Volume 16 Number 14 1988 Nucleic Acids Research

Plant intron sequences: evidence for distinct groups of introns

Brian A.Hanley and Mary A.Schuler*

Department of Plant Biology, University of Illinois, Urbana, IL 61801, USA

Received December 7, 1987; Revised and Accepted June 23, 1988

ABSTRACT

In vivo and *in vitro* RNA splicing experiments have demonstrated that the intron splicing machineries are not interchangeable in all organisms. These differences have prevented the efficient *in vivo* expression of monocot genes containing introns in dicot plants and the *in vitro* excision of some plant introns in HeLa cell *in vitro* splicing extracts. We have analyzed plant introns for sequence differences which potentially account for the functional splicing differences. Three classes of plant introns can be differentiated by the purine or pyrimidine-richness of sequences upstream from the 3' splice site. The frequency of these three types of introns in monocots and dicots varies significantly. The degree of variability in the 5' and 3' intron boundaries is evaluated for each of these classes in monocots and dicots. The 5' splice site consensus sequences developed for the monocot and dicot introns differ in their ability to base pair with conserved nucleotides present at the 5' end of many U1 snRNAs.

INTRODUCTION

In the past ten years since the discovery of intervening sequences (introns) in the pre-mRNA transcripts of eukaryotes, a significant research effort with HeLa and yeast cells (*Saccharomyces cerevisiae*) has led to some understanding of the molecular details of RNA splicing. Analysis of the biochemistry of RNA splicing has been greatly facilitated by the development of *in vitro* RNA splicing systems for HeLa and yeast cells (1-5). Through this biochemical effort and the analysis of the intron nucleotides (6-11), three sequence elements within introns have been demonstrated to be essential for splicing. These have been designated the 5' splice site, the internal splice signal, and the 3' splice site. Using *in vitro* RNA splicing systems, these conserved sequence elements have been shown to interact with trans-acting nuclear factors which mediate splicing (12).

In HeLa and yeast cells, intron splicing proceeds by a two-step mechanism: the first step is cleavage at the invariant GT dinucleotide at

the 5' splice site and formation of an intramolecular 2'-5' phosphodiester branch between the G residue at the 5' splice site and an A residue in the branch site (13). The second step entails cleavage at the AG dinucleotide at the 3' splice site, ligation of the two exons, and release of the intron as a lariat structure. Both steps require the formation of an RNA:protein complex, termed the spliceosome, which mediates the splicing events. The formation of the spliceosome requires ATP, the three conserved intron elements, and nuclear factors including the U1, U2, U5 and U4/U6 snRNPs.

Dramatic increases in the number of available plant gene sequences have enabled researchers to suggest that the conserved plant intron border elements are similar to those found in animals and yeast (14-16). In spite of these sequence similarities, a number of experiments have demonstrated that the intron splicing machineries are not interchangeable between organisms. This is particularly evident in heterologous *in vitro* splicing experiments which have demonstrated that mammalian and *Drosophila* introns are not excised by yeast cell extracts (17, 18) even though yeast introns can be spliced in mammalian nuclear extracts (19). *In vitro* experiments have also demonstrated that, although a few plant introns can be excised in HeLa cell nuclear extracts (16, 20, 21), some plant introns are not processed accurately in yeast or HeLa cell extracts (21, 22). In reciprocal experiments, four introns from the human growth hormone gene and two introns from the human α-globin gene were not excised *in vivo* in transgenic tobacco plants (21, 23).

The individual selectivities of eukaryotic RNA processing systems extend to the monocotyledonous and dicotyledonous groups of plants and have prevented the efficient *in vivo* expression of monocot genes containing introns in dicot plants. Introduction of monocot genes encoding the small subunit of ribulose 1,5-bisphosphate carboxylase of wheat or a segment of the alcohol dehydrogenase gene of maize into transgenic tobacco plants (dicots) results in the accumulation of intron-containing precursor RNAs which constitute 50% of the wheat transcript and 70% of the maize transcript produced *in vivo* (24). The maize alcohol dehydrogenase gene, which contains nine introns (25), is also inefficiently expressed in transgenic tobacco plants and in a transient tobacco protoplast assay (26). In contrast, monocot genes lacking introns, such as the wheat chlorophyll a/b binding protein gene or the maize zein gene, are efficiently expressed after transfer into dicots (27-29). Because the nucleic acid sequence within plant pre-mRNAs potentially accounts for some of these functional

splicing differences, we have analyzed the published plant intron sequences to determine if the intron splicing efficiencies can be accounted for by differences in conserved sequence elements at the 5' and 3' intron boundaries.

The first complete compilation of plant intron sequences (15) evaluated 177 monocot and dicot introns that were then available. This exhaustive study demonstrated that, in general, the plant intron borders are similar to those found in mammalian introns, but that subtle differences occur at particular positions. It was noted that in one position (-4 from the 3' splice site), plant introns preferentially contain a guanosine nucleotide whereas mammalian introns show no preference for any nucleotide (15). This study also indicated that purines were more abundant in the -5 to -15 region preceding the 3' splice site of plant introns than in mammalian introns which typically contain at least 70% pyrimidines (7).

Our closer inspection of these plant intron sequences reveals that the sequences preceding the 3' splice site in some plant introns are distinctly pyrimidine-rich whereas others are purine-rich. In this communication, we have examined all the available plant intron sequences and have subdivided them into three groups depending on the purine or pyrimidine-richness of the sequences at the 3' splice site and the monocot or dicot origin of the gene. The degree of variability in the 5' and 3' intron boundaries is evaluated for each of these groups and then compared with the results of heterologous splicing experiments. Our analysis suggests that primary sequence differences in plant introns potentially account for the differential splicing of monocot and dicot introns in transgenic plants and in mammalian splicing extracts.

