BIOINFORMATICS

Vol. 20 no. 7 2004, pages 1157–1169
DOI: 10.1093/bioinformatics/bth058

Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus

Volker Brendel[1,2,*], Liqun Xing[1,†] and Wei Zhu[1,‡]

[1] Department of Genetics, Development and Cell Biology and [2] Department of Statistics, Iowa State University, 2112 Molecular Biology Building, Ames, Iowa 50011–3260, USA

Received on July 14, 2003; revised on December 12, 2003; accepted on December 13, 2003
Advance Access publication February 5, 2004

ABSTRACT

Motivation: Accurate gene structure annotation is a challenging computational problem in genomics. The best results are achieved with spliced alignment of full-length cDNAs or multiple expressed sequence tags (ESTs) with sufficient overlap to cover the entire gene. For most species, cDNA and EST collections are far from comprehensive. We sought to overcome this bottleneck by exploring the possibility of using combined EST resources from fairly diverged species that still share a common gene space. Previous spliced alignment tools were found inadequate for this task because they rely on very high sequence similarity between the ESTs and the genomic DNA.

Results: We have developed a computer program, GeneSeqer, which is capable of aligning thousands of ESTs with a long genomic sequence in a reasonable amount of time. The algorithm is uniquely designed to tolerate a high percentage of mismatches and insertions or deletions in the EST relative to the genomic template. This feature allows use of non-cognate ESTs for gene structure prediction, including ESTs derived from duplicated genes and homologous genes from related species. The increased gene prediction sensitivity results in part from novel splice site prediction models that are also available as a stand-alone splice site prediction tool. We assessed GeneSeqer performance relative to a standard *Arabidopsis thaliana* gene set and demonstrate its utility for plant genome annotation. In particular, we propose that this method provides a timely tool for the annotation of the rice genome, using abundant ESTs from other cereals and plants.

Availability: The source code is available for download at http://bioinformatics.iastate.edu/bioinformatics2go/gs/download.html. Web servers for *Arabidopsis* and other plant species are accessible at http://www.plantgdb.org/cgi-bin/AtGeneSeqer.cgi and http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi, respectively. For non-plant species, use http://bioinformatics.iastate.edu/cgi-bin/gs.cgi. The splice site prediction tool (SplicePredictor) is distributed with the GeneSeqer code. A SplicePredictor web server is available at http://bioinformatics.iastate.edu/cgi-bin/sp.cgi

Contact: vbrendel@iastate.edu

Supplementary information: http://www.plantgdb.org/AtGDB/prj/BXZ03B

INTRODUCTION

Annotation of gene structure in eukaryotic genomes currently involves both computational and experimental approaches. Because of time and expense constraints, initial annotation mostly relies on *ab initio* gene prediction based on statistical modeling of exon and intron features. The best of these methods have been estimated to achieve about 80% sensitivity and specificity at the exon level, but the success rate is much lower at the level of entire gene structure, with typically less than half the predictions entirely accurate (Pavy *et al.*, 1999; Rogic *et al.*, 2001). In practice, a combination of different programs appears to be more successful than reliance on a single program (Pavy *et al.*, 1999; Murakami and Takagi, 1998). Spliced alignment of potential homologous protein sequences to genomic DNA is a complementary approach to *ab initio* gene prediction that gives better accuracy, provided a close enough homolog of the potential gene product is available (Gelfand *et al.*, 1996; Usuka *et al.*, 2000; Mathé *et al.*, 2002).

The most direct experimental evidence for gene structure comes from sequencing full-length cDNAs with subsequent spliced alignment of the cDNA sequences to the genomic DNA. An added advantage of this approach is that sufficient cDNA sampling under different conditions will reveal transcript isoforms arising from alternative splicing or alternative transcription start or termination points. An intermediate step in gene discovery is sequencing of expressed sequence tags (ESTs), which typically correspond to partial rather than full-length cDNAs. Clustering and assembly of ESTs to potential full-length transcripts is commonly pursued to

[*] To whom correspondence should be addressed.

[†] Current address: BASF Plant Science NC, 26 Davis Drive, Research Triangle Park, NC 27709-3528, USA.

[‡] Current address: NewLink Genetics, 2901 S. Loop Dr, Ames, IA 50010, USA.

estimate the gene space of a species, using methods that rely on pair-wise sequence similarities (Bouck *et al.*, 1999; Liang *et al.*, 2000; Pertea *et al.*, 2003; Kalyanaraman *et al.*, 2003). However, direct alignment to genomic DNA, when possible, is more accurate and informative (Zhu *et al.*, 2003).

The alignment of ESTs to genomic DNA is non-trivial for a number of reasons. ESTs are usually deposited as single-pass sequencing products, increasing the conventionally accepted rate of sequencing errors and ambiguous base determinations. ESTs are typically sampled from a large variety of origins that represent a range of subspecies, tissue types and conditions, thus leading to a heterogeneous sequence view confounded by polymorphisms and paralogous genes. In addition, sequencing artifacts (e.g. chimeras), sample contaminations and complex patterns of alternative splicing further complicate the alignment task.

A number of tools that address this alignment problem are now available and provide adequate solutions for some of these needs in more narrowly defined context. The underlying algorithms can be categorized into two groups with respect to the way they generate spliced alignments. One category involves heuristic, BLAST-like methods for the initial alignment and includes the tools sim4 (Florea *et al.*, 1998), Spidey (Wheelan *et al.*, 2001), BLAT (Kent, 2002) and Squall (Ogasawara and Morishita, 2002). Typically, these programs find matching segments at high stringency using BLAST (Altschul *et al.*, 1997) or a variant, with subsequent output parsing to favor canonical splice sites. EST_GENOME (Mott, 1997), dds/gap2 (Huang *et al.*, 1997) and GeneSeqer (Usuka *et al.*, 2000) belong to another category of programs that implement a full dynamic programming approach to derive the optimal score and spliced alignment, allowing for within-exon insertions and deletions. In GeneSeqer, potential splice sites are differentially scored according to independent splice site prediction methods. Consideration of predicted splice site strength was shown to improve the performance of the algorithm in the case of imperfect sequence matching as a result of sequencing errors or sequence polymorphisms (Usuka *et al.*, 2000).

There are several limitations in the BLAST-like spliced alignment methods. First, short exons (about 20 or fewer bases) are generally missed because they do not qualify as high-scoring segment pairs. Second, reliable alignments are limited to cognate ESTs with low sequencing error rates. For example, sim4 reports only the highest scoring match for each EST query, and TAP, a useful transcript assembly tool based on sim4 (Kan *et al.*, 2001), recommends a threshold of 92% overall identity for any such alignment to be included into the transcript assembly. In addition, the simple adjustment for exon–intron boundaries to conform to canonical splice sites whenever possible, as used in most of spliced alignment programs, further restricts application to unequivocal alignments and can lead to inconsistencies (e.g. sim4/TAP allow the standard GT–AG introns in conjunction with a complementary CT–AC intron in the same alignment, confounding assignment of the true transcript orientation). These limitations may be inconsequential when the need is for fast, reliable alignment of ESTs or cDNAs that, based on high sequence similarity, can be unambiguously assigned to a unique chromosomal locus; however they render these algorithms helpless in the situations discussed here.

EST sampling is sparse for most species when compared with the large human and mouse EST collections. However, if ESTs from related taxonomic groups could be successfully employed for gene identification, the EST resources would appear much more impressive. To date, there are well over two million ESTs from all plant species combined. Because of the inclusion of sophisticated splice site models and exhaustive alignment with a dynamic programming approach, the GeneSeqer algorithm affords a promising approach in attempts to make use of this resource. For example, GeneSeqer was recently shown to be very successful in identifying very short exons in *Arabidopsis thaliana* (Haas *et al.*, 2002) and improving *Arabidopsis* genome annotation (Zhu *et al.*, 2003). Here we report generalization of GeneSeqer to exploit heterogeneous EST sources for plant genome annotation by deriving a consensus gene structure prediction from multiple, possibly imperfect sequence alignments.

The greater accuracy afforded by the dynamic programming approach adopted in GeneSeqer is obtained at the expense of greater computational efforts. Practical implementation of the algorithm requires efficient selection of restricted genomic DNA regions and matching ESTs from a typically large EST collection in order to minimize or eliminate the computer time spent on deriving locally optimal but insignificant alignments. In this study, we present a string matching scheme based on pre-processing of the input EST dataset that allows fast target selection for detailed analysis by the dynamic programming algorithm. The previous implementation of the GeneSeqer algorithm was also modified to incorporate Bayesian statistical models for splice site prediction similar to models introduced by Salzberg (1997). We discuss applications to *A.thaliana* and rice genome annotation, which suggest that the novel algorithm approach provides a practical and powerful tool for accurate gene structure identification.

