Linux command line basics II: downloading data and controlling files

Yanbin Yin

Things you should know about programming

Learning programming has to go through the hands-on practice, a lot of practice

Hearing what I describe about a command or a program helps, but you will not be able to do it unless you type in the codes and run it to see what happens

Reading others' codes helps but often is harder than writing it by yourself from scratch

Although painful and frustrating, trouble-shooting is normal and part of the learning experience (ask experienced people or google)

To avoid errors, you have to follow rules; most errors occurred in programming are because of not knowing rules or forgetting rules

Use comments in case you forget what you've written means

write-> run -> errors -> edit-> errors -> ………………………………….. -> run -> success

Good news: finished scripts could be reused or edited for later use

Homework #7

1. Create a folder under your home called hw7

2. Change directory to hw7

3. Go to NCBI ftp site, find the genome, bacteria, ecoli MG1655 folder, and download the ptt file and the faa file in there

4. Create a copy of the ppt file, if the original file is called A.ptt, name the copied file A.ptt.bak
Do the same thing for the faa file

5. Go to https://edu.isb-sib.ch/pluginfile.php/2878/mod_resource/content/4/couselab-html/content.html and finish all quizzes in Chapters 3 to 5.

6. Use what you learned in chapter 5 to count how many protein sequences are in the faa file of step 4.
Write a report (in word or ppt) to include all the operations and screen shots.

Due on Nov 07 (send by email, if there are 2+ files, put them in a zip file; include your last name in the file name)

What we learned last class:

file system,
relative/absolute paths,
working folder, home folder

ssh, pwd, ls
cd, mkdir, rmdir, rm, man
cp, mv

If things go wrong, try:

Ctrl+c (sometimes multiple times)

q to exit from man page

The absolute basics	File control	Viewing, creating, or editing files	Misc. useful commands	Power commands	Process-related commands
ls	mv	less	man	uniq	top
cd	cp	head	chmod	sort	ps
pwd	mkdir	tail	source	cut	kill
	rmdir	touch	wc	tr	
	rm	nano		grep	
	| (pipe)			sed	
	> (write to file)				
	< (read from file)				

http://korflab.ucdavis.edu/Unix_and_Perl/unix_and_perl_v3.1.1.pdf

View files: `more,` `less,` `head,` `tail`

How you use Tab key to autocomplete

`less /home/` then hit tab twice, you will see all files/folders under /home/
`less /home/yyin/` then hit tab twice, you will see …

`less /home/yyin/U` then hit tab once, Unix_and_Perl_course will be autocompleted

`less /home/yyin/Unix_and_Perl_course/` keep doing this until you get

`less /home/yyin/Unix_and_Perl_course/Data/Arabidopsis/At_proteins.fasta`

q: quit viewing   ↑ or ↓: move up or down a line
space: next page   />: search for text '>'
B or PgUp: back a page   F or PgDn: forward a page
n: find next occurrence of 'abc'
G: go to the end   ?: find previous occuence of 'abc'

`See next page for screen shots`

How you use Tab key to autocomplete

```
yyin@ser:~$ less /home/
elfitzek/ yyin/
yyin@ser:~$ less /home/yyin/
.bash_history          .gconf/                .mission-control/
.bash_logout           .gksu.lock             .mozilla/
.bashrc                .gnome/                .pki/
.cache/                .gnome2/               .profile
.compiz-1/             .goutputstream-11LJJX  .pulse/
.config/               .gstreamer-0.10/       .pulse-cookie
current.zip            .gtk-bookmarks         .ssh/
.dbus/                 .gvfs/                 Unix_and_Perl_course/
.dmrc                  .ICEauthority          .Xauthority
Documents/             .lftp/                 .xsession-errors
Downloads/             .local/                .xsession-errors.old
.fontconfig/           __MACOSX/
yyin@ser:~$ less /home/yyin/Unix_and_Perl_course/
Applications/  Data/            .DS_Store
Code/          Documentation/ .profile
yyin@ser:~$ less /home/yyin/Unix_and_Perl_course/Data/
Arabidopsis/     GenBank/         Unix_test_files/
.DS_Store        Misc/
yyin@ser:~$ less /home/yyin/Unix_and_Perl_course/Data/
Arabidopsis/     GenBank/         Unix_test_files/
.DS_Store        Misc/
yyin@ser:~$ less /home/yyin/Unix_and_Perl_course/Data/GenBank/
E.coli.genbank    Y.pestis.genbank
yyin@ser:~$ less /home/yyin/Unix_and_Perl_course/Data/GenBank/E.coli.genbank
```

more is similar to less, but can do less than less

```
more /home/yyin/Unix_and_Perl_course/Data/Arabidopsis/At_genes.gff
```

head to dump the top few lines to the screen

```
head /home/yyin/Unix_and_Perl_course/Data/Arabidopsis/chr1.fasta

head -20 /home/yyin/Unix_and_Perl_course/Data/Arabidopsis/chr1.fasta
```

tail to dump the last few lines to the screen

```
tail /home/yyin/Unix_and_Perl_course/Data/Arabidopsis/intron_IME_data.fasta

tail -20 /home/yyin/Unix_and_Perl_course/Data/Arabidopsis/intron_IME_data.fasta
```

more, less, head, tail do not load all file content to the memory
You can edit the file content either, they are just viewers

