*Plant Molecular Biology* **48:** 49–58, 2002.



# Computational modeling of gene structure in *Arabidopsis thaliana*

Volker Brendel[1,2,*] and Wei Zhu[1]
[1]*Department of Zoology & Genetics (*author for correspondence; e-mail vbrendel@iastate.edu) and* [2]*Department of Statistics, Iowa State University, Ames, IA 50010, USA*

*Key words:* EST analysis, gene prediction, spliced alignment

**Abstract**

Computational gene identification by sequence inspection remains a challenging problem. For a typical *Arabidopsis thaliana* gene with five exons, at least one of the exons is expected to have at least one of its borders predicted incorrectly by *ab initio* gene finding programs. More detailed analysis for individual genomic loci can often resolve the uncertainty on the basis of EST evidence or similarity to potential protein homologues. Such methods are part of the routine annotation process. However, because the EST and protein databases are constantly growing, in many cases original annotation must be re-evaluated, extended, and corrected on the basis of the latest evidence. The Arabidopsis Genome Initiative is undertaking this task on the whole-genome scale via its participating genome centers. The current *Arabidopsis* genome annotation provides an excellent starting point for assessing the protein repertoire of a flowering plant. More accurate whole-genome annotation will require the combination of high-throughput and individual gene experimental approaches and computational methods. The purpose of this article is to discuss tools available to an individual researcher to evaluate gene structure prediction for a particular locus.

## Introduction

Modern DNA sequencing technology has revolutionized genetic research. Not long ago, the classical approach of isolating and characterizing a particular mutant would have reached a climax in the cloning and sequencing of the affected gene. Individual groups of researchers would contribute to our overall understanding of an organism or more general molecular mechanisms through their detailed studies of a particular gene or set of genes. This 'one gene at a time' science has now been complemented by 'high-throughput' approaches that quickly generate vast amounts of data on a large number of genes or a whole genome. Sequencing of entire genomes is the primary example of this new science, typically conducted by large research centers coordinated by national and international consortia. The sequencing of the *Arabidopsis thaliana* genome was the result of one such effort, culminating with the announcement of the complete genome in December 2000 (Arabidopsis Genome Initiative, 2000). The scope of such projects necessitates industrial approaches to data accumulation and processing, relying to a large extent on robotics and computational methods. Furthermore, this industrial approach has consequences similar to the industrialization of manufacturing: the goods delivered are produced for the entire community, and the former close connection between the craftsperson and his or her products may be lost. For genome projects, those producing the sequence can, at least initially, present only a rough overview of the features of the genome because of the scale and speed of data accumulation. The detailed understanding of particular aspects of the genome will likely have to continue to rely on the 'one gene at a time' studies.

The primary task of genome annotation involves identification of gene locations and precise gene structure in terms of promoter elements, transcription signals, exon/intron boundaries, and the translation product (or possibly multiple products in case of alternative transcription start or pre-mRNA processing sites). In the context of the discussion above, the annotation task can be seen as involving two stages. The first stage

is large-scale annotation, produced as the sequencing progresses and submitted to the community along with the publication of the genome sequence. For *Arabidopsis*, a total of about 25 500 protein-coding genes have been annotated in the five chromosomes (Arabidopsis Genome Initiative, 2000). Necessarily, a large number of these annotations are tentative and refer to hypothetical proteins or putative homologues. Thus, the second stage of annotation involves successive re-evaluation, extension, and correction of the annotation, removing many tentative assignments on the basis of novel experimental evidence.

The purpose of this article is to review options for the 'one gene at a time' biologist who wants to use the genome information for his or her detailed studies of particular genes. In this case, he or she cannot rely solely on the supplied genome annotation, which may well be incomplete or outdated. Instead, one must evaluate the sequences from scratch, using all particular information currently on hand, as, for example, EST evidence or potential protein homologues. We first review the principles of three prominent *ab initio* gene prediction programs for *Arabidopsis*, then discuss similarity-based prediction methods ('spliced alignment'), and lastly elaborate specific examples of evaluation of particular loci. The computational resources discussed in this article are summarized in Table 1.

## *Ab initio* algorithms for gene finding

A large number of gene finding algorithms have been developed that produce species-specific gene structure predictions on genomic DNA without explicit comparisons to cDNAs or protein sequences. The success of these methods depends on the applicability of extrapolation of sequence features gleaned from prior training on known gene structures. The principles of many such programs are eloquently reviewed by Claverie (1997). Recently, Pavy *et al.* (1999) evaluated programs in common use for *Arabidopsis* genome annotation and found GeneMark.hmm (Lukashin and Borodovsky, 1998) to be the most accurate program. Also in wide use are GENSCAN (Burge and Karlin, 1997) and GlimmerM (Salzberg *et al.*, 1999). All three programs are based on hidden Markov models. GENSCAN is built as an explicit state duration hidden Markov model. The algorithm explicitly scores for transcriptional and translational signals. Sequence composition is modeled by fifth-order Markov models, fitted according to exon phase and average C+G composition. GeneMark.hmm implements a similar model, although the details have not been described. GlimmerM uses dynamic programming to determine high-scoring combinations of coding exons. Exon/intron boundaries are determined from species-specific second-order Markov chain models, and exons are scored by fitting 3-periodic interpolated Markov models. On a large test set of validated multigene contigs, Pavy *et al.* (1999) reported exon level sensitivity and specificity of about 0.8 with the best *ab initio* programs. A common approach for whole-genome annotation is to increase the reliability of prediction by using the consensus prediction of a number of gene prediction algorithms. The combination of GeneMark.hmm, GENSCAN, and MZEF (Zhang, 1998) led to 97% exon level specificity on the Pavy *et al*. set, albeit, with sensitivity down to 33% (Pavy *et al*., 1999). At the whole-gene level, predicted models were found more often wrong than correct (Pavy *et al*., 1999). The main problem occurred with correct prediction of the proper gene boundaries. On balance, the *ab initio* programs are highly successful with respect to an initial annotation that can serve as a starting point for refined analysis using methods discussed in the next section, but such additional analysis remains necessary if whole-gene-level annotation accuracy is required.

