Distribution and consensus of branch point signals in eukaryotic genes: a computerized statistical analysis

Nomi L.Harris and Periannan Senapathy[1]*
Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139 and [1]Biotechnology Center, University of Wisconsin, Madison, WI 53703, USA

Received November 17, 1989; Revised and Accepted March 30, 1990

ABSTRACT

An intermediate stage in the process of eukaryotic RNA splicing is the formation of a lariat structure. It is anchored at an adenosine residue in intron between 10 and 50 nucleotides upstream of the 3′ splice site. A short conserved sequence (the branch point sequence) functions as the recognition signal for the site of lariat formation. It has been generally assumed that the branch point is recognized mainly by the presence of its unique sequence where the lariat is formed. However, the known branch point consensus sequence is found to be distributed nearly randomly throughout the gene sequence with only a slightly higher frequency in the expected lariat region. Further, the known consensus sequence is found to be clearly inadequate to specify branch points. These observations have implications for understanding the mechanism of branch point recognition in the process of splicing, and the possible evolution of the branch point signal.

INTRODUCTION

A typical eukaryotic gene consists of short coding sequences (exons) interrupted by fairly long non-coding sequences (introns) (1−3). The introns in a primary RNA (pre-mRNA) are removed by the splicing machinery in the nucleus, after which the spliced RNA (mRNA) is transported to the cytoplasm and translated. The splicing process is regulated by several sequences at the junction of exons and introns, and within introns. The splice sites on each side of an intron have been found to possess consensus sequences that are presumably recognized by the splicing machinery. The 5′ splice junction of the intron (the donor site) is marked by the 8-nucleotide conserved sequence (A/C)AG|GT(A/G)AGT (1−3). Introns are bounded at the 3′ end by an acceptor splice site, which consists of a pyrimidine-rich region of about 11 nucleotides, followed by (C/T)AG.

The first step in splicing is the assembly of a large ribonuclear protein complex called a spliceosome on the pre-mRNA (3). Before an intron is released, it forms an intermediate structure called a lariat. In the lariat form, the donor end of the intron forms a 5′-2′ phosphodiester bond with the 2′ hydroxyl group of an adenosine residue near the acceptor splice site (4). The final step of the splicing process occurs when the two exons are joined and the intron is released as a lariat RNA (5).

A short 5−8 nucleotide sequence, containing the adenosine residue at which the lariat is formed, functions as the signal for the lariat-structure formation. This signal, which is called the branch point signal, lies within the intron usually between 10 and 50 nucleotides upstream from the acceptor splice site (4). In yeast, the branch point sequence was found to be the highly conserved heptamer TACTAAC, where the last A is the site of branching. Branch point sequences of other organisms are less highly conserved than those of yeast, but 5-nucleotide consensus sequences have been found both by empirical observation (4−6) and by computer search (7). Using laboratory techniques, Ruskin et al (6) isolated several branch point signals from human beta-globin genes. The sequences they found (CTGAC, CTAAT, CTGAT, CTAAC, and CTCAC) bear a clear resemblance to the last 5 nucleotides of the yeast branch point consensus sequence, TACTAAC. Keller & Noon (7) used a computer program to search introns of various organisms (sea urchins, mice, humans, chickens, etc.) for branch points similar to the previously discovered consensus sequences. The consensus they found for Drosophila was CTAAT, and for rats and humans it was CTGAC. In all organisms examined, the T in position 2 and the A in position 4, appeared to be the most highly conserved nucleotides (7). Brown (4) searched 177 plant introns for possible branch points using a method similar to that used by Keller & Noon (7), and found a consensus sequence (C/T)T(A/G)A(T/C).

In general, the distance of the branch acceptor from the 3′ splice site is a crucial parameter in lariat formation (8,9). Further, base-pairing interactions of branch point sequences with the U2 RNA seems to be required for the correct selection of the branch acceptor nucleotide.

We describe in this paper a statistical analysis of the distribution of branch point signals in genes of different categories of organisms. Our aim was to find if the branch sequence alone was sufficient to specify the lariat site in an intron. We tested this by comparing the frequency and distribution of the branch sequence in the lariat region with those in the other regions of genes (introns and exons, and upstream and downstream of coding sequences). This study would also reveal if the known banch point consensus sequence (CTRAY, R = purine and Y = pyrimidine) is adequate to specify branch points.

