Plant Molecular Biology **36:** 573–583, 1998.

U-richness is a defining feature of plant introns and may function as an intron recognition signal in maize

Christopher H. Ko[1], Volker Brendel[2], Rebecca D. Taylor[1] & Virginia Walbot[1,*]
[1]*Department of Biological Sciences and* [2]*Department of Mathematics, Stanford University, Stanford, CA 94305-5020, USA (*author for correspondence)*

Received 13 March 1997; accepted in revised form 14 October 1997

Key words: intron processing, base composition, U-rich motif, *Zea mays*

Abstract

Using a large set of plant gene sequences we compared individual introns to their flanking exons. Both *Zea mays* and *Arabidopsis thaliana* introns are U-rich but display no apparent bias for A. We identified fifteen 11-mer U-rich motifs as frequent elements of maize introns, and these are virtually absent from exons. By mutagenesis, we show that the single U-rich motif in the *Bronze2* intron of maize plays a key role in intron processing *in vivo*.

Introduction

A fundamental process in all eukaryotes is the removal of intervening sequences (introns) from nuclear pre-mRNAs by the spliceosome, a large complex of proteins and RNAs. In mammals and fungi, the availability of *in vitro* splicing systems has permitted elucidation of two *trans*-esterification reactions involving the 5′-splice donor site, the branch point and the 3′-splice acceptor site. To date, however, a plant *in vitro* splicing system has not been developed. Plants do share the structurally conserved genes important for pre-mRNA processing, such as small nuclear RNAs and small nuclear RNPs [9, 18, 22, 35; for reviews see 37, 39]. Furthermore, genes encoding some of the SR proteins important in splice site definition in animals have recently been cloned from plants [23, 25, 26]. This suggests that the biochemistry of splicing reactions is similar in all eukaryotes.

Nearly all plant introns contain the canonical 5′ GT and 3′ AG splice sites and a loosely conserved YUNAN branch point consensus [24, 36], but they generally lack both the highly conserved polypyrimidine tract characteristic of mammalian introns and the invariant branch point consensus characteristic of yeast [5, 10, 16, 30]. When plant introns were tested in a HeLa cell *in vitro* splicing extract, only introns that fortuitously contained a polypyrimidine tract were processed efficiently [41, 43]. These results indicate that plants are likely to have unique intron recognition motifs. Not only are the central regions of maize introns sufficient to create an intron but also sizable deletions from large introns have no impact on splicing [31]. Thus, plant intron recognition motifs are located within the body of the intron and appear to be redundant.

One feature that has been postulated to be a plant intron recognition motif is A+U-richness; the A+U base content of plant introns is greater than that of many other higher organisms. For example, on average plant introns are reported to be at least 15% more A+U-rich than plant exons whereas human introns are only 2% more A+U-rich compared to human exons [8]. Based on this observation, Filipowicz and colleagues constructed a number of synthetic introns and tested their splicing efficiency in dicots; introns containing a high A+U base composition were processed more efficiently than introns that were G+C-rich [13]. Many similar experiments followed and corroborated the finding that A+U-richness is a *cis*-sequence feature important for pre-mRNA processing [31, 32].

Evidence that U-richness, rather than A+U-richness, is the key feature for intron processing has emerged from studies of maize introns. Addition of a U-rich motif to a poorly spliced intron yields a greater improvement in splicing efficiency than addition of an A-rich motif [31]. In support of these results, a survey

of maize sequences for the average frequency of each base indicated that the bias for U in introns (34% in introns vs. 20% in exons) is greater than the bias for A (26% in introns vs. 22% in exons) [30]. A specific role of U-rich tracts in 3′-splice site recognition has been demonstrated [2]. More recently, Gniadkowski *et al.* [12] demonstrated that insertion of U-rich but not A-rich segments can activate splicing of a GC-rich synthetic intron in *Nicotiana plumbaginifolia*. The working hypothesis for plant intron recognition postulates that transacting factors recognitze U- or AU-rich intron motifs [28, 30].

To determine whether a statistically significant base composition bias exists for either A or U residues in maize and *Arabidopsis* intron sequences in the context of their own exons, we compiled new databases consisting of 46 distinct maize (monocot) genes and 131 distinct *Arabidopsis* (dicot) genes. Previously sequence analyses were used to determine the average frequency of each base in the populations of introns and of exons [10, 30]. Although these analyses have been helpful in highlighting base compositional bias, their utility is limited because individual introns were not compared to their flanking exons. We also used the maize database to identify fifteen common U-rich motifs of introns.

As our new database analysis pinpoints U-richness rather than A+U-richness as characteristic of plant introns, we wished to investigate the importance of a specific U-rich motif in intron processing *in vivo*. We utilized a reporter construct containing the maize *Bronze2* (*Bz2*) intron. This 78 bp intron has only one, centrally located U-rich motif. The importance of the U-rich motif was tested directly by measuring splicing efficiency of various mutations within this motif. We show that a positive correlation appears to exist between U-richness and splicing efficiency, suggesting that the U-rich motif is a key element in splicing for maize.

Materials and methods

Database construction and sequence analysis

Genomic sequences from *Zea mays* and *Arabidopsis thaliana* were compiled from GenBank. Only completely sequenced genes for which the sequences of all introns between the translation initiation codon and the stop codon are available were included in our database. For maize, we collected 46 genes that encode distinct proteins (redundancy based on significant sequence similarity assessed as previously described [3]). These genes comprise a total of 250 exons and 204 introns. For *Arabidopsis*, we compiled a database of 131 distinct genes with a total of 709 exons and 578 introns. The compositional data displayed in Table 1 and Figure 1 were derived from exons and introns of a minimal length of 60 bp, amounting to 237 exons plus 204 introns for maize and 647 exons plus 577 introns for *Arabidopsis*. In all genes, the first exon was taken to extend from the translation initiation codon to the downstream 5′ splice junction, and the last exon was taken to extend from the upstream 3′ splice junction to the stop codon. Thus, the base composition of exons reflects the composition of coding regions only and does not include any contribution from non-coding exons. To assess compositional differences between an intron and its flanking exons, we also compiled databases consisting of 50 bp of the upstream exon, all of the intron, and 50 bp of the downstream exon. This collection was restricted to short introns of lengths 60–200 bp. These constraints resulted in 175 data points for maize and 422 data points for *Arabidopsis*. The composition of flanking exons in Figure 2 was calculated as the average composition over the upstream 50 bp of exon and the downstream 50 bp of exon.

