Nucleic Acids Research, Vol. 20, No. 16 4255–4262

Splicing signals in *Drosophila*: intron size, information content, and consensus sequences

Stephen M.Mount, Christian Burks[1], Gerald Hertz[2], Gary D.Stormo[2], Owen White[3] and Chris Fields[3]

Department of Biological Sciences, Columbia University, New York, NY 10027, [1]Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545, [2]Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309 and [3]Computing Research Laboratory, Box 300001/3CRL, New Mexico State University, Las Cruces, NM 88003-0001, USA

Received May 12, 1992; Accepted July 15, 1992

ABSTRACT

A database of 209 Drosophila introns was extracted from Genbank (release number 64.0) and examined by a number of methods in order to characterize features that might serve as signals for messenger RNA splicing. A tight distribution of sizes was observed: while the smallest introns in the database are 51 nucleotides, more than half are less than 80 nucleotides in length, and most of these have lengths in the range of 59 – 67 nucleotides. Drosophila splice sites found in large and small introns differ in only minor ways from each other and from those found in vertebrate introns. However, larger introns have greater pyrimidine-richness in the region between 11 and 21 nucleotides upstream of 3′ splice sites. The Drosophila branchpoint consensus matrix resembles C T A A T (in which branch formation occurs at the underlined A), and differs from the corresponding mammalian signal in the absence of G at the position immediately preceding the branchpoint. The distribution of occurrences of this sequence suggests a minimum distance between 5′ splice sites and branchpoints of about 38 nucleotides, and a minimum distance between 3′ splice sites and branchpoints of 15 nucleotides. The methods we have used detect no information in exon sequences other than in the few nucleotides immediately adjacent to the splice sites. However, Drosophila resembles many other species in that there is a discontinuity in A + T content between exons and introns, which are A + T rich.

INTRODUCTION

The removal of introns from the mRNA precursors of higher organisms is a complex process involving many factors (for reviews see 1–4). The splicing reaction occurs in two steps and proceeds via a branched or 'lariat' intermediate in which the 5′ end of the intron is joined, via a 2′,5′ phosphodiester bond, to a site within the intron, usually an A near the 3′ splice site. The information required for splicing appears to be limited to sequences adjacent to the three sites involved in the actual chemistry of the splicing reaction (the 5′ splice site, the 3′ splice site, and the branch point), and a pyrimidine-rich region lying between the 3′ splice site and the branchpoint. Factors responsible for the recognition of these sites have been identified. 5′ splice sites in many species fit the consensus MAG|GTRAGT (where | designates the splice site, M indicates A or C, and R indicates A or G) or some closely related variant of it (5–8). This conserved sequence is recognized by the U1 small nuclear ribonucleoprotein particle (U1 snRNP) through basepairing interactions between the 5′ end of U1 RNA and consensus nucleotides (9), and may also be recognized by additional factors (10–13). The branchpoint is recognized by the U2 snRNP (14) in a manner that also involves basepairing (10–13). In the yeast *Saccharomyces cerevisiae*, the sequence UACUAAC (in which branch formation occurs at the underlined A) is nearly invariant and plays a significant role in determining where and whether splicing will occur. While the identical sequence is preferred in mammalian splicing (15), a looser consensus sequence of UNCURAC can be found surrounding the mammalian branchpoint (16). The interaction between U2 and the branchpoint requires additional factors, including the U1 snRNP (17, 18), and at least one auxiliary factor, U2AF, that binds to the pyrimidine-rich region that lies between the branchpoint and the 3′ splice site (19). As yet, recognition of the short consensus sequence found at the 3′ splice site (YAG|G, or simply AG) has not been definitively attributed to any factor.

There is some species specificity in the interpretation of splicing signals. For example, many introns from the worm *Caenorhabditis elegans* are significantly shorter than vertebrate introns, and this species also appears to have splice site consensus sequences that differ significantly from those found in mammals, particularly at the 3′ splice site (20). As expected from these sequence features, short *C.elegans* introns are not spliced in nuclear extracts derived from human cells (21, 22). Furthermore, correlations exist between intron size and splice site sequences (23). The 5′ splice sites of *C.elegans* introns greater than 75

nucleotides have significantly more information than those of shorter introns, primarily due to conservation at intron positions 4, 5 and 6. In addition, introns in many species, including plants, *C.elegans* and Drosophila, but not mammals or the yeast *S.cerevisiae*, are significantly more A+T rich than flanking exons (24, 25), and the inability of plant introns to splice in mammalian nuclear extracts has been attributed to differences in the A+T content between species (26, 27).

Drosophila introns appear to have splice site consensus sequences much like those found in vertebrates (8) and a recognizable branchpoint consensus (28–30). In addition, known components of the splicing machinery, including U-RNAs (2, 31) and snRNP proteins (2, 31, 32), are highly conserved between Drosophila and humans. However, an earlier study of length distribution of Drosophila introns (33) found many Drosophila introns smaller than the minimum size for splicing in mammalian cells (34, 35). A cursory survey of Drosophila introns also appeared to indicate that a sizeable minority lack strongly pyrimidine-rich stretches adjacent to their 3′ splice sites (see, for example 36–38). In addition, in vitro splicing results in Drosophila Kc cell and human HeLa cell nuclear extracts indicate that the sequence requirements of these two species differ (39). In one case (30), a short (74 nucleotide) Drosophila intron was found to splice well in extracts from Drosophila cell nuclei, but showed no activity in extracts from human cells. Conversely, lengthened versions of the same intron (90 nucleotides) were observed to splice well in human extracts but were not active substrates for the Drosophila system.

