File names

DO NOT USE SPACE
Try to use ‘_’ or ‘-’ to replace spaces

Homework #8

Go to the webpage of CAZy family GH5: http://www.cazy.org/GH5_all.html

Copy & paste the entire page to an excel spreadsheet
- webpage (ctrl+a, Ctrl+c, Ctrl+v)

There are four pages; do this for all four pages, append them in one spreadsheet

Copy the col D (GenBank IDs), and paste to vi editor to save the protein ids as a file

Design command line to use sed (delete empty lines), grep (-v to remove non-id lines), sort, uniq, wc etc. to get a unique set of GenBank IDs of GH5; save as a file

Get GenPept format sequences using batch Entrez

Design command line to use grep, awk to print locus id and sequence length; save as a file

Design command line to use grep, sed (to remove characters/spaces), sort, uniq, wc to extract organism information (how many organisms have GH5 proteins, how many GH5 proteins these organisms have, the top ranked organisms etc.)

Report due April 09 (send by email)

Office hour:
Tue, Thu and Fri 2-4pm, MO325A
Or email: yyin@niu.edu

cut # extract columns from a file

```
cat file | cut -f1 = cut -f1 file   # cut the first column (default delimiter tabular key)
cat file | cut -f1 -d ' '   # specify delimiter to be regular space
cat file | cut -f1-3      # cut 1 to 3 col
cat file | cut -f1,7,10 > file.1-3-10      # cut 1, 7, 10 col and save as a new file
```

sort # sort rows in a file, default on first col in alphabetical order (0-9 then a-z, 10 comes before 9)

```
cat file | sort -k 2   # sort on 2 col
cat file | sort -k 2,2n    =    cat file | sort -k 2 -n     # sort in numeric order
cat file | sort -k 2,2nr      # sort in reverse numeric order
```

uniq # report file without repeated occurrences

```
cat file | cut -f2 | sort | uniq  = cat file | cut -f2 | sort -u   # unique text
cat file | cut -f2 | sort | uniq -c    #  count number of occurrences of unique texts
```

grep # extract lines match a given word or pattern

```
cat file | grep '>' | head   #  print only lines containing '>'
cat file | grep -v '>' | head # print lines not containing '>'
cat file | grep -n '>' | head   # also print in which lines '> is found
cat file | grep -c '>'   =   cat file | grep '>' | wc -l     # count the number of occurrences
cat file | egrep 'chr1|chr2'# print lines containing chr1 or  chr2 (multi-words or patterns)
```

sed # stream editor, modify, delete, search and replace etc

```
cat file | grep '>' | sed 's/>//'      # delete '>'
cat file | grep '>' | sed 's/>/+/'      # replace '>' with '+'
cat file | sed '/^$/d'      # delete empty line
cat file | sed '/>/d'      # delete all lines with '>'
cat file | sed -n '/>/p'      # print all lines with '>'
cat file | sed -n '101,200p'      # print selected lines (101 to 200) in the file
```

awk # give a condition, perform an action (print)

```
cat file | awk '$5=="22"'  #  $5 means the 5th col, default delimiter is regular space
cat file | awk -F "\t" '$5=="22"'      # define delimiter to be tabular space "\t"
cat file | awk '/>/'    =    cat file | grep '>'      # put pattern between //
cat file | awk '$1~/>/'      # specify the pattern appears in the 1st col
cat file | awk '{print $1,$3}'      # print the 1 and 3 cols, regular space separated
cat file | awk '{print $1,"new",$3}'      # insert a new col with text "new"
cat file | awk '{print $3,$1}'      # change the order of 1st and 3rd col
```

for loop on command line

```
for variable in `command`
do
        command 1
        command 2
done
```

The symbol on the tilde key (~)

`

http://en.wikipedia.org/wiki/Grave_accent

Or backtick

```
for x in `find ftp.ncbi.nih.gov/ -name "*faa"`
do
        echo $x
        cat $x | grep '>' | wc -l
done
```

```
for x in `find ftp.ncbi.nih.gov/ -name "*faa"`; do echo $x; cat $x | grep '>' | wc -l; done
```

```
for x in `find ftp.ncbi.nih.gov/ -name "*faa"`
do
        cat $x >> ecoli-all.faa.2
done
```

```
find ftp.ncbi.nih.gov/ -name "*faa"| xargs cat > ecoli-all.faa
```