MATERIALS AND METHODS

The 176 intron sequences used in this comparison have all been previously published. The monocot intron comparison includes maize alcohol dehydrogenase (25), glutathione-S-transferase (30), sucrose synthetase (31), waxy (32), shrunken (33), actin (34), triosephosphate isomerase (35) and alfalfa glutamine synthase (36), chalcone synthase (37) and the wheat ribulose bisphosphate carboxylase (38). The dicot comparison includes soybean actin (34, 39), leghemoglobin (Lba, Lbc1, Lbc2, Lbc3, uLb) (40-42), nodulin 24 (43), nodulin 35 (44), conglycinin (45), glycinin (46) and ribulose bisphosphate carboxylase (47); carrot extensin (48); french bean phaseolin (14) and leghemoglobin (49); pea legumin (A, D) (50, 51) and

Dicot 5' splice sites

Position	-3	-2	-1 ↓	+1	+2	+3	+4	+5	+6	+7	+8	+9
A	27	55	12	0	0	73	60	16	32	41	36	37
G	23	7	75	104	0	10	8	56	11	8	3	5
C	31	8	12	0	1	6	17	15	7	17	17	20
T	23	34	5	0	103	15	19	17	54	38	48	42
%A	26.0	52.9	11.5	0	0	70.2	57.7	15.4	30.8	39.4	34.6	35.6
%G	22.1	6.7	72.2	100	0	9.6	7.7	53.9	10.6	7.7	2.9	4.8
%C	29.8	7.7	11.5	0	1.0	5.8	16.3	14.4	6.7	16.3	16.4	19.2
%T	22.1	32.7	4.8	0	99.0	14.4	18.3	16.3	51.9	36.6	46.1	40.4
%Pur	48.1	59.6	83.7	100	0	79.8	65.4	69.3	41.4	47.1	37.5	40.4
%Pyr	51.9	40.4	16.3	0	100	20.2	34.6	30.7	58.6	52.9	62.5	59.6
CONSENSUS	N	A T	G	G	T	A	A	G	T A	A T	T A	T A
		U	C	C	A	U	U	C	A	U	A-CAP	U1snRNA

Dicot 3' splice sites

	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1 ↓
A	26	17	13	35	26	23	26	34	24	29	13	25	5	97	0
G	9	9	12	15	9	16	17	17	22	13	10	44	1	0	97
C	16	11	15	4	10	6	10	11	10	9	3	11	57	0	0
T	46	60	57	43	51	52	44	34	41	46	71	17	34	0	0
%A	26.8	17.5	13.4	36.1	27.1	23.7	26.8	35.4	24.7	29.9	13.4	25.8	5.2	100	0
%G	9.3	9.3	12.4	15.5	9.4	16.5	17.6	17.7	22.7	13.4	10.3	45.4	1.0	0	100
%C	16.5	11.3	15.4	4.1	10.4	6.2	10.3	11.5	10.3	9.3	3.1	11.3	58.8	0	0
%T	47.4	61.9	58.8	44.3	53.1	53.6	45.3	35.4	42.3	47.4	73.2	17.5	35.0	0	0
%Pu	36.1	26.8	25.8	51.6	36.5	40.2	44.4	53.1	47.4	43.3	23.7	71.2	6.2	100	100
%Pyr	63.9	73.2	74.2	48.4	63.5	59.8	55.6	46.9	52.6	56.7	76.3	28.8	93.8	0	0
	T A	T A	T A	T N	T A	T A	T A	A T	T A	T A	T	G A	C T	A	G

Monocot 5' splice sites

Position	-3	-2	-1 ↓	+1	+2	+3	+4	+5	+6	+7	+8	+9
A	26	44	8	0	0	53	33	10	14	18	11	11
G	18	9	57	75	0	9	5	51	8	12	11	11
C	24	10	6	0	0	7	14	7	20	10	20	16
T	7	12	4	0	75	6	23	7	33	35	33	22
%A	34.7	58.7	10.7	0	0	70.7	44.0	13.3	18.7	24.0	14.7	18.3
%G	24.0	12.0	76.0	100	0	12.0	6.6	68.1	10.7	16.0	14.7	18.3
%C	32.0	13.3	8.0	0	0	9.3	18.7	9.3	26.6	13.4	26.6	26.7
%T	9.3	16.0	5.3	0	100	8.0	30.7	9.3	44.0	46.6	44.0	36.7
%Pu	58.7	70.7	86.7	100	0	82.7	50.6	81.4	29.4	40.0	29.4	36.6
%Pyr	41.3	29.3	13.3	0	100	17.3	49.4	18.6	70.6	60.0	70.6	63.4
CONSENSUS	A C G	A	G	G	T	A	A T	G	T C	T A	T C	T N
		U	C	C	A	U	U	C	A	U	A-CAP	U1snRNA

Monocot 3' splice sites

	-15	-14	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1 ↓
A	6	10	9	17	9	18	10	11	14	16	10	12	3	75	0
G	13	12	17	10	14	11	14	19	16	16	7	33	1	0	75
C	23	14	15	16	16	15	16	14	14	13	10	14	56	0	0
T	33	39	34	32	36	31	35	31	31	30	48	16	15	0	0
%A	8.0	13.3	12.0	22.7	12.0	24.0	13.3	14.7	18.7	21.3	13.3	16.0	4.0	100	0
%G	17.3	16.0	22.7	13.3	18.7	14.7	18.7	25.3	21.3	21.3	9.3	44.0	1.3	0	100
%C	30.7	18.7	20.0	21.3	21.3	20.0	21.3	18.7	18.7	17.3	13.3	18.7	74.7	0	0
%T	44.0	52.0	45.3	42.7	48.0	41.3	46.7	41.3	41.3	40.1	64.1	21.3	20.0	0	0
%Pu	25.3	29.3	34.7	36.0	30.7	38.7	32.0	40.0	40.0	42.6	22.6	60.0	5.3	100	100
%Pyr	74.7	70.7	65.3	64.0	69.3	61.3	68.0	60.0	60.0	57.4	77.4	40.0	94.7	0	0
	T C	T C	T G	T A	T C	T A	T C	T G	T N	T N	T	G N	C	A	G

ribulose bisphosphate carboxylases (E9, 3A, 3C) (52, 53); broadbean legumin (54); potato patatin (55-57); tobacco ATP synthase (atp2-1) (58) and ribulose bisphosphate carboxylase (59); petunia ribulose bisphosphate carboxylase (60). The sequences upstream from the 3' splice sites were evaluated by determining the purine frequency between positions -3 and -20 and evaluating the purine distribution as outlined in Table I.

