Nucleic Acids Research, 1996, Vol. 24, No. 23 **4709–4718**

Logitlinear models for the prediction of splice sites in plant pre-mRNA sequences

Jürgen Kleffe*, Klaus Hermann, Wolfgang Vahrson, Burkhardt Wittig and Volker Brendel[1]

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

Received August 19, 1996; Revised and Accepted October 16, 1996

ABSTRACT

Pre-mRNA splicing in plants, while generally similar to the processes in vertebrates and yeast, is thought to involve plant specific *cis*-acting elements. Both monocot and dicot introns are typically strongly enriched in U nucleotides, and AU- or U-rich segments are thought to be involved in intron recognition, splice site selection, and splicing efficiency. We have applied logitlinear models to find optimal combinations of splice site variables for the purpose of separating true splice sites from a large excess of potential sites. It is shown that plant splice site prediction from sequence inspection is greatly improved when compositional contrast between exons and introns is considered in addition to degree of matching to the splice site consensus (signal quality). The best model involves subclassification of splice sites according to the identity of the base immediately upstream of the GU and AG signals and gives substantial performance gains compared with conventional profile methods.

INTRODUCTION

An important aspect of eukaryotic genome research is scanning functionally unknown sequences for potential split genes. Because experimental capacity is small compared with the rate of accumulation of genomic sequence data, the majority of new sequences must be studied by statistical, computational methods. Current approaches to finding genes and functional sites in DNA sequences have recently been reviewed (1,2). Most algorithms involve identification of potential splice sites (search by signal) as a preliminary step to the task of parsing the sequence into consistently ordered, translatable exons. Ideally, one would like splice site prediction to be based on the same signals that are recognized by the nuclear splicing machinery. Practically, most methods are based on consensus motifs and weight matrices scoring the degree of fit to some average signal pattern around known splice sites in a learning set (e.g., 3–6). Neural network applications were introduced by Brunak, Engelbrecht and Knudsen (7). The construction of decision trees based on categorical discriminant analysis by Sirajuddin *et al.* (8) has recently promised improved donor site recognition.

The general features of splicing appear to be conserved throughout all eukaryotes. The failure of accurate splicing of heterologous introns in transformed plant cells suggested that particular features of plant introns are essential for accurate pre-mRNA processing (for reviews, see 9,10). In particular, several studies have demonstrated that U-rich segments within plant introns influence splice site selection (e.g., 11–13). Most recently, it was shown that the relative contrast in U and G+C content between introns and their flanking exons correlates with splicing efficiency (Carle-Urioste, Brendel and Walbot, submitted). Contrast-enhancing changes within either introns or exons improved splicing efficiency. It was suggested that evaluation of compositional contrast could improve prediction of splice sites.

Here we present a novel algorithm for the prediction of splice sites in higher plants based upon the two variables of splice site signal strength and compositional contrast. Our model seeks to incorporate a minimal set of local sequence properties accessible to the nuclear splicing machinery, and, in particular, does not explicitly consider reading frame and codon usage information. In this way, analysis of false positive as well as false negative predictions may point to missing variables (e.g., branch point consensus, specific sequence motifs; Brendel, Kleffe and Walbot, manuscript in preparation). On the other hand, for the practical task of identifying split genes, incorporation of global coding potential assessment greatly reduces the number of falsely predicted splice sites, as demonstrated in the recent comprehensive study of splice site prediction in *Arabidopsis thaliana* by Hebsgaard *et al.* (14). From a statistical standpoint, our method is a standard application of logitlinear models. We provide a rationale for the applicability of such models for a wide variety of sequence analysis problems.

MATERIALS AND METHODS

Gene collections

Genomic sequences from *Zea mays* and *A.thaliana* were retrieved from GenBank and compiled into specifically annotated

* To whom correspondence should be addressed

non-redundant databases (redundancy as a result of significant sequence similarity was assessed as described in ref. 15). Only completely sequenced genes were included, i.e. those genes for which all introns between the start codon and the stop codon are available. For *Z.mays*, our database contains 46 genes that encode distinct proteins and comprise a total of 250 exons and 204 introns (this database denoted GBEzm). Obvious annotation errors in GenBank entries were corrected. For *Arabidopsis*, a database of 131 distinct genes was obtained with a total of 709 exons and 578 introns (GBEat). In this case, because many genes are available, we simply excluded GenBank entries with likely erroneous annotation. Korning *et al.* (16) offer a detailed account of problems with GenBank entries and provide a cleaned *Arabidopsis* gene set of similar size.

Splice site collections and control sets

Databases for donor and acceptor sites and respective control sets were derived from our gene collections in the following way. Identification of splice and control sites was restricted to the pre-mRNA portions between the known start and stop codons of each gene. This restriction accomplishes several goals. First, it limits the number of control sites, which is already large even prior to scanning the flanking regions and the opposite DNA strand. Second, correct identification of splice sites in this restricted region is of practical interest, because there are now several independent promising methods for predicting eukaryotic promoters and terminators (17–19). These methods, if applied successfully, would similarly limit the search space. Third, in terms of understanding the cellular splicing machinery, the task is, of course, to distinguish true splice sites from non-sites in pre-mRNA, not in genomic DNA.

Our donor site and corresponding control sets consist of all GU dinucleotides occurring within the prescribed sequence bounds and including 50 nt on each side. For each such site, we record the signal sequence and the percent nucleotide composition evaluated separately for the 50 nt in the 5′ and 3′ flanks. The donor signal sequence was chosen to cover the 3 nt 5′ and the 4 nt 3′ to the GU. The donor site data sets comprise 201 true sites plus 6305 control sites for *Z.mays* and 577 true sites plus 14 964 control sites for *Arabidopsis*. In three of the 204 maize introns the donor site consensus GU is replaced by GC. These exceptional sites were not considered for the purposes of this study.

For acceptor sites and their controls, we selected all AG dinucleotides with 50 nt flanks and created corresponding data sets comprising a total of 204 true sites plus 6290 control sites in maize and 577 true sites plus 15 712 control sites in *Arabidopsis*. The acceptor signal sequence was defined to cover 13 nt to the left of AG and 2 nt to the right.

Logitlinear models for splice site recognition

As evident from the sizes of the data sets described above, for every true splice site within a particular pre-mRNA there are on the order of 30 non-sites conserving the consensus GU for donors or AG for acceptors. We wish to derive rules that distinguish the real sites from the bulk of non-sites on the basis of local sequence properties. From a practical standpoint, any such rules would suffice if their general applicability could be demonstrated on a set of new sequences not involved in the derivation of these rules. Beyond this practical goal, we advocate a modeling approach that also seeks to incorporate known and presumed properties of the biochemistry underlying splice site recognition. In this way, the modeling results provide a consistency check on possible factors contributing to splice site recognition by evaluating the predictive usefulness of these factors individually and in various combinations.

