4748–4757 *Nucleic Acids Research, 1998, Vol. 26, No. 20*

Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA

Volker Brendel* and Jürgen Kleffe[1]

Department of Mathematics, Stanford University, Stanford, CA 94305, USA and [1]Freie Universität Berlin, Institut für Molekularbiologie und Biochemie, Bereich Molekularbiologie und Informatik, Arnimallee 22, 14195 Berlin, Germany

Received May 18, 1998; Revised and Accepted August 6, 1998

ABSTRACT

Prediction of splice site selection and efficiency from sequence inspection is of fundamental interest (testing the current knowledge of requisite sequence features) and practical importance (genome annotation, design of mutant or transgenic organisms). In plants, the dominant variables affecting splice site selection and efficiency include the degree of matching to the extended splice site consensus and the local gradient of U- and G+C-composition (introns being U-rich and exons G+C-rich). We present a novel method for splice site prediction, which was particularly trained for maize and *Arabidopsis thaliana*. The method extends our previous algorithm based on logitlinear models by considering three variables simultaneously: intrinsic splice site strength, local optimality and fit with respect to the overall splice pattern prediction. We show that the method considerably improves prediction specificity without compromising the high degree of sensitivity required in gene prediction algorithms. Applications to gene identification are illustrated for *Arabidopsis* and suggest that successful methods must combine scoring for splice sites, coding potential and similarity with potential homologs in non-trivial ways. A WWW version of the SplicePredictor program is available at http:/gnomic.stanford.edu/ ~volker/SplicePredictor.html/

INTRODUCTION

Accurate prediction of splice sites in pre-mRNA is prerequisite to reliable algorithms for the identification of split genes by sequence inspection (for recent reviews see 1–4). Success and failure of prediction also reflect how well the sequence requirements for faithful and efficient splicing are understood. Plants share the pattern of base preferences at the 5′ and 3′ splice sites observed for vertebrates and *Saccharomyces cerevisiae* (5,6). In addition, plant splice sites typically occur at junctions of exonic G+C-rich sequences with intronic U-rich sequences (7–11). Current methods to predict plant splice sites are based on recognition of these features (12,13). The neural network method of Hebsgaard *et al*. (NetPlantGene, 12) also evaluates coding potential at the predicted exonic side of a splice site candidate (see also method of Solovyev *et al*., 14). Consideration of coding potential is clearly useful from a practical standpoint for the purpose of identifying sites that are consistent with the assembly of an open reading frame (ORF) in the predicted mRNA. A fundamental concern may be that explicit biochemical recognition of the triplet code is probably not involved in the nuclear splicing reactions but is confined to the level of translation (although some mechanisms exist to tag mRNAs with premature translation termination codons for rapid degradation; 15,16). Scoring for typical codon bias may also mislead splice site prediction in some cases of novel genes with atypical codon usage (1). For these reasons it seems worthwhile to also study methods which do not rely on measures of coding potential.

Here we extend the method of Kleffe *et al*. (13) which is applicable also to sites in untranslated portions of the pre-mRNA. As with other methods, this method scores the quality of isolated potential sites relative to average features derived from training sets of known sites of the same type. To predict pre-mRNA processing events, however, it is more appropriate to evaluate splice site candidates in the context of other potential splice site partners (17). For example, a high quality donor site paired with a high quality acceptor site occurring downstream in the sequence at a distance corresponding to the typical intron length should be predicted with higher confidence than if it occurred in a segment devoid of good acceptor site candidates. We present a novel method of splice site prediction based on these considerations. Our method differs from global methods (most recently reviewed in 1) which evaluate complete genes by combining splice site prediction, evaluation of coding potential and similarity to known gene products into entire gene predictions. We seek to improve splice site prediction *per se* and therefore confine to a moderate sequence neighbourhood of a potential splice site that could be of direct importance for its recognition by splicing factors. NetPlantGene (12) considers similarly sized neigbourhoods of potential splice sites but uses coding information.

*To whom correspondence should be addressed at present address: Department of Zoology and Genetics, 2112 Molecular Biology Building, Iowa State University, Armes, IA 50011-3260, USA. Tel: +1 515 294 9884; Fax: +1 515 294 6755; Email: vbrendel@iastate.edu

Each potential site is assigned three scores: (i) a *P*-value measuring intrinsic splice site quality, (ii) a ρ-value measuring local optimality of the site and (iii) a γ-value measuring the contribution of the site to the predicted overall splicing pattern in the context of the flanking sequence segments. The *P*-value is calculated as previously described (13). In otherwise constant context, sites with increased *P*-value are predicted to result in more efficient splicing. A high correlation of *P*-values with experimentally measured splicing efficiencies has been demonstrated (18). Here we show that inclusion of the ρ- and γ-values significantly improves the specificity of splice site prediction.

MATERIALS AND METHODS

Gene collections

Genomic sequences from *Zea mays* and *Arabidopsis thaliana* were taken from our previously compiled non-redundant databases (13). For maize, the database contains 46 genes comprising a total of 250 exons and 204 introns. For *Arabidopsis*, the database contains 131 genes with a total of 709 exons and 578 introns. The databases include 16 pairs of highly similar genes conserved between maize and *Arabidopsis* (>40% identity on the amino acid level).

Calculation of *P*-values

P-values are calculated according to the logitlinear splice site models introduced in (13). These models assign to any GU (potential donor site) and AG (potential acceptor site) in a sequence a score between 0 and 1 based on three local sequence properties: (i) X_U, the contrast in U composition, measured as % U in the 50 bases upstream of the GU (or AG) minus % U in the 50 bases downstream; (ii) X_{GC}, the contrast in G+C composition, measured as % G+C in the 50 bases upstream of the GU (or AG) minus % G+C in the 50 bases downstream; and (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:

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

where

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

and 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} were derived as maximum likelihood estimators from training sets of known sites and non-sites (13,18). This model does not consider the small fraction (<1%) of known splice sites lacking the GU or AG consensus (19).

Calculation of ρ-values

The ρ-value of a given splice site indicates the extent to which the site occurs in a favorable sequence context. Favorable sequence context would include first, the availability of an appropriately spaced complementary splice site such that this pair of sites defines a potential intron and second, absence of nearby sites of the same type with higher *P*-values which could favorably compete with the given site for splicing factors. Numerically, ρ is calculated as follows (Fig. 1; see Table 1 for specific examples). Let d be the location of a donor site with *P*-value P^D_d. First, the downstream sequence is searched for the highest scoring acceptor site, occurring say at position a with score P^A_a. The search extends from d + 60 (60 taken as the minimal allowable intron length) to either r + 60, if there is a donor site at $r > d$ + 60 with score $P^D_r > P^D_d$, or to d + 600. The first downstream limit obtains because, in this case, acceptor sites further downstream than r + 60 would pair better with the donor site at r. The second downstream limit effectively sets the maximal size of potential introns considered to 600 bases. These size restrictions include >95% of all known plant introns (11,19). The sites d and a define an initial intron that is compared with a number of alternative introns obtained in the following way.

