J. Mol. Biol. (1998) **276**, 85–104

JMB

Prediction of Splice Sites in Plant Pre-mRNA from Sequence Properties

Volker Brendel[1]*, Jürgen Kleffe[2], Jose C. Carle-Urioste[3] and Virginia Walbot[3]

[1]*Department of Mathematics Stanford University, Stanford CA 94305-2125, USA*

[2]*Freie Universität Berlin Institut für Molekularbiologie und Biochemie, Bereich Molekularbiologie UND Informatik, Arnimallee 22 14195 Berlin, Germany*

[3]*Department of Biological Sciences, Stanford University Stanford, CA 94305-5020, USA*

**Corresponding author*

Heterologous introns are often inaccurately or inefficiently processed in higher plants. The precise features that distinguish the process of pre-mRNA splicing in plants from splicing in yeast and mammals are unclear. One contributing factor is the prominent base compositional contrast between U-rich plant introns and flanking G + C-rich exons. Inclusion of this contrast factor in recently developed statistical methods for splice site prediction from sequence inspection significantly improved prediction accuracy. We applied the prediction tools to re-analyze experimental data on splice site selection and splicing efficiency for native and more than 170 mutated plant introns. In almost all cases, the experimentally determined preferred sites correspond to the highest scoring sites predicted by the model. In native genes, about 90% of splice sites are the locally highest scoring sites within the bounds of the flanking exon and intron. We propose that, in most cases, local context (about 50 bases upstream and downstream from a potential intron end) is sufficient to account for intrinsic splice site strength, and that competition for *trans*-acting factors determines splice site selection *in vivo*. We suggest that computer-aided splice site prediction can be a powerful tool for experimental design and interpretation.

Keywords: intron; splice site prediction; compositional contrast; splice site models

Introduction

Although the basic mechanisms of pre-mRNA splicing appear to be conserved among eukaryotes, there are a number of characteristic differences between plant introns and fungal or animal introns (for recent reviews, see Brown, 1996; Simpson & Filipowicz, 1996; Filipowicz *et al*., 1995; Luehrsen *et al*., 1994). (1) Plant introns are similar in length to the exons; in particular, long introns (>5 kb) as often occur in vertebrates are almost entirely avoided. (2) Plant introns often lack a long pyrimidine tract 5′-proximal to the acceptor site, a feature which is involved in mammalian intron recognition. (3) The branchpoint motif, highly conserved in yeast introns and weakly conserved in vertebrate introns, is only marginally conserved in plant introns. (4) Plant introns are typically U-rich compared with their flanking exons, with a mean difference of about 15 percentage points compensated by increased G + C usage in exons. It has been suggested that the features characteristic of plant pre-mRNAs, including this base compositional contrast between introns and exons, could reflect adaptation to the unique challenge for plants of maintaining accurate splicing in a fluctuating thermal environment (with daily temperature ranges of more than 15°C in temperate plants; Luehrsen *et al*., 1994; Brendel *et al*., 1998).

Plant intron recognition by sequence inspection serves two primary goals. First, genome sequencing efforts produce copious raw sequence data that must initially be analyzed by computational methods. Second, the accuracy of splice site prediction from sequence tests our understanding of the signals sufficient for splice site recognition *in vivo*. Hebsgaard *et al*. (1996) presented a multilayer neural network trained for splice site recognition in *Arabidopsis thaliana*. Kleffe *et al*. (1996) introduced logitlinear models for splice site identification in *Arabidopsis* and maize. In both approaches, inclusion of the plant-specific feature of compo-

Abbreviations used: RT-PCR, reverse transcriptase polymerase chain reaction; wt, wild-type; SD, standard deviation.

0022–2836/98/060085–20 $25.00/0/mb971523

sitional contrast between exons and introns significantly decreased the number of false positive splice site predictions. However, the weakest true splice sites still scored at a level exceeded by many other apparently false positive sites. Thus, in most cases, splice site prediction based on these methods alone would be insufficient to predict entire gene structures accurately (see Burset & Guigó, 1996, for an excellent examination of the current status of gene prediction methods).

Here we pursue in more detail the predictions and limitations of the logitlinear splice site recognition models by Kleffe *et al*. (1996). These models were formulated as rules for predicting potential splice sites from the local sequence context only (encompassing 50 bases on each side of a GU dinucleotide for a potential donor site or an AG for a potential acceptor site). In particular, unlike other methods (including that of Hebsgaard *et al*., 1996; also, Solovyev *et al*., 1994, Zhang, 1997), this approach does not in any way include scoring for the triplet code signal in the coding regions. Recognition of codons by the nuclear splicing factors is unlikely because introns in untranslated regions of pre-mRNA are also accurately spliced, and there is no evidence for special splicing mechanisms for such introns. In the nucleus, competition among alternative splice sites for *trans*-acting splicing factors will play a role in splice site selection. We examine published experiments reporting alternative splice site selection and altered splicing efficiencies after sequence changes in terms of predictions derived from the splice site models. Furthermore, we discuss the extent to which true splice sites coincide with locally optimal predicted sites, and we offer a perspective on possible improvements of splice site prediction.

Results

Prediction of splice sites in maize and *Arabidopsis* genes by the method of Kleffe *et al*. (1996) is based on the parameters displayed in Tables 1 and 2. Appropriate parameter insertions into equations (2) and (1) (below) define a *P*-value for any potential splice site in a sequence. The *P*-values measure predicted intrinsic splice site strength from the combination of site quality (degree of matching to the donor or acceptor site consensus) and local compositional contrast (Table 3). The basic properties and implications of the underlying model are examined in Materials and Methods. In this section we analyze (1) the distribution of splice site scores for true and false sites in known genes, (2) the correlation of splice site scores with experimentally determined splicing efficiency, (3) model predic-

Table 1. Donor site model parameters

A. *G/GU sites* Parameter	Maize (172 sites)				*Arabidopsis* (458 sites)			
α			3.36				3.46	
δ			−8.29				−9.71	
μ			10.11				13.69	
	A	C	G	U	A	C	G	U
l_{-3}	0	−0.05	−1.17	−1.64	0	0.32	−1.20	−0.80
l_{-2}	0	−2.53	−2.91	−2.16	0	−3.21	−3.05	−2.74
l_{-1}	0	0	0	0	0	0	0	0
l_1	0	0	0	0	0	0	0	0
l_2	0	0	0	0	0	0	0	0
l_3	0	−3.38	−2.39	−3.76	0	−3.06	−2.99	−2.87
l_4	0	−0.66	−1.65	−1.50	0	−1.08	−2.11	−1.78
l_5	−2.43	−2.92	0	−2.50	−2.53	−2.82	0	−2.72
l_6	−1.26	−0.82	−1.78	0	−1.35	−0.82	−1.80	0

B. *H/GU sites* Parameter	Maize (29 sites)				*Arabidopsis* (119 sites)			
α			0.17				0.37	
δ			−17.61				−8.87	
μ			6.53				13.09	
	A	C	G	U	A	C	G	U
l_{-3}	0	−0.39	0.36	0.64	0	0.20	−0.79	−0.91
l_{-2}	0	−0.92	−2.72	−0.91	0	−0.71	−1.29	−1.26
l_{-1}	−0.45	−0.24	0	0	−0.17	−0.11	0	0
l_1	0	0	0	0	0	0	0	0
l_2	0	0	0	0	0	0	0	0
l_3	0	−3.93	−2.67	−11.61	0	−3.18	−2.67	−5.21
l_4	0	−3.92	−11.43	−4.89	0	−2.13	−4.52	−2.71
l_5	−4.86	−4.42	0	−4.04	−3.98	−5.56	0	−3.73
l_6	−2.65	−1.20	−2.57	0	−1.17	−0.90	−1.54	0

H denotes non-G. Parameters are set to 0 for residues that are prescribed by the model (including positions 1 and 2 occupied by the requisite GU and position −1 for G/GU sites) and for the consensus donor site residues. The donor site score of a given site is calculated according to equations (1) and (2) of the text.

Table 2. Acceptor site model parameters

A. *CAG/sites*

Parameter	Maize (157 sites)				*Arabidopsis* (387 sites)			
α		3.60				5.20		
δ		12.86				11.32		
μ		−10.84				−11.43		
	A	C	G	U	A	C	G	U
l_{-15}	−0.59	−0.48	−1.18	0	−0.25	−0.43	−0.45	0
l_{-14}	−0.47	−0.09	−0.28	0	−0.73	−0.38	−0.75	0
l_{-13}	−1.45	−0.14	−0.33	0	−0.87	−0.83	−1.01	0
l_{-12}	−1.11	0.03	−0.15	0	−1.02	−1.33	−1.00	0
l_{-11}	−1.23	−0.34	−0.42	0	−0.78	−1.53	−0.78	0
l_{-10}	−0.38	−0.08	−0.47	0	−0.65	−0.97	−1.32	0
l_{-9}	−0.49	0.26	−0.48	0	−1.24	−0.94	−0.99	0
l_{-8}	−0.74	0.01	−0.43	0	−0.99	−1.32	−1.11	0
l_{-7}	−1.15	−0.84	−1.35	0	−0.36	−1.43	−1.00	0
l_{-6}	−0.75	−0.95	−1.07	0	−1.06	−1.39	−0.98	0
l_{-5}	−1.48	−0.81	−1.36	0	−0.68	−1.28	−1.20	0
l_{-4}	−0.73	−1.89	0	−2.10	−0.33	−1.91	0	−1.49
l_{-3}	0	0	0	0	0	0	0	0
l_{-2}	0	0	0	0	0	0	0	0
l_{-1}	0	0	0	0	0	0	0	0
l_1	−1.26	−2.87	0	−2.79	−1.68	−2.87	0	−2.38
l_2	−1.76	−1.54	−1.56	0	−1.09	−1.05	−1.17	0

B. *DAG/sites*

Parameter	Maize (47 sites)				*Arabidopsis* (187 sites)			
α		−2.60				2.73		
δ		28.64				10.74		
μ		−11.47				−13.85		
	A	C	G	U	A	C	G	U
l_{-15}	−1.74	−0.38	−1.50	0	−1.09	−0.64	−0.40	0
l_{-14}	−1.30	−0.28	−1.65	0	−0.95	−0.02	−0.79	0
l_{-13}	−1.63	0.35	1.68	0	−1.32	−1.24	−0.48	0
l_{-12}	0.11	0.31	1.45	0	−1.08	−0.79	−1.07	0
l_{-11}	−0.45	−0.37	−1.96	0	−0.44	−0.12	−0.60	0
l_{-10}	−0.35	0.70	−0.97	0	−0.59	−0.26	−0.25	0
l_{-9}	−1.03	1.91	−0.07	0	−0.41	−1.01	−0.42	0
l_{-8}	0.10	−0.65	−0.91	0	−0.57	−0.92	−0.43	0
l_{-7}	−0.69	−1.46	−0.62	0	−1.41	−0.84	−0.84	0
l_{-6}	−0.95	−0.02	−0.35	0	−0.71	−0.93	−0.43	0
l_{-5}	−2.68	−0.42	−1.14	0	−0.83	−1.34	−1.59	0
l_{-4}	−0.55	−0.89	0	0.46	−0.63	−1.70	0	−1.58
l_{-3}	−1.93	0	−13.02	0	−2.61	0	−5.90	0
l_{-2}	0	0	0	0	0	0	0	0
l_{-1}	0	0	0	0	0	0	0	0
l_1	−3.16	−3.20	0	−5.89	−2.43	−3.18	0	−3.36
l_2	−2.61	0.20	−2.27	0	−2.07	−1.13	−1.00	0

D denotes non-C. Parameters are set to 0 for residues that are prescribed by the model (including positions −2 and −1 occupied by the requisite AG and position −3 for CAG/sites) and for the consensus acceptor site residues. The acceptor site score of a given site is calculated according to equations (1) and (2) of the text.

tions for processing of heterologous introns, (4) prediction and analysis of splicing patterns for experimentally well-studied native and mutant introns, and (5) intron recognition in pre-mRNA.