RESULTS

In several reported monocot gene transfer experiments (24), monocot genes were not efficiently expressed in transgenic dicots even though the GT-AG dinucleotide border elements in the monocot introns were identical to those found in dicot introns. To find sequences which potentially account for these in vivo expression problems, we have subdivided plant introns into different groups and intercompared their sequences. Comparison of the monocot and dicot intron sequences indicates that differences exist in the conserved sequence elements at the 5' and 3' splice sites. Although the entire pool of plant introns (15) suggests that the sequence of the plant 5' splice site consensus is similar to the highly conserved mammalian consensus sequence (CAG:GTAAGT; 7), a high degree of variability exists within the plant sequences (Fig. 1). These differences occur within the region (-3 to +7) known to interact with U1 snRNA in mammalian cells (61-64) and snR19 in yeast cells (65). The variability in this region is especially evident when the monocot and dicot 5' splice consensus sequences are compared (Fig. 1). Significant differences between monocots and dicots are observed at positions +4 through +9. At three of these positions (+4, +5, +6), the frequency of purines varies in monocot and dicot introns by more than 12%. These ratios fluctuate in part because the prevalent nucleotides used at some of these positions are not uniformly purines or pyrimidines. For example, at position +4, adenosine and thymidine are used at a high frequency in monocot introns whereas adenosine is preferentially used in dicot introns. At position +6, monocot introns contain high proportions of cytidine and thymidine and dicot introns primarily contain thymidine and adenosine. Although guanosine is

Figure 1. Monocot and dicot splice sites. Nucleotide frequencies were evaluated between positions -3 and +9 at the 5' splice sites and positions -15 and -1 at the 3' splice site. The consensus sequences derived from 104 dicot introns and 97 monocot introns are shown in bold letters. The ten conserved nucleotides found in mammalian, Drosophila, yeast, french bean and soybean U1 snRNAs (68-70) are shown below the 5' splice site consensus.

preferentially found at position +5 in monocot introns (68%), it occurs at a lower frequency (54%) in dicot introns. The most prevalent nucleotides at positions +8 and +9 are pyrimidines in monocot introns but thymidine and adenosine in dicot introns. Thus, at several internal positions, monocot and dicot introns differ in the second most prevalent nucleotide used. These differences contrast dramatically with the sequence conservation seen in intron positions +1, +2 and +3. In addition to frequency differences in the most prevalent nucleotides, particular nucleotides found prominently in monocot introns are nearly absent from dicot introns at positions +6 through +9. For example, cytidine occurs at position +6 in 27% of the monocot introns but only 7% of the dicot introns. Guanosine occurs with higher frequency (14-18%) at positions +7, +8 and +9 in monocot introns but very infrequently (3-8%) in dicot introns.

Significant differences also occur in the region upstream from the 3' splice sites of monocot and dicot introns. All plant introns contain the ubiquitous AG dinucleotide found in other 3' splice sites (7). Immediately upstream from this highly conserved dinucleotide, plant introns generally show little preference for pyrimidine or purine bases (Fig. 2). In this region, monocot introns have a much higher frequency of pyrimidine nucleotides than dicot introns (Fig. 1). Although both groups of introns contain thymidine residues as prominent nucleotides in positions -3 to -15, monocot introns have a level of cytidine (13-30%) which is higher than in dicot introns (3-16%) and very similar to the level present in mammalian introns (19-36%) (7).

Inspection of the -3 to -20 region upstream from the 3' splice indicates that many plant introns entirely lack the polypyrimidine tract which is prevalent in mammalian introns (7) and known to interact with a mammalian U5 snRNP factor (66, 67). In this region, the mammalian pyrimidine-rich sequence preceding the 3' terminal AG is frequently replaced by a preponderance of purines. Because the interaction between the mammalian snRNP and the pyrimidine-rich tract involves nucleotides between positions -2 and -19 (66), we have subdivided plant introns into three groups based on the presence of a purine-rich or pyrimidine-rich region between positions -3 and -20. Each of the classes is defined by the content and arrangement of purine/pyrimidine bases near the 3' splice site as shown in Table I. By this classification, pyrimidine-rich introns (Y, class I) contain less than five purine residues between positions -3 and -20. If more purines are present, pyrimidine-rich introns have five or

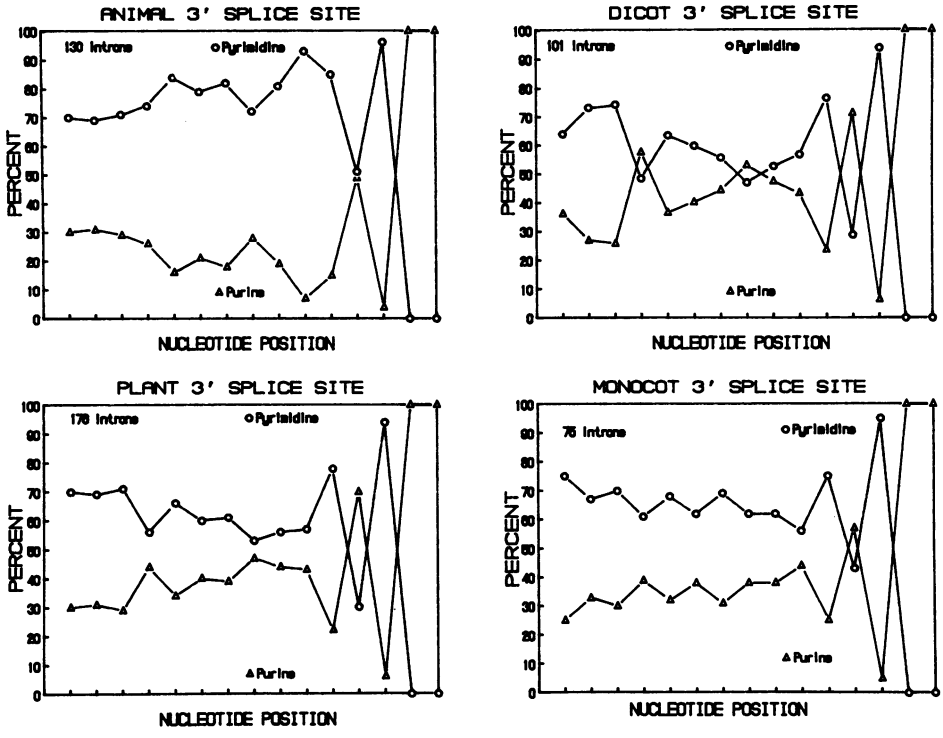

Figure 2. Pyrimidine frequencies at 3' splice site. The purine (△) and pyrimidine (○) frequencies upstream from the 3' splice site are recorded for 130 animal introns (7) and 176 plant introns (101 dicot, 75 monocot).

more consecutive pyrimidines with no distinct purine string. Purine-rich introns (U, class III) contain at least nine purine residues between -3 and -20 or, if only eight purines are present, at least three consecutive purines occur in the region upstream from the 3' splice site. The most extreme example of this intron class contains nine consecutive purines (Table I). We have established a third category containing introns that have unusually high purine frequencies but which lack consecutive strings of purines or pyrimidines. This latter group has been designated the "mixed" class (M, class II) due to the scattered pattern of purine and pyrimidine bases. In the few cases in which a polypyrimidine tract was situated near a purine tract, the introns were differentiated, as outlined in Table I, by the length of the purine tract and the overall purine content of this region. Introns have been placed in this category if three consecutive purines are located within a stretch of 6-7/18 purine residues

Table 1.