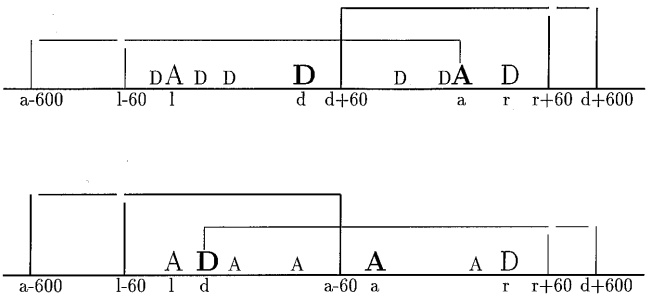

Figure 1. Calculation of ρ-values. (**Top**) The ρ-value of the donor site at position d is calculated as the weighted product of its *P*-value times the *P*-value of its maximally scoring potential complementary acceptor site at *a*. The weight is calculated as described in the text by comparison with alternative donor/acceptor site pairs. The alternative pairs of sites are determined stepwise as follows. First, the maximally scoring acceptor site complement for the donor site at *d* is determined in the sequence interval *d* + 60 to *d* + 600 or to *r* + 60 if there is a donor site at *r* > *d* + 60 with *P*-value greater that the *P*-value of the donor site at *d* (thick-lined box). Here, 60 is taken as the minimal and 600 as the maximal allowed intron length. The bound *r* + 60 is invoked in the case of a higher scoring donor site at *r* because any acceptor site to the right of this would favorably pair with this site compared to the site at *d* and could form an intron downstream of an intron starting at *d*. Alternative donor sites to the site at *d* are then searched for in the bounds *a* – 600 to *a*, where the lower bound is adjusted to *l* – 60 if there is an acceptor site at *l* < *d* with higher *P*-value than the site at *a* (in this case, any donor site to the left of *l* – 60 would favorably pair with the acceptor site at *l*; thin-lined box). The positions of five alternative donor sites are illustrated in the figure by the smaller font Ds. (**Bottom**) The ρ-value of the acceptor site at position *a* is calculated as the weighted product of its *P*-value times the *P*-value of its maximally scoring potential complementary donor site at *d*. The complementary donor site is restricted to the bounds *a* – 600 to *a* – 60 or *l* – 60 to *a* – 60 if there is an acceptor site at *l* < *a* – 60 with *P*-value greater than the *P*-value of the site at *a* (thick lined box). Alternative acceptor sites are then searched for in the bounds of *d* to at most *d* + 600 (*r* + 60 if a higher scoring donor site occurs at *r* > *a*, thin-lined box). The positions of alternative acceptor sites are illustrated in the figure by smaller font As.

Alternative donor sites are identified upstream of a. The search extends to l – 60, if there is an acceptor site at position $l < d$ with score $P^A_l > P^A_a$, or to a – 600, whichever is larger. Then, donor sites within the given limits occurring to the left of d are either paired with the acceptor site at a or, if such exists, with a higher scoring acceptor site between d and d + 60. Donor sites occurring between a – 60 and a are paired with their maximal acceptors between 60 and at most 600 bases downstream.

The limits for positions of alternative sites are as prescribed to exclude from consideration most sites that could form maximally scoring upstream or downstream introns without precluding use

of the site at *d* in a different intron (Fig. 1). Assume that within these limits there are *n* alternative donor sites (including the site at *d*) with *P*-values P_k^D and associated acceptor sites with *P*-values P_k^A, $k = 1, 2, ..., n$. Then,

$$\rho = \frac{P_d^D P_a^A}{\sum_{k=1}^{n} P_k^D P_k^A} \quad P_d^D P_a^A \qquad \mathbf{3}$$

The calculation of ρ for a given acceptor site is done in a similar way by first searching for the maximal scoring donor site upstream, and then tallying alternative donor/acceptor site pairs within limits defined analogously to the above for given donor sites. If the given site does not possess a complementary site in the prescribed limits, then ρ is set to 0. Thus, ρ is a value between 0 and 1, and high values of ρ obtain in the case of a donor/acceptor site pair with high *P*-values in a region devoid of high scoring alternative sites.

Calculation of γ-values

The ρ-value measures how well a given potential splice site can be paired up to form a potential intron. However, it is possible that the actual splicing pattern precludes splicing at that given site because usage of other site pairings is favored such that the given site is internal to an exon or intron. The γ-value reflects how well the given splice site fits in the locally predicted splicing pattern. For the given site in a sequence, look at the preceding *N* predicted sites (irrespective of whether these sites are donor or acceptor sites) and at the succeeding *N* predicted sites. Here we have used $N = 7$ (or smaller for sites at the beginning and end of the analyzed sequence). Denoting acceptor sites by A and donor sites by D, the pattern of predicted sites is a string of As and Ds of length $2N + 1$. The predicted sites are either correctly predicted or represent false sites within exons or introns (intergenic regions are assigned to either exons or introns for the purpose of this calculation). Let E and I denote exon and intron, respectively. Then all possible parsings of the sequence segment are obtained by the following string re-writing rules: (i) an A is either retained or replaced by E or I; (ii) a D is either retained or replaced by E or I; and (iii) the re-writing must not produce adjacent letter combinations other than AD, AE, DA, DI, ED, EE, IA or II. Replacement rules (i) and (ii) apply if the particular site is considered a false prediction within exon or intron. Rule (iii) assures that the resulting string represents a legitimate parse: donors are followed by intron sequence up to the next acceptor, acceptors are followed by exon sequence up to the next donor. Assign to each possible such parse a score defined as the sum of the *P*-values of all the constituent Ds and As. Find the maximal score of all parses restricted, first to parses that predict the given site to be a true splice site (i.e. the central letter of the parse string is either A or D) and second, to parses that predict the given site to be within exon or intron (i.e. the central letter of the parse string is either E or I). γ is defined as the difference of the first minus the second of these scores, or zero, if the difference is negative. Thus, if the given site is in a context that suggests preferred usage of nearby sites as splicing partners to the exclusion of the given site, its γ-value will be zero. Otherwise it will be a positive value ≤2; high values of γ would strongly suggest actual usage of the site. Examples are given in Table 1.

Table 1. Splice site prediction in part of the maize Adh1-1F genomic sequence (GenBank accession no. X04050)