Distribution of splice site scores

Table 4 establishes typical and extreme values for the splice site variable L (shifted by the corresponding α value so that site subclasses could be properly combined), the contrast variables X_U and X_{GC}, and the predicted splice site strength P. L, X_U, and X_{GC} have bell-shaped distributions (not shown) with the indicated standard deviations. The mean values of X_U and X_{GC} are in the range of values (13% to 17%) observed in comparisons with entire exons and introns (cf. Table I of Brendel *et al.*, (1998), or Table IV of Sinibaldi & Mettler, (1992)), indicating that the choice of 50 base windows for calculating the compositional contrast does not introduce special biases as a result of the particular window size. The predicted splice site strength P displays a wide range, exceptionally weak sites scoring close to zero. Note, however, that the sigmoidal transformation of θ onto the [0, 1] interval by equation (1) introduces long tails of values close to 0 or close to 1. The vast majority of potential splice sites score lower than the worst scoring true sites (i.e. they score very close to zero; Kleffe *et al.*, 1996).

Table 3. Combinatorial effect of splice site quality and compositional contrast on predicted splicining efficiency (maize models)

		$X_U = -0.230$			$\mathbf{X_U = -0.135}$			$X_U = -0.040$		
	X_{GC}=	0.23	0.13	0.03	0.23	**0.13**	0.03	0.23	0.13	0.03
A. *Donor site*	*L*									
AAG/GUUGCC	−9.15	0.17	0.07	0.03	0.09	**0.03**	0.01	0.04	0.02	0.01
CAG/GUAUCA	**−5.72**	**0.87**	**0.70**	**0.46**	**0.75**	**0.52**	**0.28**	**0.57**	**0.33**	**0.15**
CUG/GUAAGA	−3.46	0.98	0.96	0.89	0.97	**0.91**	0.79	0.93	0.82	0.63
CAG/GUAUGU	−1.55	1.00	0.99	0.98	1.00	**0.99**	0.96	0.99	0.97	0.92
AAG/GUACGU	−0.66	1.00	1.00	0.99	1.00	**0.99**	0.98	1.00	0.99	0.97
		$X_U = 25.0$			$\mathbf{X_U = 0.160}$			$X_U = 0.070$		
	X_{GC}=	0.215	−0.125	−0.03	0.215	**−0.125**	−0.03	0.215	−0.125	−0.03
B. *Acceptor site*	*L*									
UUCACGUACUACCAG/AC	−9.81	0.34	0.16	0.07	0.14	**0.06**	0.02	0.05	0.02	0.01
UAUUUGCUGGCGCAG/AA	**−6.93**	**0.90**	**0.78**	**0.57**	**0.74**	**0.52**	**0.29**	**0.48**	**0.25**	**0.11**
CUUGGGUGUUCACAG/AU	−4.74	0.99	0.97	0.92	0.96	**0.91**	0.79	0.89	0.75	0.54
UCUUAUUUGUUGCAG/GU	−2.67	1.00	1.00	0.99	1.00	**0.99**	0.97	0.99	0.96	0.90
AUCUUUCUCUUGCAG/GU	−1.31	1.00	1.00	1.00	1.00	**1.00**	0.99	1.00	0.99	0.97

Row and columns representing typical splice site values are shown in bold face.

Figure 1 illustrates the predictive power of the *P*-values. Shown are the distribution of *P*-values for all true maize donor sites and the distribution of *P*-values for all falsely predicted sites in the same genes (prediction based on a minimal score set to the minimal score of all true sites). Relative to the uniform distribution, *P*-values for the real sites are shifted towards higher values and *P*-values for the more numerous false sites are shifted to lower values (Figure 1). Thus, increasing the threshold for prediction rapidly decreases the number of false predicted sites while missing only a few true sites. For example, setting a threshold of $P = 0.25$ includes 162 of the 201 true maize donor sites while giving only 83 false positive predictions. At a threshold of $P = 0.005$, which includes all true donor sites, there are more than 500 false positive predictions. Unfortunately, in applications such as gene prediction, one cannot afford even a small number of missed true sites without affecting the entire gene parse. Similar considerations apply to maize acceptor sites and *Arabidopsis* splice sites (Kleffe *et al*., 1996).

Using a new set of 17 maize gene sequences with 98 introns deposited into GenBank more recently than those of our training set, five donor sites and three acceptor sites were missed at the cutoff that includes all sites of the training set. Thus, on the new set, the models have a reduced sensitivity of 95% for donor sites and 97% for acceptor sites. Specificity was 31% for donors and 14% for acceptors (data not shown), compared with 28% and 13% for the training set (Kleffe *et al*., 1996). It is clear that (for any statistic) as more data become available, more extremes will be found and models based on previous data will prove inadequate in these cases. On the other hand, a quantitative model seems very valuable to point out the extremes and, in some cases, may suggest re-examination of the data, particularly for possible sequencing and annotation errors.

Table 4. Average and extreme splice site variable scores

		Donor site scores		Acceptor site scores	
		Maize	*Arabidopsis*	Maize	*Arabidopsis*
$\alpha + L$[a]	Mean	−1.528	−1.830	−3.215	−2.438
	SD	1.906	1.936	2.777	2.134
	Minimum	−12.430	−8.380	−12.910	−11.040
	Maximum	3.310	3.780	2.290	3.520
X_U	Mean	−0.134	−0.137	0.160	0.175
	SD	0.097	0.101	0.092	0.097
	Minimum	−0.420	−0.420	−0.080	−0.160
	Maximum	0.160	0.200	0.480	0.440
X_{GC}	Mean	0.131	0.138	−0.126	−0.130
	SD	0.098	0.090	0.092	0.084
	Minimum	−0.220	−0.140	−0.360	−0.460
	Maximum	0.420	0.400	0.100	0.100
P	Mean	0.647	0.671	0.617	0.628
	SD	0.315	0.323	0.328	0.322
	Minimum	0.005	0.003	0.001	0.001
	Maximum	0.999	1.000	0.999	1.000

[a] The constant α was added so that the site subclasses could be combined.

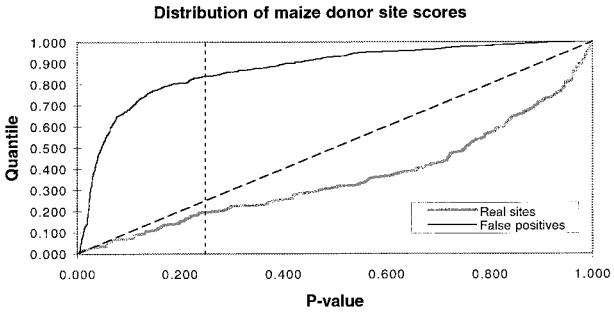

Figure 1. *P*-values were calculated for all true and predicted donor sites in a collection of 46 maize genes (201 true sites and 511 false positive predictions by the SplicePredictor program of Kleffe *et al*. (1996) set at 100% sensitivity level). The broken diagonal line corresponds to uniform distribution of *P*-values. The broken vertical line represents a cutoff for prediction set at 0.25, which would miss about 20% of the true sites but also exclude more than 83% of the false positive predictions made at the 100% sensitivity level.

As shown in Table 3, the model according to equation (2) entails that the splice site strength P is determined as an additive effect of the splice site quality L and the compositional contrast measures X_U and X_{GC} (which are largely independent of L). To test whether weakness in a particular variable can be compensated by better values in the other components, we calculated the mean variable values for different subgroups of true sites. Thus, the maize H/GU donor sites have mean values $X_U = -0.159$ and $X_{GC} = 0.145$, indicating increased compositional biases compared to the G/GU sites, which have mean values $X_U = -0.130$ and $X_{GC} = 0.129$. The *Arabidopsis* donor site subgroups are not distinguished by average X_U and X_{GC} values. More pronounced differences are seen for the acceptor sites: maize DAG/sites show mean $X_U = 0.205$ and $X_{GC} = -0.144$ compared with $X_U = 0.147$ and $X_{GC} = -0.120$ for CAG/sites, a combined increase of compositional bias of 0.082. Similarly, *Arabidopsis* DAG/sites have mean $X_U = 0.198$ and $X_{GC} = -0.153$ compared with $X_U = 0.164$ and $X_{GC} = -0.118$ for CAG/sites, a combined increase of compositional bias of 0.069. X_U and X_{GC} are negatively correlated: an increase in intron U bias ($-X_U$ at donor sites, X_U at acceptor sites) is almost certainly associated with an increase in exon G + C bias (X_{GC} at donor sites, $-X_{GC}$ at acceptor sites). Still, the mean value of X_{GC} for the 20% worst X_U donors is 0.057 for maize and 0.103 for *Arabidopsis*, and for the 20% worst X_U acceptors the values are -0.055 for maize and -0.092 for *Arabidopsis*, maintaining the sign of the compensating bias even if at reduced magnitude. The combined contrast measure $X_{GC} - X_U$ for donor sites and $X_U - X_{GC}$ for acceptor sites is lower than -0.02 in fewer than 2% of all maize and *Arabidopsis* splice sites. These results are consistent with the hypothesis of a combinatorial effect of splice site quality and compositional contrast (Carle-Urioste *et al*., 1994, 1997; Baynton *et al*., 1996).

Correlation of splice site scores with experimental measures of splicing efficiency

Quantitative comparisons of experimentally determined splicing efficiencies for different introns are problematic for reasons discussed in Materials and Methods. To evaluate the predictive power of calculated *P*-values for determination of splicing efficiency we therefore focus first on the single, particularly well-studied example of the maize *Bronze2* (*Bz2*) intron (78 bases) and 27 mutant constructs (Carle-Urioste *et al*., 1994, 1997). The mutant constructs involve sequence changes within both the intron and the flanking exons that affected, in various combinations, residues at the 5′ splice site (G to the more favorable A in the +3 position; cf. Table 1), residues at the 3′ splice site (U to the more favorable G in the +1 position; *cf*. Table 2), and the compositional contrast. We define as a score for the entire intron the product of donor and acceptor *P*-values. This simple definition merely gives some measure that incorporates the expectation that changes in either splice site should affect the splicing efficiency of the intron. The correspondence between the splicing efficiency of the various constructs and the calculated intron score is shown in Figure 2. The high correlation (0.75) suggests that the models capture *in vivo* splice site strength adequately. The correlation is even higher (0.81 to 0.88) for subsets of constructs involving changes at only one or the other splice site or affecting only compositional contrast (data not shown).

Differences between splicing in monocots and dicots

Splicing in monocots is considered to be more permissive than splicing in dicots. For example,

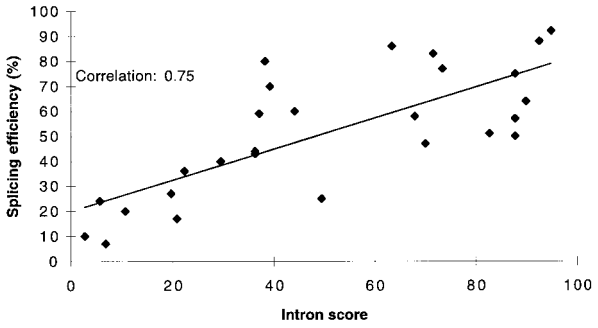

Figure 2. Prediction of splicing efficiency from splice site scores. Splicing efficiencies were taken from Carle-Urioste *et al*. (1997) for different mutants of the maize *Bronze2* intron. The intron score was calculated as the product of predicted donor and acceptor *P*-values (multiplied by 100 to simplify presentation). The continuous line was obtained by linear regression.

G + C-rich introns and synthetic introns containing stem-loop structures are efficiently spliced in maize but not in *Nicotiana plumbaginifolia* protoplasts (Goodall & Filipowicz, 1991). We determined to what extent these differences in splicing correspond to differences in splice site predictions derived from the maize as compared with the *Arabidopsis* models. Here as well as later we assume that the *Arabidopsis*-based models are representative of all dicots. This assumption remains tentative until sufficient sequences of other dicots are available as independent training sets for parameter estimations. For the available data (Goodall & Filipowicz, 1991) the results are varied. Maize *waxy* introns 9 and 10 were shown to be efficiently spliced in maize but not in *Nicotiana* protoplasts. The maize models give donor/acceptor scores 0.42/0.48 for intron 9 and 0.56/0.52 for intron 10 compared with the reduced values 0.26/0.24 and 0.40/0.43 with the *Arabidopsis* models. The synthetic 73% G + C intron syn24 was spliced in maize but not in *Nicotiana*. Consistently, whereas the maize model gives an acceptor site score above the minimal threshold, the *Arabidopsis* model score of less than 0.0005 predicts non-usage of this site. On the other hand, bean phaseolin intron 2 was efficiently spliced in both maize and *Nicotiana*, but the maize model assigns the low scores of 0.005 and 0.25 for donor and acceptor site, respectively. A tenfold better scoring donor site occurs 12 bases downstream of the assigned donor site, usage of which would insert a VLPF tetrapeptide into the translation product. Selection of this alternative site may not have been distinguished by the RNase protection analysis used (Goodall & Filipowicz, 1991).