SYSTEMS AND METHODS

Programs used

The dynamic programming subroutines of GeneSeqer were described previously (Usuka *et al.*, 2000; Usuka and Brendel, 2000). The source code of the program is available at http://bioinformatics.iastate.edu/bioinformatics2go/gs/download.html. The data and some of the figures in this article were produced with the specialized GeneSeqer web servers at http://www.plantgdb.org/cgi-bin/AtGeneSeqer.cgi (for *Arabidopsis*) and http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi (all plant species; Schlueter *et al.*, 2003).

Sim4 (Florea *et al.*, 1998) was downloaded from http://globin.cse.psu.edu/. TAP (Kan *et al.*, 2001) was obtained from http://sapiens.wustl.edu/~zkan/TAP/. The Spidey (Wheelan *et al.*, 2001) executable was obtained from http://www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/spideyexec.html. The BLAT (Kent, 2002) executable was compiled from the source code made available at Jim Kent's Web page, http://www.soe.ucsc.edu/~kent/src/

Spliced threading

The GeneSeqer algorithm solves the problem of 'threading' an EST or cDNA into a genomic DNA sequence such that each nucleotide in the matching genomic DNA segment is consistently assigned exon or intron status. The threading preferentially selects high-scoring splice sites unless strongly contradicted by sequence similarity supporting lower scoring sites. An optimal alignment score is calculated by dynamic programming as described previously (Usuka *et al.*, 2000). In similar fashion, GeneSeqer also derives the optimal threading of a protein sequence onto the inferred translation of a genomic DNA segment, allowing gene prediction by similarity to putative homologs of the given locus (Usuka and Brendel, 2000).

Scoring

A number of parameters influence the optimal alignments, including standard scores for identities, mismatches and deletions within exon alignments. In addition, persistence within and switching between exon and intron states is governed by transition probabilities derived from splice site prediction values along the genomic sequence (Usuka *et al.*, 2000; Usuka and Brendel, 2000). These values are calculated for all positions in the genomic sequence prior to the spliced alignment. Precisely, a default donor site values of 0.00005 is assigned for any GT and 0.00002 for any GC or AT (similarly, 0.00005 for any AG and 0.00002 for any AC as potential acceptor sites). The other dinucleotides have a default score 0.000001 as donor or acceptor site value. These default values are replaced by $2 \times (P - 0.5)$ whenever that value is greater, where P is the respective Bayesian a posteriori splice site probability determined from models derived as described below. In addition, sites matching the U12-type intron consensus sequence ATCCTT downstream of the GT or AT donor site dinucleotide (Zhu and Brendel, 2003) in six or five positions are scored 0.99 and 0.9, respectively, to accommodate these special cases, which are not recognized by the Bayesian models for conventional U2-type introns. Empirically, the chosen scaling seems to give a good balance between scoring for sequence similarity and scoring for splice site consensus (the balance can be changed easily by providing the GeneSeqer program at run time with other than default parameters). We should emphasize that the incorporation of good splice site models is critical in the range of applications considered here when sequence divergence is deliberately allowed to be above the level expected to result from mere sequencing errors and polymorphisms.

The quality of a particular optimal alignment is assessed by similarity and coverage scores. Similarity scores are calculated as normalized alignment scores and are derived separately for each exon, the 50-base exon flanks of each predicted intron and the entire alignment by averaging over all exons of at least 50 bases. Note that with default parameters, in the absence of insertions/deletions a similarity score of s would correspond to $0.5 \times (1 + s) \times 100\%$ sequence identity. The coverage score gives the length of the matching region relative to the entire EST length (i.e. a completely matched EST would have coverage score 1.0).

Splice site probability models

To train species-specific splice site models for use in GeneSeqer, sequences of eukaryotic genes with multiple coding exons were taken from an early version of the ExInt database (Sakharkar *et al.*, 2000; Deutsch and Long, 1999) kindly provided by the authors. This database represents a non-redundant subset of GenBank Release 106. For each annotated intron, we retrieved the flanking 50 nucleotides upstream and downstream of the consensus GT and AG dinucleotides at the $5'$ and $3'$ intron ends, respectively. Introns with non-consensus dinucleotide ends were not selected, and entries for which the sequence information for the 50 nucleotide flanks was incomplete or ambiguous were removed. All selected sequence segments were pair-wise compared, and only one sequence was retained from any set of sequences that were identical at the peptide level for the in-frame translation of the exon parts. Also excluded were sites with in-frame stop codons in the annotated exon parts. The number of sites thus obtained for 10 different species ranged from a few hundred for fungal species to many thousands for *Caenorhabdites elegans*, *A.thaliana* and human beings.

All splice sites were assigned a phase label depending on the site of disruption of the open reading frame:

Sequence	NNN	| GT . . .	AG |	NNN
Phase 1 sites	1 2 0	1	1	1 2 0
Phase 2 sites	2 0 1	2	2	2 0 1
Phase 0 sites	0 1 2	0	0	0 1 2

The donor and acceptor sites of introns not splitting codons were assigned phase 1, the sites of introns disrupting codons between the first and second codon position were assigned phase 2 and sites disrupting codons between the second and third codon position were assigned phase 0. GT and AG dinucleotides within annotated exons or internal to annotated introns were considered 'false' sites. Non-redundant sets of false sites with complete 50 nucleotide flanks were compiled in the same way as the sets of true sites. Within-exon false sites were classified by phase as displayed above.

