Genome Database Construction with GMOD 2007

Don Gilbert, Indiana University

ABSTRACT

Generic Model Organism Database (GMOD) is a federation of groups with different needs and abilities to contribute to a shared organism/genome database toolkit.

Build your own organism genome database with GMOD instructions and help. These include Chado genome database schema, with middleware for adding and extracting data, GBrowse to view genome maps, BioMart for genome data mining; Teragrid shared cyberinfrastructure for automated gene prediction and gene homology, literature and genome annotation tools, comparative maps, and biological pathway tools.

Example uses include small projects that mix public genomes with lab data, several new organism genomes, and established model organism projects.

Availability: http://www.gmod.org/ and http://iubio.bio.indiana.edu/gil/

Contact: Don Gilbert, gilbertd@indiana.edu

INTRODUCTION

Generic Model Organism Database (GMOD) is a federation of groups with different needs and abilities to contribute to a shared organism/genome database toolkit. It is inventing itself as it goes along, and welcomes new customers and contributors. GMOD is an umbrella organization that encourages sharing of lessons and expertise with genome databases.

Over 100 GMOD customers include BeeBase, BeetleBase, DictyBase, DroSpeGe, EcoCyc, FlyBase, GeneDB, Gramene, HapMap, ParameciumDB, PlasmoDB, Rat GD, Mouse GD, Saccaromyces GD, TAIR Arabidopsis, TIGR, ToxoDB, VectorBase, waterFlea Base, WormBase, and Xenopus base (iconized in Figure 1).

- Well established Model Organism Database projects with bioinformatics expertise These MODs contribute and use components to 'share the wealth' and reduce funding costs for duplicative work. MODs contribute adaptable components that work with established mix of tools and large volume data and usage, with detailed genome biology, and strict needs that imply technically complex tools.
- Many new organism database projects with limited funding, and a desire to build on established, tested MOD methods, for their similar genome database needs.
- A growing number of lab/research projects that combine public genomes with lab generated data, including microarray, functional and proteomic analyses, and genome wide surveys.

You can learn how to build your own organism genome database with GMOD instructions and installation, via the new Wiki at www.gmod.org, with active mailing list support. Example uses cover converting a GenBank Genome into a GMOD database, and adding BLAST and like genome analyses.

CURRENT CONTENTS

Available now from GMOD are the Chado genome database schema, with middleware (Perl and Java tools) for adding and extracting data. GBrowse provides for easy creation of powerful genome maps. BioMart provides full genome data mining. Literature and sequence curation tools capture improvements to genome data. CMap comparative maps and syntenic maps, and biological pathway tools add useful components.

Standard Tools and Components from GMOD

- **Chado** database schema and middleware (Chris Mungall, Dave Emmert, *et al.*)
- **GBrowse** – Web-based genome annotation viewing (Lincoln Stein, Scott Cain)
- **Apollo** – Desktop-based genome annotation editing (Nomi Harris, Michelle Clamp)
- **CMap** – Web-based comparative map viewing (Ken Clark, Ben Faga)
- **BioMart -** Genome data mining, an Ensembl/GMOD collaboration (Arek Kasprzyk *et al.*)
- **Sybil** – Web synteny maps at gene & chromosome level (Jonathan Crabtree, TIGR)
- **Turnkey** – Chado-based web site (Allen Day, Brian O'Connor)
- **Pathway Tools** – metabolic pathways (Peter Karp, BioCyc)
- **PubMed/PubFetch** – Literature management
- **Textpresso** – Automatic paper classification & searching
- **LuceGene -** Genome object/text/web search system (Don Gilbert)

Generic Genome Browser is probably GMOD's most popular component. It is easy to install, only basic command-line familiarity is required. However, the reason that GBrowse is popular is that is a very capable browser, adaptable to many organism/genome needs. GBrowse's new Gene-Balloons (Figure 2) are a good example of the expanding functionality of GMOD tools.

Figure 1. GMOD is built by and for many contributing Model Organism Databases.