Position	site	classification	*P*	ρ	γ	parse
2955	donor	intron 5 start	0.855	0.316	1.090	AEDIAEE-D-AEEDIIA
3048	acceptor	intron 5 end	0.235	0.034	0.101	EDIAEED-A-EEDIIAE
3055	acceptor	within exon	0.134	0.011	0.000	DIAEEDA-E-EDIIAEE
3058	acceptor	within exon	0.002	0.000	0.000	IAEEDAE-E-DIIIIIA
3125	donor	intron 6 start	0.750	0.386	0.742	AEEDAEE-D-IIIIIAE
3176	acceptor	within intron	0.002	0.000	0.000	IADAEED-I-IIIIAEE
3229	donor	within intron	0.008	0.000	0.000	ADAEEDI-I-IIIAEED
3262	acceptor	within intron	0.191	0.009	0.000	DAEEDII-I-IIAEEED
3366	acceptor	within intron	0.096	0.000	0.000	AEEDIII-I-IAEEDIA
3443	acceptor	within intron	0.002	0.000	0.000	AEDIIII-I-AEEDIAD
3468	acceptor	intron 6 end	0.720	0.320	0.530	ADIIIII-A-EEDIADI
3472	acceptor	within exon	0.003	0.000	0.000	DIIIIIA-E-EDIADII
3527	acceptor	within exon	0.003	0.000	0.000	ADIIIAE-E-DIADIII
3531	donor	intron 7 start	0.291	0.123	0.742	DIIIAEE-D-IADIIIA
3560	donor	within intron	0.342	0.171	0.000	IIIAEED-I-ADIIIAD
3617	acceptor	intron 7 end	0.904	0.198	0.742	IIAEEDI-A-DIIIAED
3714	donor	intron 8 start	0.831	0.002	0.742	IAEEDIA-D-IIIAEDA
3737	acceptor	within intron	0.015	0.000	0.000	AEEDIAD-I-IIAEDAE
3754	acceptor	within intron	0.972	0.100	0.000	AEDIADI-I-IAEDAEE
3769	acceptor	within intron	0.191	0.010	0.000	ADIADII-I-AEDAEEE
3804	acceptor	intron 8 end	0.031	0.002	0.742	DIADIII-A-EDAEEED
3826	donor	within exon	0.030	0.000	0.000	IADIIIA-E-DAEEEDI
3967	donor	intron 9 start	0.905	0.403	0.875	ADIIIAE-D-AEEEEED
4064	acceptor	intron 9 end	0.868	0.720	0.799	IIAEEED-A-EEEEEED
4096	acceptor	within exon	0.005	0.000	0.000	IAEEEDA-E-EEEEEDI

Examples of ρ-value calculations. (i) Acceptor site at 3527. No potential donor site is found in the limits 3408 (60 bases to the left of the higher scoring acceptor site at 3468) to 3527 – 60. Thus, by definition, ρ = 0. (ii) Donor site at 3531. The maximally scoring acceptor site in the limits 3531 + 60 to 3714 + 60 occurs at 3754. Alternative donor sites in the bounds 3754 – 600 to 3754 occur at 3229, 3560 and 3714. The first two sites also pair with the acceptor site at 3754, while the 3714 site pairs with the acceptor site at 3804 (not the site at 4064 which pairs more favorably with the donor site at 3967). By equation **3**, $\rho = (0.291 \times 0.972)^2/(0.291 \times 0.972 + 0.008 \times 0.972 + 0.342 \times 0.972 + 0.831 \times 0.031) = 0.123$. (iii) Acceptor site at 3754. The maximally scoring donor site complement occurs at 3560 with *P*-value 0.342. Alternative acceptor sites are found at positions 3617, 3737, 3769 and 3804 (sites further downstream are favorably paired to the donor site at 3967 and thus do not enter the calculations). Matching these sites with their respective most favourable donor sites leads to: $\rho = (0.972 \times 0.342)^2/(0.972 \times 0.342 + 0.904 \times 0.750 + 0.015 \times 0.342 + 0.191 \times 0.342 + 0.031 \times 0.831) = 0.100$. Example of γ-value calculations: (i) donor site at 3531. The maximal scoring assignment of donor (D) and acceptor (A) sites and exonic (E) and intronic (I) sites is shown in the last column with score: 0.008 + 0.720 + 0.291 + 0.904 + 0.831 + 0.031 = 2.785. Under the hypothesis that this site is not used, the maximal scoring parse is DIIIAEE-E-DIIIAAEE with score 2.042. Thus, γ = 2.785 – 2.043 = 0.742. (ii) Acceptor site at 3754. In this case, the maximal assignment indicates 3754 as within-intron. The maximal scoring parse displayed in the last column has score 3.833, greater than the score 3.090 of the best assignment IAEDIII-A-EEEDAEE involving 3754 as a true acceptor site. Thus, γ = 0. In this case, usage of 3754 as an acceptor site would preclude usage of 3714 as a donor site and leads to the less favorable overall splicing pattern. Note that in the entire sequence segment analyzed, the true splice sites all have γ > 0 and the false sites all have γ = 0. By *P*-value alone, the false sites 3560 and 3754 would be preferred over the nearby true sites.

RESULTS

Evaluation of local optimality can improve splice site prediction

Table 1 gives a specific example of improved splice site prediction based on the ρ- and γ-values in addition to the *P*-values. Prediction using the *P*-values alone as described by Kleffe *et al.* (13) indicates eight potential donor sites and 17 potential acceptor sites in the genomic sequence of maize *Adh1-F* from the fifth

Table 2. Sensitivity and specificity of splice site recognition

SN	*P*		ρ		γ	
	min	*SP*	min	*SP*	min	*SP*
Donor sites						
0.95	0.058	0.46	0.003	0.54	0.022	0.57
0.90	0.133	0.58	0.013	0.67	0.107	0.67
0.80	0.273	0.66	0.044	0.77	0.346	0.82
Acceptor sites						
0.95	0.049	0.39	0.001	0.40	n/a	
0.90	0.119	0.50	0.010	0.60	0.002	0.52
0.80	0.253	0.65	0.031	0.72	0.162	0.79

The 46 maize genes were scanned for splice sites between the known initiation and stop codons. Sites were accepted at the minimal *P*-values attained by the true splice sites as described previously (13). Thus identified were 201 true and 512 false donor sites and 204 true and 1359 false acceptor sites. For the γ column, only comparable sites flanked on each side by at least seven other potential sites were included (155 true and 379 false donors, 157 true and 954 false acceptors). Sensitivity and specificity are defined as $SN = TP/(TP + FN)$ and $SP = TP/(TP + FP)$, respectively, where *TP* is the number of true positives, *FN* is the number of false negatives (i.e. true sites missed) and *FP* is the number of false positives. The criterion was to accept a site as true if its value is at least the value given in the columns 'min'. For example, accepting all acceptor sites with ρ ≥ 0.031 gave 80% sensitivity and 72% specificity. n/a: any γ > 0 gave <95% acceptor site sensitivity.

intron until the translation stop codon. Ten of the 15 false sites have ρ-values of zero, indicating that these sites are locally incompatible with higher scoring alternatives. The γ-values are zero for all but the 10 true sites. In particular, based on the *P*-values alone, the false donor site at 3560 and the false acceptor site at 3754 would be problematic as both score better than the neighboring true sites. In contrast, the overall splicing pattern as measured by γ suggests the correct sites. In detail, usage of the high scoring donor site at 3714 would preclude usage of the nearby acceptor site at 3754, then the acceptor site at 3617 would be locally optimal and would be paired with the donor site at 3531, rather than with the higher scoring site at 3560, which is excluded as being too close.

Distribution of splice site scores

To show the extent of possible improvement in splice site prediction accuracy using the ρ- and γ-values, the values were calculated for all potential sites (true and false) in the maize data set predicted by the previous method at 100% sensitivity based on *P*-values only (13). The results are summarized in Table 2 in terms of prediction sensitivity and specificity relative to this set of sites. It is seen that, at sensitivity levels between 80 and 95%, prediction based on ρ- or γ-values would give increased specificity compared to prediction based on *P*-values. For example, at 80% sensitivity, specificity for both donor and acceptor site prediction could be improved to ~80% using the γ-values compared to ~65% using the *P*-values. Complementary results are shown in Table 3. For high levels of specificity, decision rules for splice site prediction based on the ρ- or γ-values show increased sensitivity levels compared to the corresponding numbers for tests based on the *P*-values. Thus, the context information measured by ρ and γ appears helpful in reducing the numbers of false positive predictions.