Based on the few examples studied, it was proposed that A + U richness of introns is an essential feature of splicing in dicots but not of splicing in monocots; in monocots, A + U richness might compensate for suboptimal splice sites (Goodall & Filipowicz, 1991). In our databases, 18% of maize introns have G + C content exceeding 50% compared with fewer than 4% of *Arabidopsis* introns exceeding even the lower value of 40%. As previously reported (Carle-Urioste *et al*., 1997; Brendel *et al*., 1998), base compositional contrast is equally stringently maintained in both species; in particular, the G + C-rich maize introns are typically flanked by even more G + C-rich exons (see also Table 4: X_U and X_{GC} averages are very close comparing maize and *Arabidopsis*). Application of the heterologous models to the entire maize and *Arabidopsis* gene sets gave average splice site *P*-values that differed at most by 0.11 in the case of maize acceptors scored by the *Arabidopsis* model (average 0.51 compared with 0.62 with the maize model). Interestingly, the differences in the *waxy* introns 9 and 10 donor site scores for maize versus *Arabidopsis* are among the largest differences observed: only six of the 201 maize donor sites in our collection had a larger drop in *P*-value when scored with the dicot rather than the maize model.

Further evidence for a role of base composition in splice site selection in dicots derives from studies of the expression of maize *Adh1* intron mutants in tobacco nuclei (Lou *et al*., 1993a). Several alternative 3′ splice site choices were observed for intron 3 (see below). Splice site choice was shown to be mediated by base compositional context, which was manipulated by means of deletion mutations and fusion constructs of introns 1 and 3. Our model predictions are consistent with the experimental results. In all constructs the unique 5′ site scores above 0.98, and there are no other high scoring alternatives. The native 3′ site scores are 0.31 for intron 1 and 0.66 for intron 3; these values are not exceeded within the respective introns. A site at position −105 relative to the native 3′ end of intron 3 was used as the major alternative acceptor site in constructs incorporating the wild-type intron. The splicing efficiency at the −105 site was 13% relative to 69% at the usual −1 site. In this context, the −105 site is the second highest scoring acceptor site with a score of 0.47. In fusion constructs linking N-terminal sequences of intron 3 with C-terminal sequences of intron 1 the −105 site scores 0.41 and was used with 22% efficiency. A 107 base deletion at the C terminus of intron 3 eliminated the usual −1 acceptor site. In this case, the −105 site score improves to 0.53, and this site was used with 66% efficiency.

Intron 1 was spliced with only 28% efficiency. Mutation of the 5′ splice site to perfectly match the consensus resulted in improved splicing efficiency (74%). The improved site has $L = 0.00$ and $P = 1.00$ compared with $L = -4.96$ and $P = 0.99$ for the native site. Unlike the *L*-values, the *P*-values do not seem to differ much. However, it is important to note the non-linearity of the transformation equation (1), which primarily emphasizes differences in the mid-range (see Discussion).

Splicing of genes of animal origin in plant cells

Assays of intron splicing in heterologous species suggested that, whereas plant introns are usually faithfully spliced in mammalian cells, mammalian introns are inefficiently and inaccurately spliced in plant cells (reviewed by Luehrsen *et al*., 1994). The only established case of processing of a mammalian genomic intron in plant cells is intron 2 of the human β-globin gene, which is spliced at an alternative 3′ splice site in tobacco and *Orychophragmus violaceus* protoplasts (Wiebauer *et al*., 1988). Sequence inspection of this gene with the *Arabidopsis* splice site models shows that both β-globin introns have plant-compatible donor sites (scoring 0.82 and 0.84, respectively) and poorly scoring acceptor sites (0.07 and 0.09, respectively). The only high scoring acceptor site (0.95) is the site used in plant cells.

The failure of splicing of mammalian introns in plant cells is generally attributed to the lack of base compositional contrast at the intron/exon borders that have been examined. The only heter-

ologous intron that is known to be accurately spliced in plants (tobacco) is the 66 base A + U-rich intron of the primate virus SV40 t-antigen mRNA (Hunt *et al*., 1991). Our dicot models give a high donor score of 0.995 (with $X_U = -0.16$ and $X_{CG} = 0.20$) paired with a high acceptor score of 0.85 ($X_U = 0.26$, $X_{CG} = -0.22$), predicting that this viral intron would be correctly recognized in tobacco.

Analysis of intron mutation and creation experiments

For a number of genes, alternative splicing patterns have been observed in response to substitution, insertion, or deletion mutations introduced into the native intron. For each gene, we discuss the correspondence of our splice site predictions with the observed splicing patterns (Tables 5 to 10). With very few exceptions, the experimentally determined preferred sites correspond to the highest scoring sites predicted by the model. Examination of the exceptional cases showed that these frequently involved special experimental circumstances, notably low overall transcript levels. This suggests that our SplicePredictor program can be a powerful tool for the interpretation and design of plant pre-mRNA processing experiments.

Insertions of long intron and non-intron sequences into maize introns

Luehrsen & Walbot (1992) investigated the impact on splicing of insertions of intron and non-intron sequences into maize *Adh1-S* intron 1 and actin intron 3 in a series of 17 constructs. Their control plasmid pAL61 bears the full-length *Adh1-S* intron 1 (534 bases) flanked upstream by the native *Adh1-S* exon 1 and downstream by five bases of *Adh1-S* exon 2 translationally fused with the firely luciferase gene. The intron was spliced from this construct with 60 to 70% efficiency. Its donor site AAG/GUCCGC scores the high value 0.96 whereas the acceptor site CCUGGACCCGUG-CAG/CU scores a moderate value of 0.55, less than in the native *Adh1-S* context (0.93). There are no higher scoring donor or acceptor sites nearby, and thus the predicted splicing is consistent with the observed transcripts. Insertions of actin intron 3 sense or anti-sense sequences into the *Adh1-S* intron were not detrimental to splicing. There is one high scoring (0.72) potential acceptor site of the DAG/class within the actin intron 3 sense strand (the high score resulting from highly favorable contrast, $X_U = 0.22$, $X_{GC} = -0.24$; this site also occurs in the construct pALXI31 given Table 5). Apparently, this site is not used. It is possible that the predominant CAG/type acceptors are preferentially selected even when compositional contrast favors an alternative DAG/acceptor.

Insertion of *Adh1-S* cDNA fragments in either orientation into *Adh1-S* intron 1 resulted in decreased splicing efficiency for the longer introns. Sequence inspection reveals two potential alternative 3′ splice sites. The cDNA sense fragments contain a site scoring 0.55. In its native context, this site occurs towards the end of *Adh1-S* exon 5 and scores insignificantly as a result of unfavorable compositional context ($X_U = -0.02$, $X_{GC} = -0.02$); although exon 5 is U-rich (32.5%), the downstream intron is even more U-rich (37.2%), resulting in negative X_U value. In the cDNA context, however, the site is put upstream of exon 6 (only 18.4% U), creating the favorable context responsible for the

Table 5. Splice site analysis in maize *Adh1-S* intron 1 mutant constructs

Construct[a]	Intron length	Acceptor site					Splicing efficiency (%)[a,b]	Spliced *versus* pAL61[a,c]
		Sequence	*L*	X_U	X_{GC}	*P*		
pAL61	534	CCTGGACCCGTGCAGCT	−6.03	0.12	−0.10	0.550	57	+ + + + +
pAL6103	244	CCTGGACCCGTGCAGCT	−6.03	0.12	−0.10	0.550	55	+ + + + +
pAL6116	136	CCTGGACCCGTGCAGCT	−6.03	0.12	−0.10	0.550	63	+ + + + +
pAL6113	104	CCTGGACCCGTGCAGCT	−6.03	0.10	−0.06	0.379	15–17	++
pAL6110	57	CCTGGACCCGTGCAGCT	−6.03	0.12	−0.10	0.550	9	
pALT3	103	AATTTTTATTTTTAGCT	−5.69	0.32	−0.44	0.997	43	+ + + + +
	132	CCTGGACCCGTGCAGCT	−6.03	0.26	−0.22	0.964	12	++
pALA1	132	CCTGGACCCGTGCAGCT	−6.03	0.04	−0.22	0.616	35	+ + +
pALC6	103	CGGGCCCCCGCCCAGCT	−8.88	−0.04	0.02	0.002		+ + +
	132	CCTGGACCCGTGCAGCT	−6.03	−0.06	0.20	0.005		Site not used
pALG3	132	CCTGGACCCGTGCAGCT	−6.03	−0.06	0.18	0.006		Site not used
pALX1	104	TTGGTTTTTTTTCAGCT	−5.45	0.28	−0.24	0.987		++
pALXT3	103	AATTTTTATTTTTAGCT	−5.69	0.14	−0.26	0.214		Site not used
	132	TTGGTTTTTTTTCAGCT	−5.45	0.44	−0.40	1.000		+ + + +
pALXA5	132	TTGGTTTTTTTTCAGCT	−5.45	0.22	−0.40	0.995		+ + + +
pALXI14	420	TTTCTCGGACGTAAGGC	−3.10	0.00	−0.06	0.007		+ + +
	454	TTGGTTTTTTTTCAGCT	−5.45	0.24	−0.18	0.960		++
pALXI31	249	TCTTCTGTTCTAAAGGG	−5.49	0.22	−0.24	0.723		Site not used
	414	TTGGTTTTTTTTCAGCT	−5.45	0.22	−0.30	0.986		+ + + + +

[a] From Luehrsen & Walbot (1994a).
[b] Method: RNase protection analysis.
[c] Method: RT-PCR analysis; + symbols show splicing efficiency relative to the control plasmid pAL61.

high score. Another high-scoring potential acceptor site is created by the cDNA insertions at the *Adh1-S* intron 1 *Stu*I AGG|CCT site (0.70 for the sense insertion, 0.72 for the antisense insertion). The insertions provide a favorable downstream context, which dramatically raises the score at this site from its value of 0.03 in pAL61. Insertion of *Adh1-S* cDNA fragments into actin intron 3 resulted in reduced transcript abundance but not in reduced splicing efficiency at the native splice sites. These results could be explained by decreased stability of alternatively spliced transcripts.

Insertions of A + T-rich fragments of bacteriophage λ DNA also proved detrimental to luciferase expression. Our analysis reveals three high scoring potential acceptor sites (0.996, 0.86 and 0.93) in this fragment. Splicing at any of these sites (not determined) would lead to multiple in-frame stop codons that would explain the decreased expression.

Maize Adh1-S *intron 1*

Luehrsen & Walbot (1994a) derived 14 variants of maize *Adh1-S* intron 1 of different lengths and base composition. Splicing efficiency and 3′ splice site selection were shown to depend on the nature of the insertions. In all spliced transcripts the 5′ splice site was the native donor site (see above). Selected acceptor sites and their predicted scores are displayed in Table 5. pAL61 has the full-length *Adh1-S* intron 1, which was correctly spliced with 57% efficiency. Internal deletions of moderate size did not affect splicing efficiency (pAL6103 and pAL6116). The context of the 3′ splice site is changed only for longer deletions. pAL6113 has unfavorable 3′ context (acceptor site score reduced from 0.550 to 0.379) and is spliced with only 15% efficiency. The acceptor site of pAL6110 is in the same 50 base context as in the native intron (pAL61). The very short length of the predicted intron (57 bases, shorter than for any native intron in our collection; Table 11) is likely responsible for the poor splicing.

Derivatives of pAL6113 with inserts of A, C, G, or T-rich oligomers displayed different splicing patterns. The T-rich insert in pALT3 created a new acceptor splice site that was used preferentially over the native site. Both sites score very high, the new site near perfect (0.997), a bit higher than the native site (0.964). The A-rich insert (pALA1) also improved splicing efficiency, which according to our model might result from increased G + C contrast relative to the downstream exon. The C and G-rich inserts abolished splicing at the native acceptor site, consistent with very low scores. The C-rich insert in pALC6 created a new site corresponding to a 103 base intron. This site scores very poorly in all respects, and we cannot explain its usage with our model. Note, however, that transcript levels for the vectors with C and G-rich inserts were dramatically lower than for the other constructs (Luehrsen & Walbot, 1994a), and interpretation of the splicing pattern therefore must remain tentative.

