A Genome Grid for Finding new Bug Genes

Don Gilbert, Indiana University, gilbertd@indiana.edu

Abstract

Accurate gene prediction and automated annotation is lagging behind needs for the rapidly increasing number of new genomes. Prediction tools are increasingly sophisticated and accurate. These draw on the range of available gene evidence and improved modeling of gene structures. Yet they are sensitive to the availability of genome data and expected structures. Detection of novel and diverged genes remains problematic. Even for species clades with a well-characterized model such as Drosophila, gene finding is an urgent task. Next-generation genomics technology, such as genome-wide expression technology, finds thousands of genes have been found with new methods. Combining bioinformatics with prediction tools for both new and well-characterized Arthropod genomes unovers 10% to 50% new species genes and diverged genes.

We envision Genome Grid as pipelines that many scientists can use, installed on NSF TeraGrid and other shared cyberinfrastructure, as part of a Generic Model Organism Database project (GMO, gmod.org/Genome_models). Middleware and science gateway methods to use computational grids for genome analyses are in development as an open source project. Genome community software is now available, at TerraGrid ta/communities/genomes, and for local use. These tools help genome analyses keep pace with rapid expansion of genomics information.

Contact: Don Gilbert, gilbertd@indiana.edu

Genome Grid Overview

Genome informatics still needs to provide biocurators with;
(1) effective use of randomly changing, growing complex genome data;
(2) rapid analyses of data from Next-Gen genome technology;
(3) frequent re-analyses encompassing expanded new evidence;
(4) easy access for many scientists to informatics methods

Clusters, grids and clouds of computers now provide infrastructure for these. Available genome processing tools can put these resources to use for the benefit of any scientist. The focus of this project is to enable a scientist or group with genome sequence and evidence (related proteins, ESTs, tile array expression, etc.) to analyze these in a rapid and effective manner, without large infrastructure cost.

Computational engineering of this effort is on middleware for parallelizing genome data and collating results on grid systems. Genome informatics is known for its rapidly changing, parallel, data-based paradigms, where genome sequence and annotations can be subdivided as desired for many single-run analyses on cluster or grid systems. Parallel results are then re-assembled at genome locations. Large volume data now emerging from next-generation genomics technologies such as genome-wide tile expression arrays and short-read sequencing are incorporated in analyses in a similar data parallelization.

Genome Grid Software

A basic component of this genome grid framework is genome partitioning and result assembly. It is based on the design and Perl scripts of EvidenceModeler [Haas et al. 2008]. Analyses are run in parallel on many genome parts. Results are collated to full genome data sets. Many genome informatics tools work well on genome parts.

Components of this package now available in include genome partitioning, grid job submission for standard genome applications, and re-assembling of GenBank/EST and GFF formats. Applications used include Augustus, NCBI Blast, EvidenceModeler, Exonerate, and SNAP. Planned enhancements include more genome applications, Ergatis workflow (O Iris and colleagues), BioMart for genome analysis data, and PASA EST assembly.

“Instant” results web access is planned via a TeraGrid science gateway. This will provide genome results through BLAST sequence searches, Gilibrow maps, and annotation test expression reports. This type of access helps immensely in assessing value, checking, and correcting errors.

Table 1. TerraGrid usage steps.

<table>
<thead>
<tr>
<th>Step</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preparation</td>
<td>One time</td>
</tr>
<tr>
<td>1. TerraGrid account</td>
<td></td>
</tr>
<tr>
<td>2. Establish certificates</td>
<td>Grid security, local workstation certificate</td>
</tr>
<tr>
<td>3. Locate bio software</td>
<td></td>
</tr>
<tr>
<td>Install applications (tg/community/genomes)</td>
<td></td>
</tr>
<tr>
<td>4. Repetitive runs</td>
<td></td>
</tr>
<tr>
<td>Collect, partition data</td>
<td>Copy to shared disk, Partition &amp; randomize</td>
</tr>
<tr>
<td>Parallel runs analyses</td>
<td>Run scripts, check errors, re-run as needed</td>
</tr>
<tr>
<td>Cellular results</td>
<td>Post-process to combine results from nodes</td>
</tr>
</tbody>
</table>

Basic steps in Table 1 for using TerraGrid for genome analyses are not complicated, but require learning for a new user. Web documentation is sufficient for those with cluster or grid experience. Data selection, preparation, transport to TerraGrid, and return of results can be automated with data grid and workflow tools. A TerraGrid science gateway provides genome project infrastructure to rapidly produce single or multi-species analyses, with standard bioinformatics tools. Genome grid applications are available at TerraGrid site in TG_COMMUNITY/genomes. Any US scientist can obtain a start-up allocation in two weeks to analyze genomes with these tools.