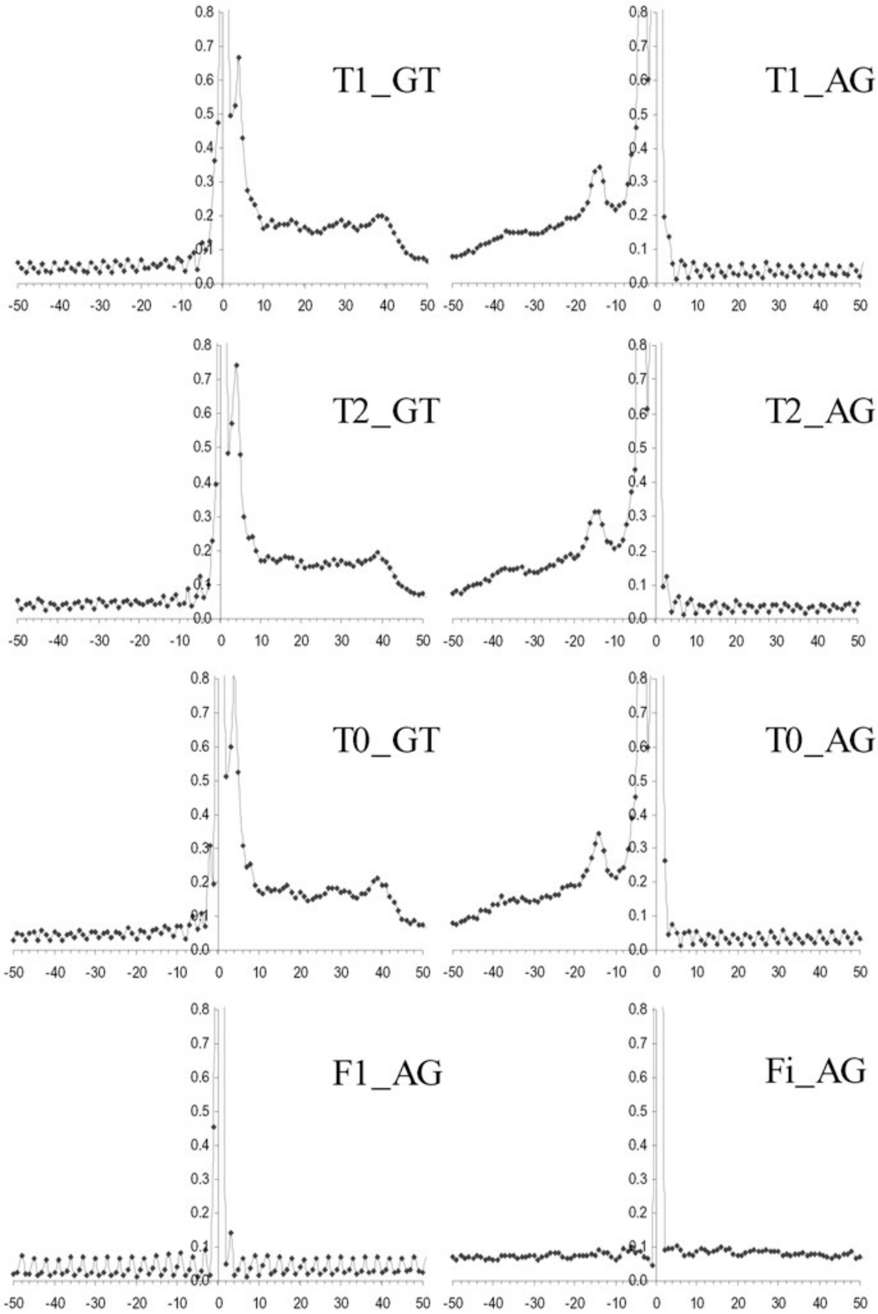

Fig. 1. Information plot around true and false splice sites in *C.elegans*. The information content I_i was calculated according to Equation (1) in the text. The central consensus dinucleotides have information content 2 (plot curtailed). Dataset designations: T1, T2, T0, true splice sites in phase 1, 2 and 0, respectively; F1, within-exon sites in phase 1; Fi, within-intron sites; GT, true and false donor sites; AG, true and false acceptor sites.

For the purpose of model training, the sets of within-exon false sites were reduced by random sampling to contain the same number of sites as the corresponding sets of true sites. The set of within-intron false sites was reduced to the largest size of the three phase-specific sets of true sites.

The extent of the splice signal was determined for each set of sites on the basis of information content plots (White *et al.*, 1992; Fig. 1). Precisely, for each position i for the aligned sequences, we calculated the quantity

$$I_i = 2 + \sum_{B \in (A,C,G,T)} f_{iB} \log_2 (f_{iB}), \qquad (1)$$

here, f_{iB} is the frequency of nucleotide B in position i. Thus, a position in which a single nucleotide occurs exclusively has maximal information content 2, whereas a position with equal representation of all nucleotides has minimal information content 0.

To avoid small sample effects or potential sampling biases, 100 bootstrap samples were derived from each dataset, each sample consisting of 500 sites drawn randomly with replacement from the original set. For each sample, the mono- and dinucleotide frequencies were determined in each position. The frequencies, f_{iB} were obtained as the averages of the frequencies observed in the 100 samples.

The splice signal extent was defined as the maximal continuous window around the GT and AG consensus dinucleotides for which no three consecutive positions have information content,

$$I_i \leq \overline{I} + 1.96\,\sigma_{\overline{I}}, \tag{2}$$

here, $\overline{I}$ is the average information content over all positions i more than 20 nucleotides away from the consensus dinucleotide (considering the base value, independent of the actual splice signal), and $\sigma_{\overline{I}}$ is the average sample standard deviation of $\overline{I}$.

As seen in Figure 1, the information content reflects both the splice site signal and the 3-base periodicity in coding regions. Because of the degeneracy of the genetic code, the third codon position is much less restricted by amino acid coding constraints. Instead, this position tends to be highly biased toward bases consistent with the overall G + C-content of the cognate genome or isochore (e.g. Sharp and Matassi, 1994). In our final models for splice site prediction, the splice site signal window was extended in both directions in order to include the 3-base periodicity signature to more clearly distinguish false within-exon and within-intron sites from true splice sites (see also Hebsgaard *et al.*, 1996).

Dinucleotide frequencies in each position of the signal windows were used as parameters in a Bayesian probability model for signal-window sized sequences. Let $S = s_{-l}s_{-l-1}, \ldots, s_{-1}\,\mathrm{GT}\,s_1s_2, \ldots, s_r$ denote the sequence around a potential donor site GT. We consider seven possible classifications of this sequence: the GT could be a true donor site in phase 1, 2 or 0, or the GT could occur within a coding exon in any of the three phases, or the GT could occur internal to an intron. Let these hypotheses be denoted by $T1, T2, T0, F1, F2, F0$ and Fi, respectively. Then

$$P\{H|S\} = \frac{P\{H\}P\{S|H\}}{\sum_H P\{H\}P\{S|H\}}, \tag{3}$$

where H indexes the seven hypotheses. The likelihoods $P\{S|H\}$ were calculated according to a first-order Markov model with transition probabilities $\tau_{iAjB} = f_{iAjB}/f_{iA}$, where f_{iAjB} is the frequency of dinucleotide *AB* in positions *ij*. Additionally, we explored a two-class model with cases $T = T1 + T2 + T0$ and $F = F1 + F2 + F0 + Fi$. In each case, a site S was classified according to the maximal value among the posterior probabilities $P\{H|S\}$. Prior probabilities were set to equal weights for all cases.

Performance statistics were derived following Brunak *et al.* (1991), using the notation of Snyder and Stormo (1995). Let 'positive' denote acceptance and 'negative' rejection of a site as splice site. The number of predicted positives, PP, consists of TP, true positives (real sites) and FP, false positives (non-sites). Similarly, the number of predicted negatives, PN, comprises FN, false negatives (real sites of low probability score) and TN, true negatives (non-sites). Let $\mathrm{AP} = \mathrm{TP} + \mathrm{FN}$ be the number of actual positives (true sites), and let $\mathrm{AN} = \mathrm{FP} + \mathrm{TN}$ be the number of actual negatives (non-sites). Thus, $\alpha = \mathrm{FN}/\mathrm{AP}$ and $\beta = \mathrm{FP}/\mathrm{AN}$. $Sn = \mathrm{TP}/\mathrm{AP} = 1 - \alpha$ measures the sensitivity of the method: what fraction of the real sites are correctly predicted? $Sp = \mathrm{TP}/\mathrm{PP}$ measures the specificity of the method: what fraction of the predicted positives are real sites?

Note that $S_p = 1 - (\mathrm{AN}/\mathrm{PP})\,\beta$ will typically be a better measure of performance than β because in practical applications (prediction of sites in genomic DNA) the population size of false sites would generally be much larger than the number of true sites, which will inflate even seemingly small values of β. On the other hand, values of *Sp* are hard to compare across datasets with widely differing sample sizes of actual negatives. For this reason, we use the normalized specificity.

$$\sigma = \frac{1 - \alpha}{1 - \alpha + \beta}. \tag{4}$$

σ is identical to *Sp* for balanced datasets with $\mathrm{AN} = \mathrm{AP}$. In general,

$$S_P = \frac{1 - \alpha}{1 - \alpha + r\beta}, \quad r = \frac{AN}{AP}. \tag{5}$$

A common decision strategy for accepting the null hypothesis, $H_0 = T$, is based on the Bayes Factor (BF),

$$\mathrm{BF} = \frac{P\{T|S\}}{(1 - P\{T|S\})} \Bigg/ \frac{P\{T\}}{(1 - P\{T\})}, \tag{6}$$

(for recent review, see Kass and Raftery, 1995). In words, BF is equal to the ratio of the posterior odds of H_0 to its prior odds. For the 2-class model, BF is computationally equivalent to the likelihood ratio $P\{S|T\}/P\{S|F\}$. For the 7-class model,

$$\mathrm{BF} = \frac{\sum_{x\in(1,2,0)} P\{T_x\}P\{S|T_x\}}{\sum_{x\in(1,2,0)} P\{T_x\}} \Bigg/ \frac{\sum_{x\in(1,2,0,i)} P\{F_x\}P\{S|F_x\}}{\sum_{x\in(1,2,0,i)} P\{F_x\}}, \tag{7}$$

which is the ratio of the average likelihood of the observed sequence, S under the two alternative hypotheses. Kass and Raftery (1995) suggest a critical value,