* To whom correspondence should be addressed

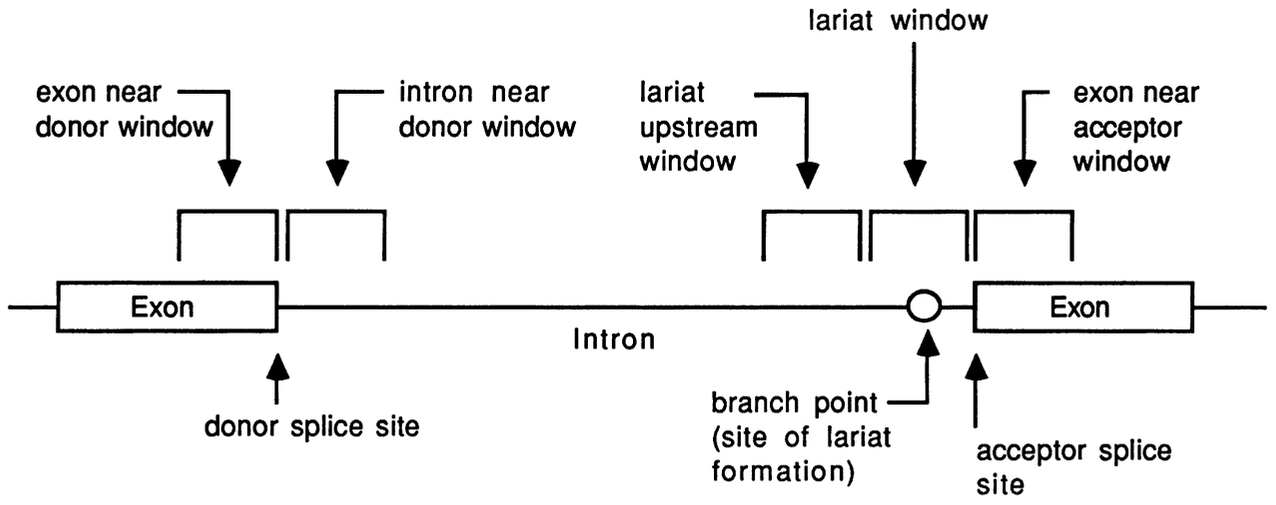

Figure 1. DNA 'windows' in introns and exons analyzed for branch point sequences. Different windows were analyzed for the frequency of occurrences of branch-point-like sequences and for the nucleotide frequencies in these sequences. The windows we examined are: (i) lariat window (−1 to −50 nucleotides upstream of the acceptor splice site in introns); (ii) lariat upstream window (−51 to −100 nucleotides upstream of the acceptor splice site); (iii) exon near acceptor window (1 to 50 nucleotides downstream of the acceptor splice site); (iv) exon near donor window (−1 to −50 nucleotides upstream of the donor splice site); (v) intron near donor window (1 to 50 nucleotides downstream of the donor splice site). Exons are indicated by boxes and the branch point site by an open circle.

MATERIALS AND METHODS

The computer programs used in this study analyzed both GenBank data (release 56.0) and computer-generated random nucleotide sequences. Of the GenBank data, totally 1965 introns and exons were examined. The seven GenBank sequence data files used were primate, rodent, mammal, vertebrate, invertebrate, plant and viral. The programs were written in C and run on a Sun workstation under the UNIX operating system.

The scoring system we used for rating potential branch point sequences was similar to that used by Keller & Noon (7). In their program, introns were examined between 10 and 60 nucleotides upstream from the acceptor splice site. If more than one branch sequence was found in an intron, the sequence closest to the splice site was chosen as the 'primary' signal.

The computer program we developed searched for branch points in a specified window with respect to the acceptor site or the donor site. For example, to search for branch point sequences directly upstream of the acceptor site, the window coordinates are entered as −50, 0. When random sequences were searched, the program simply processed strings of random nucleotides of the specified window size. Random sequences were generated by the computer as described earlier (10).

