Bulk data files // TeraGrid uses for Genome Databases

GMOD meet, June 2006

Don Gilbert, gilbertd@indiana.edu
### Index of /genome

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td>30-Dec-2005 15:23</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Drosophila_melanogaster/</td>
<td>20-Mar-2004 14:22</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Drosophila_pseudobscura/</td>
<td>25-Dec-2005 22:10</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Saccharomyces_cerevisiae/</td>
<td>27-Dec-2005 21:27</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Genome Data Sets

In order to simplify the retrieval of common datasets, the Generic Model Organisms Database (GMOD) community has implemented a series of standard URLs. Each MOD (will|should) have an index page like this one describing the species and datasets that are available. Read more about the GMOD project at [www.gmod.org](http://www.gmod.org).

Use below links to choose among Sequence (FasA), Features (GFF), Summary tables and other bulk genome data. Select among species, feature type, chromosomes, data release versions.

**Available URLs**

<table>
<thead>
<tr>
<th>URL</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>/genome</td>
<td>Display this HTML-formatted index page that contains links to each of the species available through common URLs.</td>
</tr>
<tr>
<td>/genome/Binomial_name</td>
<td>An index page for species &quot;Binomial_name&quot;. This will be an HTML-format page containing links to each of the genome releases.</td>
</tr>
<tr>
<td>/genome/Saccharomyces_cerevisiae/</td>
<td>Leads to index for the named release. It should be an HTML-format page containing links to each of the data sets described below.</td>
</tr>
<tr>
<td>/genome/Binomial_name/release</td>
<td>Leads to an index of the most current release, symbolic link style.</td>
</tr>
<tr>
<td>/genome/Binomial_name/current/dna</td>
<td>Returns a FASTA file containing big DNA fragments (e.g. chromosomes). MIME type is application/x-fasta.</td>
</tr>
<tr>
<td>/genome/Saccharomyces_cerevisiae/current/dna</td>
<td></td>
</tr>
<tr>
<td>/genome/Binomial_name/current/mrna</td>
<td>Returns a FASTA file containing spliced mRNA transcript sequences. MIME type is application/x-fasta.</td>
</tr>
<tr>
<td>/genome/Saccharomyces_cerevisiae/current/mrna</td>
<td></td>
</tr>
<tr>
<td>/genome/Binomial_name/current/cdna</td>
<td>Returns a FASTA file containing non-coding RNA sequences. MIME type is application/x-fasta.</td>
</tr>
<tr>
<td>/genome/Saccharomyces_cerevisiae/current/cdna</td>
<td></td>
</tr>
<tr>
<td>/genome/Binomial_name/current/protein</td>
<td>Returns a FASTA file containing all the protein sequences known to be encoded by the genome. MIME type is application/x-fasta.</td>
</tr>
<tr>
<td>/genome/Saccharomyces_cerevisiae/current/protein</td>
<td></td>
</tr>
<tr>
<td>/genome/Binomial_name/current/feature</td>
<td>Returns a GFF3 file describing genome annotations. MIME type is application/x-gff3.</td>
</tr>
<tr>
<td>/genome/Saccharomyces_cerevisiae/current/feature</td>
<td></td>
</tr>
</tbody>
</table>
Bulk genome data

Database: FlyBase Chado DB r4
Species: Drosophila melanogaster
Release: r4.3.0, dated 200602017 as dmel_r4.3_20060217

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td>06-Jun-2006 10:29</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Release.txt</td>
<td>29-Mar-2006 00:22</td>
<td>1k</td>
<td></td>
</tr>
<tr>
<td>blast/</td>
<td>28-Mar-2006 13:38</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>fasta/</td>
<td>28-Mar-2006 13:44</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>gff/</td>
<td>28-Mar-2006 13:40</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>reimotes.txt</td>
<td>28-Mar-2006 00:22</td>
<td>1k</td>
<td></td>
</tr>
<tr>
<td>tables/</td>
<td>28-Mar-2006 00:14</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