Table 3. Specificity and sensitivity of splice site recognition

SP	*P*		ρ		γ	
	min	*SN*	min	*SN*	min	*SN*
Donor sites						
0.90	0.760	0.49	0.174	0.59	0.600	0.63
0.80	0.520	0.68	0.054	0.75	0.300	0.83
0.65	0.240	0.81	0.010	0.91	0.085	0.92
0.50	0.075	0.94	0.002	0.97	0.001	0.96
Acceptor sites						
0.90	0.710	0.51	0.148	0.57	0.570	0.54
0.80	0.500	0.63	0.063	0.71	0.192	0.79
0.65	0.250	0.81	0.015	0.87	0.071	0.84
0.50	0.110	0.91	0.003	0.95	0.001	0.92

The data were derived from the 46 maize genes as described in the legend to Table 2. Abbreviations are as in Table 2.

Empirical rules for splice site prediction

The three scores calculated for a splice site are positively correlated: a site with high *P*-value is likely to be locally optimal and to fit in the locally maximal splicing pattern leading to high ρ- and γ-values, respectively. It is possible for only one or the other score to be high. Examples of sites with high *P*-value but γ = 0 were discussed in the legend to Table 1. Similarly, a site with small *P*-value may have a relatively high γ-value if the site is the optimal partner of a high scoring complementary site (e.g. site 3804 in Table 1). We therefore propose to evaluate splice sites on the basis of all three variables. A variety of statistical techniques could be used to derive simple multivariate functions that discriminate between true and false sites in the training set, including the logitlinear models, neural networks, and discriminant analysis used in previous approaches based on other variables (12–14). These techniques would only optimize the average accuracy of prediction and the relative weighting of the three variables would be difficult to interpret biologically. We therefore pursued a more empirical approach.

Our new SplicePredictor program prints out for each site the *P*-, ρ- and γ-value as well as the optimal parse associated with the γ-value. To assess quickly the overall quality of a site we implemented a * grading system: the values of *P*, ρ and γ are labeled 5*, 4*, 3* or 2* if they match or exceed the threshold values for 90, 80, 65 and 50% prediction specificity on the training set, 1* otherwise (see Table 3 for the threshold values for maize). The sum of the *-values (σ, attaining values between 3* and 15*) serves as a simple combined measure. σ0* is assigned to sites scoring below the *P*-value threshold that selects the sites displayed in the SplicePredictor output. The extreme cases in the σ*-system separate true and false predictions dramatically. For example, for the maize set, the ratio of true to false sites is 1/150 for σ3* donors and 2/722 for σ3* acceptors, compared to 60/2 for σ15* donors and 67/1 for σ15* acceptors. Combining different σ-values, we classify sites as follows:

σ3* – 4*	doubtful	(specificity <3%)
σ5* – 7*	uncertain	(specificity 10–20%)
σ8* – 10*	possible	(specificity 35–45%)
σ11* – 13*	likely	(specificity 60–80%)
σ14* – 15*	highly likely	(specificity >85%)

Table 4. Improved specificity of splice site prediction using the multi variate σ* scoring system

	Maize					*Arabidopsis*				
σ	*TP*	*FP*	*SN*	*SP*	*P-SP*	*TP*	*FP*	*SN*	*SP*	*P-SP*
Donor sites										
14*	105	4	0.55	0.96	0.89	363	43	0.65	0.89	0.87
11*	148	25	0.77	0.86	0.71	460	121	0.82	0.79	0.73
8*	177	85	0.92	0.68	0.55	509	293	0.91	0.63	0.53
5*	190	226	0.99	0.46	0.31	540	667	0.97	0.45	0.34
3*	192	534	1.00	0.26	0.28	558	2734	1.00	0.17	0.18
Acceptor sites										
14*	99	5	0.51	0.95	0.87	280	24	0.50	0.92	0.89
11*	139	30	0.72	0.82	0.76	419	87	0.75	0.83	0.72
8*	166	80	0.86	0.67	0.55	487	265	0.87	0.65	0.56
5*	187	217	0.96	0.46	0.36	537	695	0.96	0.44	0.44
3*	194	1433	1.00	0.12	0.13	561	4720	1.00	0.11	0.12

Sites were evaluated in the coding regions of the non-redundant gene sets described in (13) extended by 500 bases preceding and succeeding the start and stop codons, respectively. Sites with insufficient context for γ-value determination were excluded as in Table 2. Performance values are cumulative; for example, there were 148 true maize donor sites with σ* value at least 11*. *TP*, true positives, *FP*, false positives; *SN*, sensitivity; *SP*, specificity based on the σ* predictions; *P-SP*, specificity based on *P*-value predictions at the same sensitivity level.

(specificity calculated non-cumulatively within the indicated σ*-value class only; e.g. about 10–20% of σ5* – 7* sites are true sites rather than false positives). Table 4 shows the improved cumulative specificity for the σ*-system compared to prediction based on *P*-values alone. The inclusion of the ρ- and γ-values is seen to yield gains in specificity of up to 15%, particularly for the medium scoring site classes.

Conundrums and limitations: poorly scoring true splice sites and highly scoring false splice sites

The overall performance of SplicePredictor appears satisfactory given its statistical nature, the simplicity of the underlying model, and the lack of knowledge about the precise features and mechanisms of splice site recognition *in vivo*. It should be instructive to investigate more closely the instances where model predictions and annotated splicing patterns appear at variance. For our purposes, 'true' splice sites were initially defined as sites reported in the literature and GenBank annotation based on experimental evidence. The predicted gene products were checked for absence of internal stop codons, and GenBank entries with unclear annotation were not used (20). Discrepancies between model predictions and the annotated splicing patterns may arise from a variety of reasons. For example, it is possible that high scoring alternative splice sites are indeed used in a minor fraction of transcripts. Processing of some pre-mRNAs may involve specific splicing factors that positively select one site over another or mask alternative sites. More likely is failure of the prediction algorithm because of its oversimplicity. Detailed examination of all cases would clearly be beyond our capacities. But we hope that theoretical predictions of splice sites will become a common tool in the hands of biologists to analyze and interpret their experimental results; confirmed shortcomings of current models will point to necessary refinements in the models.

These issues are illustrated in Figure 2 with two specific examples. Analysis of the maize *Aux311* gene, encoding an auxin-binding protein, reveals the occurrence of high scoring alternative sites within the exceptionally long first and last introns (1612 and 1525 bases, respectively). The first intron contains a σ13* acceptor site at position 1251. The model cannot explain why this site is apparently not used in conjunction with the native donor site at 1010. If it were, the native acceptor site at 2621 would still have reasonably scoring possible donor site partners at locations 1353, 1987, 2180 and 2276. Similarly, the fourth intron would be predicted to contain an exon at 3922–4094, flanked by a σ14* acceptor site and a σ12* donor site. The auxin-binding proteins in maize are encoded by a multigene family, the known members of which conserve the number and length of the exons but differ considerably in intron lengths (21). It is possible that the differences in introns and alternative splicing patterns may play a role in differential expression of these genes. One use of SplicePredictor may be to suggest suitable probes for the detection of alternatively spliced transcripts in RNase protection experiments.

