1247$) and exons (220 bp, $n = 1872$) within the database and taking the logarithm of this value to be $R_{frequency}$, since this is the positional uncertainty that must be overcome by the spliceosome to locate an intron end. This $R_{frequency}$ value is valid for both donor *and* acceptor sites, because $R_{frequency}$ is based on the frequency of a particular site within the transcribed genome, and the frequencies of sequenced donor and acceptor intron ends are approximately equal. (We had 1799 donor sites and 1744 acceptor sites.) We also expect $R_{frequency}$ to equal $R_{sequence}$ for both sites (Schneider *et al.*, 1986; Schneider, 1988; Schneider & Stormo, 1989; Penotti, 1990). Although $R_{frequency}$ is biased to be small because longer introns are not consistently reported in GenBank, measures of $R_{sequence}$ are also biased to be small because of site positioning errors in GenBank. Despite these caveats, $R_{sequence} \approx R_{frequency}$ for the acceptor junction. In contrast, $R_{frequency}$ does *not* approximate $R_{sequence}$ for the donor junction. From the large size of our database, the observed difference of more than 1 bit in information content between the two sites and between the donor $R_{sequence}$ and $R_{frequency}$ is highly significant ($p < 10^{-12}$).

5. Discussion

(a) Mechanics of Splice Junction Location

Alternative splicing could account for a 1 bit difference between $R_{sequence}$ for donor and acceptor sites. However, this would require that there be twice as many donor sites as acceptor sites. In contrast, the data strongly suggest that these numbers are similar. In our database, which had no selection for or against either site so far as we know, the difference can account for only $\log_2(1799) - \log_2(1744) = 0.04$ bits. We assume that most investigators would have found alternative donor sites in almost every gene if they exist. Thus, alternative splicing is unlikely to explain the difference.

A surprisingly simple explanation begins with the observation that once the acceptor site has been located, it is no longer necessary to look at sequences on one side of the site. If only half of the RNA needs to be searched, one bit less information would be needed to find the donor sites, because one bit resolves the choice between two equally likely possibilities, and this choice is made once the acceptor has been bound. In other words, once an acceptor site has been found *the search problem has been cut in half.* This implies that the acceptor sites must be bound first. Indeed, this has been observed by others experimentally (Steitz *et al.*, 1988; Frendewey & Keller, 1985; Lamond *et al.*, 1987; Robberson *et al.*, 1990; Talerico & Berget, 1990). This explanation is still not sufficient to explain the difference of 1.43 bits between $R_{sequence}$ and $R_{frequency}$ for the donor sites. However, if the spliceosome did a linear scan for the donor site after binding the acceptor, the information requirement for the donor site might be reduced still further (in preparation). Scanning models for splicing have a somewhat confusing history in that some groups found evidence for scanning, whereas other groups found evidence against scanning (Lang & Spritz, 1983; Kühne *et al.*, 1983; Reed & Maniatis, 1986; Robberson *et al.*, 1990; Green, 1991). Since our data are consistent with scanning in either the $5' \rightarrow 3'$ direction or the $3' \rightarrow 5'$ direction starting from the acceptor site, they support the exon definition model. This model proposes that the spliceosome binds the acceptor first and then searches $5' \rightarrow 3'$ across the exon for the next donor. Once the two exons have been located, the intervening intron is removed (Robberson *et al.*, 1990; Talerico & Berget, 1990).

The U1 snRNP, which is known to be involved in donor binding, may also be required for binding of spliceosomes to the acceptor sites (Rosbash & Séraphin, 1991; Goguel *et al.*, 1991). This suggests that U1 is part of the acceptor recognition complex. These data do not exclude the possibility proposed above, that the acceptor site is bound first. In fact, we can imagine a system in which a single protein has two nucleic acid binding sites, one of which must be bound first. A simple model for this can be made by crossing two sticks and pushing the crossed point into a glob of clay. Once this mold has been formed, there is a required order for the sticks to be inserted, provided that the second stick is not threaded from one end.

(b) Morphology of Splice Junction Patterns

The information curves clearly demonstrate that most information resides on the intron side of each splice junction (Fig. 1). On the donor side, 82% of the information present (6.50 bits) is on the intron side of the splice point, while the acceptor side has 97% of the information present (9.05 bits) on the intron side. This bias was also observed by C. Fields for *C. elegans* splice sites (Fields, 1990). Since a certain amount of pattern must exist near each splice point for the points to be located precisely, and any pattern on the exon side of the splice junction constrains protein coding, it is sensible that most of the splice pattern resides on the intron, where it does not limit the choice of codons. We imagine that this situation has arisen through the co-evolution of the splice sites and the splice site recognition components in the spliceosome. The overall effect may have been that during evolution the splice patterns were both "pushed" from the exons onto the introns.

(c) Comparison of Donor and Acceptor Splice Patterns

We aligned the two logos in Fig. 1 by their respective splice points and found that the only region of overlap is from 3 bases in front of the splice points to 2 bases after the splice points. There is a correlation between the ends of this overlap and the dip at position +2 in the donor curve as well as with the zero point at position -3 on the acceptor curve. That is, on both graphs the region of overlap is defined by the edge of an information curve on one side and a dip in the information curve on the other side. In the region of overlap, the consensus sequences (created from the most frequent base at each position) are the same: CAG|GT. Related consensus sequences appear in many species (Csank *et al.*, 1990).

This result, although derived from positions that have small information contents, is highly significant. We cannot neglect the T at the $3'$ side of the acceptor site, although it is not even visible in Fig. 1 at position +2. For the 1737 bases at this point, the expected number of each base is $1737/4 = 434$, so the observed numbers of C's and T's at this position (308 and 590, see Fig. 1) contribute 37 and 56 respectively to a total χ^2 at this position of 94. With only 3 degrees of freedom, the significance is $p < 10^{-6}$. The significance of the C on the $5'$ side of the donor site is even greater, contributing 111 to a total χ^2 of 328 ($p \ll 10^{-6}$).