Replacement of residues upstream of the pAL6113 acceptor site with T residues raised the site score to 0.987 and increased splicing efficiency (pALX1). Insertion of the T-rich oligomer into this construct (pALXT3) created the upstream site that was used preferentially in pALT3. In this vector, however, this upstream site was not used. These results can be explained by compositional contrast: whereas in pALT3 the site is in favorable context and scores 0.997, in pALXT3 the context is much less favorable as a result of the improved context for the nearby downstream native site, and the score drops to 0.214.

Surprisingly, insertion of *Adh1-S* intron 1 internal sequences (pALXI14) resulted in splicing at a site upstream of the improved native acceptor site slightly more frequently than at the native site. This new site scores very poorly in our model, and the model cannot explain why predominant splicing at the high scoring native site 34 bases downstream is not maintained (see also Luehrsen & Walbot (1994a) for a discussion of this site as exceptional). By contrast, a high scoring intron-internal site in pALXI31 was not used, and all splicing occurred at the improved native acceptor site.

Maize Adh1-S *intron 3*

Splicing patterns of the maize *Adh1* intron 3 with several alternative 3′ splice site choices were studied by Lou *et al*. (1993b). Splicing occurred with different efficiencies from the native 5′ site (score 0.94) to more than five acceptor sites within the intron. Observed splice site usage and splice site scores for the series of 30 constructs are given in Table 6. In 19 of 25 constructs with alternative 3′ sites, splicing occurred predominantly at the highest scoring site. The most notable difference between prediction based on scores and observed splicing efficiencies occurs in construct 3.Δ77, in which 27 bases of the intron and 50 bases of the downstream exon were replaced by a CAG trinucleotide. The trinucleotide insertion creates a high-scoring potential acceptor site (0.97) 80 bases downstream from the reportedly used −105 site. Base substitutions in construct AT3 + 4 reduced U richness in the 5′ proximal segments of sites −1 and −105, but also improved the base usage contrast at site −139 by increasing the G + C content in its downstream flanking 50 base window. Although scoring 0.82, this site is inefficiently used. In contrast, increased G + C content in the downstream flanking sequences improved splicing efficiency of the maize *Bz2* intron (Carle-Urioste *et al*., 1997). The generalization implied in the logit-linear model by inclusion of the X_{GC} variable may be too simplistic: some specificity for particular G + C-rich motifs (or avoidance of others) rather than pure G + C content *per se* may be required for enhanced site recognition (Carle-Urioste *et al*., 1997).

Table 6. 3′ splice site selection in maize *Adh1-S* intron 3 mutant constructs

Site	−285		−199		−139		−105		−1		CAG[c]	
Construct[a]	*P*	*%spl*[b]	*P*	*%spl*	*P*	*%spl*	*P*	*%spl*	*P*	*%spl*	*P*	*%spl*
Wild-type	0.04	**4**	0.01	5	0.31		0.47	11	**0.66**	**57**	–	–
3.Δ96	0.04		0.01		0.31		**0.47**	**70**	–	–	–	–
3.Δ133	0.04		0.01		0.31		**0.78**	**74**	–	–	–	–
3.Δ145	0.04	1	0.01		0.23		**0.87**	**74**	–	–	–	–
3.Δ199	**0.04**	**44**	0.01	11	–	–	–	–	–	–	–	–
3.Δ241	0.04	**45**	**0.18**	22	–	–	–	–	–	–	–	–
3.Δ315	0.36	63	–	–	–	–	–	–	–	–	–	–
3.Δ384[d]	–	–	–	–	–	–	–	–	–	–	–	–
3.Δ101*[e]	0.04	4	0.01	5	0.31	6	**0.47**	9	0.29	**37**	–	–
3.Δ50	0.04	5	0.01	6	0.31		0.47	7	**0.60**	**63**	0.56[f]	
3.Δ77	0.04	2	0.01		0.31	6	0.47	**61**	–	–	**0.97**	
3.Δ179	0.04	27	0.01		**0.96**	**42**			–	–	0.97	
3.Δ215	0.04	**42**	**0.08**	12	–	–	–	–	–	–	0.17	
3.Δ248	0.04	**60**	**0.08**	5	–	–	–	–	–	–	0.38[f]	
3.Δ306	**0.24**	**57**	–	–	–	–	–	–	–	–	0.20	
3.Δ334	0.17	61	–	–	–	–	–	–	–	–	0.98[f]	
3.Δ344	–	–	–	–	–	–	–	–	–	–	0.96	63
3.Δ354	–	–	–	–	–	–	–	–	–	–	0.69	56
3.Δ379	–	–	–	–	–	–	–	–	–	–	0.98	64
AT1	0.04	3	0.01		0.31		**0.69**	**61**	0.56	14	–	–
AT2	0.04	5	0.01	4	0.31		**0.47**	**37**	0.29	33	–	–
AT3	0.04	7	0.01		0.31	6	**0.47**	**43**	0.01	13	–	–
AT1 + 2	0.04	7	0.01		0.31	5	**0.69**	**47**	0.20	6	–	–
AT2 + 3	0.04	7	0.01		0.31	5	**0.47**	**51**	0	3	–	–
AT1 + 3	0.04	6	0.01		0.31		**0.69**	**66**	0	1	–	–
AT3 + 4	0.04	25	0.01	6	**0.82**	9	0	11	0.01	7	–	–
AT2 + 10.A	0.04	6	0.01	6	0.31		0.47	15	**0.83**	**52**	–	–
AT2 + 10.T	0.04	6	0.01	6	0.31		0.47	14	**0.86**	**53**	–	–
AT20	0.04	13	0.01	8	0.31		**0.47**	**36**	0.11	9	–	–
−105. − 285	**0.36**	**48**	0.01		0.23		0.02		–	–	–	–
−285. − 1. − 105	0.03	8	0.01		0.23		**0.95**	**68**	0.03		–	–

[a] From Lou *et al*. (1993b).
[b] Blank cells correspond to no detectable levels of splicing (spl). Sites missing from a construct are indicated by –. In case of alternative sites, the highest scoring sites and the higher efficiencies are indicated in bold face. Method: RT-PCR analysis.
[c] The CAG trinucleotide was inserted into the indicated constructs at the beginning of the downstream exon.
[d] No splicing occurred in this construct.
[e] In this construct splicing also occurred at site +132 (score: 0.707).
[f] In these constructs the CAG site is only three residues downstream of the rightmost site in the row. It is not clear whether the experimental evidence distinguished between the sites.

Maize Bronze2 *intron*

Pre-mRNA processing of insertion mutants of the *Bz2* gene was first studied by Luehrsen & Walbot (1994b). The wild-type gene contains a single intron (78 bases), which is spliced with >90% efficiency *in vivo*. Four of the 15 insertion mutants studied were shown to be alternatively spliced. Table 7 gives the splice site scores of the observed transcripts for the wild-type, the alternatively spliced constructs, and one further insertion mutant (the remaining constructs were controls that were unspliced if the native intron was deleted and spliced normally otherwise). The *Bz2* wild-type donor and acceptor sites (pCABz2i) score the high values of 0.815 and 0.966, respectively, consistent with the observed high efficiency of splicing *in vivo* and in transient assays. Insertion of a 113 base-pair fragment from the C-terminal end of *Adh1-S* intron 1 41 bases downstream of the *Bz2* intron (i-s113 S) resulted in skipping of the *Bz2* acceptor site in favor of the *Adh1-S* derived 3′ site. Because the insertion is within the second exon of *Bz2*, the compositional context for the new 3′ site is highly favorable, resulting in a score of 0.998 for this site compared with a score of 0.869 for the upstream *Bz2* acceptor site. Thus, acceptor site choice in this case is consistent with selection of the locally optimal scoring site. Interestingly, some splicing was observed from an alternative 5′ site, which is internal to the *Bz2* wild-type intron and is apparently not used in the wild-type gene (a context in which splicing to the native acceptor site would result in an unacceptably short intron). The score of this site (0.293) is moderate relative to the score distribution of known maize donors (Figure 1). The same splicing pattern was observed when a 346 base-pair fragment internal to *Adh1-S* intron 1 was inserted instead of the 113 base-pair fragment (i-s346 S). In this case, the insertion results in a near perfectly scoring acceptor site right at the transition point of *Adh1-S* intron sequences to *Bz2* exon sequences. Again, favorable compositional context can explain use of the site in this construct while it is not used in its normal context within the *Adh1-S* pre-mRNA; in the normal context the score is an insignificant 0.002. Deletion of the *Bz2* intron sequences apart from the immediate context of the 5′ splice site (x-s346 S) from

Table 7. Splice site analysis in maize *Bronze2* mutant constructs

Construct[a]	[a,b]	Position[c]	Donor site					Acceptor site				
			Sequence	L	X_U	X_{GC}	P	Sequence	L	X_U	X_{GC}	P
pCABz2i		328	AAGGTGAGC	−3.21	−0.04	0.10	0.815					
		405						CGTGCCTTTCTGCAGTT	−5.08	0.14	−0.28	0.966
i-s113S		328	AAGGTGAGC	−3.21	−0.04	0.10	0.815					
		372	AAGGTAACG	−4.70	−0.08	−0.02	0.293					
		405						CGTGCCTTTCTGCAGTT	−5.08	0.06	−0.24	0.869
		559						CCTGGACCCGTGCAGGG	−4.72	0.32	−0.32	0.998
i-s346S		328	AAGGTGAGC	−3.21	−0.04	0.10	0.815					
		372	AAGGTAACG	−4.70	−0.08	−0.02	0.293					
		405						CGTGCCTTTCTGCAGTT	−5.08	0.08	−0.24	0.896
		791						TTTCTCGGACGTAAGGG	−5.56	0.36	−0.38	0.999
x-s346S		328	AAGGTGAGC	−3.21	−0.02	−0.08	0.377					
		727						TTTCTCGGACGTAAGGG	−5.56	0.36	−0.38	0.999
c-a346S		401	CGAGTTAGT	−15.17	−0.16	0.32	0.000					
		492						TTCATTAAAGAACAGGG	−8.47	−0.02	−0.10	0.017
		557	ACGGTAATA	−6.29	−0.04	0.12	0.200					
		652						TCACCACGATTGCAGGA	−5.30	0.14	−0.08	0.724
i-a346S		328	AAGGTGAGC	−3.21	−0.04	0.10	0.815					
		405						CGTGCCTTTCTGCAGTT	−5.08	0.08	−0.24	0.896
		531						TTGCAACATTCTTAGAT	0.07	0.04	0.02	0.165
		570						TTCATTAAAGAACAGGG	−8.47	−0.02	−0.10	0.017
		635	ACGGTAATA	−6.29	−0.04	0.12	0.200					
		730						TCACCACGATTGCAGGA	−5.30	0.14	−0.08	0.724

[a] From Luehrsen & Walbot (1994b).
[b] Alternative splicing events are indicated in column two. Method: RT-PCR.
[c] Numbering is from the initiator AUG in the pre-mRNA. Some sites are apparently not used (see the text).

i-s346 S did not change the splicing pattern. This deletion reduces the donor site score to 0.377.

Surprising splicing patterns were found in constructs that contain the 346 base-pair *Adh1-S* intron fragment inserted in the antisense direction. Two novel introns of lengths 92 and 96 bases were spliced from the insert in the construct without the *Bz2* intron (c-a346 S). Whereas the second intron has reasonably scoring splice sites, the first intron has very poorly scoring sites. In particular, the donor site does not conform at all to the maize donor site consensus and is in a bad GC context. Its U context, however, is better than average ($X_U = -0.16$). Transcript levels for this construct were low, and most of the transcripts were unprocessed. The same insertion in the construct containing the *Bz2* intron (i-a346 S) yielded alternatively spliced transcripts in which the *Bz2* donor site was paired with either the wild-type *Bz2* acceptor site or any of the two novel sites seen in c-a346 S.

