BIOINFORMATICS APPLICATIONS NOTE Vol. 18 no. 5 2002
Pages 761–762

# Gene structure identification with MyGV using cDNA evidence and protein homologs to improve ab initio *predictions*

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

[1]*Department of Zoology and Genetics and* [2]*Department of Statistics, Iowa State University, Ames, IA 50011, USA*

Received on October 2, 2001; revised on December 4, 2001; accepted on December 7, 2001

## ABSTRACT

**Summary:** MyGV is an application to visualize (potentially genome-scale) gene structure annotation and prediction. The output of any external gene prediction program can be easily converted to a generalized format for input into MyGV. The application displays all input simultaneously in graphical representation, with a toggle option for a text-based view. Zooming capabilities allow detailed comparisons for specific genome locations. The tool is particularly helpful for refinement of *ab initio* predicted gene structures by spliced alignment with cDNA or protein homologs.

**Availability:** The program was written in Java and is freely available to non-commercial users by electronic download from http://bioinformatics.iastate.edu/bioinformatics2go/MyGV.

**Contact:** vbrendel@iastate.edu

Accurate and comprehensive genome annotation remains the foremost challenge in the post-sequencing genome era. The recent recognition of extensive alternative splicing of mammalian genes underscores the importance of the annotation task, because in most cases transcript isoforms are expected to reflect functional diversity. The theoretical foundations of gene structure are presently only partially understood. Exon prediction methods are largely based on statistical approaches. Different programs often give conflicting predictions. Large collections of Expressed Sequence Tags (ESTs) and, increasingly, full-length cDNAs can provide evidence for certain exons and thus improve *ab initio* gene prediction methods (Kan *et al.*, 2001; Gemünd *et al.*, 2001). Additionally, spliced alignment with selected protein targets can often identify the gene structure of a homologous gene locus. In practice, successful gene annotation relies on careful comparison of multiple sources of prediction. MyGV was developed in response to such needs on the premise that such comparisons are most efficiently evaluated by a combination of graphical representation and analytical detail (see also Harris, 1997; Kent and Zahler, 2000; Rutherford *et al.*, 2000).

## INPUT: SEQUENCE FILES AND EXTERNAL PROGRAM RESULTS

MyGV accepts as sequence input representation of a DNA molecule in common GenBank or FASTA file format. The program provides an annotation overview panel in which CDS feature entries of GenBank files are graphically represented by solid arrows extending from first to last exon and pointing in the direction of transcription. Detailed gene structure is displayed in a second, scalable view panel. Currently, no other annotation features in the sequence input files are being used. Additional input consists of formatted output of gene prediction programs, which is similarly displayed. This input is generated by piping the output of external programs (run on the same sequence input) through format converters. The current release includes format converters for Fgenesh (FGH; Salamov and Solovyev, 2000), GeneMark.hmm (GM; Lukashin and Borodovsky, 1998), GeneSeqer (Usuka *et al.*, 2000; Usuka and Brendel, 2000), GENSCAN (GSN; Burge and Karlin, 1997), and GlimmerM (Salzberg *et al.*, 1999), but others can easily be written by the user according to need.

## MyGV DISPLAY

Figure 1 illustrates the application with analysis of a segment of the *Drosophila melanogaster* genome. The input sequence file was GenBank AE002638, representing about 4.9 Mb at the terminal tip of the left arm of chromosome 2. The detailed view covers a small region including the annotated genes CG15386, CG7074, and CG7082 and represents the output of the *ab initio* programs GSN, FGH, and GM as well as the EST spliced alignments generated by GeneSeqer (EST and AGS). The display is divided into five regions:

(1) Toolbar. This section of the display controls a number of pull-down menus that are used to open and close files, execute gene prediction programs, select the zoom level, and similar functions.

© Oxford University Press 2002

(2) Annotation List Tree (ALT) panel. All displayed items are listed with checkboxes that allow selection and de-selection of individual items.

(3) Annotation Overview (AO) panel. Annotated and predicted genes are represented by arrows from 5′- to 3′-extent of the coding region. Different programs are distinguished by the color scheme, e.g. GenBank, blue, GSN, cyan. The vertical green lines delineate the region of the input sequence analyzed in detail in the

(4) Annotation Scalable View (ASV) panel. The color scheme in the ASV panel is the same as in the AO panel, except that exon quality scores are color-coded whenever assigned by a program. Introns are shown as horizontal lines connecting the exon boxes. Vertical lines of proportional lengths flanking the introns indicate splice site scores given by GSN and GeneSeqer.