- Save history of your commands:
history | less
history > hist1

- Send message to other online users
write username (ctrl+c to exit)

- Change your password
passwd

Ctrl+c to tell Shell to stop current process
Ctrl+z to suspend
bg to send to background
Ctrl+d to exit the terminal (logout)

Install external command line softwares

Yanbin Yin

Spring 2013

Programs/tools we often use

- BLAST
- FASTA
- HMMER
- EMBOSS
- lftp
- bioperl
- R
- Galaxy
- Clustalw
- MAFFT
- MUSCLE
- SRA toolkit
- weblogo
- PhyML
- FastTree
- RaxML
- USEARCH
- …

http://gacrc.uga.edu/

Linux-based program types

- Source codes in C, C++, Java, Fortran etc.
  - Need to be compiled before execute the command
- Precompiled executables or binary codes
- Source codes in scripting languages (perl, python, R etc.)
  - Can execute directly

On your own ubuntu machine …

- You are the root and using the sudo command you can install anything you want into the system directory (/usr/bin/, /bin/, /lib/ etc.)
  - apt-get (Advanced Packaging Tool) can do many installations for you from source or binary codes

- On glu, you are not the root and you can only install things under your home using the "hard" or the "most common" way
  - Download->unpack->install->edit PATH environmental variable
  - Make sure you create folders for each tools, e.g. home/tools/fasta

On MAC

http://www.digimantra.com/howto/apple-aptget-command-mac/
http://superuser.com/questions/173088/apt-get-on-mac-os-x

http://www.macobserver.com/tmo/article/install_the_command_line_c_compilers_in_os_x_lion
Install Xcode, then C compiler, then you can install Mac port
http://www.macports.org/install.php
With mac port, you can install
wget: `sudo port install wget`
lftp : `sudo port install lftp`
hmmer: `sudo port install hmmer`
emboss: `sudo port install emboss`
R: `sudo port install R`
blast:
http://www.blaststation.com/freestuff/en/howtoNCBIBlastMac.html

Install BLAST on your own machine

[test if installed]

```
sudo apt-get install blast2
```

[test if installed]

```
blastall
```

[where it installed]

```
which blastall
```

[if you want to uninstall]

```
sudo apt-get remove blast2
```

[test if it's gone]

```
blastall
```

[new version of blast]

```
sudo apt-get install ncbi-blast+
```

[Run the following before and after installation]

```
ls /usr/bin/| wc
```

Use apt-get to install

lftp, emboss, hmmer, bioperl, clustaw, muscle, R

```
sudo apt-get install xxx
```

[to test if installed, type in the command]

fasta

[not available in the package list]

Install FASTA using the common way

[download]
http://fasta.bioch.virginia.edu/

```
wget -q
http://faculty.virginia.edu/wrpearson/fasta/CURRENT/fasta-
36.3.5e.tar.gz
```

[be organized]

```
mkdir tools
cd tools
mkdir fasta
mv fasta-36.3.5e.tar.gz fasta
cd fasta
```

[unpack]

```
tar zxf fasta-36.3.5e.tar.gz
```

[compile/install]

```
cd fasta-36.3.5e/
ls -l
ls -l ../bin/
less README
cd src
make -f ../make/Makefile.linux_sse2
all
cd ../bin/
ls -l
ssearch
cd
ssearch
```

[edit path variable]

```
vi .bashrc
export PATH="absolute path to fasta
bin folder";
. .bashrc [execute the script]
ssearch
```

[add alias of a command ll]

```
vi .bashrc
alias ll='ls -l'
alias lt='ls -lt'
```

Install BLAST using the common way

```
lftp ftp.ncbi.nih.gov:/blast/executables/LATEST> get ncbi-blast-2.2.27+-ia32-linux.tar.gz
tar -zxf ncbi-blast-2.2.27+-ia32-linux.tar.gz
ll
cd ncbi-blast-2.2.27+/bin
ll
./blastp -h
```

Download ncbi-blast-2.2.27+-x64-linux.tar.gz if your machine is 64 bit, to find out

```
uname -a
```