For the motif search based on Table 2, we first defined all 11-mers occurring in the constructs between positions 25 and 45 (Table 2). There are 80 distinct 11-mers that occur only in the constructs that are spliced as well or better than the wild type (>30%), and there are 115 distinct 11-mers characteristic of the down mutation (<11%). Matches against these motif templates were defined as 11-mers that are identical to the template in at least 9 positions. For the general search for 11-mer maize intron motifs listed in Table 3, we analyzed all intron and exon sequences for occurrence and listed the top fifteen motifs preferentially found in introns.

Over-representation of U-tracts and other patterns were assessed by the method of Schbath *et al.* [34]. In brief, for each pattern $W = N_1\ N_2\ ...\ N_h$, where N_i is one the four letters A, C, G or U, a z-score is defined as $Z(W) = (o(W) - e(W))/s(W)$, where o(W) is the observed count of W in the data set, e(W) is the expected count calculated for different order Markov models, and s(W) is the standard deviation. The z-scores are asymptotically normally distributed, and z-scores exceeding 3.0 were taken to indicate significant over-representation. Here, we applied the method to patterns of size $h <= 6$ and Markov orders zero to

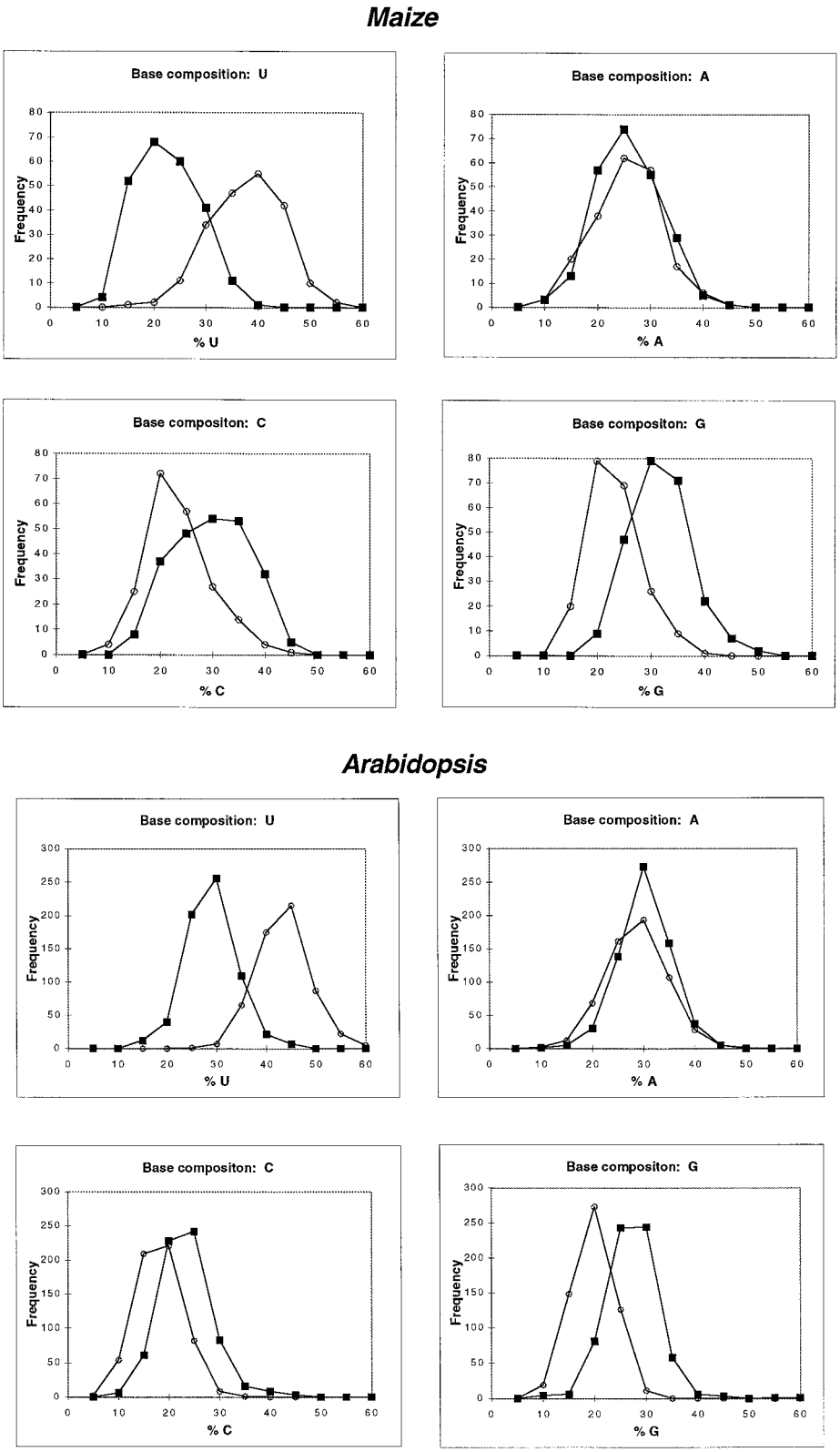

Figure 1. Average base contents. Distribution of base content percentages for maize and *Arabidopsis* introns and exons. Introns are denoted by open circles and exons are denoted by filled squares.

