Genomic Feature Database

Bio::SeqFeature::Store

Why Use It?

- Get the coordinates of all annotated genes

- Get the sequence of all genes

- Get the sequence of all genes with CDS that are greater than 500bp

- Get a list of all features in which a TE has inserted

- Get promoter sequence of all genes

- Get all of the above for a genes with specific descriptions

- Gbrowse uses it for all of the above and much more.

To use it:
Build a database with a standardized database schema for genomic feature data

Simply:

- use a BioPerl script (bp_seqfeature_load.pl) with a GFF file to create and load the database

Creating and loading a database of sequence features

```
bp_seqfeature_load.pl --help
Unknown option: help
Usage: /usr/local/bin/bp_seqfeature_load.pl [options] gff_file1 gff_file2...
  Options:
          -d --dsn            The database name (dbi:mysql:test)
          -s --seqfeature     The type of SeqFeature to create (Bio::DB::SeqFeature)
          -a --adaptor        The storage adaptor to use (DBI::mysql)
          -v --verbose        Turn on verbose progress reporting
             --noverbose      Turn off verbose progress reporting
          -f --fast           Activate fast loading (only some adaptors)
          -T --temporary-directory  Specify temporary directory for fast loading (/tmp)
          -c --create         Create the database and reinitialize it (will erase contents)
          -u --user           User to connect to database as
          -p --password       Password to use to connect to database
          -i --ignore-seqregion
                              If true, then ignore ##sequence-region directives in the
                              GFF3 file (default, create a feature for each region)
```

GFF: Generic Feature Format.

GFF files are plain text, 9 column, tab-delimited files

Column 1: "seqid"

Column 2: "source"

Column 3: "type"

Columns 4 & 5: "start" and "end"

Column 6: "score"

Column 7: "strand"

Column 8: "phase"

Column 9: "attributes"

```
##gff-version 3
ctg123 . exon 1300 1500 . + . ID=exon00001
ctg123 . exon 1050 1500 . + . ID=exon00002
ctg123 . exon 3000 3902 . + . ID=exon00003
ctg123 . exon 5000 5500 . + . ID=exon00004
ctg123 . exon 7000 9000 . + . ID=exon00005
```

Query the Database

- Now that the database is built and filled with data, scripts can be written to query the database

use Bio::DB::SeqFeature::Store;

```perl
use Bio::DB::SeqFeature::Store;

# Create an object with info about the DB and
# all the tools to interact with it
my $db = Bio::DB::SeqFeature::Store->new(
                    -adaptor => 'DBI::SQLite',
                    -dsn     => '/path/to/database.db'
              );
```

Documentation

```
http://gmod.org/wiki/GFF
http://www.sqlite.org/sqlite.html
http://search.cpan.org/~cjfields/BioPerl-1.6.901/Bio/DB/SeqFeature/Store/DBI/SQLite.pm
http://search.cpan.org/~cjfields/BioPerl-1.6.901/Bio/DB/SeqFeature/Store.pm
```