## Spliced alignment

Currently the most successful and direct method for gene identification in genomic DNA relies on cDNA sequencing with subsequent sequence alignment to the corresponding genomic DNA region. Because complete cDNA sequencing can be time-consuming and costly, high-throughput EST (expressed sequence tag) sequencing has become the practical alternative to whole-genome sequencing efforts. The publicly available EST collections (GenBank dbEST, http://www.ncbi.nlm.nih.gov/dbEST/) range in size from over 3.5 million entries for human to several thousands for more than 40 other species. Efficient data mining of this resource requires fast and accurate algorithms to screen an appropriate EST collection for matches against a query genomic DNA input.

The alignment of ESTs (or complete cDNAs) to eukaryotic genomic DNA typically involves long gaps corresponding to the intervening sequences that are spliced from the pre-mRNA transcript. In the absence

*Table 1.* Some resources for computational gene structure prediction in *Arabidopsis thaliana*.

| Program | Web site | Reference |
|---|---|---|
| Ab inito *prediction* | | |
| GeneMark.hmm | http://dixie.biology.gatech.edu/GeneMark/eukhmm.cgi | Lukashin and Borodovsky, 1998 |
| GENSCAN | http://genes.mit.edu/GENSCAN.html | Burge and Karlin, 1997 |
| GlimmerM | http://www.tigr.org/tdb/glimmerm/glmr_form.html | Salzberg *et al.* |
| *Spliced alignment*: | | |
| GeneSeqer | http://bioinformatics.iastate.edu/bioinformatics2go/gs.cgi | Usuka and Brendel, 2000; Usuka *et al.*, 2000 |
| NAP | http://bioinformatics.iastate.edu/aat/aat.html | Huang and Zhang, 1996; Huang *et al.*, 1997 |
| PROCRUSTES | http://www-hto.usc.edu/software/procrustes/qpn.html | Gelfand *et al.*, 1996 |
| Sim4 | http://globin.cse.psu.edu/globin/html/docs/sim4.html | Florea *et al.*, 1998 |

of sequencing errors, alignment of a cognate EST to its genomic DNA source is straightforward, and a general alignment tool such as BLASTN (Altschul *et al*., 1997) would suffice in principle. Because EST sequences are generally less reliable, specialized algorithms also take into account consensus splice site sequences to identify introns correctly even in the presence of mismatches and insertions/deletions in the alignment. The sim4 program (Florea *et al*., 1998) implements an efficient algorithm for such alignments under the restriction of gap-free matching in presumed exons. Introns are identified by adjusting the ends of consecutive 'exon cores' (consistently ordered, close, high-scoring gap-free alignment blocks) to match the consensus 5′- and 3′-splice site signals GT and AG, respectively (or the complementary dinucleotides CT and AC).

The recent GeneSeqer algorithm (Usuka *et al*., 2000) implements a full dynamic programming approach to derive the optimal score and spliced alignment. The within-exon alignment may contain insertions and deletions, and potential splice sites are differentially scored according to independent splice site prediction methods. Consideration of predicted splice site strength was shown to improve the performance of the algorithm in the case of imperfect sequence matching (as a result of sequencing errors or alignment of non-cognate, but homologous ESTs). The power of such 'spliced alignment' with protein (rather than cDNA) targets was first demonstrated by Gelfand *et al*. with their PROCRUSTES program (Gelfand *et al*., 1996) and by Huang *et al*. with their AAT software (Huang and Zhang, 1996; Huang *et al*., 1997). The GeneSeqer algorithm was also extended to alignment of protein sequences with genomic DNA by maximizing similarity of the inferred translation product with the target protein (Usuka and Brendel, 2000).

## Case studies

The individual *Arabidopsis* researcher interested in a particular gene or gene family has unprecedented resources because of the completed sequencing of the *Arabidopsis* genome. In principle, each gene can now be uniquely identified on the chromosomes and studied in its genomic context. Because the genome annotation is as yet incomplete, the initial part of such individual research essentially involves re-annotation of the particular loci of interest. The published database annotation will provide a good starting point, but it may not have been updated since the database entry was originally submitted and thus it may be outdated or incomplete. The current *ab initio* gene prediction programs provide a second resource for such re-annotation. But if one is interested in particular loci, knowing that the average exon prediction accuracy of these programs is about 80% is of little comfort. For a five-exon predicted gene structure, one may suspect that one of the exons is incorrectly predicted – but which one? Or maybe this particular prediction is accurate above or below average. Thus, as a third resource, one must look at the latest evidence provided by more recently submitted matching ESTs or potential protein homologues that may not have been available at the time that the original annotation was performed. This additional evidence may not always solve the entire annotation problem, but may at least substantiate or refute some of the predicted exons.