Table 1. Average intron and exon base composition analysis.

	Base	Average base composition (%)		
		[intron]	*[exon]*	*[intron]-[exon]*
Maize	U	35.2	20.1	15.1
	A	23.2	23.4	−0.2
	C	20.9	27.2	−6.3
	G	20.7	29.3	−8.6
Arabidopsis	U	41.0	26.4	14.6
	A	26.2	27.8	−1.6
	C	15.6	20.9	−5.3
	G	17.2	24.9	−7.7

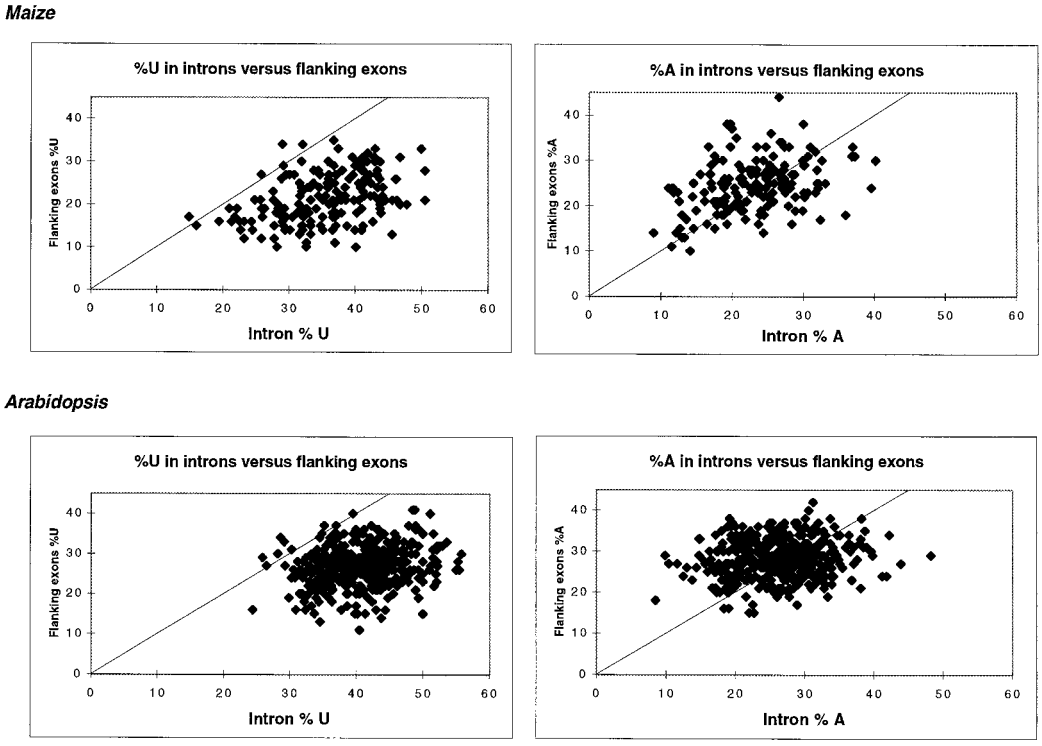

Figure 2. Comparison of individual introns to their flanking exons. Maize: %U in individual introns vs. %U in the flanking exons (left) and %A in individual introns vs. %A in the flanking exons (right). *Arabidopsis*: %U in intron vs. %U in the flanking exons (left) and %A in intron vs.%A in the flanking exons (right). The bisecting lines represent equal usage in an intron and its flanking exons. Points above the line correspond to introns for which the flanking exons have higher U content (left) or A content (right), and points below the line correspond to introns for which the flanking exons have lower U content (left) or A content (right).

two. For example, the expected frequency of U_4 under the zero order Markov assumption is given by $f(U)^4$ and by $f(U)\,[f(UU)/f(U)]^3$ under the first-order Markov assumption, where f(U) and f(UU) are the observed frequencies of U and UU.

z-scores were calculated by the programs of Schbath *et al.* [34] available at http://www-bia.inra.fr/J/AB/genome/RMES/welcome.html>
The databases used in this study are available via anonymous ftp to gnomic.stanford.edu, see file pub/README.PlantSpl.

Table 2. Relative splicing efficiency of introns with mutations in the U-rich motif.

Construct	Motif	U residues total	U residues run	Relative splicing efficiency (%)
pSU	GCAAGUUUUUUUUUUCAAGG	11	11	103
pS31	GCAAGUGUUUGUUCUUCAAGG	8	3	71
pS29	GCAAGUGUUUUUUCGGCAAGG	7	6	65
pS32	GCAAGUGUCGUUUUUUCAAGG	8	6	50
pS101	GCAAGUUGUUGGGUUUCAAGG	7	3	50
pS43	GCAAGGUUCUUUUCUUCAAGG	8	4	43
pS90	GCAAGUUUUGUUGUUUCAAGG	9	4	34
pSWT	GCAAGUGUCUUUUCUUCAAGG	8	4	30
pS62	GCAAGUGUCUUGUUGUCAAGG	7	2	11
pS68	GCAAGUGUCGGGGUUUCAAGG	5	3	7
pS64	GCAAGUGUCUGGUUGGCAAGG	5	2	5
pS95	GCAAGUGGCUGUUCUGCAAGG	5	2	5
pS94	GCAAGUGUCGUGUCGGCAAGG	4	1	5
pS40	GCAAGGUGUUUGUCUGCAAGG	6	3	4
pS76	GCAAGUGGCUUGGCGGCAAGG	3	2	4
pS21	GCAAGUGUUGUGGCGUCAAGG	5	2	4
pS52	GCAAGUGUGGUGGCGUCAAGG	4	1	3
pSG	GCAAGGGGGGGGGGGGCAAGG	0	0	3
pS74	GCAAGGGUCGGGGUUGCAAGG	3	2	2

Table 3. U-rich maize intron motifs (11-mer).