Mount noticed a four base similarity in the two splice patterns, which he interpreted as reflecting its functional characteristics (Mount, 1982). Instead, we suggest that both donor and acceptor sites have evolved from a single "proto-splice site" ancestor, with the information at each site having been shifted to maximize pattern on the intron side of both splice sites, as discussed above. Fig. 2 shows how this shift may have occurred. The bottom two ⇐Fig 2 sequence logos in Fig. 2 show the patterns at the modern donor and acceptor splice sites. According to this model, the proto-splice site "cag" (at the top of Fig. 2) lost emphasis to become a smaller "cag" on the donor side, while the "gt" became a more strongly conserved "GT". On the acceptor side, the proto-splice site "cag" gained the emphasis to a "CAG" while the "gt" became even smaller. Although both the donor and the acceptor sites have the *same consensus*, it is clear from Fig. 2 that the *emphasis* on individual bases within the modern regions is now different.

(d) Evolution of Splice Junctions and the Spliceosome

A common origin for donor and acceptor sites implies that ancient eukaryotes once possessed a single splice pattern located by a single type of splice site recognizer. Perhaps the pattern looked like the logo at the top of Fig. 2. In such a system, introns could not be distinguished from exons, and when splicing took place, the same pattern would be reconstructed (Hickey *et al.*, 1989). This could have led to uncontrolled "recursive splicing". Because of this difficulty, perhaps a third element eventually appeared, guaranteeing that only intervening sequences would be removed. One candidate for this third element might be the sequences around the branch point, which is between the donor and the acceptor[1] while another possibility is the polypyrimidine tract. Alternatively, the system could avoid having a third element if the ends of the intron were distinct, as they are today. One advantageous way to do this is to shift the patterns onto the intron, so that less pattern remains on the exon after splicing.

If both splice site patterns were shifted into the intron during the course of evolution, several things would have occurred simultaneously.

First, the donor and acceptor patterns would have become differentiated. This would distinguish the introns from the exons. If the branch point's original function was to identify the inside of the intron, then the pattern around the branch site could have decayed as the donor and acceptor sites became more distinct. The modern human branch point contains no more than 3 to 7 bits (Guthrie & Patterson, 1988; Penotti, 1991). In contrast, the branch point of yeast, which has a splicing mechanism similar to that of humans (Pikielny *et al.*, 1983), is probably higher (Green, 1986; Keller & Noon, 1984), suggesting that some organisms still use the branch point for distinguishing the intron from the exon.

[1] The branch point is a base within the intron RNA to which the first chemical bond is made from the donor site. In the second step of splicing, the donor becomes directly connected to the acceptor.

Secondly, the splice pattern differentiation would have been accompanied by divergent evolution of the splice site recognizers. The splicing reaction, originally needing only one kind of recognizer, came to require several different recognizers acting in tandem. This explains the multi-part structure of the modern spliceosome (Reed *et al.*, 1988; Guthrie & Patterson, 1988; Zieve & Sauterer, 1990). Because of this, the spliceosome elements should also have a common ancestor (Reddy & Busch, 1988), which we call the "proto-spliceosome". This implies that modern spliceosome components should have sequences in common.

Small nuclear ribonuclear particles (snRNPs) in the spliceosome do indeed share similar structures (Padgett *et al.*, 1986; Steitz *et al.*, 1988; Guthrie & Patterson, 1988; Zieve & Sauterer, 1990; Kastner *et al.*, 1990; Sillekens *et al.*, 1987) The U1 snRNP recognizes and binds to the donor site (Padgett *et al.*, 1986; Sharp, 1987; Mount *et al.*, 1983; Zieve & Sauterer, 1990), and possibly also the acceptor (Grabowski *et al.*, 1991; Rosbash & Séraphin, 1991; Goguel *et al.*, 1991). U2 appears to bind in the branch point / polypyrimidine region (Green, 1991). U4, U5, and U6 also participate in both parts of spliceosome formation (Guthrie & Patterson, 1988; Zieve & Sauterer, 1990; Lührmann, 1988; Newman & Norman, 1992). In mammals, these snRNPs have identical cores of six or seven proteins (Sharp, 1987; Zieve & Sauterer, 1990; Kastner *et al.*, 1990; Sillekens *et al.*, 1987; Lührmann, 1988; Guthrie & Patterson, 1988). Not only do the protein components have common structures, but there also exists a common structural domain among the snRNAs (Reddy & Busch, 1988; Branlant *et al.*, 1982). Thus, the common sequence structure found at the ends of the introns may well be reflected in both the proteins and RNAs that make up the spliceosome.

The similarity between donor and acceptor sites and the predominance of information residing on the intron side might also be explained if introns were created by the insertion of transposable elements (Rogers, 1985; Cavalier-Smith, 1985; Hickey & Benkel, 1986; Hickey *et al.*, 1989). We do not have strong evidence to choose between this model and ours, but they are not mutually exclusive, so combined models are possible.

(e) Energetics of Splice Junction Location

In this paper we investigated the fine structure of a set of RNA patterns bound by the spliceosome and found that they are not arbitrarily constructed objects, but rather that they reflect both the functions performed and the evolutionary history they have passed through. The function of splice sites is to allow the binding of the spliceosome while avoiding interference with coding regions. For this a pattern is needed whose conservation is sufficient for the sites to be located in the pre-mRNA. Evidently that pattern was modified during evolution to place the majority of the conservation on the intron, where it would not restrict the coding freedom of the proteins.

The constraint on total conservation at both donor and acceptor sites reflects the requirement for precise location of the sites. Without the ability to do this, the wrong proteins would be made. Alternative splicing should also be considered precise since the alternative products are used by the cell.

