doi:10.1006/jmbi.2000.3641 available online at http://www.idealibrary.com on IDEAL® *J. Mol. Biol.* (2000) **297**, 1075–1085

JMB

Gene Structure Prediction by Spliced Alignment of Genomic DNA with Protein Sequences: Increased Accuracy by Differential Splice Site Scoring

Jonathan Usuka[1] and Volker Brendel[2]*

[1]*Department of Chemistry Stanford University, Stanford CA 94305, USA*

[2]*Department of Zoology and Genetics, Iowa State University, 2112 Molecular Biology Building, Ames IA 50011-3260, USA*

**Corresponding author*

Gene identification in genomic DNA from eukaryotes is complicated by the vast combinatorial possibilities of potential exon assemblies. If the gene encodes a protein that is closely related to known proteins, gene identification is aided by matching similarity of potential translation products to those target proteins. The genomic DNA and protein sequences can be aligned directly by scoring the implied residues of in-frame nucleotide triplets against the protein residues in conventional ways, while allowing for long gaps in the alignment corresponding to introns in the genomic DNA. We describe a novel method for such spliced alignment. The method derives an optimal alignment based on scoring for both sequence similarity of the predicted gene product to the protein sequence and intrinsic splice site strength of the predicted introns. Application of the method to a representative set of 50 known genes from *Arabidopsis thaliana* showed significant improvement in prediction accuracy compared to previous spliced alignment methods. The method is also more accurate than *ab initio* gene prediction methods, provided sufficiently close target proteins are available. In view of the fast growth of public sequence repositories, we argue that close targets will be available for the majority of novel genes, making spliced alignment an excellent practical tool for high-throughput automated genome annotation.

© 2000 Academic Press

Keywords: target protein; intron; spliced alignment; dynamic programming; Hidden Markov Model

Introduction

Gene identification by sequence inspection is a difficult but necessary task at the annotation step of genome sequencing projects (for reviews, see Claverie, 1997; Burge & Karlin, 1998). Experimental evidence for exon assignments may derive from cDNA sequencing or reverse transcriptase polymerase chain reaction (RT-PCR) (Sze *et al.*, 1998). Typically, the cDNA sequences will come from independently sequenced cDNA libraries, and assignment of a cDNA to its cognate gene will be on the basis of sequence identity. In the presence of sequencing errors, or in the case of non-cognate but related cDNA from a homologous locus in a different species or a different member of the same gene family, the assessment of significant sequence similarity must be error-tolerant. Several algorithms have been proposed specifically for the alignment of genomic DNA with cDNA (Gotoh, 1982; Huang, 1994; Huang *et al.*, 1997; Florea *et al.*, 1998). These alignments allow for long gaps in the cDNA that would typically correspond to introns in the genomic DNA. More recently, we introduced a novel algorithm based on a Hidden Markov Model that explicitly assigns exon or intron status to each residue in the genomic DNA (Usuka *et al.*, 2000). Simultaneous scoring for sequence similarity and intrinsic quality of the implied splice sites was shown to significantly improve the accuracy of the alignments for non-cognate cDNAs.

Another frequent situation in practice occurs when a predicted gene product is found to be similar to some protein in the existing databases, e.g., on the basis of a BLAST search (Altschul *et al.*,

Abbreviations used: BAC, bacterial artificial chromosome; HMM, Hidden Markov Model; NCBI, National Center of Biotechnology Information (USA); RT-PCR, reverse transcriptase polymerase chain reaction; TP, true positive; TN, true negative; FN, false negative; FP, false positive.

E-mail address of the corresponding author: vbrendel@iastate.edu

0022-2836/00/051075–11 $35.00/0

© 2000 Academic Press

1997). In that case, the gene prediction can often be confirmed or refined by a spliced alignment of the genomic DNA with the target protein (Gelfand *et al*., 1996). Several algorithms for such alignments have been proposed (States & Botstein, 1991; Posfai & Roberts, 1992; Xu *et al*., 1995; Birney *et al*., 1996; Gelfand *et al*., 1996; Guan & Uberbacher, 1996; Huang, 1996; Huang & Zhang, 1996; Rogozin *et al*., 1996; Zhang *et al*., 1997; Brown *et al*., 1998). Here, we extend our Hidden Markov Model for alignment of a genomic DNA with a cDNA or EST to accommodate matching of protein targets. We present the subroutine *sahmtP* (Spliced Alignment Hidden Markov Tool for Proteins) which implements a dynamic programming algorithm to efficiently calculate the optimal scoring alignment between a template DNA and a target protein sequence. The novelty in our approach compared to previous algorithms consists in the simultaneous assessment of the significance of the sequence alignment and the intrinsic quality of the implied splice sites.

We chose applications to sequence data from *Arabidopsis thaliana* illustration and evaluation of the algorithm for several reasons. First, there is a pressing need for automated genome annotation as the completion of the sequencing of the entire genome approaches (Meinke *et al*., 1998). Second, splice site prediction is well modeled specifically for this plant (Hebsgaard *et al*., 1996; Kleffe *et al*., 1996; Brendel & Kleffe, 1998). Third, the availability of an independently derived set of genes with known exon/intron structure allows unbiased performance comparisons for different algorithms. Extensions to gene identification in other organisms are assessed in the Discussion.

Procedures

Programs

We implemented our algorithm in a program named GeneSeqer as described below. Spliced alignments were also computed with the DPS, EXT, and NAP tools of the AAT package kindly provided by X. Huang (Huang *et al*., 1997). Briefly, DPS compares a DNA sequence to a set of target proteins, producing a file of high-scoring chains of segment pairs. This output file is then reformatted by EXT for use in the NAP program, which produces a spliced alignment. Because our target protein selection was made independent of the AAT default selection, Huang's program will be referred to as NAP, with the caveat that this designates the slightly modified NAP program of the AAT package rather than the previous stand-alone NAP program (Huang & Zhang, 1996). Code for the PROCRUSTES program (Gelfand *et al*., 1996) was not available; sample results using the PROCRUSTES Web service indicated poor performance relative to GeneSeqer and NAP. GENSCAN (Burge & Karlin, 1997) and GeneGenerator (Kleffe *et al*., 1998) were used to represent *ab initio* gene finding programs that do not rely on spliced alignment. Of these, GENSCAN has the capability of predicting multiple genes in either orientation for a given input sequence. For two genes in this study, GENSCAN correctly predicted a partial second gene within the input sequence. For the performance evaluations in this study, these predictions were interpreted as correctly identified non-coding regions with respect to the central gene.

Arabidopsis thaliana gene set

Program performance was evaluated on a set of 50 *A. thaliana* genes with known cDNAs derived from 27 genomic BACs. The set comprises the first 50 genes with unambiguous annotation from a list kindly provided by Larry Parnell, Cold Spring Harbor Laboratory. For the purposes of this study, the extent of each gene was defined as the BAC segment from 500 nt upstream of the translation start codon to 500 nt downstream of the translation stop codon, collinear with the RNA transcript. A complete description of the gene set with GenBank accessions is available in the Supplementary Material.