Motif	# of Us	Longest U-run	Occurrence (%) Intron	Exon
UGAUUUUUUUU	9	8	26	0
UUGAUUUUUUU	9	7	24	1
UUCUUUGUUUU	9	4	23	1
UUCUUUUUUUU	10	8	23	0
UUUAUUUUUUU	10	7	23	0
UUUUCUUUGUU	9	4	23	1
UUUUUUUCUUU	10	7	23	0
UGUUUUUUUUU	10	9	22	1
UUAUUUUUUUU	10	8	22	0
UUUCUUUGUUU	9	3	22	0
UUUGUUUAUUU	9	3	22	1
UUUGUUUCUUU	9	3	22	1
UUUUUCUUUCU	9	5	22	0
UUUUUUCUUUG	9	6	22	1
UUUUUUUUGU	10	9	22	1

Plasmid construction

The construction of plasmid pSWT (previously called pSuccess) has been described [6]. This plasmid contains the *Bronze2* (*Bz2*) intron fused to the firefly luciferase gene; the luciferase gene is expressed only when the intron is spliced accurately because the ATG is upstream of the intron. All additional plasmids contained mutations in the U-rich motif (positions +30 to +40 in the intron; see Figure 3) and were constructed by the double-stranded site-directed mutagenesis method according to the manufacturer's instructions

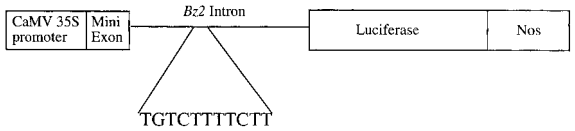

Figure 3. A schematic diagram of plasmid pSWT. The U-rich motif from position numbers +30 to +40 is highlighted (+1 is the G residue at the start of the intron). Not drawn to scale.

(Clonetech). For mutagenesis, a degenerate primer CGGCAGCAAGt/gt/gt/gt/gt/ct/gt/gt/gt/ct/gt/g CAAGGTAACGTG was used to generate all mutants except pSU and pSG. The plasmids pSU and pSG were derived from pSWT by replacing TGTCTTTTCTT with 11 Us or Gs, respectively. For RNA probe synthesis, the plasmid pBSWT was constructed by inserting a 110-mer oligonucleotide fragment that contains the *Bz2* intron into the *Sma*I site of pBluescript. The plasmids pBSSG and pBSSU were constructed by the double-stranded site directed mutagenesis method described above.

Transient assays

Electroporation and reporter assays were performed as previously described using maize Black Mexican Sweet (BMS) suspension tissue culture cells [29]. Protoplasts were generated by incubating cells with 0.5% Rhozyme (Interspec) and 0.5% Cellulase RS (Yakult Honsha Co.). A GUS expression plasmid was used as an internal standard to correct for transfection efficiency. Each construct was tested in duplicate samples in at least three independent experiments. The averages from these experiments are shown in Table 2. Standard deviations were less than 15% of the averages within an experiment.

RT-PCR analysis

Reverse transcription-PCR was performed as previously described [6]. The PCR products were fractionated on a 2% NuSieve agarose gel, transferred onto Hybond membrane, and probed with a ^{32}P-labeled oligonucleotide complementary to exon sequences.

Maize BMS Nuclear Extracts

Nuclear extracts were prepared from maize BMS cells by using an *Arabidopsis* nuclear extract protocol [11], except that cells were sonicated and a 100X protease cocktail [1] was substituted for antipain and leupeptin. The extracts were resuspended to a final concentration of 5 mg/mL.

Results

Base composition analysis of plant intron and exon sequences

It has been known for nearly a decade that, on average, plant introns are more A+U-rich than plant exons (recently reviewed in [37]. The current availability of many additional fully sequenced plant genes enabled us to study this compositional bias in more detail. Table 1 shows that average A usage is essentially the same in plant introns and exons and does not differ much from 25%. On the other hand, the average U content of introns exceeds the average U content of exons by 15.1% in maize and by 14.6% in *Arabidopsis*. Thus, the 'A+U-richness' of plant introns is actually U-richness (see also the discussion of the importance of U in [37]). The under-representation of C and G accompanying the U-richness of the introns occurs to about an equal extent, such that neither C or G appears to be preferentially avoided in introns. Table 1 also shows that both introns and exons of *Arabidopsis* genes have an average U content that is about 6% higher than that of maize genes.

The distributions of base content percentages for large collections of maize and *Arabidopsis* introns and exons are displayed in Figure 1. All the distributions are unimodal and more or less bell-shaped. Most striking is the difference between U and A in both species: the distributions of U-content are clearly separated between introns and exons whereas the distributions of A content are virtually identical in introns and exons.

Figure 1 demonstrates that a typical plant intron will be more U-rich than a typical plant exon. There are, however, some introns that are of relatively low U content, and some exons of relatively high U content. For example, in *Arabidopsis* 31 of 577 introns contain less than 33% U and 44 of 647 exons have a U content of at least 33%. Thus, the significantly large overlap in the U distribution (Figure 1) indicates some probability of randomly choosing an intron that is less U-rich than an exon.