$$c = 2 \ln \mathrm{BF}, \tag{8}$$

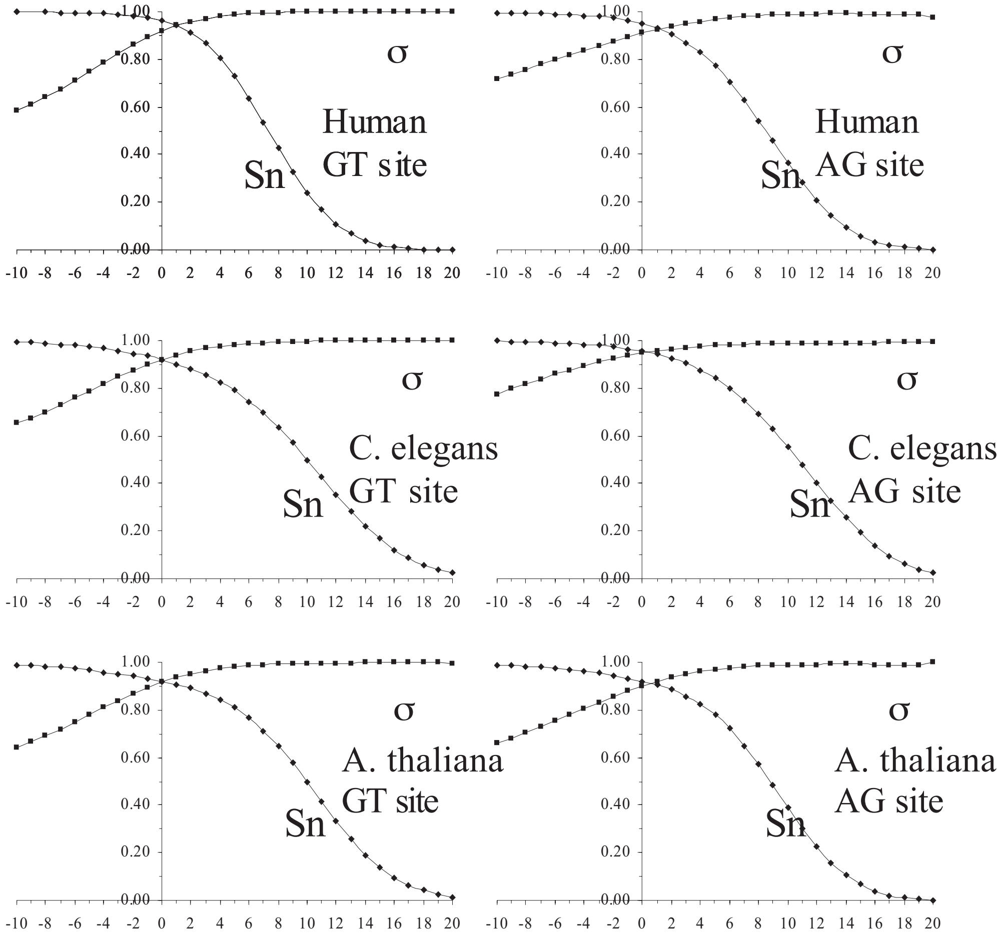

Fig. 2. Sensitivity (S_n) and normalized specificity (σ) as a function of BF. The BF may be used as a threshold for accepting a predicted splice site as true or not. Higher BF values increase specificity at the expense of sensitivity. The values were derived on the entire non-redundant sets of true and false sites. X-axis values are $c = 2 \ln \mathrm{BF}$.

for evaluation of evidence in favor of H_0, with approximate interpretation of 'positive evidence for H_0' for values in the range 2–6, 'strong support for H_0' for values in the range 6–10 and 'very strong support for H_0' for values exceeding 10. Sample performance statistics in dependence on c are shown in Figure 2.

As an independent test of the accuracy of the method, we also implemented the models in our SplicePredictor tool along with the earlier logitlinear models (Kleffe and Brendel, 1998). Consideration of local optimality of splice site scores significantly increases prediction specificity (Brendel and Kleffe, 1998; Pertea *et al.*, 2001). A simple method for local pruning of suboptimal predicted sites was implemented as the default option in SplicePredictor. Briefly, within the vicinity of high scoring sites (Bayes factor at least 6.0, no well scoring sites of the opposite type close by) lower scoring sites remain unreported. We assessed the accuracy of splice site prediction using this method on a subset of 329 *A.thaliana* genes from the set compiled by Pertea *et al.* (2001) by restricting the set to only those genes with EST confirmation for all splice sites and confining the non-coding gene flanks to at most 500 nucleotides on both ends. SplicePredictor performance was comparable with GeneSplicer (Pertea *et al.*, 2001), which (with default settings) achieves somewhat lower sensitivity but higher specificity (program kindly supplied by the authors).

Evaluation

To benchmark the prospects and limits of gene prediction by spliced alignment, we evaluated the GeneSeqer performance on the AraSet *Arabidopsis* gene set distributed for such purposes by Pavy *et al.* (1999), available at http://www.psb.ugent.be/bioinformatics/GeneComp/. This set consists of 74 contigs comprising 2–4 genes each, 168 genes and 859 introns in total. Spliced alignments were based on the mapping of 176 195 *Arabidopsis* ESTs that were downloaded from the NCBI dbEST database (Boguski *et al.*, 1993, http://www.ncbi.nlm.nih.gov/dbEST/).

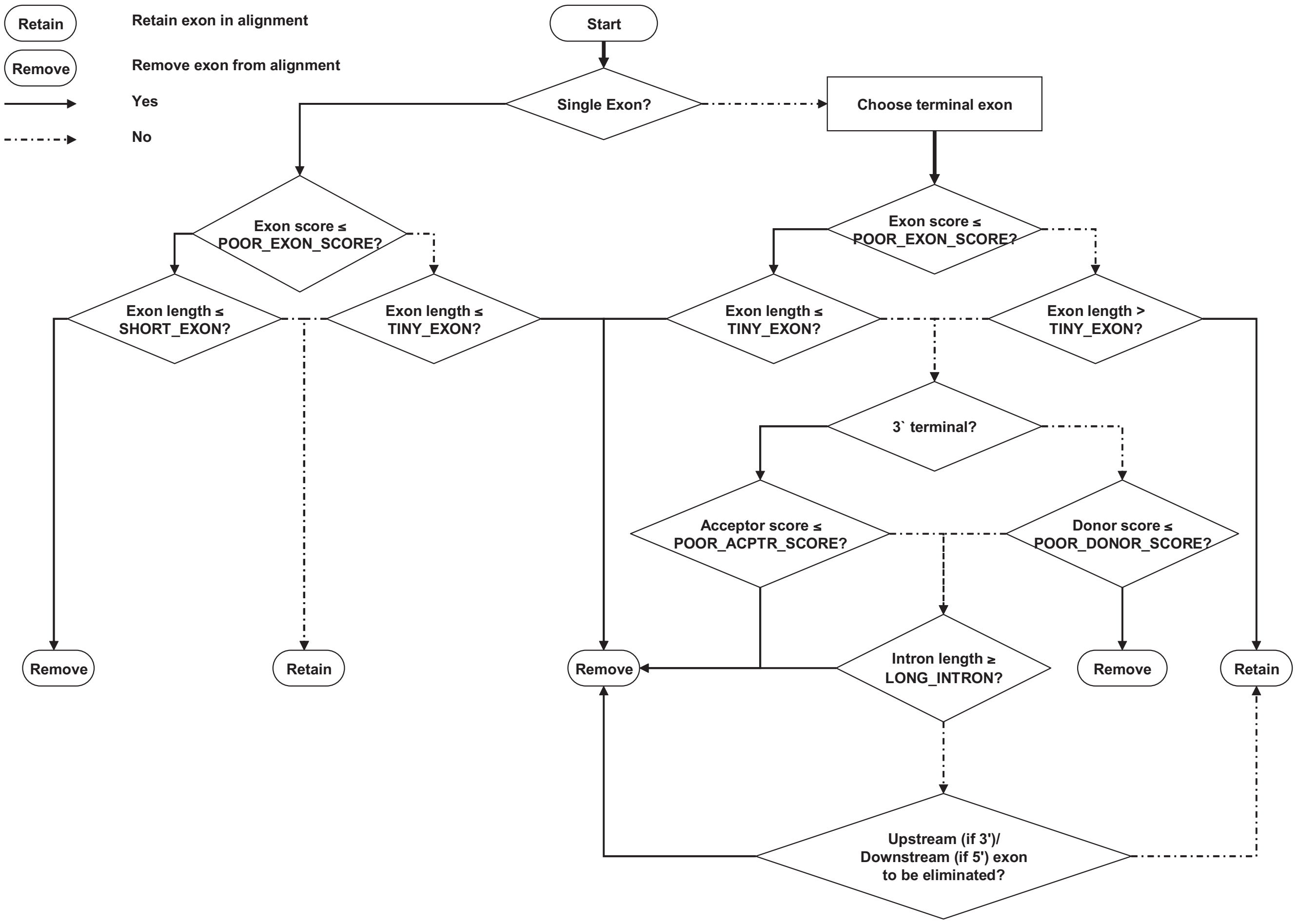
Retain
Retain exon in alignment
Remove
Remove exon from alignment
Yes
No
Start
Single Exon?
Choose terminal exon
Exon score ≤ POOR_EXON_SCORE?
Exon score ≤ POOR_EXON_SCORE?
Exon length ≤ SHORT_EXON?
Exon length ≤ TINY_EXON?
Exon length ≤ TINY_EXON?
Exon length > TINY_EXON?
3` terminal?
Acceptor score ≤ POOR_ACPTR_SCORE?
Donor score ≤ POOR_DONOR_SCORE?
Intron length ≥ LONG_INTRON?
Upstream (if 3')/ Downstream (if 5') exon to be eliminated?
Remove
Retain
Remove
Remove
Retain

Fig. 3. Decision tree for post-processing of GeneSeqer spliced alignments. Capitalized items are parameters supplied in a parameter file. Weakly supported terminal exons are removed successively from both ends.