Create or edit files

<u>Text editors:</u>

nano
pico
vi

Suppose you are at your home:

Write the top part of the intAt_genes.gff file to a new file

```
head -20 /home/yyin/Unix_and_Perl_course/Data/Arabidopsis/At_genes.gff > head
```

Try nano (Intuitive user interface)
```
nano head
```

Try vi (command-driven interface, but much more power)
```
vi head
```

Create a file from scratch using vi.
1) you type `vi filename` and hit *enter*
2) after you are in *vi*, type *i* to get into edit mode and copy & paste content in *vi*
3) hit *Esc* to exit edit mode and then *:x* to save the file and exit *vi*.

Input and output redirection: **the greater-than sign**

Unix has a special way to **direct input and output** from commands or programs.

By default, the input is from keyboard (called standard input, *stdin*): you type in a command and Shell takes the command and executes it.

The standard output by default is to the terminal screen (*stdout*);

if the command or program failed, you will also have standard errors dumped to the terminal screen (*stderr*).

However, if you do not want the output dumped to the screen, you can use "**>**" to redirect/write the output into a file. For example, try

```
ls /home/yyin
ls /home/yyin > list
ls /home/yyim
ls /home/yyim 2> err
```

"*2>*" to dump the error message
No space here!

vi basics

command mode → **i** → edit mode
← Esc

The following commands **operate in command mode (hit Esc before using them)**

x	delete one character at cursor position
u	undo
dd	delete the current line
G	go to end of file
1G	go to beginning of file
10G	go to line 10
$	go to end of line
1	go to beginning of line
:q!	exit without saving
:w	save (but not exit)
:wq or :x	save and exit
Arrow keys:	move cursor around (in both modes)

http://cbsu.tc.cornell.edu/ww/1/Default.aspx?wid=36

Search and substitution in vi

In command mode, you can do a number of fancy things. The most useful are:

- Search: hit slash ("/") to get the cursor to the left-bottom corner; you can type any word or letter to search it; type *n* to go to the next instance

- Replace: hit *Esc* (at any time, hitting *Esc* to get back to the default status is the safest thing to do) and type "*:1,$s/+/pos/g*" and then enter will replace all "*+*" to "pos".

Try this in
`vi head`

1) hit *Esc* to exit edit mode and then *:q!* to NOT save the file and exit *vi*.

From the first line to the last

all instances in a row

The first field: to be replaced

$$:1,\$s/+/pos/g$$

Substitution

The second field: to replace with

Ready to type in command
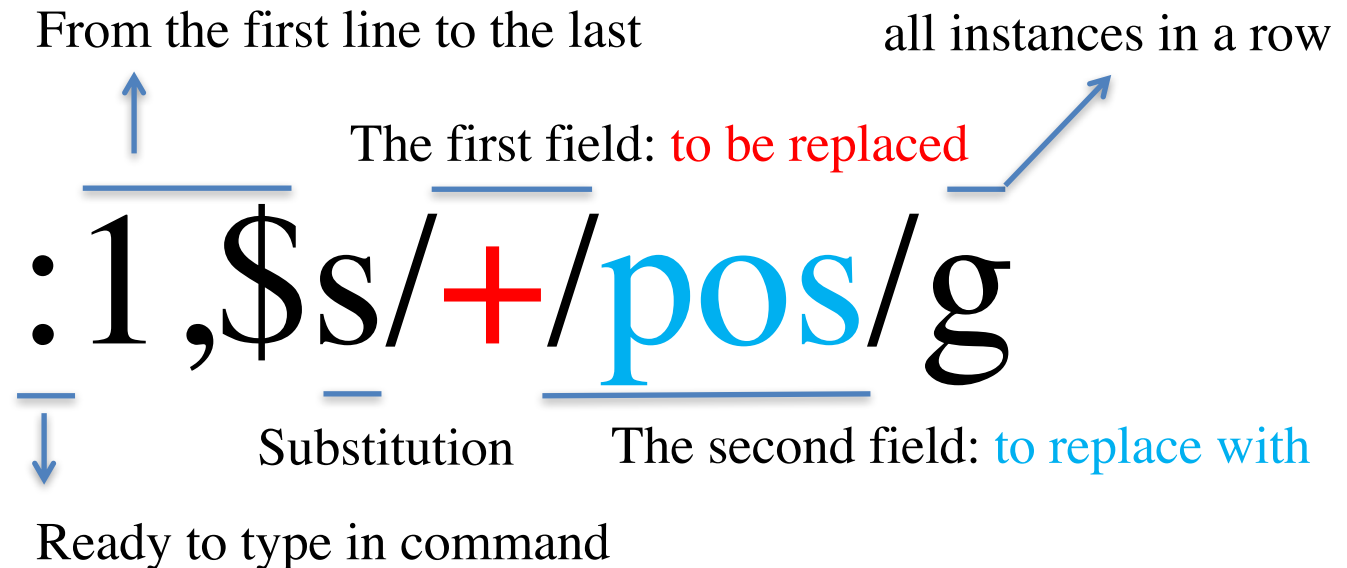
Wild cards and regular expression