We have undertaken a thorough examination of sequences from Drosophila introns and flanking exons using a number of computational methods, including information content analysis (23, 40), two independent consensus-search methods, CONSENSUS (41, 42) and RTIDE (43, 44), and simple assessment of the frequencies and distributions of subsequences from individual nucleotides to hexanucleotides. Our goal was to further define the relevant consensus sequences and explore the possibility that there are size-dependent sequence features. We have confirmed the very sharp length distribution previously reported by Hawkins (33). We have also observed that the highest information content is at the 5′ and 3′ splice sites, and that these sites differ in only minor ways between large and small introns and from vertebrate splice sites. However, large introns do tend to have a greater density of pyrimidines in the region upstream of the 3′ splice site. A putative branchpoint consensus, C T A A T, differs from the mammalian signal primarily in the absence of G at the position immediately preceding the branchpoint. The distribution of occurrences of this sequence in Drosophila suggests a minimum distance between 5′ splice sites and branchpoints of about 38 nucleotides and a minimum distance between 3′ splice sites and branchpoints of about 15 nucleotides.

DATABASE AND METHODS

Data sets

Database entries containing Drosophila nucleotide sequences were taken from the invertebrate division of Release 64.0 of GenBank (45). This data set was manually edited to remove redundant copies of the same gene, tRNA-coding genes, and other questionable entries. Subsets of sequence data were automatically extracted from the collection of Drosophila entries using the program ExtractGenBank (R.Farber, Los Alamos National Laboratory), keying off the annotated IVS regions in the FEATURES table. These subsets consisted of whole introns, 101 nucleotide-long windows approximately centered on 5′ splice sites, and 101 nucleotide-long windows approximately centered on 3′ splice sites. This procedure resulted in splice sites from introns whose sequences were determined in their entirety. The spans listed with the IVS annotation were used to calculate intron lengths. The data set was divided into short and long introns. 80 nucleotides was chosen as the cut-off value for this division after examining the distribution of lengths shown in Figure 1. This resulted in the data sets listed in Table 1. The number of large introns is greater than the number of 5′ splice sites or 3′ splice sites from large introns because of cases of alternative splicing in which the individual splice sites may be used to generate more than one intron.

Methodology

Frequency matrices and information content were determined as described in Fields (23). Statistical uncertainties were calculated using the 'exact' method described in Schneider *et al.* (40), except that the observed frequencies at each position are used in the numerical expansion, rather than the genomic probabilities.

The identification of conserved sequences, with emphasis on definition of the branchpoint consensus, was performed with three different approaches. The RTIDE program (43, 44) takes as input approximately aligned sequences and finds the most common 'word', allowing for mismatches, within a specified 'window' of sequence. The word length, window size and amount of allowed mismatch are all user specified parameters. The CONSENSUS program (41, 42, 46) takes as input unaligned sequences and attempts to find an alignment that maximizes the information content of the identified sites. The output includes a 'specificity matrix' which can be used by the program PATSER to find matches to the matrix in any sequence (42, 47). Finally, occurrences of oligonucleotide sequences of various lengths within these data sets (or subsets of them), were determined.

RESULTS AND DISCUSSION

Size of introns

We first examined the distribution of sizes among the introns in our sample (Figure 1). As had been noted by Hawkins (33), the majority of Drosophila introns are relatively small. In our data set, the median length is 79 nucleotides, and there is a sharp distribution of sizes around a modal length of approximately 63 nucleotides (Figure 1B). This is in marked contrast to the situation in mammals, where introns of less than 70 nucleotides are extremely rare and do not splice well in vitro (48) or in vivo (34). Our data sets were restricted to completely sequenced introns. This restriction introduces a bias against larger introns

Table 1. Drosophila intron data sets

category	parent intron sizes	number of examples
complete introns	small (51–80)	107
	large (81–5392)	102
5′ splice sites	small (51–80)	107
	large (81–5392)	99
3′ splice sites	small (51–80)	107
	large (81–5392)	98

(Figure 1A). Many Drosophila introns larger than 6,000 nucleotides have been described, yet none have been completely sequenced. However, this bias should not greatly affect the apparent distribution of sizes within the set of small introns (Figure 1B). For example, the occurrence of 32 introns in the range of 61–65 nucleotides but only five introns in the range of 76–80 nucleotides cannot be explained by the difficulties associated with sequencing or reporting an additional 15 nucleotides. A similar distribution of intron sizes has been described for *C.elegans*, except that the modal size in the worm is approximately 50 nucleotides (20, 23).

Splice site consensus and information content

A frequency matrix was prepared from 5′ splice sites and 3′ splice sites in each of the two size classes. These nucleotide frequencies are graphed in a number of ways in Figure 2, and values for nucleotides near the splice sites are reported in Table 2. We then applied the information content measure of Schneider *et al.* (23, 40), which reflects the extent of deviation from random base composition, to the data in the frequency matrices. The resulting distributions are shown in Figure 3.

A comparison of the nucleotide frequency distribution within 5′ splice sites in Drosophila with all of those compiled by Senapathy *et al.* (8), most of which are from vertebrates, shows considerable similarity (Table 2A). In each case the consensus MAG|GTRAGT is valid and intron position 5 is the most highly conserved nucleotide outside of the invariant GT. One minor difference between Drosophila and vertebrates is that position 6 is less variable in Drosophila (68%T) than it is in the total set (50%T). Short and long Drosophila introns are likewise similar. There are two cases of deviation from the rule that introns begin with GT in this data set. They are the *Gpdh* gene intron C, which begins with GC, in the sequence context AAGGCAAGT, and *rudimentary* intron E, which begins CT in GCGCTGAGA. The one case of deviation from the rule that introns end with AG is *perB* intron E, which ends CG. The phenomenon of rare nonconsensus sites has been described previously (8), but the frequency of exceptions is low in all data sets, and many apparent exceptions have proven to be errors in one way or another (49). In each of the cases cited here, the exception was based on cDNA sequence information and the authors noted the exception in a refereed journal article.