For each intron, every 5-nucleotide sequence within the window being considered was evaluated as a potential branch point signal. A sequence was assigned a score (between 0 and 100) that reflected how closely it resembled known branch point sequences. We constructed a composite weight table for branch point sequences based on data for Keller and Noon (7) and Brown (4) derived from plant, rat, human, chicken, and Drosophila DNA (Table 1). The weight table shows the percentage of occurrence of each nucleotide at each of the five positions in experimentally found branch point sequences. Each of the five nucleotides in the sequence being evaluated was assigned weight for matching the known consensus. In the present analysis, scores of >96% were taken to represent branch point sequences which correspond to just 4 sequences (CTGAC, CTAAC, CTGAT, CTAAT).

For each window examined, the 5-nucleotide sequence with the highest score was selected as the best potential branch point sequence in that window. After the best branch point for each window was chosen, various statistics about the branch points were compiled. The positions of the best branch points (in each window) and the distribution of scores were recorded. A weight table was also printed for the branch-point sequences in each window category. The intent was to compare the results from actual genes with those from random sequences.

Table 1. The "standard" branch point sequence weight-table.

Nucleotide	Frequency of Occurrence (percent)				
	Sequence Position				
	-3	-2	-1	0	1
A	1	0	39	99	11
C	76	8	15	1	45
G	2	0	42	0	6
T	21	91	4	0	38

This is a composite weight-table derived from experimental data for plant, rat, human, chicken and drosophila DNA published earlier (5,12). The values from these published tables were combined and the percentage for each nucleotide occurring at each position is given here. This table is used as the standard weight table to score potential branch point sequences in the analysis described in this paper.

RESULTS AND DISCUSSION

Consensus and distribution of branch point sequences in eukaryotic genes

Because the site of lariat formation is in the region 10−50 nucleotides upstream of the 3′ splice site, earlier investigators

had looked for branch point sequences only in this region using the computer. This approach fails to account for the possibility that this sequence is not uniquely confined to this region. If the branch point sequences were found only in this particular region, one could presume that this information by itself could be sufficient for the splicing process to detect the lariat site. Otherwise, it would suggest that the branch sequence alone is insufficient for this recognition.

In order to test the above hypothesis, we computed the frequency of the best ranking branch point sequences in a number of different windows (as described under Methods). This analysis would also show if the consensus sequence we have used (and so far known) is adequate to specify the branch points. Figure 1 illustrates the different windows we analyzed for branch point sequences. The window between −50 and −1 nucleotides from the acceptor site, which we refer to as the lariat window, is where branch points have been empirically observed. We also examined the region between −100 and −51 nucleotides from the acceptor site (the lariat upstream window). A region downstream of the acceptor site (which is at the beginning of exons) called the exon near acceptor window, was also analyzed. Two windows near the donor site were searched as well: the exon near donor window, which ran from 50 nucleotides before the donor site up to the 0 position of the donor site (which is at the end of exons), and the intron near donor window, which is the portion of the intron up to 50 nucleotides from the donor site.

In order to obtain a consensus weight matrix, we computed the composite frequencies of each of the 4 nucleotides found at the consensus branch point sequences of plant, rat, human, chick and Drosophila, based on data from Keller and Noon (12) and Brown (5). The resulting 'standard' composite data is given in Table 1.

Based on the standard table, we compared the different windows (see Figure 1) with respect to the frequency of the high-scoring branch point sequence having a score of at least 97% (one of the 4 possible sequences of CTRAY). The frequency (percent) of windows containing branch point sequences were determined for each window category described in Figure 1, as well as for each of the 7 GenBank categories of organisms. The results (Table 2) indicate that the lariat window has the highest percentage of high-scoring branch point sequences. In order to compare these results with a window containing a purely random nucleotide sequence (with 1/4 probability for each nucleotide), we generated 200 different random sequences each 50-nucleotide long, and performed the same analysis. The results (Table 2) show that in the case where the middle nucleotide was constrained to be a G/A as in the standard table (Table 1), the lariat window in invertebrate, mammal, plant, primate, rodent, vertebrate and viral categories of genes contained, respectively, 30, 46, 48, 24, 36, 28 and 14 percent of sequences that matched 97% or better with the standard weight-table, whereas the random sequence contained only 17% of windows that similarly matched. When the middle nucleotide was not constrained to be G/A, a higher percentage of sequences in each window category, as with the random sequence, matched the standard table. However, the overall pattern of frequencies with high scoring branch point sequences in different window categories remained similar in all GenBank groups of organisms.