To evaluate prediction accuracy at the intron level, we define correct introns, overlapping introns, wrong introns and missed introns as in Pavy *et al.* (1999). Thus, a predicted intron identical to an annotated intron is classified as a 'correct intron'. An 'overlapping intron' refers to a predicted intron overlapping with some annotated intron, but with a different 5′ and/or 3′ splice site. A 'wrong intron' refers to a predicted intron overlapping with annotated exons, but not with annotated introns. Both overlapping introns and wrong introns are counted as incorrect (false positive) predictions (note that this assumes a lack of alternative splicing in the test set). 'Missed introns' are annotated introns that are not overlapped by any predicted intron (false negatives). Because only introns in coding sequences (CDS) are annotated in AraSet, introns predicted by spliced alignment outside of CDS cannot be evaluated. Thus, sensitivity at the intron level is defined as (number of correct introns/number of annotated introns), and specificity is determined as (number of correct introns/number of predicted introns in CDS).

ALGORITHM

Quality adjustments

By default, GeneSeqer will align any EST to a genomic locus with which it shares at least partial significant similarity as determined in the fast screen for matching loci described below. This may result in optimal scoring, but clearly poor alignments over the entire EST when the significant similarity is limited to disjoint segments of the EST. While such alignments can still be useful to indicate exon potential in the matching genomic segments (if not an entire gene structure), we have now implemented in the GeneSeqer program a post-processing step that quality-adjusts such alignments based on user-specified parameters. Briefly, a predicted gene structure is assessed exon by exon, starting with the terminal exons, with weakly matching terminal exons recursively being eliminated. The elimination process involves a decision tree (Fig. 3). For example, the 3′-most exon in a multiple-exon predicted gene structure is quality-adjusted as follows: (1) Is the exon score below the parameter POOR_EXON_SCORE (default: 0.7)? If yes, and (2) the exon length is at most TINY_EXON bases (default: 20), the exon is removed from the alignment. If the exon is longer and (3) the acceptor site score is at most POOR_ACPTR_SCORE (default: 0.5) or (4) the length of the intron is at least LONG_INTRON bases (default: 300), the exon is removed. If conditions (3) and (4) for elimination are not met, the exon is retained unless the upstream exon is to be eliminated by the same criteria. To complete the decision tree, exons that (1) score above POOR_EXON_SCORE are retained if, (2), they are of length greater than TINY_EXON. However, they are eliminated if they are shorter and successively either (3) the acceptor score is poor, (4) the intron is long, or the upstream exon is weak. Predicted 5′ gene structure ends are similarly adjusted.

The parameters for this quality adjustment are entirely empirical and should be changed appropriate to a specific alignment task. For example, if the alignment is of a cognate cDNA to its genomic origin, short terminal exons may be missed with default parameters that are selected to avoid chance matches in large-scale mapping of non-cognate ESTs. To provide most flexibility, the GeneSeqer text display of an alignment includes all the exons, but only the quality-trimmed parts are used for consensus gene prediction (see below).

Strand selection

Based on sequence similarity alone, a spliced alignment could be made equally with either strand of a genomic DNA. For multi-exon alignments, GeneSeqer orients the alignment to maximize the average splice site score. For example, the alignments for the example displayed in Figure 4 assign high splice site scores for introns two and three, thus aiding in the detection of the first intron as an AT–AC intron in the same orientation. In ambiguous cases, occasional retention of a poly-A tag in the EST sequence may indicate the direction of transcription. For single exon alignments, GeneSeqer assigns a putative transcription orientation based on overlap with multi-exon alignments as described next. In general, no attempt is made to use annotated orientation, if available, because we have found such annotation not always reliable. However, a particular alignment orientation can be enforced at run time.

Consensus gene structures

A critical step in our strategy to predict gene structure by spliced alignment is the derivation of a consensus gene structure prediction from multiple, possibly low scoring, overlapping spliced alignments. If the resulting gene structure spans multiple-exons and contains an open reading frame across these multiple-exons, confidence in the prediction should be very high because the GeneSeqer algorithm (unlike *ab initio* gene prediction programs) does not score in any way for coding frame consistency in the initial alignment step. Figure 4 provides a typical example, discussed below.

Determination of consensus gene structures in our algorithm is a multi-step process. First, all EST alignments are clustered into predicted gene locations (PGLs) based on genomic location. This clustering is achieved by going through all the alignments by increasing left-point coordinate. Clusters are separated by gaps of at least JOIN_LENGTH bases, a parameter that can be changed at run time (default: 300). An exception to this is made if a new alignment is of opposite orientation compared with the current PGL; in this case, a new PGL is assigned. Single exon alignments are displayed in the orientation of their associated PGL. If a PGL consists entirely of single exon alignments, then the orientation is determined first by the presence of any potential poly-A tags and second by choosing the orientation that gives the longest open reading frame. It is clear that intergenic regions less than JOIN_LENGTH may cause problems, but empirically these rules seem to work very well [see Zhu *et al.* (2003), for extensive applications to *Arabidopsis*].

Within each PGL, alternative splicing would result in inconsistent predicted gene structures (PGSs) from individual ESTs. This is represented in the GeneSeqer output by multiple alternative gene structures (AGSs) within a single PGL. An example is given in Figure 5. Assembly of AGSs proceeds left to right, with each PGS added into the current AGS as long as its exon/intron assignments are consistent with the current AGS. Otherwise, a new AGS is started. The alignment ends of an AGS may be slightly adjusted to fit a PGS. This adjustment eliminates wrong alternative splicing predictions that would otherwise result from weak, random matching of EST end sequences, which are typically of lower sequence quality. The GeneSeqer output only indicates the alternative transcript isoform fragments confirmed by spliced alignment but does not further process these fragments to assemble all potential full-length transcript isoforms. However, the output could easily be parsed and re-formatted for input into the TAP program (Kan *et al.*, 2001) for this purpose (currently, TAP uses sim4 spliced alignments by default). Haas *et al.* (2003) recently introduced an alternative algorithm to generate maximal alignment assemblies.

Fast screen for matching ESTs

Efficient use of EST evidence for genome annotation requires mapping large EST collections onto BAC-size genomic DNA segments. Because dynamic programming is computationally prohibitive for such large problems, a fast screen must be implemented to select promising EST matches for gene-sized genomic segments. In the absence of very long introns, the dynamic programming algorithm can then be applied to the selected DNA input (the case of long introns can be handled by more sophisticated screening that eliminates presumed intron-internal sequences; not pursued here). For GeneSeqer, we have implemented the suffix array method of Manber and Myers (1993) for pre-processing of the EST database. Note that for applications in which the genomic

