EvidentialGene: Perfect Genes Constructed from Gigabases of RNA

Gilbert, Don
Indiana University, Biology, Bloomington, IN 47405; gilbertd@indiana.edu

Gene construction, not prediction. The decade of gene prediction is over; gene construction from transcriptome sequence now surpasses predictions for biological validity. To paraphrase a recent paper: ".. over half the computational gene predictions were imperfect, with missing exons, false exon predictions, wrong intron boundaries, fused and fragmented genes". Gene construction from RNA is not without similar and other problems, but these can be reliably resolved with attention to detail and a river of sequence data.

Improved gene annotations are essential for their uses in genome biology. How many gene comparison studies have significant artifacts of quality? Most, probably: in a recent review of gene orthology, "genome annotation emerged as the largest single influencer, affecting up to 30%" of the discrepancies among orthology assessments (1).

Our results with *The. cacao* tree gene annotation finds protein homology discrepancy between competing gene sets to be the same range as among related plant species. A gene average difference of 6-13 aminos separates two cacao gene sets, while an average 10 aminos separate each of cacao, poplar, castor bean and grape gene sets. Our recent EvidentialGene construction for *Nasonia* jewel wasp improved its gene orthology to the highest in the hymenoptera clade [2]. This wasp and pea aphid gene sets produced with EvidentialGene surpass the RefSeq set produced by NCBI for same loci, and built with same gene evidence.

Transcript assemblies can be more accurate than predictions, but but contain similar errors and take effort to resolve conflicts. RNA data quality sets limits, and software struggles at both ends of the data river. Sensible data reduction is a major gene construction task, where 10^9 RNA reads are assembled to 10^6 of competing transcripts, and those filtered with multiple criteria for the closest approach to 10^4.5 biological genes. Species expressed genes often differ from those mapped from other species. Transcript evidence can give truer answers for phylogeny, gene function, etc., than when one protein is fit to many species.

Assembling a transcriptome can be more costly than genome assembly, and less useful without one. Alternate transcripts, tandem duplicate paralogs and fused genes, and bad guesses need to be resolved with reference to a genome backbone.

Gene construction software and methods continue to improve, but are imperfect. A current best strategy employed with EvidentialGene uses several gene modelling and assembly methods, extracting the best of their many results. This is consistent with recent results of others, pertaining to transcriptome assembly [3,4]. Rough edges need smoothing: predictor models and transcript assemblies each have qualities the other lacks, for coding sequences and sequence signals, gene holes and mash-ups. Multiple lines of

gene evidence can score the quality of competing gene constructions to select a best, if not yet perfect, gene set.

Genome/transcriptome informatics now uses computing clusters "wastefully" to produce best results, in that many complete gene prediction and transcript assembly runs are used, varying parameters and data slices, to produce a superset of models that contain the most accurate subset. Genomics computing is embarrassingly parallel in large part, where an entire genome/gene data set is subdivided, by genes, locations and other ways, as many times as computing resources permit. A current typical set of 1 billion transcriptome reads plus 200,000 homologous proteins for a eukaryote can be analyzed once with around 1000 cpu hours, but the multiple analyses needed to obtain that perfect subset of genes will be 10+ times higher.

A critical component of this approach to perfecting gene sets is the ability to select biologically valid models from a large superset that includes fragments, fusions and complete fabrications by the gene assembly/prediction components. EvidentialGene software for this uses extensive evidence annotation and maximization. It relies on deterministic evidence scoring, giving same result for one locus or 50,000. It is not a majority vote among alternates, as some others, but the single best scoring model is chosen. The algorithm for evidence scoring attempts to match expert choices, using base-level and gene model quality metrics. Currently scoring is fine-tuned to each particular species and evidence sets. This differs from peer methods of Glean, Evigan, EvidenceModeler and others [5,6] for its deterministic evidence scoring, detailed per gene annotations, and single-best model/locus approach.

EvidentialGene construction steps
http://arthropods.eugenes.org/EvidentialGene/about/genes2annot-diagram2.jpg

1. produce several predictions and transcript assembly sets with quality models, from observation that no single set is best for all loci, but enough variants often have best among them.
2. annotate gene models with all evidence, including gene model quality (transcript introns, exons, prot. homology, transposons, ...)
3. score models from weighted sum of evidence.
4. remove models below minimum score
5. select from overlapped models/locus the model with highest score including metrics of joined genes, so that highest score is not nesc. longest model.
7. Re-evaluate results
8. Iterate over 3..7 with alternate scoring to refine a final best set.

EvidentialGene components and notes (Sep 2011):
AUGUSTUS: http://augustus.gobics.de/ : make several gene sets with alternate training/evidence mixes. Training and preparing evidence sets require iterative work.
GMAP-GSNAP: http://research-pub.gene.com/gmap/ : use for long and short rna genome mapping; recent updates are best. Accuracy for splice read mapping is higher than other such methods by recent comparisons. Allows for SNP-compliant mapping.