Select Format, Feature Type and Sequence Region below and click on a Download button to get data

Format
- Sequence (Fasta)
- Features (GFF)

Feature Type (for FastA)
- Protein-coding Gene
- Gene region +/- 2000bp
- Transposable Element
- Transcript
- 5' UTR
- tRNA
- Protein
- 3' UTR
- misc RNA
- CDS
- Intron
- Pseudogene

Sequence Region
Chromosomes arm
- 2L
- 2R
- 3L
- 3R
- 4
- U
- X
- mt or
- All

URL:
Bulkfile output of Chado DB

- Any Genome DB wants genome outputs
  http://gmod.cvs.sourceforge.net/gmod/schema/GMODTools/
- Generate public releases of genome
  - Fasta, GFF with project-standard formats, IDs
  - Database summary tables
  - Web-usable, standard-url “/genome/” folders per species, release
- Usage
  - Extensively configurable via XML
    - Chado SQL calls, Perl post-processors, DB-Public mappings,
  - Extensible for new outputs (e.g. Biomart tables)
  - Tested with Yeast, Fruitfly Chado DBs (others??)
GFF 2 BioMart 4 data miners

http://gmod.cvs.sourceforge.net/gmod/schema/GMODTools/bin/gff2biomart5.pl

SCRIPT USE

• Simple Perl transformer: feed GFF, Fasta
• Creates tables for BioMart (v0.3 now): .sql, .txt and .xml

IN BIOMART

• filter (include, exclude) features that exist in regions, including joint filters (has predicted gene but no homology)
• output 4 kinds of attributes: a feature table, per-feature sequence, region table, per-region sequence
• E.g., http://insects.eugenes.org/BioMart/martview
Gff2Biomart Outputs

- Region Table: chromosome in 1 Kb bins. Features that overlap bin are tabulated.
- Feature Table: per-feature tables store all GFF attributes (id, dbxref, match stats, ..)
- DNA Table: for fasta output
- Config. Table: main_biomart.xml and sequence_biomart.xml for web form interface.
TeraGrid Summary

- **PROBLEM** in bioinformatics: enable use of large biology data analyses on shared cyberinfrastructure.
- **SOLUTION**: Parallelize data access rather than applications for Grid use of existing and new biology analyses.
- **RESULTS**: New insect and crustacean genomes have been analyzed on TeraGrid to assess data grid methods in genome informatics. Rapid Grid analyses have facilitated rapid biology discoveries in these genomes.
New Fly, waterFlea genomes

- **Biologists Need rapid access:** to new genomes for *Daphnia pulex* and twelve *Drosophila*

- **Find the Genes:** Compare to 9 proteomes: fly, worm, mouse, yeast, human, ...

- **Generic Model Organism Database (GMOD) tools** organize TeraGrid results for public:
  - genome maps (GBrowse), web BLAST, data mining (BioMart), genome summaries
  - wifleabase.org (Daphnia), insects.euGenes.org (Drosophila)
Proteome Annotations

Found Eukaryote Genes

- Daphnia
- Drosophila
- Mosquito

Total protein genes

Percent of proteome

Thousands
Phylogeny / Gene Similarity
Possible gene gain/loss
## TeraGrid usage steps

<table>
<thead>
<tr>
<th>Step</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Preparation</strong></td>
<td><strong>One time</strong></td>
</tr>
<tr>
<td>1. Obtain TeraGrid account</td>
<td>Via web <a href="http://www.teragrid.org/userinfo/">http://www.teragrid.org/userinfo/</a></td>
</tr>
<tr>
<td>2. Establish certificates</td>
<td>Grid-security entries; test proxy; local workstation certificate</td>
</tr>
<tr>
<td>3. Locate biology software</td>
<td>Find and compile parallel applications</td>
</tr>
<tr>
<td><strong>Processing</strong></td>
<td><strong>Per analysis</strong></td>
</tr>
<tr>
<td>4. Locate and prepare data</td>
<td>partition, shred &amp; randomize</td>
</tr>
<tr>
<td>5. Transfer data to TeraGrid</td>
<td>FTP, secure-shell, other</td>
</tr>
<tr>
<td>6. Configure and run analysis</td>
<td>Globus run scripts, attention to errors, queuing</td>
</tr>
<tr>
<td>7. Return and collate results</td>
<td>Post-process to combine results from nodes; e.g. to-GFF for map view of genome blast.</td>
</tr>
</tbody>
</table>
Data grid methods