Shannon's channel capacity theorem (Shannon, 1949) applies here. When translated into molecular terms, this theorem states that so long as a certain specific minimum energy is dissipated during binding, the sites may be located *to any arbitrarily low* (but not zero) *degree of precision* (Schneider, 1991*a*). The precision obtained depends on the detailed coding of the binding interaction, *not* on the total energy dissipated. This mathematical result, which is contrary to most molecular biologist's intuition, was also surprising to Shannon and to many workers in the communications field. As Ninio pointed out (Ninio, 1975) accuracy may be obtained by either using a sophisticated binding site, or by a kinetic proofreading system (Hopfield, 1974; Ehrenberg & Blomberg, 1980; Blomberg *et al.*, 1980; Blomberg & Ehrenberg, 1981; Savageau & Lapointe, 1981). Shannon's theorem applies to both, but we would expect proofreading only when molecules being distinguished are so small that a large number of contacts cannot be formed. This may well be the case in nucleic-acid polymerization and translation, but in cases where a large number of contacts can evolve—such as the recognition of RNA binding sites—precision should not be limited by energetics because large numbers of energetically tiny interactions can lead to sharp discriminations without large energy dissipation. Under these conditions, Shannon's theorem allows us to state that splicing could be as precise as is necessary for survival of the organism, irrespective of the binding energy.

Since we know the frequency of splice junctions, we can calculate the required decrease in entropy necessary to place the spliceosome at the acceptor junction, $R_{frequency}$. Knowing that the acceptor is bound first immediately implies that the entropy reduction needed to locate the donor sites must be at least one bit smaller than that needed for the acceptor. Since this was observed, our measurements of $R_{sequence}$ show that the appropriate amounts of information are available at both donor and acceptor sites, so we know that—according to Shannon's theorem—it is possible for the sites to be located precisely. The missing fact in this logic is experimental data on the average binding energies of the spliceosome to binding sites and to non-specific RNA regions, since the difference between these would reveal whether the minimum energy demanded by theory is indeed dissipated. From the Second Law of Thermodynamics—which is a simplified version of the channel capacity theorem under isothermal conditions (Schneider, 1991*b*)—we predict that at least $\mathcal{E}_{min} = k_{\mathrm{B}} T \ln(2)$ joules must be disspated away from the molecular machine for each bit gained (where k_{B} is Boltzmann's constant, T is the temperature in kelvin). Thus binding the donor requires at least $8\mathcal{E}_{min}$ and the acceptor $9\mathcal{E}_{min}$ joules of dissipation. We can show that the appropriate energy dissipation measure in this case is the standard free energy, ΔG° (in preparation). Since $\Delta G^{\circ} = k_{\mathrm{B}} T \ln(K_{spec})$ joules per bit where K_{spec} is the specific binding constant (between non-specific and specific binding) and since $\Delta G^{\circ} \geq R_{frequency}\mathcal{E}_{min}$ (according to the Second Law), the specific binding constant, K_{spec} should be at least 2^8 and 2^9 for the donor and acceptor sites, respectively. Data of this magnitude (10^2 to 10^3) have already been reported for yeast donor sites (Goguel *et al.*, 1991). However, since yeast splice site patterns differ significantly from human ones (Maniatis & Reed, 1987), further experimental work is needed to determine the corresponding numbers for human spliceosomes.

We thank E. Brody, D. Halverson, P. N. Hengen, N. Herman, W. Kasprzak, D. Landsman, P. Lemkin, S. Lesher, D. McPheeters, J. Mack, J. Maizel, H. Martinez, R. A. Padgett, J. Strathern, P. Rogan, D. Rubens, O. White, and M. Yarmolinsky for reading and commenting on the manuscript, W. G. Alvord for statistical advice, and the Advanced Scientific Computing Laboratory for computational resources and their support. We thank M. Yarmolinsky for suggesting the possibility that spliceosomal scanning explains the donor/acceptor discrepancy. R. M. Stephens was supported by the NCI/FCRDC Student Intern Program, the NIH/FAES Mones Berman Memorial Fund, and the NIH Student Research Training Program.

APPENDIX 1 Variation of $R_{sequence}$ Determined by a Monte Carlo Method

An estimate of the variation of $R_{sequence}$ was obtained by a Monte Carlo simulation. Data corresponding to Table 1 of (Schneider *et al.*, 1986) were: n_{site}: the number of binding site sequences available; L: the number of bases across the binding site (from the "range"); $R_{sequence}$: information content of the site; SD: variation of $R_{sequence}$ due to small sample size.

In the first step of this method, a random number with a flat-distribution between 0 and 32 is chosen. For 10,000 values of this "$R_{sequence-desired}$", the following steps are carried out:

A matrix of frequencies, $f(b,l)_{true}$, is generated using a random number generator. Index b is the base (a, c, g or t) and l is the position in the matrix ($l = 1 \ldots L$). This matrix represents the "true" probabilities that would be found in an infinite set of binding-site sequences. By setting all values of the $f(b,l)_{true}$ matrix to 0.25, the initial information content is set to zero. Entries of the matrix are then modified randomly one by one until the information content reaches the current $R_{sequence-desired}$ value. If a modification does not increase the information content, it is rejected and another is tried. The effect is that the matrix is "evolved" to have the desired information content. Thus the final set of 10,000 $f(b,l)_{true}$ matrixes have a flat distribution of information contents ($R_{sequence-true}$) from 0 to 32 bits.

For each $f(b,l)_{true}$ matrix, the following is carried out: A set of n_{site} binding site sequences is generated, each having length L, with frequencies defined by $f(b,l)_{true}$. These are gathered together to produce an $f(b,l)_{simulated}$ frequency matrix, and its information content, $R_{sequence-simulated}$ is calculated (with small sample bias correction) according to (Schneider *et al.*, 1986). Although it is possible to generate several $R_{sequence-simulated}$ values from each $f(b,l)_{true}$, this results in stripes on a plot of $R_{sequence-simulated}$ versus $R_{sequence-true}$. Since this may affect the results, only one $R_{sequence-simulated}$ is calculated per $f(b,l)_{true}$.

If $R_{sequence-simulated}$ falls into the range $R_{sequence} \pm SD$, then the corresponding $R_{sequence-true}$ is collected. The distribution of these collected $R_{sequence-true}$ values is determined. In general it is a normal distribution, with mean close to $R_{sequence}$. One standard deviation of the distribution of $R_{sequence-true}$ replaces SD as the measure of variation. By this means we work backwards from the measurement ($R_{sequence} \pm SD$) and a simulated set of measurements ($R_{sequence-simulated}$) to find the likely population distribution

($R_{sequence-true} \pm \sigma_{population}$).

