Short Specialist Review

Gene structure prediction in plant genomes

Volker Brendel
Iowa State University, Ames, IA, USA

1. Introduction

In eukaryotes, the presence of intervening sequences (introns) within most genes makes the problem of computational gene structure prediction distinct from (and harder than) the same problem in prokaryotes. However, even among eukaryotes, the problem is varied beyond the basic need for species-specific training of algorithm parameters. For example, introns are rare in yeast genes, and, where found, they typically bear highly conserved signature patterns for splicing signals that make identification easy. The purpose of this short review is to discuss approaches to gene identification in plants. There are two reasons why this task is somewhat different from the same task in animals. The first reason is biological. Introns in plant genes have a similar length distribution as exons (Brendel *et al*., 1998). In particular, there appear to be none of the very long introns that confound gene prediction in vertebrates. The second reason is pragmatic. Expressed Sequence Tag (EST) sequencing and whole-genome as well as genome-survey sequencing is in progress for dozens of plant species, all of which appear closely enough related that gene identification in any given species can very efficiently leverage annotation and sequence information from all the other species. In practice, a combination of ab initio gene prediction, spliced alignment, and expert annotation will, in most cases, produce highly reliable protein-coding gene structures. Identification of noncoding RNA genes such as *snoRNA* and *miRNA* genes might prove a much harder problem (Chen *et al*., 2003; Reinhart *et al*., 2002; Llave *et al*., 2002; Marker *et al*., 2002).

1.1. Ab initio algorithms for plant gene finding

Most of the experience with computational gene structure prediction in plants was derived from annotation and reannotation of the Arabidopsis genome. The initial annotation (Arabidopsis Genome Initiative, 2000) was achieved with specially trained gene-finding programs such as GENSCAN (Burge and Karlin, 1997), GeneMark.hmm (Lukashin and Borodovsky, 1998), and GlimmerA (Salzberg *et al*., 1999). These programs derive their predictions “ab initio”: the input consists only of the genomic DNA to be annotated, and gene structures are derived on the

basis of models for exons, introns, and transcription and splicing signals (reviewed in Mathé *et al*., 2002). The underlying methods, also called "intrinsic methods", are similar between the various programs; however, combinations of the programs have been shown to perform better than any single program (Murakami and Takagi, 1998; Pavy *et al*., 1999). The performance of these methods has been thoroughly tested (Pavy *et al*., 1999). At best, exon level sensitivity and specificity were found around the 80% mark, and fewer than half of the predicted gene structures were completely correct. While unsatisfactory in terms of current understanding of gene structure and transcriptional and posttranscriptional control, in practice, such approximate annotations are highly useful because they immediately provide a glimpse of the gene space. Biologists can often find their genes of interest by a simple BLAST search (Altschul *et al*., 1997) against the annotated genes and then refine the respective gene models by more detailed analysis. On a large scale, reannotation efforts (Wortman *et al*., 2003) rely on approaches discussed in the next two sections.

1.2. Spliced alignment with cDNAs and ESTs

Full-length cDNA sequencing provides the most direct way for gene structure identification, because by definition the cDNA residues derive from exons in the genomic sequence. The problem of threading a cDNA sequence back into its corresponding genomic DNA is known as *spliced alignment*. The alignment task is straightforward in the absence of significant differences between the cDNA and the genomic DNA (some level of mismatch would simply result from sequencing errors or differences between DNA sources). A number of programs are available to this end, including dds/gap2 (Huang *et al*., 1997), sim4 (Florea *et al*., 1998), GeneSeqer (Usuka *et al*., 2000), and BLAT (Kent, 2002). In practice, large-scale EST-sequencing projects often provide the first view of a plant's gene space. If not accompanied by a genome-sequencing project, the ESTs are typically clustered and assembled on the basis of sequence similarity to derive nonredundant sets of putative transcripts ("unigenes"; e.g., Quackenbush *et al*., 2001; Dong *et al*., 2004). Otherwise, inherent problems of EST clustering (limited power to detect overlap and to distinguish duplicated genes, chimeric clones) can be largely eliminated by spliced alignment of ESTs onto the genome and subsequent assembly of these spliced alignments to complete gene structures (Zhu *et al*., 2003; Haas *et al*., 2003). In many cases, incorporation of splice site prediction into the derivation of optimal spliced alignments allows the use of nonnative ESTs for gene structure prediction (Brendel *et al*., 2004). This approach appears very promising, given the large combined EST resources across all plant species.

1.3. Spliced alignment with proteins

Although gene order is not necessarily conserved between even closely related plant species (Tarchini *et al*., 2000; Fu and Dooner, 2002), the proteomes of different plants are estimated to be more than 90% conserved (Bennetzen, 2000).