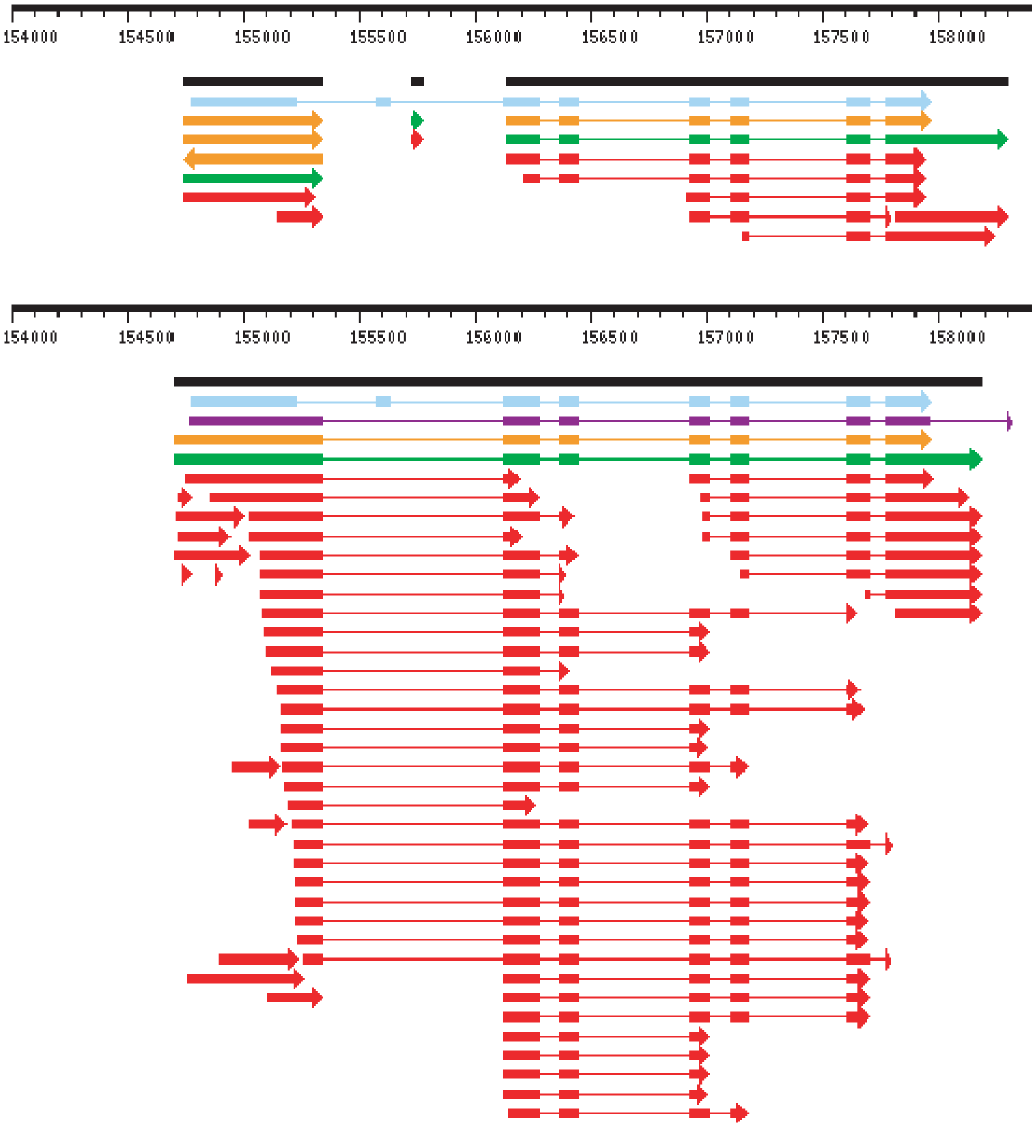

Fig. 4. Gene structure annotation for a putative rice gene on chromosome one. The schematic displays of the GeneSeqer spliced alignments were generated with the GeneSeqer Web server at the PlantGDB site (http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi; Schlueter *et al.*, 2003). The scale refers to the numbering of the BAC sequence deposited in GenBank as accession AP003271. GenBank CDS annotation is shown in light blue, with solid boxes corresponding to exons and thin lines corresponding to introns. The arrow indicates the direction of transcription. The same convention is used for EST spliced alignments (red), alternative gene structures (green) derived from consistently overlapping EST spliced alignments, long open reading frames (orange) and protein spliced alignments (purple). Upper panel: Spliced alignment of nine rice ESTs confirms the five 3′ most exons of the annotated gene structure, but is inconclusive with respect to the 5′ end of the gene. Lower panel: Spliced alignment with 53 barley ESTs suggests a seven-exon gene structure (green), which encodes a single long open reading frame (orange). The translation product is highly similar to the *Arabidopsis* gene At3g53520 product, a UDP-glucuronic acid decarboxylase, and direct spliced alignment of the *Arabidopsis* protein supports the same gene structure (purple). A protein database search showed that the rice homolog has been deposited as GenBank accession BAB84333.

DNA query is fixed (e.g. annotation of a complete genome), additional pre-processing the genomic DNA sequence may be considered.

Three parameters determine the outcome of the initial screen for matching ESTs. The GeneSeqer *-x* wsize option specifies the minimal exact match size for successful extension (typically, *wsize* is set to 12–16; higher values allow much faster screening for high quality matches only). Precisely, the genomic DNA query is processed along the 5′ to 3′ direction, with each consecutive wsizemer match against the EST

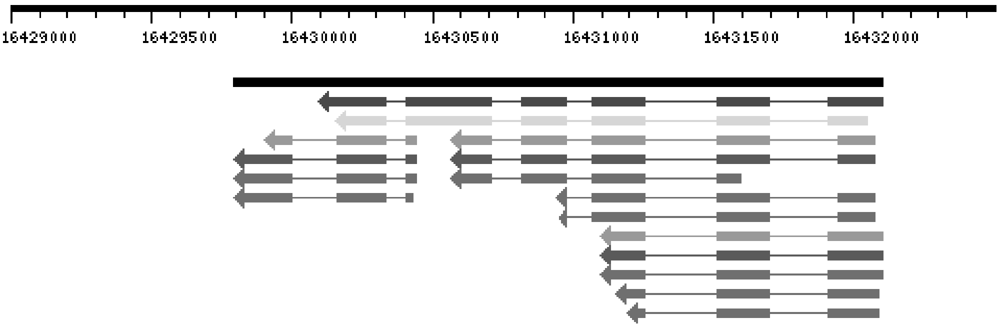

Fig. 5. Alternative gene structure prediction for AraSet entry seq62 representing the *Arabidopsis* At4g37070 gene. Symbols and colors are as explained in the legend to Figure 4. In addition, GenBank mRNA annotation is shown in dark blue. The eight matching ESTs (red) were assembled into three consistent transcript fragments (green). The first intron has two alternative donor sites, supported by two and three ESTs, respectively. Note that the GeneSeqer program does not attempt to display all possible full-length transcript isoforms. However, inspection of the open reading frames (orange) suggests that the gene may have two transcript isoforms differing only in the first donor site, but maintaining the reading frame such that the two protein isoforms differ only by an additional 11 amino acids in the longer protein.

database added to a set of linked lists that store match information for each specific EST. As the linked lists grow, the matches from each individual EST are continuously merged into high-scoring segment pairs (HSPs) that allow for small insertions and deletions in both genomic DNA and EST. The *-y minq*HSP option sets the minimal score for HSPs to be considered. Related HSPs are then further chained together to define matching regions between the genomic DNA and the specific EST using the algorithm of Pearson and Lipman (1988), with a minor penalty for long gaps in the genomic region (possible introns). These two steps are analogous to the first two steps in the algorithm applied by sim4 (Florea *et al.*, 1998). However, sim4 only utilizes the best scoring chain for each EST, whereas multiple non-intersecting chains with significant scores higher than a cutoff value would be selected in GeneSeqer. This allows a single EST to be matched to different genomic loci. This property is crucial for the applications discussed here. The cognate EST location is easily identified as the highest scoring match, but, in addition, an EST can often be successfully used to identify gene structure in a duplicated locus, in particular a locus with potentially low cognate EST representation (Zhu *et al.*, 2003). The cutoff value for successful HSP chains is specified by the GeneSeqer *-z minq*HSP*c* argument. Each promising region is then slightly expanded to allow for uncertainties at the ends, and the full dynamic programming alignment is applied to this genomic DNA region and the entire EST sequence.

Complexity

A typical application of GeneSeqer is to map a large EST collection (total sequence length M) to a single genomic sequence of length n. The whole process of EST mapping consists of three parts: construction of the suffix array for the EST sequences, genomic localization (fast screen with GeneSeqer option *-x* wsize), and spliced alignment. The run time for building the suffix array for ESTs is $O(M \cdot \log M)$ using the algorithm of Manber and Myers (1993). This computational time is typically negligible because a large number of ESTs are usually pre-processed to build the suffix array, which avoids potential overhead in repeated small-scale analyses. The genomic localization step is very fast with run time $O[n \cdot (\text{wsize} + \log M)]$, based on a search algorithm for suffix arrays using longest common prefixes (Gusfield, 1997). Therefore, the computation for large-scale mapping is dominated by the cost for the spliced alignment part and thus is linearly proportional to the expected number of alignments and the square of the average alignment length.

