Flat Databases HOWTO

Lincoln Stein
Cold Spring Harbor Laboratory (http://www.cshl.org)

lstein-at-cshl.org

Brian Osborne
Cognia Corporation (http://www.cognia.com)

brian-at-cognia.com

Heikki Lehväslaiho
European Bioinformatics Institute (http://www.ebi.ac.uk)

heikki-at-ebi.co.uk

The Open Biological Database Access (OBDA) standard specifies a way of generating indexes for entry-based sequence files (e.g. FASTA, EMBL) so that the entries can be looked up and retrieved quickly. These indexes are created and accessed using the `Bio::DB::Flat` module.

1. Creating OBDA-Compliant Indexed Sequence Files

`Bio::DB::Flat` has the same functionality as the various `Bio::Index` modules. The main reason to use it is if you want to use the BioSequence Registry system (see the OBDA Access HOWTO at http://bioperl.org/HOWTOs), or if you want to share the same indexed files among scripts written in other languages, such as those written with BioJava or BioPython.

There are four steps to creating a `Bio::DB::Flat` database:

1. Select a Root Directory

Select a directory in which the flat file indexes will be stored. This directory should be writable by you, and readable by everyone who will be running applications that access the sequence data.

2. Move the Flat Files Into a Good Location

   The indexer records the path to the source files (e.g. FASTA, or local copies of GenBank, Embl or SwissProt). This means that you must not change the location or name of the source files after running the indexer. Pick a good stable location for the source files and move them there.

3. Choose a Symbolic Name for the Database

   Choose a good symbolic name for the database. For example, if you are mirroring GenBank, "genbank" might be a good choice. The indexer will create files in a subdirectory by this name located underneath the root directory.

4. Run the bioflat_index.pl script to load the sequence files into the database.

   The final step is to run the bioflat_index.pl script. This script is located in the BioPerl distribution, under scripts/db. For convenience, you may want to copy it to /usr/bin or another system-wide directory.

2. Choosing Your Options

The first time you run the script, the typical usage is as follows:

```
bioflat_index.pl -c -l /usr/share/biodb -d genbank -i bdb -f fasta data/*.fa
```

The following command line options are required:

Table 1.

Option	Describtion
-c	create a new index
-l	path to the root directory
-d	symbolic name for the new database
-i	indexing scheme (discussed below)
-f	source file format

The `-c` option must be present to create the database. If the database already exists, `-c` will reinitialize the index, wiping out its current contents.

The `-l` option specifies the root directory for the database indexes.

The `-d` option chooses the symbolic name for the new database. If the `-c` option is specified, this will cause a new directory to be created underneath the root directory.

The `-i` option selects the indexing scheme. Currently there are two indexing schemes supported: "bdb" and "flat." "bdb" selects an index based on the BerkeleyDB library. It is generally the faster of the two, but it requires both the BerkeleyDB library and the Perl DB_File module to be installed on your system. "flat" is a sorted text-based index that uses a binary search algorithm to rapidly search for entries. Although not as fast as bdb, the flat indexing system has good performance for even large databases, and it has no requirements beyond Perl itself. Once an indexing scheme has been selected there is no way to change it other than recreating the index from scratch using the `-c` option.

The `-f` option specifies the format of the source database files. It must be one of the formats that BioPerl supports, including "genbank", "swiss", "embl" or "fasta". Consult the `Bio::SeqIO` documentation for the complete list. All files placed in the index must share the same format.

The indexing script will print out a progress message every 1000 entries, and will report the number of entries successfully indexed at the end of its run.

To update an existing index run bioflat_index.pl without the `-c` option and list the files to be added or reindexed. The `-l` and `-d` options are required, but the indexing scheme and source file format do not have to be specified for updating as they will be read from the existing index.

For your convenience, bioflat_index.pl will take default values from the following environment variables:

Table 2.

ENV variable	describtion
OBDA_FORMAT	format of sequence file (`-f`)
OBDA_LOCATION	path to directory in which index files are stored (`-l`)
OBDA_DBNAME	name of database (-d)
OBDA_INDEX	type of index to create (`-i`)

3. Moving Database Files

If you must change the location of the source sequence files after you create the index, there is a way to do so. Inside the root directory you will find a subdirectory named after the database, and inside that you will find a text file named "config.dat." An example config.dat is shown here:

```
index flat/1
fileid_0 /share/data/alnfile.fasta 294
fileid_1 /share/data/genomic-seq.fasta 171524
fileid_2 /share/data/hs_owlmonkey.fasta 416
fileid_3 /share/data/test.fasta 804
fileid_4 /share/data/testaln.fasta 4620
primary_namespace ACC
secondary_namespaces ID
format URN:LSID:open-bio.org:fasta
```

For each source file you have moved, find its corresponding "fileid" line and change the path. Be careful not to change anything else in the file or to inadvertently replace tab characters with spaces.