A.

No. purines	Arrangement	Intron Category
1-5/18	---	Pyrimidine-rich (Y)
6/18	5 consecutive pyrimidines no strings of 3 purines	Pyrimidine-rich (Y)
6-7/18	3 consecutive purines	Mixed (M)
6/18	4 consecutive purines	Purine-rich (U)
6-8/18	no string of consecutive purines or pyrimidines	Mixed (M)
8/18	3 or more consecutive purines	Purine-rich (U)
9-18/18	---	Purine-rich (U)

B.

	$^{-20}$ $^{-1}$	Purines	
PAT 31	TACTTTTCTTTTCGAGTCAG	4/18	
SAC 1	ACCGCAACGTGTCCTTTCAG	6/18	Pyrimidine-rich (Y)
MWAX 12	TCGTCGTCCTCTCTTCCCAG	2/18	
AGS 11	GTGGCTTGTGTTATTTGAAG	8/18	
CONGL 4	TTCTTTGTTCACAAATATAG	6/18	Mixed (M)
PAT 42	GGTTACATTATATTATGCAG	8/18	
PAT 23	TTTTTTAAAAAAAAGTGCAG	10/18	
SAC 22	ATCAATTCCTTTTAAAACAG	7/18	Purine-rich (U)
MWAX 2	TTGTTCGGGCATGCATGCAG	8/18	

Definition of intron classes. (A) Intron classes are first defined by determining the number of purine nucleotides between -3 and -20 upstream from the 3' splice site (column 1). If 6-8 purine nucleotides occur in this region, the arrangement of purine and pyrimidine residues within this region are evaluated as outlined in column 2. (B) Examples of the three intron classes.

or if purine-rich sequences containing 6/18, 7/18 or 8/18 purines lack consecutive purine or pyrimidine strings. The purine or pyrimidine-richness of the introns within these classes is especially evident when diagrammed as shown in Fig. 4. The purine-rich introns contain predominant purines at -4 and -6 through -10 and a very low

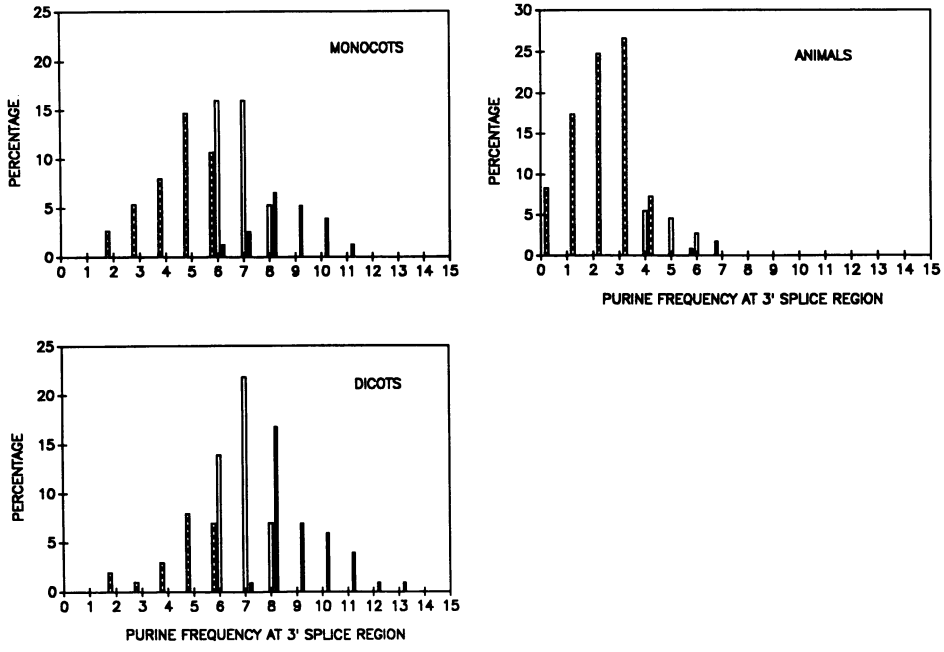

Figure 3. Distribution of monocot and dicot introns. The frequency of monocot and dicot intron classes are recorded relative to the purine frequency in the 18 nt. between -3 and -20. The frequencies of the intron classes are recorded for the 109 vertebrate and viral introns listed by Mount using the 13 nt. between -3 and -15 (7). Crosshatched boxes (pyrimidine-rich), open boxes (mixed), filled boxes (purine-rich).

pyrimidine content. The pyrimidine-rich introns contain predominant pyrimidines at -3 and -5 through -15. Of the 176 published plant intron sequences that we have included in this comparison, 52 (30%) are classified as pyrimidine-rich, 53 (30%) are purine-rich and 71 (40%) are mixed. For comparison, we have classified 109 of the vertebrate and viral introns described by Mount (7). In this group, 93/109 (85%) are pyrimidine-rich, 13/109 (12%) are mixed, and 3/109 (3%) are purine-rich. In chi-square analysis, $p = 0.000$ for this data. Thus the abundance of introns in the mixed and purine-rich classes is significant, in view of the virtual absence of these classes in mammalian genes.