The *in vivo* mechanisms of splice site selection are undoubtedly quite complex and details are still largely unknown. Thus, any modeling attempt will necessarily involve considerable simplifications. Here, we consider splice site recognition in terms of a classical dose-response assay. In this case, ‘dose’ refers to measurable characteristics of a site (degree of matching to the extended signal consensus, compositional contrast between the upstream and downstream signal flanking regions, possibly other features). The response is whether or not splicing occurs at that site under standard conditions.

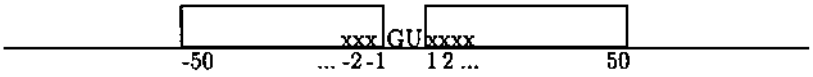
For a given site, let P denote the probability that splicing succeeds. Experimentally P could be measured in terms of splicing efficiency expressed as the proportion of pre-mRNA transcripts that are spliced at that site. All other things being equal, we assume that P depends only on the dose of measured sequence characteristics. For simplicity, first assume that splice site sequence characteristics can be measured in terms of a single variable, x, such that high positive values of x correspond with P-values close to 1 and low negative values of x correspond with P-values close to 0. It is convenient and customary in this case to utilize a sigmoidal curve of the explicit form:

$$P(x) = \frac{e^{\alpha + x}}{1 + e^{\alpha + x}} \qquad \mathbf{1}$$

where $x = -\alpha$ is the dose that gives half-maximal response. More generally, various characteristics of a site are measured simultaneously such that a specific site, i, is represented by the vector $\vec{x}_i' = (1, x_{i1}, x_{i2}, \ldots\ldots, x_{ik})$ of observables, and, in the simplest generalization of **1**, $\alpha + x$ is replaced by the linear combination $\vec{x}_i'\vec{\beta}$, where $\vec{\beta} = (\alpha, \beta_1, \beta_2, \ldots\ldots, \beta_k)'$ is a set of parameters to be estimated. Parameter estimation for this logitlinear model can be carried out in standard fashion as follows (e.g., 20,21). Let Y be the indicator variable, set to 1 for true splice sites and to 0 for non-sites. Let $\vec{y}$ be the vector of observed Y values, consisting of AP ones for true sites and AN zeros for false sites in a given training set (AP and AN stand for actual positives and actual negatives, a notation also used below). Then the joint loglikelihood of the observations $\vec{y}$ is given by

$$L = L(\vec{\beta}) = \sum_{i=1}^{N} \left(y_i \ln P(\vec{x}_i'\vec{\beta}) + (1 - y_i) \ln(1 - P(\vec{x}_i'\vec{\beta})) \right) \qquad \mathbf{2}$$

where $N = AP + AN$ is the total number of sites in the training set. The maximum likelihood estimator for β is found numerically as the unique solution to the system of non-linear equations $X(\vec{y} - \vec{P}) = \vec{0}$, where X is the $(1 + k) \times N$ matrix with columns equal to the vectors $\vec{x}_i$ (X is assumed to be of full rank), and $\vec{P}$ is the vector of probabilities $P(\vec{x}_i'\vec{\beta})$. Many computer programs are available to obtain such parameter estimates numerically, e.g., by application of the Newton–Raphson algorithm (20,21). We used corresponding functions from StatUnit by Tue Tjur (22) as implemented in our freely distributed DNASTAT package (23). Convergence was always achieved after a few steps, but the



$$\delta_{ib} = \begin{cases} 1 & \text{if base at position } i \text{ is } b, \ b = A, C, G, \text{ or } U \\ 0 & \text{otherwise} \end{cases}$$

$$W_s = \sum_{i=-3}^{4} \sum_{b} \delta_{ib} \ln f_{ib} \qquad \text{or} \qquad L = \sum_{i=-3}^{4} \sum_{b} \delta_{ib} l_{ib}$$

$$X_U = \left(\sum_{i=-50}^{-1} \delta_{iU} - \sum_{i=1}^{50} \delta_{iU} \right) /50$$

$$X_{GC} = \left(\sum_{i=-50}^{-1} (\delta_{iG} + \delta_{iC}) - \sum_{i=1}^{50} (\delta_{iG} + \delta_{iC}) \right) /50$$

Figure 1. Definition of donor site variables. The signal sequence is taken to extend 3 nt upstream and 4 nt downstream of the consensus GU. δ_{ib} is an incidence variable, set to one if the nucleotide at position i is b and zero otherwise. f_{ib} is the frequency of nucleotide b in signal position i in the training set of true donor sites. The l_{ib} are signal sequence weights that are estimated as discussed in the text. Base compositional contrast is measured in terms of percent U (X_U) and percent G+C (X_{GC}) usage difference over 50 base flanks upstream and downstream of the GU. Variables for true acceptor sites and for non-sites were assessed similarly as discussed in the text.

calculations were lengthy due to the large number of observations and parameters.

A cautionary note concerning interpretation of the model is in order. Assuming the validity of the model, P would truly be an estimate of splicing efficiency if the training data consisted of repeated observations for each particular combination of site characteristics. In other words, the indicator variable Y would have to be observed repeatedly for each site, and P is simply the estimated success probability for this binomial variable. Such data are commonly not available. Instead, we sample over distinct sites in the pre-mRNA and make only one observation per site (true splice site or non-site). The validity of this sampling scheme as an approximation to the repeated observations sampling scheme rests on the continuity of the P function in its dependence on $\vec{x}'\vec{\beta}$. Thus, we replace sampling of repeated experiments on the same site by sampling of other sites with similar characteristics $\vec{x}$. Sites with scores $\vec{x}'\vec{\beta}$ similar to those of many non-sites are considered poor splice sites; sites with exceptionally high scores are considered efficient splice sites.

There is experimental evidence in support of these modeling assumptions for our choice of independent variables. Thus, it is well documented that improved matching of the splice site consensus increases splicing efficiency and that base compositional changes in both introns and exons affect splice site selection and splicing efficiency in the predicted manner (e.g., 11–13,24–26).

Selection of independent variables

In this paper we consider the prediction of splice sites on the basis of three local sequence properties: (i) the signal sequence, (ii) the U content of 50 nt sections upstream and downstream of the consensus donor GU or acceptor AG, and (iii) the G+C content of these sections. The latter two properties were quantified as the difference in percent nucleotide content and denoted by X_U and X_{GC}, respectively (Fig. 1). These variables are clearly not independent but emphasize different aspects of a potential splice site and its context.

