Annual Report for Period: 06/2011 - 06/2012
Principal Investigator: Gilbert, Donald G.
XSEDE Award: MCB10014, Genome Informatics for animals and plants
related NSF Award: 0640462, Shared genome database informatics and cyberinfrastructure

Progress Summary

The primary research focus is developing for and using shared cyberinfrastructure for assembly, annotation and comparative analysis of new organism genomes. This fourth year effort has refined the genome/gene annotation software pipeline *EvidentialGene*, introduced the in prior annual report. In conjunction with this software engineering, it has been used to produce well-annotated gene sets for *Daphnia magna* waterflea, pea aphid, *Nasonia* jewel wasp, and *Theobroma cacao* chocolate tree. This work uses the NSF-funded shared cyberinfrastructure of Teragrid/XSEDE and National Center for Genome Analysis Support.

EvidentialGene Methods.

Gene construction software and methods continue to improve, but are imperfect. Gene construction from transcript sequence can surpass predictions for biological validity. To paraphrase others, *".. over half the gene predictions were imperfect, with missing exons, false exons, wrong intron ends, fused and fragmented genes"*, with respect to a 2006 gene set compared to a 2011 transcript assembly. But, gene assembly from RNA has similar and other problems. Perfecting this means using all of the best data and tools, plus evidence quality tests, to build accurate genes.

A current best strategy employed with EvidentialGene uses several gene modeling and assembly methods, extracting the best of their many results. This is consistent with recent results of others, pertaining to transcriptome assembly [1,2,3]. 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 will score the quality of competing gene constructions to select a best gene set.

Genome/transcriptome informatics now uses computing clusters "wastefully" to produce best results. Many complete gene prediction and transcript assembly sets are generated, varying parameters and data slices, to produce a superset of models that contain an accurate subset. A current set of 5 billion transcriptome reads plus 200,000 homologous proteins for a eukaryote can be analyzed once with around 3000 cpu hours, but multiple analyses needed to obtain that accurate subset of genes will be 10+ times higher. Current transcriptome data sets want more compute cores and big memory systems to fully assemble with current methods. This project is developing methods that reduce such memory requirements but they remain in the 100GB to 500GB range.

A critical component of this approach is the ability to select biologically valid models from a large superset that includes fragments, fusions and fabrications of the gene assembly and 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.

EvidentialGene construction steps

1. produce several predictions and transcript assembly sets with quality models. No single method/set is best at all loci, variants often have best among them.
2. Annotate models with all evidence, esp. gene model qualities (transcript introns, exons, homology, transposons)
3. Score models from weighted sum of evidence.
4. Remove models below minimum evidence score
5. Select from overlapped models/locus the highest score, include fusion metrics (longest is not always best)
6. Evaluate results, genome-wide averages and with inspection (map views of errors)
7. Iterate 3..7 with alternate scoring to refine final best set.

Evidence evaluation criteria for genes are, in part, protein homology, coding/non-coding ratio, RNA read coverage, read intron recovery, and transcript assembly equivalence.

EvidentialGene Results.

Gene sets for *Acyrthosiphon* pea aphid and *Nasonia* jewel wasp, built with EvidentialGene incorporating RNA assembly and gene prediction directed by evidence, prove superior on several evidence metrics to those of NCBI RefSeq, that were built using same available evidence.

Pea aphid v2, 2011 June

Evidence	Evigene	RefSeq	OGS1
Introns	70%	68%	52%
EST cover	79%	69%	49%
RNA assembl	49%	43%	27%
Protein score	76%	46%	47%

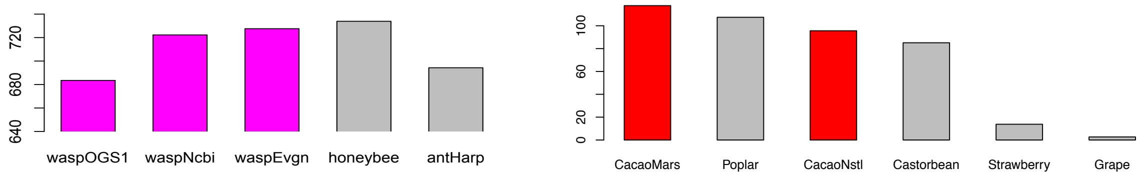
Nasonia jewel wasp v2, 2012 January

Evidence	Evigene	RefSeq	OGS1
Introns	97%	90%	85%
EST cover	72%	67%	51%
RNA assembl	63%	36%	29%
Homology	679	635	--