We discuss three typical examples drawn from the very well annotated 1.9 Mb *A. thaliana* chromosome

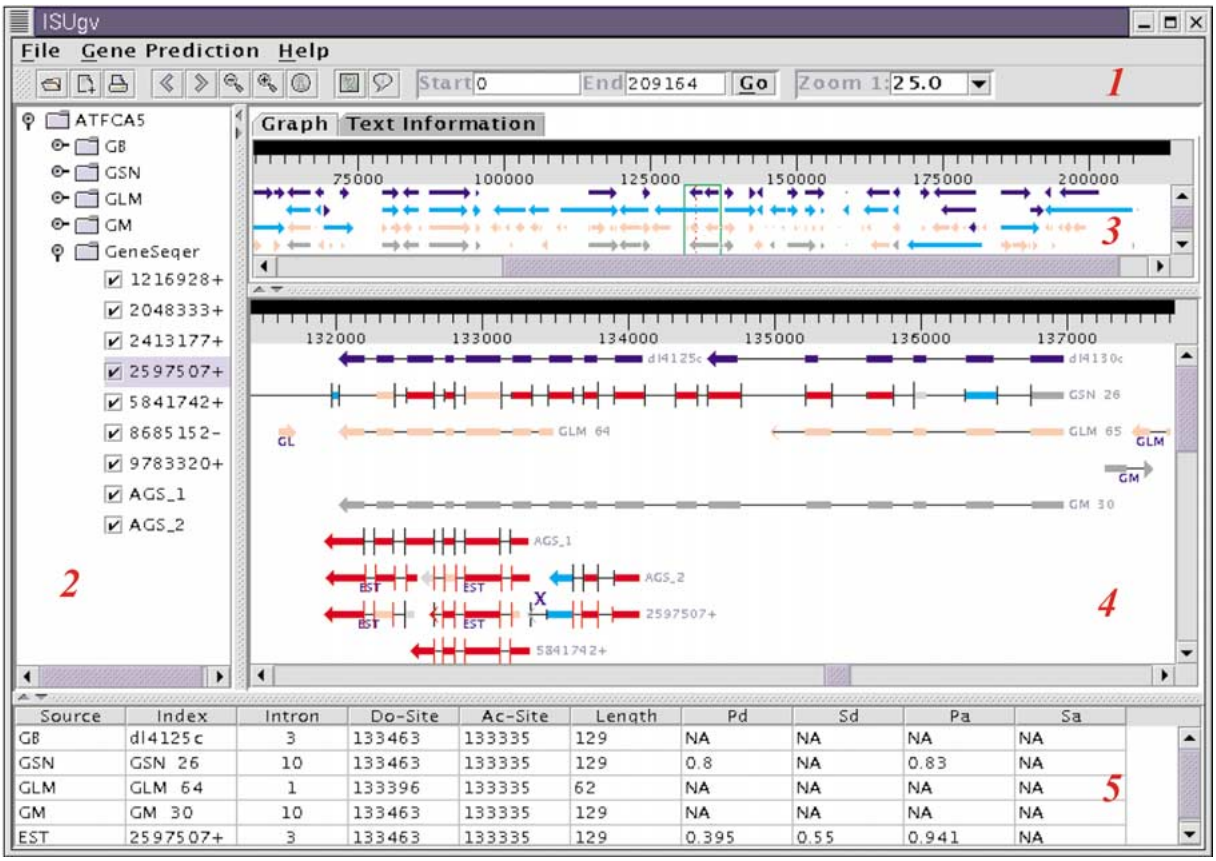


*Figure 1.* Genome annotation for a segment of *A. thaliana* chromosome 4 (GenBank locus ATFCA5, accession Z97340) based on *ab initio* gene structure prediction programs and spliced alignment of ESTs. Results are displayed with ISUgv, a Java tool for visualization of gene structure annotation and prediction (Zhu and Brendel, unpublished). The display is divided into five regions. 1, Toolbar. 2, Annotation List Tree (ALT) panel. The checked boxes correspond to the GenBank GI identifiers of aligned ESTs. A '+' following the GI identifier indicates alignment of the strand corresponding to the GenBank entry, whereas a '−' indicates alignment of the complementary strand. AGS, Alternative Gene Structures, represent the consensus of overlapping ESTs, after removal of more tentative exon predictions. Details will be presented elsewhere. 3, Annotation Overview (AO) panel. Annotated and predicted genes are represented by arrows from 5′ to 3′ extent of the coding region. Color scheme: GenBank (GB), blue; GENSCAN (GSN), cyan; GlimmerM (GLM), pink; GeneMark.hmm (GM), gray. The vertical green lines delineate the region of the input sequence analyzed in detail in the Annotation Scalable View (ASV) panel. 4, the color scheme in the ASV panel is the same as in the AO panel, except that exon quality scores assigned by GENSCAN and GeneSeqer are color-coded. For both programs, the quality scores are normalized to a maximal value of 1.0. Exons are represented by colored boxes as follows: red, score >0.9; pink, score >0.8; cyan, score >0.7; light gray, score >0.6; gray, otherwise. Introns are shown as horizontal lines connecting the exon boxes. Splice site scores given by GENSCAN and GeneSeqer are indicated by vertical lines of proportional lengths flanking the introns. 5, Text Data Overview (TDO) panel. This panel tabulates details of the (predicted) exon or intron marked by the blue cross in the ASV panel. Pd, donor site score; Sd, similarity score for the donor site flanking the 50 nucleotide exon region; Pa, acceptor site score; Sa, similarity score for the acceptor site flanking the 50 nucleotide exon region. The evidence of seven overlapping EST spliced alignment supports the GenBank annotation for dl4125c. The EST-derived annotation agrees with the GeneMark.hmm exon assignments in this region, but the GeneMark.hmm prediction extends 5′ into the dl14130c region.