The signal information was evaluated in several alternative ways. First, a nucleotide frequency profile was derived in the usual way from all true donor (acceptor) sites. For a given site, the variable W_s was defined as the sum of log-frequencies taken from the profile, where summation extends over the entries associated with the nucleotides occurring in the given site (Fig. 1). A similar variable W_n was constructed using the profile derived from all non-sites. A logitlinear model based on the four measurements W_s, W_n, X_U and X_{GC} is given by

$$\ln \frac{P}{1-P} = \alpha + \beta W_s + \gamma W_n + \delta X_U + \mu X_{GC} \qquad \mathbf{3}$$

where α, β, γ, δ and μ are the components of the parameter vector $\vec{\beta}$ in the general formulation given above. A special case is $\gamma = -\beta$ in which case the signal variable is the log-likelihood ratio $W_{sn} = W_s - W_n$.

Alternatively, the variables W_s and W_n were replaced by a set of individual factor variables $L_1,\ldots, L_m$, one for each signal position. Each of these factor variables assumes one of the four levels l_{iA}, l_{iC}, l_{iG} or l_{iU} depending on the observed nucleotide in signal position i. The corresponding logitlinear model is

$$\ln \frac{P}{1-P} = \alpha + L + \delta X_U + \mu X_{GC}, \quad L = \sum_{i=1}^{m} L_i \qquad \mathbf{4}$$

The values l_{ib} are unknown parameters to be estimated. In each of the positions i, one of the l_{ib} must be set to some arbitrary value (different choices of this level are easily seen to be equivalent after commensurate changes in the constant α). We initially set $l_{iU} = 0$, calculated the remaining parameters, and then re-parametrized by adjusting the largest parameter in each position to zero. In this way, the consensus residue in each position may be interpreted as a standard, with alternative residues assigned penalties $l_{ib} < 0$. Note that the apparently simpler model **3** based on the variables W_s and W_n actually involves a similar number of

parameters, because in this case the profile frequencies are also derived from the training data and thus should be counted among the parameters.

Splice site prediction and evaluation of models

Given the above interpretation of P, prediction of splice sites may be based on the following simple rule:

decision: true site if $P > c$
decision: non-site if $P \leq c$ **5**

where c is an appropriately chosen constant between 0 and 1. The problem of splice site prediction then consists of optimal choices for P as a function of a set of observed site characteristics and for c given P in order to minimize prediction errors. For any particular function P and threshold c, the described classification splits a sample of splice sites S into the two sub-samples $S_{>c}$ and $S_{\leq c}$ of sites with score $>c$ or $\leq c$, respectively. In general, both sub-samples contain true sites and non-sites. The prediction method is better the more the distributions of true sites and non-sites are biased towards true sites in $S_{>c}$ and non-sites in $S_{\leq c}$.

Our evaluation of the performance of the various models follows the treatment of Brunak *et al.* (7), using the notation of Snyder and Stormo (5). Thus, let the number of sites in $S_{>c}$ be predicted positives (PP), consisting of TP true positives (real sites) and FP false positives (non-sites). Let $S_{\leq c}$ consist of FN false negatives (real sites of low score) and TN true negatives (non-sites), adding up to a total of PN predicted negative sites. Let $AP = TP + FN$ be the number of actual positives (true sites), and let $AN = FP + TN$ be the number of actual negatives (non-sites). Then $Sn = TP/AP$ measures the sensitivity of the method: what fraction of the real sites are correctly predicted? $Sp = TP/PP$ measures the specificity of the method: what fraction of the predicted positives are real sites? An overall measure of the quality of the method is given by the Kendall tau rank correlation coefficient for dichotomous classifications (e.g., 27),

$$\tau = \frac{(TP)(TN) - (FP)(FN)}{\sqrt{(PP)(PN)(AP)(AN)}} \quad \textbf{6}$$

τ is 1 for a completely accurate prediction and –1 for a completely erroneous prediction.

High specificity is important for gene prediction programs based on signal identification. Each false positive splice site prediction may generate a large number of false exon–intron structures. However, high sensitivity may be even more important. Each true site in $S_{\leq c}$ that is discarded by the gene prediction algorithm causes it to miss the true parsing of the gene. We therefore focused in this study on site prediction methods that keep the rate of false negative predictions very low while minimizing the rate of false positive predictions. For each model we report sensitivity, specificity, and τ for three choices of c (relative to the training set): (i) the highest response value P for non-sites giving $Sn = 1$ (predicting all true sites correctly), (ii) the highest response value P for non-sites giving $Sn \geq 0.95$ (allowing 5% of real sites to be missed), and (iii) the smallest value of c that maximizes τ. Note that there are always different choices of c giving the same predictor performance for the training sample. In each case the response value close to an appropriately chosen true site value would work as well. Restriction to response values of non-sites as candidates for the threshold c gives more reliable results, because the sample of non-sites is much larger than the sample of true sites.

When comparing the performance of different models it is convenient to employ the monotonicity of $P(\vec{x}_i'\vec{\beta})$ and re-scale the P-values such that the decision threshold is at a fixed value. For example, changing the constant α to $\alpha - \ln[c/(1-c)]$ adjusts the decision threshold to 0.5.

RESULTS

Splice site profiles

All true splice sites were initially aligned with respect to the consensus GU or AG, respectively, and the nucleotide distribution was determined for each column of this alignment. Let f_{ib} denote the frequency of nucleotide b in position i, where i extends over all signal positions. The entropy values $I(i) = -\Sigma_b f_{ib} \ln f_{ib}$ provide a means of defining the extent of the signal sequence (28). $I(i)$ attains its minimal value 0 in the extreme case that one of the bases b in position i is completely conserved and its maximal value $-\ln 0.25 = 1.386$ in the uniform case $f_{ib} = 0.25$. Selecting the region around the splice sites with significantly lower than base level entropy, we defined the extent of the donor signal from –3 to +6 (negative integers extending into the exon, positive integers extending into the intron) and the acceptor signal from –15 to +2 (negative integers extending into the intron, positive integers extending into the exon). The donor site signal corresponds to the well known region of complementarity to U1-snRNA, and the acceptor site signal contains the conserved U-rich tract upstream of position –4 (for recent reviews see 9,10).

None of the 204 sequences in the maize acceptor site set has a G in position –3. If regarded as an absolute requirement, then the search algorithm for potential acceptor sites would have to discard any GAG/.. triplet site, no matter how well the other positions of the consensus are matched. A more sanguine approach would consider the complete lack of G as reflecting the relatively small sample size of the training set and allow for a small proportion of GAG/.. sites for consideration. We opted for the latter (particularly because in the *Arabidopsis* set of 578 sites there are three sequences with G in position –3) and replaced the observed frequencies in the profiles with probabilities estimated from pseudocounts derived as in (29) based on a Bayesian principle with uniform prior distribution.