The percentages of lariat windows with high-scoring branch point sequences in all GenBank categories, with the exception of viruses, are higher than those of random sequences. Furthermore, the low frequency (24−48%) of high scoring branch point sequences even in the lariat window (when the middle nucleotide was constrained), indicates that the majority of real branch point sequences varies from the standard branch point sequence (CTRAY) used in the current study. It is also clear from Table 1 that approximately more than 50% of branch points must have sequences different from this consensus sequence. Thus the consensus sequence we have used is clearly inadequate to specify branch points. Further analysis is needed to identify the other branch point sequences.

Since the above data indicate that branch point sites are (1) nearly randomly distributed in exons and introns and (2) some exons may lack a branch point site upstream in the lariat region, it was interesting to see the distribution of the branch point sequences in individual genes. We also wanted to see how the top-ranking 5′ and 3′ splice sites were distributed. The distribution of the branch points and the top 30 ranking 5′ and 3′ splice sites are shown for two genes in Figure 2. It shows that the splice sites and branch point sites are almost randomly

Table 2. Branch point sequence frequency in different "windows" of gene sequences.

Category of organisms	Window	Frequency of windows containing high-scoring branch point sequences (percent)	
		Unconstrained middle nt	Constrained middle nt
Invertebrate	exon near acceptor	21	4
	lariat	48	30
	lariat upstream	28	11
	exon near donor	18	11
	intron near donor	25	13
Mammal	exon near acceptor	20	11
	lariat	57	46
	lariat upstream	35	14
	exon near donor	25	13
	intron near donor	34	19
Plant	exon near acceptor	27	14
	lariat	62	48
	lariat upstream	32	18
	exon near donor	30	14
	intron near donor	37	24
Primate	exon near acceptor	24	11
	lariat	46	24
	lariat upstream	26	13
	exon near donor	36	17
	intron near donor	31	13
Rodent	exon near acceptor	15	9
	lariat	68	36
	lariat upstream	24	4
	exon near donor	34	15
	intron near donor	31	12
Vertebrate	exon near acceptor	24	19
	lariat	47	28
	lariat upstream	37	20
	exon near donor	30	11
	intron near donor	35	18
Viral	exon near acceptor	31	9
	lariat	39	14
	lariat upstream	29	12
	exon near donor	32	0
	intron near donor	26	8
Random	(50 nt)	32	17

Frequency of windows with high-scoring (97% or higher match) branch point sequence in genes of different categories of organisms was computed as follows. Potential branch point sequences were rated and scores were assigned according to how well they matched the standard (composite) consensus sequence we made from the published data (see Table 1). The 5-nucleotide sequence, among all the possible 5-nucleotide sequences in a window, that best matched the standard consensus was taken as the branch point sequence in that window. If the score of this branch point sequence was 97% or greater, then the window was counted as having a branch point sequence. The frequency of such windows containing a branch point is given in the table for each window described in Figure 1 and for each of the 7 categories of organisms. The middle position of the branch point sequence was either constrained to match the 'standard' consensus (G/A) or left unconstrained -- when constrained, the branch point sequence should best match all the 5 nucleotides in the consensus sequence; when it was not constrained, the branch point sequence should best match all these positions except the 4th position, i.e., the A or G.

