BIOINFORMATICS

Vol. 16 no. 3 2000
Pages 203–211

Optimal spliced alignment of homologous cDNA to a genomic DNA template

Jonathan Usuka[1], Wei Zhu[2] and Volker Brendel[2,*]

[1]Department of Chemistry, Stanford University, Stanford, CA 94305, USA and [2]Department of Zoology and Genetics, Iowa State University, 2112 Molecular Biology Building, Ames, IA 50011-3260, USA

Received on March 28, 1999; revised on July 30, 1999; accepted on August 28, 1999

Abstract

Motivation: *Supplementary cDNA or EST evidence is often decisive for discriminating between alternative gene predictions derived from computational sequence inspection by any of a number of requisite programs. Without additional experimental effort, this approach must rely on the occurrence of cognate ESTs for the gene under consideration in available, generally incomplete, EST collections for the given species. In some cases, particular exon assignments can be supported by sequence matching even if the cDNA or EST is produced from non-cognate genomic DNA, including different loci of a gene family or homologous loci from different species. However, marginally significant sequence matching alone can also be misleading. We sought to develop an algorithm that would simultaneously score for predicted intrinsic splice site strength and sequence matching between the genomic DNA template and a related cDNA or EST. In this case, weakly predicted splice sites may be chosen for the optimal scoring spliced alignment on the basis of surrounding sequence matching. Strongly predicted splice sites will enter the optimal spliced alignment even without strong sequence matching.*

Results: *We designed a novel algorithm that produces the optimal spliced alignment of a genomic DNA with a cDNA or EST based on scoring for both sequence matching and intrinsic splice site strength. By example, we demonstrate that this combined approach appears to improve gene prediction accuracy compared with current methods that rely only on either search by content and signal or on sequence similarity.*

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. Both programs are also implemented as a Web service at http://gremlin1.zool.iastate.edu/cgi-bin/sp.cgi and http://gremlin1.zool.iastate.edu/cgi-bin/gs.cgi, respectively.*

Contact: *vbrendel@iastate.edu*

Introduction

Global sequencing efforts are currently producing vast amounts of raw genomic sequence data for many different organisms. The pace of sequencing necessitates that the sequence annotation, in particular with respect to gene structure, be largely based on computational algorithms for automated sequence interpretation [for a recent review see Claverie (1997)]. Experimental evidence for exon assignments may derive from cDNA or EST sequencing. 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 simplest, unambiguous case, the alignment will consist of (1) perfectly matching segments corresponding to the exons, and (2) deletions in the cDNA corresponding to introns in the genomic template. In practice, matching may be less than perfect due to either sequencing errors or, more importantly, due to matching of genomic sequences with non-cognate cDNA. The non-cognate cDNAs derive not from the given genetic locus but from homologous loci, for example, the corresponding locus in a related species or a duplicated locus representing a different member of the same gene family. In this case, the alignment will generally have to include mismatches and gaps, but may still strongly support a particular gene structure prediction at the locus being analyzed.

We present the subroutine sahmtD (Spliced Alignment Hidden Markov Tool for cDNA) which implements a dynamic programming algorithm to efficiently calculate the optimal scoring alignment between an assumed template DNA and a second sequence representing a related collinear spliced product. The novelty in our approach compared to previous algorithms (Gotoh, 1982; Florea *et al.*, 1998; Huang, 1994; Huang *et al.*, 1997; Mott, 1997) consists (1) in the simultaneous assessment of the significance of the sequence alignment and the intrinsic

[*]To whom correspondence should be addressed.

quality of the implied splice sites, and (2) in the explicit assignment of exon or intron status to each nucleotide in the genomic DNA. The algorithm is considerably more reliable in cases where global sequence similarity is weak or compromised by regions of poor local similarity. Applications are illustrated in the context of resolution of multiple hits in cDNA database searches with genomic sequence queries and the study of a hypothetical novel *Arabidopsis thaliana* gene family.

System and methods

We pose the problem of finding an optimal alignment of a genomic nucleotide sequence $G_1, G_2, \ldots, G_N$ of length N with a cDNA or EST nucleotide sequence $C_1, C_2, \ldots, C_M$ of length M. A precise definition will be given later of optimality relative to a scoring system that simultaneously evaluates the pairwise sequence similarity and the quality of predicted splice sites in the genomic sequence. Both sequences consist of letters from the alphabet $\mathbf{A} = \{\mathrm{A, C, G, T, N}\}$ where A, C, G, T denote the nucleotides adenine, cytosine, guanine, and thymine, respectively, and N denotes an undetermined nucleotide. 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. We use the notation $\mathbf{A}^+$ for the alphabet superset $\{\mathrm{A, C, G, T, N, \_}\}$ and $\mathbf{A}^-$ for the subset $\{\mathrm{A, C, G, T}\}$. All possible alignments may be viewed as outputs of a Hidden Markov Model (HMM). The HMM defines a probability space consisting of all possible "threadings" of cDNA sequences of length M over the alphabet $\mathbf{A}$ into the given genomic sequence. The formulation of the algorithm in terms of a HMM is merely for convenience of presentation. The coding of the algorithm involves log probabilities that are in practice replaced by any suitable additive weights without loss of generality.

The state sequence underlying a given alignment will be denoted as $Q = q_1 q_2 \ldots q_L$, where $\max\{M, N\} \leq L \leq M + N$. The set of states of the HMM consists of the exon states e_n, $n = 1, 2, \ldots, N$, with output $\begin{matrix} X \\ Y \end{matrix}$, $X, Y \in \mathbf{A}^+$, and the intron states i_n, $n = 1, 2, \ldots, N$, with output $\begin{matrix} G_n \\ \star \end{matrix}$, where the $\star$ symbol serves as a placeholder for the spliced sequence parts. Transitions between the states are limited to the following transitions with non-zero transition probabilities $\tau_{q_l, q_{l+1}}$ (Figure 1):

$$\begin{array}{llll}
\tau_{e_n, e_{n+1}} = (1 - P_{\Delta G})(1 - P_{D(n+1)}) & & & \\
\tau_{i_n, e_{n+1}} = P_{A(n)}(1 - P_{\Delta G}) & & & \\
\tau_{e_n, e_n} = P_{\Delta G} & & \tau_{i_n, e_n} = P_{A(n)} P_{\Delta G} & \\
\tau_{e_n, i_{n+1}} = (1 - P_{\Delta G}) P_{D(n+1)} & & \tau_{i_n, i_{n+1}} = 1 - P_{A(n)} &
\end{array}$$

for $n = 1, 2, \ldots, N$ (third line) or $N - 1$ (other lines). Here $P_{D(n)}$ and $P_{A(n)}$ are the pre-determined probabilities 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 for plant gene identification discussed here, these values are set equal to the P-values calculated by the SplicePredictor program (Brendel and Kleffe, 1998; Kleffe *et al.*, 1996). Sites that are not scored by SplicePredictor are given small positive probabilities so that non-consensus sites supported by surrounding exon sequence matching are not excluded *a priori*. Other assignments could be made, for example derived from NetPlantGene output (Hebsgaard *et al.*, 1996) or (in the absence of models appropriate for the given species) generic assignments (distinguishing only between GT, GC, and other potential donor sites, and between AG and other potential acceptor sites). $P_{\Delta G}$ is a parameter that denotes the probability of inserting a gap symbol in the genomic sequence.