Inspection of splice site scores in the maize *waxy* locus suggests the possibility of alternative transcripts lacking exons 7 and 8 because the acceptor sites of introns 6 and 7 are very low scoring. All other introns are well defined by high scoring sites (including the first intron in the 5′ untranslated part of the pre-mRNA). Experimentally, no alternatively spliced cDNAs have been reported for the wildtype *waxy* gene, but exon skipping occurred in mutants containing retrotransposon insertions into introns (22): *wx-Stonor* (insertion into intron 5) gave rise to some transcripts linking exon 5 to either exon 8, 9 or 12; *wx-B5* (insertion into intron 2) displayed transcripts linking exon 1 to exons 3 or 4; and *wx-G* (insertion into intron 8) had transcripts linking exons 6 or 7 to either 9 or 12. The apparent local effect of the retrotransposon insertions may be related to alterations in the sequence of splicing events during pre-mRNA processing (22). The accumulation of incompletely spliced transcripts in plant tissues has been reported for a number of plant genes and probably reflects inefficient normal splicing (23).

Prediction of introns in untranslated pre-mRNA

Our gene collections contained three occurrences in maize and several in *Arabidopis* of introns within 5′ or 3′ untranslated segments of the pre-mRNA. We were interested in the scores of splice site prediction methods for these introns, anticipating failure of methods that rely much on evaluation of coding potential. For maize, the *waxy* first intron has donor and acceptor sites in the σ9* class (Fig. 2), the *shrunken-2* first intron has σ13 and 15* sites, and the sucrose synthetase first intron has σ4* sites. The data for *Arabidopsis* are displayed in Table 5. It is evident that these sites are predicted well with the general model, suggesting that there are no special rules governing intron recognition in non-coding pre-mRNA. In particular, the typical base compositional contrast seems to be generally conserved also for comparisons of exons and introns in untranslated regions. With the exception of the 3′ intron in alternatively spliced transcripts of ATRPB1, all listed *Arabidopsis* sites also appear in the NetPlantGene prediction lists. This attests to the rather clever ability of the trained neural networks to weigh different features appropriately in such a way that the lack of a strong coding signal apparently does not prevent site prediction (12). The GENSCAN algorithm (24), on the other hand, does not incorporate prediction of introns in non-coding regions. GENSCAN output for the given GenBank files includes

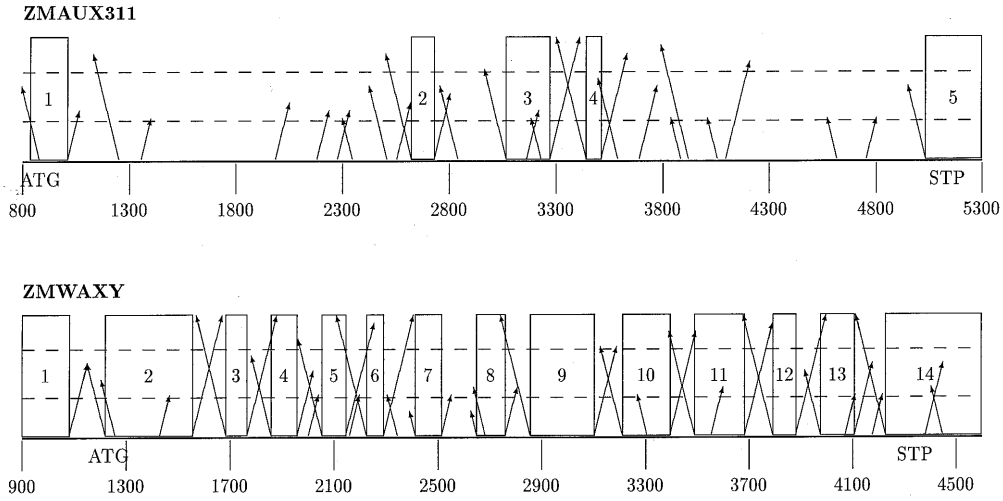

Figure 2. Splice site prediction for the maize *Aux311* gene encoding an auxin-binding protein (GenBank ZMAUX311) and the maize *waxy* locus (GenBank ZMWAXY). Exons are indicated by the numbered boxes. The location of the translation start and stop codons are labeled ATG and STP respectively. The locations of predicted donor sites are indicated by arrows pointing to the right and the locations of predicted acceptor sites are indicated by arrows pointing to the left. The length of an arrow is proportional to the σ* score. The dashed lines correspond to the thresholds σ5* and σ11*. Predicted splice sites scoring less than σ5* were omitted for clarity, with the exception of the true acceptor sites in waxy introns 6 and 7.

Table 5. Splice site prediction in untranslated segments of *A.thaliana* genes

LOCUS	donor site						acceptor site					
	pos	*P*	*L*	X_U	X_{GC}	σ*	pos	*P*	*L*	X_U	X_{GC}	σ*
Introns in the 5′ untranslated region												
ATANT	218	0.962	-3.52	-0.14	0.14	15	675	0.821	-5.23	0.22	-0.12	15
ATH3G	692	0.916	-3.63	-0.18	0.06	15	1100	0.976	-6.27	0.22	-0.20	15
ATHETR1A	366	0.935	-0.87	-0.18	0.12	15	734	0.627	-5.12	0.08	0.04	14
ATHPHYTOA	2001	0.664	-3.91	-0.06	0.04	14	2922	0.971	-8.27	0.38	-0.20	12
Introns in the 3′ untranslated region												
ATHATSA1	3205	0.070	-7.22	0.02	0.10	6	3298	0.967	-3.85	0.26	0.08	15
ATPOSF21	2049	0.980	0.10	-0.24	0.10	15	2175	0.008	-9.32	0.00	0.06	3
							2134	0.817	-5.66	0.18	-0.18	15
ATRPCL9G	2128	0.885	-4.50	-0.12	0.14	15	2375	0.674	-6.51	0.16	-0.02	13
ATRPB1	7671	0.235	-5.42	-0.08	0.00	8	7758	0.035	-8.49	0.06	0.06	5

Positions are given relative to the GenBank files indicated in the LOCUS column. The 3′ site in ATPOSF21 is given as 2175 in the GenBank annotation based on cDNA data; also listed is a predicted site at 2134, scoring much better than 2175. The intron in ATRPB1 occurs in alternatively spliced transcripts. See GenBank for references.

some of the splice sites in the untranslated regions as part of wrongly predicted coding exons (data not shown).

Applications to *Arabidopsis* genes

Using our database of 131 non-redundant *Arabidopsis* genes we derived data corresponding to Tables 2 and 3 for *Arabidopsis* as a model dicot (data not shown). The contributions of the ρ- and γ-values to reduce false positive prediction rates were found to be of similar magnitude as those reported for the maize data (Table 4). The *-system threshold values specific to *Arabidopsis* are incorporated into the SplicePredictor program and are invoked by the appropriate command line species choice.