Signal site extent and the derived profile frequencies do not differ significantly from previously published tabulations (9,28,30) and are therefore not shown here. The most frequent nucleotides in each position yield the familiar (A or C)AG/**GU**AAGU donor site and U_{11}GC**AG**/GU acceptor site consensuses. The control set profiles in each position closely reflect the average genomic base composition, as expected, because the non-sites are sampled equally from exons and introns.

Compositional contrast

The average mononucleotide composition in windows of length 50 nt flanking both sides of the donor GU and acceptor AG signals is displayed in Table 1. The intron internal flanks display a U percentage ~13.5 and 16.5 points higher than that in the exons around the donor and acceptor sites, respectively. The exon parts are 8–10.5 percentage points higher in G content, and to a lesser extent higher in C and, for acceptor sites, also in A. The Wilcoxon signed-rank test for comparison of matched pairs was used for

testing the statistical significance of the content differences (31). The contrast in C, G and U content is statistically significant for both donor and acceptor sites, whereas the A content differs significantly only for acceptor sites. Note that we cannot similarly specify a typical compositional contrast for the set of non-sites, because this set contains sites with flanking regions entirely within exons, or entirely within introns, or involving exons and introns in either order. In particular, the AG immediately upstream of the GU in the donor site consensus gives rise to a fair proportion of non-acceptor sites with donor-like contrast. Similarly, the GU immediately downstream of AG in the acceptor site consensus gives rise to non-donor sites with acceptor-like contrast.

Splice site prediction

We initially studied 12 different models for the prediction of splice sites. Results for the maize data are shown in Table 2; the *Arabidopsis* set gave similar results (not shown). For each model, the sensitivity, specificity, and τ value were derived for different levels of the threshold *c* in equation **5**. The models are represented by their defining sets of variables given in column one of Table 2. For instance, the set of variables W_s, W_n, X_U, X_{GC} denotes model equation **3**, and the set W_s, W_n, X_U denotes model equation **3** without considering the variable X_{GC}, i.e., setting the parameter $\mu = 0$.

Table 1. Compositional contrast around maize and *Arabidopsis* splice sites.

	Donor sites			Acceptor sites		
	exon	intron	Δ	intron	exon	Δ
Maize						
U	21.1	34.5	–13.4	37.6	21.5	16.1
A	24.0	23.7	0.3	21.5	25.0	–3.5
G	26.9	18.9	8.0	19.2	29.5	–10.3
C	28.0	22.9	5.1	21.7	24.0	–2.3
Arabidopsis						
U	27.3	41.0	–13.7	43.4	26.0	17.4
A	26.9	27.0	–0.1	24.5	29.0	–4.5
G	23.0	14.6	8.4	16.9	26.4	–9.5
C	22.8	17.4	5.4	15.2	18.6	–3.4

Displayed are the average percent base frequencies in 50 nucleotide flanks upstream and downstream of the conserved donor GU and acceptor AG, respectively. For convenience, the differences between upstream and downstream percentages are given in columns 4 and 7.

Table 2. Prediction of splice sites in maize pre-mRNAs

	FN = 0, Sn = 100%			FN = 10, Sn = 95%			tau maximal				
Variables	FP	Sp (%)	tau	FP	Sp (%)	tau	FN	Sn (%)	FP	Sp (%)	tau
Donor sites											
W_{sn}	**2945**	**6**	0.18	**629**	**23**	0.44	46	77	171	48	0.59
W_{sn}, X_U	2108	9	0.24	289	40	0.60	55	73	72	67	0.69
W_{sn}, X_{GC}	2479	8	0.21	384	33	0.54	56	72	60	71	0.71
W_{sn}, X_U, X_{GC}	**2342**	**8**	0.22	**298**	**39**	0.59	52	74	67	69	0.71
W_s, W_n	2926	6	0.19	627	23	0.44	47	77	165	48	0.59
W_s, W_n, X_U	2094	9	0.24	292	40	0.60	61	70	62	69	0.69
W_s, W_n, X_{GC}	2518	7	0.21	383	33	0.54	57	72	58	71	0.71
W_s, W_n, X_U, X_{GC}	2375	8	0.22	293	39	0.60	53	74	65	69	0.71
L	2983	6	0.18	590	24	0.46	44	78	164	49	0.60
L, X_U	1382	13	0.31	306	38	0.59	52	74	59	72	0.72
L, X_{GC}	1900	10	0.26	346	36	0.56	39	81	80	67	0.73
L, X_U, X_{GC}	**1497**	**12**	0.30	**284**	**40**	0.60	38	81	70	70	0.74
Acceptor sites											
W_{sn}	**3197**	**6**	0.17	**1373**	**12**	0.30	100	51	127	45	0.46
W_{sn}, X_U	2597	7	0.21	923	17	0.37	75	63	75	63	0.62
W_{sn}, X_{GC}	2471	8	0.22	731	21	0.42	82	60	83	60	0.58
W_{sn}, X_U, X_{GC}	**2716**	**7**	0.20	**689**	**22**	0.43	70	66	68	66	0.65
W_s, W_n	3195	6	0.17	1340	13	0.30	94	54	134	45	0.48
W_s, W_n, X_U	2699	7	0.20	610	24	0.45	58	72	109	57	0.63
W_s, W_n, X_{GC}	2536	7	0.21	653	23	0.44	66	68	102	57	0.61
W_s, W_n, X_U, X_{GC}	2836	7	0.19	493	28	0.49	63	69	74	66	0.66
L	2533	7	0.21	1225	14	0.32	88	57	134	46	0.50
L, X_U	1506	12	0.30	529	27	0.48	55	73	86	63	0.67
L, X_{GC}	1526	12	0.30	502	28	0.49	64	69	94	60	0.63
L, X_U, X_{GC}	**1516**	**12**	0.30	**411**	**32**	0.53	33	84	133	56	0.67

Predictions are based on 201 true and 6305 false donor sites and on 204 true and 6290 false acceptor sites. FN, number of false negatives; FP, number of false positives; Sn, sensitivity; Sp, specificity; tau, correlation coefficient, equation **6**. The 12 different models are based on equations **3** and **4** of the text with only the indicated variables included. For example, W_s, W_n, X_{GC} refers to model equation **3** with $\delta = 0$. Numbers that highlight performance differences between selected models appear in bold face.

As a standard model for comparisons we chose prediction based on the usual profile scores, W_{sn}. Requiring 100% sensitivity, such that no true splice sites are overlooked, forces choice of the predictor threshold c of equation **5** below the level of the weakest true site in the training set. This strict requirement results in a very large number of false splice site predictions. In the case of the standard model, falsely predicted splice sites outnumber the true ones in a ratio of about fifteen to one. Allowing 5% of true sites to be missed greatly reduces the number of false positive predictions. Specificity improves about 4-fold to 23% for donor sites and ~2-fold to 12% for acceptor sites in the standard model (Table 2).