To obtain a more detailed picture of the contrast in U content between introns and exons, we determined the difference in U content between introns and their immediately flanking exons. This analysis is especially important, because it analyzes the sequence of pre-

mRNAs in the way individual transcripts are presented to the spliceosome *in vivo*. These results are displayed in Figure 2. The plotted line in each graph represents equal base content (U or A) for an intron and its flanking exons; points above the line represent cases for which the flanking exons have higher U or A content than the intron, and points below the line represent introns for which the flanking exons have lower U or A content. It is clear that the contrast in U content between introns and exons is strongly conserved in the comparison of individual introns and their flanking exons. The differences in U content of each intron and its flanking exons in both maize and *Arabidopsis* are essentially normally distributed, with a mean of 14% and a standard deviation of 7%. Only 4 of 175 maize introns (2.3%) have a lower U content than the flanking exons; in *Arabidopsis* the analogous fraction is 9 out of 422 introns (2.1%). The differences in A content of an intron and its flanking exons are also normally distributed, but in contrast to U content, the mean difference is -1.5% in maize and -2.0% in *Arabidopsis* (see Figure 2: the points of intron %A vs. flanking exons %A are scattered about evenly on both sides of the bisecting line). These results unambiguously demonstrate that individual plant introns are U-rich, not A+U-rich, compared to their flanking exons.

The results also suggest that the contrast in U content between introns and exons may reflect selection on the average base content of introns and exons, for example for codon usage. More importantly, it appears that there is a local constraint which acts specifically to maintain the local contrast between an intron and its flanking exons. We tested this hypothesis of local constraint in two ways. First, we calculated the proportion of negative values for intron U content minus exon U content, assuming independent pairing of all intron and exon combinations. Specifically, for the 175 maize intron/flanking exon U-content values we formed all independent combinations ($175 \times 175 = 30\,625$) of intron U content with exon U content. A total of 2097 (6.8%) of these combinations gave negative values, a proportion about three times higher than the one observed. For *Arabidopsis*, 4839 of 178 084 possible combinations (2.7%) gave negative values, slightly above the observed value. These calculations indicate that the authentic intron and exon combinations are more likely to have a high U content difference than combinations formed by chance. Second, in order to test the statistical significance of the above results, we performed a randomization test by taking 100 random samples, each consisting of 175 data points for maize and 422 data points for *Arabidopsis* as in the original data sets. Intron and exon U content values were selected independently with replacement from the sets of observed values. For maize, only 2 of the 100 random samples contained as few as 4 data points with lower U-content in intron than exon; for *Arabidopsis*, 24 of the 100 random samples contained 9 or fewer exons with higher U content than the intron. We conclude from both simulations that there is evidence for a local constraint, particularly in maize, that enforces a higher U content in introns compared to their flanking exons. Specific examples of maize genes with overall high and low GC content but with the characteristic compositional contrast between introns versus exons are discussed elsewhere [4].

In vivo *analysis of splicing efficiency*

From prior experimental work and the demonstration of base compositional bias, we hypothesize that U-rich motifs could serve as internal intron recognition motifs in plants. To determine the impact of a single U-rich motif on splicing processes *in vivo*, we used the plasmid pSWT, which contains an exon and the maize *Bz2* intron fused to the luciferase gene. We had previously demonstrated that the pSWT plasmid expresses luciferase activity only when accurate splicing occurs and that only 30% of messages are spliced; the low splicing efficiency was attributed to the high G+C content of this intron relative to its flanking exons [6, 7]. A schematic diagram of the pSWT plasmid is shown in Figure 3. The *Bz2* gene contains a 78 base intron with a single, centrally located U-rich motif (8 U residues out of 11 bases, position numbers +30 to +40; see Figure 3). Using site-directed mutagenesis we generated a number of alterations in this motif (Table 2). The effect of each mutation was assessed by electroporating the altered plasmid into maize BMS protoplasts and then measuring luciferase expression. Gene expression was normalized to a GUS expression plasmid used as an internal standard.

As shown in Table 2 for mutants we have tested, reasonable splicing efficiency appears to be depend on a minimal U-rich motif within the intron. All of the mutants that spliced as well as the native *Bz2* intron or at higher levels ($>30\%$) have 7 or more Us in the 11-mer motif and a run of at least 3 contiguous Us. In contrast, the mutants that spliced less efficiently than the wild type ($<11\%$), generally had a lower number of Us (0–7) and/or lacked a contiguous U-run. Both high U content (7 or more) and the contiguous U-run

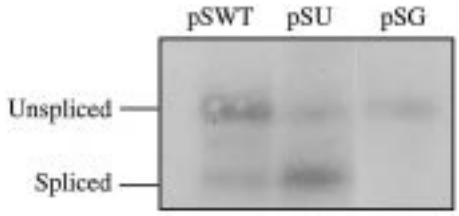

Figure 4. RT-PCR analysis of transcripts pSWT, pSU, and pSG. Each plasmid was electroporated into maize BMS cells and its RNAs, after DNaseI treatment, were analyzed by RT-PCR. The upper band is the amplified product of unspliced RNAs and the lower band is the amplified product of spliced RNAs.

appear to be important features because pS62 has 7 Us but only U dimers and showed 11% splicing and pS40 has a U trimer but only 6/11 Us and splices at 4%. The relative splicing efficiencies of plasmid pSU, which contains 11 U residues, and the plasmid pSG, which contains no U residues, are 103% and 3%, respectively.

RT-PCR experiments performed with total RNAs extracted from BMS protoplasts containing a test plasmid were in agreement with reporter assay results. A representative autoradiograph of an RT-PCR experiment with pSWT, pSU, and pSG is shown in Figure 4. We have previously established that RT-PCR data corroborate the results of our reporter assays [6, 7].

Motif analysis

To test whether the differences in relative splicing efficiency of these constructs can be attributed to sequence motifs found in actual maize introns and lacking in exons, we searched a maize database restricted to short introns (60–200 bases; 188 introns) and short exons (60–200 bases; 177 exons) for matches against motifs derived from the constructs (see Materials and methods for details). The restriction to short introns is not severe in that 80% of maize introns are at most 200 bases [4]. It is not known whether the mechanisms of intron recognition are different for longer introns, but for our purposes of generalization of results for the *Bz2* intron (78 bases) it seemed appropriate to limit the database comparisons to introns of typical lengths. Exon lengths were chosen in the same range to provide a comparable data set. The motifs present in constructs that exhibited a relative splicing efficiency of 30% or higher were represented in 73% of introns but were found in only 31% of exons. In contrast, the motifs in constructs that spliced poorly were represented in only 55% of introns and were actually more commonly found (63%) in exons. These results suggest that U-rich motifs that contributed to splicing success in our mutagenesis survey are preferentially found in maize introns.