The frequencies of the purine and pyrimidine-rich classes of introns vary significantly in monocots and dicots (chi-square $p = 0.0075$). 16/75 (21%) of the monocot introns are purine-rich, 31/75 (41%) are pyrimidine-rich and 28/75 (37%) introns are mixed. 37/101 (37%) of the

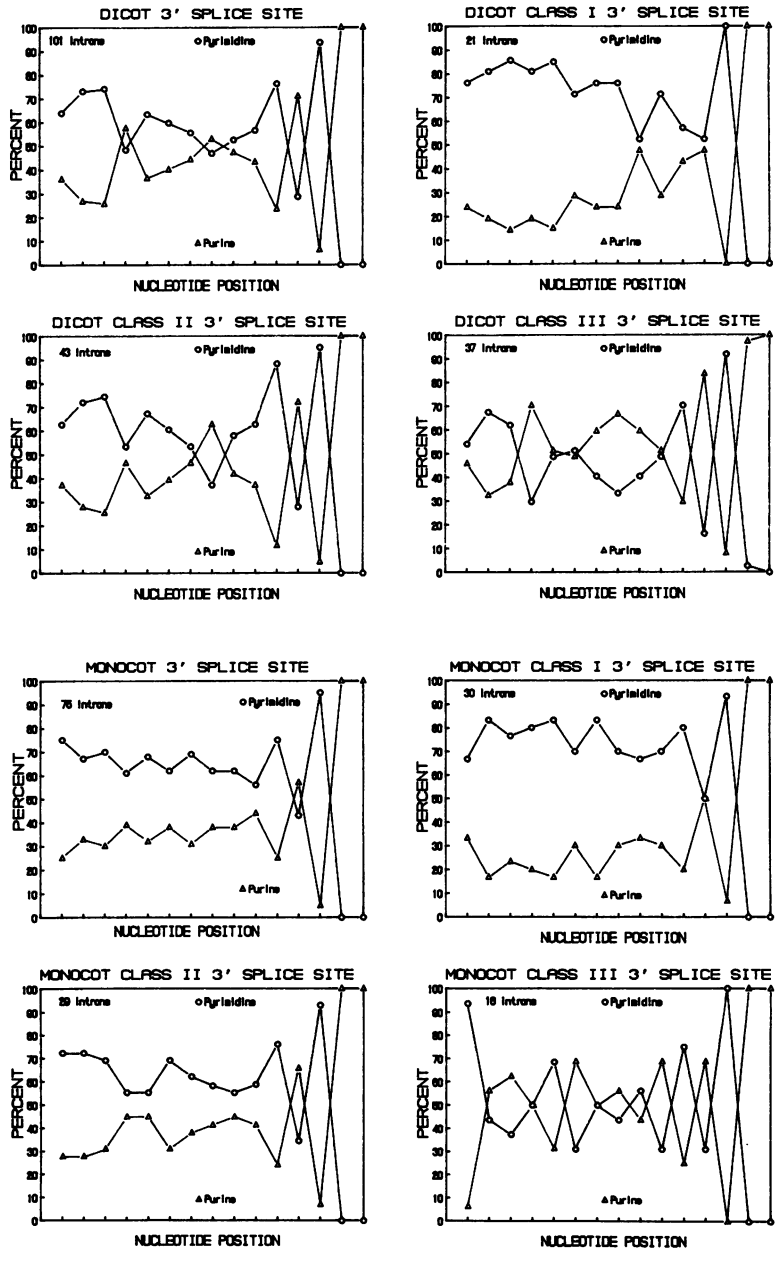

Figure 4. Nucleotide frequencies in the mixed, purine and pyrimidine-rich intron classes. The purine (Δ) and pyrimidine (O) frequencies at the 3' splice site are recorded separately for the pyrimidine-rich (Y), the mixed (M) and purine-rich (U) introns in dicots and monocots. The number of introns used in each evaluation are shown in the upper left of each panel.

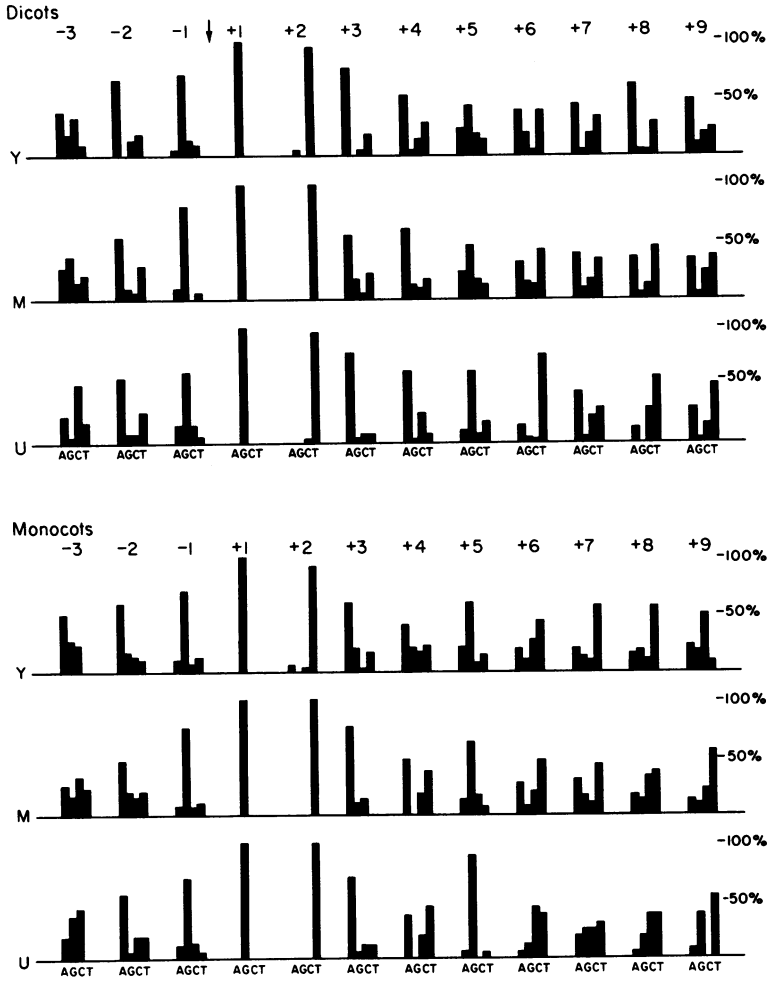

Figure 5. The 5' splice site boundaries present in each intron class. The nucleotide frequencies occurring at 5' splice sites between positions -3 and +9 are recorded for the pyrimidine-rich (Y), mixed (M) and purine-rich (U) introns.

dicot introns are purine-rich, 21/101 (21%) are pyrimidine-rich and 43/101 (42%) occur in the mixed category. The distribution of these classes relative to the purine frequency is diagrammed in Fig. 3. Although this scheme subdivides the introns into smaller groups, nucleotide patterns appear even within these limited groups (Fig. 4). The length and pyrimidine-richness of the tracts present in the pyrimidine-rich monocot

introns resemble those found in mammalian introns (7). In contrast, dicot pyrimidine-rich introns contain more abbreviated pyrimidine tracts (Fig. 4) with prominent purines at positions -4, -5 and -7 which are absent from pyrimidine-rich monocot and mammalian introns. The dicot purine-rich introns contain purines as the predominant nucleotides at positions -4 and -6 through -12 but predominant pyrimidines at -3, -5, -13, and -14. Their monocot counterparts contain predominant purines at -4, -6, -8, -10, -13 and -14.