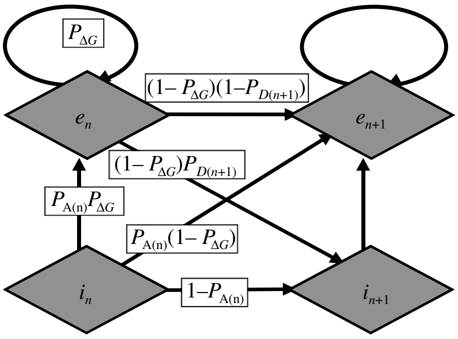

$P_{\Delta G}$
$(1- P_{\Delta G})(1-P_{D(n+1)})$
e_n
e_{n+1}
$(1- P_{\Delta G})P_{D(n+1)}$
$P_{\mathrm{A(n)}}P_{\Delta G}$
$P_{\mathrm{A(n)}}(1- P_{\Delta G})$
i_n
$1-P_{\mathrm{A(n)}}$
i_{n+1}

Fig. 1. States and transitions of the Hidden Markov Model. States are represented by diamonds. The model involves exon (e) and intron (i) states. The index n represents the position in the genomic sequence assigned to the state. Transitions between the states are indicated by arrows. The transition probabilities are shown for transitions from states e_n and i_n. $P_{\Delta G}$ is the probability of a nucleotide deletion in the genomic sequence. $P_{D(n)}$ and $P_{A(n)}$ are the probabilities of position n in the genomic DNA to be a donor or acceptor site, respectively.

The output weights in the exon states e_n are set to

$$\log P_{e_n}\begin{pmatrix} G_n \\ X \end{pmatrix} = \begin{cases} \sigma & \text{if } G_n = X \\ \mu & \text{otherwise} \end{cases}$$

$$\log P_{e_n}\begin{pmatrix} \mathrm{N} \\ X \end{pmatrix} = \nu$$

$$\log P_{e_n}\begin{pmatrix} G_n \\ \mathrm{N} \end{pmatrix} = \nu$$

$$\log P_{e_n}\begin{pmatrix} \mathrm{N} \\ \mathrm{N} \end{pmatrix} = \nu$$

$$\log P_{e_n}\begin{pmatrix} G_n \\ \_ \end{pmatrix} = \delta$$

$$\log P_{e_n}\begin{pmatrix} \mathrm{N} \\ - \end{pmatrix} = \delta$$

for $G_n \in \mathbf{A}^-$, $X \in \mathbf{A}^-$, where σ, μ, ν, and δ represent the weights for identities, mismatches, alignment positions involving undetermined characters, and cDNA deletions, respectively. The output weights corresponding to genomic sequence deletions are also set uniformly to

$$\log P_{e_n}\begin{pmatrix} - \\ X \end{pmatrix} = \delta,\ \ X \in \mathbf{A}.$$

Note that for a strict HMM formulation genomic sequence deletions would be output from additional 'delete' states. However, because the transitions from the delete states are exactly like the transitions from the corresponding exon states, our formulation is more efficient (in the coding detailed below, the output weights are always assigned in conjunction with the transition probabilities so that it is always clear whether e_n corresponds to a delete state or not). For the intron states i_n,

$$\log P_{i_n}\begin{pmatrix} G_n \\ \star \end{pmatrix} = 0.$$

Thus the output probabilities involve only four parameters (see Implementation).

With the above formulation, optimal alignments are precisely defined as state sequences $Q = q_1q_2 \ldots q_L$ with associated output S^N_M (representing a sequence alignment of $G_1G_2 \ldots G_N$ with $C_1C_2 \ldots C_M$) such that the joint probability $P(Q, S^N_M)$ is maximal over all possible Q and S^N_M. This maximal probability is calculated in standard fashion as

$$P = \max\{E^N_M, I^N_M\}\ ,$$

where

$$E^n_m = \max P(Q = q_1q_2 \ldots q_l, q_l = e_n, S^n_m)\ ,$$

and

$$I^n_m = \max P(Q = q_1q_2 \ldots q_l, q_l = i_n, S^n_m)\ ,$$

for $n = 1, 2, \ldots, N$, $m = 1, 2, \ldots, M$, $\max\{m, n\} \leq l \leq m + n$, and maximization is over all possible Q and S^n_m representing alignments of $G_1, G_2, \ldots, G_n$ with $C_1, C_2, \ldots, C_m$.

Let $\tau_{e_0,e_1} = \tau_{i_0,e_1} = \tau_{e_1}$ and $\tau_{e_0,i_1} = \tau_{i_0,i_1} = \tau_{i_1} = 1 - \tau_{e_1}$, where τ_{e_1} is the initial exon state probability. Then E^N_M and I^N_M are found from the following recursion:

$$\begin{aligned}
&E^n_0 = I^n_0 = 1,\\
&E^0_m = 1, \qquad I^0_m = 0, \qquad n = 0, 1, \ldots, N, m = 1, 2, \ldots, M,\\
&E^n_m = \max\left\{\max\left\{E^{n-1}_m \tau_{e_{n-1},e_n},\ I^{n-1}_m \tau_{i_{n-1},e_n}\right\} P_{e_n}\begin{pmatrix} G_n \\ - \end{pmatrix},\right.\\
&\qquad \max\left\{E^{n-1}_{m-1} \tau_{e_{n-1},e_n},\ I^{n-1}_{m-1} \tau_{i_{n-1},e_n}\right\} P_{e_n}\begin{pmatrix} G_n \\ C_m \end{pmatrix},\\
&\qquad \left.\max\left\{E^n_{m-1} \tau_{e_n,e_n},\ I^n_{m-1} \tau_{i_n,e_n}\right\} P_{e_n}\begin{pmatrix} - \\ C_m \end{pmatrix}\right\},\\
&I^n_m = \max\left\{E^{n-1}_m \tau_{e_{n-1},i_n},\ I^{n-1}_m \tau_{i_{n-1},i_n}\right\},\\
&\qquad n = 1, 2, \ldots, N, m = 1, 2, \ldots, M.
\end{aligned}$$

At each maximization step, the state transition and output yielding the maximum are stored to facilitate the backtracing of the optimal alignment(s). See Figure 2 for a hypothetical example.

Implementation

Given the P-values according to the SplicePredictor, few parameters need to be specified for a complete implementation of the algorithm. The following default values worked well for a large range of applications we examined:

$$\begin{aligned}
&\tau_{e_1} = 0.5\\
&P_{\Delta G} = 0.03\\
&\sigma = 2.0 \qquad \mu = -2.0 \qquad \nu = 0.0 \qquad \delta = -4.0.
\end{aligned}$$

In a typical application the genomic DNA template will extend 5′ and 3′ of the cDNA ends. Setting E^n_0 and I^n_0 to one for all n amounts to no end gap penalties at the 5′ end. To symmetrize with respect to 3′ end gap penalties, in the programmed updating of E^n_m the output weight

$$\log P_{e_n}\begin{pmatrix} G_n \\ - \end{pmatrix}$$

is set to zero for $m = M$. Similarly,

$$\log P_{e_n}\begin{pmatrix} - \\ C_m \end{pmatrix}$$

is set to zero for $n = N$. Note that within-exon gaps in the genomic DNA are more costly than gaps in the cDNA because $\log P_{\Delta G} < \log(1 - P_{\Delta G})$, a desirable setting as cDNA or EST sequences are typically less reliably determined in practice.