[edit path variable]

```
vi .bashrc
export PATH="absolute path to blast bin folder";
. .bashrc
blastp
```

Install HMMER

```
sudo apt-get install hmmer
```

Hard way:

http://hmmer.janelia.org/software

bioperl

http://www.bioperl.org/wiki/Installing_BioPerl_on_Ubuntu_Server

```
sudo apt-get install bioperl
```

The hard way to install bioperl

```
wget -q http://bioperl.org/DIST/current_core_unstable.tar.bz2
tar -xjvf current_core_unstable.tar.bz2
cd bioperl-*
perl Build.PL   # choose the defaults
./Build test
./Build install
```

http://www.bioperl.org/wiki/Installing_BioPerl_on_Ubuntu_Server

Install MAFFT the hard way

```
wget -q
http://mafft.cbrc.jp/alignment/software/mafft-
7.029-with-extensions-src.tgz

tar xzf mafft-7.029-with-extensions-src.tgz

cd mafft-7.029-with-extensions/core/

sudo make
sudo make install
```

http://mafft.cbrc.jp/alignment/software/source.html

Install Galaxy

http://wiki.galaxyproject.org/Admin/Get%20Galaxy

```
sudo apt-get install mercurial
hg clone https://bitbucket.org/galaxy/galaxy-dist/
hg update stable
cd galaxy-dist
sh run.sh
```

http://localhost:8080

Edit universe_wsgi.ini file to allow access from other computers

Setup admin user:
http://wiki.galaxyproject.org/Admin/Interface
edit universe_wsgi.ini file

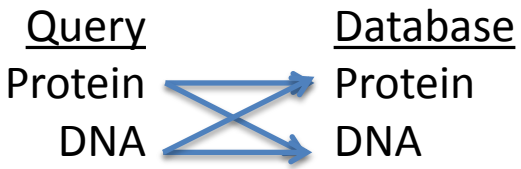
Run BLAST and HMMER in command line

Yanbin Yin

Spring 2013

BLAST

```
blastall - | less
-p # specify blastp, blastn, blastx, tblastn, tblastx

More commands in blast package

formatdb (format database)
megablast (faster version of blastn)
rpsblast (protein seq vs. CDD PSSMs)
impala (PSSM vs protein seq)
bl2seq (two sequence blast)
blastclust (given a fasta seq file, cluster them based on sequence similarity)
blastpgp (psi-blast, iterative distant homolog search)
```