Regular expression (regex or regexp) is a very powerful tool for text processing and widely used in text editors (e.g. vi) and programming languages (e.g. Shell commands: sed, awk, grep and perl, python, PHP) to automatically edit (match and replace strings) texts.

Finding and replacing exact words or characters are simple, e.g. the vi example shown above

However, if you want to match multiple words or characters, you will need wildcards or patterns.

a list of commonly used wildcards and patterns:

*   any numbers of letters, numbers and characters except for spaces and special characters, e.g. ()[]+\/$@#%;,?

.   any single letter, number and character including special characters

^   start of a line   caret

$ end of a line

^$ an empty line, i.e. nothing between ^ and $

[]   create your own pattern, e.g. [ATGC] matches one of the four letters only, [ATGC]{2} matches two such letters; [0-9]: any numbers

\w   any letter (a-z and A-Z)

\d   any number (0-9)

+   previous items at least one times, e.g. \w+ matches words of any sizes

{n}   previous items n times, e.g. \w{5} matches words with exactly five letters

\s   space   Curly brackets

\t   tabular space

\n   new line

http://www.bsd.org/regexintro.html

14

Use regex inside vi

```
head -20 /home/yyin/Unix_and_Perl_course/Data/Arabidopsis/At_proteins.fasta > head

vi head
```

Inside vi, try
```
:1,$s/ *//g
```

Hit u to undo

What about
```
:1,$s/ .*//g
```

1) hit *Esc* to exit edit mode and then *:x* to save the file and exit *vi*.

Get data from remote ftp/http website

ftp
lftp
sftp
ncftp

```
lftp addr      command to connect to a remote ftp server
cd dir         change to the directory
cd ..          change to the upper folder (..)
ls             list files and folders in the current directory at once
ls dir         list files and folders in dir at once
ls | less      list page by page (good if the list is too long)
get file       get a file
mirror dir     get a folder
zless file     view the file content
by or bye      exit lftp
```

NCBI ftp site :