(5) Text Data Overview (TDO) panel. This panel tabulates details of the (predicted) exon or intron marked by a blue cross in the ASV panel. Normalized similarity and splice site scores generated by the corresponding programs are displayed whenever applicable.

The AO panel can be toggled to 'text' which will display the program output of the selected item in the ASV panel. The text information can be edited in the AO panel, with an update function redrawing the graphical display in the ASV panel accordingly. This function is useful for manual refinement of the computed gene predictions. Other features of the program are described in the software documentation.

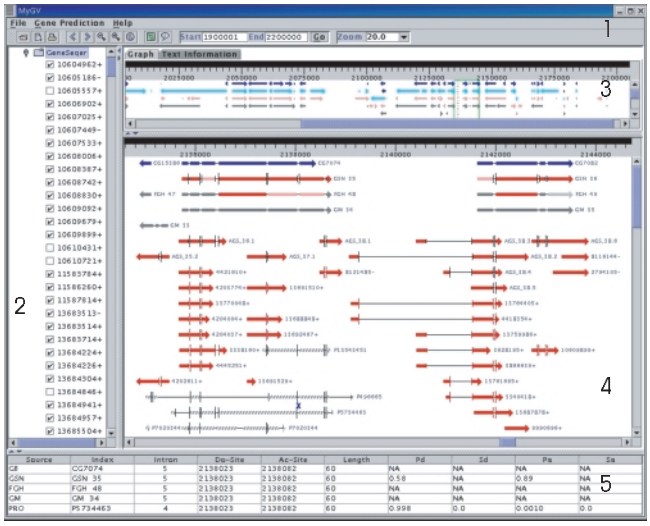


**Fig. 1.** Genome annotation for a segment of *D. melanogaster* chromosome 2 (GenBank AE002638) based on *ab initio* gene structure prediction programs and spliced alignment of ESTs and proteins. The five numbered MyGV display regions are described in the text. Gene prediction results are shown in panel 4. A single EST (GenBank GI: 4202611) confirms intron 1 but not intron 2 of the GM prediction (GM 33) upstream of the GenBank annotated CG15386 gene. GSN, FGH, and GM give a consistent gene prediction that agrees with and extends CG7074. Introns 1–4 and 6 are confirmed by multiple EST evidence, summarized by GeneSeqer as AGS_36.1, AGS_37.1, and AGS_38.1. The blue cross in panel 4 selects intron 4 of the GeneSeqer spliced alignment with a putative *Schizosaccharomyces pombe* protein (GI: 5734463) for display in panel 5. EST evidence confirms the coding exon assignments for CG7082 and supports four isoforms of the 5′-terminal untranslated region (AGS_38.2, AGS_38.3, AGS_38.4, and AGS_38.5), differing in the location of an upstream exon.

## ACKNOWLEDGEMENTS

This work was supported in part by NIH grant 5R44HG01850-03. W.Z. is grateful to the Bioinformatics and Computational Biology graduate program at Iowa State University for a J.Cornette Fellowship during the first half of 2001.

## REFERENCES

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

Gemünd,C., Ramu,C., Altenberg-Greulich,B. and Gibson,T.J. (2001) Gene2EST: a BLAST2 server for searching Expressed Sequence Tag (EST) databases with eukaryotic gene-sized queries. *Nucleic Acids Res.*, **29**, 1272–1277.

Harris,N.L. (1997) Genotator: a workbench for sequence annotation. *Genome Res.*, **7**, 754–762.

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.

Kent,W.J. and Zahler,A.M. (2000) The Intronerator: exploring introns and alternative splicing in *Caenorhabditis elegans*. *Nucleic Acids Res.*, **28**, 91–93.

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

Rutherford,K., Parkhill,J., Crook,J., Horsnell,T., Rice,P., Rajandream,M.-A. and Barrell,B. (2000) Artemis: sequence visualization and annotation.

Salamov,A.A. and Solovyev,V.V. (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res.*, **10**, 516–522.

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.

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

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