Download and Sync with the Entire PDB Database

Tianyi Shi

2020-10-17

Contents

All the PDB files are available via PDB's FTP server. Simply `cd` to a directory and use `wget` to download all of them.

```
mkdir pdb; cd pdb; mkdir zipped unzipped; cd zipped
wget ftp://ftp.wwpdb.org/pub/pdb/data/structures/all/pdb/*
```

Depending on the traffic, The first download will take 2~4 days. After the first download, to sync the local database with the FTP server, add the `-N` option to `wget` (essentially this means downloading "new" files only).

The `*.ent.gz` files downloaded are zipped. You can use `gunzip` to decompress them to ready-to-use `*.ent` files (which is written in PDB format), but that will remove the `*.gz` files, which is inconvenient if you want to keep your database in sync with the remote. Use `gzcat` (or equivalently `gunzip -c`) instead to preserve the `*.gz` files while decompressing:

```
find . -name "*.gz" | xargs -I{} -n1 bash -c '
src={}; dst=../unzipped/$(basename ${src%*.gz})
[ -f $dst ] && echo "$dst already exists, skipping..." || ( gzcat -cv $src > $dst )'
```

You can monitor the progress from the verbose output of `gzcat`:

```
./pdb1byf.ent.gz:       78.1%
./pdb1byg.ent.gz:       76.3%
./pdb1byh.ent.gz:       77.3%
./pdb1byi.ent.gz:       75.2%
...
```