This suggests a successful complementary strategy to gene identification based on spliced threading of protein sequences into genomic DNA thought to encode a homologous protein. The implicit conservation of the reading frame across predicted coding exons makes the algorithms for such alignment tasks more complex than in the cDNA/EST spliced alignment case. Programs such as AAT (Huang *et al*., 1997) and GeneSeqer (Usuka and Brendel, 2000) can be used, provided close-enough homologs can be identified as probes. In practice, good results can be achieved with a combination of cDNA/EST and protein spliced alignments (Schlueter *et al*., 2003). This derives from the fact that EST coverage is mostly from the 5′- and 3′-ends of a gene, regions that are the most difficult to predict by protein spliced alignment because of natural variation in N- and C-termini of proteins, whereas the EST-sparse internal exons are best predicted with homologous proteins (Usuka and Brendel, 2000). The limitations of this approach are, first, that this method obviously does not apply to genes that are specific to a given species; and, second, that poor choices of the protein probe will lead to unreliable predictions. The latter problem could be exacerbated if the probes themselves are erroneously predicted proteins, thus potentially propagating annotation errors (Gilks *et al*., 2002).

1.4. Combined approaches and user-contributed annotations

The problem of automated computational gene structure prediction remains challenging. In particular, it is difficult to weight different sources of prediction and evidence to derive a consistent prediction. A human expert can often easily enough distinguish solid and plausible evidence from the spurious or mistaken. Examples of this include an unmasked poly-A tail in an EST sequence matched erroneously to an adenine-rich segment of genomic DNA in a spliced alignment, a missed U12-type intron, or a chance alignment of a low-complexity protein region. But large-scale applications of computational gene structure prediction will be hard-pressed to find program parameter settings that work for the special cases. It is to be hoped that wide community participation in expert-based gene structure annotation and editing will result in highly reliable annotations for the fully sequenced plant genomes, providing a foundation for further development of computational approaches (allowing model training on error-free data) and of homology-based annotation of subsequently sequenced genomes.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Bennetzen JL (2000) Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. *The Plant Cell*, **12**, 1021–1029.

Brendel V, Carle-Urioste JC and Walbot V (1998) Intron recognition in plants. In *A Look Beyond Transcription: Mechanisms Determining mRNA Stability and Translation in Plants*, Bailey-Serres J and Gallie DR (Eds.), American Society of Plant Physiologists: Rockville, pp. 20–28.

Brendel V, Xing L and Zhu W (2004) Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics*, **20**, 1157–1169.

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

Chen CL, Liang D, Zhou H, Zhuo M, Chen YQ and Qu LH (2003) The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. *Nucleic Acids Research*, **31**, 2601–2613.

Dong Q, Schlueter SD and Brendel V (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Research*, **32**, D354–D359.

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

Fu H and Dooner HK (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 9573–9578.

Gilks WR, Audit B, De Angelis D, Tsoka S and Ouzounis CA (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.

Haas B, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, *et al* (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, **31**, 5654–5666.

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

Kent WJ (2002) BLAT–the BLAST-like alignment tool. *Genome Research*, **12**, 656–664.

Llave C, Kasschau KD, Rector MA and Carrington JC (2002) Endogenous and silencing-associated small RNAs in plants. *The Plant Cell*, **14**, 1605–1619.

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

Marker C, Zemann A, Terhorst T, Kiefmann M, Kastenmayer JP, Green P, Bachellerie JP, Brosius J and Hüttenhofer A (2002) Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Current Biology*, **12**, 2002–2013.

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

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

Pavy N, Rombauts S, Déhais P, Mathé C, Ramana DV, 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.

Quackenbush J, Cho J, Lee D, Liang F, Holt I, Karamycheva S, Parvizi B, Pertea G, Sultana R and White J (2001) The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Research*, **29**, 159–164.

Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B and Bartel DP (2002) MicroRNAs in plants. *Genes and Development*, **16**, 1616–1626. Erratum in: *Genes and Development*, **16**, 2313.

Salzberg SL, Pertea M, Delcher AL, Gardner MJ and Tettelin H (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics*, **59**, 24–31.

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

Tarchini R, Biddle P, Wineland R, Tingey S and Rafalski A (2000) The complete sequence of 340 kb of DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. *The Plant Cell*, **12**, 381–391.

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. *Journal of Molecular Biology*, **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.

Wortman J, Haas BJ, Hannick LI, Smith RK Jr, Maiti R, Ronning CM, Chan AP, Yu C, Ayele M, Whitelaw CA, *et al* (2003) Annotation of the *Arabidopsis* genome. *Plant Physiology*, **132**, 461–468.

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