4 region originally described by Bevan *et al.* (1998, coordinates 7.0–8.9 Mb on the chromosome). The examples illustrate several possibilities that arise when comparing the given annotation (in this case, the existing but out-dated GenBank annotation) or the *ab initio* predicted gene annotation with evidence from spliced threading. The alignment of one or several more recent ESTs may provide evidence for the correctness of the given gene annotation, it may suggest re-assignment of exon and intron boundaries, or it may indicate a novel gene annotation in a previously not annotated region. The examples argue for ongoing annotation efforts that reflect current resources, including better annotation tools, vastly increased EST collections, and larger protein repositories.

*New EST evidence confirms the original gene annotation*

Figure 1 gives an example of supporting EST evidence displayed by the ISUgv genome annotation viewer (Zhu and Brendel, unpublished). The example derives from the 130–137 kb region of GenBank locus ATFCA5 (accession Z97340). The GenBank annotation according to Bevan *et al*. (1998) indicates two genes in this region, dl4125c and dl4130c. The aggregate of seven overlapping ESTs confirms the dl4125c exon/intron assignments. Interestingly, the GeneSeqer alignment for EST GenBank index (GI) 2597507 predicts the third intron (133 463 to 133 335) on the basis of a short, weakly matching 3′-most exon segment (133 334 to 133 318). In this case, the strong acceptor site score at 133 335 (score 0.94 on a scale of 0 to 1) drives the optimal alignment to this solution, and the 10-nucleotide overlap with the central ESTs GI:5841742 and GI:1216928 results in the consensus gene prediction consistent with the dl4125c annotation. In contrast, the predictions from both GENSCAN and GeneMark.hmm additionally combine several exons of the upstream dl4130c annotated gene with dl4125c into a single-ene prediction (the GENSCAN gene model also extends considerably in the 3′ direction with five additional exons up to position 126 113). No ESTs match dl4130c, and no protein homologues map to this region. It is possible that all matching ESTs derive from the 3′ end of a long transcript originating in the dl4130c region. Alternatively, the lack of ESTs for dl4130c may reflect the low abundance of distinct transcripts from a second gene. Without such extra evidence, one cannot distinguish the possibilities for the N-terminal exon assignments. Compared to GeneMark.hmm and GENSCAN, GlimmerM appears to optimize for smaller gene models. Here, the GlimmerM model conformed to the downstream six exons of dl4125c, but failed to identify the upstream exons revealed by EST GI:2597507.

*New EST evidence is in conflict with earlier gene annotation*

A second case is displayed in Figure 2. EST evidence in the 190–200 kb region of GenBank locus ATFCA0 (accession Z97335) suggests a gene structure quite different from the original GenBank annotation, but confirms introns 1 and 6–9 of the GeneMark.hmm prediction. There are three ESTs (GIs 8698471, 8682984, 8695751) that contradict the prediction of the third intron of the GeneMark.hmm gene structure. All of these ESTs give perfect alignment over their entire length (intron-flanking alignment displayed in the upper panel in Figure 2) and match uniquely to this location in the genome. Open reading frames are stopped in all three frames in the upstream exon for the predicted direction of transcription. Thus, a likely interpretation is that these ESTs correspond to the 3′ end of a transcript and that the predicted intron is in the 3′-untranslated region of such transcript. Because the *ab initio* gene prediction programs predict coding exons only, this intron could not have been predicted by any of these programs. On the basis of the EST evidence, we consider the GeneMark.hmm prediction of exons 1–3 most likely correct, with the exception of the GeneMark.hmm predicted 3′ end of the third exon, which should be replaced by the assignment given by the EST alignment. Note that EST GI:8689419 supports the GeneMark.hmm and GlimmerM annotated start codon (perfect matching extending 17 bases upstream of the ATG) and contradicts the GenBank annotation and GENSCAN prediction. Interestingly, ESTs GI:8721769 (sampled from root tissue) and GI:9786549 (sampled from developing seed) are in conflict with respect to the first intron assignment. It is possible that the seed EST reflects inefficient or alternative splicing of the transcript.

The second gene in this region is supported by a single EST (GI:935155). A BLASTX database search revealed similarity of the EST-derived translation product to the *Arabidopsis* 22 kDa peroxisomal membrane protein GI:11282649, encoded at about 2.2 Mb on chromosome 4. Spliced alignment of this protein sequence with the genomic DNA identifies this locus as a homologue. The protein sequence alignment is shown in Figure 3. Both proteins have seven exons, intron positions are conserved, and strong similarity extends over all exons. Compared to this standard, the GlimmerM model correctly predicts exons 1–5 and 7, misses exon 6, and predicts an extra exon in intron 3.