Target proteins

The target proteins for spliced alignment were acquired by querying GenBank *via* an automated, gapped BLASTP search (Altschul *et al*., 1997). For each gene, the query consisted of the cognate protein sequence translated from the coding region. The 20 protein sequences with the highest degree of similarity to the cognate protein were saved as initial target protein sets (including as the top scoring target the cognate protein itself). For some genes, these 20 targets included sequences with such marginal similarity to the query that the DPS program of AAT did not report any high-scoring chains of segment pairs. Those targets were eliminated. The final set of target proteins consisted of 905 sequences. Thirtyseven genes had a full set of 20 targets, and only three genes had less than ten targets.

For comparison, target protein were alternatively derived from BLASTP searches with the GENSCAN predicted peptides as queries. With one exception, at least three of the top five targets were identical whether they were drived from the cognate protein sequence or the GENSCAN prediction as query.

For each gene, all target protein sequences were aligned with the cognate protein sequence by the CLUSTALW program with default parameters (Thompson *et al*., 1994). The CLUSTALW reported alignment score was used as the measure of similarity to evaluate spliced alignment accuracy as a function of the similarity of the input protein target. The alignment score is percent sequence similarity relative to the shorter sequence, with adjustment for any gaps introduced in the alignment (Thompson *et al*., 1994).

Algorithm

We pose the problem of finding an optimal alignment of a genomic sequence $G_1, G_2, \ldots, G_N$ of length N with a target protein sequence $A_1, A_2, \ldots, A_M$ of length M. Optimality will be defined precisely below relative to a scoring system that simultaneously evaluates the pairwise sequence similarity of the translation product of the genomic sequence with the protein sequence and the quality of predicted splice sites in the genomic sequence. The genomic sequence consists of letters from the alphabet {A, C, G, T, N} where A, C, G, T denote the nucleotides adenine, cytosine, guanine, and thymine, respectively, and N denotes an undetermined nucleotide. The protein sequence is drawn from the 20-letter alphabet representing the naturally occurring amino acid residues in the standard one-letter-code. An alignment between the sequences may include gaps in either sequence, indicated by the additional gap symbol (-) juxtaposed to each of the letters comprising the corresponding insertion in the other sequence. Conceptually, an alignment may be viewed as an output of a Hidden Markov Model (HMM). The HMM defines a probability space consisting of all possible "threadings" of protein sequences of length M onto the given genomic sequence. The actual coding of the algorithm involves log probabilities that can be replaced by any additive weights without loss of generality.

The state sequence underlying a given alignment will be denoted as $Q = q_1 q_2 \cdots q_L$, where $\max\{3M, N\} \leqslant L \leqslant 3M + N$. The set of states of the HMM consists of "exon states" e and "intron states" i_0, i_1, and i_2. The three intron states represent introns in different coding frame phases: i_0 introns do not disrupt codons, i_1 introns split a codon between codon positions one and two, and i_2 introns split a codon between codon positions two and three. Transitions between the states are determined by the probabilities $P_{\mathrm{D(n)}}$ and $P_{\mathrm{A(n)}}$ that G_{n} in the genomic sequence is the first base (donor site) or last base (acceptor site) of an intron, respectively. In the applications discussed here, these values were set equal to the P-values calculated by the SplicePredictor program (Kleffe *et al*., 1996; Brendel & Kleffe, 1998). The output probabilities in the exon states may be set proportional to conventional amino acid substitution scores (see below). In addition, the algorithm allows for complete codon insertions and deletions as well as frameshift mutations due to single or double nucleotide insertions and deletions.

Optimal alignments are precisely defined as state sequences $Q = q_1 q_2 \cdots q_L$ with associated output $S_{\mathrm{M}}^{\mathrm{N}}$ (representing a sequence alignment of $G_1 G_2 \cdots G_N$ with $A_1 A_2 \cdots A_{\mathrm{M}}$) such that the joint probability $P(Q, S_{\mathrm{M}}^{\mathrm{N}})$ is maximal over all possible Q and $S_{\mathrm{M}}^{\mathrm{N}}$. This maximal probability is calculated in standard fashion as:

$$P = \max\{E_{\mathrm{M}}^{\mathrm{N}},\ (I_0)_{\mathrm{M}}^{\mathrm{N}},\ (I_1)_{\mathrm{M}}^{\mathrm{N}},\ (I_2)_{\mathrm{M}}^{\mathrm{N}}\},$$

where

$$E_{\mathrm{m}}^{\mathrm{n}} = \max P(Q = q_1 q_2 \cdots q_l, q_l = e,\ S_{\mathrm{m}}^{\mathrm{n}}),$$

and

$$(I_{\mathrm{x}})_{\mathrm{m}}^{\mathrm{n}} = \max P(Q = q_1 q_2 \cdots q_l, q_l = i_{\mathrm{x}},\ S_{\mathrm{m}}^{\mathrm{n}}),$$

for $x = 0,1,2$, $n = 1,2,\ldots,N$, $m = 1,2,\ldots,M$, $\max\{3m,\ n\} \leqslant l \leqslant 3m + n$, and maximization is over all possible Q and $S_{\mathrm{m}}^{\mathrm{n}}$ representing alignments of $G_1, G_2, \ldots, G_n$ with $A_1, A_2, \ldots, A_m$.

$E_{\mathrm{M}}^{\mathrm{N}}$ and $(I_{\mathrm{x}})_{\mathrm{M}}^{\mathrm{N}}$ are found from the following recursion:

$$
\begin{aligned}
E_0^{\mathrm{n}} &= (I_0)_0^{\mathrm{n}} = (I_1)_0^{\mathrm{n}} = (I_2)_0^{\mathrm{n}} = 1, && n = 0, 1, \ldots, N,\\
E_{\mathrm{m}}^{\mathrm{n}} &= 1, \quad (I_0)_{\mathrm{m}}^{\mathrm{n}} = (I_1)_{\mathrm{m}}^{\mathrm{n}} = (I_2)_{\mathrm{m}}^{\mathrm{n}} = 0, && n = 0, 1, 2, \quad m = 1, 2, \ldots, M,\\
E_{\mathrm{m}}^{\mathrm{n}} &= \max\Bigg\{ E_{m-1}^{n-3}\,(1 - P_{\mathrm{D(n-2)}})\,P\begin{pmatrix} G_{\mathrm{n-2}}G_{\mathrm{n-1}}G_{\mathrm{n}} \\ A_{\mathrm{m}} \end{pmatrix},\\
&\qquad E_{m-1}^{n-2}\,(1 - P_{\mathrm{D(n-1)}})\,P\begin{pmatrix} G_{\mathrm{n-1}}G_{n}- \\ A_{\mathrm{m}} \end{pmatrix},\\
&\qquad E_{\mathrm{m-1}}^{\mathrm{n-1}}\,(1 - P_{D(\mathrm{n})})\,P\begin{pmatrix} G_{\mathrm{n}}-- \\ A_{\mathrm{m}} \end{pmatrix},\\
&\qquad E_{\mathrm{m}}^{\mathrm{n-3}}\,(1 - P_{D(\mathrm{n-2})})\,P\begin{pmatrix} G_{\mathrm{n-2}}G_{\mathrm{n-1}}G_{\mathrm{n}} \\ - \end{pmatrix},\\
&\qquad E_{\mathrm{m}}^{\mathrm{n-2}}\,(1 - P_{\mathrm{D(n-1)}})\,P\begin{pmatrix} G_{\mathrm{n-1}}G_{\mathrm{n}}- \\ - \end{pmatrix},\\
&\qquad E_{\mathrm{m}}^{\mathrm{n-1}}\,(1 - P_{\mathrm{D(n)}})\,P\begin{pmatrix} G_{\mathrm{n}}-- \\ - \end{pmatrix},\\
&\qquad (I_0)_{\mathrm{m-1}}^{\mathrm{n-3}}\,P_{\mathrm{A(n-3)}}\,P\begin{pmatrix} G_{\mathrm{n-2}}G_{\mathrm{n-1}}G_{\mathrm{n}} \\ A_{\mathrm{m}} \end{pmatrix},\\
&\qquad (I_1)_{\mathrm{m-1}}^{\mathrm{n-2}}\,P_{\mathrm{A(n-2)}}\,P\begin{pmatrix} G_{\mathrm{x}}G_{\mathrm{n-1}}G_{n} \\ A_{\mathrm{m}} \end{pmatrix},\\
&\qquad (I_2)_{\mathrm{m-1}}^{\mathrm{n-1}}\,P_{\mathrm{A(n-1)}}\,P\begin{pmatrix} G_{\mathrm{x}}G_{x+1}G_{\mathrm{n}} \\ A_{\mathrm{m}} \end{pmatrix} \Bigg\},\\
(I_0)_{\mathrm{m}}^{\mathrm{n}} &= \max\left\{(I_0)_{\mathrm{m}}^{\mathrm{n-1}}\,(1 - P_{\mathrm{A(n-1)}}),\ E_{\mathrm{m}}^{\mathrm{n-1}}\,P_{\mathrm{D(n)}}\right\},\\
(I_1)_{\mathrm{m}}^{\mathrm{n}} &= \max\left\{(I_1)_{\mathrm{m}}^{\mathrm{n-1}}\,(1 - P_{\mathrm{A(n-1)}}),\ E_{\mathrm{m}}^{\mathrm{n-2}}\,P_{\mathrm{D(n)}}\right\},\\
(I_2)_{\mathrm{m}}^{\mathrm{n}} &= \max\left\{(I_2)_{\mathrm{m}}^{\mathrm{n-1}}\,(1 - P_{\mathrm{A(n-1)}}),\ E_{\mathrm{m}}^{\mathrm{n-3}}\,P_{\mathrm{D(n)}}\right\}, && n = 3, 4, \ldots, N,\\
& && m = 1, 2, \ldots, M.
\end{aligned}
$$

G_x and G_{x+1} denote upstream nucleotides of split codons that are recorded at the appropriate intron opening transitions. To facilitate the backtracing of the optimal alignment(s), the algorithm stores the state transition and output yielding the maximum score at each maximization step.

Implementation

Given the *P*-values according to the SplicePredictor program, few parameters need to be specified for a complete implementation of the algorithm. The output weights (logarithms of the output probabilities in the recursion equations above) are set by default to scaled values of the BLOSUM62 amino acid substitution scoring matrix (Henikoff & Henikoff, 1992). All deletions (one to three nucleotide deletions in the genomic DNA or an amino acid deletion in the protein sequence) are given the same deletion penalty corresponding to twice the lowest mismatch score in the BLOSUM62 matrix. The lack of a gap opening penalty may introduce a larger number of smaller gaps in weakly similar exons compared to conventional alignment methods with affine gap penalties. Naturally occurring introns exceed a minimal length of about 55 to 60 bases (for plants, see Brendel *et al*., 1998). To avoid solutions with unacceptably small intron assignments, implementation of the algorithm includes a ''short intron penalty'' as described by Usuka *et al*. (2000).

The algorithm was implemented as the C subroutine *sahmtP* in our previous SplicePredictor program (Brendel & Kleffe, 1998) as well as in the GeneSeqer program (Usuka *et al*., 2000). Limitations on the maximal lengths of the genomic DNA and protein sequence depend on the memory of the CPU. Our server is set to align genomic DNA segments up to 13 kb against a protein of up to 1600 residues. Runtime is proportional to the product of the sequence lengths. More space-efficient, but slower variants of dynamic programming alignment methods using a divide and conquer approach have been described by Hirschberg (1975), Myers & Miller (1988) and Huang & Miller (1991) but have not been implemented in our programs. Because the internal parts of introns play no role in the spliced alignment once the highest scoring exons and splice sites have been correctly identified, our algorithm can be modified to accommodate alignment of genes with long introns without increased space requirements.

Output of the program is illustrated in Figure 1. The optimal alignment of a tomato MADS-box protein to the *Arabidopsis* gene encoding the MADS-box protein AGL9 gives a prediction that correctly identifies five of eight exons. The right border of exon two is shifted upstream by six nucleotides to maximize sequence similarity to the target protein that lacks the extra two amino acids. At the C terminus, the algorithm suggests a frameshift in exon seven at position 73,319 to maximize similarity to the target protein that terminates after another 19 residues. Actually, exon seven continues in frame until 73,285, with another intron from 73,284 to 73,193 followed by the terminal exon of 91 nucleotides. The last intron has 5′ and 3′ splice sites scoring the high scores of 0.91 and 0.99, respectively. Thus, in this case, direct gene prediction from the genomic DNA without target protein information would have indicated the correct exon assignment based on splice site and open reading frame considerations. The spliced alignment is of interest in suggesting possible mutation events that may have led to the divergence of the two proteins at the C terminus.

Scoring the alignment

The program scores each predicted exon separately by tallying up the output weights corresponding to the alignment of the exon and protein sequence. This value is normalized by the equivalent sum of weights assuming perfect matching to the genomic DNA. From our experience, the optimal alignment of unrelated proteins to a genomic DNA produces typical exon quality values less than 0.10.