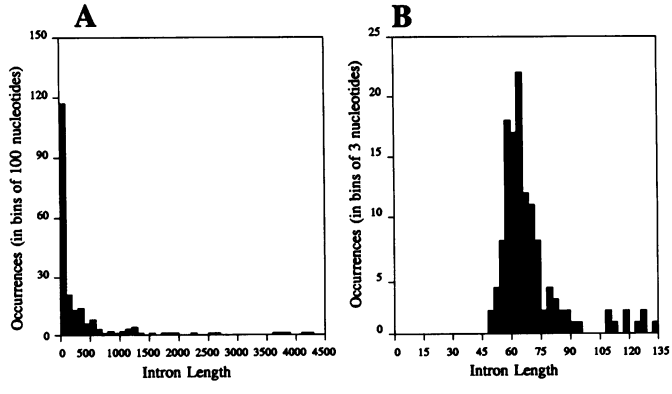

Figure 1. Size distribution of Drosophila introns. A. All introns with lengths less than 6,000 nucleotides are included, and the number of examples in each bin of 100 nucleotides is plotted. B. The distribution of sizes among introns of less than 135 nucleotides are plotted with a bin size of 3.

As expected, there is a peak of information content at each splice site, corresponding to the traditional splice site consensus sequences, and the total information in both 5′ and 3′ splice sites from the large and small data sets are comparable. A large peak of information around position +40 of short intron 5′ splice sites is the only significant difference in information content between large and small introns, and corresponds to the branchpoint (see below).

Nucleotide frequencies

Both large and small introns are characterized by higher A+T content than exons. Base composition in the region between −51 to −42 relative to 3′ splice sites and +21 and +30 relative to 5′ splice sites (areas relatively devoid of information; see Figure 2) is 32% A, 19% C, 16% G and 33% T. This intron A+T content of 65% compares with an A+T content of 48% in the 50 nucleotide window of flanking exonic sequence contained in our database. This difference of 17% in A+T content between introns and exons is reminiscent of plant (particularly dicot) introns (50, 51), in which high A+T content has been shown to be a signal for splice site recognition (24, 26, 27). More recently, Csank *et al.* (25) have observed that high A+T content is a general property of introns in many species. Ironically, the two most highly studied groups, the yeast *Saccharomyces cerevisiae* and mammals are among the few exceptions.

Of particular interest is the pyrimidine-rich region associated with 3′ splice sites, which has been shown to play a critical role in branchpoint recognition during mammalian splicing (19) (Reviewed in 3). In our Drosophila data set, it appears that the pyrimidine-rich region of vertebrate introns is replaced by a large T-rich region and a much smaller C-rich region. Examination of Figure 2 shows that the region between −30 and −10 relative

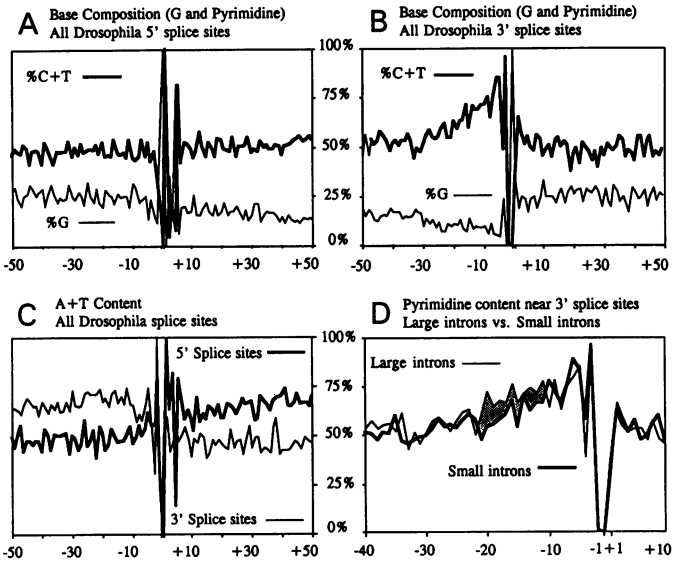

Figure 2. Nucleotide frequencies. C + T and G content is plotted for a 100 nucleotide window around each type of splice site (**A**: 5′ splice sites; **B**: 3′ splice sites). **C**: The frequency of A+T across 5′ and 3′ splice sites are superimposed to emphasize the uniformity of exonic and intronic A+T content. **D**: The frequency of pyrimidines is separately plotted for the intronic region adjacent to 3′ splice sites in large (thin line) and small (thick line) introns. The stippled area emphasizes the greater pyrimidine content of long introns.

Table 2A. 5′ splice site sequences

Drosophila (all introns)

	−5	−4	−3	−2	−1	1	2	3	4	5	6	7	8
A	33	34	37	52	9	0	0	60	71	9	11	39	27
C	24	21	29	15	8	0	0	1	9	2	14	13	21
G	14	23	15	11	71	100	0	35	9	82	6	19	20
T	29	22	19	21	12	0	100	4	11	6	68	29	32
consensus:			M	A	G	G	T	R	A	G	T	W	

Total (all organisms from Senapathy *et al.*, dominated by mammals)

	−5	−4	−3	−2	−1	1	2	3	4	5	6	7	8
A			32	60	9	0	0	59	71	7	16		
C			37	13	5	0	0	3	9	6	16		
G			18	12	79	100	0	35	11	82	18		
T			13	15	7	0	100	3	9	6	50		
consensus:			M	A	G	G	T	R	A	G	T		

Drosophila (short introns)

	−5	−4	−3	−2	−1	1	2	3	4	5	6	7	8
A	36	28	33	53	7	0	0	61	69	7	10	36	26
C	22	21	29	14	7	0	1	2	11	3	14	14	16
G	15	29	17	12	74	100	0	34	10	86	5	14	24
T	27	21	21	21	13	0	99	4	9	4	71	36	34
consensus:				A	G	G	T	R	A	G	T	W	

Drosophila (long introns)