This example demonstrates how the latest available evidence must be considered to give a reliable annotation. The derived annotation of two genes, one encoding a peroxisomal protein and the other a protein of unknown function, is much different from the GenBank annotation, citing a hypothetical protein of 12 exons with weak similarity to mouse laminin chain B1 precursor extending from coordinate 199 892 to 191 737. Correct and wrong annotations both lead to entries in the public protein databases. Because the protein databases are in turn used for gene prediction, the urgent need for more accurate database annota-

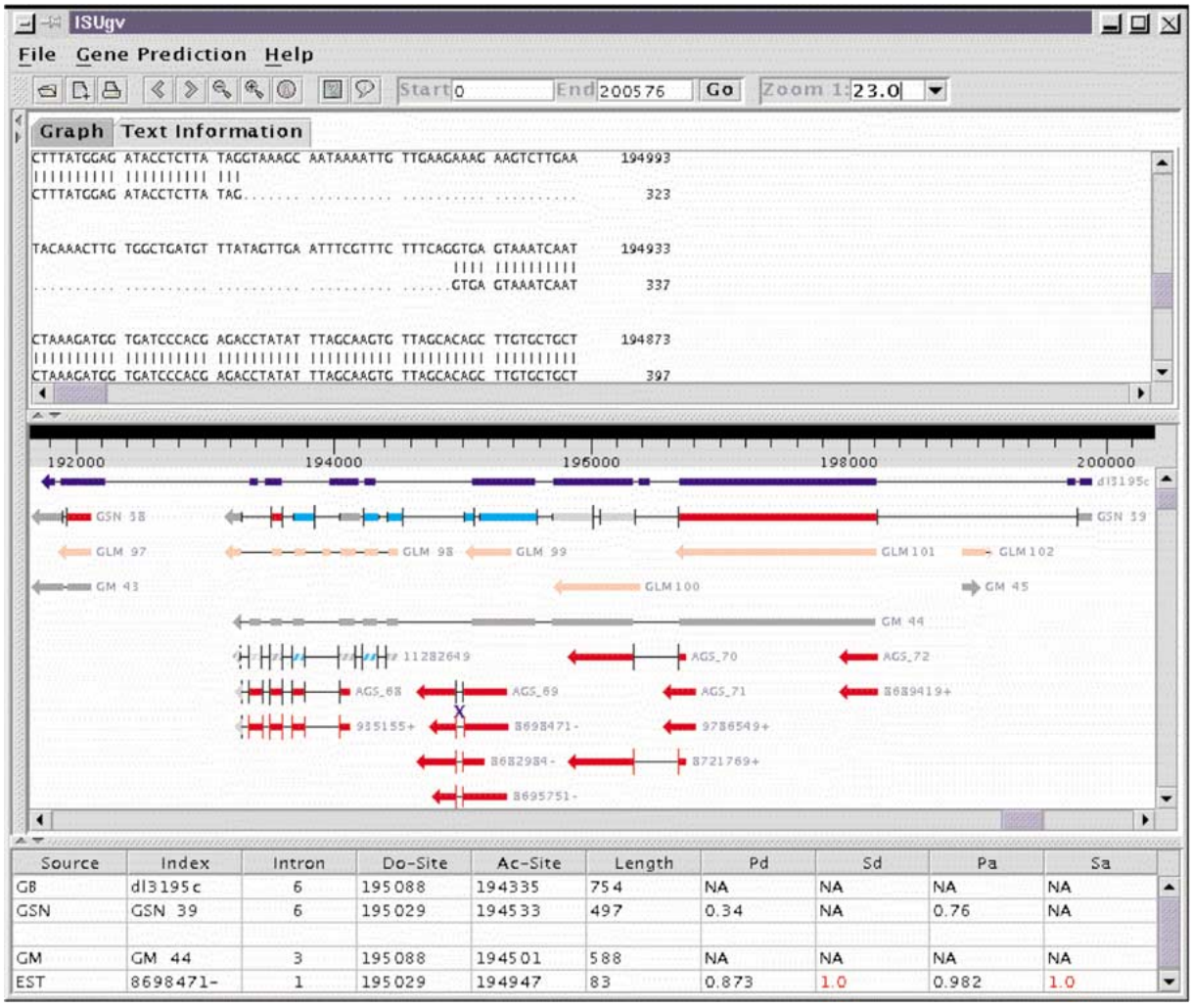


*Figure 2.* EST and protein spliced alignment contradict GenBank annotation. The displayed region corresponds to 191–200 kb of GenBank LOCUS ATFCA0 (accession Z97335). Symbols are as in Figure 1. The AOV panel is toggled to display text corresponding to the alignment in the region selected by the blue cross in the ASV panel. The alignment is supported by three different ESTs. Neither GenBank annotation nor any of the three *ab initio* programs predict the displayed intron (GENSCAN predicts the donor site but not the acceptor site). Further analysis suggests two genes in this region, one encoding a peroxisomal protein homologous to the pmb22 peroxisomal protein (GenBank GI:11282649), and the second in the downstream region encoding a protein of unknown function; see text for discussion.