RESULTS AND DISCUSSION

Spliced alignment with heterologous ESTs

Figure 4 illustrates the application of spliced alignment for gene structure annotation. The upper panel shows nine PGSs with rice ESTs (red) that result in three disjoint AGSs (green). The complete alignments are available as Supplementary information at http://www.plantgdb.org/AtGDB/prj/BXZ03B/atac/gs_sorted-output-Ex1_top.html. The three AGSs are supported by similarity scores of about 0.8, 0.9 and 0.95, respectively. While the 3′-terminal exons of the annotated gene structure are confirmed by spliced alignment, contradictory results are obtained at the 5′ end. This issue is resolved when ESTs from plants other than rice are added. Using the GeneSeqer Web service at PlantGDB (Schlueter *et al.*, 2003), a total of 266 ESTs could be significantly aligned in this region. The lower panel in Figure 4 depicts the results for a subset of these ESTs, all derived from barley. Several of these ESTs bridge all coverage gaps and predict a single gene structure with seven exons (green). An open reading frame (orange) spans all the exons, and its translation identifies the gene as coding for a UDP-glucuronic acid decarboxylase. Of particular interest is identification of the first intron (785 bases) as a U12-type intron with AT–AC borders. The intron is in a coding region that is highly conserved with an *Arabidopsis* homolog (At3g53520), and the *Arabidopsis* gene also has a U12-type intron in the same position (although none but the U12 signatures are preserved in the intron sequences; see Zhu *et al.*, 2003).

By including splice site scoring and preferences, GeneSeqer can use even quite diverged ESTs to predict the correct gene structure (in this case, the barley ESTs have an average similarity score of only 0.72). For comparison, none of the other programs we tried (sim4, BLAT, Spidey) produced any alignments for the same genomic DNA and EST (results shown as Supplementary information at http://www.plantgdb.org/AtGDB/prj/BXZ03B/atac/gs_sorted-output-Ex1_top.html).

Evaluation of spliced alignment accuracy

We have recently reported on the utility of spliced alignment in correcting and refining *A.thaliana* genome annotation (Zhu *et al.*, 2003). As an independent assessment of the applicability and performance quality of GeneSeqer, here we evaluate its accuracy relative to the AraSet test set compiled by Pavy *et al.* (1999). All available *Arabidopsis* ESTs were mapped onto the AraSet contigs using GeneSeqer default parameters. Post-screening of the reported alignments was used to select subsets of alignments satisfying more stringent match criteria. Because the alignments with terminal ESTs correspond to predicted transcript ends rather than coding region ends as in the AraSet annotation, evaluations were made entirely on the intron level, using standard performance measures (Pavy *et al.*, 1999).

The results are summarized in Table 1. With default parameters, the spliced alignment indicated 782 introns (compared to 859 annotated introns in AraSet). Of these, 625 introns coincided with annotated introns for a sensitivity of 0.728. Assessment of specificity is less straightforward. First, spliced alignment, unlike *ab initio* programs tested on AraSet, can reveal introns in untranslated regions (UTRs). Here, 76 introns were predicted outside of the CDS bounds annotated in AraSet. A careful inspection indicated that this set contains both UTR introns and introns of genes that were omitted in the AraSet annotation (see below). A second problem is that some of the overlapping introns may correspond to correctly predicted alternative transcripts. Thus, the listed specificity of 0.885 may be underestimating the actual specificity.

In order to clearly separate errors of the spliced alignments from errors in the AraSet annotation, we evaluated a subset of all predicted introns that satisfy very stringent alignment quality criteria. Let Pd (Pa) and Sd (Sa) denote the splice site score and local similarity score for each donor (acceptor) site, respectively. Requiring $\text{Pd} > 0.00002, \text{Pa} > 0.00002$ and $\text{Sd} > 0.95, \text{Sa} > 0.95$ selects only introns with canonical splice sites supported by EST matching with more than 97.5% identity in the flanking 50 exon bases. For this subset, 463 of the 471 predicted introns within CDS bounds coincide with the AraSet annotation. The remaining eight introns were further scrutinized, and all seem authentic (Table 2). In three cases (seq53, seq62 and seq72), the annotated introns are supported by other ESTs, and thus the two conflicting coordinate sets represent alternative splicing events. In the other five cases, there is no EST support for the annotation, and thus the EST-supported coordinates may be assumed to be the correct annotation. With that correction, the specificity of GeneSeqer high-quality intron prediction is 100%, as expected. Sensitivity in this case dropped to just over 50%. For comparison, exon level sensitivity and specificity were estimated at just above and below 80%, respectively, for the best *ab initio* gene prediction programs (Pavy *et al.*, 1999).

Table 1. GeneSeqer intron level performance evaluation relative to AraSet (859 annotated introns)

	Default*	Canonical sites*	High quality*
Predicted introns	782	684	499
Predicted introns in UTR[†]	76	42	28
Predicted introns in CDS	706	642	471
Correct introns	625	609	463
Overlapping introns	64	32	8[‡]
Wrong introns	17	1	0
Missed introns	188	235	391
Specificity	0.885	0.949	0.983
Corrected specificity	≥0.895	≥0.961	1.000
Sensitivity	0.728	0.709	0.539

*Default, GeneSeqer default parameters; Canonical sites, predicted canonical introns only; High quality, canonical introns with high sequence similarity of EST to flanking exons; see text for details.

[†]Some of these introns are actually from unannotated genes; see text for details.

[‡]Listed in Table 2.

Table 2. Annotated introns in AraSet contradicted with high quality intron predictions derived from EST spliced alignments

SeqID	Annotated intron		Predicted intron		EST evidence*	Alternative splicing
	5′ss	3′ss	5′ss	3′ss		
seq06	5753	5885	5764	5885	gi:8695314	N
seq53	3795	3708	3795	3735	gi:8715801	Y
seq62	2139	2351	2106	2351	gi:1054038	Y[†]
seq72	6486	6656	6481	6656	gi:4714042	Y
seq73	2232	2078	2258	2091	gi:19828992	N
seq73	3515	3398	3515	3407	gi:19868516	N
seq81	4016	3985	4088	3985	gi:14580187	N
seq84	4759	5173	4759	5170	gi:19865385	N

*Only one EST is listed for each predicted intron; for details, see Supplementary information at http://www.plantgdb.org/AtGDB/prj/BXZ03B/AraSet/AraSet-AtGDB.php

[†]See Figure 5.

Sensitivity for the spliced alignment approach depends mostly on the availability of ESTs. However, when using non-cognate ESTs, we are also assessing the ability of the program to use such data for accurate prediction. As displayed in Table 1, with GeneSeqer default parameters a gain of about 20% in sensitivity is accompanied by a drop in specificity of about 10%. Restriction of the predicted intron set to only canonical introns (without the additional requirement for high-quality flanking exon matching) gives intermediate values.

There were 28 introns in the high-quality subset that are not located within the annotated CDS bounds and are thus potential UTR introns. Further analysis indicates that some of these introns are actually from genes that are not annotated in AraSet. For example, three genes are annotated in AraSet contig seq25, with a 4.4 kb 'intergenic region' between the second and the third genes. The most recent *Arabidopsis* genome annotation suggests that there is a gene At5g63670 with five exons in the 'intergenic region', supported by three full-length cDNAs and three ESTs. Similar situations also occur in the AraSet contigs seq30, seq41 and seq69. Supporting data for all these cases are available as Supplementary information at http://www.plantgdb.org/AtGDB/prj/BXZ03B/AraSet/AraSet-AtGDB.php

Applications to rice genome annotation

To test the utility of GeneSeqer for annotation of the rice genome, we analyzed a randomly selected rice BAC (GenBank accession AP002487) in detail. Spliced alignment results of the central 44 000 bases of the sequence are displayed as Supplementary information at http://www.plantgdb.org/AtGDB/prj/BXZ03B/OsBAC/gs_sorted-output-Ex2_top.html. Overall, spliced alignment confirmed six genes, only one of which agrees with the current gene annotation provided in the GenBank file. For each gene, a sufficient number of ESTs from heterogeneous sources could be found to give a complete tiling of the gene, supported by open reading frames spanning all exons and showing high similarity to known *Arabidopsis* gene products.