	−5	−4	−3	−2	−1	1	2	3	4	5	6	7	8
A	30	39	41	51	12	0	0	60	72	11	12	42	27
C	26	21	29	16	9	1	0	0	8	3	15	12	28
G	12	17	13	10	69	99	0	36	7	78	8	24	15
T	31	22	16	22	10	0	100	4	13	8	65	21	29
consensus:			M	A	G	G	T	R	A	G	T	A	

Table 2B. 3′ splice site sequences

Drosophila (all introns)

	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	1	2	3
A	21	21	22	20	19	19	24	19	10	11	28	5	100	0	33	17	18
C	21	23	16	24	24	37	28	36	28	20	23	68	0	0	15	21	32
G	8	10	9	9	10	6	11	6	5	4	23	0	0	100	34	19	25
T	49	45	53	47	47	39	37	40	57	64	25	27	0	0	18	43	25
	T	T	T	T	T	Y	Y	Y	T	T		C	A	G	R	T	

Total (all organisms, from Senapathy *et al.*, dominated by mammals)

	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	1	2	3
A	11	11	10	8	11	10	11	11	7	8	25	3	100	0	27		
C	29	33	30	30	32	34	37	38	39	36	26	75	0	0	14		
G	14	12	10	10	9	11	10	9	7	6	26	1	0	100	49		
T	46	44	50	52	48	45	42	43	47	51	23	21	0	0	10		
	T	Y	Y	Y	Y	Y	Y	Y	Y	Y		C	A	G	G		

Drosophila (short introns)

	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	1	2	3
A	25	28	24	23	16	20	23	19	7	13	23	4	100	0	36	19	21
C	17	22	11	19	22	32	32	42	32	16	27	64	0	0	16	17	35
G	5	9	9	10	7	7	10	6	5	3	19	0	0	100	30	19	23
T	53	40	55	48	55	41	35	34	57	68	31	33	0	0	19	46	21
	T	T	T	T	T	Y	Y	Y	Y	T		C	A	G	R	T	

Drosophila (long introns)

	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	1	2	3
A	16	15	19	15	22	17	26	18	13	8	33	6	99	0	30	14	13
C	27	24	20	32	27	43	24	29	23	26	19	73	1	0	14	26	30
G	12	10	9	7	13	4	11	6	6	6	29	0	0	100	38	19	28
T	45	50	51	46	38	36	39	47	57	60	19	20	0	0	18	41	30
	T	T	T	Y	Y	Y	T	Y	Y	T		C	A	G	R	T	

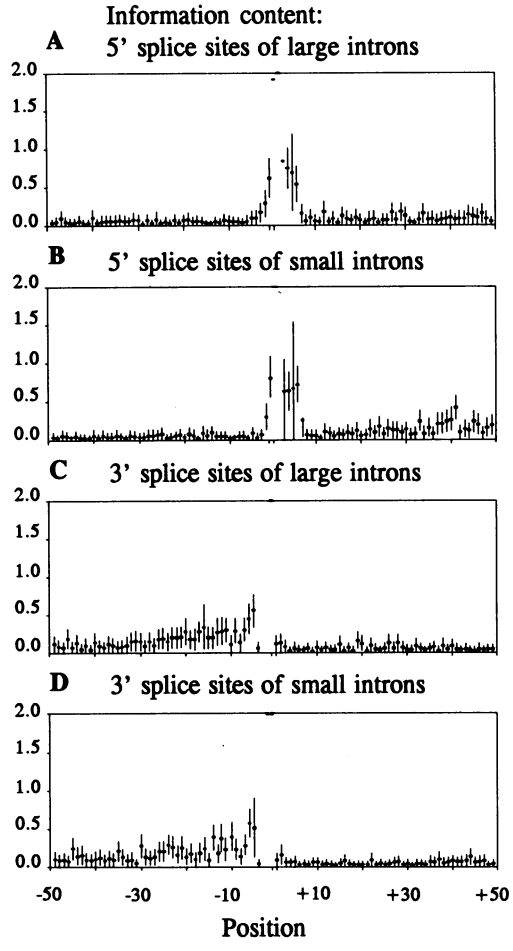

Figure 3. Information content (23, 40) at nucleotide positions between −50 and +50 surrounding the 5′ splice sites (A and B) and 3′ splice sites (C and D) of long (A and C) and short (B and D) Drosophila introns. Error bars indicate a statistical uncertainty of two standard deviations. Because we chose to use the exact method (see 51), calculation of error would have been computationally intensive for positions with high information content, and was not performed. Positions with higher information content generally have lower errors.

to 3′ splice sites of introns in both size classes is distinctive in that the frequency of A and G declines and the frequency of T increases as the 3′ splice site is approached. Overall frequency of A+T does not change, but T increases to approximately 50%. In both large and small introns, the frequency of C is not significantly higher in the −30 to −10 region than it is in the −50 to −30 region, but C is very common at positions −9, −8 and −7. The position with maximal C content is −9 in the case of large introns (41%) and −7 in the case of small introns (44%). Interestingly, this C-rich region is in precisely the location of a peak of pyrimidine-richness in yeast introns (52), where the pyrimidine stretch plays a less significant role than it does in vertebrates (53). However, this small peak of C-richness is not seen in all Drosophila introns, and many introns lack C's in this region.

The most significant difference we have observed between large and small introns is that large introns have greater pyrimidine content near the 3′ splice site (Figure 2D). This is particularly true in the −21 to −11 region, where large introns have a 10% higher pyrimidine content overall (70% vs. 60%). Only 27/107 (25%) of 3′ splice sites from small introns have a stretch of 8 consecutive pyrimidines, while 50/98 (51%) of large introns do. Similarly, 55/107 (51%) of 3′ splice sites from small introns have a stretch of 12 nucleotides with 10 or more pyrimidines, while 71/98 (72%) of large introns do. In this regard, it is large Drosophila introns that most resemble mammalian introns, which typically exceed 80% pyrimidines in this region (8). It is interesting to compare this difference between large and small introns to that observed in *C.elegans* (23). In each case, large introns were observed to carry additional signals. In the case of worms, larger introns were observed to have greater information content at the 5′ splice site. In the case of flies, we observe that larger introns have a stronger pyrimidine stretch.