The memory requirements of the program are minimal because the updating of the E^n_m and I^n_m matrices at a given index pair only involves cells at most one row and column up and to the left. In practice, for given index n the program simply fills out one of two row arrays of size M (labeled $n \mod 2$), using information from the previous calculations stored in the array labeled $(n - 1) \mod 2$. In the next step the then unnecessary information in array $(n - 1) \mod 2$ is overwritten. For convenience, we store the maximal scoring state transitions for backtracing an optimal path. This allows rapid recovery of an optimal alignment at the cost of extra storage.

index *n*	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
genomic DNA	**T**	**-**	**C**	**A**	**G**	**G**	**T**	**A**	**A**	**G**	**T**	**C**	**A**	**A**	**A**	**T**
EST	**T**	**T**	**C**	**A**	**N**	*	*	*	*	*	**T**	**C**	**-**	**-**	**C**	**T**
index *m*	1	2	3	4	5	5	5	5	5	5	6	7	7	7	8	9
state sequence	e_1	e_1	e_2	e_3	e_4	i_5	i_6	i_7	i_8	i_9	e_{10}	e_{11}	e_{12}	e_{13}	e_{14}	e_{15}
transition probabilities	τ_{e1}	τ_{e1e1}	τ_{e1e2}	τ_{e2e3}	τ_{e3e4}	τ_{e4i5}	τ_{i5i6}	τ_{i6i7}	τ_{i7i8}	τ_{i8i9}	τ_{i9e10}	τ_{e10e11}	τ_{e11e12}	τ_{e12e13}	τ_{e13e14}	τ_{e14e15}
output weights	σ	δ	σ	σ	ν	0	0	0	0	0	σ	σ	δ	δ	μ	σ

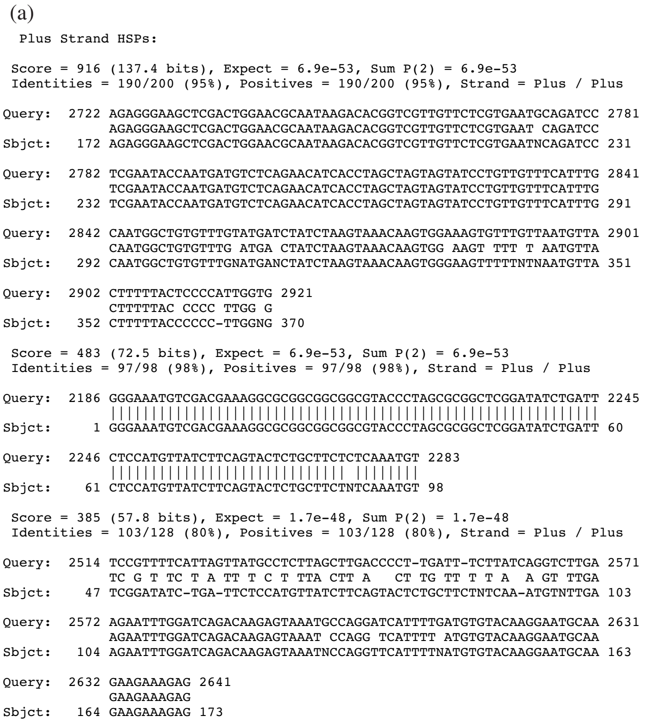
Fig. 2. Hypothetical alignment of a genomic DNA with an EST sequence. The genomic sequence (second row) comprises 15 nucleotides. The EST sequence (third row) is of length nine nucleotides, including in position 5 a non-determined base (N). An alignment is shown that assigns intron status to the genomic DNA positions 5–9. The underlying state sequence is displayed in the fifth row. Indeces *n* (first row) and *m* (forth row) record the position in the genomic DNA and EST sequences, respectively. A deletion in the genomic sequence is accommodated by the transition from state e_1 into itself. The transition probabilities and output weights (bottom two rows) are assigned as described in the text. The algorithm maximizes the sum of log transition probabilities plus output weights over all possible spliced alignments.

The algorithm was implemented as the C subroutine sahmtD (for Spliced Alignment Hidden Markov Tool for cDNA) in our previous SplicePredictor program (Brendel and Kleffe, 1998; Kleffe *et al.*, 1996). Limitations on the maximal lengths of the genomic DNA and cDNA depend on the memory of the CPU. Our WWW server is currently set up to align genomic DNA segments up to 13 kb against a cDNA of up to 7 kb. If the input exceeds these limits, an exit warning is displayed and recompilation suggested with increased limits. For plant genes that typically lack long exons and introns (90% of maize and *Arabidopsis* exons and introns are less than 510 nucleotides; Brendel *et al.*, 1998) these limits appear adequate in practice. Detection of long introns would not be explicitly feasible by our approach. However, a long intron would be a reasonable interpretaton if the 5′ and 3′ ends of a single EST matched dispersed regions in a genomic DNA.

One possible use of the algorithm is to screen a novel large genomic contig against an entire EST database. This could be achieved by pre-selecting matching ESTs with a fast screening program like BLAST (Altschul *et al.*, 1997) or its derivatives; e.g. http://genome-www2.stanford.edu/cgi-bin/AtDB/nph-blast2atdb. This strategy has been pursued by Florea *et al.* (1998) and Mott (1997). Alternatively, we have also implemented the sahmtD subroutine in a standalone program called GeneSeqer. In GeneSeqer, each EST is initially fast-screened against the genomic DNA for a matching region of specific quality. A matching region defines the core of a larger segment that will produce a significant spliced alignment. Our implementation is based on the initial identification of exactly matching 12mers by the suffix array method of Manber and Myers (1993). Matching 12mers are first maximally extended and then assembled into matching regions allowing for small insertions and deletions in both genomic and cDNA and longer gaps in the genomic DNA (possible introns). These regions in the genomic DNA are then extended by typically several hundred nucleotides to define the segment to which the sahmtD algorithm is applied. Details will be presented elsewhere (W. Zhu and V. Brendel, in preparation). For our Web server we are periodically pre-processing the major publicly available plant EST collections. However, the pre-processing for user-specific EST collections can also be achieved in reasonable time. For example, pre-processing of 37 745 *Arabidopsis thaliana* ESTs from GenBank took 9 min 22 s on our server in single-user mode. Matching these ESTs onto a 107 kb contig (GenBank U89959) produced 89 spliced alignments in 8 min 29 s. For comparison, the sim4 program of Florea *et al.* (1998) took 18 min 35 s on the same data (note that this program produces all exact matches of length at least 12 and thus the output is much less specific than the GeneSeqer output; concerning sensitivity, see Figure 6 and Discussion).

