This HOWTO describes a method for loading sequence annotation data in GFF3 format into a Chado database.

Download the GFF3 Files

An easy way to load data into the database is to use a GFF3 file and the script `load/bin/gmod_bulk_load_gff3.pl`. A good set of sample data is the GFF3 file prepared by the nice folks at the Saccharomyces Genome Database:

ftp://ftp.yeastgenome.org/pub/yeast/data_download/chromosomal_feature/saccharomyces_cerevisiae.gff

This file contains Gene Ontology (GO) annotations, so if you didn't load GO when you executed `make ontologies` you will get many warning messages about being unable to find entries in the dbxref table. If you want to load GO you should be able to execute `make ontologies` and select **Gene Ontology** for installation.

Add an Entry for Your Organism

You will need to have an entry for your species in the Chado organism table. If you are unsure if this entry exists log into your database and execute this SQL command:

```
SELECT common_name FROM organism;
```

If you do not see your organism listed, execute a command equivalent to this:

```
INSERT INTO organism (abbreviation, genus, species, common_name)
               VALUES ('S.cerevisiae', 'Saccharomyces', 'cerevisiae', 'yeast');
```

Substitute in the appropriate values for your own organism if it's not *yeast*.

Load the GFF3

Unless your GFF3 is sorted by location with grouped gene models (gene, mRNA, CDS/exon/UTR), you must first do this. Use this gmod_gff3_preprocessor.pl.

```
> gmod_gff3_preprocessor.pl --gfffile saccharomyces_cerevisiae.gff --outfile saccharomyces_cerev
```

Then execute `gmod_bulk_load_gff3.pl`:

```
>gmod_bulk_load_gff3.pl --organism yeast  --gfffile saccharomyces_cerevisiae.sorted.gff
```

This loads the GFF3 file. The loading script requires GFF3 as it has tighter control of the syntax and requires the use of a controlled vocabulary (from Sequence Ontology Feature Annotation (SOFA)), allowing mapping to the relational schema. In addition to supplying the location of the file with the `--gfffile` flag, the `--organism` tag uses the common name (`common_name` field) from the Chado organism table. Do `perldoc gmod_bulk_load_gff.pl` for more information on adding other organisms and databases, as well as other available command line flags.

Note that `gmod_load_gff3.pl` is also available, but is limited in how much it has been supported and in how flexible it currently is. It is a good example of how to write code using Class::DBI classes that are created at the time of install. For more information on using these classes, see Modware for a Class::DBI-based middleware/API.

Creating GFF3 from UniProt/SwissProt Files

A recent update (April 2007) to `bp_genbank2gff3.pl` extends it to handle Swiss and EMBL format input, along with GenBank. You can now create GFF3 entries of UniProt sequences suited to loading into Chado, including most of the protein description, Dbxref, and related fields useful in annotating genome matches. Use the `--format Uniprot` flag to specify this input format (`--format EMBL` can also be useful).

```
>bp_genbank2gff3.pl --noCDS --in uniprot-subset.dat --format Uniprot
>gmod_bulk_load_gff3.pl --database mygenome --gff  uniprot-subset.dat.gff --organism fromdata
```

Use the `--organism fromdata` flag to load UniProt with many organisms.

> **This code needs to be tested.** Please help improve this section with your tests.

More Information

See the related HOWTO Load RefSeq Into Chado.

Please send questions to the GMOD developers list:

gmod-devel@lists.sourceforge.net

Or contact the GMOD Help Desk

Authors

- Scott Cain
- Brian Osborne