Performance statistics

Program performance was evaluated by the standard measures for prediction accuracy per nucleotide and per exon (Burset & Guigó, 1996; Huang *et al*., 1997; Burge & Karlin, 1997). To assess the contribution of the sophisticated splice site prediction scores in GeneSeqer compared to NAP, the exon measures were also applied to introns. At the nucleotide level, all nucleotides were classified as true positives (TP), true negatives (TN), false positives (FP), or false negatives (FN) depending on the agreement of predicted gene structure with actual gene structure. Thus, TP is the number of nucleotides correctly assigned exon status, TN is the number of nucleotides correctly assigned noncoding status, FP is the number of nucleotides incorrectly assigned exon status, and FN is the number of nucleotides incorrectly assigned noncoding status. Note, that in this context, exon status refers to the coding portion of exons only. Performance at the nucleotide level is summarized by the sensitivity $S_n = TP/(TP + FN)$, specificity $S_p = TP/(TP + FP)$, and the approximate correlation $AC = 0.5((TP/(TP + FN) + TP/(TP + FP) + TN/(TN + FP) + TN/(TN + FN)) - 1$. Following Burset & Guigó (1996), predicted exons and introns are considered correct only if they exactly match actual exons or introns, with correct splicing boundaries. Exon level performance is summarized by sensitivity (proportion of actual exons that is correctly predicted), specificity (proportion of predicted exons that is correctly predicted), missing exons (proportion of actual exons without overlap to predicted exons), and wrong exons (proportion of predicted exons without overlap to actual

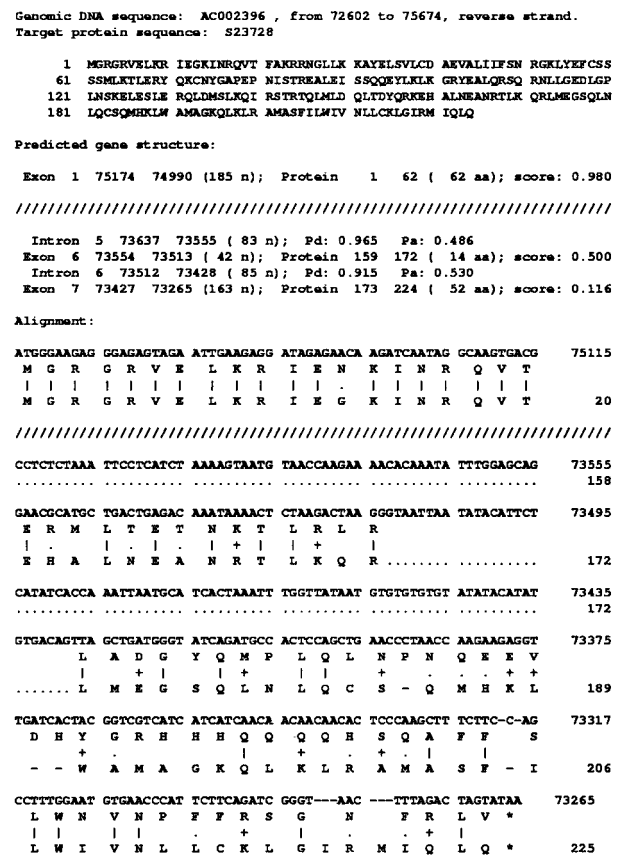

Figure 1. Sample output for spliced alignment with GeneSeqer. The genomic sequence input is the segment of the *Arabidopsis thaliana* chromosome I BAC F3I6 (GenBank Accession AC002396) encoding the MADS-box protein AGL9. The segment is colinear with the coding region and extends 500 nt upstream of the ATG translation start codon and 500 nucleotides downstream of the stop codon. The target protein sequence is the tomato TDR5 protein (GenBank Accession S23728) which shares about 67 % global sequence similarity with AGL9. The algorithm correctly identifies five of the eight exons of the AGL9 gene. The discrepancies are discussed in the text. The ///../// lines indicate parts of the output that were deleted for brevity. The predicted gene structure is described by the predicted exons (positions given relative to the input GenBank file) and matching target protein segments. The scores are normalized similarity scores for the sequence comparisons of implied exon translation and matching protein segment (1.00 would be perfect identity). For the introns, the donor (Pd) and acceptor site (Pa) scores are displayed. The alignment gives the genomic DNA sequence, its inferred protein translation (one-letter-code), and the matching parts of the target protein sequence. Identical residues are linked by (|), positively scoring substitutions by (+), and zero scoring substitutions by (.) according to the amino acid substitution scoring matrix used in the alignment, here BLOSUM62 (Henikoff & Henikoff, 1992). Gaps are represented by (-), introns by dots. Numbers at the right indicate the positions of the last nucleotide of the genomic DNA and of the last amino acid of the target protein displayed on the respective lines.

exons). Intron level performance is assessed similarly.

Prediction confidence

In practical applications, the alignment score of target protein to actual gene product is unknown. Target proteins would typically be selected on the basis of high local similarity to one or more predicted exons, and the spliced alignment would be used to confirm and extend the initial (partial) gene structure prediction. To assess prediction reliability in this situation, we evaluated program performance exon by exon based on the assigned confidence in the prediction. For GENSCAN, the confidence score was taken to be the probability value of the exons reported by the program. For NAP, the confidence score for an exon was taken to be the average of the 5′ and 3′ confidence values reported by the program. For GeneSeqer, the confidence score was set to the alignment score calculated for each exon. This choice does not reflect adjustment of the confidence depending on splice site scores, which would seem difficult to capture in a single value.

Results

Gene identification by sequence inspection necessarily relies on prior knowledge of typical gene structure that gets incorporated into a program in some formal way. The *ab initio* methods rely on training of model parameters on sets of established genes. At minimum, this typically involves Markov models for exon and intron sequences. GENSCAN (Burge & Karlin, 1997), currently the most successful method, is based on a comprehensive probabilistic model including profiles for transcriptional, translational, and splice signals. By contrast, previous spliced alignment methods have not involved sequence models but predict genes entirely on the basis of similarity to other genes or gene products. In either approach, success of the method depends on the accuracy and applicability of prior knowledge. *Ab initio* methods are biased to succeed for genes similar to those in the training set of the method, while spliced alignment depends on the suitability of the chosen target sequence. The GeneSeqer algorithm represents a hybrid approach, combining evaluation of sequence similarity with evaluation for predicted splice site strength. We examined performance of the algorithm relative to *ab initio* and similarity-based spliced alignment methods.

Nucleotide level performance

GENSCAN and GeneGenerator perform similarly in terms of nucleotide sensitivity (0.90-0.91) and specificity (0.95-0.96; Table 1). Both NAP and GeneSeqer perform better when the target proteins match the gene product with alignment scores of approximately 60 and better. The poor results with

low similarity targets are expected because the spliced alignment algorithms optimize matching to any supplied target, independent of its relatedness to the actual gene product. It is noteworthy, however, that GeneSeqer outperforms NAP for low similarity targets.

For comparison, we also ran GeneSeqer with generic rather than *Arabidopsis* specific splice site scores. In this case, the performance statistics of GeneSeqer were only slightly better than those of NAP (Table 1). Thus, most of the improvement with our algorithm seems to derive from the inclusion of differential scoring for predicted splice site strength.

Exon and intron level performance

The goal of gene prediction tools is to correctly identify the translation start and stop signals and the precise borders of exons and introns. Therefore, a more thorough test of program performance is to assess the proportion of precisely identified actual exons and introns as well as the error rates for predicted exons and introns. GENSCAN achieved exon and intron sensitivity of only about 0.75 (Table 1). Thus, 25 % of the actual exons in the data set were not correctly identified. GeneGenerator performed even worse, although it erred less in terms of missed exons. NAP achieved better exon sensitivity only with targets with CLUSTALW similarity scores $\geqslant 80$ and better intron sensitivity with targets with score $\geqslant 60$. GeneSeqer produced better results at lower similarity score thresholds (60 and 30, respectively). Averages of sensitivity and specificity are displayed in Figures 2 and 3. GeneSeqer again persistently outperformed NAP. Remarkably, for intron prediction, GeneSeqer outperformed GENSCAN even for spliced alignments in the $\geqslant 30$ similarity score range.