```
Pmb22       1  MGSSPPKKTTLQRYLSQLQQHPLRTK=AITAGVLSGVSDVVSQKLSGIQKIQLRRVLLKY=I
Predicted   1  --MSDLAKDAWRKYLIQLQAHPLRTK=AITAGVLAGCSDATAQKISGVKRICFRRLLLLM=L

Pmb22      61  FAGGFLGPAGHFFHTYLDKFFKGKKDTQTVAKK=VILEQLTLSPLNHLLFMIYYGVVIE=RT
Predicted  59  YGFAIGGPFGHFFHKLMDTIFKGKKGNSTVAKK=VLLEQLTSSPWNNFLFMSYYGIVVE=GR

Pmb22     121  PWTLVRERIKKTYPTVQLTAWT=FFPVVGWINYKYVPLHFRVILHSLVAFFW=GIFLTLRAR
Predicted 119  PWKLVKHKLGKDYPTIQLTAWK=FWPIVGWVNYGYVPLQFRVLFSSFVASCW=SIFLNLRAR

Pmb22     181  SMTLALAKAK  190
Predicted 179  SPVIKNA---  185
```

*Figure 3.* Protein sequence alignment of the *A. thaliana* 22 kDa peroxisomal membrane protein (pmb22, GI:11282649, 2.2 Mb region of chromosome 4) with the predicted protein in the 194 kb region of ATFCA0 (Figure 2). Intron positions are indicated by '—'. Identical residues are on black background, and conservative substitutions are on gray background.

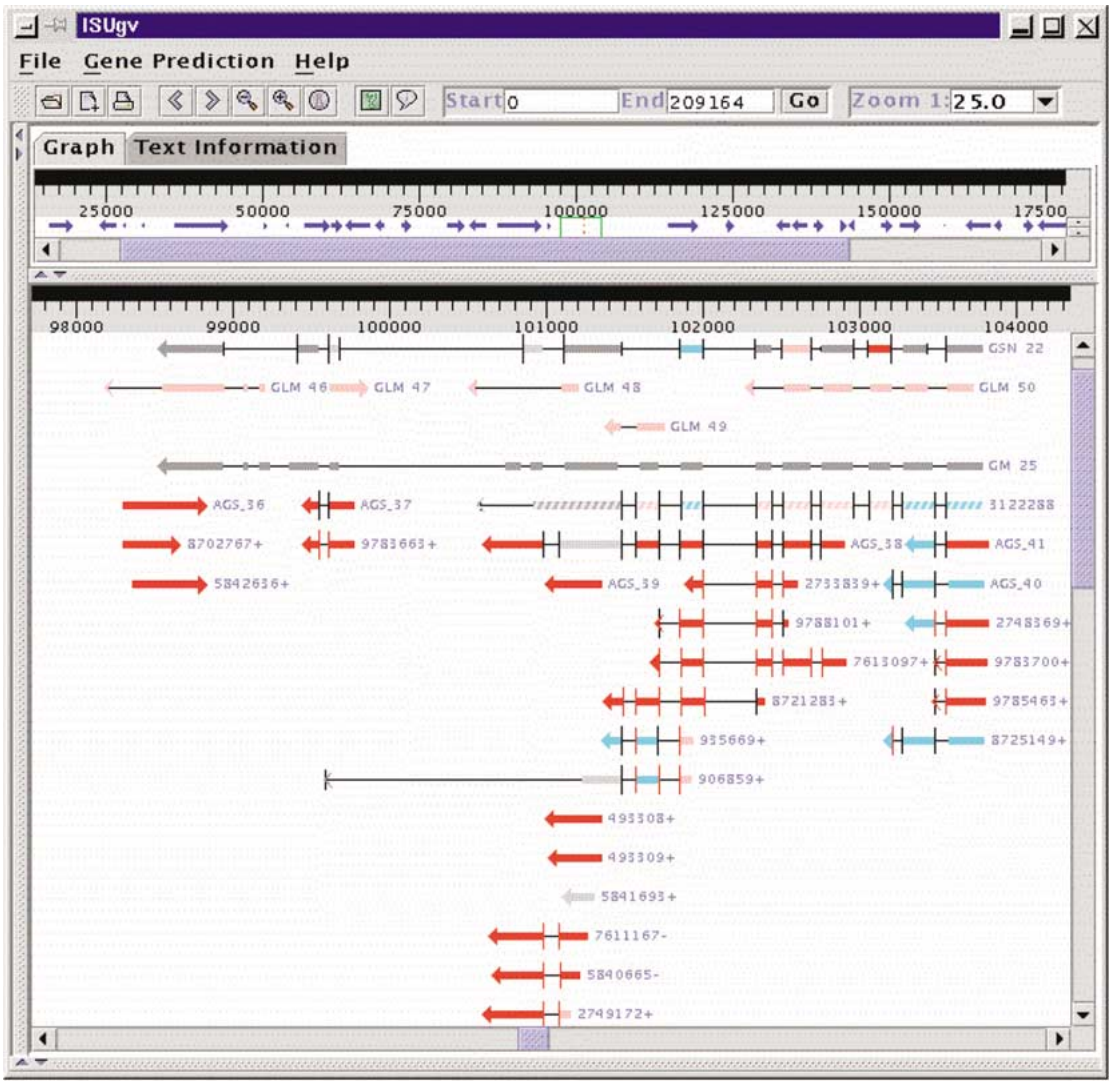


*Figure 4.* Gene discovery by ESTs. Two EST clusters align with the *A. thaliana* BAC GenBank ATFCA5 in the 100–104 kb region. Spliced alignment with the importin $\alpha$-1 subunit (GenBank GI:3122288) suggests a 10-exon gene structure consistent with the EST evidence. Details of the alignments are discussed in the text.