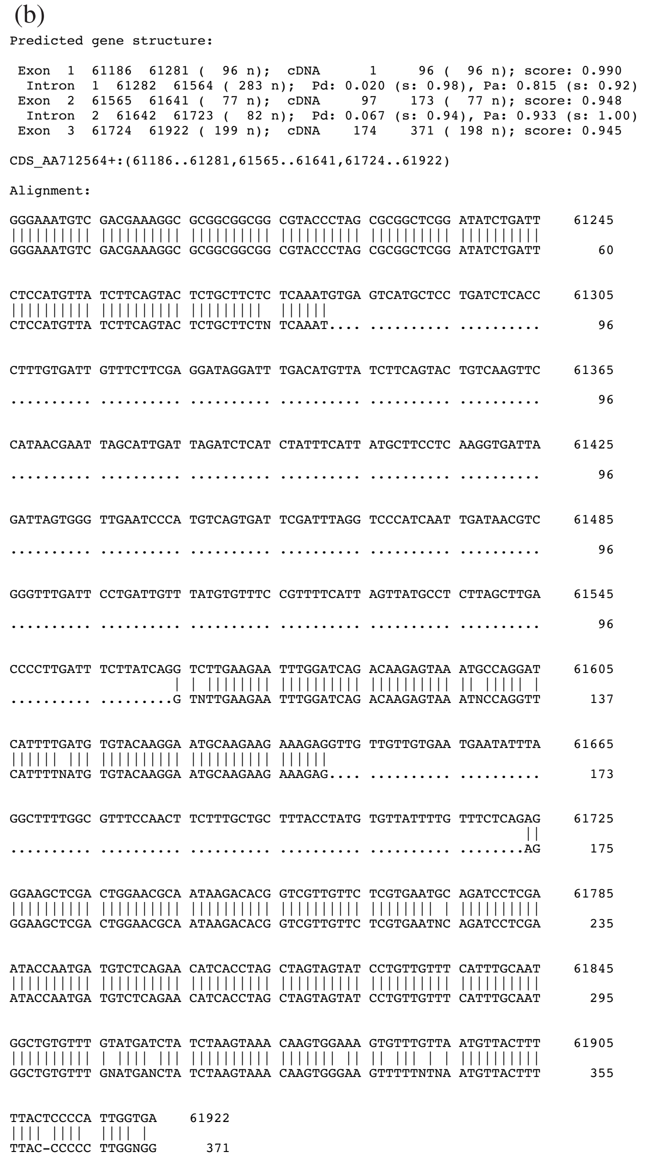
Minimal intron length

The algorithm as given does not impose any restrictions on exon and intron lengths. Naturally occurring introns exceed a minimal length of about 55–60 bases (for plants, see Brendel *et al.*, 1998). To avoid solutions with unacceptably small intron assignments, the algorithm was modified to include a ‘short intron penalty’. This penalty is levied upon intron to exon state transitions depending on whether or not the then closed intron exceeds the required minimum length. The implementation is straightforward: at the maximization step for I_m^n, a variable **intronstart[n][m]** is set to *n* (beginning of a new intron) or carried over from **intronstart[n-1][m]** (continuation of an existing intron; the index *n* of **intronstart** can again be replaced by *n* mod 2). The weight is added during the E_m^n maximization if the current index *n* does not exceed the **intronstart** value by more than the defined minimal intron length. Because of the left to right directionality of the maximization algorithm, the modified procedure is not guaranteed to find the optimal score of all alignments satifying the minimal intron length constraint. In practice, this poses no problem. The different donor site assignments are most likely well distinguished by the



```
(a)
  Plus Strand HSPs:

 Score = 916 (137.4 bits), Expect = 6.9e-53, Sum P(2) = 6.9e-53
 Identities = 190/200 (95%), Positives = 190/200 (95%), Strand = Plus / Plus

Query:  2722 AGAGGGAAGCTCGACTGGAACGCAATAAGACACGGTCGTTGTTCTCGTGAATGCAGATCC 2781
             AGAGGGAAGCTCGACTGGAACGCAATAAGACACGGTCGTTGTTCTCGTGAAT CAGATCC
Sbjct:   172 AGAGGGAAGCTCGACTGGAACGCAATAAGACACGGTCGTTGTTCTCGTGAATNCAGATCC 231

Query:  2782 TCGAATACCAATGATGTCTCAGAACATCACCTAGCTAGTAGTATCCTGTTGTTTCATTTG 2841
             TCGAATACCAATGATGTCTCAGAACATCACCTAGCTAGTAGTATCCTGTTGTTTCATTTG
Sbjct:   232 TCGAATACCAATGATGTCTCAGAACATCACCTAGCTAGTAGTATCCTGTTGTTTCATTTG 291

Query:  2842 CAATGGCTGTGTTTGTATGATCTATCTAAGTAAACAAGTGGAAAGTGTTTGTTAATGTTA 2901
             CAATGGCTGTGTTTG ATGA CTATCTAAGTAAACAAGTGG AAGT TTT T AATGTTA
Sbjct:   292 CAATGGCTGTGTTTGNATGANCTATCTAAGTAAACAAGTGGGAAGTTTTTNTNAATGTTA 351

Query:  2902 CTTTTTACTCCCCATTGGTG 2921
             CTTTTTAC CCCC TTGG G
Sbjct:   352 CTTTTTACCCCCC-TTGGNG 370

 Score = 483 (72.5 bits), Expect = 6.9e-53, Sum P(2) = 6.9e-53
 Identities = 97/98 (98%), Positives = 97/98 (98%), Strand = Plus / Plus

Query:  2186 GGGAAATGTCGACGAAAGGCGCGGCGGCGGCGTACCCTAGCGCGGCTCGGATATCTGATT 2245
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct:     1 GGGAAATGTCGACGAAAGGCGCGGCGGCGGCGTACCCTAGCGCGGCTCGGATATCTGATT 60

Query:  2246 CTCCATGTTATCTTCAGTACTCTGCTTCTCTCAAATGT 2283
             ||||||||||||||||||||||||||||| ||||||||
Sbjct:    61 CTCCATGTTATCTTCAGTACTCTGCTTCTNTCAAATGT 98

 Score = 385 (57.8 bits), Expect = 1.7e-48, Sum P(2) = 1.7e-48
 Identities = 103/128 (80%), Positives = 103/128 (80%), Strand = Plus / Plus

Query:  2514 TCCGTTTTCATTAGTTATGCCTCTTAGCTTGACCCCT-TGATT-TCTTATCAGGTCTTGA 2571
             TC G T TC T A TT T C T TTA CTT A   CT TG TT T T A  A GT TTGA
Sbjct:    47 TCGGATATC-TGA-TTCTCCATGTTATCTTCAGTACTCTGCTTCTNTCAA-ATGTNTTGA 103

Query:  2572 AGAATTTGGATCAGACAAGAGTAAATGCCAGGATCATTTTGATGTGTACAAGGAATGCAA 2631
             AGAATTTGGATCAGACAAGAGTAAAT CCAGG TCATTTT ATGTGTACAAGGAATGCAA
Sbjct:   104 AGAATTTGGATCAGACAAGAGTAAATNCCAGGTTCATTTTNATGTGTACAAGGAATGCAA 163

Query:  2632 GAAGAAAGAG 2641
             GAAGAAAGAG
Sbjct:   164 GAAGAAAGAG 173
```