A number of Drosophila introns lack strongly pyrimidine-rich regions near the 3′ splice site. For example, the second *white* intron has less than 50% pyrimidines (14/31), and no stretch of 12 consecutive nucleotides with more than seven pyrimidines between the 3′ splice site and the branchpoint (30). However, the generality of this observation is hard to assess from the database, because shorter stretches, although not more abundant than would be expected by chance, can usually be found and could play a functional role (for example, only seven of 204 3′ splice sites lack a stretch of 8 nucleotides with 6 or more pyrimidines in the −51 to −3 region). One possibility suggested by our data is that other aspects of base composition serve as a component of the splicing signal. For example, G-poorness contributes more to the information content of this region than does pyrimidine-richness, with the percentage of G residues being only 8.3% between −5 and −17 (see Figures 2 and 3). Conceivably, this G-poorness could reflect differences between Drosophila and mammals in the specificity of RNA binding by U2AF.

Branchpoint consensus

Because splice sites are readily identified by comparing genomic and cDNA sequences, the consensus sequences at the splice sites are easily identified and tabulated (see Table 2). However, definitive localization of a branchpoint requires analysis of splicing intermediates or excised introns, which are usually obtained only by means of in vitro splicing. Only five wild type Drosophila introns have been analyzed in this way (see Table 3), and considerable experimental work will be required to establish a statistically meaningful data set of confirmed branchpoint sequences. However, computational analyses can yield information about the branchpoint consensus. For example, a previous analysis using the consensus search program known as Makecons with a data set of 24 introns identified C T A A T as a putative Drosophila branchpoint consensus (Keller and Noon, 1984). We therefore applied a variety of computational methods to the problem of identifying signals within introns or within the flanking exons. Unlike this previous study, which involved refinement of an *a priori* consensus matrix, each of the methods reported here involved an unbiased search for a consensus.

Initially, we analyzed our data set with the RTIDE program (43, 44) and the CONSENSUS program (41, 42). The RTIDE method takes as input approximately aligned sequences and reports the most frequent 'word' in a window of variable size, allowing for mismatches between the words. In addition, the frequency of the most frequent word is an additional measure of information content. We examined data sets consisting of

Table 3. Branchpoint sequences

Mammalian examples (Actual numbers, from Nelson and Green, 1989, reference 16):									
						BP			
	A	3	10	0	8	10	29	1	2
	C	8	9	20	6	4	1	15	11
	G	5	7	3	0	13	0	7	2
	T	15	5	8	17	4	1	8	16
Consensus:		T	N	C	T	R	A	C	Y
Yeast sequence:		T	A	C	T	A	A	C	T
Examples from Drosophila:									Reference
ftz:		A	G	C	T	A	A	C	C 29
white:		T	C	T	T	A	A	T	A 30
Myosin HC exon 19:		T	T	T	T	A	A	T	C Bernstein*
Myosin HC exon 19:		A	A	C	T	A	A	T	T Bernstein*
Myosin HC exon 6:		T	C	C	T	A	A	T	G Bernstein*
Drosophila branchpoint matrix as determined by CONSENSUS (percentages):									
	A	42	39	0	24	73	88	16	9
	C	18	0	71	25	17	8	12	37
	G	8	21	27	2	9	0	0	19
	T	31	40	1	49	0	4	72	35
Consensus:			W	C	T	A	A	T	Y
Information:		0.05	0.33	1.55	0.14	0.85	1.13	0.47	0.27

*Diane Hodges and Sanford I. Bernstein, 1992, Mech. Dev., in press.

sequences 101 nucleotides long centered on splice sites from large and small introns (see Table 1). Word length was varied from 4 to 6, and window size from 8 to 15, allowing 0,1 or 2 mismatches, with consistent results. The distribution of numerical scores that resulted when a window of eight nucleotides was used to look for 5 letter words, allowing one mismatch, is shown in Figure 4. As expected, these distributions are similar to those resulting from information-content calculations (Figure 3). The most frequent 'word' at the 5′ splice site, GTAAG, agrees well with the consensus data above. At 3′ splice sites, scores are lower, and the highest scoring words have the form TNCAG. The intron region adjacent to 3′ splice sites (roughly −10 to −30) shows much higher scores than intron sequences in general. Analysis of the words that contribute these high scores is indicative of signals for the branchpoint and pyrimidine stretch. The highest scoring word for windows centered between −21 and −26 resembles the branchpoint consensus in each case (TAATT, TTAAT or CTAAT are observed). Windows centered between −8 and −15 have words consisting of entirely pyrimidines (TTTCT, TCTTT, TCCTT and CTTTC), and TACAG is the highest scoring word for each of the four eight-nucleotide windows that include the −5 to −1 region. In the region between the pyrimidine and branchpoint regions, words such as TTTAT, which can become either branchpoint (TTAAT) or pyrimidine stretch (TTTTT) with the variation of single nucleotide, score highest.