Availability of close target proteins

As shown above, the performance of the spliced alignment methods depends dramatically on the availability of close target sequences. Table 2 shows that the availability of close targets does not appear to be a practical limitation, at least for the gene set examined in this comparison. For as many as 60 % of the genes there was at least one (non-cognate) target in the databases that matched the cognate gene product with perfect similarity score, and 90 % of the genes had at least one very close homolog (score $\geqslant 98$). For all genes in the study, non-cognate targets were available at score levels for which GeneSeqer outperforms the *ab initio* programs. A caveat is that this gene set may be biased for genes of general interest, as reflected in the availability of experimentally verified annotation.

Prediction confidence

The performance evaluation statistics for the spliced alignment methods as a function of the target sequence similarity can only be used indirectly in novel gene finding applications. Because the gene product is unknown, similarity comparisons must necessarily be between predicted gene products and the target sequences. Nonetheless, this approach can be useful to provide a useful check on the gene prediction. If the genomic DNA input can be shown to encode a gene product with alignment score at least about 60, then the GeneSeqer spliced alignment would typically be quite reliable. In the case that several target sequences are available, presumably the highest scoring pair of predicted gene and target would be the correct assignment.

A complementary approach is afforded by confidence values assigned to each exon prediction by the programs we compared. Also, in this way, certain exons can be identified as highly likely correct, whereas others may be identified as highly tentative. Table 3 shows that more than 90 % of the nucleotides predicted to be exon are correctly predicted by GeneSeqer for exons of confidence score $\geqslant 0.20$. At confidence scores $\geqslant 0.70$, all three programs attain very low error rates. Figure 4 shows approximate linearity of confidence values with exon specificity. The GENSCAN data are more variable, possibly at least in part because of the much smaller data set (one prediction per gene compared to predictions for all targets for the other programs).

Issues of automated genome annotation

To test the benefits and limitations of GeneSeqer in a practical genome annotation task, we picked the first *Arabidopsis* BAC deposited in GenBank this year (accession AC006932; 89,479 nt) for re-annotation. The GenBank annotation gives 23 coding sequences, whereas GENSCAN predicts 19

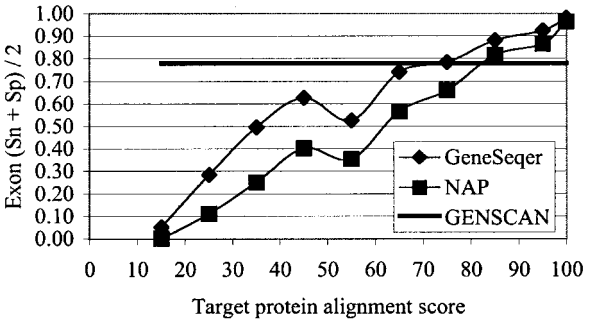

Figure 2. Performance comparison at the exon level. For GeneSeqer and NAP, the performance values were averaged for all spliced alignments with target proteins matching the cognate gene product at the indicated alignment score. The alignment scores were rounded such that score 15 represents all scores in the 10 to 19 range, score 25 represents all scores in the 20 to 30 range, etc. The performance of the *ab initio* gene finding program GENSCAN is represented by a straight line because this program does not incorporate protein alignment scores.

Table 1. Performance evaluation of gene prediction programs

			Nucleotide			Exon				Intron			
	Sim	No.	Sn	Sp	AC	Sn	Sp	ME	WE	Sn	Sp	MI	WI
GENSCAN	n/a	50	0.90	0.96	0.89	0.74	0.81	0.10	0.02	0.73	0.80	0.01	0.01
GeneGenerator	n/a	50	0.91	0.95	0.89	0.64	0.57	0.05	0.17	0.72	0.61	0.01	0.03
NAP			0.41	0.92	0.39	0.00	0.00	0.34	0.22	0.33	0.20	0.08	0.41
GS-generic	**15**	14	0.52	0.88	0.44	0.01	0.01	0.18	0.23	0.32	0.13	0.02	0.40
GeneSeqer			**0.60**	**0.93**	**0.54**	**0.06**	**0.04**	**0.19**	**0.17**	**0.38**	**0.19**	**0.00**	**0.43**
NAP			0.67	0.89	0.65	0.10	0.12	0.22	0.17	0.26	0.29	0.12	0.14
GS-generic	**25**	77	0.80	0.88	0.72	0.19	0.16	0.07	0.22	0.45	0.32	0.13	0.12
GeneSeqer			**0.86**	**0.92**	**0.80**	**0.29**	**0.28**	**0.10**	**0.14**	**0.59**	**0.55**	**0.11**	**0.09**
NAP			0.81	0.91	0.78	0.24	0.26	0.15	0.07	0.51	0.54	0.04	0.06
GS-generic	**35**	87	0.83	0.91	0.80	0.40	0.37	0.10	0.13	0.68	0.61	0.01	0.07
GeneSeqer			**0.86**	**0.94**	**0.85**	**0.50**	**0.49**	**0.06**	**0.07**	**0.78**	**0.73**	**0.01**	**0.04**
NAP			0.86	0.93	0.83	0.37	0.43	0.14	0.02	0.60	0.69	0.05	0.04
GS-generic	**45**	136	0.89	0.93	0.85	0.50	0.51	0.08	0.06	0.71	0.72	0.04	0.04
GeneSeqer			**0.90**	**0.95**	**0.88**	**0.62**	**0.63**	**0.05**	**0.04**	**0.82**	**0.84**	**0.03**	**0.03**
NAP			0.87	0.97	0.87	0.34	0.37	0.08	0.01	0.69	0.76	0.00	0.02
GS-generic	**55**	84	0.88	0.97	0.87	0.48	0.46	0.05	0.08	0.79	0.75	0.01	0.04
GeneSeqer			**0.88**	**0.97**	**0.88**	**0.52**	**0.53**	**0.05**	**0.05**	**0.84**	**0.82**	**0.00**	**0.03**
NAP			0.93	0.96	0.91	0.56	0.57	0.04	0.01	0.81	0.83	0.01	0.02
GS-generic	**65**	100	0.96	0.97	0.94	0.71	0.70	0.02	0.04	0.87	0.85	0.00	0.00
GeneSeqer			**0.96**	**0.98**	**0.95**	**0.74**	**0.74**	**0.02**	**0.02**	**0.90**	**0.90**	**0.00**	**0.00**
NAP			0.95	0.99	0.96	0.65	0.67	0.04	0.00	0.84	0.88	0.00	0.01
GS-generic	**75**	94	0.95	0.99	0.95	0.74	0.76	0.04	0.03	0.91	0.88	0.00	0.03
GeneSeqer			**0.96**	**0.99**	**0.96**	**0.77**	**0.79**	**0.03**	**0.02**	**0.95**	**0.95**	**0.00**	**0.00**
NAP			0.95	0.99	0.95	0.81	0.82	0.03	0.00	0.89	0.92	0.00	0.00
GS-generic	**85**	82	0.95	0.99	0.95	0.85	0.86	0.03	0.01	0.91	0.93	0.00	0.00
GeneSeqer			**0.95**	**0.99**	**0.95**	**0.88**	**0.89**	**0.03**	**0.01**	**0.96**	**0.97**	**0.00**	**0.00**
NAP			0.99	0.99	0.98	0.86	0.87	0.03	0.01	0.90	0.93	0.01	0.00
GS-generic	**95**	112	0.99	0.99	0.98	0.87	0.88	0.03	0.01	0.91	0.93	0.01	0.00
GeneSeqer			**0.99**	**0.99**	**0.99**	**0.92**	**0.93**	**0.01**	**0.01**	**0.96**	**0.97**	**0.01**	**0.00**
NAP			1.00	1.00	1.00	0.96	0.97	0.01	0.00	0.97	0.99	0.00	0.00
GS-generic	**100**	119	1.00	1.00	1.00	0.97	0.97	0.01	0.00	0.98	0.99	0.00	0.00
GeneSeqer			**1.00**	**1.00**	**1.00**	**0.98**	**0.98**	**0.00**	**0.00**	**0.99**	**0.99**	**0.00**	**0.00**