If n_{site} is small, the distribution diverges significantly from $R_{sequence}$ because of the small sample bias correction, and the results cannot be used. In these cases the value of SD can be used since it gives a minimum for the variation of $R_{sequence}$.

The collected distribution of $R_{sequence-true}$ represents information contents of frequency matrixes which *could have* produced the information content observed from natural sequences. As such, it is not a measure of the variation of $R_{sequence}$, but rather indicates the range of possible $R_{sequence}$ values that might be obtained if a larger number of example binding sites were available.

The program, Rsim version 1.92, is written in Pascal (Jensen & Wirth, 1975) and is available as part of the Delila system (Schneider *et al.*, 1982; Schneider *et al.*, 1984; Schneider & Stephens, 1990), see Materials and Methods.

References

Adobe Systems Incorporated (1985*a*). *PostScript Language Reference Manual.* Addison-Wesley Publishing Company, Reading, MA.

Adobe Systems Incorporated (1985*b*). *PostScript Language Tutorial and Cookbook.* Addison-Wesley Publishing Company, Reading, MA.

Aebi, M., Hornig, H. & Weissmann, C. (1987). 5′ cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5′ splice region, not by the conserved 5′ GU. *Cell,* **50**, 237–246.

Aebi, M. & Weissmann, C. (1987). Precision and orderliness in splicing. *Trends in Genetics,* **3**, 102–107.

Beacham, I. R., Schweitzer, B. W., Warrick, H. M. & Carbon, J. (1984). The nucleotide sequence of the yeast ARG4 gene. *Gene,* **29**, 271–279.

Berg, O. G. (1988). Selection of DNA binding sites by regulatory proteins. Functional specificity and pseudosite competition. *J. Biomol. Struct. Dyn.* **6**, 275–297.

Berg, O. G. & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins, statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.* **193**, 723–750.

Berg, O. G. & von Hippel, P. H. (1988*a*). Selection of DNA binding sites by regulatory proteins. *TIBS,* **13**, 207–211.

Berg, O. G. & von Hippel, P. H. (1988*b*). Selection of DNA binding sites by regulatory proteins. II. The binding specificity of cyclic AMP receptor protein to recognition sites. *J. Mol. Biol.* **200**, 709–723.

Blomberg, C. & Ehrenberg, M. (1981). Energy considerations for kinetic proofreading in biosynthesis. *J. Theor. Biol.* **88**, 631–670.

Blomberg, C., Ehrenberg, M. & Kurland, C. G. (1980). Free-energy dissipation constraints on the accuracy of enzymatic selections. *Quart. Rev. Bioph.* **13**, 231–254.

Branlant, C., Krol, A., Ebel, J. P., Lazar, E., Haendler, B. & Jacob, M. (1982). U2 RNA shares a structural domain with U1, U4, and U5 RNAs. *EMBO J.* **1**, 1259–1265.

Breathnach, R. & Chambon, P. (1981). Organization and expression of eucaryotic split genes for coding proteins. *Annu. Rev. Biochem.* **50**, 349–393.

Brillouin, L. (1951). Physical entropy and information. II. *J. of Applied Physics,* **22**, 338–343.

Brody, E. & Abelson, J. (1985). The "spliceosome": yeast pre-messenger RNA associates with a 40S complex in a splicing-dependent reaction. *Science,* **228**, 963–967.

Brunak, S., Engelbrecht, J. & Knudsen, S. (1990). Cleaning up gene databases. *Nature,* **343**, 123.

Brunak, S., Engelbrecht, J. & Knudsen, S. (1991). Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.* **220**, 49–65.

Cantor, C. R. (1990). Orchestrating the human genome project. *Science,* **248**, 49–51.

Cavalier-Smith, T. (1985). Selfish DNA and the origin of introns. *Nature,* **315**, 283–284.

Csank, C., Taylor, F. M. & Martindale, D. W. (1990). Nuclear pre-mRNA introns: analysis and comparison of intron sequences from *Tetrahymena thermophila* and other eukaryotes. *Nucl. Acids Res.* **18**, 5133–5141.

Ehrenberg, M. & Blomberg, C. (1980). Thermodynamic constraints on kinetic proofreading in biosynthetic pathways. *Biophys. J.* **31**, 333–358.

Eiglmeier, K., Honoré, N., Iuchi, S., Lin, E. C. C. & Cole, S. T. (1989). Molecular genetic analysis of FNR-dependent promoters. *Mol. Microb.* **3**, 869–878.

Fichant, G. & Gautier, C. (1987). Statistical method for predicting protein coding regions in nucleic acid sequences. *CABIOS,* **3**, 287–295.

Fields, C. (1990). Information content of *Caenorhabditis elegans* splice site sequences varies with intron length. *Nucl. Acids Res.* **18**, 1509–1512.

Frendewey, D. & Keller, W. (1985). Stepwise assembly of a pre-mRNA splicing complex requires U-snRNPs and specific intron sequences. *Cell,* **42**, 355–367.

García-Blanco, M. A., Jamison, S. F. & Sharp, P. A. (1989). Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. *Genes Dev.* **3**, 1874–1886.

Gelfand, M. S. (1990). Computer prediction of the exon-intron structure of mammalian pre-mRNAs. *Nucl. Acids Res.* **18**, 5865–5869.

Goguel, V., Liao, X., Rymund, B. C. & Rosbash, M. (1991). U1 snRNP can influence 3′-splice site selection as well as 5′-splice site selection. *Genes Dev.* **5**, 1430–1438.

Goodrich, J. A., Schwartz, M. L. & McClure, W. R. (1990). Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). *Nucl. Acids Res.* **18**, 4993–5000.

Grabowski, P. J., Nasim, F. H., Kuo, H. & Burch, R. (1991). Combinatorial splicing of exon pairs by two-site binding of U1 small nuclear ribonucleoprotein particle. *Mol. Cell. Biol.* **11**, 5919–5928.