Splice site clusters identify potential genes

For both the maize and the *Arabidopsis* training sets, σ14–15* splice sites are highly likely to be true sites (>85% specificity in transcribed regions). Furthermore, >50% of the true splice sites in the training sets are in this score class. These numbers suggest that clusters of high scoring predicted splice sites would likely correspond to split genes. Assume a gene with four exons. Taking a length of 300 bases for each exon and intron (>80% of maize and *Arabidopsis* exons and introns are at most 300 bases; 11,19), the splice sites would occur within a 1500 base segment comprising all three introns and the two interior exons. We would expect three of the six splice sites to be in the σ14–15* score class

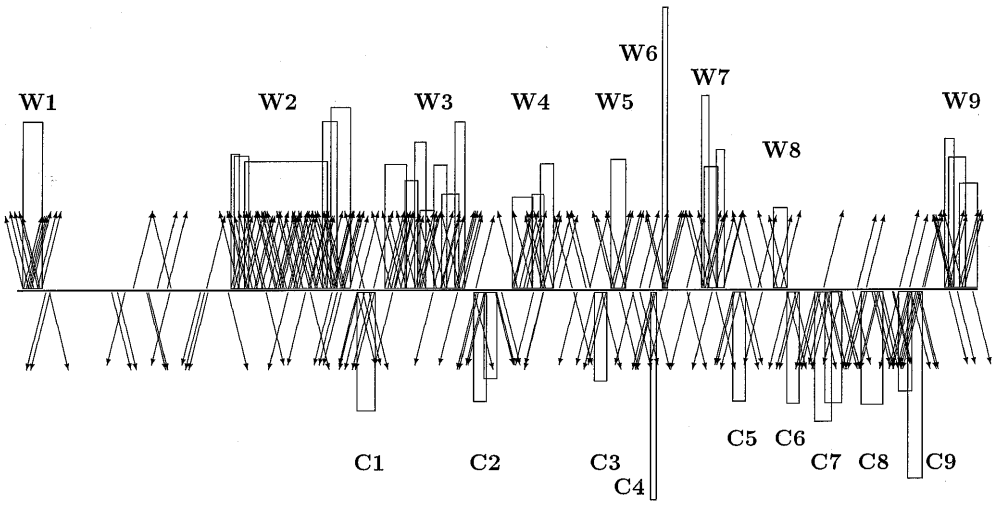

Figure 3. Clusters of predicted splice sites in the *A.thaliana* contig BAC T7123 (106 973 bases; GenBank accession no. U89959). The locations of predicted donor sites are indicated by arrows pointing to the right on the + strand and to the left on the – strand and the locations of predicted acceptor sites are indicated by arrows pointing to the left on the + strand and to the right on the – strand (i.e. on each strand, predicted splice sites are pointing into potential introns). Only highly likely sites ($\sigma \geq 14$*) are shown. Clusters on the + strand are indicated by boxes above the line, and clusters on the – strand are indicated by boxes below the line. The area of a box is proportional to the number of sites in the corresponding sequence segment. Continuous regions of overlapping clusters are labeled W1–W9 on the + strand and C1–C9 on the – strand.

and maybe one additional false site. Thus, a criterion of four or more σ14–15* splice sites within a sequence segment of at most 1500 bases is taken to define a splice site cluster.

The SplicePredictor program optionally prints all splice site clusters satisfying the above (or another, user-defined) criterion. Overlapping clusters are merged unless the start site of the downstream cluster is within the last 750 bases of the upstream cluster and the endpoints of the two clusters are different. For this definition, splice site clusters are identified in 75 of 80 *Arabidopsis* genes with at least four exons searched in the region of 500 bases upstream of the start codon to 500 bases downstream of the stop codon. As a control, when the same sequence segments were searched on the complementary strand, only one cluster was identified. This number may not be representative for intergenic regions.

To test the use of splice site clusters for gene identification we applied the SplicePredictor program with a σ14* threshold to the BAC T7I23 *Arabidopsis* contig (106 973 bases of genomic DNA; GenBank accession no. U89959). On each strand, nine continuous regions of splice site clusters were identified (Fig. 3). There are no overlaps between clusters on opposite strands. The clusters may serve as a rough segmentation of the genomic sequence into potential transcription units. In general, our GeneGenerator program (25) may be used to produce a number of alternative gene predictions in a tentative transcription unit. In many cases, these predictions can be adjusted and confirmed by comparison with existing protein and cDNA databases, for example with a combination of the BLAST (26) and PROCRUSTES (27) programs. For the T7I23 contig, we found the highest scoring potential split gene in each of the splice site cluster segments (the boundaries of each segment were extended by about 2000 bases beyond the ends of the identified cluster, giving limits which should in most cases include the entire gene). The predicted translation products were then compared against the databases using the BEAUTY database search server (28). The gene predictions were refined by further processing the segments in the order determined by the highest scoring database similarities. This 'divide and conquer' strategy seeks to establish first the genes for which there is the most evidence (i.e. database similarities that may include previous determination of the particular gene, either as genomic DNA or cDNA, or of homologs in other species). The established predictions then reduce the boundaries for further gene searches.

The predictions for the central region of the T7I23 contig encompassing splice site clusters C2 to W6 of Figure 3 are displayed in Table 6. The displayed region is bounded by the well established genes encoding the *ara-5* and CER1-like/CER1 proteins in clusters C2 and W5/W6, respectively. The second predicted gene in C2 is confirmed beyond doubt by very strong similarity to other *mago nashi* proteins. The other gene predictions are more tentative. However, several exons are strongly suggested by sequence similarity. Sequence alignments supporting the Table 6 predictions are available on the WWW page http:/gnomic.stanford.edu/~volker/Arabidopsis/BAC-T7I23/BAC-T7I23.html/ Other exons are predicted with confidence as a result of excellent splice site scores. In particular, the displayed assignments include all but three of the σ14* and 15* sites on the predicted coding strands in this region (note however, that no genes were predicted for splice site clusters C3 and C4 on the opposite strand).

A stochastic gene assembly program (*Exdomino*, V.Brendel and J.Kleffe, unpublished) was used to find the best gene structures incorporating all strongly predicted exons and splice sites. Current programs for gene and splice site prediction failed to identify several of the exons that are strongly suggested by sequence similarity. For example, GENSCAN (24) identifies only the first two exons of the *mago nashi* gene, and given sufficient sequence context the program links these exons to exons 3–8 of the preceding *ara-5* gene. Similarly, GENSCAN combines exons

Table 6. Gene prediction for the central region of the *A.thaliana* contig BAC T7I23