Performance is evaluated in terms of nucleotide sensitivity (Sn), specificity (Sp), and approximate correlation (AC), exon sensitivity and specificity, missing (ME) and wrong exons (WE), intron sensitivity and specificity, and missing (MI) and wrong (WI) introns. For GeneSeqer and NAP, performance values are averaged over all spliced alignments with target proteins of similarity score indicated in column two. Similarity scores were rounded such that 15 represents the range 10 to 19, 25 represents the range 20 to 29, etc. The number of data points for nucleotide and exon evaluation in each range are given in column three; because of the inclusion of intron-less genes and gene predictions, the numbers are slightly different for intron evaluation (not shown). Comparing the bold faced entries shows the improvement of GeneSeqer performance with increased similarity of the target protein with the cognate gene product. The values in the row GS-generic were obtained with GeneSeqer using generic rather than *Arabidopsis* specific splice site scores.

genes. The GENSCAN predicted peptides were used as queries of a BLAST search against the current NCBI non-redundant protein database. For each query, the six top scoring database entries at threshold *E*-value 10^{-10} were used for spliced alignment against the genomic DNA in the region of the corresponding GENSCAN prediction. All but three of the GENSCAN predicted peptides gave a full six targets at this stringency. Figures 5 and 6 show typical results for the spliced alignment.

In the first example (Figure 5), GENSCAN predicts a single gene spanning the GenBank annotated T27G7.3 and T27G7.2 genes. The BLAST and GeneSeqer outputs reveal strong similarity to the basic leucine zipper protein (blzp) of maize (GenBank accession AAC39351) and its homolog bZIP in tobacco (GenBank accession AAF06696). Inspection of the spliced alignments indicated frameshifts in the N-terminal part of the blzp alignment. Strong conservation in the predicted exons from the bZIP alignment suggests the displayed predicted gene structure gs_PGS. While there should be no doubt that this region of the *Arabidopsis* genome encodes a member of the basic leucine zipper family of transcription factors, the correct assignment of exons and introns remains tentative in the absence of EST evidence. The spliced alignment alone is not necessarily conclusive because a given target may represent a paralog rather than an ortholog of the genomic DNA locus. The frame-

Table 2. Similarity scores of most closely related target proteins

	Target				
% Genes	1	2	3	4	5
20	100	100	99	96	94
40	100	100	94	87	82
60	100	98	82	71	69
80	99	86	57	42	42
90	98	61	38	35	31

Target proteins were ordered by closest similarity to the cognate gene product. For example, 20% of the genes in the set had at least five targets with similarity score $\geqslant 94$, and 90% of the genes had at least one target with score $\geqslant 98$.

shifts in the blzp alignment give reason to be more confident in the bZIP-based prediction gs_PGS.

In general, one would also have to investigate the validity of the selected targets, which in turn may be based on prediction from sequence only. In this context, one of the BLAST derived targets of the Figure 5 GENSCAN prediction was a predicted peptide from a different *Arabidopsis* BAC (GenBank accession AC011438) that matched the AC006932 genomic DNA perfectly, but differed from the AC006932 GenBank annotation. As it were, AC011438 and AC006932 are BACs that overlap in the first 14,386 nt of AC006932, were deposited in GenBank by the same authors within a period of three weeks, but were annotated differently in the same region.

Figure 6 gives a second example of genome annotation by spliced alignment. Again, the GENSCAN prediction and GenBank annotation overlap, but differ. Spliced alignment gives yet a third prediction based on strong similarity to the bacterial *nifS* gene. Intriguingly, putative nifS homologs have been reported for human (Land & Rouault, 1998) and other eukaryotes, but the predicted *Arabidopsis* nifS gene product is much more similar to the bacterial proteins than to the putative eukaryotic homologs. The role of nifS in plants remains to be explored.

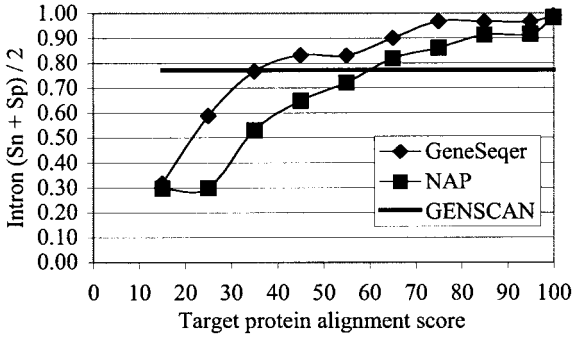

Figure 3. Performance comparison at the intron level. Performance values were averaged over all spliced alignments with similar alignment score of target protein and cognate gene product as described in the legend to Figure 2.