Fig. 3. Resolution of EST hits from a BLAST search. (a) The 3kb region 59001–62000 of the *Arabidopsis thaliana* contig U89959 ('Query') was subjected to a BLAST search against the *Arabidopsis* EST database using the Stanford Genome Center server (http://genome-www2.stanford.edu/cgi-bin/AtDB/nph-blast2atdb). Three hits are reported with the plus strand of the 371 base cDNA clone AA712564 ('Subject'). (b) SplicePredictor resolves the three BLAST hits into a single consistent spliced alignment consisting of three exons and two introns. The EST represents a full or almost full-length cDNA encoding a 71 amino acid polypeptide (start codon at 61191–61193, stop codon at 61769–61771). The scores for the predicted exons were calculated as described in the text. The predicted donor and acceptor sites are scored by *P*-value (Pd and Pa, respectively, Kleffe *et al.*, 1996) and by the similarity score (s) calculated for the proximal 50 exon bases. Alignment positions that align identical letters are indicated by vertical bars.



```
(b)
Predicted gene structure:

 Exon  1  61186  61281 (  96 n);  cDNA      1     96 (  96 n); score: 0.990
  Intron  1  61282  61564 ( 283 n);  Pd: 0.020 (s: 0.98), Pa: 0.815 (s: 0.92)
 Exon  2  61565  61641 (  77 n);  cDNA     97    173 (  77 n); score: 0.948
  Intron  2  61642  61723 (  82 n);  Pd: 0.067 (s: 0.94), Pa: 0.933 (s: 1.00)
 Exon  3  61724  61922 ( 199 n);  cDNA    174    371 ( 198 n); score: 0.945

CDS_AA712564+:(61186..61281,61565..61641,61724..61922)

Alignment:

GGGAAATGTC GACGAAAGGC GCGGCGGCGG CGTACCCTAG CGCGGCTCGG ATATCTGATT    61245
|||||||||| |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
GGGAAATGTC GACGAAAGGC GCGGCGGCGG CGTACCCTAG CGCGGCTCGG ATATCTGATT       60


CTCCATGTTA TCTTCAGTAC TCTGCTTCTC TCAAATGTGA GTCATGCTCC TGATCTCACC    61305
|||||||||| |||||||||| |||||||||  ||||||
CTCCATGTTA TCTTCAGTAC TCTGCTTCTN TCAAAT.... .......... ..........       96


CTTTGTGATT GTTTCTTCGA GGATAGGATT TGACATGTTA TCTTCAGTAC TGTCAAGTTC    61365

.......... .......... .......... .......... .......... ..........       96


CATAACGAAT TAGCATTGAT TAGATCTCAT CTATTTCATT ATGCTTCCTC AAGGTGATTA    61425

.......... .......... .......... .......... .......... ..........       96


GATTAGTGGG TTGAATCCCA TGTCAGTGAT TCGATTTAGG TCCCATCAAT TGATAACGTC    61485

.......... .......... .......... .......... .......... ..........       96


GGGTTTGATT CCTGATTGTT TATGTGTTTC CGTTTTCATT AGTTATGCCT CTTAGCTTGA    61545

.......... .......... .......... .......... .......... ..........       96


CCCCTTGATT TCTTATCAGG TCTTGAAGAA TTTGGATCAG ACAAGAGTAA ATGCCAGGAT    61605
                    | | |||||||| |||||||||| |||||||||| || ||||| |
.......... .........G TNTTGAAGAA TTTGGATCAG ACAAGAGTAA ATNCCAGGTT      137


CATTTTGATG TGTACAAGGA ATGCAAGAAG AAAGAGGTTG TTGTTGTGAA TGAATATTTA    61665
|||||| ||| |||||||||| |||||||||| ||||||
CATTTTNATG TGTACAAGGA ATGCAAGAAG AAAGAG.... .......... ..........      173


GGCTTTTGGC GTTTCCAACT TCTTTGCTGC TTTACCTATG TGTTATTTTG TTTCTCAGAG    61725
                                                               ||
.......... .......... .......... .......... .......... ........AG      175


GGAAGCTCGA CTGGAACGCA ATAAGACACG GTCGTTGTTC TCGTGAATGC AGATCCTCGA    61785
|||||||||| |||||||||| |||||||||| |||||||||| |||||||| | ||||||||||
GGAAGCTCGA CTGGAACGCA ATAAGACACG GTCGTTGTTC TCGTGAATNC AGATCCTCGA      235


ATACCAATGA TGTCTCAGAA CATCACCTAG CTAGTAGTAT CCTGTTGTTT CATTTGCAAT    61845
|||||||||| |||||||||| |||||||||| |||||||||| |||||||||| ||||||||||
ATACCAATGA TGTCTCAGAA CATCACCTAG CTAGTAGTAT CCTGTTGTTT CATTTGCAAT      295


GGCTGTGTTT GTATGATCTA TCTAAGTAAA CAAGTGGAAA GTGTTTGTTA ATGTTACTTT    61905
|||||||||| | |||| ||| |||||||||| ||||||| || || ||| | | ||||||||||
GGCTGTGTTT GNATGANCTA TCTAAGTAAA CAAGTGGGAA GTTTTTNTNA ATGTTACTTT      355


TTACTCCCCA TTGGTGA    61922
|||| ||||  |||| |
TTAC-CCCCC TTGGNGG      371
```

Fig. 3. cont.

combination of *P*-value and alignment quality, and the intron length restriction would mainly serve to eliminate alignment paths that display within likely introns short stretches of relatively high sequence similarity that can be expected at random.

Scoring the alignment

The program scores each predicted exon separately by tallying up the output weights corresponding to the alignment of exon and cDNA. Only matches and penalties for gaps in the genomic DNA are counted. This value is normalized by the equivalent sum of weights assuming perfect matching to the genomic DNA. For ungapped alignments, this score is correlated with percentage identity. From our experience, the optimal alignment of unrelated ESTs to a genomic DNA rarely produces exon quality values above 0.4 for exons of lengths at least 60 nucleotides (data not shown).