Unlike RTIDE, the program CONSENSUS looks for consensus patterns within sequences without regard to any pre-existing alignment. Rather, the matrix with maximal information content derived using one contribution from each of the sequences is determined. As expected, the 5′ splice site data sets yielded a matrix that was essentially identical to the 5′ splice site consensus. The 3′ splice site data sets yielded matrices that were A+T rich, pyrimdine rich, resembled the branchpoint, or had some combination of these features. In order to look for a branchpoint consensus matrix without interference from other sequence elements (such as the pyrimidine stretch or splice site),

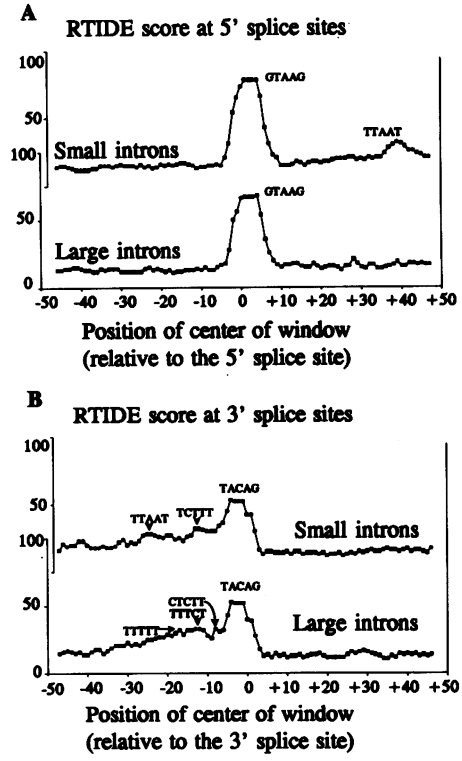

Figure 4. RTIDE analysis of Drosophila introns. Scores from the program RTIDE (43, 44) using 5 nucleotide words in an 8 nucleotide-long window are plotted vs. position, and the highest scoring 'word' corresponding to each peak is indicated.

a modified data set consisting of the 42-nucleotide region between −51 to −10 was used. The program was run with a window size of 8, using an order-independent algorithm, and saving a maximum of 1000 matrices per cycle. The matrix initially derived by the CONSENSUS program was used to rescan the data and

generate a new alignment. Information content was calculated as in Hertz *et al.* (42)using as an *a priori* base probability the actual values from the modified data set (A=0.318, C=0.181, G=0.133 and T=0.368). The rescanning procedure was performed six times before the alignment became self-generating; the improvement in information content was 0.0795 bits. This pattern includes one position with very little information, and can be thought of as a seven nucleotide matix beginning with position 2, roughly corresponding to the consensus WCTAATY (see Table 3). Numerous CONSENSUS runs were performed with varying parameters. The low frequency of T in the third position of the matrix (three nucleotides before the branchpoint) was not consistently observed. However, the low frequency of G in the fifth position of the matrix (immediately adjacent to the branchpoint) was a consistent feature of all derived branchpoint matrices. Because of its similarity to the branchpoint consensus sequence described in other organisms, it is extremely likely that this sequence is the Drosophila branchpoint consensus. Although similar to the mammalian branchpoint consensus, this Drosophila sequence differs in a number of ways, most notable among them being the lack of G in the position immediately preceding the branchpoint (see Table 3).

In order to investigate the branchpoint consensus further, the distribution (relative to splice sites) of a large number of tetranucleotides was examined. These distributions were generally in agreement with the consensus sequence determined as described above. Frequency data for CTAA and CTGA are presented in Figure 5. As expected from its similarity to the branchpoint consensus, CTAA occurs frequently in the region between −50 and −18 relative to 3′ splice sites. These distributions match those expected if CTAA were indeed a common tetranucleotide at Drosophila branchpoints, with larger introns showing a tendency for branchpoints slightly further upstream. The two occurrences of CTAA at −18 and three at −20 indicate use of a branchpoint A at −15 in two cases and −17 in three others. This suggests a minimum distance between the branchpoint and 3′ splice site of either 15 or 17 nucleotides, a number that is in reasonably good accord with results from mammalian systems, where a minimum distance of 18 nucleotides is indicated by the available data (16, 48).

The distribution of CTGA tetranucleotides relative to 3′ splice sites stands in sharp contrast to that of CTAA tetranucleotides. In fact, no significant enrichment of CTGA is observed in the branchpoint region. This is consistent with the consensus sequence derived above, and indicates a difference between Drosophila and mammals, where G is the predominant nucleotide in this position (reference (16, 48) and Table 3).

The distribution of CTAA relative to 5′ splice sites in small introns can also be used as an indication of the minimal distance between 5′ splice sites and branchpoints. There are two occurrences of CTAA at position 35 and seven at position 36, indicating possible use of a branchpoint A at position 38 in two cases and at position 39 in seven others. In fact, there are 27 occurrences of CTAA that would indicate branchpoint A's between position 39 and position 43 in this data set of 107 sequences, clearly indicating a very strong tendency for branchpoints to occur within a narrow range. Further support for this observation is found in the peak of information at this position in short introns (see Figure 3). The 5′ splice site to branchpoint distance implied by these results is considerably less than that in mammalian introns. For example, manipulation of the 5′ splice site to branchpoint distance of the 66 nucleotide SV40

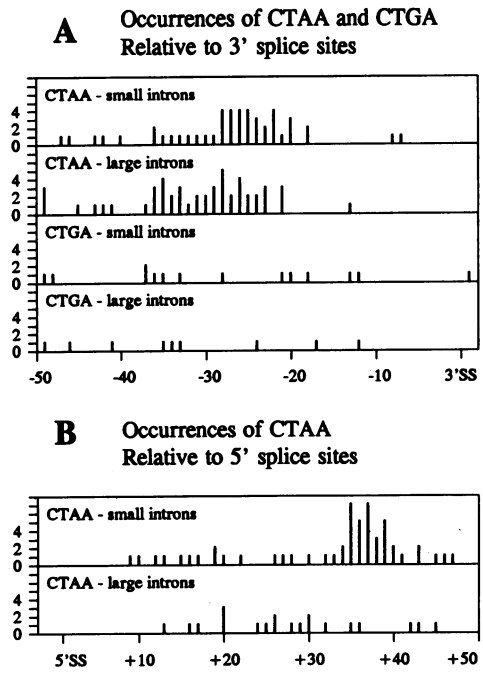

Figure 5. Occurrences of selected tetramers relative to splice sites. All occurrences of the indicated tetranucleotide in all introns within the indicated database are plotted relative to the splice site. See text.

small-t intron (48) indicated that the wild type distance in that case (48 nucleotides) is minimal; an intron with a distance of 46 nucleotides showed no splicing, while a distance of 53 nucleotides showed significantly increased splicing. In a study of the α-tropomyosin gene, Smith and Nadal-Ginard (35) found a 5′ splice site to branchpoint distance of 51 nucleotides too short, but 59 sufficient, for in vitro splicing. In addition, a distance of 49 nucleotides between the 5′ splice site and branchpoint was found too short to allow U4-U5-U6 binding to an adenovirus E1A pre-mRNA in HeLa cell nuclear extracts (54).