Discussion

With unabated advances in DNA sequencing technology and increased allocation of resources to sequencing efforts, public and private sequence repositories continue to grow greatly and rapidly. Exploring this wealth of information at commensurate pace requires largely automated sequence annotation. The primary task is to identify the protein coding genes and their translation products. For all higher eukaryotic organisms this involves identification of translation start and stop signals as well as the splice sites that define the extent of coding exons. The precise molecular mechanisms for transcription and pre-mRNA processing are not understood well enough to model these processes accurately. Thus, gene identification by sequence inspection currently relies on statistical methods that evaluate potential gene structures in a given sequence on the basis of prior knowledge of sequence features of known genes. *Ab initio* methods identify the most likely gene structures by evaluating the probability of potential exon and intron assignments using sequence models derived

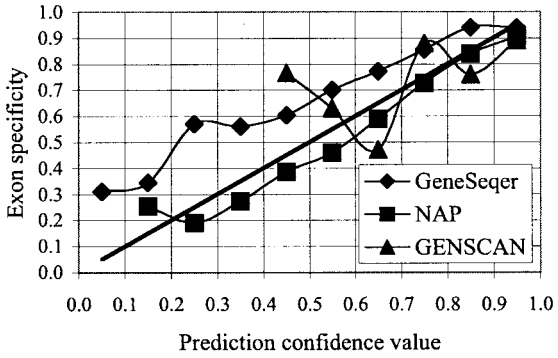

Figure 4. Accuracy of confidence assessments. For each program, predicted exons were assigned a confidence value as described in the text. The plot shows the proportion of correctly predicted exons for each set of predicted exons at the indicated level of confidence. Confidence scores were rounded such that score 0.05 represents all scores in the 0.00 to 0.09 range, score 0.15 represents all scores in the 0.10 to 0.19 range, etc.. Averages over less than 15 scores were omitted. Perfect correlation of probability of correct prediction with confidence score (diagonal line).

Table 3. Evaluation of prediction confidence

Prediction confidence	Nucleotide specificity		
	GeneSeqer	NAP	GENSCAN
0.05	0.76		
0.15	0.86	9.76	
0.25	0.90	0.74	
0.35	0.94	0.88	
0.45	0.96	0.92	1.00
0.55	0.96	0.94	0.94
0.65	0.98	0.95	0.87
0.75	0.99	0.98	1.00
0.85	0.98	0.98	0.91
0.95	0.99	0.99	0.95

Nucleotide specificity is the proportion of nucleotides predicted to be exon that are correctly predicted. Values were averaged over all predicted exons with the indicated confidence value. Averages over less than 15 data points were omitted. Confidence values for the different programs were calculated as described in the text and rounded such that 0.05 represents the range 0.00 to 0.09, 0.15 represents the range 0.10 to 0.19, etc.

from the analysis of a training set of known genes. Spliced alignment methods predict gene structure by maximizing sequence similarity of potential translation products to known ''target'' proteins.

We have compared the performance of the *ab initio* methods GENSCAN and GeneGenerator with that of the spliced alignment methods NAP and our novel GeneSeqer. On average, the spliced alignment methods give more accurate predictions whenever the target protein is sufficiently similar to the encoded gene product. The situation is reversed when the target protein is a poor match. In this case, the spliced alignment methods fail to identify some exons by maximizing similarity to the target, whereas the *ab initio* methods recognize these exons based on intrinsic sequence features. Thus, spliced alignment will be successful in practice only if, (1) the pool of target proteins is large enough that close targets to the given genomic DNA sequence exist, (2) such targets can be identified in the absence of prior knowledge of the encoded gene product, and (3) the spliced alignment method is sufficiently robust to sequence variation such that distant homologs can still be successfully used as targets. The first requirement is well met by the continuing sequencing and annotation efforts. As more and more of the natural protein repertoire is becoming known, the chances of finding new genes with no homologs in the repositories will become increasingly remote. For the *A. thaliana* gene set we analyzed, very close homologs were available for more than 90 % of the genes (Table 2).

The second requirement, identification of suitable targets, should also not be a practical limitation. A typical strategy would be to use an initial gene prediction by any method to search the protein databases for at least partially matching entries. The gene prediction can then be confirmed or refined by spliced alignment with those targets.

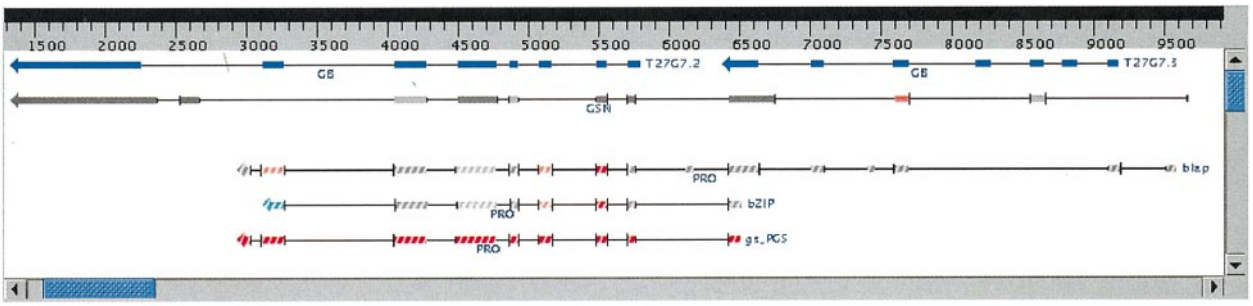

Figure 5. Gene prediction by spliced alignment. The Figure gives a screenshot of our GeneSeqer graphical user interface (Zhu & Brendel, unpublished results) for the region 1000 to 10,000 of GenBank accession AC006932 *Arabidopsis* BAC. Exons are indicated by colored boxes, introns by lines, predicted splice sites by vertical bars. The first line represents the GenBank annotation (genes T27G7.2 and T27G7.3, both encoded on the complementary strand). The second line gives the GENSCAN prediction. The following two lines are the schematic spliced alignments with two target proteins, and the last line displays our predicted gene structure (gs_PGS) based on evaluation of the spliced alignment quality (see the text for details). The different colors represent ranges of prediction confidence; very high, (red); moderate, (green); weak, (gray). Gs_PGS exon assignments differ from GenBank T27G7.2 as follows: novel exon 1 from 6514 to 6437, novel intron 1 from 6436 to 5758 (splice site scores 0.83 and 0.51), novel final exon from 3034 (3′ splice site score: 0.60) to 2947.

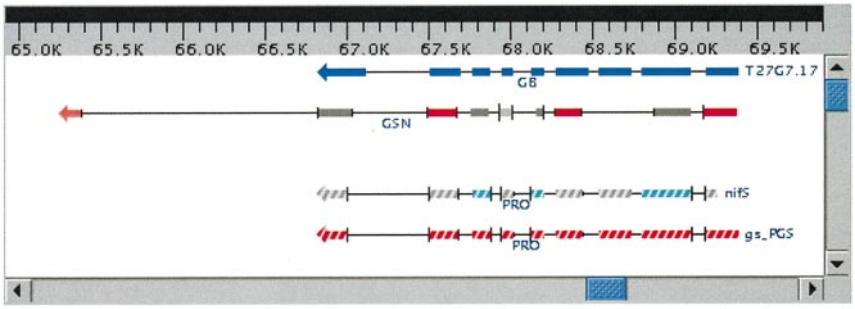

Figure 6. Gene prediction by spliced alignment. A nifS homolog is predicted in the region 65,000 to 70,000 of the GenBank accession AC006932 *Arabidopis* BAC. Symbols are as described in the legend to Figure 5. Gs_PGS exon assignments differ from GenBank T27G7.17 as follows: exon 4, 68,434 to 68,289 instead of 68,470 to 68,280, exon 8, 67,671 to 67,513 instead of 67,695 to 67,513, and exon 9,66,990 to 66,820 instead of 67,110 to 66,820.