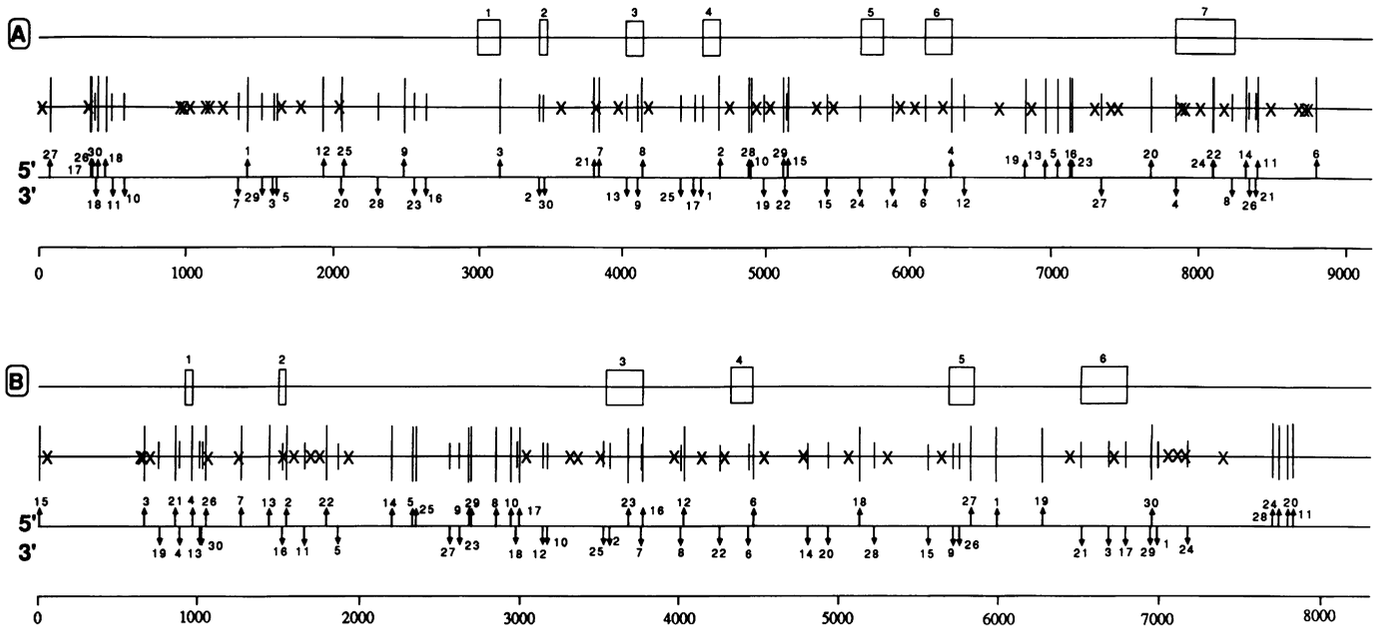

Figure 2. The nearly random distribution of splice sites and branch point sites in eukaryotic gene sequences. Eukaryotic gene sequences containing at least 5 exons were chosen randomly to analyze the distribution of splice signals in them. The sequences were analyzed by our computer program RATE, which ranks each sequence location based on the splice-site scoring matrix (as described in reference 2). In each sequence the top 30 ranking locations are shown ('|' for 3′ ss and '|' for 5′ ss). Similarly, we found the locations matching with the branch point sequences, CTRAY, using the computer (marked 'x'). In the figure, the distribution of the 5′ ss, 3′ ss and branch point sites are shown for 2 genes: (A) Chicken Ovalbumin gene and (B) Human Interleukin-1 beta gene. The exon locations are indicated by numbered boxes on the first line of each gene. The second line shows the distribution of splice sites and branch point sites. On the third line, the up arrows indicate the 5′ ss (with their corresponding ranks), while the 3′ ss are indicated by the down arrows (along with their ranks). The fourth line is a reference scale in number of nucleotides.

distributed in exons and introns as well as the untranslated upstream and downstream sequences. Only 5 out of 11 exons (leaving the two first-exons) in the two genes shown contained a branch point sequence CTRAY in the expected region. Many branch point sites occur within exons and in other intron regions. Analysis with a large number of other genes indicated a similar pattern. Thus, the problem of 'selecting' the right sites that circumscribe the exons by the cellular machinery seems to be very complex. One way to identify the correct branch point sequence(s) is to look for other consensus sequences within lariat regions where branch point sites are missing.

Implications of the random distribution of branch point sequences throughout genes

The fact that branch point-like sequences are found in such abundance in eukaryotic genes implies that the splicing machinery must have some mechanism for recognizing the correct branch point site. The process of splicing seems to recognize exons and introns sequentially with a 5′ to 3′ scanning mechanism. The present results indicate that the scanner could not identify the donor splice site, branch point, and acceptor splice site within an intron in a sequential manner, because of the random occurrence of the branch point sequence in introns. Thus, even a scanning model, in which a donor site is identified first and then the intron scanned for a branch point site which helps determine the acceptor site, is untenable.