Pea rbcS3A *intron 1*

Elements critical for splice site selection in wild-type and mutant constructs of the pea *rbcS3A* first intron were studied by McCullough *et al*. (1993) and Baynton *et al*. (1996). A first series of experiments showed that splicing could be initiated from three sites: an upstream cryptic exonic 5′ site, a downstream cryptic intronic 5′ site, and the usual (+1) site, depending on mutations that altered the match with the plant donor site consensus. The *P*-scores for the three alternative sites in these various constructs are shown in the upper half of Table 8. The predominantly selected splice sites are indicated in bold face. In the 13 constructs reported, the maximal scoring site is almost always the predominantly used site (the only exceptions involving small differences between two almost equally good sites, e.g. construct 3A1.-2C, + 3A;-57E). An interesting comparison is 3A1.-2C;-57E *versus* 3A1.-2 T; + 106E. In the first case, the upstream exonic cryptic site is enhanced to an almost perfect score (0.995) and was used exclusively over the native +1 site. In the second construct, the downstream intronic cryptic site is enhanced to a score of 0.98 but in this case substantial splicing still occurred at +1. This example suggests caveats to predicting splicing efficiencies entirely from the *P*-values.

The −57E site is AAG/GUAAGU with a perfect *L*-score of zero, the +106E site is GAG/GUAAGU with a nearly perfect *L*-score of −1.20, and the +1 site is CAG/GUCAGA with $L = -4.09$ (−6.83 for the -2 T mutant and −7.30 for the -2C mutant; see Table 1). Preferential use of the +1 site over +106E in 3A1wt; + 106E provided compelling evidence for a role of compositional contrast in splice site selection (McCullough *et al*., 1993). This also justifies our modeling approach involving the X_U and X_{GC} variables in addition to the *L* variable; the favorable context of the +1 site compensates for the lower *L*-score relative to the +106E site such that both sites score almost equally high (Table 8).

Table 8. 5′ splice site selection in pea *rbcS3A* intron 1 mutant constructs

Site Construct[a]	−57 *P*[b]	+1 *P*	+106 *P*
3A1wt	0.182	**0.976**	0.356
3A1.1A	**0.182**	0.000	**0.356**[c]
3A1.2A	**0.182**	0.000	**0.356**[c]
3A1. − 2T	0.182	**0.680**	0.356
3A1. + 5A	0.182	**0.808**	0.356
3A1. − 2T, + 5A	**0.182**	0.182	**0.356**[c]
3A1wt; − 57E	**0.995**	**0.976**	0.356
3A1. − 2C; − 57E	**0.995**	0.680	0.356
3A1. − 2C, + 3A; − 57E	***0.995***	**0.983**	0.356
3A1. + 5C; − 57E	**0.995**	***0.704***	0.356
3A1.1A; − 57E	**0.995**	0.000	0.356
3A1wt; + 106E	0.182	**0.976**	***0.980***
3A1. − 2T; + 106E	0.182	**0.680**	**0.980**[c]
3A1.1A; − 57E; + 106E	**0.995**	0.000	***0.980***
3A1.; + 106E;ex3a	0.182	0.115	**0.998**
3A1.; + 106E;in4	0.182	**0.890**	0.770
3A1.; + 106E;in4a	0.182	**0.924**	0.986
3A1. − 2T; + 106E;ex3a	0.182	0.007	**0.998**
3A1. − 2T; + 106E;in4	0.182	**0.301**	0.770
3A1. − 2T; + 106E;in4a	**0.182**	**0.391**	**0.986**[d]
3A1.; − 57E;ex3a	***0.980***	**0.985**	0.356
3A1.; − 57E;in4	**1.000**	0.329	0.356
3A1. − 57E;in4a	**1.000**	0.911	0.356

[a] From McCullough *et al*. (1993).

[b] Predominantly used sites are indicated in bold face. Sites in bold face italics were used to a lower extent, and sites in regular type were not used in detectable amounts. Method: RT-PCR and RNase protection analyses.

[c] In these constructs there seems to be an additional site, probably at +169 (score: 0.455).

[d] Another alternative site is at +35 (score: 0.078).

In a second series of experiments, McCullough *et al.* (1993) replaced sequences upstream or downstream from the +1 site with heterologous exonic or intronic sequences (lower half of Table 8). For seven of the nine constructs the observed splicing patterns are according to expectation derived from splice site scores. The complex splicing patterns of constructs involving sequence replacements between the wild-type −2T and +106E sites (constructs 3A1.-2T; + 106E;in4 and 3A1.-2T; + 106E;in4a) cannot be readily explained.

Baynton *et al.* (1996) constructed ten mutants of the pea *rbcS3A* intron 1 involving U or A-rich tracts upstream of the normal 3′ splice site and upstream of a cryptic site at base position 62 in the downstream exon. Sufficiently long U-tracts were shown to activate the cryptic site and yield correspondingly spliced transcripts in transfected tobacco leaf disc nuclei. Table 9 gives the observed splicing efficiencies at the two sites as well as the associated splice site variable and *P*-values calculated with the *Arabidopsis*-based dicot models. With one exception, predominant splicing occurred at the highest-scoring site. The wild-type was efficiently spliced (72%) with splicing entirely at the −1 site. The model predictions are consistent in that the −1 site scores highly (0.882) whereas the +62 site scores very poorly (0.001). Replacement of nine U−residues in the wild-type acceptor upstream −17 to −4 region (construct -1A) reduced splicing to 10%, which was found to be distributed evenly over three alternative sites at −1, +62, and +98. All sites score poorly. A series of constructs that introduced U-tracts of various lengths upstream of the +62 site was shown to activate this site to different degrees depending on the length of the U-tract. 6UG improves the +62 score to 0.329, but it also raises the −1 site score because it increases the U-contrast in the 50 base window around the −1 site. Consistently, splicing efficiency was 82%, up 10% relative to the wild-type with all splicing occurring at the −1 site. The 10U and 14U constructs resulted in competition between the −1 and +62 sites. The *P*-value of the −1 site is decreased to 0.842 in the 10U constructs and to 0.630 in the 14U constructs, values lower than the +62 site values. However, the distribution of splicing events at the two sites is only partially explained by the model predictions. For example, for 10UC the −1 site score is lower than the +62 site score (0.939) but most splicing still occurs at −1. It is possible that the −1 site is favored as a result of variables not included in the model, e.g. the branchpoint quality or location. 14UG and 14UA both render the +62 site almost perfect with a score of 0.998. However, 14UG had no splicing at −1, whereas 14UA had 37% splicing at −1 compared with 48% at +62. While the model parameters (Table 2) indicate that G is the preferred base over A at acceptor site position −4, the model cannot differentiate between these choices in the context of the highly favorable base choices in the other positions. Insertion of A instead of U-tracts upstream of +62 (constructs 14A) did not activate this site; the *P*-values are very low.

Splicing of synthetic introns

The requirements for intron recognition in plants were also studied for synthetic introns inserted into a plant expression vector transcribed in tobacco protoplasts (Goodall & Filipowicz, 1989; Gniadkowski *et al.*, 1996). Splicing efficiencies and splice site scores for 29 representative constructs are listed in Table 10. Because of the small size of the basic intron construct (85 bases), changes of bases within the intron can affect both splice site scores. Changes in splicing efficiency again are reflected in commensurate changes in splice site scores, with one notable exception: acceptor sites preceded by C-rich segments are scored very low by our models but in some cases were used efficiently (e.g. syn18, syn31, ClaU2/SacU2, MluU2). Recall the same problem with respect to pALC6 (Table 6). A possible explanation is that cytosine can serve as a pyrimidine substitute for uracil in acceptor sites but C-rich acceptors are avoided for some other reasons in natural genes (and thus score low by our models which were trained on natural genes). No splicing was detected in the

Table 9. 3′ splice site selection in pea *rbcS3A1* intron 1/exon 2 mutant constructs

	Site −1					Site +62				
Construct[a]	*L*	X_U	X_{GC}	*P*	% Splicing[a,b]	*L*	X_U	X_{GC}	*P*	% Splicing[a,b]
wt	−4.01	0.10	−0.16	**0.882**	**71.9 ± 4.9**	−12.84	0.02	−0.02	0.001	0
−1A	−8.84	−0.08	−0.16	0.008	3.8 ± 0.1	−12.84	0.02	−0.02	0.001	3.6 ± 0.1
6UG	−4.01	0.16	−0.16	**0.934**	**81.8 ± 2.4**	−7.05	0.04	−0.06	0.329	0
10UG	−4.01	0.12	−0.12	0.842	17.3 ± 2.2	−3.29	0.12	−0.12	**0.990**	**75.6 ± 1.4**
10UC	−4.01	0.12	−0.12	0.842	**70.0 ± 2.8**	−5.20	0.12	−0.12	**0.939**	16.6 ± 2.0
14UG	−4.01	0.04	−0.10	0.630	0	−2.85	0.20	−0.14	**0.998**	**88.5 ± 2.4**
14UC	−4.01	0.04	−0.10	0.630	34.4 ± 1.7	−4.76	0.20	−0.14	**0.987**	**52.5 ± 5.8**
14UA	−4.01	0.04	−0.10	0.630	36.9 ± 3.3	−3.18	0.20	−0.16	**0.998**	**48.4 ± 4.3**
14UU	−4.01	0.04	−0.10	0.630	33.4 ± 2.0	−4.34	0.22	−0.16	**0.994**	**49.1 ± 5.1**
14AC	−4.01	0.16	−0.10	**0.861**	**87.2 ± 1.9**	−13.39	−0.08	−0.14	0.001	0
14AG	−4.01	0.16	−0.10	**0.861**	**81.6 ± 4.3**	−11.48	−0.08	−0.14	0.004	0

The highest scoring sites and the highest efficiencies are indicated in bold face.
[a] From Baynton *et al.* (1996).
[b] Method: RT-PCR.

Table 10. Splice site analysis in *syn* intron mutant constructs

		Donor site				Acceptor site						
Construct[a]	Sequence	L	X_U	X_{GC}	P	Sequence	L	X_U	X_{GC}	P	Intron score[b]	% Splicing[a]
syn35	GAGGTAAGA	−2.55	0.44	0.34	1.000	TTTGTTTTCTTGCAGGT	−2.43	0.52	−0.32	1.000	1.000	86
syn18	GAGGTAAGA	−2.55	−0.34	0.36	1.000	CCGGGCCCGCCGCAGGT	−10.50	0.18	−0.08	0.087	0.087	83
syn7	GAGGTAAGA	−2.55	−0.34	0.36	1.000	TATGHATATCATGCAGGT	−6.24	0.34	−0.32	0.998	0.998	82
syn17	GAGGTAAGA	−2.55	0.00	0.36	0.997	TATGATATCATGCAGGT	−6.24	0.06	−0.32	0.964	0.961	82
syn31	GAGGTAAGA	−2.55	−0.34	0.36	1.000	CCGGGCCCGCCGCAGGT	−10.50	0.10	0.02	0.012	0.012	81
syn25	GAGGTAAGA	−2.55	−0.36	0.38	1.000	TATGATATCATGCAGGT	−6.24	0.38	−0.32	0.999	0.999	81
syn15	GAGGTAAGA	−2.55	−0.34	0.36	1.000	TATGATATCATGCAGGT	−6.24	0.36	−0.32	0.999	0.999	80
syn16	GAGGTAAGA	−2.55	−0.36	0.38	1.000	TATGATATCATGCAGGT	−6.24	0.38	−0.32	0.999	0.999	80
syn29-5X	GAGGTAAGA	−2.55	−0.32	0.42	1.000	TATGATATCATGCAGGT	−6.24	0.24	−0.24	0.988	0.988	80
syn29R	GAGGTAAGA	−2.55	0.00	0.24	0.985	TATGATATCATGCAGGT	−6.24	0.24	−0.24	0.988	0.973	80
syn28L	GAGGTAAGA	−2.55	0.00	−0.08	0.453	TATGATATCATGCAGGT	−6.24	0.34	−0.32	0.998	0.452	80
syn26	GAGGTAAGA	−2.55	−0.36	0.38	1.000	TTTGTTTTCTTGCAGGT	−2.43	0.50	−0.32	1.000	1.000	61
syn36	GAGGTAAGA	−2.55	−0.44	0.34	1.000	TTTGTTTTCTTGCAGGT	−2.43	0.52	−0.32	1.000	1.000	54/67
syn23	GAGGTAAGA	−2.55	0.08	−0.14	0.144	TATGATATCATGCAGGT	−6.24	0.06	−0.04	0.524	0.075	47
syn13	GAGGTAAGA	−2.55	0.00	−0.08	0.453	TATGATATCATGCAGGT	−6.24	0.24	−0.24	0.988	0.448	23
syn19	GAGGTAAGA	−2.55	0.00	−0.08	0.453	CCGGGCCCGCCGCAGGT	−10.50	0.08	0.00	0.012	0.005	<10
syn24	GAGGTAAGA	−2.55	0.08	−0.14	0.144	CCGGGCCCGCCGCAGGT	−10.50	−0.10	0.20	0.000	0.000	<10
ClaU2/SacU2	GAGGTAAGT	−1.20	−0.10	0.04	0.978	TTATGCGGCCCGCAGGT	−8.82	0.10	−0.02	0.094	0.092	66
MluU2*m	GAGGTAAGT	−1.20	0.02	−0.08	0.725	CCCCGCGGACTGCAGGT	−8.57	−0.06	0.16	0.003	0.002	66
SacU2m	GAGGTAAGT	−1.20	0.10	−0.18	0.236	TTATGCGGACTGCAGGT	−6.48	0.12	−0.06	0.683	0.161	63
ClaU/SacU2	GAGGTAAGT	−1.20	0.00	−0.06	0.808	TTATGCGGCCCGCAGGT	−8.82	0.10	−0.02	0.094	0.076	57
SacU3	GAGGTAAGT	−1.20	0.10	−0.18	0.236	TTATGCGGCCCGCAGGT	−8.82	0.20	−0.08	0.391	0.092	56
MluU2	GAGGTAAGT	−1.20	−0.06	0.00	0.945	CCCGCGGCCCGCAGGT	−10.91	−0.08	0.18	0.000	0.000	55
MluU2*	GAGGTAAGT	−1.20	0.02	−0.08	0.725	CCCCGCGGCCCGCAGGT	−10.91	−0.08	0.20	0.000	0.000	42
ClaU3	GAGGTAAGT	−1.20	−0.20	0.16	0.998	CCCCGGCCCGCAGGT	−10.91	−0.10	0.20	0.000	0.000	41
SacU2	GAGGTAAGT	−1.20	0.10	−0.18	0.236	TTATGCGGCCCGCAGGT	−8.82	0.10	−0.02	0.094	0.022	41
ClaU	GAGGTAAGT	−1.20	0.00	−0.06	0.808	CCCCGCGGCCCGCAGGT	−10.91	−0.10	0.20	0.000	0.000	15
ClaA2	GAGGTAAGT	−1.20	0.10	0.04	0.863	CCCCGCGGCCCGCAGGT	−10.91	−0.10	0.20	0.000	0.000	<5
MluA2	GAGGTAAGT	−1.20	0.12	0.00	0.749	CCCCGCGGCCCGCAGGT	−10.91	−0.10	0.18	0.000	0.000	<5