If a target gives a strong global match to a predicted gene product, this would most likely correspond to a correct gene prediction of a homolog of that target protein. Excluding the cognate protein as a target, for our dataset of 50 genes the best targets gave 41 entirely correct predictions with GeneSeqer (38 with NAP), compared to only 13 for GENSCAN and three for GeneGenerator. In case of partial matching, confidence assignments to the predicted exons can suggest the extent of reliable predictions (Huang *et al.*, 1997; Burge & Karlin, 1997; Figure 4).

For the third requirement, we have shown that the novel GeneSeqer algorithm significantly improves prediction accuracy with more distant homologs by incorporating intrinsic splice site strength evaluation into the spliced alignment (Figures 2 and 3). Moreover, such spliced alignment may also suggest mutation events that could explain protein divergence (Figure 1).

For now, GeneSeqer has only been applied to gene identification in plant genomic DNA. Extensions to other organisms should be straightforward and involve mostly incorporation of species-specific splice site prediction methods. Currently, the algorithm as described cannot predict entire gene structures for genes including very long introns (although, in this case, the output would likely show two or more distinct partial alignments corresponding to the different exons separated by long introns). This limitation can be overcome by more sophisticated pre-processing (unpublished results). Comparison of Figures 2 and 3 clearly shows a particular problem for spliced alignment to correctly predict the initial and terminal exons. This results from the fact that most sequence variation of homologous proteins occurs at the N and C termini, so that lower scoring targets would most often correspond to sequences that have diverged in those parts. It may be possible to improve prediction accuracy in these cases by incorporating specific scores for transcription and translation start and termination signals as done in GENSCAN (Burge & Karlin, 1997).

At present, spliced alignment may be the most powerful tool for genome annotation, but by itself it is still insufficient for accurate automated annotation. Figures 5 and 6 provide typical examples for which alternative predictions have to be resolved by expert human interpretation. Further improvements towards automated annotation may have to involve simultaneous spliced alignment and phylogenetic reconstruction of target protein relationships that distinguishes true orthologs among several matching targets.

Program Availability

The algorithm is available as a C subroutine and is implemented in the SplicePredictor and GeneSeqer programs. The source code is available *via* anonymous ftp from ftp.zmdb.iastate.edu. SplicePredictor and GeneSeqer are also implemented as Web services at http://gremlin1.zool.iastate.edu/cgi-bin/sp.cgi and http://gremlin1.zool.iastate.edu/cgi-bin/gs.cgi, respectively.

Acknowledgments

V.B. was supported in part by NSF grant no. 9872657. J.U. was supported in part by NSF grant no. 9734893. The authors thank H.C. Andersen, D. Brutlag, and W. Zhu for critical reading of the manuscript and W. Zhu for help with Figures 5 and 6.

References

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

Birney, E., Thompson, J. D. & Gibson, T. J. (1996). PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucl. Acids Res.* **24**, 2730-2739.

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

Brendel, V., Carle-Urioste, J. C. & Walbot, V. (1998). Intron recognition in plants. In *A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants* (Bailey-Serres, J. & Gallie, D. R., eds), pp. 20-28, American Society of Plant Physiology, Rockville, MD.

Brown, N. P., Sander, C. & Bork, P. (1998). Frame: detection of genomic sequencing errors. *Bioinformatics,* **14**, 367-371.

Burge, C. & Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268**, 78-94.

Burge, C. & Karlin, S. (1998). Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**, 346-354.

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

Claverie, J.-M. (1997). Computational methods for the identification of genes in vertebrate genomic sequences. *Hum. Mol. Genet.* **6**, 1735-1744.

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

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

Gotoh, O. (1982). An improved algorithm for matching biological sequences. *J. Mol. Biol.* **162**, 705-708.

Guan, X. & Uberbacher, E. C. (1996). Alignments of DNA and protein sequences containing frameshift errors. *Comput. Appl. Biosci.* **12**, 31-40.

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

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915-10919.

Hirschberg, D. S. (1975). A linear space algorithm for computing maximal common subsequences. *Commun. Assoc. Comput. Mach.* **18**, 341-343.

Huang, X. (1994). On global sequence alignment. *Comput. Appl. Biosci.* **10**, 227-235.

Huang, X. (1996). Fast comparison of a DNA sequence with a protein sequence database. *Microbial Compar. Genomics,* **1**, 281-291.

Huang, X. & Miller, W. (1991). A time-efficient, linear-space local similarity algorithm. *Advan. Appl. Math.* **12**, 337-357.

Huang, X. & Zhang, J. (1996). Methods for comparing a DNA sequence with a protein sequence. *Comput. Appl. Biosci.* **12**, 497-506.

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

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

Kleffe, J., Hermann, K., Vahrson, W., Wittig, B. & Brendel, V. (1998). GeneGenerator: a flexible algorithm for gene prediction and its application to maize sequences. *Bioinformatics,* **14**, 232-243.

Land, T. & Rouault, T. A. (1998). Targeting of a human iron-sulfur cluster assembly enzyme, nifs, to different subcellular compartments is regulated through alternative AUG utilization. *Mol. Cell,* **2**, 807-815.

Meinke, D. W., Cherry, J. M., Dean, C., Rounsley, S. D. & Koornneef, M. (1998). *Arabidopsis thaliana*: a model plant for genome analysis. *Science,* **282**, 662-682.

Myers, E. W. & Miller, W. (1988). Optimal alignments in linear space. *Comput. Appl. Biosci.* **4**, 11-17.

Posfai, J. & Roberts, R. J. (1992). Finding errors in DNA sequences. *Proc. Natl Acad. Sci. USA,* **89**, 4698-4702.

Rogozin, I. B., Milanesi, L. & Kolchanov, N. A. (1996). Gene structure prediction using information on homologous protein sequence. *Comput. Appl. Biosci.* **12**, 161-170.

States, D. J. & Botstein, D. (1991). Molecular sequence accuracy and the analysis of protein coding regions. *Proc. Natl Acad. Sci. USA,* **88**, 5518-5522.

Sze, S.-H., Roytberg, M. A., Gelfand, M. S., Mironov, A. A., Astakhova, T. V. & Pevzner, P. A. (1998). Algorithms and software for support of gene identification experiments. *Bioinformatics,* **14**, 14-19.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673-4680.

Usuka, J., Zhu, W. & Brendel, V. (2000). Optimal spliced alignment of homologous cDNA to a genomic DNA template. *Bioinformatics,* **16**, In the press.

Xu, Y., Mural, R. J. & Uberbacher, E. C. (1995). Correcting sequencing errors in DNA coding regions using a dynamic programming approach. *Comput. Appl. Biosci.* **11**, 117-124.

Zhang, Z., Pearson, W. & Miller, W. (1997). Aligning a DNA sequence with a protein sequence. *J. Comp. Biol.* **4**, 339-349.

Edited by G. von Heijne

(Received 22 November 1999; received in revised form 25 February 2000; accepted 25 February 2000)

http://www.academicpress.com/jmb

Supplementary material comprising 111 Figures is available from JMB Online.