1. @virtualdata = biodirectory("find protein coding sequences for Drosophila species"),
2. @realdatalist = biodirectory("get locators for @virtualdata split n ways"), for n compute nodes
3. for i (1.. n) { copy(realdatalist[i], gridcpu[i]); results[i] = runapp(gridcpu[i]) }
4. result_table = collate( @results );

These steps will work for gene finders, homology comparison, multiple alignment tools, and phylogenetic comparison.
• TeraGrid for genomes interest group?
  • Every genome DB could use TeraGrid (US) or DEISA (Europe) or other for: comparative genome analyses, gene finding, phylogenetics.
  • Learning curve, DG will help, build generic tools

• Genome/organism public discussion lists:
  • Bionet/BIOSCI is available: www.bio.net, Usenet bionet.*
  • ~50 active lists: arabidopsis, worms(2), yeast, fly, corn, medicago, molec. methods, bio-soft, others

• Contact: gilbertd@indiana.edu
Thanks to these folks

- IU and national TeraGrid group for the CPUs
- NIH for Fruitfly genomes; JGI and DGC for Daphnia genome
- GMOD project developers for the tools
Genome Annotations

- **Gene Homology**
  - Nine well-annotated proteomes: Yeast, Worm, Mosquito, Fruitfly, Bee, Zebrafish, Mouse, Human, Arabidopsis
  - BLAST the 13+ genomes at TeraGrid.org

- **Gene Predictions**
  - SNAP - good ab-initio predictor, best finding new Dros. Reproductive genes.

- Collate to Gene Finding Format for map views, BioMart, sharing
Gff2biomart Example

% $b/gmod/biomart/gff2biomart5.pl -db=drospege_mart_caf1b \
   -dataset=$bmid -species=Drosophila_${species} -version=$dpid \
   -output biomart-$dp -table=cross_genome_match_dmelchr,HSP_modDM \
   -fasta $scd/${dpid}.fa.gz \n   $gff1/${dp}-chromosomes.gff $gff1/${dp3}-markers.gff.gz \n   $gff1/${dp3}-dmel-algn.gff.gz \n   $sc/caf1a/dgil/${dp}prot9-hsp.gff.gz $sc/caf1a/oliv/${dp}.caf1.gff.gz \n   $sc/caf1a/ncbi/${dp}_caf1_NCBI_GNO.gff.gz \n   ... etc ...

# INSTALL IN BioMart DB:
% mysqladmin create drospege_mart_caf1b
% cat biomart-dana/*.sql | mysql drospege_mart_caf1b
% mysqlimport drospege_mart_caf1b `pwd`/biomart-dana///*.txt
% cat biomart-dana/dana_meta.sql_example | mysql drospege_mart_caf1b
Select the Attribute Page

FEATURE TABLE

Common Features
- ✔ Chromosome
- ✔ Biotype
- ✔ End
- ✔ Score
- ✔ Name
- ✔ Note
- ✔ Source
- ✔ Start
- ✔ Strand
- ✔ ID
- ✔ Dbxref
- ✔ Other Attributes

Select the output format:
- ✔ HTML
- ✔ Text, comma separated
- ✔ MS Excel

File compression:
- ✔ None
- ✔ gzip (.gz)

Enter a name for this result set:
Name: daph-genepred-noest

Enter a value to open results in new window (REQUIRES POP-UP UNBLOCKING), or to provide a name for file download.