[a] From Goodall & Filipowicz (1989), upper half, and Gniadkowski *et al.* (1996), lower half; constructs are ordered by decreasing splicing efficiency. Method: RNase protection analysis.
[b] Intron score = $5'P \times 3'P$.

Table 11. Length distribution of plant introns and internal exons

	Introns		Internal exons	
	Maize	*Arabidopsis*	Maize	*Arabidopsis*
Minimum	64	59	9	9
10% quantile	80	79	76	61
20% quantile	86	84	87	76
Median	107	97	126	119
80% quantile	184	181	224	228
90% quantile	511	303	319	302
Maximum	5056	1737	1410	3062

The data are based on 201 maize and 578 *Arabidopsis* introns and 156 maize and 457 *Arabidopsis* internal exons.

ClaA2 and MluA2 constructs for which A-rich insertions destroy the favorable U context of the donor site. The donor site scores are still high because of the near perfect matching to the donor consensus. In some cases, the observed combinatorial effects of the different splice site variables are clearly more complex than modeling by the simple linear relationship equation (2) implies.

The relatively inefficient splicing of syn26 and syn36 despite perfectly scoring splice sites is likely the result of elimination of a possible branchpoint motif (Goodall & Filipowicz, 1989). Scoring for presence and location of branchpoint motifs is not included in our acceptor site prediction algorithm because the sequence requirements for branchpoints in plant introns are poorly understood at present (Liu & Filipowicz, 1996; Simpson *et al*., 1996).

Intron and exon definition by splice site recognition

Now that we have shown that the logitlinear models for plant splice sites give reasonable predictions of experimentally measured splice site strengths we turn our attention to the problem of identification of introns in pre-mRNA. The initial challenge is that with a threshold set low enough not to miss any true sites there is an excess of false positive predictions ranging from 2.5-fold for maize donors to 7.5-fold for *Arabidopsis* acceptors (Kleffe *et al*., 1996). These ratios are overly pessimistic for prospects of intron recognition by sequence inspection because (1) in reality one is looking for appropriately spaced pairs of potential donors and downstream acceptors, and (2) in the cellular context there would presumably be competition among nearby splice sites for the same pool of splicing factors. Both considerations bear on the gene prediction problem.

To address the first point, we determined typical intron and exon lengths (Table 11). It is noteworthy that plant introns and internal exons have similar length distributions: approximately 80% of introns and exons range from about 60 to 200 bases. Table 12 displays the average splice site scores defining introns and exons of particular lengths. In maize, long introns show much stronger acceptor sites than mid-size introns, mostly as a result of increased U contrast. The short introns have higher scoring donor sites. Long introns in *Arabidopsis* show increased contrast values for both splice sites, resulting in much elevated intron scores (defined as above as the product of donor and acceptor *P*-values). The same pattern is seen for long maize exons, but in *Arabidopsis* long exons show increased *P*-values for acceptor sites only. We speculate that longer segments must be delineated by stronger splice sites to compensate for larger numbers of competing alternative sites (see below). We found no correlation between the scores of the donor and acceptor site of introns nor between the scores of the acceptor and donor sites flanking internal exons (data not shown).

The second point concerns the issue of locally optimal splice sites. Do true splice sites score maximally in their local context? We consider two scan-

Table 12. Average splice site scores for different length classes of introns and internal exons

	No.	$5'X_U$	$5'X_{GC}$	$5'L$	$5'P$	$3'X_U$	$3'X_{GC}$	$3'L$	$3'P$	Intron score[a]
A. *Maize intron length*										
< 100	88	−0.145	0.138	−4.427	0.686	0.164	−0.130	−5.340	0.622	0.430
100–200	74	−0.126	0.123	−4.457	0.616	0.139	−0.112	−5.567	0.546	0.359
> 200	39	−0.123	0.129	−4.371	0.619	0.188	−0.137	−5.234	0.714	0.426
B. Arabidopsis *intron length*										
< 100	306	−0.127	0.132	−4.649	0.646	0.173	−0.118	−6.880	0.612	0.393
100–200	167	−0.139	0.139	−4.647	0.678	0.166	−0.132	−6.966	0.607	0.390
> 200	103	−0.161	0.153	−4.659	0.731	0.187	−0.151	−6.581	0.687	0.511
C. *Maize exon length*										
< 100	48	−0.133	0.125	−4.135	0.680	0.160	−0.100	−4.875	0.596	0.410
100–200	70	−0.106	0.111	−4.451	0.587	0.154	−0.120	−5.438	0.608	0.371
> 200	38	−0.161	0.166	−4.488	0.709	0.184	−0.165	−5.533	0.690	0.511
D. Arabidopsis *exon length*										
< 100	170	−0.143	0.128	−4.600	0.648	0.157	−0.111	−6.802	0.601	0.394
100–200	178	−0.123	0.130	−4.547	0.673	0.163	−0.122	−6.991	0.573	0.404
> 200	109	−0.119	0.142	−4.572	0.670	0.191	−0.138	−6.553	0.666	0.451

[a] Intron score is defined as the product of $5'P$ times $3'P$. The scores for the exons are derived from the exon defining splice site *P*-values.

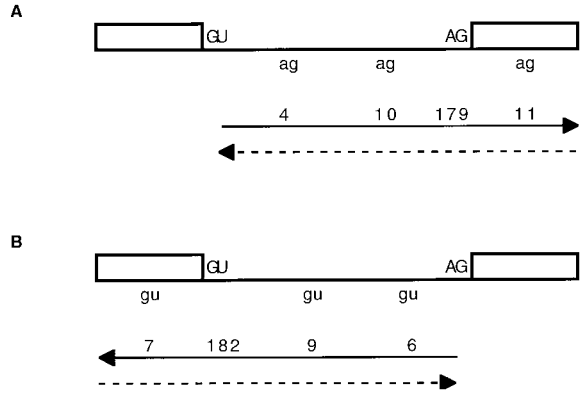

Figure 3. Relative position of locally optimal scoring splice sites in maize genes. True splice sites are indicated by capital letters and predicted splice sites are indicated by lower case letters. Continuous and broken arrows represent scanning according to intron and exon definition, respectively. (A) Acceptor sites. Within the bounds of an upstream donor site (GU) and the next downstream donor site (or the end of the gene), the correct acceptor site (AG) scored maximal for 179 of the 204 considered intron-exon pairs. In four cases the maximal scoring acceptor site was found within the true intron and within 60 bases of the upstream donor site, in ten cases it was found within the intron but further downstream than 60 bases from the donor site, and in 11 cases it occurred downstream from the true acceptor site within the exon. (B) Donor sites. Within the bounds of an upstream acceptor site (or the beginning of the gene) and the next downstream acceptor site (AG), the correct donor site (GU) scored maximal for 182 of the 204 considered intron-exon pairs. In six cases the maximal scoring donor site was found within the true intron and within 60 bases of the downstream acceptor site, in nine cases it was found within the intron but further upstream than 60 bases from the acceptor site, and in seven cases it occurred upstream from the true donor site within the exon.

ning mechanisms (Figure 3). Assume a given donor site has been correctly identified (Figure 3(A)). We consider potential association of this site with either the downstream acceptor site (model of intron definition) or with the upstream acceptor site (model of exon definition; Berget, 1995). Scanning in either direction up to the next donor site, we determined whether the highest scoring acceptor site is either: (1) the true acceptor site, (2) within the intron but located such that the predicted alternative splicing would lead to an intron of unacceptably small length (less than 60 bases); (3) otherwise within the intron (predicting an intron shorter than the true intron), or (4) within the exon (predicting an intron larger than the true intron). Alternatively, we assumed the acceptor site correctly identified and scanned for associated donor sites (Figure 3(B)). The numbers of locally optimal splice sites in the different relative regions defined above are displayed in Figure 3 for the maize set of 204 introns. The true splice sites are locally optimal within the prescribed bounds for almost 90% of the introns, starting either from donor sites scanning for the appropriate acceptor or from acceptor sites scanning for the appropriate donor. This result fits with a possible model for intron recognition in which splicing proceeds from identification of the best splice sites in a gene *via* correct pairing of donor and acceptor sites according to the criterion of locally optimal sites. The exceptional sites are not preferentially associated with any particular class of genes but tend to occur in the longer introns (data not shown).

Discussion

The principles of plant pre-mRNA splicing have been extensively studied in the last ten years (for recent reviews see Brown, 1996; Simpson & Filipowicz, 1996; Filipowicz *et al*., 1995). Of initial interest was the sensitivity of plant intron splicing to disruptions by transposon insertions in either exons or introns (reviewed by Luehrsen *et al*., 1994). In addition, efforts have been directed at elucidating the factors that contribute to the ability of plants to facilitate splicing with sufficient accuracy and efficiency over a wide temperature range. Whereas in yeast the splice signals are highly conserved sequence motifs, the corresponding signals in plants are more relaxed in terms of sequence requirements, in particular for the branchpoint (Simpson *et al*., 1996). The characteristic feature of plant genes is a distinct base compositional contrast of relatively G + C-rich exons flanking U-rich introns.

Recently, two methods have been proposed to identify potential plant splice sites by sequence inspection based on sequence conservation around the 5′ and 3′ splice sites and compositional contrast between site upstream and downstream sequences (Hebsgaard *et al*., 1996; Kleffe *et al*., 1996). Here we have studied the method of Kleffe *et al*. (1996) in terms of its value to predict and interpret splice site choices in native and mutated plant genes. This method is based on logitlinear models, which were trained to discriminate true from false splice sites in known maize and *Arabidopsis* genes. Sites are evaluated by means of *P*-values calculated according to equations (1) and (2). We have shown that the predicted *P*-values correlate well with measured splicing efficiency over a wide range of sequence constructs (Figure 2). In addition, for all well-studied plant introns and the more than 170 associated altered constructs, efficiently used splice sites almost invariably score high and correspond to the predicted sites based on comparisons among scores of nearby alternative sites (Results; Tables 5 to 10). The few exceptions often correspond to constructs that resulted in low RNA yields or very complex splicing patterns. Thus, we conclude that the model captures the essential features of splice site selection and usage *in vivo*.