tion is clear. A conservative approach, adopted by many genome centers, is to use only experimentally proven gene products for genome annotation based on similarity. However, this approach may be too conservative because similarity on the peptide level between two inferred translation products predicted from different loci is most parsimoniously explained as resulting from correct prediction of two members of a gene family (see Brendel and Kleffe, 1998 and Usuka *et al.*, 2000 for examples and further discussion). In fact, gene structure prediction based on assignment of conserved regions as exons and variable regions as introns in comparisons of genomic DNA from distantly related but syntenic plant species may be the most powerful method for identifying unknown genes (Bennetzen, 2000).

*New EST evidence leads to novel gene annotation*

Figure 4 gives an example of gene discovery by ESTs. Four clusters of ESTs match significantly in an unannotated region of GenBank locus ATFCA5. GENSCAN and GeneMark.hmm both predict one gene in this region, GlimmerM predicts five. Figure 5 shows the EST alignments in the 99–104 kb region displayed by the GeneSeqer web server. A convenient feature of

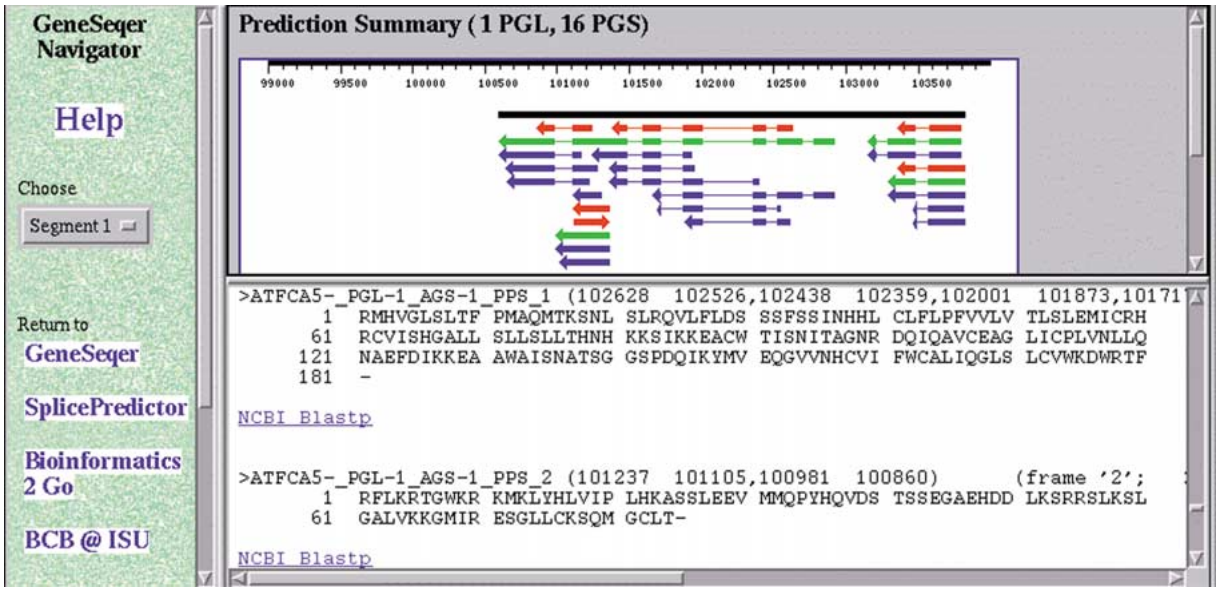


*Figure 5.* Application of the GeneSeqer web service. The server returns the EST alignments (upper panel, blue) that are displayed in more detail in Figure 4. The consensus gene structure prediction (green) allows two long open reading frames (red) in the 100–103 kb region. The corresponding translation product is shown in the lower panel. A BLASTP query with predicted protein fragments revealed the similarity to the importin $\alpha$ protein that resulted in the gene prediction shown in Figure 4.

this interface is that the EST-predicted consensus gene structures are scanned for long open reading frames and the corresponding peptide sequences are linked as queries to NCBI BLASTP. In this example, a 180 amino acid predicted protein fragment showed strong similarity to importin $\alpha$ proteins from a number of different animal and plant species. The spliced alignment of the *Arabidopsis chromosome* 3 encoded importin $\alpha$ (GI:3122288; chromosomal coordinates 2120 569 to 2123 844) is shown in Figure 4. (To complicate matters further, GI:3122288 was derived from a cDNA with several differences to the chromosomal sequence. Translation of the genomic DNA results in a translation stop at the end of the penultimate exon, consistent with sequences of importin $\alpha$ proteins from tomato, *Drosophila*, and mouse.). This alignment was initially puzzling because it suggests extension of the open reading frame beyond the N-terminal stop indicated in Figure 5. Closer sequence inspection resolved this puzzle as resulting from a likely error in the genomic sequence: all four ESTs – GI:2733839, GI:9788101, GI:8721283, and GI:7613097 – match perfectly to the genomic DNA except for a single nucleotide insertion of a G at position 102 360 in the ATFCA5 sequence. The insertion leads to the frameshift that shortens the open reading frame. This example illustrates the additional power of spliced alignment algorithms that do not require continuous open reading frames and thus can detect frameshift errors or polymorphisms. At the predicted 3′ end of the gene, the five strongly matching ESTs split into two groups of two and three ESTs. The second group appears to define an additional intron in the 3′-untranslated region for some of the transcripts of this gene.