Inclusion of the contrast variables X_U and X_{GC} significantly improves prediction quality. For example, for the standard model at 95% sensitivity the number of false positive donor sites decreases from >600 to <300 when either X_U alone or both X_U and X_{GC} are considered in the model. Similarly, the number of false positive acceptor sites drops from nearly 1400 to <700 upon inclusion of both contrast variables.

Replacing the W_{sn} variable by the separate terms W_s and W_n for the most part does not change the predictor performance very much. An exception are the acceptor site models involving X_U, for which the W_s, W_n models are clearly superior to the models at the 95% sensitivity level (but not in terms of comparing the τ values). A much stronger and consistent improvement is obtained with the L models involving the contrast variables. Specificity at the 95% sensitivity level is up to 40% for donor sites and 32% for acceptor sites for the L, X_U, X_{GC} model.

It is noteworthy that, at the 100% sensitivity level, the gains obtained with the L models compared with the standard profile models are at best slight when the contrast variables are not considered. However, the improvements are substantial for the full models including the contrast variables: donor site specificity increases from 6% to 12% for the L model as a result of adding the variables X_U and X_{GC} compared with an increase from 6% to only 8% for the profile models, and acceptor site specificity increases from 7% to 12% compared with an increase for the profile models from 6% to 7%. A distinction between the L variable and the profile variables W_{sn} or W_s and W_n is that, for the former, weights in each signal position are derived *ab initio* during the training, whereas for the latter the positional weights are fixed as log-frequencies and only the relative contribution of their overall sum per site is estimated in the regression model. The greater flexibility of the L variable appears advantageous. Moreover, the large performance differences between the models upon inclusion of the contrast variables suggests that signal strength and compositional contrast do not contribute independently to splice site recognition. This apparent correlation is obvious for acceptor sites, because the U-rich section of the acceptor signal sequence clearly contributes to the U-content of the upstream flank, but it is less obvious for donor site prediction.

To further investigate these issues, we categorized the splice sites according to presence or absence of the most conserved signal residue apart from the donor GU and acceptor AG. For donor sites this residue is G in position –1 immediately upstream of the GU. Eighty-six percent of the donor sites in the maize set and 79% in the *Arabidopsis* set conserve this G. For acceptor sites, signal position –3 immediately upstream of the AG is C in 77% of the maize introns and in 68% of the *Arabidopsis* introns. Table 3 and Table 4 display the predictor performance of the L, X_U, X_{GC} model trained separately on the respective subsets of splice sites. For maize donor sites, great improvement is obtained at the 100% sensitivity level: compared with the non-categorized model, the number of false positive predictions is reduced nearly 3-fold. Improvement is also substantial for *Arabidopsis* donor sites at 100% sensitivity. A small but consistent improvement is evident for acceptor sites in both species, and for all comparisons with the 95% sensitivity and maximal τ criteria.

While 100% sensitivity is a desired goal for splice site prediction in gene finding algorithms, this is met with considerable difficulty for methods based entirely on local sequence characteristics. This is particularly evident for *Arabidopsis* acceptor site prediction. There are three sites in our training set which score so low in all models that their inclusion forces acceptance of an exceedingly large number of false sites (data not shown). One of these sites precedes a nine base exon (exon 2 of *Atbfruct1*, GenBank accession number X74515) and accordingly displays atypical compositional contrast as the downstream 50 base flank is essentially all intron. The second site occurs in the second intron of the cytochrome c gene (GenBank accession number M85253). This intron is very short (59 bases) and only 13.6% U, resulting in poor profile and contrast scores. The third site (intron 10 of GenBank accession number U05599) is one of only three *Arabidopsis* acceptor sites featuring G in position –3 and there are only two U nucleotides in the –15 to –5 region. Exclusion of these troublesome sites from the training set restores normal levels of predictor performance (Table 4), thus clearly identifying these sites as outliers.

Table 3. Prediction of splice sites in maize pre-mRNAs upon subclassification of splice sites according to extended consensus

Set	FP	Sp (%)	tau	FN	Sn (%)	FP	Sp (%)	tau	FN	Sn (%)	FP	Sp (%)	tau
5′ GGU	361	32	0.49	8	95	166	50	0.64	53	69	24	83	0.73
HGU	148	16	0.40	1	97	119	19	0.42	5	83	11	69	0.75
combined	**509**	**28**	0.51	9	96	**285**	**40**	0.60	58	71	35	80	0.75
CAG	857	15	0.27	7	96	232	39	0.56	32	80	61	67	0.70
DAG	453	9	0.29	2	96	69	39	0.61	12	74	5	88	0.81
combined	**1310**	**13**	0.33	9	96	**301**	**39**	0.60	44	78	66	71	0.74

Results are shown for the L, X_U, X_{GC} model, trained separately on the indicated subsets of donor and acceptor site samples. Notation is as in Table 2. Prediction criteria were set to 100% sensitivity (columns 2–4), ≥95% sensitivity (columns 5–9), and maximal tau (columns 10–14). The set 5′ GGU consists of 172 true and 1466 false donor sites, all with G preceding the GU donor consensus. H denotes non-G. 5′ HGU consists of 29 true and 4839 false donor sites. D denotes non-C. The set 3′ CAG consists of 157 true and 1645 false acceptor sites, and 3′ DAG consists of 47 true and 4645 false acceptor sites. The combined sets show improved prediction compared with the non-categorized models (cf. corresponding bold face entries in Table 2).

Table 4. Prediction of splice sites in *Arabidopsis* pre-mRNAs

Set	FP	Sp (%)	tau	FN	Sn (%)	FP	Sp (%)	tau	FN	Sn (%)	FP	Sp (%)	tau
5′ all	3594	14	0.32	28	95	819	40	0.60	101	82	205	70	0.75
5′ GGU	1509	23	0.37	22	95	395	52	0.66	52	89	158	72	0.77
5′ HGU	1011	11	0.31	5	96	362	24	0.47	31	74	38	70	0.72
combined	2520	19	0.39	27	95	757	42	0.61	83	86	196	72	0.77
3′ all	4417	12	0.29	28	95	754	42	0.61	122	79	216	68	0.72
3′ CAG	1855	17	0.27	19	95	323	53	0.67	79	80	102	75	0.75
3′ DAG	2289	8	0.25	9	95	305	37	0.58	56	70	72	65	0.67
combined	4144	12	0.30	28	95	628	47	0.65	135	76	174	72	0.73

Results are shown for the L, X_U, X_{GC} model, trained separately on the indicated subsets of donor and acceptor site samples. Notation is as in Table 2. Prediction criteria were set to 100% sensitivity (columns 2–4), ≥95% sensitivity (columns 5–9), and maximal tau (columns 10–14). The 5′ all set consists of 577 true and 14 964 false donor sites. 5′ GGU consists of 458 true and 3729 false donor sites, all with G preceding the GU donor consensus. H denotes non-G. 5′ HGU consists of 119 true and 11 235 false donor sites. The 3′ all set consists of 574 true and 15 712 false acceptor sites. Three poorly scoring true acceptors were excluded as discussed in the text. D denotes non-C. The set 3′ CAG consists of 387 true and 3240 false acceptor sites, and 3′ DAG consists of 187 true and 12 472 false acceptor sites. The rows labeled 'combined' give the overall values for prediction based on the subclassifications.