To assess motif bias in a more general way we next searched our database for any 11-mer motif of any composition that was preferentially found in introns but not in exons. The results are shown in Table 3. The top fifteen motifs with the highest bias for presence in introns and absence in exons are (expectedly) all very U-rich. These motifs are each present in 22–26% of maize short introns but are found in only 0–1% of short exons. Although in our study of *in vivo* splicing the plasmid pSU with 11U residues is spliced with the highest efficiency, the intron motif search indicates that the most commonly found U-rich motif is interrupted by other bases (G>C>A). It is interesting that five of the common intron motifs have 9 Us but a contiguous run of only 3 or 4 Us, similar in structure to the motifs that improved splicing efficiency of the *Bz2* intron (Table 2).

The relation between statistical and biological significance of sequence patterns is complex because, for the most part, natural sequences do not fit mathematically tractable models (e.g. [19, 20]). However, even simple models may aid in data interpretation by providing some benchmarks for the generality of certain observations. In this context, we compared counts of U-tracts in the aggregate of short maize introns with expected counts calculated by the method of Schbath *et al.* [34] for Markov models up to order two. Table 4 shows that the UU dinucleotide is slightly under-represented (in accord with earlier data of White *et al.* [42]) but U_{3-6} are all over-represented, most significantly for U_4 and U_5 according to the Markov order one model. Thus, short U-tracts appear to be both abundant and over-represented even after accounting for the high U base content of maize introns.

Interestingly, short A-tracts are also highly over-represented, particularly A_{4-6} (Table 4), but at only one half to one third of the representation of the corresponding U-tract counts. For example, 61% of short maize introns contain at least one U_4 whereas only 27% contain at least one A_4.

For comparison, we also determined the same statistics for all 32 pentamers represented by the branch point motif YUNAN [36]. Of these, only CUGAU, CUGAC, and CUGAA are significantly over-represented according to any of the Markov models (Table 4). Noteworthy, C in position one, G in position three, and non-G in position five are the highest frequency residues in the standard branch point

Table 4. Occurrence of U- and A-tracts and specific branch point motifs in short introns of maize.

Motif	Occurrence (%) in introns[1]	Observed counts	Expected counts[2]		
			M0	M1	M2
UU	99	2309	2243	n/a	n/a
UUU	91	885	854	**764**	n/a
UUUU	61	342	299	**253**	339
UUUUU	31	152	**105**	**84**	130
UUUUUU	17	76	**37**	**28**	**50**
AA	96	1070	1023	n/a	n/a
AAA	57	301	**232**	**253**	n/a
AAAA	27	110	**52**	**60**	**85**
AAAAA	11	49	**12**	**14**	**24**
AAAAAA	5	28	**3**	**3**	**7**
CTGAC	18	35	**15**	**19**	24
CTGAA	14	34	**16**	21	27
CTGAU	20	43	**25**	31	41

[1] Data set: 188 maize introns of 60–200 bases. For example, 17% of these introns contain at least one incidence of UUUUUU.
[2] M0, M1, M2: Markov models of order zero, one, and two, respectively. Entries in bold face indicate significant over-representation by a z-score of at least 3.0.

sequence weight table [17]. Thus, in this case, over-representation by statistical criteria appears to reflect biological significance.

Discussion

Our refinement of analysis of base composition biases to compare introns to their flanking exons demonstrates that maize and *Arabidopsis* introns are U-rich and exons are G+C-rich. Our finding is entirely consistent with the observations that insertion of U-rich sequences into introns are generally much more effective than A-rich sequences in increasing splicing efficiency *in vivo* [individual experiments are reviewed in 37]. Even in many studies designed to study the impact of A+U-richness, most A+U insertions or deletions tested were predominantly U-rich [12, 13, 27, 32].

The U-richness of introns is a feature of both maize and *Arabidopsis*, model organisms representing monocots and dicots, the two branches of flowering plants. In a previous study, it was reported that some monocot introns were not accurately processed in dicots; these experiments have been interpreted to mean that even among angiosperms, intron recognition motifs have diverged [14, 21]. Nearly all studies of splicing in transient assays or in transgenic plants have used chimeric constructs in which a test intron was flanked by novel exons. It is not clear whether inaccurate processing of specific plant introns reflects a peculiarity of the individual introns tested, the context of the intron, which is often altered in constructs, or the existence of a diverged intron recognition motif. The possibility of an auxiliary splicing factor unique to major taxa of flowering plants cannot be ruled out, but we consider it unlikely.

We propose that efficient splicing in synthetic constructs as well as in transcripts of native genes depends on the contrast in U-richness between an intron and its flanking exons. Our database analysis demonstrates that this bias is approximately 15% for both maize and *Arabidopsis*. In a separate study, we have shown that manipulating either intron or exon G+C content to decrease that bias decreases splicing efficiency and this base changes that increase this bias favor splicing [7].

It is likely that specific U-rich motifs within introns are the essential features of U-richness. Our *in vivo* data assessing splicing of mutations in the 8/11 U motif of the *Bz2* intron demonstrated that this motif contributes

to splicing success in maize; generally, splicing efficiency was improved by an elevated U content and by a longer run of contiguous Us. Although it is difficult to define precisely the minimum number of Us necessary for intron recognition, mutants with at least a run of 4 Us in a region with additional Us nearby are spliced efficiently in our experiments.