For donor and acceptor sites, the program displays the *P*-values and evaluates the exon quality for the adjacent 50 exon bases. For non-cognate, but homologous ESTs, these values may indicate high conservation around the splice sites, even though the central parts of long exons may have diverged considerably.

```
Predicted gene structure:

 Exon  1  62556  62646 (  91 n);  cDNA    316    406 (  91 n); score: 0.637
  Intron  1  62647  62847 ( 201 n);  Pd: 0.059 (s: 0.64), Pa: 0.523 (s: 0.64)
 Exon  2  62848  63178 ( 331 n);  cDNA    407    737 ( 331 n); score: 0.656
  Intron  2  63179  63270 (  92 n);  Pd: 0.278 (s: 0.74), Pa: 0.982 (s: 0.56)
 Exon  3  63271  63391 ( 121 n);  cDNA    738    854 ( 117 n); score: 0.612
  Intron  3  63392  63661 ( 270 n);  Pd: 0.390 (s: 0.56), Pa: 0.001 (s:  n/a)
 Exon  4  63662  63683 (  22 n);  cDNA    855    877 (  23 n); score: 0.591
  Intron  4  63684  63754 (  71 n);  Pd: 0.996 (s:  n/a), Pa: 0.996 (s:  n/a)
 Exon  5  63755  63773 (  19 n);  cDNA    878    896 (  19 n); score: 0.684

CDS_AB008268+:(62556..62646,62848..63178,63271..63391,63662..63683,63755..63773)

Alignment:

CTCAAAGGCT GCATCAACGA CGCCAAGTGC ATGCG-TCAC CTTCTCATCA ACAAATTCAA    62614
 | ||||| | ||||||  ||  || |||| | ||| | ||     |    ||| ||||| |
TTGAAAGGTT GCATCAGTGA TGCTAAGTCC ATGAGATCTT TATTGGTTCA ACAAA-TGGG      374

ATTCTCCCCA GATTCAATTC TCATGCTTAC CGGTACAGAG TATTTCTATC TTTTCAAATG    62674
 ||| |     || || |||| ||||||| ||  |
TTTCCCTATT GACTCTATTC TCATGCTCAC AG........ .......... ..........      406

CCTATGTTTG CTACTATACT ACTATTCCTT GGATTTTGAA TACAATTTTC CTTGGCCTCT    62734

.......... .......... .......... .......... .......... ..........      406

TCAATCTGAT AAACACACAT TCCAAGTTAC CATTTCGAAC CACTTTGATA AAAATGTGTT    62794

.......... .......... .......... .......... .......... ..........      406

GCATTCCATA GCTGACTAAC TAATTGTTCA TCATGGATGG TTTTCATTCT CAGAGGAAGA    62854
                                                          | || ||
.......... .......... .......... .......... .......... ...AAGATGA      413

AACTGATCCA TATCGTATCC CGACCAAGCA AAACATGAGG ATGGCATTGT ATTGGCTCGT    62914
| |    ||   |  | || | |||| |||   |||||| ||| | |||  ||    ||| | ||
AGCCAGCCCG CAGAGAATAC CGACGAAGAG AAACATTAGG AAGGCGATGA GATGGTTAGT      473

ACAGGGATGC ACAGCAGGCG ACTCACTTGT CTTCCACTAC TCTGGTCATG GTTCGCGTCA    62974
  | ||   | | |||| | | ||||||| ||  ||||| | | |||||||||| | || |  ||
TGAAGGGAAC AGAGCAAGGG ACTCACTAGT GTTCCATTTC TCTGGTCATG GATCTCAGCA      533

AAGAAACTAC AACGGTGATG AAGTTGATGG CTATGATGAA ACACTCTGTC CTCTGGATTT    63034
 |   ||||| ||||| || | |  | |||||   | ||||||  |  | || | || | ||
GAATGACTAC AACGGAGACG AGATCGATGG TCAAGATGAA GCCTTGTGCC CTTTAGACCA      593

TGAAACTCAG GGGATGATTG TAGACGATGA GATCAACGCA ACCATTGTAC GCCCTCTTCC    63094
||||||  |  || |  ||   | || || || ||| |||    |   | ||   | |||||
TGAAACAGAA GGAAAAATCA TTGATGACGA GATTAACCGG ATACTCGTGA GGCCTCTCGT      653

ACATGGTGTC AAGCTCCATT CAATTATCGA TGCTTGCCAT AGTGGTACCG TTCTGGATTT    63154
 ||||| |   ||||| ||   |  | |||||  || ||  |  || || || | | || |||||
CCATGGAGCT AAGCTTCACG CTGTCATCGA CGCCTGTAAC AGCGGGACTG TCCTTGATTT      713

ACCCTTCCTA TGCAGAATGA ACAGGTTATT AGTCCCTCAA CCGCTTCTAA AAGGGATGTT    63214
||||||| |  ||||| |||  | ||
ACCCTTCATT TGCAGGATGG AGAG...... .......... .......... ..........      737

GCTTACCTCT CTCGTTATAT TTAACATACA TCCATTTTTT TTTTTAATTG AAACAGAGCT    63274
                                                                |
.......... .......... .......... .......... .......... ......GAAT      741

GGGCAGTATG TGTGGGAGGA TCATCGGCCT AGGTCAGGTT TGTGGAAAGG AACTGCTGGT    63334
||    ||||   ||||| ||  ||| |  |  | |||||    | |  ||||| ||| | ||||
GGTTCTTATG AATGGGAAGA CCATAGATC- A-GTCAGAGC T-TACAAAGG AACAGATGGT      798

GGAGAAGCCA TTTCAATTAG TGGATGTGAT GATGATCAGA CTTCGGCCGA CACATCAGTA    63394
|||| |||   | |   | || ||  |||||  |||||| |   |    |   | ||| ||
GGAGCAGCTT TCTGTTTCAG TGCTTGTGAC GATGATGAAT CCAGTGGTTA CAC-TCC...      854

AGTAGAACGA CTCTAATCAT ACGTCTTGCT GTTGTAGTTG GTTCCTCCTC TCATGATTAA    63454

.......... .......... .......... .......... .......... ..........      854

AACACATACA CAGGCGCTGT CGAAGATCAC GTCTACGGGT GCTATGACTT TCTGTTTTAT    63514

.......... .......... .......... .......... .......... ..........      854

TCAAGCAATT GAACGCAGCG CACAAGGCAC AACCTATGGA AGCCTTCTGA ATTCTATGCG    63574

.......... .......... .......... .......... .......... ..........      854

CACCACAATA AGGAATACAG GGAATGATGG TGGTGGTAGT GGTGGAGTTG TGACGACTGT    63634

.......... .......... .......... .......... .......... ..........      854

GCTGAGCATG CTTCTGACAG GGGGAAGTGC GAT-TGGGGG ATTAAGACAG GTAAAATTCT    63693
                             ||  | |   |||| |  || ||||
.......... .......... .......TGT GTTCACGGGG AAGAACACAG ..........      877

TTCTTGCTCT CTTGTGTTGA TACAGATCGA TAAATGTTTT CTTAAATCTG TTTTTTGACA    63753

.......... .......... .......... .......... .......... ..........      877

GGAGCCTCAA CTGACTGCTT    63773
 |||||   | || |  ||||
.GAGCCATGA CTTATAGCTT      896
```

Fig. 4. Spliced alignment of a hypothetical cDNA from *Arabidopsis thaliana* contig GenBank AB008268 with the 62001 to 64000 segment of contig GenBank U89959. The cDNA was derived from the predicted gene from positions 201025 to 22033 of contig AB008268 (Table 1), curtailed to include from exon 1 only the last 100 nucleotides (the alignment with the full-length cDNA is unchanged for the displayed segments but meaningless for the divergent 5′-terminal nucleotides; cf. Figure 5).

Applications and discussion

We illustrate performance of the algorithm with examples that arose in our attempts to annotate a segment of the *Arabidopsis thaliana* chromosome 1 contig GenBank U89959 (106973 nucleotides) with the help of SplicePredictor (Brendel and Kleffe, 1998). Figure 3 shows the results of a search against the *Arabidopsis* EST database with the segment from positions 59001 to 62000, a region that our initial gene finding algorithms had difficulty resolving. The EST approach proved successful as we recovered clear evidence for a complete gene in that region. The predicted gene product of 71 amino acids shows partial similarity to yeast protein COX17 and the hypothetical yeast protein YHR6 (Brendel and Kleffe, 1998).

In our second example we analyze the region 62001 to 64000 on the same contig. Initial gene prediction for that region had suggested a gene product with similarity to another hypothetical gene on a different contig (GenBank AB008268) on chromosome 5. In Figure 4 we show the spliced alignment of the hypothetical cDNA of the AB008268 protein with the U89959 genomic DNA segment. Remarkably, the five exon structure of the AB008268 gene is conserved in the alignment, with part of exon 1 and exons 2 and 3 significantly similar even at the nucleotide level. Thus, it is likely that these are two genes of a gene family that may have arisen from gene duplication. Intron 4 is strongly predicted by very high donor and acceptor scores, but the 3′ end of intron 3 seems tentative. Closer inspection of the coding potential suggests that the genes have diverged at the 3′ end by an insertion in U89959 relative to AB008268 (or, equivalently, a deletion in AB008268 relative to U89959) according to the exon/intron assignments displayed in Table 1. The alignment of the putative gene products (Figure 5) strongly supports these assignments.

This example also illustrates limitations of the algorithm. The long insertion in the U89959-encoded exon 4 relative to the AB008268-encoded exon 4 is associated with too high a gap penalty compared with the alternative alignment picked by the algorithm (Figure 4) to outweigh the benefits of the much better scoring of exon 5 of the presumably correct alignment given in Table 1 and Figure 5. It is clear that the default parameters of the algorithm specifying the relative weights of splice site scores, sequence matching, and gap penalties will not be uniformly optimal.