The diagrammatic analysis used in Fig. 3 and 4 suggests that the dicot purine-rich introns contain more prominent purine tracts than the monocot purine-rich introns. Inspection of individual introns reveals that a higher proportion of the dicot introns contain more than four consecutive purines in the region between -3 and -20 (19% dicots, 9% monocots).

Because intron recognition sequences occur at both ends of the intron, the purine and pyrimidine-rich classes of introns were evaluated for differences near the 5' splice sites in the region between -3 and +7 which hypothetically base pairs with U1 snRNA. The sequence compilations in Figure 5 indicate that each class has nucleotide preferences in positions -3 through +7. Although the 5' terminal sequences of only two dicot U1 snRNAs are known (68, 69), the first eleven nucleotides of these plant snRNAs are identical to the U1 snRNAs in mammals, Drosophila and yeast (65, 70). Potential base pairing structures can be identified between this sequence and the region near the 5' splice site. When the less stable guanosine-uridine base pairings are included, a large proportion of the pyrimidine-rich introns in dicots can not base pair with U1 snRNA at positions -3, +5 and +6, while those in monocots can not base pair at positions -3, +6 and +7. A high percentage of the 5' splice sites present in the dicot purine-rich introns base pair with the U1 snRNA at all of these positions but those present in the monocot purine-rich introns base pair less effectively with the U1 snRNA.

DISCUSSION

The purpose of this study is to define differences between plant intron sequences to gain insight into the different splicing efficiencies observed in vitro and in vivo.

Evaluation of the 5' splice sites has indicated that this region in plant introns is much more variable than in mammalian introns (15). Comparison of the 5' splice sites in dicot and monocot introns (Fig. 1)

indicates that particular positions in this region are more divergent than previously suggested (15). The most apparent differences occur between positions +4 and +9 in the intron. In addition, the second nucleotide preferences for the monocot and dicot 5' splice sites vary considerably. The dicot 5' splice consensus sequence has increased proportions of purines at +4 and +7, and thymidine at position +6. The monocot and dicot nucleotide preferences generate consensus sequences which vary in the region (-3 to +6) potentially base pairing with a plant U1 snRNA. Although no sequence is available for the 5' end of any monocot U1 snRNA, the 5' sequences for two dicot U1 snRNAs (Phaseolus vulgaris, Glycine max) are identical to the U1 snRNA sequence found in other organisms (65, 68-71). When the monocot and dicot 5' splice consensus sequences are evaluated with respect to this highly conserved sequence, a higher proportion of the dicot introns base pair with U1 snRNA at +4 and +6 and fewer base pair at position +5. A large proportion of monocot introns can not base pair with U1 snRNA at +4, +6 or +7. The large number of plant introns (36%) which have less than 6/9 nucleotides complementary to the U1 snRNA consensus suggests that alternate U1 snRNAs exist in plant nuclei. In support of this, multiple major and minor forms of pea U1 snRNA have recently been resolved on two-dimensional gels (72). Northern analysis using oligonucleotide probes specific for U1 snRNA has demonstrated that three to five abundant forms of U1 snRNA exist in all monocot and dicot nuclei examined (73). Whether any possess alternate 5' sequences has not yet been determined.

Because sequences near the 3' splice site are involved in the first stages of the mammalian intron recognition and cleavage at the 5' splice site (74-76), we have separated the plant 3' splice sites into groups containing pyrimidine or purine-rich tracts between nucleotides -3 and -20. The high frequency of purine-rich and mixed classes of introns in the plant intron pool (30% purine, 40% mixed) is significantly different ($p=0.000$) from the frequency of purine-rich and mixed introns in vertebrate and viral genes (3% purine, 12% mixed). Limited but less extensive purine-tracts are also prevalent upstream from the 3' splice boundaries of Drosophila introns (11). The proportion of purine and pyrimidine-rich introns in monocots and dicots varies significantly ($p=0.0075$). Whereas 37% of the dicot introns are purine-rich and 21% are pyrimidine-rich, 21% of the monocot introns are purine-rich and 41% are pyrimidine-rich. Unlike the other intron classes, the purine-rich introns have a highly conserved 5' splice sequence. The

distribution of these intron types is not limited to a particular class of genes. Some genes with multiple introns contain a single type of intron but the majority contain two or three types.

In contrast to previous evaluations (7, 15, 16), we have classified introns using an extended 3' splice region (-3 to-20) that includes the entire region (-2 to -19) necessary for binding of the mammalian U5 snRNP (66). The three intron classes can also be defined using the abbreviated -3 to -15 region (not shown). With this more limited region, a higher proportion are designated as mixed introns because few have distinctive pyrimidine or purine tracts within this subset of nucleotides. Nevertheless, a higher proportion of dicot introns are purine-rich and a higher proportion of monocot introns are pyrimidine-rich (dicots: 29% purine, 47% mixed, 24% pyrimidine; monocots: 15% purine, 48% mixed, 37% pyrimidine) ($p=0.0493$).

The relevance of these observations lies in the importance of the splice sites in plant intron excision. In mammalian intron recognition, specific sequences near the 3' splice site are required for the initial 5' cleavage events (66, 67, 74-76), but in yeast, 5' cleavage can occur in the absence of a functional 3' splice site (77). The variety of 3' splice sites found in plant introns suggest that different intron recognition mechanisms and/or factors exist within plant nuclei. If, as in mammalian cells, the 3' splice site sequences are required for the recognition of plant introns, different factors must associate with the pyrimidine and purine tracts. If the 3' splice site is dispensible for the initial cleavage events then sequences in this region may be highly diverged and other sequences may be well conserved. In this regard, the pyrimidine-rich plant introns resemble mammalian introns having pyrimidine tracts and relatively unconserved 5' splice sequences (7). The purine-rich introns resemble yeast introns which lack pyrimidine tracts and have strictly conserved 5' splice sequences (13). Both classes are abundant in plants suggesting that both the mammalian and yeast intron recognition mechanisms exist in plant nuclei. Clearly, the plant snRNA profiles are complex enough (72, 73) to accommodate both mechanisms.