Table 2. Bug genomes analyzed on TerraGrid with genome grid methods.

<table>
<thead>
<tr>
<th>Genomes</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daphnia waterlilia</td>
<td>Full genome assembly, analysis, annotation and tile array analyses. Plans to repeat for many species of Daphnia.</td>
</tr>
<tr>
<td>Drosophila fruittella</td>
<td>12 fractilt species gene predictions, homology, plus genome tile expression analysis.</td>
</tr>
<tr>
<td>Ayltheroposipra pea aphid</td>
<td>Full genome analysis and annotation; plan tile array analyses</td>
</tr>
<tr>
<td>Nasomia jewel wap</td>
<td>Gene predictions; plan tile array analyses</td>
</tr>
<tr>
<td>Isobod tick</td>
<td>Comparative transcript analyses.</td>
</tr>
</tbody>
</table>

Dozens of Bug, or Arthropod, genomes have been analyzed with these methods (Table 2), with a variety of interesting outcomes (see below). The most recent genome analyzed, pea aphid, included PASA EST assembly, Arthropod proteome mapping with BLAST and exonerate, gene prediction with Augustus and SNAP, and gene annotation with RepProt and UniProt databases. The first pass analysis took 12 days. First analyses and new results were detailed in a refined set of analyses, a “complete” genome annotation, at the cost of one informatician’s half-time effort over six weeks. Results are at http://insects.eugenics.org/Drosophila/data/aprid/.

Interesting Genome Biology Results

Daphnia has lots of tandem genes, and gene finders make many mistakes with them. The same prediction errors occur in Drosophila and other genomes, but are less obvious.

Duplicate genes are frequent, and very near (1kb) tandem duplicates are especially common in Daphnia, exceeding the duplicate rich Cae. elegans. One aspect of genome biology that is difficult to model is a cluster of nearby duplicate genes. Nearby near-identical exons can confuse computational methods that use alignment, including BLAST, GeneWise and similar gene mappers that align a protein to find genes. -tmRNA prediction also can fail to dual genes especially different to nearby genes. The initial set of Daphnia gene predictions had many errors finding these, with 5,000 predicted genes spanning two or more distinct matches to the same protein.

Duplicate genes help find mistakes

Expressed genes are poorly found as homology with D. melanogaster declines. Novel genes are poorly predicted, as protein homology and prediction trained with Dmel will mislead. Figure 2 summarizes species group percentages for ESTs and duplicate genes that are missed by gene predictions. Most misses are these lacking Dmel homology.

These methods of gene duplicate detection have been applied to predictions for 12 Drosophila species genomes. It is one way to independently check predictions without reliance on comparison the reference species genomes. This example is taken from a case in the Dros. willistoni genome, where no single predictor correctly called all four Cytochrome P450 genes. However, among 13 predictors were cases of a true model for each gene.

Genome Grid, 2008.07

References


Figure 1. Tandem gene prediction errors.

Figure 2. Novel Drosophila species genes missed by prediction.

Figure 3. Gene prediction species clines: biology or computational artifact?

In Figure 3B, two gene predictions show a lower rate of tandem genes predicted for the D.melanogaster species. Other predictions show nocline, or a reverse cline comparable to that found for non-Dmel organism gene sets. The dilemma expressed in Figure 3B, of inconsistent predicted clines in duplicates, can be explained in large part by prediction errors, with results shown in Figure 4. This species-bias error is eliminated by training the predictors with same or near species gene data, as shown in Figure 3F for two gene calling methods (SNAP, Exonerate). The bar graphs of Figure 3 show gene counts for each of 10 species, arranged phylogenetically in hist colors from near-Dmel (red) to far-Dmel (yellow).

Genome tile expression finds novel genes

Genome grid methods have turned genome tile array expression to gene predictions, for Daphnia and Drosophila, finding many new genes.

Genes called from tile expression experiments find 5,000 to 10,000 new genes above the 30,000 predicted for Daphnia. Figure 4 shows one such new gene and tile evidence. The analysis approach combines genome prediction software (Augustin) with tile transcription evidence, much like EST evidence. A similar amount of total new gene expression for Dros. melanogaster was found by Manak et al. 2006.

Figure 4. Tissue expression finds missing genes. Daphnia example

What does tile expression uncover? Among new tile expression genes, 10% have protein homology, and 19% have EST support (25% have one of these). This is a beginning to understand novel genes. But why have they been missed by current gene prediction? Using many treatment groups from cell lines and development stages, the modelCODE project seeks detailed answers. A set of Drosophila melanogaster gene predictions have been produced using Affymetrix tile expression data (see Hsu et al. 2006), that include 33 treatment groups. There is high concordance (83%) between tile transcription fragments and predicted exons.