Grabowski, P. J., Seiler, S. R. & Sharp, P. A. (1985). A multicomponent complex is involved in the splicing of messenger RNA precursors. *Cell,* **42**, 345–353.

Green, M. R. (1986). Pre-mRNA splicing. *Annu. Rev. Genet.* **20**, 671–708.

Green, M. R. (1991). Biochemical mechanisms of constitutive and regulated pre-mRNA splicing. *Annu. Rev. Cell Biol.* **7**, 559–599.

Guthrie, C. & Patterson, B. (1988). Spliceosomal snRNAs. *Annu. Rev. Genet.* **22**, 387–419.

Hartz, D., McPheeters, D. S., Traut, R. & Gold, L. (1988). Extension inhibition analysis of translation initiation complexes. *Meth. Enzym.* **164**, 419–425.

Herman, N. D. & Schneider, T. D. (1992). High information conservation implies that at least three proteins bind independently to F plasmid *incD* repeats. *J. Bact.* **174**, 3558–3560.

Hickey, D. A. & Benkel, B. (1986). Introns as relict retrotransposons: implications for the evolutionary origin of eukaryotic mRNA splicing mechanisms. *J. Theor. Biol.* **121**, 283–291.

Hickey, D. A., Benkel, B. F. & Abukashawa, S. M. (1989). A general model for the evolution of nuclear pre-mRNA introns. *J. Theor. Biol.* **137**, 41–53.

Hopfield, J. J. (1974). Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity. *Proc. Natl. Acad. Sci. USA,* **71**, 4135–4139.

Hopfield, J. J. & Tank, D. W. (1986). Computing with neural circuits: a model. *Science,* **233**, 625–633.

Jensen, K. & Wirth, N. (1975). *Pascal User Manual and Report.* Springer-Verlag, New York.

Kanehisa, M. (1988). *IDEAS - 88: Integrated Database and Extended Analysis System for Nucleic Acids and Proteins. User Manual.* revised edition, Advanced Scientific Computing Laboratory, Frederick Cancer Research Facility, MD.

Kastner, B., Bach, M. & Lührmann, R. (1990). Electron microscopy of small nuclear ribonucleoprotein (snRNP) particles U2 and U5: evidence for a common structure-determining principle in the major U snRNP family. *Proc. Natl. Acad. Sci. USA,* **87**, 1710–1714.

Keller, E. B. & Noon, W. A. (1984). Intron splicing: a conserved internal signal in introns of animal pre-mRNA's. *Proc. Natl. Acad. Sci. USA,* **81**, 7417–7420.

Krämer, A. (1987). Analysis of RNase-a-resistant regions of adenovirus 2 major late precursor-mRNA in splicing extracts reveals an ordered interaction of nuclear components with the substrate RNA. *J. Mol. Biol.* **196**, 559–573.

Kudo, M., Lida, Y. & Shimbo, M. (1987). Syntactic pattern analysis of $5'$-spice site sequences of mRNA precursors in higher eukaryote genes. *CABIOS,* **3**, 319–324.

Kühne, T., Wieringa, B., Reiser, J. & Weissmann, C. (1983). Evidence against a scanning model of RNA splicing. *EMBO J.* **2**, 727–733.

Lamond, A. I., Konarska, M. M. & Sharp, P. A. (1987). A mutational analysis of spliceosome assembly: evidence for splice site collaboration during spliceosome formation. *Genes Dev.* **1**, 532–543.

Lang, K. M. & Spritz, R. A. (1983). RNA splice site selection: evidence for a $5' \to 3'$ scanning model. *Science,* **220**, 1351–1355.

Lührmann, R. (1988). snRNP proteins. In *Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles*, (Birnstiel, M. L., ed.), vol. 1, pp. 71–99, Springer-Verlag, Berlin.

Lukashin, A. V., Engelbrecht, J. & Brunak, S. (1992). Multiple alignment using simulated annealing: branch point definition in human mRNA splicing. *Nucl. Acids Res.* **20**, 2511–2516.

Maniatis, T. & Reed, R. (1987). The role of small nuclear ribonucleoprotein particles in pre-mRNA splicing. *Nature,* **325**, 673–678.

Mars, M. & Beaud, G. (1987). Characterization of vaccinia virus early promoters and evaluation of their informational content. *J. Mol. Biol.* **198**, 619–631.

Mount, S. M. (1982). A catalogue of splice junction sequences. *Nucl. Acids Res.* **10**, 459–472.

Mount, S. M., Pettersson, I., Hinterberger, M., Karmas, A. & Steitz, J. A. (1983). The U1 small nuclear RNA-protein complex selectively binds a $5'$ splice site *in vitro*. *Cell,* **33**, 509–518.

Nakata, K., Kanehisa, M. & DeLisi, C. (1985). Prediction of splice junctions in mRNA sequences. *Nucl. Acids Res.* **13**, 5327–5340.

Newman, A. J. & Norman, C. (1992). U5 snRNA interacts with exon sequences at $5'$ and $3'$ splice sites. *Cell,* **68**, 743–754.

Ninio, J. (1975). Kinetic amplification of enzyme discrimination. *Biochimie,* **57**, 587–595.

Ohshima, Y. & Gotoh, Y. (1987). Signals for selection of a splice site in pre-mRNA: computer analysis of splice junction sequences and like sequences. *J. Mol. Biol.* **195**, 247–259.

Olsen, G. J. (1983). *Comparative Analysis of Nucleotide Sequence Data.* PhD thesis, University of Colorado Health Science Center.

O'Neill, M. C. (1991). Training back-propagation neural networks to define and detect DNA-binding sites. *Nucl. Acids Res.* **19**, 313–318.

Padgett, R. A., Grabowski, P. J., Konarska, M. M., Seiler, S. & Sharp, P. A. (1986). Splicing of messenger RNA precursors. *Annu. Rev. Biochem.* **55**, 1119–1150.

Penotti, F. E. (1990). Human DNA TATA boxes and transcription initiation sites: a statistical study. *J. Mol. Biol.* **213**, 37–52.

Penotti, F. E. (1991). Human pre-mRNA splicing signals. *J. Theor. Biol.* **150**, 385–420.