From examination of a database of 46 maize genes we established that about 90% of native

splice sites are the locally highest scoring sites within the bounds of the flanking exon and intron (Figure 3). These results, as well as the analysis of the experimental results for a few particular introns, suggest that splice site selection involves competition of alternative sites for required factors and that the sequence-derived splice site scores are good indicators of competitive splice site strength.

It is remarkable that the models for splice site prediction are entirely based on local sequence information within 50 base windows of GU or AG dinucleotides for donor and acceptor sites, respectively. Thus, we hypothesize that local features are sufficient to account for intron recognition in plants. In recent experiments we have shown that compositional changes in exons appear to affect splicing only if the changes are in the proximal 50 bases but not if they are further upstream or downstream (M. Latijnhouwers, V. Brendel, V. Walbot & J. C. Carle-Urioste, unpublished results). The compositional difference between plant exons and introns has been conjectured to reflect either positive selection for U-rich motifs in the introns (Luehrsen & Walbot, 1992; Lou *et al*., 1993a) or avoidance of stable secondary structures in the introns (Goodall and Filipowicz, 1991). Based on our results here it seems equally plausible that the base compositional contrast is essential only at the splice sites, possibly mediating interactions between intronic U-rich motif and exonic G + C-rich motif binding proteins. From this perspective, the homogeneity of base composition throughout the introns would reflect negative selection against fortuitous activation of cryptic splice sites by the presence of G + C-rich elements. Candidate intron recognition factors include two proteins from *Nicotiana plumbaginifolia* with poly(U) affinity (Gniadkowski *et al*., 1996). Members of the serine-arginine-rich protein family of splicing factors have also been identified in plants (Lazar *et al*., 1995; Lopato *et al*., 1996), although their exact roles in pre-mRNA splicing are unclear presently. Our analysis of experimental studies of splicing of animal genes in plants suggests that the plant-specific compositional contrast between exons and introns is sufficient to explain success and failure of splicing. It seems likely that species-specific differences between *trans*-acting splicing factors can account for the splicing failure of most animal genes in plants as well as for the reported differences between monocots and dicots.

Comparisons with the experimental data also suggest some limitations inherent to the current splice site prediction models. Discrimination between alternative high scoring sites (and also between alternative low scoring sites) appears difficult. One example is the 14UG *versus* 14UA +62 acceptor site comparison in Table 9: these sites both score 0.998 but exhibited distinctly different usage *in vivo*. Artificial acceptor sites involving C-rich segments upstream of the AG appear to function in several studied constructs (Tables 5 and 10). However, such sites are rare in natural genes: only three maize acceptors have as many as seven C residues in the −15 to −5 acceptor region (average: 2.4, compared with 4.9 U residues), and only three *Arabidopsis* acceptors have five C residues (average: 1.3, compared with 5.7 U residues), and thus score low by our model. It is possible that splicing factors recognize any pyrimidine-rich segment upstream of acceptors and that selective pressures not directly related to splicing result in the preferred U-usage compared with C-usage in native genes. Caveats also apply to comparisons of *P*-values across the different subclasses of sites (G/GU versus H/GU donors, and CAG/*versus* DAG/acceptors) and usage of the maize and *Arabidopsis* models for other monocots and dicots, respectively.

The model predicts that increased G + C content in the flanking exons will improve splicing efficiency of an intron. Such an effect was demonstrated for the maize *Bz2* intron (Carle-Urioste *et al*., 1997), but the generalization is as yet unproven. Construct AT3 + 4 in Table 6 has a site (−139), which scores well as a result of improved G + C contrast but was poorly used *in vivo*. A convincing demonstration of the effect of compositional contrast on splice site selection is given by the recent experiments by McCullough & Schuler (1997). In mutant constructs of the soybean β-conglycinin α′ subunit intron 4 containing two alternative donor sites with identical nine base signal sequence, donor site selection was shown to strongly depend on the composition of the surrounding sequences. The observed splicing patterns correspond to selection of the highest scoring sites predicted by the model (Table 13), with one exception. The exception is construct -81.-2 T, + 5A/int5 S/ + 1wt in which the distal donor site was weakened by a double mutation in positions −2 and +5. The favorable context of this site relative to the proximal site gives it a better score than the exclusively used proximal site. Some exceptions such as this are unavoidable in any statistical method, but overall it seems that the weighting of signal strength and compositional contrast as trained in the model works remarkably well.

How can plant splice site prediction be improved further? Some improvement can be anticipated simply from increased data sets available for model training. Given that our modeling approach is basically sound, the increased data sets should provide more accurate parameter estimates. Moreover, finer subclassifications of sites can be studied, as well as more detailed models with more parameters (for example, di- rather than mononucleotide models that should better reflect RNA-RNA interactions; e.g. see White *et al*., 1992). Ultimately, the models should incorporate more sophisticated representations of binding preferences of the hypothesized *trans*-acting splicing factors. Scoring for putative branchpoint consensus sequences may improve acceptor site identification. Following Simpson *et al*. (1996) we searched for the patterns CTNA and TTNA 21 to 60 bases upstream

Table 13. Splice site analysis in soybean β-conglycinin α′ subunit intron 4 mutant constructs

Construct[a]	Position[b]	Donor site				
		Sequence	L	X_U	X_{GC}	P
wt	+1	GAGGTAAGC	−2.02	0.00	0.16	0.97
int5S	**−79**	**GAGGTAAGC**	**−2.02**	**−0.20**	**0.14**	**1.00**
	+1	GAGGTAAGC	−2.02	0.14	−0.04	0.39
int5AS	**−79**	**GAGGTAAGC**	**−2.02**	**0.00**	**0.20**	**0.99**
	−59	CATGTAATT	−3.52	−0.08	0.30	0.81
	+1	GAGGTAAGC	−2.02	−0.12	0.00	0.93
ex6S	−80	GAGGTAAGC	−2.02	0.00	0.00	0.81
	+1	**GAGGTAAGC**	**−2.02**	**−0.10**	**0.22**	**1.00**
ex6AS	***−80***	***GAGGTAAGC***	***−2.02***	***0.00***	***−0.04***	***0.71***
	+1	**GAGGTAAGC**	**−2.02**	**−0.06**	**0.18**	**0.99**
Δ-81/int5S/ + 1wt	+1	GAGGTAAGC	−2.02	0.14	−0.04	0.39
−81. + 1A/int5S/ + 1wt	+1	GAGGTAAGC	−2.02	0.14	−0.04	0.39
−81. − 2T, + 5A/int5S/ + 1wt	−79	GTGGTAAAC	−7.29	−0.18	0.16	0.53
	+1	**GAGGTAAGC**	**−2.02**	**0.14**	**−0.04**	**0.39**
−81. − 2T, + 5A/int5S/ + 1wt	**−79**	**GTGGTAAGC**	**−4.76**	**−0.18**	**0.14**	**0.91**
	+1	GAGGTAAGC	−2.02	0.14	−0.04	0.39
−81. + 1A/int5S/ + 1wt	**−79**	**GAGGTAAAC**	**−4.55**	**−0.20**	**0.16**	**0.95**
	+1	GAGGTAAGC	−2.02	0.14	−0.04	0.39
ex6S,truncated	+1	GAGGTAAGC	−2.02	−0.10	0.22	1.00
int5S,truncated	+1	GAGGTAAGC	−2.02	0.14	−0.04	0.39
int5S + AU	**−95**	**GAGGTAAGC**	**−2.02**	**−0.08**	**0.08**	**0.97**
	+1	GAGGTAAGC	−2.02	0.16	−0.10	0.19
int5S + AG	**−95**	**GAGGTAAGC**	**−2.02**	**−0.04**	**0.02**	**0.89**
	+1	***GAGGTAAGC***	***−2.02***	***0.06***	***0.04***	***0.80***
−81. − 2T/INT5S + AU	**−95**	**GTGGTAAGC**	**−4.76**	**−0.06**	**0.08**	**0.59**
	+1	GAGGTAAGC	−2.02	0.16	−0.10	0.19
−81. − 2T/int5S + AG	***−95***	***GTGGTAAGC***	***−4.76***	***−0.02***	***0.02***	***0.30***
	+1	**GAGGTAAGC**	**−2.02**	**0.06**	**0.04**	**0.80**
int5S,truncated + AG	+1	GAGGTAAGC	−2.02	0.06	0.04	0.80

In case of alternative sites, the predominant site is indicated in bold face. Sites in bold face italic were used to a lower extent. All processed transcripts are spliced at the same acceptor site (P-value 0.71). Method: RT-PCR.

[a] From McCullough & Schuler (1997).

[b] +1 is the wild-type site. Negative numbers refer to 5′-distal sites; e.g., −79 extends the intron upstream by 79 bases.

of true and putative acceptor sites. Similarly to their numbers, we found the first pattern in 60% of monocot and 66% of dicot introns, and the second pattern occurred in an additional 29% and 30% of monocot and dicot introns, respectively. These numbers are consistently higher than the corresponding numbers for occurrences upstream of non-sites (including only intron and exon-internal sites that score above the level of the minimal true acceptor site scores). For example, CTNA was found upstream of only 41% of maize and 48% of *Arabidopsis* non-sites. Tolstrup *et al*. (1997) recently reported a twofold reduction in the number of false positive acceptor site predictions in *A. thaliana* upon inclusion of scoring for potential branch-points in their neural networks.

Prediction of gene structure can rely on assessment of coding potential in addition to splice site recognition. We have recently developed the algorithm GeneGenerator that produces alternative predictions of gene structure for genomic sequence input (Kleffe *et al*., 1998). Essentially, potential exons and introns are delimited by predicted splice sites and arranged in combinations that give translation products of favorable codon usage. Simultaneous evaluation of splice site strength and composition of the translation products appears very successful for gene prediction.

From a methodological standpoint, we wish to emphasize the new possibilities of computer-aided experimental design afforded by the growing molecular databases and availability of advanced sequence analysis programs. We propose that experiments involving manipulation of molecular sequences can be modeled expediently *in silico*. Experimental results not predicted by the theoretical studies should help to refine our sequence models, which can then be more confidently applied to the large amounts of sequence data now accumulating from the genome projects.

Materials and methods

Gene collections

Genomic sequences from *Zea mays* and *Arabidopsis thaliana* were retrieved from GenBank and compiled into specifically annotated non-redundant databases as described (Kleffe *et al*., 1996). For maize, our database contains 46 genes comprising a total of 250 exons and 204 introns. For *Arabidopsis*, a database of 131 distinct genes was obtained with a total of 709 exons and 578 introns.

Logitlinear models for splice site prediction

Our analyses are based on the splice site models introduced by Kleffe *et al*. (1996). The rationale of the method

and specifics of parameter derivation (training) are described in that paper. The models assign to any GU or AG in a sequence a score between 0 and 1 based on three local sequence properties: (i) the contrast in U composition, measured as $X_U = \%$ U in the 50 bases upstream of the GU (or AG) minus % U in the 50 bases downstream; (ii) the contrast in G + C composition, measured as $X_{GC} = \%$ G + C in the 50 bases upstream of the GU (or AG) minus % G + C in the 50 bases downstream; (iii) splice site quality, measured as a sum of position and base-specific weights such that high scores reflect base choices generally consistent with the most frequent (consensus) bases in each position. Specifically, the score of a given site is calculated as:

$$P = \frac{\exp(\theta)}{1 + \exp(\theta)} \quad (1)$$

where

$$\theta = \alpha + \delta X_U + \mu X_{GC} + L, \qquad L = \sum_i \sum_b \delta_{ib} l_{ib} \quad (2)$$

here δ_{ib} is 1 if the base in position i is b, and 0 otherwise. Summation extends over nine positions for potential donor sites and over 15 positions for potential acceptor sites. The parameters α, δ, μ, and l_{ib} are given in Tables 1 and 2.