Given that the run of 11 Us yielded the highest splicing efficiency, it is tempting to speculate that such a homouridine run is the intron recognition motif in plants. However, a search of our databases for homouridine runs of at least 11 showed that only 1 maize intron and 13 of 578 *Arabidopsis* introns contain such a motif. Therefore, it is unlikely that the plant intron recognition motif is limited to long homouridine motifs. We hypothesize that the actual intron recognition motif resembles the U-rich motifs in Table 3, a U-rich region punctuated by a few non-U residues. We speculate that such motifs, despite inherent degeneracy, could be accurately and efficiently found by RNA binding proteins. A single protein with broad sequence specificities, such as the human intron recognition factors $U2AF^{65}$ and polypyrimidine tract-binding protein [38] may exist. Alternatively, there may be several U-motif-binding proteins with distinct specificities.

The contribution of U-rich motifs to splicing has also been studied in non-plant organisms. The most detailed study has been performed in *Saccharomyces cerevisiae*. In this yeast, a U-rich motif plays a critical role in the selection of 3′-splice acceptor sites. This finding is supported by database analysis in which it is observed that U-rich motifs are concentrated just proximal to authentic 3′-splice sites [15, 40]. It is not known whether U-rich motifs play a similar, specific role in 3′-splice site selection in plants. U-rich motifs are distributed uniformly over the span of plant introns, suggesting that the role of a U-rich motif is not limited to 3′-splice site selection [32].

By what mechanism could the cell distinguish base compositional contrast of a high U content in introns and a high G+C content of exons to facilitate intron recognition? It is formally possible that general RNA stability during or after processing can explain the differences in yield of spliced and unspliced products observed in our study; however, a direct measurement of *Bz2* message half-life in the cytoplasm indicated that the spliced and unspliced mRNAs have an identical half-life of about 60 min [33]. In addition, RT-PCR yield (after correcting for transfection efficiency) is similar in various constructs containing the *Bz2* intron; some of these constructs produced pre-mRNAs designed so that only unspliced mRNA yields luciferase activity, suggesting that mRNAs with an intron are as stable in maize as spliced messages [6, 7]. Collectively, these results suggest that differential RNA half-life cannot explain the differences in the amount of spliced product or luciferase activity we measure.

We and others have previously proposed that U-rich motifs serve as an intron internal recognition motif. We hypothesize that *trans*-acting factors actually mediate the distinction between an intron and its flanking exons. Based on our analysis of introns and their flanking exons, proteins that recognize G+C-rich exons, U-rich introns, or a combination of both types of proteins could mediate intron recognition. In mammals and yeast, many *trans*-acting factors that bind to introns have been identified. It is likely that the same and perhaps additional factors operate in plants. Gniadkowski *et al.* reported two proteins with apparent molecular weights of 50 and 54 kDa that bind to poly(U) in *Nicotiana plumbaginifolia* [12]. Using standard procedures for nuclear protein separation [1, 11] we also find several proteins in maize nuclei that crosslink specifically to U-rich substrates (C.H. Ko, unpublished data). In particular, a 35 kDa protein binds preferentially to the *Bz2* intron native 8/11 U motif or to poly(U), and it can be purified by poly(U) affinity chromatography (R.D. Tayler, unpublished data). Ascribing a specific role to this and other intron- or exon-binding proteins in plants will await development of an *in vitro* splicing assay based on plant components.

Acknowledgements

We thank Mark Alfenito, Janie Hershberger, Ann Stapleton and Pepe Carle-Urioste for comments on the manuscript. C.H.K. was supported by a Plant Biology Postdoctoral Fellowship from the NSF. This work was supported by grants from NSF (IBN-96-03927) to V.W. and from NIH (2R01HG00335-09) to V.B.

References

1. Ausubel F, Brent R, Kingston R, Moore D, Seidman J, Smith J, Struhl K, (eds) Current Protocols in Molecular Biology, pp. 13.13.1–13.13.9. John Wiley, New York (1987).
2. Baynton CE, Potthoff SJ, McCullough AJ, Schuler MA: U-rich tracts enhance 3′ splice site recognition in plant nuclei. Plant J 10: 703–711 (1996).
3. Brendel V: PROSET a fast procedure to create non-redundant sets of protein sequences. Math Comput Mod 16 (6/7): 37–43 (1992).