Other tetranucleotides from the branchpoint consensus (TAAT, TAAC, AACT and AATC) are distributed similarly to CTAA (data not shown). Thus, the distribution of branchpoint-related sequences indicates that the smaller size of Drosophila introns is associated with a decreased distance between 5′ splice sites and branchpoints rather than a decreased distance between branchpoints and 3′ splice sites.

Mechanistic implications

Several of the results obtained here make it interesting to speculate that small and large introns may differ with respect to the mechanism by which branchpoints are recognized.

First, our observations on intron size and the distribution of branchpoint-like sequences in Drosphila argue that mammals and fruit flies differ considerably with regard to the range of acceptable distances between the 5′ splice site and the branchpoint. There is experimental support for this conclusion. For example, the second Drosophila *white* intron (74 nucleotides) has a 5′ splice site to branchpoint distance of 43 nucleotides (less than the mammalian minimum) and is efficiently spliced in nuclear extracts from Drosophila, but not human, cells (30). It is likely that this minimum distance reflects the spatial requirements for spliceosome assembly in mammals, and indeed,

Himmelspach *et al.* (54) have observed defective spliceosome assembly on experimentally shortened introns. However, because the size of U RNAs (2, 5) and snRNP proteins (32, 55) is comparable in Drosophila and mammals, it seems unlikely to us that the shorter minimum distance could be explained by flies simply having smaller snRNPs. Second, it is striking that the separation between the 5′ splice site and branchpoint in many short Drosophila introns is just over than the minimum distance. These observations indicate that branchpoint recognition in small introns may be facilitated by direct interaction between a factor at the 5′ splice site (the U1 snRNP, a larger complex including the U1 snRNP, or some other factor) and a factor at the branchpoint (the U2 snRNP, a larger complex including the U2 snRNP, or some other factor). Normally, association between the U2 snRNP and the branchpoint is promoted by the binding of U2AF to the pyrimidine stretch. Thus, the reduced pyrimidine content of small Drosophila introns relative to large Drosophila introns (which are more like mammalian introns in their pyrimidine content) also supports the hypothesis that branchpoints in small introns might be recognized by a mechanism involving the 5′ splice site. Such a mechanism of branchpoint recognition may be corroborated by the observation that, in yeast, the branchpoint as well as the 5′ splice site has been shown to have a role in the formation of early complexes including U1 snRNP (56). We (SM, unpublished results) have begun to explore these ideas experimentally.

ACKNOWLEDGEMENTS

We are grateful to G.Hartzell for systems support, to R. Farber for providing us with an updated version of the ExtractGenBank software, to Ted Dunning at the Computer Research Laboratory of New Mexico State University for advice on statistical uncertainties, and to Nicole Kawachi for assistance with the tetranucleotide analysis. S.M. was supported by NIH grant GM 37991, by a NSF Presidential Young Investigator award, and by Basil O'Conner Starter Scholar Research award 5-630 from the March of Dimes Birth Defects Foundation. G.S. and G.H. were supported by NIH grants GM 28755 and HG 00249, O.W. and C.F. were supported by U.S. Department of Energy Genome Program Grant 89ER60865, and C.B was supported by NIH grant GM 37812. This work was in part done under the auspices of the Aspen Center for Physics under a grant from the NSF.