Cluster	Gene prediction exon	from	to	σ* 3′	σ* 5′	%idty	Identity/Evidence/Comments
C2.1	1	52176	52163	-	**14** .	100	**ara-5**, small GTP-binding protein (GenBank
	2	52055	51983	10 .	**14** .	96	Accession D01027); high similarity to the
	3	51896 .	51849 .	10 .	12 .	100	mRNA of a pea homolog (GenBank D12549)
	4	51758 .	51711 .	12 .	9 .	100	
	5	51623 .	51552 .	11 ..	13 .	100	
	6	51474 .	51319 .	8 .	**15** ..	100	
	7	51216 .	51113 .	7 .	**15** ..	100	
	8	50834 .	50738 .	**15** .	-	100	
C2.2	1	54789 .	54574 .	-	13 .	74	**mago nashi**; high similarity to a *Drosophila*
	2	54335 .	54159 .	7 .	5 .	75	homolog
	3	53914	53855	**14** .	-	95	
W4.1	1	55015	55057	-	**15** ..	-	significant similarity to the hypothetical proteins
	2	55145	55230	12 .	**15** ..	50	ZC513.5 of *C. elegans* (GenBank U53155)
	3	55300	55341	0	8 .	36	and yeast YNR030w (GenBank Z71645)
	4	55453	55498	10 .	5 .	87	
	5	55618	55704	**15** .	10 .	-	
	6	55818	55873	10 .	**15**	-	
	7	56036	56153	3	13	44	
	8	56281	56339	12 .	**14**	32	
	9	56449	56520	3	4	38	
	10	56635	56698	11	**15** .	-	
	11	56803	56906	7	0	44	
	12	57109	57180	3	11 .	50	
	13	57291	57377	**15** .	9	59	
	14	57468	57535	6 .	9	64	
	15	57709	57777	7 .	9	-	
	16	57981	58018	11 .	9 .	-	
	17	58110 .	58209 .	10 .	**15**	-	
	18	58461 .	58526 .	**14** .	**14** .	-	
	19	58796	58799	**15** .	-	-	
W4.2	1	59167	59526 .	-	**15** ..	20	significant similarity to the predicted *Arabidopsis* protein
	2	59607 .	60821 .	**15** ..	-	29	F8A5.28 (GenBank AC002292) and other sequences; cDNA: exon 2 (ATTS1091, ATTS1259, ATTS1092)
W4.3	1	61191 .	61281 .	-	4 .	-	similarity to yeast proteins COX17 (GenBank L75948) and
	2	61565 .	61641 .	13 .	9	56	hypothetical yeast protein YHR6 (GenBank 731704);
	2	61724	61771	**14** ..	-	-	cDNA: exons 1–3 (AA712564)
W4.4	1	62310	62646 .	-	5 .	29	high similarity to another putative *Arabidopsis* gene on
	2	62848 .	63178 .	10 .	12 .	69	chromosome 5 (GenBank AB008268) 20662-22033; exon 2 is
	3	63271 .	63391 .	**15** ..	9	32	49% identical to the central part of yeast YORF197w
	4	63468	63683 .	3	**15** ..	28	(GenBank Z75105); cDNA: exon 1 (H76651)
	5	63755	63814	**15** .	-	58	
W5	1	65542	66058 .	-	**15** .	97	CER1-like protein; GenBank ATHCER1L19, ATCER1L,
	2	66130 .	66349 .	**15** .	**15** .	94	and ATHCER1LB; the 67349 acceptor site
	3	66433 .	66823 .	**14** .	13 ..	100	scores below minimal threshold; the program
	4	67131 .	67226 .	8 .	**15** ..	100	predicts the site 67406 instead
	5	67350 .	67550 .	0 .	**15** ..	100	
	6	67636 .	67908 .	**15** ..	13 ..	100	
	7	68011	68184	12 .	-	100	
W6	1	69206	69265	-	**15** .	100	CER1, maize *gl1* homolog; GenBank ATU40489 and
	2	69375	69598	0	13 .	100	ATHCER1; the 69374 acceptor site scores below the default
	3	69685	69917	6 .	12 .	100	threshold; the program predicts a different first
	4	70063	70282 .	11 .	**15** ..	100	exon, 69312-69598
	5	70902	71277	12 .	7 .	99	
	6	71372 .	71479	7 .	6 .	100	
	7	71579 .	71779 .	11 .	**15** .	100	
	8	71914 .	72021 .	**14** .	**15** ..	100	
	9	72123	72296 .	**15** .	**15** ..	98	
	10	72369 .	72542	10 .	- .	96	

Potential genes were initially identified in the extended splice site cluster regions 55000–48000 (C2), 53000–62000 (W4), 68000–62000 (C3), 64000–70000 (W5), 73000–68000 (C4) and 70000–74000 (W6). Sites labeled with dots in columns three and four are also predicted by the GENSCAN program (24). Sites labeled with dots in columns five and six are also predicted by the NetPlantGene program (12); two dots indicate the designation H for highly probable sites according to this program. %idty refers to the percent amino acid identity with a corresponding segment of the same length in one of the target proteins listed in the last column. Alternative gene structure predictions were resolved by maximizing similarity to the target proteins and usage of high-scoring splice sites as explained in the text.

17 and 18 of the W4.1 gene of Table 6 with most of the two exon W4.2 gene into one predicted gene, but does not involve any of the other exons. The *Exdomino* prediction for the C-terminal exons of the W.1 gene was driven to include the high scoring splice sites defining the putative exons 17–19. In this case, the different possibilities are not clearly resolved by similarity to other sequences, and in general the initial and terminal exons are most difficult to predict due to a lack of adequate models for intergenic regions (1,2). This presents a particular problem in species like *Arabidopsis* for which genes are typically closely spaced.

The value of using sequence similarity for improved predictions were clearly demonstrated by Gelfand *et al*. (27) with their PROCRUSTES algorithm for spliced alignment. For the potential genes of Table 6 and indicated protein targets, the PROCRUSTES program (default settings) gave concordant results for *mago nashi* and the W4.4 gene. For the W4.1 gene, the global spliced alignment failed to indicate significant similarity, whereas the local alignment identified exons 7, 8, 12 and 14.

DISCUSSION

Our understanding of genome organization is challenged by the current limits of predicting gene structure and expression from sequence inspection. Thus, the elucidation of sequence features contributing to accurate and efficient pre-mRNA splicing is of fundamental interest. Moreover, these issues are also of great practical importance for genome projects (sequence annotation) and genetic engineering (design of mutants and transgenic organisms). In plants, cryptically spliced transcripts produce truncated and mutant proteins and reduce protein expression, whereas inclusion of an accurately spliced intron enhances protein expression compared to intronless transcription units (29–31). Little is known, however, about the quantitative aspects of intron enhancement and how to maximize active protein by modulating intron structure.

Prediction of splice sites in plant pre-mRNA by our or other methods includes false positives (high scoring sites that occur within annotated exons and introns) and false negatives (true sites that are missed). Besides the limitations inherent in the statistical methods, there are other complicating factors which contribute to such results. Alternative splicing may validate some of the apparently false positive predictions, and some of the false negatives may include inefficient sites (32–34). Tissue- or gene family-specific regulation of pre-mRNA processing may be obscured in pooling all sequences for training of statistical splice site prediction methods. Alternative processing may also result from changes in physiological state, e.g. regulated lack of splicing of the maize *Bz2* intron after cadmium exposure (35).

A difficulty in the design of successful algorithms for the prediction of splice sites from sequence inspection is the inclusion and proper weighting of the multiple contributing variables to *in vivo* recognition of a particular site. The neural network approach of Hebsgaard *et al*. (12) for splice site prediction in *Arabidopsis* provides an elegant solution to this problem. We have pursued an alternative and complementary approach that extends the previously described logitlinear models (13). The key element of the extensions is the vector representation of splice site scores incorporating three elements: intrinsic splice site quality (*P*-value), local optimality (ρ-value) and fit with respect to locally predicted exon/intron structure (γ-value). The additional variables can in many cases correctly classify potential sites that would be confused on the basis of *P*-values alone (detailed example discussed in Table 1). In particular, true sites with low scores in all three components are highly exceptional, whereas high scores in all components produce very reliable predictions (Table 4).