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

Pikielny, C. W., Teem, J. L. & Rosbash, M. (1983). Evidence for the biochemical role of an internal sequence in yeast nuclear mRNA introns: implications for U1 RNA and metazoan mRNA splicing. *Cell,* **34**, 395–403.

Quinqueton, J. & Moreau, J. (1985). Application of learning techniques to splicing site recognition. *Biochimie,* **67**, 541–547.

Reddy, R. & Busch, H. (1988). Small nuclear RNAs: RNA sequences, structure, and modifications. In *Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles*, (Birnstiel, M. L., ed.), vol. 1, pp. 1–37, Springer-Verlag, Berlin.

Reed, R., Griffith, J. & Maniatis, T. (1988). Purification and visualization of native spliceosomes. *Cell,* **53**, 949–961.

Reed, R. & Maniatis, T. (1986). A role for exon sequences and splice-site proximity in splice-site selection. *Cell,* **46**, 681–690.

Robberson, B. L., Cote, G. J. & Berget, S. M. (1990). Exon definition may facilitate splice site selection in RNAs with multiple exons. *Mol. Cell. Biol.* **10**, 84–94.

Rogers, J. (1985). Exon shuffling and intron insertion in serine protease genes. *Nature,* **315**, 458–459.

Rosbash, M. & Séraphin, B. (1991). Who's on first? the U1 snRNP-5′ splice site interaction and splicing. *TIBS,* **16**, 187–190.

Rothstein, J. (1951). Information, measurement, and quantum mechanics. *Science,* **114**, 171–175.

Savageau, M. A. & Lapointe, D. S. (1981). Optimization of kinetic proofreading: a general method for derivation of the constraint relations and an exploration of a specific case. *J. Theor. Biol.* **93**, 157–177.

Schneider, T. D. (1988). Information and entropy of patterns in genetic switches. In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, (Erickson, G. J. & Smith, C. R., eds), vol. 2, pp. 147–154, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Schneider, T. D. (1991*a*). Theory of molecular machines. I. Channel capacity of molecular machines. *J. Theor. Biol.* **148**, 83–123.

Schneider, T. D. (1991*b*). Theory of molecular machines. II. Energy dissipation from molecular machines. *J. Theor. Biol.* **148**, 125–137.

Schneider, T. D. & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucl. Acids Res.* **18**, 6097–6100.

Schneider, T. D. & Stormo, G. D. (1989). Excess information at bacteriophage T7 genomic promoters detected by a random cloning technique. *Nucl. Acids Res.* **17**, 659–674.

Schneider, T. D., Stormo, G. D., Gold, L. & Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* **188**, 415–431.

Schneider, T. D., Stormo, G. D., Haemer, J. S. & Gold, L. (1982). A design for computer nucleic-acid sequence storage, retrieval and manipulation. *Nucl. Acids Res.* **10**, 3013–3024.

Schneider, T. D., Stormo, G. D., Yarus, M. A. & Gold, L. (1984). Delila system tools. *Nucl. Acids Res.* **12**, 129–140.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423, 623–656.

Shannon, C. E. (1949). Communication in the presence of noise. *Proc. IRE,* **37**, 10–21.

Shannon, C. E. & Weaver, W. (1949). *The Mathematical Theory of Communication.* University of Illinois Press, Urbana.

Sharp, P. A. (1987). Splicing of messenger RNA precursors. *Science,* **235**, 766–771.

Sillekens, P. T. G., Habets, W. J., Beijer, R. P. & van Venrooij, W. J. (1987). cDNA cloning of the human U1 snRNA-associated A protein: extensive homology between U1 and U2 snRNP-specific proteins. *EMBO J.* **6**, 3841–3848.

Steitz, J. A., Black, D. L., Gerke, V., Parker, K. A., Krämer, A., Frendewey, D. & Keller, W. (1988). Functions of the abundant U-snRNPs. In *Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles*, (Birnstiel, M. L., ed.), vol. 1, pp. 115–154, Springer-Verlag, Berlin.

Stephens, R. M. & Schneider, T. D. (1992). Features of spliceosome evolution and function inferred from an analysis of the information at human splice sites. *J. Mol. Biol.* **228**, 1124–1136.

Stormo, G. D. (1990). Consensus patterns in DNA. *Meth. Enzym.* **183**, 211–221.

Stormo, G. D., Schneider, T. D., Gold, L. & Ehrenfeucht, A. (1982*a*). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli. Nucl. Acids Res.* **10**, 2997–3011.

Stormo, G. D., Schneider, T. D. & Gold, L. M. (1982*b*). Characterization of translational initiation sites in *E. coli. Nucl. Acids Res.* **10**, 2971–2996.

Talerico, M. & Berget, S. M. (1990). Effect of $5'$ splice site mutations on splicing of the preceding intron. *Mol. Cell. Biol.* **10**, 6299–6305.

Tribus, M. & McIrvine, E. C. (1971). Energy and information. *Sci. Am.* **225** (3), 179–188. (Note: the table of contents in this volume incorrectly lists this as volume **224**).

Wang, X. & Padgett, R. A. (1989). Hydroxyl radical "footprinting" of RNA: application to pre-mRNA splicing complexes. *Proc. Natl. Acad. Sci. USA,* **86**, 7795–7799.