One aim of our intron comparisons is to identify monocot and dicot RNA processing signals which determine the processing efficiency of an intron in a foreign environment. The purine-rich or pyrimidine-rich divisions of plant introns described in this paper emphasize one of the major differences between plant and animal introns. Even though the splicing

efficiency for only a few plant introns have been defined (24), these, and future, in vivo expression studies can be evaluated in light of these intron classes. In a variety of plant gene transfer experiments, transgenic tobacco nuclei efficiently excised dicot introns but were limited in their ability to excise monocot introns. The two dicot introns with well-defined splicing efficiencies (24) fall in the mixed and pyrimidine-rich categories. The mixed intron (rbcS-E9-intron 2) is processed efficiently in transgenic tobacco cells despite limited complementary (4/9 matches) with nucleotides 2 to 10 in the U1 snRNA consensus sequence. The pyrimidine-rich intron (rbcS-3A-intron 1) which is excised efficiently has good complementarity with U1 snRNA (7/9 matches). The only monocot introns for which any in vivo splicing efficiencies have been reported (24) fall in the pyrimidine-rich class. Both introns, intron 1 of the wheat rbcS gene and intron 6 of the maize Adh-1S gene, contain multiple nucleotides used infrequently in dicot 5' splice sites. Thus, monocot introns may be ineffectively processed in dicot nuclei because the dicot splicing machinery has a limited ability to recognize the usual 5' splice sites found in monocot introns. The frequency of the purine-rich and mixed introns in dicots suggest that dicot cells are more capable of processing purine-rich introns which have distinct 5' splice site sequences. Monocot cells, which have abundant pyrimidine-rich introns, may be more capable of excising pyrimidine-rich introns.

Based on the above analysis, we conclude that the splicing deficiencies of monocot pre-mRNAs in dicot nuclei probably arise from a combination of factors including the presence of multiple unusual nucleotides in the monocot 5' splice sites and the relative absence of purine-rich regions near the 3' splice site. Pyrimidine-rich introns can be spliced in dicot nuclei, but as deviations from the dicot 5' splice consensus accumulate, the splicing efficiency declines. Although site-directed mutagenesis, in vitro and in vivo processing experiments must be used to evaluate the importance of these parameters, our observations provide some guidelines for predicting and evaluating the splicing potential for plant introns in heterologous systems.

ACKNOWLEDGEMENTS

We gratefully acknowledge Warren Lamboy for statistical evaluations. This work has been supported by a grant from the National Science Foundation (DCB-8402792).

*To whom correspondence should be addressed.

REFERENCES

1. Padgett,R.A., Hardy,S.F. and Sharp,P.A. (1983) Proc. Natl. Acad. Sci. USA 80, 5230-5234.
2. Hernandez,N. and Keller,W. (1983) Cell 35, 87-99.
3. Ruskin,B., Krainer,A.R., Maniatis,T. and Green,M.R. (1984) Cell 38, 317-331.
4. Lin,R.-J., Newman,A.J., Cheng,S.C. and Abelson,J. (1985) J. Biol. Chem. 260, 14780-14792.
5. Newman,A.J., Lin,R.-J., Cheng,S-C. and Abelson,J. (1985) Cell 42, 335-344.
6. Breathnach,R., Benoist,C., O'Hare,K., Gannon,F. and Chambon,P. (1978) Proc. Natl. Acad. Sci. USA 75, 4853-4857.
7. Mount,S.M. (1982) Nucl. Acids Res. 10, 459-472.
8. Langford,C.J. and Gallwitz,D. (1983) Cell 33, 518-527.
9. Pikielny,C.W., Teem,J.L. and Rosbash,M. (1983) Cell 34, 395-403.
10. Keller,E.B. and Noon,W.A. (1984) Proc. Natl. Acad. Sci. USA 81, 7417-7420.
11. Keller,E.B. and Noon,W.A. (1985) Nucl. Acids Res. 13, 4971-4981.
12. Maniatis,T. and Reed,R. (1987) Nature 325, 673-678.
13. Padgett,R.A., Grabowski,P.J., Konarska,M.M., Seiler,S. and Sharp,P.A. (1986) Ann. Rev. Biochem. 55, 1119-1150.
14. Slightom,J.L., Sun,S.M. and Hall,T.C. (1983) Proc. Natl. Acad. Sci. USA 80, 1897-1901.
15. Brown,J.W.S. (1986) Nucl. Acids Res. 14, 9549-9559.
16. Brown,J.W.S., Feix,G. and Frendeway,D. (1986) EMBO J. 5, 2749-2758.
17. Langford,C., Nellen,W., Niessing,J. and Gallwitz,D. (1983) Proc. Natl. Acad. Sci. USA 80, 1496-1500.
18. Watts,F., Castle,C. and Beggs,J. (1983) EMBO J. 2, 2085-2091.
19. Ruskin,B., Pikielny,C.W., Rosbash,M. and Green,M.R. (1986) Proc. Natl. Acad. Sci. USA 83, 2022-2026.
20. Hartmuth,K. and Barta,A. (1986) Nucl. Acids. Res. 14, 7513-7527.
21. Van Santen,V. and Spritz,R. (1987) Gene 56, 253-265.
22. Cramer,J.H., Lea,K. and Slightom,J.L. (1985) Proc. Natl. Acad. Sci. USA 82, 334-338.
23. Barta,A., Sommergruber,K., Thompson,D., Hartmuth,D., Matzke,M.A. and Matzke,A.J.M. (1986) Plant Mol. Biol. 6, 347-357.
24. Keith,B. and Chua,N.-H. (1986) EMBO J. 5, 2419-2425.
25. Dennis,E.S., Gerlach,W.L., Pryor,A.J., Bennetzen,J.L., Inglis,A., Llewellyn,D., Sachs,M.M., Ferl,R.J. and Peacock,W.J. (1984) Nucl. Acids Res. 12, 3983-4000.
26. Callis,J. (1986) Ph.D. Thesis, Stanford University, CA.
27. Lamppa,G., Nagy,F. and Chua,N.-H. (1985) Nature 316, 750-752.
28. Matzke,M.A., Susani,M., Binns,A.N., Lewis,E.D., Rubenstein,I. and Matzke,A.J.M. (1984) EMBO J. 3, 1525-1531.
29. Goldsbrough,P.B., Gelvin,S.B. and Larkins,B.A. (1986) Mol. Gen. Genetics 202, 374-381.
30. Shah,D.M., Hironaka,C.M., Wiegand,R.C., Harding,E.I., Krivi,G.G. and Tiemeier,D.C. (1986) Plant Mol. Biol. 6, 203-211.
31. Werr,W., Frommer,W.-B., Maas,C. and Starlinger,P. (1985) EMBO J. 4, 1373-1380.
32. Klosgen,W.B., Gierl,A., Schwarz-Sommer,S. and Saedler,H. (1986) Mol. Gen. Genet. 203, 237-244.
33. Sheldon,H., Ferl,R., Fedoroff,N. and Hannah,L.C. (1983) Mol. Gen. Genet. 190, 421-426.