EXAMPLE GMOD DATABASES

Many organism database projects are contributing to and/or adopting GMOD components. These include from Indiana VectorBase (Notre Dame), Purdue's EcoliHub (with Jim Hu's EcoliWiki) and Soybean genome projects, Indiana University's FlyBase, DroSpeGe and wFleaBase genome projects. The *Daphnia pulex* genome is hosted at wFleaBase.org and Joint Genome Institute (JGI). Automated and curated contents include a TeraGrid-computed gene homology and predictions, cDNA/EST assemblies with the TIGR-developed PASA pipeline, a Chado-based genome database, GBrowse genome maps, BioMart data mining and more.

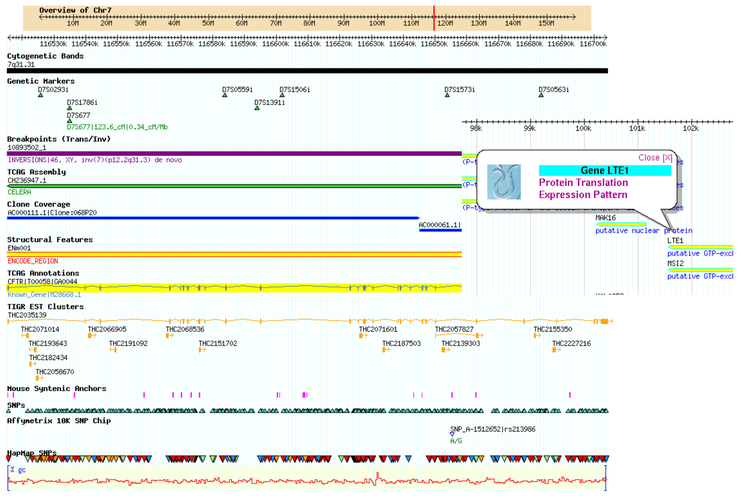

Figure 2. GBrowse from Human chromosome 7 database, www.chr7.org, shows a wealth of information in complex genome data. Gene-Balloons allow views of detailed gene information.

HOW-TO BUILD A CHADO GENOME DATABASE

Modularity is inherent in the GMOD Chado database schema, with a core module and several biology groupings, with common structure. **Ontologies**, organizing standard vocabularies in biology, are at the core of Chado's design, making it excellent for annotation of biology data. **Associated Software** for Chado includes middleware in Perl (BioPerl) and Java for managing data, and stand-alone programs with Chado adaptors such as Gbrowse, and Apollo. **Complexity and Detail** is inherent in genome data, and Chado embraces this with room to grow without sacrificing long-term stability of the database and its interfaces. **Data Integration** is another key component of Chado, where public and lab data sets can be combined in a common warehouse. **Support** is actively provided as a shared responsibility among the GMOD user community. There are several useful, worked examples documented at GMOD.org, such as this recipe, http://www.gmod.org/Load_RefSeq_Into_Chado to load a Genome of your favorite organism from GenBank into a Chado database.

Chado Modules

- **CV**: Controlled vocabularies and ontologies
- **Sequence**: Biological sequences and objects which can be localized on them
  - Companalysis: Adjunct to sequence module for in-silico analysis
  - Map: Adjunct to sequence module for non-sequence localization
- **Expression**: Transcript and protein expression events
- **Genetics**: Genetic/phenotypic interactions in genotypic/environmental context
- **Library**: for descriptions of molecular libraries
- **Mage**: for microarray data
- **Organism**: Taxonomy / species information
- **Phenotype**: for phenotypic data
- **Phylogeny**: for organisms and phylogenetic trees
- **Pub**: Publication / Biblio. / Reference information
- **Stock**: for specimens and biological collections
- **Contact**: for people, groups, and organizations
- **General**: General information / database cross-references

GENOME ASSEMBLY, ANNOTATION, ACCESS

Creating and collecting genome data is the start of a genome database. This includes genome assembly (from WGS and 454 technology), automated annotation and analysis for finding model organism gene homology, EST/cDNA collections, gene predictors (GenScan, TwinScan, GeneWise, and many more). All this evidence must be combined intelligently for a full gene catalog, including function (Gene Ontology), pathway (KEGG), homology, EST expression and related knowledge. One such tool is the Program to Assemble Spliced Alignments (PASA) for EST and cDNA data, from TIGR. Teragrid shared cyberinfrastructure for automated gene prediction, gene homology and annotation is another area where new, shared genome methods are becoming available.