For comparison, we also used the GAP2 program of Huang *et al.* (1997), the EST_GENOME program of Mott (1997), and the *ab initio* gene prediction algorithm GenScan of Burge and Karlin (1997) on the same data. GAP2 predicts the gene structure (62557 . . **62646,62848** . . 63180,63273 . . 63402,63489 . .

Table 1. Predicted gene structure of two closely related *Arabidopsis thaliana* genes. The two potential genes located on chromosomes 1 (contig GenBank U89959) and 2 (contig GenBank AB008268) were predicted as described in the text. The 'from/score' and 'to/score' columns give the starting and ending positions of the exons or the splice site scores of the introns (see Brendel and Kleffe, 1998; 15* is the optimal score). The 'size' column refers to the lengths of the exons and introns in number of nucleotides. The column 'sim' gives the similarity score comparing the corresponding exons from U89959 and AB008268 as derived from the spliced alignment (Figure 4). The similarity score for the first exon (shown in parenthesis) refers to the score for the 3′ end of the exons aligned as in Figure 4. Exons 4 and 5 as given by the coordinates in this Table were not aligned by the algorithm

		U89959			AB008268			
	#	from/score	to/score	size	from/score	to/score	size	sim
exon	1	62310	62646	337	20620	21025	406	(0.60)
intron	1	5*	10*	201	15*	15*	89	
exon	2	62848	63178	331	21115	21445	331	0.66
intron	2	12*	15*	92	15*	6*	81	
exon	3	63271	63391	121	21527	21644	118	0.60
intron	3	9*	3*	76	5*	15*	87	
exon	4	63468	63683	216	21732	21905	174	
intron	4	15*	15*	71	15*	10*	68	
exon	5	63755	63814	60	21974	22033	60	

```
                          8                      32
U89959         mlvncsg    CRTPLQLPSGARSIRCALCQAVTHI adprtappp
                          |   + +   ||+++|+ |..||.+
AB008268       masrrevrcr CGRRMWVQPDARTVQCSTCHTVTQL yslvdiarganriihgfqqllrqhqpqhheq
                          11                     35

                 42           56     60
U89959           QPSSAPSPPPQIHAP pgq LPHPHGRKRAVICGISYRFSRHELKGCINDAKCMRHLLINKFKFSP
                 | .   .|||++  |     || | |+||||+||++|+ . +.|||||+||| || ||+.+. |
AB008268       q QQQMMAQPPPRLLEP     LPSPFGKKRAVLCGVNYKGKSYSLKGCISDAKSMRSLLVQQMGFPI
                 68           82     83

U89959         DSILMLT=EEETDPYRIPTKQNMRMALYWLVQGCTAGDSLVFHYSGHGSRQRNYNGDEVDGYDETLCP
               ||||||| |+|..| |||||+|+| |+ |||+|  | ||||||+|||||+|.+|||||+|| ||.|||
AB008268       DSILMLT=EDEASPQRIPTKRNIRKAMRWLVEGNRARDSLVFHFSGHGSQQNDYNGDEIDGQDEALCP

                                                                          233
U89959         LDFETQGMIVDDEINATIVRPLPHGVKLHSIIDACHSGTVLDLPFLCRMNR=AGQYVWEDHR prsgl
               || ||+| |+|||||  +|||| ||.|||++||||+|||||||||+|||.|  |.| |||||
AB008268       LDHETEGKIIDDEINRILVRPLVHGAKLHAVIDACNSGTVLDLPFICRMER=NGSYEWEDHR svra
                                                                          256

                239                 262     265                                30
U89959          WKGTAGGEAISISGCDDDQTSADT s=a LSKITSTGAMTFCFIQAIERSAQGTTYGSLLNSMRTTI
                +||| || |. .|.||||++|. |     +    +|||||+ ||+|++ +. . ||| ||| | +.|
AB008268        YKGTDGGAAFCFSACDDDESSGYT p=  VFTGKNTGAMTYSFIKAVKTAGPAPTYGHLLNLMCSAI
                261                 284     286                                32

               3                                   336             354
U89959         R ntgndgggsggvvttvlsmlltggsaigglrq= EPQLTACQTFDVYAKPFTL
               |                                   || ||+ + ||+||  |.|
AB008268       R eaqsrlafngdytssdasa=              EPLLTSSEEFDLYATKFVL
               4                                   344             362
```

Fig. 5. Alignment of the predicted protein sequences encoded by the exons from the *Arabidopis thaliana* contigs GenBank U89959 and GenBank AB008268 as assigned in Table 1. The alignment was produced by the PPAT algorithm (V. Brendel, unpublished) which is an extension of the published SSPA algorithm (Karlin *et al.*, 1995). Introns are indicated by '='. The alignment was scored with the BLOSUM62 amino acid substitution scoring matrix (Henikoff and Henikoff, 1992). Aligned residues are connected with a vertical bar if identical, by '+' if positively scoring in the BLOSUM62 matrix, by '.' if scoring 0, and by a blank if scoring negatively. Residues that could not be aligned in significantly scoring alignment blocks are given in lower case. It is seen that strong conservation extends over the C-terminal parts of exons 1, all of exons 2 and 3, the N-terminal parts of exons 4, and all of exons 5.