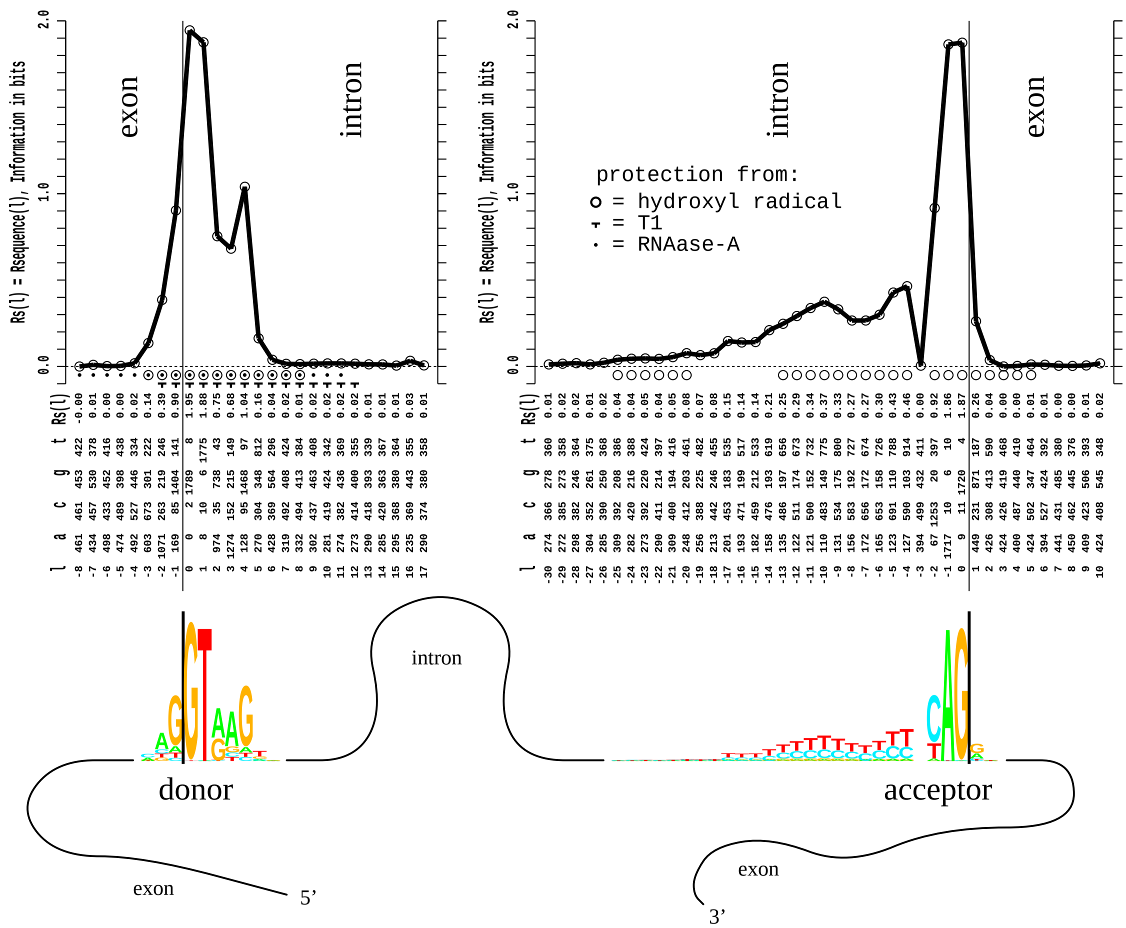
Watson, J. D. (1990). The human genome project: past, present and future. *Science,* **248**, 44–49.

Watson, J. D., Hopkins, N. H., Roberts, J. W., Steitz, J. A. & Weiner, A. M. (1987). *Molecular Biology of the Gene.* fourth edition, The Benjamin/Cummings Publishing Co., Inc., Menlo Park, California.

Weber, I. T. & Steitz, T. A. (1984). A model for the non-specific binding of catabolite gene activator protein to DNA. *Nucl. Acids Res.* **12**, 8475–8487.

Zieve, G. W. & Sauterer, R. A. (1990). Cell biology of the snRNP particles. *CRC Crit. Rev. Bioch. Mol. Biol.* **25**, 1–46.