PASA: http://pasa.sourceforge.net/ : can use to assemble EST + RNA-seq pre-assemblies, correct final gene predictions, add alternate transcripts from evidence. Assembly errors can be significant, with current cufflinks, velvet and others preferred now to PASA.

Cufflinks: http://cufflinks.cbcb.umd.edu/ : mapped rna-seq assembly, use version 0.8 rather than newer. In my hands, versions later than 0.8 drop 3/4 of valid rna assemblies, due to over agressive rejection from low read score metrics. Assemblies from 0.8 and newer can be combined to select best.

Velvet/Oases: http://www.ebi.ac.uk/~zerbino/velvet/ : denovo rna-seq assembly. Make several assemblies w/ alternate kmer parameters and pick best, use long + short reads, especially paired end, for best results. See also Trinity/Inchworm, Abyss, Mira, etc. - moving target for which is best, try several.

exonerate: http://www.ebi.ac.uk/~guy/exonerate/ : homology protein gene models from protein2genome:bestfit refinement after tblasn

NCBI BLAST: blast.ncbi.nlm.nih.gov/ : tblastn/blastx, blastp, blastn, ...

Evigene scripts : http://arthropods.eugenes.org/EvidentialGene/evigene/

Summary of EvidentialGene sets and transcript evidence

	Daphnia_magna		Daphnia_pulex		Pea_aphid	
Evid.	Nevd	Dmag11	Nevd	Dplx10	Nevd	Aphid11
EST	26mb	0.822	12mb	0.884	36mb	0.813
Pro	27mb	0.819	21mb	0.831	27mb	0.786
RNA	32mb	0.684	42mb	0.677	55mb	0.428
Intron	89k	0.937	68k	0.963	127k	0.690

Coding span	31mb	48mb	42mb
Exon span	53mb	71mb	74mb
Gene count	34614	47712*	32967
Genome size	131mb*	227mb	541mb

Daphnia magna gene set, 2011mar, is preliminary on an incomplete genome assembly. Daphnia pulex 2010 gene set is beta status, remains to be updated. D. pulex gene counts include fragments, transposons and otherwise poor gene models. Nevd for EST, Pro, RNA is total span of non-overlapping reads or alignments, but count for good, unique Intron sites from spliced reads.

Nasonia jewel wasp, 2012jan

	Evigene	RefSeq2	OGS1.2
Introns	97%	90%	85%
EST coverage	72%	67%	51%
RNA assembly	63%	36%	29%
Homolog score	679	635	--
Coding span	28 mb	10 mb	28 mb
Exon span	45 mb	24 mb	29 mb
Gene count	24560	12989	18941
Alternate tr.	7839	1475	7
Genome size	295 mb		

Theobroma cacao chocolate tree, 2012

Evidence	Evigene	Cirad
Introns	91%	82%
EST coverage	57%	48%
RNA assm.	67%	32%
Homolog score	549	522
Coding bases	35 mb	34 mb
Exon bases	54 mb	48 mb
Gene count	29283	29484
Alternate tr.	14920	0
Genome size	347 mb	

Introns : match to EST/RNA spliced introns
EST coverage : overlap with EST exons
RNA assembly : >=66% equivalence with RNA/EST assemblies
Homolog score : blastp bitscore average for found homologs

References:

1. Trachana K, Larsson TA, Powell S, Chen W-H, Doerks T, Muller J, and Bork P. 2011. Orthology prediction methods: A quality assessment using curated protein families Bioessays 33: 769–780, 2011 WILEY Periodicals, Inc. DOI 10.1002/bies.201100062

2. Gilbert D, 2010. EvidentialGene: Evidence Directed Gene predictions for Eukaryotes. http://arthropods.eugenes.org/EvidentialGene/ and "Perfect(ing) Arthropod Genes with Next Gen Informatics". 4th Arthropod Genomics Symposium. http://arthropods.eugenes.org/EvidentialGene/about/PerfectGenes2010.pdf

3. Zhao et al. 2011.

Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study
BMC Bioinformatics, 12(Suppl 14):S2 http://www.biomedcentral.com/1471-2105/12/S14/S2

4. Kumar and Blaxter, 2010.
Comparing de novo assemblers for 454 transcriptome data
BMC Genomics 11:571 http://www.biomedcentral.com/1471-2164/11/571

5. Liu Q, Mackey A, Roos D, Pereira F, 2008.
Evigan: A Hidden Variable Model for Integrating Gene Evidence for Eukaryotic Gene Prediction
Bioinformatics 2008, 24(5):597-605.

6. Haas, B. et al. 2008
Automated eukaryotic gene structure annotation using
EVidenceModeler and the Program to Assemble Spliced Alignments.
Genome Biology, 9:R7 doi:10.1186/gb-2008-9-1-r7.