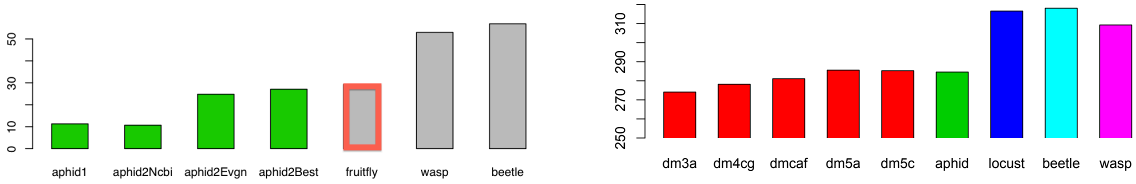
How many gene comparison studies have significant artifacts of quality? In a recent review of gene orthology, "genome annotation emerged as the largest single influencer, affecting up to 30%" of the discrepancies among orthology assessments [4]. Gene function, derived from orthology or by experiment, is sensitive to imperfections. Differential expression measures are muddled on imperfect genes. Many biology studies that use genome-wide constructed genes hinge on the gene quality.

Gene set quality affects species orthology rankings, as indicated in below graphs of the average homology scores. Gene sets produced here are compared with others of same species and related species. Our results with cacao tree find protein homology discrepancy between competing gene sets is in the same range as between related plant species. Our gene set for jewel wasp improved its orthology to the highest in the hymenoptera clade. The pea aphid Evigene set improved in this 1 year of work to same level as fruitfly genes improved in 10 years of work. This indicates mainly that gene evidence is increasing, but also that EvidentialGene is capable of using these new data effectively.

Wasp genes (left, purple) improve to level of honeybee, and cacao genes (right, red) improve above poplar tree.



Aphid genes reach fruitfly homology quality level in 1 year (left, green), versus 10 years of gradual improvement in fruitfly genes (right, red).



Outcomes of this project's annual work include the genome informatics software, available at http://arthropods.eugenes.org/EvidentialGene/. Gene sets have been provided to collaborating genome projects, and are either now publicly available or are in progress to that with the genome project. This pea aphid set is at the community genome database, http://www.aphidbase.com/ as ACYPI v2.1, also at http://arthropods.eugenes.org/EvidentialGene/pea_aphid2/ . The jewel wasp gene set is at http://arthropods.eugenes.org/EvidentialGene/nasonia/ and gene submission to NCBI GenBank is in progress. The cacao gene set is available to Phytozome, NCBI and the Cacao Genome project, with genome publication in preparation. The *Daphnia magna* gene set remains preliminary, pending a genome assembly update and addition of 5 billion RNA-Seq reads, on top of an already impressive 1.5 billion reads. Daphnia species gene numbers remain at top of animals, with extensive gene duplications, but these require extra analysis effort.

Discussion.

EvidentialGenes results are not perfect, but this approach appears to be working. A major remaining need is in tuning for problem cases. Expert inspection combined with evidence rescoring reduces these, but the last 10% require effort similar to the first 90%.

The data explosion from low cost, high throughput sequencers now surpasses abilities of software and shared cyberinfrastructure to fully and properly analyze these data. Transcriptome assembly and gene construction is now engineering rather than art/science; the methods work when properly applied to produce biologically valid genes. However, current software for this, shared hardware for data storage, network transfers, memory requirements, are all stressed by the large size of these data sets. Software in this area is improving on a monthly basis. Cyberinfrastructure resources are available and improving. But analyses of these data sets now require much effort in slicing and reducing data in creative ways to enable software and hardware to process, often with more failed attempts than successful ones, or with partial results that do not fully use the available data.

Biologists are capable of analyzing genomes and gene sets now without a large effort from expert bioinformaticians. The availability of genomics software, in shared-computing environments (e.g XSEDE) is growing, along with easier to use, web-enabled interfaces (e.g. Galaxy). Documentation is well enough developed that biology graduate students can and do learn requisite genome informatics methods, with some script programming for data processing, to produce genome annotations as part of their research. There remain aspects that need improvement in the software and cyberinfrastructure so that these biology students can turn their data into biologically valid models.

References:

1. Zhao *et al*. 2011. Optimizing de novo transcriptome assembly from short-read RNA-Seq data: a comparative study. BMC Bioinformatics, 12:S2. doi:10.1186/1471-2105-12-S14-S2
2. Schulz MH, DR Zerbino, M Vingron and E Birney, 2012. Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012 Apr 15;28(8):1086-92 PMID:22368243
3. Martin, JA and Zhong Wang, 2011. Next-generation transcriptome assembly. Nature Reviews | Genetics, doi:10.1038/nrg3068
4. 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. doi:10.1002/bies.201100062