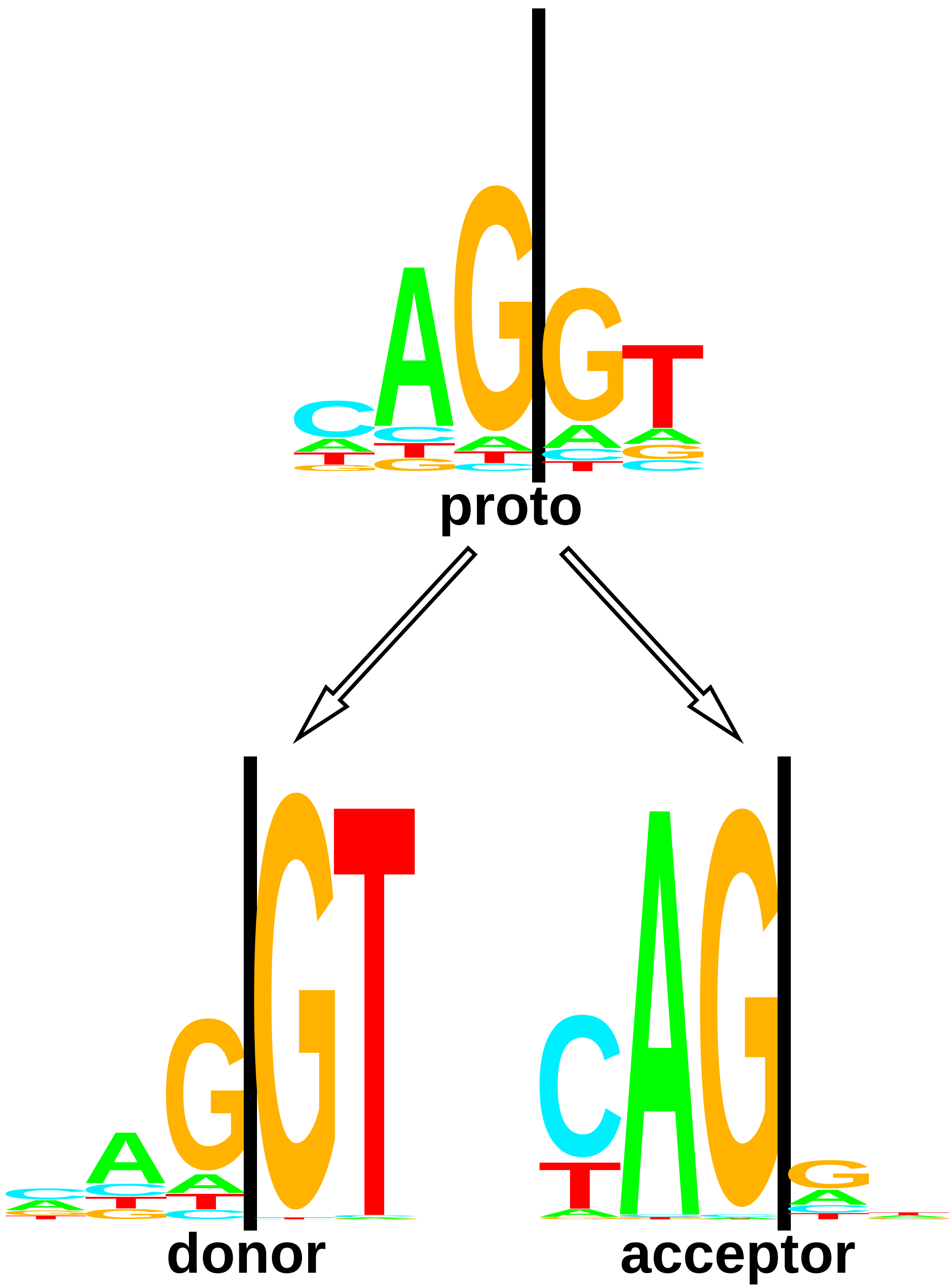
Figure 1: Information curves and sequence logos for human spliceosome binding sites.
The left half of the figure shows the donor splice sites from position -8 to position +17, while the right half shows the -30 to +10 region around the acceptor sites. Position zero on both curves is the point on the intron adjacent to the splice point, *i.e.* on the $5'$ side, the intron is cut immediately before position zero while on the $3'$ side it is cut immediately after position zero. (These are the coordinates provided by GenBank.)
In the matrix corresponding to each graph, the bottom row, labeled l, contains the position on the sequences relative to the splice points. The next four rows are the numbers of a's, c's, g's, and t's (labeled as such) found at each position. These were used to create the frequency matrix for the analysis. For random sequences the frequencies at a position in the matrix should be about equal, and examination of the matrixes at the edges shows this to be the case. Examination of the matrixes around the zero points, however, shows a decided inequality in the numbers of the various bases. This means that the sequences around these zero positions are not random, and therefore there is information (conservation) at these points (the spikes on the graph). The top row of the matrixes, labeled $R_s(l)$ $(= R_{sequence}(l))$, is the amount of information present at position l on the sequences. The symbols found between this row and the graph represent those positions apparently protected by the spliceosome in protection experiments (Mount *et al.,* 1983; Wang & Padgett, 1989; R. A. Padgett, personal communication).
The curve and the matrix are summarized by the sequence logos (Schneider & Stephens, 1990) at the bottom of the figure. In a logo, the total height of the stack of letters at each position is the amount of information present at that position. The heights of the individual letters are proportional to their frequencies at that position. The letters are ordered with the most frequent on top, so the most common base appears on the top of the logo and one may read the consensus sequence directly from the figure. The vertical bars are 2 bits high; the region between them is removed during splicing. Error bars for the heights of the information curves and sequence logos are not shown in the figure because they are below the resolution of the printer.



l	a	c	g	t	Rs(l)
-8	461	461	453	422	-0.00
-7	434	457	530	378	0.01
-6	498	433	452	416	0.00
-5	474	489	398	438	0.00
-4	492	527	446	334	0.02
-3	603	673	301	222	0.14
-2	1071	263	219	246	0.39
-1	169	85	1404	141	0.90
0	0	2	1789	8	1.95
1	8	10	6	1775	1.88
2	974	35	738	43	0.75
3	1274	152	215	149	0.68
4	128	95	1468	97	1.04
5	270	304	348	812	0.16
6	428	369	564	296	0.04
7	319	492	408	424	0.02
8	332	494	413	384	0.01
9	302	437	463	408	0.02
10	281	419	424	342	0.02
11	274	382	436	369	0.02
12	273	414	400	355	0.02
13	290	418	393	339	0.01
14	285	420	363	367	0.01
15	295	368	380	364	0.01
16	235	369	443	355	0.03
17	290	374	380	358	0.01

l	a	c	g	t	Rs(l)
-30	274	366	278	360	0.01
-29	272	385	273	358	0.02
-28	298	382	246	364	0.02
-27	304	352	261	375	0.01
-26	285	390	250	368	0.02
-25	309	392	208	386	0.04
-24	282	420	216	388	0.04
-23	273	392	220	424	0.05
-22	290	411	214	397	0.04
-21	309	400	194	416	0.05
-20	248	412	203	461	0.08
-19	256	388	225	482	0.07
-18	213	442	246	455	0.08
-17	201	453	183	535	0.15
-16	193	471	199	517	0.14
-15	182	459	212	533	0.14
-14	158	476	193	619	0.21
-13	135	486	197	656	0.25
-12	122	511	174	673	0.29
-11	121	500	152	732	0.34
-10	110	483	149	775	0.37
-9	131	534	175	800	0.33
-8	156	583	192	727	0.27
-7	172	656	172	674	0.27
-6	165	653	158	726	0.30
-5	123	691	110	788	0.43
-4	127	590	103	914	0.46
-3	394	499	432	411	0.00
-2	67	1253	20	397	0.92
-1	1717	10	6	10	1.86
0	9	11	1720	4	1.87
1	449	231	871	187	0.26
2	426	308	413	590	0.04
3	424	426	419	468	0.00
4	400	487	440	410	0.00
5	424	502	347	464	0.01
6	394	527	424	392	0.01
7	441	431	485	380	0.00
8	450	462	445	376	0.00
9	409	423	506	393	0.01
10	424	408	545	348	0.02

Figure 2: Sequence logos for proto-, donor, and acceptor splice sites. The logo at the top represents a reasonable guess for the original proto-splice site. It was created by combining the donor and acceptor frequency matrixes into a single data set consisting of 3543 sequences, and so it represents the "average" pattern. The bottom two logos correspond to parts of those in Fig. 1. The vertical bars are 2 bits high.


proto
donor
acceptor