4. Brendel V, Carle-Urioste JC, Walbot V: Intron recognition in plants. In: Bailey-Serres J, Gallie DR (eds) A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants. American Society for Plant Physiology, Rockville, MD, in press (1997).
5. Brown JWS: A catalogue of splice junction and putative branch point sequences from plant introns. Nucl Acids Res 14: 9549–9559 (1986).
6. Carle-Urioste JC, Ko CH, Benito M-I, Walbot V: *In vivo* analysis of intron processing using splicing-dependent reporter gene assays. Plant Mol Biol 26: 1785–1795 (1994).
7. Carle-Urioste JC, Brendel V, Walbot V: A combinatorial role for exon, intron and splice site sequences in splicing in maize. Plant J 11: 1253–1263 (1997).
8. Csank C, Taylor FM, Martindale DW: Nuclear pre-mRNA introns: analysis and comparison of intron sequences from *Tetrahymena thermophila* and other eukaryotes. Nucl Acids Res 18: 5133–5141 (1990).
9. Egeland DB, Sturtevant AP, Schuler MA: Molecular analysis of dicot and monocot small nuclear RNA populations. Plant Cell 1: 633–643 (1989).
10. Filipowicz W, Gniadkowski M, Klahre U, Liu H-X: Pre-mRNA splicing in plants. In: Lamond A (ed) Pre-mRNA Processing, pp. 65–77. Landes Company, Austin, TX (1995).
11. Foster R, Gasch A, Kay S: Analysis of protein, DNA interactions. In: Koncz C, Chua N-H, Schell J (eds) Methods in *Arabidopsis* Research, pp. 378–392. World Scientific, Singapore (1993).
12. Gniadkowski M, Hemmings-Mieszczak M, Klahre U, Liu H-X, Filipowicz W: Characterization of intronic uridine-rich sequence elements acting as possible targets for nuclear proteins during pre-mRNA splicing in *Nicotiana plumbaginifolia*. Nucl Acids Res 24: 619–627 (1996).
13. Goodall GJ, Filipowicz W: The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. Cell 58: 473–483 (1989).
14. Goodall GJ, Filipowicz W: Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. EMBO J 10: 2635–2644 (1991).
15. Guthrie C: Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein. Science 253: 157–163 (1991).
16. Hanley BA, Schuler MA: Plant intron sequences: evidence for distinct groups of introns. Nucl Acids Res 16: 7159–7176 (1988).
17. Harris NL, Senapathy P: Distribution and consensus of branch point signals in eukaryotic genes: a computerized statistical analysis. Nucl Acids Res 18: 3015–3019 (1990).
18. Hodges PE, Jackson SP, Brown JD, Beggs JD: Extraordinary sequence conservation of the PRP8 splicing factor. Yeast 11: 337–342 (1995).
19. Karlin S, Brendel V: Chance and statistical significance in protein and DNA sequence analysis. Science 257: 39–49 (1992).
20. Karlin S, Brendel V: Patchiness and correlations in DNA sequences. Science 259: 677–680 (1993).
21. Keith B, Chua N-H: Monocot and dicot pre-mRNAs are processed with different efficiencies in transgenic tobacco. EMBO J 5: 2419–2425 (1986).
22. Kulesza H, Simpson GG, Waugh R, Beggs JD, Brown JWS: Detection of a plant protein analogous to the yeast spliceosomal protein, PRP8. FEBS Lett 318: 4–6 (1993).
23. Lazar G, Schaal T, Maniatis T, Goodman HM: Identification of a plant serine-arginine-rich protein similar to the mammalian splicing factor SF2/ASF. Proc Natl Acad Sci USA 92: 7672–7676 (1996).
24. Liu H-X, Filipowicz W: Mapping of branch point nucleotides in mutant pre-mRNAs expressed in plant cells. Plant J 9: 381–389 (1996).
25. Lopato S, Mayeda A, Krainer AR, Barta A: Pre-mRNA splicing in plants: characterization of SR splicing factors. Proc Natl Acad Sci USA 93: 3074–3079 (1996).
26. Lopato S, Waigmann E, Barta A: Characterization of a novel Arginine/Serine-rich splicing factor in *Arabidopsis*. Plant Cell 8: 2255–2264 (1996).
27. Lou H, McCullough AJ, Schuler MA: 3′ splice site selection in dicot plant nuclei is position dependent. Mol Cell Biol 13: 4485–4493 (1993).
28. Lou H, McCullough AJ, Schuler MA: Expression of maize Adh1 intron mutants in tobacco nuclei. Plant J 3: 393–403 (1993).
29. Luehrsen KR, de Wet JR, Walbot V: Transient expression analysis in plants using firefly luciferase reporter gene. Meth Enzymol 216: 397–414 (1992).
30. Luehrsen KR, Taha S, Walbot V: Nuclear pre-mRNA splicing in higher plants. Prog Nucl Acids Res Mol Biol 47: 149–193 (1994).
31. Luehrsen KR, Walbot V: Addition of A- and U-rich sequence increases the splicing efficiency of a deleted form of a maize intron. Plant Mol Biol 24: 449–463 (1994).
32. Luehrsen KR, Walbot V: Intron creation and polyadenylation in maize are directed by AU-rich RNA. Genes Devel 8: 1117–1130 (1994).
33. Nash J: *Bronze-2* gene of maize: analysis of transcription and splicing patterns. Ph.D. Thesis, Stanford University, Stanford, CA (1992).
34. Schbath S, Prum B, Turckheim E de: Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences. J Comp Biol 2: 417–437 (1995).
35. Simpson GG, Clark CP, Rothnie HM, Boelens W, van Venrooij W, Brown JWS: Molecular characterization of the spliceosomal proteins U1A and U2B″ from higher plants. EMBO J 14: 4540–4550 (1995).
36. Simpson CG, Clark G, Davidson D, Smith P, Brown JWS: Mutation of putative branch point consensus sequence in plant introns reduces splicing efficiency. Plant J 9: 369–380 (1996).
37. Simpson GG, Filipowicz W: Splicing of precursors to mRNA in higher plants: mechanism, regulation and sub-nuclear organisation of the spliceosomal machinery. Plant Mol Biol 32: 1–41 (1996).
38. Singh R, Valcárcel J, Green MR: Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science 268: 1173–1176 (1995).
39. Solymosy F, Pollák T: Uridylate-rich small nuclear RNAs (UsnRNAs), their genes and pseudogenes, and UsnRNPs in plants: structure and function. A comparative approach. Crit Rev Plant Sci 12: 275–369 (1993).
40. Umen JG, Guthrie C: A novel role for a U5 snRNP protein in 3′ splice the selection. Genes Devel 9: 855–868 (1995).
41. van Santen SV, Spritz RA: Splicing of plant pre-mRNAs in animal systems and vice versa. Gene 56: 253–265 (1987).
42. White O, Soderlund C, Shanmugan P, Fields C: Information contents and dinuclotide compositions of plant intron sequences vary with evolutionary origin. Plant Mol Biol 9: 1057–1064 (1992).
43. Wiebauer K, Herrero J-J, Filipowicz W: Nuclear pre-mRNA processing in plants: Distinct modes of 3′-splice site selection in plants and animals. Mol Cell Biol 8: 2042–2051 (1988).