The importance of reducing the uncertainty in gene prediction schemes is readily appreciated when one is trying to annotate a really novel sequence. Even if a program claims a success rate of, say, 70% correctly identified exons in a training set, for a novel gene prediction of, for example, a gene with eight exons this would leave an expected 8-choose-2 to 8-choose-3 possible combinations for exactly which two or three of the eight predicted exons may be wrong! Any exons or introns that can be assumed with great confidence would reduce these combinatorial possibilities considerably. Potential exons and introns flanked by highly reliable splice site predictions would be the obvious candidates for definite inclusion into entire gene structure predictions. On the other hand, gene prediction algorithms cannot afford to overlook any potential splice sites because of the risk of missing correct assemblies of genomic segments into ORFs. Our proposed solution retains a relatively large set of potential splice sites based on minimal *P*-value requirements, but further differentiates sites within this set based on the described context variables.

The annotation of a typical *Arabidopsis* genomic DNA contig in Table 6 illustrates these considerations. Strong support for the displayed gene predictions comes from the combination of splice site prediction, evaluation of coding potential, cDNA matching and similarity to potential homologs. Methods that rely entirely or mostly on only part of the different types of support are clearly not successful. For example, similarity to a target protein provides confirmation for some weak or possibly even non-consensus splice sites. Conversely, there are also examples of strongly predicted splice sites and introns that define exons with no significant similarity to the target proteins. Such exons may, however, be parts of an entire predicted gene product that displays overall significant similarity to a target protein (e.g. exons 5, 6 and 10 of the W4.1 gene in Table 6), and simply represent less conserved segments of homologous proteins. We are currently developing a flexible algorithm (*Exdomino*) for the prediction of alternative gene structures satisfying any specific constraints, e.g. inclusion of particular splice sites, exons or introns. Identification of suitable constraints, at present, seems to require expert input and can only partly be solved by programming.

Program availability

The databases and the SplicePredictor program, which implements our current algorithm for splice site prediction in plant genes, are available electronically from either V. Brendel (volker@gnomic.stanford.edu) or J. Kleffe (jkleffe@euler.grumed.fu-berlin.de). SplicePredictor is also implemented as a Web service at http://gnomic.stanford.edu/ ~volker/SplicePredictor.html

ACKNOWLEDGEMENTS

The authors wish to thank Prof. Virginia Walbot for discussions and critical reading of the manuscript. V.B. was supported in part by NIH grant 5R01HG00335-10.

REFERENCES

1 Burge,C.B. and Karlin,S. (1998) *Curr. Opin. Struct. Biol.*, **8**, 346–354.
2 Claverie,J.-M. (1997) *Hum. Mol. Genet.*, **6**, 1735–1744.
3 Fickett,J.W. (1996) *Trends Genet.*, **12**, 316–320.
4 Burset,M. and Guigó,R. (1996) *Genomics*, **34**, 353–367.
5 Brown,J.W.S. and Simpson,C.G. (1998) *Ann. Rev. Plant Physiol. Plant Mol. Biol.*, **49**, 77–95.
6 Simpson,G.G. and Filipowicz,W. (1996) *Plant Mol. Biol.*, **32**, 1–41.
7 Goodall,G.J. and Filipowicz,W. (1989) *Cell*, **58**, 473–483.
8 Lou,H., McCullough,A.J. and Schuler,M.A. (1993) *Mol. Cell. Biol.*, **13**, 4485–4493.
9 McCullough,A.J., Lou,H. and Schuler,M.A. (1993) *Mol. Cell. Biol.*, **13**, 1323–1331.
10 Luehrsen,K.R., Taha,S. and Walbot,V. (1994) *Prog. Nucleic Acids Res. Mol. Biol.*, **47**, 149–193.
11 Brendel,V., Carle-Urioste,J.C. and Walbot,V. (1998) In Bailey-Serres,J. and Gallie,D.R. (eds), *A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants*. Am. Soc. Plant Physiol., Rockville, MD, pp. 20–28.
12 Hebsgaard,S.M., Korning,P.G., Tolstrup,N., Engelbrecht,J., Rouzé,P. and Brunak,S. (1996) *Nucleic Acids Res.*, **24**, 3439–3452.
13 Kleffe,J., Hermann,K., Vahrson,W., Wittig,B. and Brendel,V. (1996) *Nucleic Acids Res.*, **24**, 4709–4718.
14 Solovyev,V.V., Salamov,A.A. and Lawrence,C.B. (1994) *Nucleic Acids Res.*, **22**, 5156–5163.
15 Thermann,R., Neu-Yilik,G., Deters,A., Frede,U., Wehr,K., Hagemeier,C., Hentze,M.W. and Kulozik,A.E. (1998) *EMBO J.*, **17**, 3484–3494.
16 Nagy,E. and Maquat,L.E. (1998) *Trends Biochem. Sci.*, **23**, 198–199.
17 Shapiro,M.B. and Senapathy,P. (1987) *Nucleic Acids Res.*, **15**, 7155–7174.
18 Brendel,V., Kleffe,J., Carle-Urioste,J.C. and Walbot,V. (1998) *J. Mol. Biol.*, **276**, 85–104.
19 Brown,J.W.S., Smith,P. and Simpson,C.G. (1996) *Plant Mol. Biol.*, **32**, 531–535.
20 Korning,P.G., Hebsgaard,S.M., Rouzé,P. and Brunak,S. (1996) *Nucleic Acids Res.*, **24**, 316–320.
21 Schwob,E., Choi,S.Y., Simmons,C., Migliaccio,F., Ilag,L., Hesse,T., Palme,K. and Söll,D. (1993) *Plant J.*, **4**, 423–432.
22 Varagona,M.J., Purugganan,M. and Wessler,S.R. (1992) *Plant Cell*, **4**, 811–820.
23 Nash,J. and Walbot,V. (1992) *Plant Physiol.*, **100**, 464–471.
24 Burge,C. and Karlin,S. (1997) *J. Mol. Biol.*, **268**, 78–94.
25 Kleffe,J., Hermann,K., Vahrson,W., Wittig,B. and Brendel,V. (1998) *Bioinformatics*, **14**, 232–243.
26 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
27 Gelfand,M.S., Mironov,A.A. and Pevzner,P.A. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 9061–9066.
28 Worley,K.C., Wiese,B.A. and Smith,R.F. (1995) *Genome Res.*, **5**, 173–184.
29 Callis,J., Fromm,M. and Walbot,V. (1987) *Genes Dev.*, **1**, 1183–1200.
30 Sinibaldi,R.M. and Mettler,I.J. (1992) *Prog. Nucleic Acids Res. Mol. Biol.*, **42**, 229–257.
31 Schuler,M.A. (1998) In Bailey-Serres,J. and Gallie,D.R. (eds), *A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants*. Am. Soc. Plant Physiol., Rockville, MD, pp. 1–19.
32 Jarvis,P., Belzile,F. and Dean,C. (1997) *Plant J.*, **11**, 921–931.
33 Martin,D.J., Firek,S., Moreau,E. and Draper,J. (1997) *Plant J.*, **11**, 933–943.
34 Nishihama,R., Banno,H., Kawahara,E., Irie,K. and Machida,Y. (1997) *Plant J.*, **12**, 39–48.
35 Marrs,K.A. and Walbot,V. (1997) *Plant Physiol.*, **113**, 93–102.