We propose that the splicing machinery first locates the donor splice site for an intron. In a second step, it locates the first downstream acceptor site which has a good branch point sequence within the first 50 nucleotides upstream of this site. In other words, neither a branch point site nor an acceptor site sequence, occurring downstream of a 5′ ss, can be a real site independent of each other. Both have to occur together in the 5′ branch point−3′ ss orientation (within 10−50 nucleotides) for both to be real sites. This argument is supported by the fact that there exist many non-functional 3′ splice sites scoring higher than the real 3′ splice site in the introns (2,13). This hypothesis is also supported by the observation that, if an intron happens to have more than one sequence that could function as branch point, only the sequence closest to the acceptor splice site appears to serve this function in most cases (6). When the branch point sequence in intron 1 of the human beta-globin gene was removed, it was found that splicing was not prevented; rather, a cryptic branch point sequence upstream of the deleted one was activated. The observation that introns lacking a viable branch point fail to splice out normally confirms the importance of branch points in the splicing process (7).

The nearly random occurrence of branch point sequences in genes may have some implications in understanding their evolution. This observation is consistent with the hypothesis that the very first genes evolved from random primordial sequences by a gene-search mechanism evolving the split-gene architecture in the first genes (10−12); the mechanism selected exons, introns and splice-signals from the pre-existing, primordial, random sequences. Consistent with this is our observation that many good splice-junction sequences and branch point sequences in genes do not function as real splice sites (Figure 2). These findings suggest that real splicing signals may exist in a specific positional context in the gene sequence.

We tabulated the codon frequencies at each of the 5 nucleotide positions of the highest-scoring branch point sequences of the

lariat window. When the middle position was constrained to be a G or an A, we found that the codon at the third position was almost always a stop-codon (data not shown). This is consistent with the hypothesis that branch point sequences evolved from stop codons (10,11). This hypothesis suggests that the splicing mechanism for removing introns was developed in order to overcome the problem of randomly distributed stop codons. A stop-codon scanning mechanism may have been responsible for the evolution of the splice junction signals and the branch point signal from stop codons. Thus, how a particular consensus sequence came to serve this special function of signalling lariat formation may be explained by its possible mode of evolution. However, further analysis is needed to get to the question of what information (other than the known consensus) is required for a site to function as a branch point. One way perhaps is to experimentally look for branch sites in lariat regions which lack the known consensus sequence.

REFERENCES

1. Green, M.R. (1986) Annu. Rev. Genet. *20*, 671–693.
2. Shapiro, M.B. and Senapathy, P. (1986) Nucl. Acids Res. *15*, 7155–7174.
3. Sharp, P.A. (1987) Science *235*, 766–771.
4. Brown, J.W.S. (1986) Nucl. Acids Res. *14*, 9549–9558.
5. Brown, J.W.S., Felix, G. and Frendewey, D. (1986). The EMBO Journal, *5*, 2749- 2758.
6. Ruskin, B., Krainer, A.R., Maniatis, T. and Green, M.R. (1984) Cell, *38*, 317–331.
7. Keller, E.B. and Noon, W.A. (1984). Proc. Natl. Acad. Sci. USA *81*, 7417–7420.
8. Hartmuth, K. and Barta, A. (1988) Mol. Cell. Biol. *8*, 2011–2020.
9. Parker, R., Siliciano, G. and Gutherie, C. (1987) Cell *49*, 229–239.
10. Senapathy, P. (1986) Proc. Nat. Acad. Sci. USA *83*, 2133–2137.
11. Senapathy, P. (1988) Proc. Natl. Acad. Sci. USA *85*, 1129–1133.
12. Senapathy, P. (1988) Mol. Gen. (Life Sci. Adv.) *7*, 53–65..
13. Senapathy, P., Shapiro, M.B. and Harris, N. (1990). In Methods in Enzymology, 'Computer Analysis of Protein and Nucleic Acid Sequences'. R.F. Doolittle Ed. *183*, 252–278.