Interpretation of model parameters

The parameter values estimated in the wake of fitting the models to the training sets should reflect the relative importance of the different variables and thus may guide future modeling as well as interpretations in terms of the underlying splicing process. We discuss these possibilities for the *Arabidopsis* donor site model with subclassification. The estimated model parameters are given in Table 5. Several observations stand out: (i) The constant term α is ~9-fold higher for the GGU sites than for the HGU sites. Inserting into equation **4** we calculate a *P*-value of 0.97 for a perfectly matching GGU donor site $L = 0$ assuming no effect of compositional contrast $X_U = X_{GC} = 0$, compared with a *P*-value of only 0.59 for a perfectly matching HGU donor site. Thus, our model predicts that splice site quality is considerably reduced if the consensus G in position –1 is replaced. (ii) From Table 1, the expected contrast values are $-X_U \approx X_{GC} \approx 0.14$. Inserting these values (with L remaining zero) results in *P*-values of 1.00 and 0.97 for GGU and HGU sites, respectively. Thus, the model predicts that average or better compositional contrast can fully restore splice site quality in the HGU class of sites. (iii) The most negative weights occur in positions –2, 1 and 3 for GGU sites, and in positions 1–3 for HGU sites (positions counted relative to GU, cf. Fig. 1). Note that the penalties in positions 1–3 are more severe for the HGU set compared with the GGU set, indicating that further mismatches to the consensus are probably ill tolerated in the former set. The importance of G at position 3 was also confirmed by Hebsgaard *et al.* (14) and likely reflects a requirement of hybridization by U1-RNA (9,10).

Validation tests for the splice site predictor

We tested in several ways how well the prediction rule **5** works to identify true splice sites in genomic DNA. Standard cross-validation techniques proved all models robust so that over-fitting during training could be ruled out (data not shown). For an additional test, we compiled independent test sets consisting of five maize genes (originally excluded from our training set as a result of missing sequence information for one intron in each case) and for *Arabidopsis* comprising 65 sequences contained in the Korning *et al.* (16) set, but not in our training set. Splice sites in these sets were predicted with the predictor threshold c set to the values derived in the training. As shown in Table 6, the prediction quality on the test sets is entirely comparable with the results obtained for the training sets (Table 3 and Table 4). Prediction with the thresholds that maximize the τ correlation measure on the training sets is the least stable. However, this is of little concern because in typical applications the threshold will be set to the more stable extremes that give either high sensitivity or high specificity. It is of particular significance to note that the L, X_U, X_{GC} model with subclassification outperforms the other models on the test sets as it did on the training sets. Thus, the displayed performance values seem to reflect the best possible accuracy for splice site prediction based on the scope of models we investigated.

Table 5. *Arabidopsis* donor site model parameters

Parameter	GU set (458 sites)				HGU set (119 sites)			
α	3.46				0.37			
δ	–9.71				–8.87			
μ	13.68				13.09			
	A	C	G	U	A	C	G	U
l_{-3}	0	0.32	–1.20	–0.80	0	0.20	–0.79	–0.91
l_{-2}	0	–3.22	–3.05	–2.74	0	–0.71	–1.29	–1.26
l_{-1}	0	0	0	0	–0.17	–0.11	0	0
l_1	0	–3.01	–2.99	–2.87	0	–3.18	–2.67	–5.21
l_2	0	–1.08	–2.11	–1.78	0	–2.13	–4.52	–2.71
l_3	–2.53	–2.82	0	–2.72	–3.98	–5.56	0	–3.73
l_4	–1.35	–0.82	–1.80	0	–1.17	–0.90	–1.54	0

Parameters are given for the L, X_U, X_{GC} model (equation **4**) with subclassification. Parameters set to 0 correspond to the consensus donor site residues.

Table 6. Validation test of the splice site predictor

	c (Sn = 100%)					*c* (Sn = 95%)					*c* (tau maximal)				
Set	FN	Sn (%)	FP	Sp (%)	tau	FN	Sn (%)	FP	Sp (%)	tau	FN	Sn (%)	FP	Sp (%)	tau
Maize															
Donor sites															
W_{sn}	0	100	444	7	0.2	3	91	90	26	0.46	13	63	25	47	0.52
W_{sn}, X_U, X_{GC}	0	100	349	9	0.24	5	86	41	42	0.58	13	63	11	67	0.63
L, X_U, X_{GC}	0	100	216	14	0.33	3	91	45	42	0.60	10	71	9	74	0.71
GGU/HGU	1	97	70	33	0.54	3	91	39	45	0.63	17	51	4	82	0.64
Acceptor sites															
W_{sn}	0	100	503	7	0.18	3	92	216	14	0.30	20	46	23	43	0.42
W_{sn}, X_U, X_{GC}	0	100	436	8	0.20	2	95	108	24	0.45	15	59	13	63	0.60
L, X_U, X_{GC}	0	100	216	15	0.34	4	89	68	33	0.51	11	70	27	49	0.57
CAG/DAG	0	100	185	17	0.37	6	84	51	38	0.54	12	68	18	58	0.61
Arabidopsis															
Donor sites															
L, X_U, X_{GC}	0	100	1389	17	0.36	14	95	315	46	0.64	57	79	74	75	0.76
GGU/HGU	0	100	969	22	0.43	14	95	289	48	0.65	46	83	78	75	0.78
Acceptor sites															
L, X_U, X_{GC}	3	99	1710	14	0.31	19	93	301	46	0.64	63	77	84	72	0.74
CAG/DAG	3	99	1636	14	0.32	27	90	247	51	0.66	73	74	70	75	0.73

GGU/HGU and CAG/DAG denote the L, X_U, X_{GC} models with subclassification. The three levels of the predictor threshold were set in accord with the respective training data (Tables 3 and 4). Notation is as in Table 2. The maize test set consists of 35 true and 951 false donor sites and 37 true and 933 false acceptor sites. The *Arabidopsis* test set consists of 277 true and 6101 false donor sites and 280 true and 6022 false acceptor sites.