63580); exon boundaries in agreement with Table 1 are printed in bold face. This assignment includes non-consensus splice sites at 63181 (TA donor), 63403 (GA donor), and 63488 (CT acceptor). EST_GENOME (version 4, obtained from ftp.sanger.ac.uk/pub/pmr) gives an alignment with only two introns, (62554..**62646**,**62848**..**63178**,**63271**..63370). GenScan predicted the 5′ incomplete gene structure (62294..62481,62507..**62646**,**62848**..63138,63312..**63391**,63432..**63683**,**63755**..**63814**). This assignment includes the unrealistically small intron 63392 to 63431 of 40 nucleotides only. These comparisons suggest that our algorithm can usefully extend gene prediction in the presence of weak sequence similarity information. We

```
Predicted gene structure:

 Exon  1    749    772 (  24 n);  cDNA      1     24 (  24 n); score: 0.833
  Intron  1    773    879 ( 107 n);  Pd: 0.820 (s:  n/a), Pa: 0.440 (s: 0.82)
 Exon  2    880    939 (  60 n);  cDNA     25     84 (  60 n); score: 0.817
  Intron  2    940   1034 (  95 n);  Pd: 0.629 (s: 0.80), Pa: 0.852 (s: 0.80)
 Exon  3   1035   1307 ( 273 n);  cDNA     85    357 ( 273 n); score: 0.791
  Intron  3   1308   1382 (  75 n);  Pd: 0.757 (s: 0.70), Pa: 0.706 (s: 0.68)
 Exon  4   1383   1457 (  75 n);  cDNA    358    432 (  75 n); score: 0.653
  Intron  4   1458   1553 (  96 n);  Pd: 0.166 (s: 0.58), Pa: 0.844 (s:  n/a)
 Exon  5   1554   1582 (  29 n);  cDNA    433    461 (  29 n); score: 0.793

CDS_2317492+:(749..772,880..939,1035..1307,1383..1457,1554..1582)

Alignment:

GGTAGAAATG GCCAACCAAA ACAGGTTTTT GATTCTTCTT CTTTGCAAAA AACTGTTCTT      808
|| ||||||| | || || || ||||
GGCAGAAATG GTCAGCCGAA ACAG...... .......... .......... ..........       24


TTAATCTTCT TAAGTTAGTT GATTGGTTAG TGTTTCGTGT TTTTATTAAT TTTTCTACAA      868

.......... .......... .......... .......... .......... ..........       24


TGTCAATGCA GACAATTAGC TACATGGCTG AGCGTGTTGT TGGTCACGGA TCTTTTGGTG      928
            ||| ||||| |||||||||| ||||| ||||  || || ||  || |||||
.......... .ACAGTTAGC TACATGGCTG AGCGTATTGT AGGGCAAGGT TCATTTGGGA       73


TTGTGTTCCA AGTGAGTTTT TGTTAGCTTA TCATCTCTCA CGTTTTACAT CATGTATCTA      988
|||| |||||
TTGTCTTCCA G......... .......... .......... .......... ..........       84


TCAATTGACT GAATTGGATT CCTCTGTATT TTCTTATGTG ATTTAGGCGA AATGTCTTGA     1048
                                                  || | ||||||| ||
.......... .......... .......... .......... ......GCAA AATGTCTGGA       98


GACAGGAGAA ACTGTTGCGA TAAAGAAAGT TTTACAAGAT AGGAGGTACA AGAACCGTGA     1108
|||||| ||  || ||||| | | ||||| || | | || ||| | | | |||| ||||||||||
GACAGGTGAG ACAGTTGCTA TCAAGAAGGT TCTTCAGGAT AAGCGCTACA AGAACCGTGA      158


GCTTCAAACC ATGAGGCTAC TTGACCATCC TAATGTTGTG TCTCTGAAAC ATTGTTTCTT     1168
|||||| ||| ||| | || | ||||||| ||  ||||||||   ||||||| | | ||||||||
GCTTCAGACC ATGCGCCTTC TTGACCACCC AAATGTTGTA GCTCTGAAGC ACTGTTTCTT      218


CTCAACCACT GAAAAAGATG AGCTTTACCT CAATCTTGTT CTTGAGTACG TTCCAGAAAC     1228
||| || ||| || || |||| | || || ||  ||  | ||| |||||||| | | || |||||
CTCTACAACT GAGAAGGATG AACTGTATCT AAACTTGGTT CTTGAGTATG TGCCTGAAAC      278


TGTTCATCGT GTTATCAAAC ACTACAACAA ACTGAATCAG AGAATGCCTC TTATATACGT     1288
|||||||||| ||| | || | | ||||||||   |||| |||  | ||||| | |||| || ||
TGTTCATCGT GTTGTGAAGC ATTACAACAA GATGAACCAG CGTATGCCAC TTATCTATGT      338


CAAACTTTAC ACTTATCAGG TATCAATTAT TACATTTCTG TTAGATTTAG AGTTATGTTT     1348
 || || ||  |  || |||
GAAGCTGTAT ATGTACCAG. .......... .......... .......... ..........      357


TGGATCTCTT ATCGGATTTC TCTATGTTTG GCAGATTTTT AGAGCCTTAT CTTACATTCA     1408
                                     |||| | || || |||  ||||||| ||
.......... .......... .......... ....ATTTGT AGGGCATTAG CTTACATCCA      383


CCGATGCATT GGTGTGTGTC ATCGTGACAT AAAACCTCAA AACTTGTTGG TATGTACAAG     1468
     ||||  || || || | |  | || ||  || || ||  ||  |  ||
TAATAGCATC GGAGTTTGCC ACAGAGATAT CAAGCCACAG AATCTTCTG. ..........      432


TTAAATAAAC AGGAGCTCAC AGTATACCCG GGAATATACT TTTCTTCATT GTCTAATGCT     1528

.......... .......... .......... .......... .......... ..........      432


ACTTTTTTCA CCTGTAATGG AACAGGTAAA TCCGCACACT CATCAAGTAA AGCT     1582
                           |||||  || || ||  |||||| | | ||||
.......... .......... .....GTAAA CCCACATACC CATCAACTCA AGCT      461
```

Fig. 6. Spliced alignment of a rice EST (GenBank Accession 2317492) with an *Arabidopsis thaliana* protein similar to shaggy related protein kinase (GenBank Accession AAB61055). All predicted introns are correct. Exons 1, 4, and 5 contain only short stretches of identities, which cause other algorithms based on mostly sequence similarity scoring to miss such exons (discussed in the text). Strong splice site scores at the exon borders nonetheless push this alignment to the optimal score for sahmtD.

think that the combined scoring for good splice sites and sequence matching is at the core of these advances. Indeed, when the algorithm was re-run with all potential donor and acceptor sites given a generic score (data not shown), then the spliced alignment failed to recognize the strong acceptor site of intron 4 in Table 1 that was previously part of the optimal alignment (Figure 4).

Inclusion of sophisticated rules for splice junctions into the alignment algorithm was suggested by Florea *et al.* (1998) as a possible extension for their sim4 program. To further test whether our implementation of such extension yields practical benefits we ran GeneSeqer and sim4 on a set of 50 *Arabidopsis thaliana* genes of known exon/intron structure with an EST database consisting of more than 45 000 rice ESTs from GenBank. Eight of these genes gave significant GeneSeqer alignments comprising a total of 84 predicted exons, only two of which proved wrong. Figure 6 gives an exemplary output. Note that none of exons 1, 4, and 5 contain runs of identity longer than eight nucleotides. All these exons are missed by sim4 with default parameters (matching word size $W = 12$; values of $W = 10$ and smaller give a display of disjoint fragments instead of exon/intron structure). EST_GENOME and GAP2 detect also exon 4 but still miss exons 1 and 5.

The above examples illustrates the potential use of the algorithm for comparing closely related genes. The template cDNA may derive from another member of the same gene family, or from a homologous locus in a different species. As long as the sequences have not diverged by substantial insertions and deletions, the spliced alignment will work fine and help predict gene structure. Because the divergence at the nucleotide level is generally much larger than on the amino acid level, in general it will be more promising to make the spliced alignment of the genomic DNA directly with a protein template by maximizing the similarity of the predicted amino acid translation with the protein template (Birney *et al.*, 1996; Gelfand *et al.*, 1996; Huang and Zhang, 1996). An extension of our algorithm that accommodates this task is presented elsewhere (Usuka and Brendel, 2000).

Acknowledgments

V.B. was supported in part by NSF grant #9872657. J.U. was supported in part by NSF grant #9734893. The authors thank J. Kleffe and T. Wu for critical reading of the manuscript.

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. *Nucl. Acids Res.*, **25**, 3389–3402.

Birney,E., Thompson,J.D. and Gibson,T.J. (1996) Pair-wise 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. and Kleffe,J. (1998) Prediction of locally optimal splicesites 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. and Walbot,V. (1998) Intron recognition in plants. In Bailey-Serres,J. and Gallie,D.R. (eds), *A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants* American Society of Plant Physiology, Rockville, MD, pp. 20–28.

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

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. and 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. and 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.

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. *Nucl. Acids Res.*, **24**, 3439–3452.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices fromprotein blocks. *Proc. Natl. Acad. Sci. USA*, **89**, 10915–10919.

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

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

Karlin,S., Weinstock,G. and Brendel,V. (1995) Bacterial classifications derived from RecA protein sequence comparisons. *J. Bacteriol.*, 6881–6893.

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

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

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

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.