A powerful feature of the GeneSeqer spliced alignment method is that the concurrent optimization for sequence similarity and splice site scores allows effective use of heterologous ESTs in gene structure prediction. Here, ESTs GI:935669, GI:906859, and GI:8725149 derive from the paralogous importin $\alpha$ gene on chromosome 3, yet predict four introns consistent with the cognate ESTs.

## Perspective

In their recent careful evaluation of gene prediction programs for *Arabidopsis*, Pavy *et al*. (1999) showed that even the best method, GeneMark.hmm (Lukashin and Borodovsky, 1998) found the correct gene model in only 67 of 168 known genes analyzed. Prediction of mammalian gene structure appears similarly challenging (Rogic *et al*., 2001). These studies strongly suggest that our theoretical understanding of both transcription and RNA-processing signals remains incomplete.

Predictions based on the consensus of several different methods increases the specificity of the predictions but at the cost of much reduced sensitivity (Pavy *et al*., 1999). The fact that different programs perform better or worse for particular genes indicates that the current models for gene prediction are too general and might be improved if the models were trained on specific subsets of genes. Some improvement was in fact observed for *Arabidopsis* after separating two classes of genes on the basis of codon usage (Mathé *et al*., 2000).

Here we have demonstrated by examination of a number of typical examples that additional analysis for a particular locus may significantly increase the odds of correct gene prediction relative to the average performance of *ab initio* gene prediction methods. In particular, spliced alignment with ESTs or potential protein homologs can provide substantial evidence in favor of one or another exon/intron assignment. Current methods for mammalian genome annotation seek to automate some of these additional analyses (Kan *et al*., 2001; Yeh *et al*., 2001). Driven by these needs, genome annotation facilitates a transition of modern molecular biology. Increasingly, high-throughput and individual gene experimental approaches as well as computational methods converge to increase our detailed understanding of complex biological processes. Within the next quarter century, we anticipate an interplay of theoretical and experimental research in biology similar to the synergistic pursuit of theoretical and experimental physics in the 20th century. For a recent example, Shoemaker *et al*. (2001) used microarray technology to experimentally validate and refine computational gene predictions for human chromosome 22. Similar steps for better gene prediction in *Arabidopsis* are reviewed elsewhere (Cho and Walbot, 2001).

With continuing increases in DNA sequencing capacities, much insight may be expected from comparative sequence analysis. Studies of genomic microcolinearity in plants that have diverged over five million years or more suggests that only genic regions are highly conserved, thus providing another means of identifying genes (Bennetzen, 2000). The next generation of biologists will be well trained in bioinformatics as well as genomics approaches and be able to view biological problems from a much wider, multifaceted perspective. Such expanded view will constitute a much better approximation to biological reality than afforded within current paradigms.

## Acknowledgements

V.B. was supported in part by NSF grant DBI-9872657. W.Z. was supported by a J. Cornette Fellowship from the Bioinformatics and Computational Biology graduate program at Iowa State University. The authors wish to thank Virginia Walbot for critical comments on 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.

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–813.

Bennetzen, J.L. 2000. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. Plant Cell 12: 1021–1029.

Bevan, M. *et al.* 1998. Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. Nature 391: 485–488.

Brendel, V. and Kleffe, J. 1998. Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in *Arabidopsis thaliana* genomic DNA. Nucl. Acids Res. 26: 4749–4757.

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

Cho, Y. and Walbot, V. 2001. Computational methods for gene annotation: the *Arabidopsis* genome. Curr. Opin. Biotechnol. 12: 126–130.

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.

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

Huang, X. and Zhang, J. 1996. Methods for comparing a DNA sequence with a protein sequence. Comput. Appl. Biosci. 12: 497–506.

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

Lukashin, A.V. and Borodovsky, M. 1998. GeneMark.hmm: new solutions for gene finding. Nucl. Acids Res. 26: 1107–1115.

Mathé, C., Déhais, P., Pavy, N., Rombauts, S., Van Montagu, M. and Rouzé, P. 2000. Gene prediction and gene classes in *Arabidopsis thaliana*. J. Biotechnol. 78: 293–299.

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

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

Salzberg, S.L., Pertea, M., Delcher, A.L., Gardner, M.J. and Tettelin, H. 1999. Interpolated Markov models for eukaryotic gene finding. Genomics 59: 24–31.

Shoemaker, D.D. et al. 2001. Experimental annotation of the human genome using microarray technology. Nature 409: 922–927.

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

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.

Yeh, R.-F., Lim, L.P. and Burge, C.B. 2001. Computational inference of homologous gene structures in the human genome. Genome Res. 11: 803–816.

Zhang, M.Q. 1998. Identification of protein coding regions in *Arabidopsis thaliana* genome based on quadratic discriminant analysis. Plant Mol. Biol. 37: 803–806.