DISCUSSION

Elucidation of the *trans*-acting factors and *cis*-acting elements involved in splice site selection and determination of splicing efficiency in plant pre-mRNAs has been hampered by the absence of a plant *in vitro* splicing system. Similarities to yeast and vertebrates in terms of splice site consensuses and the sequences of the small nuclear ribonucleic acids are juxtaposed by plant specific features. Thus, plant introns generally lack a polypyrimidine run upstream of the 3′ splice site, a feature that is conserved in most yeast and vertebrate introns, nor do they contain a characteristic branchpoint sequence. On the other hand, plant introns are typically distinguished from the flanking exons by a strong bias towards U bases (Table 1), and this bias plays a role in accurate pre-mRNA processing (for reviews see 9,10).

Here we pursued the challenge of predicting splice sites in plant pre-mRNA from local sequence inspection. A successful solution to this problem would both suggest that our understanding of splice site recognition variables is sound and also be of considerable practical importance in the context of gene identification algorithms. Specifically, we attempted to derive rules that would distinguish true donor and acceptor splice sites from the large excess of alternative sites that minimally conserve only the characteristic GU and AG consensuses, respectively. For each site we evaluated (i) the particular sequence in the signal positions that are typically conserved in true splice sites, and (ii) the compositional contrast between the flanking 50 bases upstream and downstream of the sites (Fig. 1). We explored logitlinear models to map linear combinations of the values of these variables onto the [0,1] interval, where 1 indicates strong prediction of a splice site and 0 indicates strong rejection of that hypothesis.

Profile methods

The strongest evidence that the considered variables truly determine splice site selection would derive from a clear separation of the scores for true splice sites from those of control non-sites in both training and test sets. The degree of separation is most stringently assessed by how many false positive predictions are made when the predictor is trained to 100% sensitivity, i.e., not to reject any true site. For example, just defining the splice sites by the consensus GU and AG dinucleotides leads to ~30 false positive predictions for every truly identified site. Our results (Table 2) show that discrimination based on the usual profile scores W_{sn} decrease the number of false positive predictions only ~2-fold. For every true donor site in the maize training set the model would also accept an average of 15 false sites, and for every true acceptor site it would include ~16 non-acceptor sites.

One hundred percent sensitivity is, of course, a very strict requirement, and one that is easily foiled by a few exceptional sites in the training set, including possible cases of erroneous splice site determinations or annotations. Some such problems occurred with our original *Arabidopsis* acceptor site training set (Table 4). But even allowing 5% of true sites to be overlooked by the predictor (i.e., training the predictor to 95% sensitivity), the profile method would still accept an average of three or seven

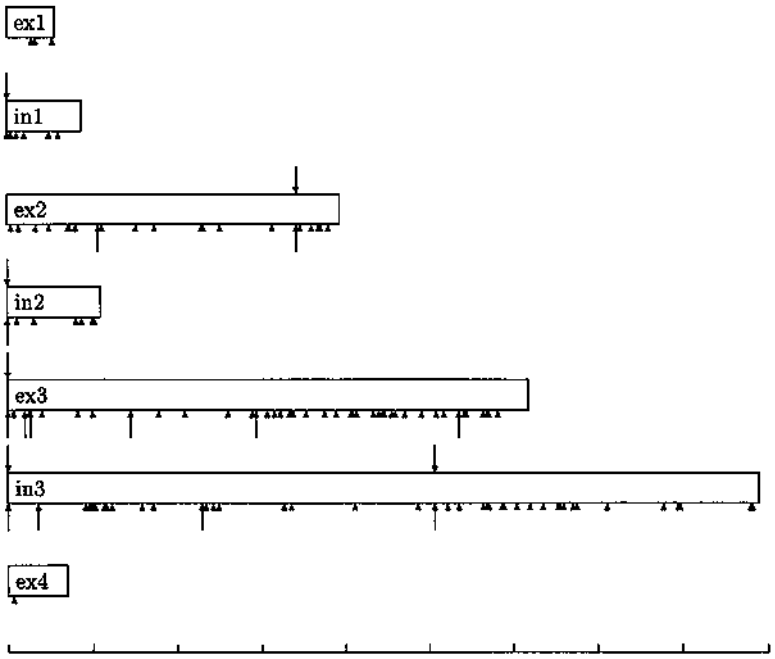


Figure 2. Donor site prediction for the maize actin gene sequence (GenBank accession no. J01238). The four exons and three introns are drawn to scale. Each unit on the scale on the bottom represents 100 nt. Small triangles from below indicate positions at which GU is found in the sequence. Arrows from below mark GU locations which are accepted as donor sites by the standard donor site profile model W_{sn}. Arrows from above indicate GU positions accepted by the logitlinear model L, X_U, X_{GC} with sub-classification. For each model, the prediction threshold was set at the 95% sensitivity level as derived from the training set. All true donor sites are strongly predicted. Note the false positive donor site prediction caused by the AG/GU intron 2/exon 3 boundary which in this sequence is not distinguished by compositional contrast.

false sites for every correctly predicted maize donor or acceptor site, respectively (Table 2).

Improved methods

The above numbers demonstrate the need for improved splice site prediction methods. We showed that including compositional contrast variables and modeling the signal sites by the more general weights L (Fig. 1) greatly enhanced the performance of the predictor: at 100% sensitivity, the ratio of false positive sites to correctly identified sites decreased to ~7, and at 95% sensitivity, this ratio decreased to 2 or less (Table 2).

The result that the logitlinear models involving L instead of the usual profile scores performed so much better was initially surprising. Our interpretation is that the equal weighting of the different signal positions in a common profile application is inadequate. For example, in the maize donor sites the position immediately upstream of GU is 86% G. However, in the non-sites exceeding the minimal profile score of the true sites this position is only 33% G. Thus, the majority of these non-sites compensate for the lack of this G with a more favorable match to the consensus in other positions. But the compensation in terms of scoring apparently does not reflect splice site functionality.

As a refinement of the model, we therefore subclassified true splice sites and control non-sites according to the presence or absence of the most strongly conserved splice site bases apart from the characteristic GU and AG. This approach succeeded in further lowering the numbers of false positive predictions, both in the training sets (Table 3 and Table 4) and in independent test sets (Table 6). The improvements in splice site prediction are illustrated for a typical gene in Figure 2. The best model (L, X_U, X_{GC} with subclassification) is built into our **SplicePredictor** program.

We would note that our performance statistics are conservatively estimated by disregarding the relative location of the false positive predictions with respect to the known or potential splice products. For example, the intron 2/exon 3 boundary in the maize actin gene is AG/GU and results in both a correct acceptor site prediction and a false positive donor site prediction (cf. Fig. 2). However, as long as the splicing machinery efficiently recognizes the upstream true donor site, this false positive site is irrelevant. Hebsgaard *et al.* (14) recently demonstrated that incorporation of this type of global coding potential assessment drastically improves exon/intron prediction.