Figure 3. Biologist's Desire: Search millions of organism genes in current databases around the world, simply and quickly, finding the best answer directly. Simple and powerful User Interface designs are one goal of GMOD to facility genome data access.

MOD USER INTERFACES

The user interface (UI) is the most visible aspect of a model organism database (MOD), and arguably has the most direct impact on the satisfaction of its users. It also drives the design of many aspects of any genome database project. A recent caucus of MOD projects shared these UI insights. *General lessons learned*: Clarity in actions required of users, and clarity and reliability of results of these are important to users. Appearance is less important to users than functionality and responsiveness. Developing good UIs takes sustained work, including feedback and community testing. *Complexity is an inherent problem*: MODs deal with rich, complex data that is constantly expanding and changing. A central challenge is to make common tasks easy and complex tasks possible. There is a need at many MODs for broader availability of power-user interfaces for complex queries, for uploading and operating on sets of genes in one step, and for flexible configuration of data output formats. *Good new ideas:* Wikipedia provides an excellent example for science community participation that several MODs are adopting. More dynamic web content and graphical summaries can help manage information Google can be harnessed to aid (Figure 3), but is not solely sufficient for, searching MOD data.

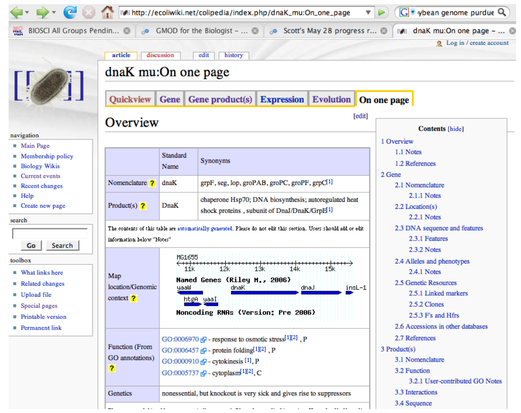

Figure 4. WikiGenomes will facilitate direct community annotations (e.g. ecoliwiki.net), with a standard collaboration interface popularized by Wikipedia.

COMING ATTRACTIONS

Coming attractions from GMOD include 'GoogleGene' and 'GoogleGenomeMaps' (Figure 3), to search the world of genome data, including interactive maps. WikiGenome will aid community annotations of new and old genomes (Figure 4) in an established collaborative paradigm, e.g. the EcoliWiki project. TeraGenome plans to provide quick and easy whole genome analyses with TeraGrid shared cyberinfrastructure.

CONTACT FOR COLLABORATION

The Genome Informatics Lab at Indiana University includes Don Gilbert and associates working to facilitate generic organism/genome databases. We welcome new collaborations; please contact Don at gilbertd@indiana.edu, or see http://iubio.bio.indiana.edu/gil/

References

Colbourne, J.K., Singan, V.R., Gilbert, D.G. 2005. wFleaBase: the Daphnia genome database, **BMC Bioinformatics**, 6:45 doi:10.1186/1471-2105-6-45 URL: wfleabase.org

Gilbert, D.G. 2007. DroSpeGe: rapid access database for new *Drosophila* species genomes. *Nucleic Acids Res.* **35**(Database issue): D480-D485; doi:10.1093/nar/gkl997

Stein L.D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J.E., Harris, T.W., Arva, A., and Lewis, S. 2002. The generic genome browser: a building block for a model organism system database. **Genome Res.** 12: 1599-610. URL: www.gmod.org

Wang L et al. (2007) BeetleBase: the model organism database for Tribolium castaneum. Nucleic Acids Res 35: D476-9

See also http://www.gmod.org/wiki/index.php/Publications