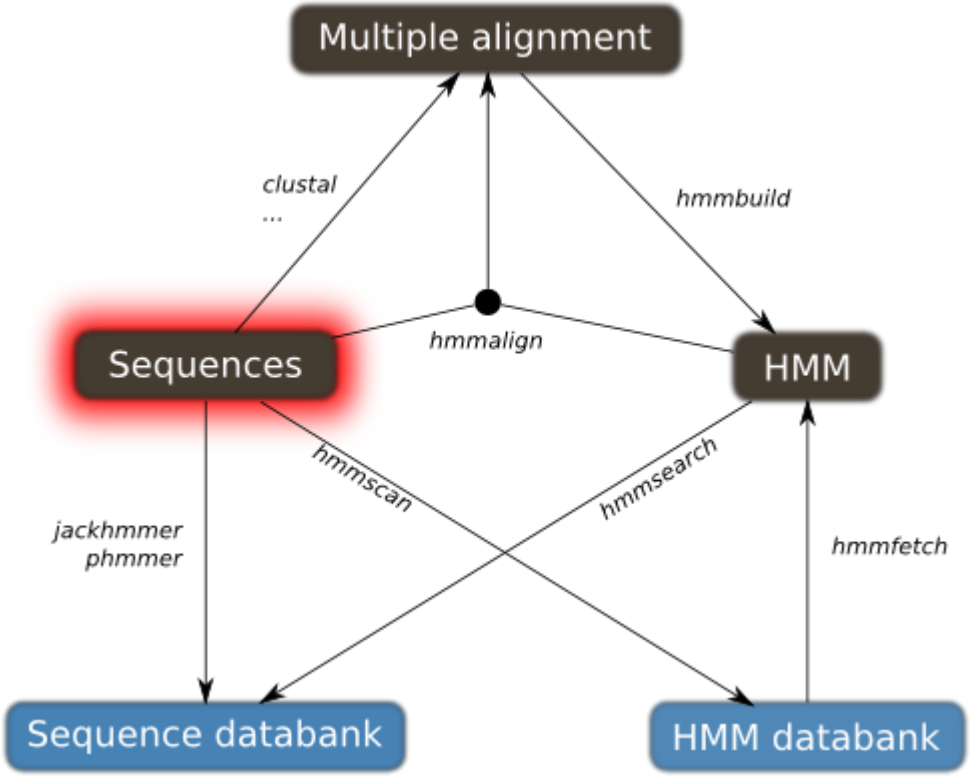


```
formatdb -i ecoli-all.faa
formatdb - # see the options, for nt db, use -p F
less ecoli-all.faa # select the 3rd protein (YP_488309.1)
vi test-query.fa # create a file to store this protein seq

[now blast, which is in your path alreay]
blastall -p blastp -i test-query.fa -d ecoli-all.faa
blastall -p blastp -i test-query.fa -d ecoli-all.faa > test-qery.fa.out

[-m 9, the tabular format output without alignment, easy to parse]
blastall -p blastp -i test-query.fa -d ecoli-all.faa -m 9
blastall -p blastp -i test-query.fa -d ecoli-all.faa -m 9 > test-
qery.fa.out.m9

[-e 1e-2, showing only hits with evalue < 1e-2]
blastall -p blastp -i test-query.fa -d ecoli-all.faa -m 9 -e 1e-2

[Now try something big (and slow)]
blastall -p blastp -i test-query.fa -d
/home/yyin/work/class/metagenemark_predictions.faa -m 9 -e 1e-2 > test-
qery.fa.cowrumen.out.m9 &

[Do some parsing]
less test-query.fa.cowrument.out.m9 | cut -f1,2,3,7- | less
less test-query.fa.cowrument.out.m9 | cut -f1,2,3,7- | grep -v '^#' |
cut -f2 | sort -u | head
```

HMMER: http://hmmer.janelia.org/

What is HMMER? ftp://selab.janelia.org/pub/software/hmmer3/3.0/Userguide.pdf
HMMER is a software package that is used for searching sequence databases for homologs, making protein sequence alignments, and making **profile hidden Markov models** (profile HMMs)**.** It implements methods using probabilistic models called **profile hidden Markov models**, mathematically representing multiple sequence alignments.

Compared to BLAST, FASTA, and other sequence alignment and database search tools based on older scoring methodology, HMMER aims to be significantly *more* accurate and *more* able to detect remote homologs because of the strength of its underlying mathematical models. In the past, this strength came at significant computational expense, but in the new HMMER3 project, HMMER is now essentially **as fast as** BLAST



http://drmotifs.genouest.org/2010/10/sequence-hammering/

```
Go to http://cys.bios.niu.edu/dbCAN/family.php?ID=GH5  and download
wget -q http://cys.bios.niu.edu/dbCAN/data/aln/cazy-family/aln/GH5.aln
less GH5.aln

hmmbuild # list options
hmmbuild -h  # list complete options
hmmbuild --informat afa GH5.hmm GH5.aln # build model, afa: aligned fasta format,
see User Guide page 16 footnote
less GH5.hmm # profile HMM file is a text file

hmmsearch
hmmsearch -h
hmmsearch --domtblout GH5.hmm.cowrumen.dm GH5.hmm metagenemark_predictions.faa >
GH5.hmm.cowrumen.out & # save parseable table of per-domain hits to file

[a little parsing, alignment in GH5.hmm.cowrumen.out]
less GH5.hmm.cowrumen.dm | grep -v '^#' | awk '{print
$1,$3,$6,$7,$12,$13,$16,$17,$18,$19}' | less
less GH5.hmm.cowrumen.dm | grep -v '^#' | awk '{print
$1,$3,$6,$7,$12,$13,$16,$17,$18,$19}' | awk '$6<1e-2&&($8-$7)/$3>.8' | less

Extracting domain regions is easy if using perl and bioperl
```