34. Shah,D.M., Hightower,R.C. and Meagher,R.B. (1983) J. Mol. Appl. Genet. 2, 111-126.
35. Marchionni,M. and Gilbert,W. (1986) Cell 46, 133-141.
36. Tischer,E., DasSarma,S. and Goodman,H.M. (1986) Mol. Gen. Genet. 203, 221-229.
37. Sommer,H. and Saedler,H. (1986) Mol. Gen. Genet. 202, 429-434.
38. Broglie,R., Coruzzi,G., Lamppa,G., Keith,B. and Chua,N.-H. (1983) Biotech. 1, 55.
39. Shah,D.M., Hightower,R.C. and Meagher,R.B. (1982) Proc. Natl. Acad. Sci. USA 79, 1022-1026.
40. Hyldig-Nielsen,J.J., Jensen,E.O., Paludan,K., Wiborg,O., Garrett,R., Jorgensen,P. and Marcker,K.A. (1982) Nucl. Acids Res. 10, 689-701.
41. Wiborg,O., Hyldig-Nielsen,J.J., Jensen,E.O., Paludan,K. and Marcker,K.A. (1982) Nucl. Acids Res. 10, 3487-3493.
42. Wiborg,O., Hyldig-Nielsen,J.J., Jensen,E.O., Paludan,K. and Marcker,K.A. (1983) EMBO J. 2, 449-452.
43. Katinakis,P. and Verma,D.P.S. (1985) Proc. Natl. Acad. Sci. USA 82, 4157-4161.
44. Nguyen,T., Zelechowska,M., Foster,V., Bergmann,H. and Verma,D.P.S. (1985) Proc. Natl. Acad. Sci. USA 82, 5040-5044.
45. Schuler,M.A., Schmitt,E. and Beachy,R.N. (1982) Nucl. Acids. Res. 10, 8225-8244.
46. Marco,Y.A., Thanh,V.H., Tumer,N.E., Scallon,B.J. and Nielsen,N.C. (1984) J. Biol. Chem. 259, 13436-13441.
47. Berry-Lowe,S.L., McKnight,T.D., Shah,D.M. and Meagher,R.B. (1982) J. Mol. Appl. Genet. 1, 483-498.
48. Chen,J. and Varner,J.E. (1985) EMBO J. 4, 2145-2150.
49. Lee,J.S. and Verma,D.P.S. (1984) EMBO J. 3, 2745-2752.
50. Lycett,G.W., Croy,R.R.D., Shirsat,A.H. and Boulter,D. (1984) Nucl. Acids Res. 12, 4493-4506.
51. Bown,D., Levasseur,M., Croy,R.R.D., Boulter,D. and Gatehouse,J.A. (1985) Nucl. Acids Res. 13, 4527-4537.
52. Coruzzi,G., Broglie,R., Edwards,C. and Chua,N.-H. (1984) EMBO J. 3, 1671-1679.
53. Fluhr,R., Moses,P., Morelli,G., Coruzzi,G. and Chua,N.-H. (1986) EMBO J. 5, 2063-2071.
54. Baumlein,H., Wobus,U., Pustell,J. and Kafatos,F.C. (1986) Nucl. Acids Res. 14, 2707-2720.
55. Rosahl,S., Schmidt,R., Schell,J. and Willmitzer,L. (1986) Mol. Gen. Genet. 203, 214-220.
56. Pikaard,C.S., Mignery,G.A., Ma,D.P., Stark,V.J. and Park,W.D. (1986) Nucl. Acids Res. 14, 5564-5566.
57. Bevan,M., Barker,R., Goldsbrough,A., Jarvis,M., Kavanagh,T. and Iturriaga,G. (1986) Nucl. Acids Res. 14, 4625-4638.
58. Boutry,M. and Chua,N.-H. (1985) EMBO J. 4, 2159-2165.
59. Mazur,B.J. and Chui,C.-F. (1985) Nucl. Acids Res. 13, 2373.
60. Turner,N.E., Clark,W.G., Tabor,G.J., Hironaka,C.M., Fraley,R.T. and Shah,D. (1986) Nucl. Acids Res. 14, 3325.
61. Mount,S.M., Petterson,I., Hinterberger,M., Karmas,A. and Steitz, J.A. (1983) Cell 33, 509-518.
62. Kramer,A.R., Keller,W., Appel,B. and Luhrmann,R. (1984) Cell 38, 299-307.
63. Black,D.L., Chabot,B. and Steitz,J.A. (1985) Cell 42, 737-750.
64. Zhuang,Y. and Weiner,A.M. (1986) Cell 46, 827-835.
65. Siliciano,P.G., Jones,M.H. and Guthrie,C. (1987) Science 237, 1484-1487.

66. Tazi,J., Alibert,C., Temsamani,J., Reveilland,I., Cathala,G., Brunel,C., and Jeanteur,P. (1986) Cell 45, 755-766.
67. Gerke,V. and Steitz,J.A. (1986) Cell 47, 973-984.
68. Van Santen,V. and Spritz,R. (1987) Proc. Natl. Acad. Sci. USA 84, 9094-9098.
69. Spritz, R. Personal communication.
70. Reddy,R. (1985) Nucl. Acids Res. 13, r155-r163.
71. Brown,D.T., Morris,G.F., Chodchoy,N., Sprecher,C., and Marzluff,W.F. (1985) Nucl. Acids Res. 13, 537-556.
72. Tollervey,D. (1987) J. Mol. Biol. 196, 355-361.
73. Egeland,D.B., Pizanis,A., and Schuler,M.A. (1988) Manuscript in preparation.
74. Reed,R. and Maniatis,T. (1985) Cell 41, 95-105.
75. Frendeway,D. and Keller,W. (1985) Cell 42, 355-367.
76. Ruskin,B. and Green,M.R. (1985) Cell 43, 131-142.
77. Rymond,B. and Rosbash,M. (1985) Nature 317, 735-736.