CONCLUSIONS

After genome sequencing and assembly, genome annotation is the most critical task in the characterization of the genetic blueprint of an organism. For all eukaryotic model organisms that have been sequenced, the annotation efforts have continued and are continuing for years after the initial sequence release. Thus, the human genome is still being evaluated, and in particular, the abundance of alternative splicing of human genes has only recently been appreciated (Mironov *et al.*, 1999; Modrek and Lee, 2002; Brett *et al.*, 2002). The annotation tasks for plant genomes currently pose distinct challenges compared with vertebrate genome annotation. First, EST and full-length cDNA availability is much less for plants than for human beings and mouse. Currently, there are 501 000 wheat ESTs as the largest plant collection, compared to more than five million for human beings and 4 million for the mouse (see http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). Only 131 000 rice ESTs were publicly available at the time of preparing this manuscript, less than the 179 000 *Arabidopsis* ESTs for an about 3-fold smaller genome. Second, all plant genomes surveyed to date are replete with gene duplications as a result of both polyploidization and random segmental duplications (e.g. Gaut, 2001; Blanc *et al.*, 2003).

We have recently reported the mapping of all *Arabidopsis* ESTs onto the *Arabidopsis* genome using GeneSeqer and showed that about 65% of annotated gene locations had EST evidence, with full coverage for about 23% of the genes (Zhu *et al.*, 2003). Here we have presented details of the GeneSeqer algorithm with respect to the derivation of consensus gene structures from multiple ESTs from potentially hetereogeneous, diverged sources. A number of key differences in the algorithm compared with other programs geared toward fast alignment of cognate ESTs allow efficient use of non-native EST resources. We believe this will greatly aid in the annotation of plant genomes, particularly rice and maize. The GeneSeqer Web service at PlantGDB (http://www.plantgdb.org/cgi-bin/GeneSeqer.cgi) should allow any member of the plant research community easy access to the annotation tools, and we hope that such community input will quickly improve the status of plant genome annotation.

ACKNOWLEDGEMENTS

We would like to thank Dr Qunfeng Dong, Shannon D.Schlueter and Michael E.Sparks for critically reading the manuscript and for helpful discussions. V.B. was supported in part by NSF grants DBI-9872657 and DBI-0110254.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Blanc,G., Hokamp,K. and Wolfe,K.H. (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. *Genome Res.*, **13**, 137–144.

Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST–database for 'expressed sequence tags'. *Nat. Genet.*, **4**, 332–333.

Bouck,J., Yu,W., Gibbs,R. and Worley,K. (1999) Comparison of gene indexing databases. *Trends Genet.*, **15**, 159–162.

Brendel,V. and Kleffe,J. (1998) Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. *Nucleic Acids Res.*, **26**, 4748–4757.

Brett,D., Pospisil,H., Valcarcel,J., Reich,J. and Bork,P. (2002) Alternative splicing and genome complexity. *Nat. Genet.*, **30**, 29–30.

Brunak,S., Engelbrecht,J. and Knudsen,S. (1991) Prediction of human mRNA donor and acceptor sites from the DNA sequence. *J. Mol. Biol.*, **220**, 49–65.

Deutsch,M. and Long,M. (1999) Intron–exon structures of eukaryotic model organisms. *Nucleic Acids Res.*, **27**, 3219–3228.

Florea,L., Hartzell,G., Zhang,Z., Rubin,G.M. and Miller,W. (1998) A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.*, **8**, 967–974.

Gaut,B.S. (2001) Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res.*, **11**, 55–66.

Gelfand,M.S., Mironov,A.A. and Pevzner,P.A. (1996) Gene recognition via spliced sequence alignment. *Proc. Natl Acad. Sci., USA*, **93**, 9061–9066.

Gusfield,D. (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York, pp. 152–155.

Haas,B.J., Delcher,A.L., Mount,S.M., Wortman,J.R., Smith,R.K.,Jr, Hannick,L.I., Maiti,R., Ronning,C.M., Rusch,D.B., Town,C.D., Salzberg,S.L. and White,O. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.*, **31**, 5654–5666.

Haas,B.J., Volfovsky,N., Town,C.D., Troukhar,M., Alexandrov,N., Feldmann,K.A., Flavell,R.B., White,O. and Salzberg,S.L. (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* **3**, research 0029.1–0029.12.

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

Huang,X., Adams,M.D., Zhou,H. and Kerlavage,A.R. (1997) A tool for analyzing and annotating genomic sequences. *Genomics*, **46**, 37–45.

Kalyanaraman,A., Aluru,S., Kothari,S. and Brendel,V. (2003) Efficient clustering of large EST data sets on parallel computers. *Nucleic Acids Res.*, **31**, 2963–2974.

Kan,Z., Rouchka,E.C., Gish,W.R. and States,D.J. (2001) Gene structure prediction and alternative splicing analysis using genomically aligned ESTs. *Genome Res.*, **11**, 889–900.

Kass,R.E. and Raftery,A.E. (1995) Bayes Factors. *J. Am. Stat. Assoc.*, **90**, 377–395.

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Liang,F., Holt,I., Pertea,G., Karamycheva,S., Salzberg,S. and Quackenbush,J. (2000) An optimized protocol for analysis of EST sequences. *Nucleic Acid Res.*, **28**, 3657–3665.

Manber,U. and Myers,G. (1993) Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.*, **22**, 935–948.

Mathé,C., Sagot,M.F., Schiex,T. and Rouzé,P. (2002) Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.*, **30**, 4103–4117.

Mironov,A.A., Fickett,J.W. and Gelfand,M.S. (1999) Frequent alternative splicing of human genes. *Genome Res.*, **9**, 1288–1293.

Modrek,B. and Lee,C. (2002) A genomic view of alternative splicing. *Nat. Genet.*, **30**, 13–19.

Mott,R. (1997) EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.*, **13**, 477–478.

Murakami,K. and Takagi,T. (1998) Gene recognition by combination of several gene-finding programs. *Bioinformatics*, **14**, 665–675.

Ogasawara,J. and Morishita,S. (2002) Fast and sensitive algorithm for aligning ESTs to human genome. In *Proceedings of the First IEEE Computer Society Bioinformatics Conference*. Stanford, California, pp. 43–53.

Pavy,N., Rombauts,S., Déhais,P., Mathé,C., Ramana,D.V., Leroy,P. and Rouzé,P. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics*, **15**, 887–899.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci., USA*, **85**, 2444–2448.

Pertea,G., Huang,X., Liang,F., Antonescu,V., Sultana,R., Karamycheva,S., Lee,Y., White,J., Cheung,F., Parvizi,B., Tsai,J. and Quackenbush,J. (2003) TIGR gene indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. *Bioinformatics*, **19**, 651–652.

Pertea,M., Lin,X. and Salzberg,S.L. (2001) GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Res.*, **29**, 1185–1190.

Rogic,S., Mackworth,A.K. and Ouellette,F.B. (2001) Evaluation of gene-finding programs on mammalian sequences. *Genome Res.*, **11**, 817–832.

Sakharkar,M., Long,M., Tan,T.W. and de Souza,S.J. (2000) ExInt: an exon/intron database. *Nucleic Acids Res.*, **28**, 191–192.

Salzberg,S.L. (1997) A method for identifying splicing sites and translational start sites in eukaryotic mRNA. *Comput. Appl. Biosci.*, **13**, 365–376.

Schlueter,S.D., Dong,Q. and Brendel,V. (2003) GeneSeqer@PlantGDB—gene structure prediction in plant genomes. *Nucleic Acids Res.*, **31**, 3597–3600.

Sharp,P.M. and Matassi,G. (1994) Codon usage and genome evolution. *Curr. Opin. Genet. Dev.*, **4**, 851–860.

Snyder,E.E. and Stormo,G.D. (1995) Identification of protein coding regions in genomic DNA. *J. Mol. Biol.*, **248**, 1–18.

Usuka,J. and Brendel,V. (2000) Gene structure prediction by spliced alignment of genomic DNA with protein sequences: increased accuracy by differential splice site scoring. *J. Mol. Biol.*, **297**, 1075–1085.

Usuka,J., Zhu,W. and Brendel,V. (2000) Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics*, **16**, 203–211.

Wheelan,S.J., Church,D.M. and Ostell,J.M. (2001) Spidey: a tool for mRNA-to-genomic alignments. *Genome Res.*, **11**, 1952–1957.

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

Zhu,W. and Brendel,V. (2003) Identification, characterization, and molecular phylogeny of U12-dependent introns in the *Arabidopsis thaliana* genome. *Nucleic Acids Res.*, **31**, 4561–4572.

Zhu,W., Schlueter,S.D. and Brendel,V. (2003) Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. *Plant Physiol.*, **132**, 469–484.