As shown by Kleffe *et al*. (1996), prediction of splice sites improved upon subclassification of sites, distinguishing GU sites as G/GU or H/GU (H denoting non-G) and AG sites as CAG/or DAG/(D denoting non-C). Thus, parameters must be taken from Tables 1 and 2 appropriate to the subclassification of the given site. For example, maize *waxy* intron 12 has the donor site GAC/GUAAGC with an *L*-score of −1.08 (upper left panel of Table 1). The compositional contrast values are $X_U = -0.34$ and $X_{GC} = 0.26$, and equations (2) and (1) give $\theta = 7.73$ and $P = 0.999$. The acceptor site is GUCCUCUCUUCCCAG/UG with an *L*-score of −8.42 (upper left panel of Table 2). Wth compositional contrast values $X_U = 0.20$ and $X_{GC} = -0.08$, equations (2) and (1) give $\theta = -1.38$ and $P = 0.20$. These calculations are straightforward and can easily be carried out on a pocket calculator or programmed. Our program *SplicePredictor* is available at http://gnomic.stanford.edu/~volker/SplicePredictor.html.

The interpretation of $P(\theta)$ requires careful consideration of the underlying assumptions and training of the model (cf. Kleffe *et al*., 1996). Ideally, $P(\theta)$ would represent the probability that splicing occurs at a given site. This interpretation would hold if the training data consisted of repeated observations of splicing success or failure at a given site, and if the dependence of P on θ were exactly of the form 1. The actual training data, however, are different, representing single observations of success at known splice sites and failure at all other potential sites. Another complication arises from the dependence of splice site strength on the global context of the site. For example, a highly efficient *in vivo* donor site could be rendered mute by mutation of its corresponding acceptor site. Similarly, a relatively high scoring site may be skipped *in vivo* in favor of an even higher scoring site close by. We show in Figure 2 that, in otherwise constant context, *P*-scores of mutated sites generally correlate well with experimentally measured splicing efficiencies. Moreover, in most cases, high scoring sites are selected preferentially over lower scoring alternative sites.

$P(\tau)$ increases with θ: good splice sites with high values of P correspond to high values of θ. Consistently, the parameter δ is negative for donor sites and positive for acceptor sites, whereas μ is positive for donor sites and negative for acceptor sites. A typical donor site will have negative X_U and positive X_{GC} , thus giving a positive value of $\delta X_U + \mu X_{GC}$ for δ and μ as in Table 1, and a typical acceptor site will have positive X_U and negative X_{GC} , also giving a positive value of $\delta X_U + \mu X_{GC}$ for δ and μ as in Table 2. For each position i, the values of l_{ib} are arbitrary up to an additive constant, because changes in this constant can be absorbed by corresponding changes in the global constant α. For the canonical representation given in Tables 1 and 2, the value 0 is assigned to the consensus base in each position (as well as to those bases excluded from occurring by specification of the subclass of the site). In this case, the value of the constant term α determines the base *P*-value for a consensus splice site (AAG/GUAAGU for G/GU donor sites and AAU/GUAAGU for the H/GU donor sites, U_{11}GCAG/GU for CAG/acceptor sites and U_{11}GUAG/GU for DAG/acceptor sites) in the absence of any compositional contrast. The predominant donor class G/GU has a much higher base value than the minor subclass: 0.97 (maize and *Arabidopsis*) *versus* 0.54 (maize) and 0.59 (*Arabidopsis*). Interestingly, the base *P*-values of CAG/ and DAG/*Arabidopsis* acceptor sites are similar (0.99 *versus* 0.93), but for maize the value for DAG/sites (0.07) is very much lower than the value for the CAG/sites (0.97). The high value of δ for the maize DAG/sites suggests that lack of the predominant C in position −3 must be compensated by good U contrast.

For the most part, the scores l_{ib} are negative and the closer to zero the more frequent b is in position i in the collection of true splice sites used as a training set. Some exceptions are notable. For example, G and to a lesser extent C are higher scoring bases at maize DAG/acceptor site positions −13 and −12. While U is the most frequent base, on average, in these positions, true acceptor site, rarely have long uninterrupted runs of U; the positive weights for G and C probably help to discriminate against false positive sites in the training set. However, the small sample size of DAG/sites available for training (Kleffe *et al*., 1996) suggests caveats to parameter interpretations in this case.

The model predicts that sites with weak quality (low *L*-values) can be partially rescued in a context of favorable compositional contrast, and, *vice versa*, sites of high quality can overcome unfavorable compositional contrast. This is illustrated in Table 3, which gives the *P*-values of maize donor and acceptor sites of different quality in dependence on the values of X_U and X_{GC}. Splice sites of typical quality (represented in the Table by a donor site with an *L*-score of −5.72 and an acceptor site with an *L*-score of −6.93) are the most sensitive to contrast changes, whereas sites of low quality remain weak even in very favorable context, and sites of high quality are predicted to be efficient splice sites even in unfavorable compositional context.

Experimental determination of splice site selection and splicing efficiencies

Experimentally, splice site selection and splicing efficiency have been determined by a variety of techniques, including primer extension, RNase protection assays, reverse transcriptase PCR (RT-PCR), and reporter gene expression. Merits and limitations of these

methods have been reviewed by Luehrsen *et al*. (1994). Some of the differences between the model predictions and the interpretations of experimental data discussed above may in part reflect methodological limitations. For example, primer extension and RNase protection may miss unanticipated splice products and may inadequately resolve splice site selection of nearby sites. RT-PCR quantification assumes equal amplification of different products. All techniques rely on unchanged stability of the various transcripts analyzed. Reporter gene expression also assumes equal translation efficiency for alternatively spliced mRNAs. These assumptions have been shown to hold for particular genes, but the generalization is tentative, especially with respect to some of the more unusual constructs discussed above.

Acknowledgments

V. B. was supported in part by NIH grants 2R01HG00335-09 and 5R01GM10452-32. Contributions by J. C.-U. and V. W. were supported by NIH grant GM49681.

References

Baynton, C. E., Potthoff, S. J., McCullough, A. J. & Schuler, M. A. (1996). U-rich tracts enhance 3′ splice site recognition in plant nuclei. *Plant J.* **10**, 703–711.

Berget, S. M. (1995). Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270**, 2411–2414.

Brendel, V., Carle-Urioste, J. C. & Walbot, V. (1998). Intron recognition in plants. In *A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants* (Bailey-Serres, J. & Gallie, D. R., eds), American Society Plant Physiology, Rockville, MD, in the press.

Brown, J. W. S. (1996). *Arabidopsis* intron mutations and pre-mRNA splicing. *Plant J.* **10**, 771–780.

Burset, M. & Guigó, R. (1996). Evaluation of gene structure prediction programs. *Genomics,* **34**, 353–367.

Carle-Urioste, J. C., Ko, C. H., Benito, M.-I. & Walbot, V. (1994). *In vivo*analysis of intron processing using splicing-dependent reporter gene assays. *Plant Mol. Biol.* **26**, 1785–1795.

Carle-Urioste, J. C., Brendel, V. & Walbot, V. (1997). A combinatorial role for exon, intron and splice site sequences in splicing in maize. *Plant J.* **11**, 1253–1263.

Filipowicz, W., Gniadkowski, M., Klahre, U. & Liu, H.-X. (1995). Pre-mRNA splicing in plants. In *Pre-mRNA Processing* (Lamond, A. I., ed.), pp. 65–77, R. G. Landes Publishers, Georgetown, TX.

Gniadkowski, M., Hemmings-Mieszczak, M., Klahre, U., Liu, H.-X. & Filipowicz, W. (1996). Characterisation 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.

Goodall, G. J. & Filipowicz, W. (1989). The AU-rich sequences present in the introns of plant nuclear pre-mRNAs are required for splicing. *Cell,* **58**, 473–483.

Goodall, G. J. & Filipowicz, W. (1991). Different effects of intron nucleotide composition and secondary structure on pre-mRNA splicing in monocot and dicot plants. *EMBO J.* **10**, 2635–2644.

Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J., Rouzé, P. & Brunak, S. (1996). Splice site prediction in *Arabidopsis thaliana* pre-mRNA by combining local and global sequence information. *Nucl. Acids Res.* **24**, 3439–3452.

Hunt, A. G., Mogen, B. D., Chu, N. M. & Chua, N.-H. (1991). The SV40 small t intron is accurately and efficiently spliced in tobacco cells. *Plant Mol. Biol.* **16**, 375–379.

Kleffe, J., Hermann, K., Vahrson, W., Wittig, B. & Brendel, V. (1996). Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences. *Nucl. Acids Res.* **24**, 4709–4718.

Kleffe, J., Hermann, K., Vahrson, W., Wittig, B. & Brendel, V. (1998). GeneGenerator–a flexible algorithm for gene prediction and its application to maize sequences. *Bioinformatics,* **14**, in the press.

Lazar, G., Schaal, T., Maniatis, T. & Goodman, H. M. (1995). Identification of a plant serine-arginine-rich protein similar to the mammalian splicing factor SF2/ASF. *Proc. Natl Acad. Sci. USA,* **92**, 7672–7676.

Liu, H.-X. & Filipowicz, W. (1996). Mapping of branch-point nucleotides in mutant pre-mRNAs expressed in plant cells. *Plant J.* **9**, 381–389.

Lopato, S., Mayeda, A., Krainer, A. R. & Barta, A. (1996). Pre-mRNA splicing in plants: characterization of Ser/Arg splicing factors. *Proc. Natl Acad. Sci. USA,* **93**, 3074–3079.

Lou, H., McCullough, A. J. & Schuler, M. A. (1993a). Expression of maize Adh1 intron mutants in tobacco nuclei. *Plant J.* **3**, 393–403.

Lou, H., McCullough, A. J. & Schuler, M. A. (1993b). 3′ splice site selection in dicot plant nuclei is position dependent. *Mol. Cell. Biol.* **13**, 4485–4493.

Luehrsen, K. R. & Walbot, V. (1992). Insertion of non-intron sequence into maize introns interferes with splicing. *Nucl. Acids Res.* **20**, 5181–5187.

Luehrsen, K. R. & Walbot, V. (1994a). 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.

Luehrsen, K. R. & Walbot, V. (1994b). Intron creation and polyadenylation in maize are directed by AU-rich RNA. *Genes Dev.* **8**, 1117–1130.

Luehrsen, K. R., Taha, S. & Walbot, V. (1994). Nuclear pre-mRNA processing in higher plants. *Prog. Nucl. Acids Res. Mol. Biol.* **47**, 149–193.

McCullough, A. J. & Schuler, M. A. (1997). Intronic and exonic sequences modulate 5′ splice site selection in plant nuclei. *Nucl. Acids Res.* **25**, 1071–1077.

McCullough, A. J., Lou, H. & Schuler, M. A. (1993). Factors affecting authentic 5′ splice site selection in plant nuclei. *Mol. Cell. Biol.* **13**, 1323–1331.

Simpson, C. G., Clark, G., Davidson, D., Smith, P. & Brown, J. W. S. (1996). Mutation of putative branch-point consensus sequences in plant introns reduces splicing efficiency. *Plant J.* **9**, 369–380.

Simpson, G. G. & Filipowicz, W. (1996). Splicing of precursors to mRNA in higher plants: mechanisms, regulation and sub-nuclear organisation of the spliceosomal machinery. *Plant Mol. Biol.* **32**, 1–41.

Sinibaldi, R. M. & Mettler, I. J. (1992). Intron splicing and intron-mediated enhanced expression in monocots. *Prog. Nucl. Acids Res. Mol. Biol.* **42**, 229–257.

Solovyev, V. V., Salamov, A. A. & Lawrence, C. B. (1994). Predicting internal exons by oligonucleotide

composition and discriminant analysis of spliceable open reading frames. *Nucl. Acids Res.* **22**, 5156–5163.

Tolstrup, N., Rouzé, P. & Brunak, S. (1997). A branch point consensus from *Arabidopsis* found by non-circular analysis allows for better prediction of acceptor sites. *Nucl. Acids Res.* **25**, 3159–3163.

White, O., Soderlund, C., Shanmugan, P. & Fields, C. (1992). Information contents and dinucleotide compositions of plant intron sequences vary with evolutionary origin. *Plant Mol. Biol.* **19**, 1057–1064.

Wiebauer, K., Herrero, J.-J. & Filipowicz, W. (1988). Nuclear pre-mRNA processing in plants: distinct modes of 3′-splice-site selection in plants and animals. *Mol. Cell. Biol.* **8**, 2042–2051.

Zhang, M. Q. (1997). Identification of protein coding regions in the human genome by quadratic discriminant analysis. *Proc. Natl Acad. Sci. USA,* **94**, 565–568.

Edited by F. E. Cohen

(Received 15 May 1997; received in revised form 6 November 1997; accepted 6 November 1997)