Programming for Bioinformatics
BIOL 8803 B
August 31st, 2015

Commands for today:

`wget` – get a remote file

`curl` – capture a URL, in this case a file

`ftp` – File transfer protocol interface

`scp` – Secure copy, copy a file to a different computer

`cut` – Get certain columns from a file

`paste` – Merge files together by columns

`join` – Join files by a common column

`touch` – change the attributes of a file

File Transfer Exercises

1.) Downloading UCSC Genome Browser files with `wget`

Toy exercise – the objective here is only to learn how to use wget. You need not to wait for each file to finish downloading, you can quit the download if you figured you are using the right command.

a.) What is an ENCODE repository? Download a file (pick a small one of your choice) from the ENCODE repository on the UCSC genome browser with `wget` and understand what data it contains

b.) The database folder located in hg19 in goldenPath is an important folder. We are going to fetch four files `knownGene.txt.gz`, `knownGene.sql`, `kgXref.txt.gz`. and `kgXref.sql`. Copy their link address, paste in a file. Now get `wget` to download the files by reading the URL from the file you just created. In other words, this time the input URL comes from a file, and there are multiple of them. **Hold on to these files, we will use these later on.**

c.) Download all of the Pol2b binding data available in one line with `wget`. Hint: regex.

d.) Test if the URL http://www.ncbi.nlm.nih.gov/SNP/snp_ref.cgi?searchType=adhoc_search&type=rs&rs=rs12345 is correct using `wget`. Hint: Spiderman.

2.) Downloading UCSC Genome Browser files with `curl`

a.) Repeat the last question's part a) but this time using `curl`

b.) Download two files with `curl` (this time give the URL on the command line and not from a file)

c.) What are the differences between `wget` and `curl`?

3.) Downloading lots of files with `ftp`

a.) Go to the UCSC Genome Browser download site (ftp mirror)
http://hgdownload.cse.ucsc.edu/downloads.html. Look around

b.) ftp into `hgdownload.cse.ucsc.edu` (user name is `anonymous`)

c.) Use ftp to download multiple files, perhaps the ENCODE GIS-PET RNA clusters for the hg18 assembly of the human genome?

4.) Write one-liners to perform the following actions with wget/curl:

a.) Download the file taxdb.tar.gz from ftp.ncbi.nih.gov/blast/db/

b.) Download all files that start with "blast" from ftp.ncbi.nih.gov/blast/documents/

c.) Download all "ppt" files from ftp.ncbi.nih.gov/blast/demo/

5.) Use `wget`, `curl` and/or `ftp` to download files from the NCBI

a.) Download a Genbank sequence file of interest – e.g. NM_006565.3, using the eutils. Google and figure out how you can do this. It's easy.

b.) Download multiple sequences of interest using a list of accessions (3-10 accessions for your favorite genes or proteins), similarly to the previous question.

c.) Lets `ftp` this time to the NCBI server. Figure out the address for the server (it's on NCBI). The login credentials are `anonymous` and password is your email address. Navigate to genbank/genomes/Bacteria/Neisseria_meningitidis_FAM18_uid255 and get me the gff and ptt files. What are these two formats?

6.) E-utils

The documentation on how to use E-utils can be found here:
http://www.ncbi.nlm.nih.gov/books/NBK25500/)

Downloading a random genome sequence for *Neisseria meningitidis*. To do so, your approach will be the following:

a.) Retrieve the Genome ID for *N. meningitidis*

b.) Retrieve the Nucleotide ID linked to the Genome ID and limit the search to only RefSeq Genome Sequences

c.) Download the genome sequence

7.) More utilitarian usage with `scp`

a.) Copy a file to a remote server (such as biocluster or whatever you have access on)

b.) Copy a directory

d.) Copy them both back to the machine you are on

8.) `touch`ing things around

a.) Create a test file by the name of efg.txt using `touch`

b.) Open it up in `emacs` and write something in it

c.) Save it and close it

d.) `touch` the file again, did anything change?

e.) List all the attributes of abc.txt (created in 2a) and pay attention to time and date of access

f.) `touch` abc.txt and list all the attributes again, what changes?

9.) Happy together!

a.) Create `1.txt` with the following content:

```
1 abc
2 lmn
3 pqr
```

b.) Create `2.txt` with the following content:

```
1 abc
3 lmn
9 opq
```

c.) `join` the two files by **column 1**

d.) `join` the two files by **column 2**

e.) Repeat part (d) but this time also include all records from the **first** file. This is referred to as left outer join

f.) Repeat part (d), this time also include all records from the **second** file. This is referred to as right outer join

g.) Repeat part (d) one last time and include all records from **both** files. This is referred to as full outer join

10.) File handling/manipulation

a.) Generate two files by using the following commands

```
cat /dev/urandom | tr -dc 'ACGT' | fold -w 50 | head -50 > r1.fa
```

```
cat /dev/urandom | tr -dc 'acgt' | fold -w 50 | head -50 > r2.fa
```

b.) Display the first 5 characters in each of the first 5 lines of r1.fa

c.) Combine (horizontally) the first 5 lines of r1.fa and r2.fa into a new file. That is, the resulting file's first 5 lines will be from r1.fa and the next 5 lines will be from r2.fa.

d.) Combine (vertically) the last 5 lines of r1.fa and r2.fa into a new file. That is, the resulting file's first column will be last 5 lines of r1.fa and the second column will be last 5 lines will be from r2.fa.