Perspective

Despite the substantial improvements in splice site prediction with the refined models, the success rate of the predictor is not entirely satisfactory. This may reflect inherent limitations within the current framework. First, there are other factors influencing splice site selection that have not been considered. For example, recent studies have identified branchpoints in some plant pre-mRNAs (32,33). The exact sequence requirements for plant branchpoints are unclear, but, when present, the branchpoint likely has a role in 3′ splice site selection. There are also suggestions of specific intron and exon recognition factors (9,10), which may actively contribute to true splice site selection or otherwise mask high scoring non-sites by binding to pre-mRNA.

Also not considered in our current model are changes in the local site attributes during the sequential process of splicing. For example, the splice sites delineating very short exons would typically display poor compositional contrast. Once one of the introns is removed, the remaining intron will be flanked by the fusion of the short exon with its adjacent exon, which should restore the typical compositional contrast. It is also possible that relatively high scoring non-sites are locally suboptimal compared with nearby true sites and are irrelevant because splice site selection is accomplished by scanning for the locally best matching site. Moreover, interaction of adjacent splice sites in either 5′ to 3′ ('intron definition') or 3′ to 5′ polarity ('exon definition'; 34) may indicate selection of splice sites in pairs rather than individually.

The above issues will have to be addressed by further modeling studies. The logitlinear modeling approach provides a very flexible tool in this context which is statistically very well understood and, as discussed, thoroughly motivated for sequence signal recognition problems.

PROGRAM AVAILABILITY

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

ACKNOWLEDGEMENTS

We thank C. Burge, J. C. Carle-Urioste, R. Pincus, and V. Walbot for helpful discussions. V.B. was supported in part by NIH grants 2R01HG00335-09 and 5R01GM10452-32. K.H was supported by Deutsche Forschungsgemeinschaft project KL 760/1-3 awarded to J.K.

REFERENCES

1 Gelfand, M.S. (1995) *J. Comp. Biol.* **2**, 87–115.
2 Burset, M. and Guigó, R. (1996) *Genomics* **34**, 353–367.
3 Fields, C.A. and Soderlund, C.A. (1990) *Comp. Appl. Biol. Sci.* **6**, 263–270.
4 Guigó, R., Knudsen, S., Drake, N. and Smith, T. (1992) *J. Mol. Biol.* **226**, 141–157.
5 Snyder, E.E. and Stormo, G.D. (1993) *Nucleic Acids Res.* **21**, 607–613.
6 Dong, S. and Searls, D.B. (1994) *Genomics* **23**, 540–551.
7 Brunak, S., Engelbrecht, J. and Knudsen, S. (1991) *J. Mol. Biol.* **220**, 49–65.
8 Sirajuddin, K. Nagashima, T. and Ono, K. (1995) *Comp. Appl. Biol. Sci.* **11**, 349–359.
9 Filipowicz, W., Gniadkowski, M., Klahre, U. and Liu, H.-X. (1995) Pre-mRNA splicing in plants. In Lamond, A.I. (ed.) *Pre-mRNA Processing*. R.G. Landes Publishers, Georgetown, TX, pp. 65–77.
10 Luehrsen, K.R., Taha, S. and Walbot, V. (1994) *Prog. Nucleic Acids Res. Mol. Biol.* **47**, 149–193.
11 Goodall, G.J. and Filipowicz, W. (1989) *Cell* **58**, 473–483.
12 Lou, H., McCullough, A.J. and Schuler, M.A. (1993) *Mol. Cell. Biol.* **13**, 4485–4493.
13 Luehrsen, K.R. and Walbot, V. (1994) *Genes Dev.* **8**, 1117–1130.
14 Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouzé, P. and Brunak, S. (1996) *Nucleic Acids Res.* **24**, 3439–3452.
15 Brendel, V. (1992) *Mathl. Comput. Modelling* **16** (6/7), 37–43.
16 Korning, P.G., Hebsgaard, S.M., Rouzé, P. and Brunak, S. (1996) *Nucleic Acids Res.* **24**, 316–320.
17 Kondrakhin, Y.V., Kel, A.E., Kolchanov, N.A., Romashchenko, A.G. and Milanesi, L. (1995) *Comp. Appl. Biol. Sci.* **11**, 477–488.
18 Chen, Q.K., Hertz, G.Z. and Stormo, G.D. (1995) *Comp. Appl. Biol. Sci.* **11**, 563–566.
19 Prestridge, D.S. (1995) *J. Mol. Biol.* **249**, 923–932.
20 Santner, T.J. and Duffy, D.E. (1989) *The Statistical Analysis of Discrete Data*. Springer-Verlag, New York.
21 Christensen, R. (1990) *Log-linear Models*. Springer-Verlag, New York.
22 Tue Tjur (1992) StatUnit: Turbo Pascal unit for statistical analysis. Institute of Mathematical Statistics, University of Copenhagen. http://www.math.ku.dk/~tuetjur.
23 Kleffe, J., Hermann, K., Gunia, W., Vahrson, W. and Wittig, B. (1995) *Comp. Appl. Biol. Sci.* **11**, 449–455.
24 McCullough, A.J., Lou, H. and Schuler, M.A. (1993) *Mol. Cell. Biol.* **13**, 1323–1331.
25 Lou, H., McCullough, A.J. and Schuler, M.A. (1993) *Plant J.* **3**, 393–403.
26 Luehrsen, K.R. and Walbot, V. (1994) *Plant Mol. Biol.* **24**, 449–463.
27 Kendall, M. and Gibbons, J.D. (1990) *Rank Correlation Methods*. 5th edition, Edward Arnold, London.
28 White, O., Soderlund, C., Shanmugan, P. and Fields, C. (1992) *Plant Mol. Biol.* **19**, 1057–1064.
29 Tanner, M. and Wong, W.H. (1987) *J. Am. Stat. Assoc.* **82**, 528–540.
30 Sinibaldi, R.M. and Mettler, I.J. (1992) *Prog. Nucleic Acids Res. Mol. Biol.* **42**, 229–257.
31 Woolson, R.F. (1987) *Statistical Methods for the Analysis of Biomedical Data*. John Wiley & Sons, New York.
32 Simpson, C.G., Clark, G., Davidson, D., Smith, P. and Brown, J.W.S. (1996) *Plant J.* **9**, 369–380.
33 Liu, H.-X. and Filipowicz, W. (1996) *Plant J.* **9**, 381–389.
34 Berget, S.M. (1995) *J. Biol. Chem.* **270**, 2411–2414.