REFERENCES

1. Green, M.R. (1986) *Annu. Rev. Genet.*, **20**, 671–708.
2. Guthrie, C. and Patterson (1988) *Ann. Rev. Genetics*, **22**, 387–419.
3. Smith, C.W.J., Patton, J.G. and Nadal-Ginard, B. (1989) *Annu. Rev. Genet.*, **23**, 527–577.
4. Green, M.R. (1991) *Annu. Rev. Cell Biol.*, **7**, 559–600.
5. Mount, S.M. (1982) *Nucleic Acids Res.*, **10**, 459–472.
6. Shapiro, M.B. and Senapathy, P. (1986) *Nucleic Acids Res.*, **15**, 7155–7174.
7. Jacob, M. and Gallinaro, H. (1989) *Nucleic Acids Res.*, **17**, 2159–2180.
8. Senapathy, P., Shapiro, M.B. and Harris, N.L. (1990) *Methods Enzymol.*, **183**, 252–278.
9. Zhuang, Y. and Weiner, A.M. (1986) *Cell*, **46**, 827–835.
10. Zapp, M.L. and Berget, S.M. (1989) *Nucleic Acids Res.*, **17**, 2655–2674.
11. Bruzik, J.P. and Steitz, J.A. (1990) *Cell*, 889–899.
12. Seraphin, B. and Rosbash, M. (1990) *Cell*, **63**, 619–629.
13. Newman and Norman (1991) *Cell*, **65**, 115–123.
14. Black, D.L., Chabot, B. and Steitz, J.A. (1985) *Cell*, **42**, 737–750.
15. Zhuang, Goldstein and Weiner (1989)
16. Nelson, K.K. and Green, M.R. (1989) *Genes Dev.*, **3**, 1562–1571.
17. Barabino, S.L., Blencowe, B.J., Ryder, U., Sproat, B.S. and Lamond, A.I. (1990) *Cell*, **63**, 293–302.
18. Rosbash, M. and Seraphin, B. (1991) *Trends Biochem. Sci.*, **16**, 187–190.
19. Ruskin, B., Zamore, P.D. and Green, M.R. (1988) *Cell*, **52**, 207–219.
20. Blumenthal, T. and Thomas, J. (1988) *Trends Genet.*, **4**, 305–308.
21. Kay, R.J., Russnak, R.H., Jones, D., Mathias, C. and Candido, P. (1987) *Nucleic Acids Res.*, **9**, 3723–3741.
22. Ogg, S.C., Anderson, P. and Wickens, M.P. (1990) *Nucleic Acids Res.*, **18**, 143–149.
23. Fields, C. (1990) *Nucleic Acids Res.*, **18**, 1509–1512.
24. Weibauer, K., Herrero, J.-J. and Filopowicz, W. (1988) *Mol. Cell. Biol.*, **8**, 2042–2051.
25. Csank, C., Taylor, F.M. and Martindale, D.W. (1990) *Nucleic Acids Res.*, **18**, 5133–5141.
26. Goodall, G.J. and Filopowicz, W. (1989) *Cell*, **58**, 473–483.
27. Goodall, G.J. and Filopowicz, W. (1991) *EMBO J.*, **10**, 2635–2644.
28. Keller, E.B. and Noon, W.A. (1984) *Proc. Natl. Acad. Sci. USA*, **81**, 7417–7420.
29. Rio, D.C. (1988) *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 2904–2909.
30. Guo, M., Lo, P. and Mount, S. (1992) *submitted*,
31. Mount, S.M. and Steitz, J.A. (1981) *Nucleic Acids Res.*, **9**, 6351–6368.
32. Paterson, T., Beggs, J., Finnegan, D. and Luhrmann, R. (1991) *Nucleic Acids Res.*, **19**, 5877–5882.
33. Hawkins, J.D. (1988) *Nucleic Acids Res.*, **16**, 9893–9905.
34. Wieringa, B., Hofer, E. and Weissman, C. (1984) *Cell*, **37**, 915–925.
35. Smith, C.W.J. and Nadal-Ginard, B. (1989) *Cell*, **56**, 749–758.
36. Falkenthal, S., Parker, V.P. and Davidson, N. (1985) *Proc. Natl. Acad. Sci. U.S.A.*, **82**, 449–453.
37. Bernstein, S.I., Hansen, C.J., Becker, K.D., Wassenberg, D.R., Roche, E.S., Donady, J.J. and Emerson, C.P. (1986) *Mol. Cell. Biol.*, **6**, 2511–2519.
38. Eveleth, D.D., Gietz, R.D., Spencer, C.A., Nargang, F.E., Hodgetts, R.B. and Marsh, J.L. (1986) *EMBO J.*, **5**, 2663–2672.
39. Siebel, C.W. and Rio, D.C. (1989) *Science*, **248**, 1200–1208.
40. Schneider, T.D., Stormo, G.D., Gold, L. and Ehrenfeucht, A. (1986) *J. Mol. Biol.*, **188**, 415–431.
41. Stormo, G.D. (1990) *Methods Enzymol.*, **183**, 211–221.
42. Hertz, G.Z., III, G.W.H. and Stormo, G.D. (1990) *Comput. Appl. Biosci.*, **6**, 81–92.
43. Galas, D.J., Waterman, M.S. and Eggert, M. (1985) *J. Mol. Biol.*, **186**, 117–128.
44. Waterman, M.S. and Jones, R. (1990) *Methods Enzymol.*, **183**, 221–237.
45. Burks, C., Cinkosky, M.J., Gilna, P., Hayden, J.E.-D., Abe, Y., Atencio, E.J., Barnhouse, S., Benton, D., Buenafe, C.A., Cumella, K.E., Davison, D.B., Emmert, D.B., Faulkner, M.J., Fickett, J.W., Fischer, W.M., Good, M., Horne, D.A., Houghton, F.K., Kelkar, P.M., T.A., K., Kelly, M., King, M.A., Langan, B.J., Lauer, J.T., Lopez, N., Lynch, C., Lynch, J., Marchi, J.B., Marr, T.G., Martinez, F.A., McLeod, M.J., Medvick, P.A., Mishra, S.K., Moore, J., Munk, C.A., Mondragon, S.M., Nasseri, K.K., Nelson, D., Nelson, W., Nguyen, T., Reiss, G., Rice, J., Ryals, J., Salazar, M.D., Stelts, S.R., Trujillo, B.L., Tomlinson, L.J., Weiner, M.G., Welch, F.J., Wiig, S.E., Yudin, K. and Zins, L.B. (1990) *Meth. Enzymol.*, **183**, 3–22.
46. Stormo, G.D. and Hartzell, G. W. III (1989) *Proc. Natl. Acad. Sci. U.S.A.*, **86**, 1183–1187.
47. Stormo, G.D. (1988) *Annu. Rev. Biophys. Biophys. Chem.*, **17**, 241–263.
48. Fu, X.-Y., Colgan, J. and Manley, J.L. (1988) *Mol. Cell. Biol.*, **8**, 3582–3590.
49. Jackson, I.J. (1991) *Nucleic Acids Res.*, **19**, 3795–3798.
50. Hanley and Schuler, M. (1988) *Nucleic Acids Res.*, **16**, 7159–7176.
51. White, O., Soderland, C., Shanmugam, P. and Fields, C. (1992) *Plant Molec. Biol.*, **in press**,
52. Parker, R. and Patterson, B. (1987) In M. Inouye and Dudock (ed.), Molecular Biology of RNA: New Perspectives, Academic Press, New York. pp. 133–149.
53. Patterson, B. and Guthrie, C. (1991) *Cell*, **64**, 181–187.
54. Himmelspach, M., Gattoni, R., Gerst, C., Chebli, K. and Stevenin, J. (1991) *Mol. Cell. Biol.*, **11**, 1258–1269.
55. Mancebo, R., Lo, P.C.H. and Mount, S.M. (1990) *Mol. Cell. Biol.*, **10**, 2492–2502.
56. Séraphin, B. and Rosbash, M. (1991) *EMBO J.*, **10**, 1209–1216.