`lftp ftp.ncbi.nih.gov`

The prompt will change to:
`lftp ftp.ncbi.nih.gov:/>`

After '>' you can type in command and hit enter:
`lftp ftp.ncbi.nih.gov:/>ls`

The ftp site can also be accessed through a web browser

```
lftp ftp.ncbi.nih.gov:/> ls
dr-xr-xr-x    3 ftp      anonymous        4096 May 27  2011 1000genomes
-r--r--r--    1 ftp      anonymous 10738466816 Dec   6 14:25 10GB
-r--r--r--    1 ftp      anonymous  1074790400 Dec   6 14:25 1GB
-r--r--r--    1 ftp      anonymous        1868 Dec   5 21:45 README.ftp
lr--r--r--    1 ftp      anonymous          29 Dec   5 22:08 asn1-converters -> toolbox/ncbi_tools/converters
dr-xr-xr-x   12 ftp      anonymous      184320 Dec 14 23:00 bigwig
dr-xr-xr-x    4 ftp      anonymous        4096 Jan 28 16:10 bioproject
dr-xr-xr-x    2 ftp      anonymous        4096 Jan 28 09:53 biosample
dr-xr-xr-x   10 ftp      anonymous        4096 May 24  2012 blast
dr-xr-xr-x    3 ftp      anonymous        4096 Sep 13  2004 cgap
dr-xr-xr-x    4 ftp      anonymous        4096 May 25  2011 cn3d
dr-xr-xr-x   30 ftp      anonymous        4096 Jan 14 20:20 dbgap
dr-xr-xr-x   11 ftp      anonymous        4096 Jun  4  2006 entrez
dr-xr-xr-x    7 ftp      anonymous        4096 Oct 13  2011 epigenomics
dr-xr-xr-x    6 ftp      anonymous        4096 Aug  4  2006 fa2htgs
-r--r--r--    1 ftp      anonymous        3262 Dec   5 21:45 favicon.ico
dr-xr-xr-x   12 ftp      anonymous       65536 Dec 21 18:02 genbank
dr-xr-xr-x    6 ftp      anonymous        4096 Dec 13 19:24 gene
dr-xr-xr-x   97 ftp      anonymous        4096 Jan  9 22:54 genomes
dr-xr-xr-x 1073741824 ftp        anonymous       0 Jan 28 20:06 geo
dr-xr-xr-x   25 ftp      anonymous        4096 Sep 20  2011 hapmap
dr-xr-xr-x   13 ftp      anonymous        4096 Jan 27  2012 mmdb
dr-xr-xr-x    6 ftp      anonymous       49152 Dec 21 17:05 ncbi-asn1
dr-xr-xr-x  164 ftp      anonymous        8192 Dec   2 17:49 pub
dr-xr-xr-x   11 ftp      anonymous        4096 Nov 29 13:17 pubchem
dr-xr-xr-x    2 ftp      anonymous        4096 Jan 28 09:03 pubmed
dr-xr-xr-x   15 ftp      anonymous        4096 Jan 14 18:59 refseq
dr-xr-xr-x   57 ftp      anonymous        4096 Aug 20  2008 repository
-r--r--r--    1 ftp      anonymous          26 Dec   5 21:45 robots.txt
dr-xr-xr-x    5 ftp      anonymous        4096 Nov 14 14:07 sequin
dr-xr-xr-x    9 ftp      anonymous        4096 May 24  2010 sky-cgh
dr-xr-xr-x   17 ftp      anonymous        4096 Jan 18 18:23 snp
dr-xr-xr-x   13 ftp      anonymous        4096 May 16  2012 sra
dr-xr-xr-x    2 ftp      anonymous        4096 Sep 29  2004 tech-reports
dr-xr-xr-x   12 ftp      anonymous        4096 Jun 29  2011 toolbox
dr-xr-xr-x    5 ftp      anonymous        4096 Apr 24  2009 tpa
dr-xr-xr-x    4 ftp      anonymous        4096 Sep 13 15:46 variation
lftp ftp.ncbi.nih.gov:/> 
```

ls command:

list files and folders

Where bacterial genomes are in the ftp site?

The end of the page after ls

```
dr-xr-xr-x    2 ftp      anonymous       4096 Dec  6  2010 _Nostoc_azollae__0708_uid49725
-r--r--r--    1 ftp      anonymous  455294518 Jan 28 08:08 all.GeneMark.tar.gz
-r--r--r--    1 ftp      anonymous  102841288 Jan 28 08:14 all.Glimmer3.tar.gz
-r--r--r--    1 ftp      anonymous  227990948 Jan 28 12:02 all.Prodigal.tar.gz
-r--r--r--    1 ftp      anonymous 4532882964 Jan 28 08:42 all.asn.tar.gz
-r--r--r--    1 ftp      anonymous 1515030808 Jan 28 09:03 all.faa.tar.gz
-r--r--r--    1 ftp      anonymous 2190975020 Jan 28 09:20 all.ffn.tar.gz
-r--r--r--    1 ftp      anonymous 2416170041 Jan 28 09:55 all.fna.tar.gz
-r--r--r--    1 ftp      anonymous    9095708 Jan 28 10:29 all.frn.tar.gz
-r--r--r--    1 ftp      anonymous 6720488727 Jan 28 10:45 all.gbk.tar.gz
-r--r--r--    1 ftp      anonymous  548255240 Jan 28 11:37 all.gff.tar.gz
-r--r--r--    1 ftp      anonymous  182962489 Jan 28 11:41 all.ptt.tar.gz
-r--r--r--    1 ftp      anonymous    2600313 Jan 28 11:43 all.rnt.tar.gz
-r--r--r--    1 ftp      anonymous     357580 Jan 28 11:44 all.rpt.tar.gz
-r--r--r--    1 ftp      anonymous 4107354627 Jan 28 11:53 all.val.tar.gz
dr-xr-xr-x    2 ftp      anonymous       4096 Sep 25 04:11 alpha_proteobacterium_HIMB59_uid175778
dr-xr-xr-x    2 ftp      anonymous       4096 Sep 25 04:12 alpha_proteobacterium_HIMB5_uid175779
dr-xr-xr-x    2 ftp      anonymous       4096 Dec 20 05:32 bacterium_BT_1_uid184079
dr-xr-xr-x    2 ftp      anonymous       4096 Dec  6  2010 cyanobacterium_UCYN_A_uid43697
dr-xr-xr-x    2 ftp      anonymous       4096 Dec  6  2010 gamma_proteobacterium_HdN1_uid51635
dr-xr-xr-x    2 ftp      anonymous       4096 Sep  1  2011 halophilic_archaeon_DL31_uid72619
dr-xr-xr-x    2 ftp      anonymous       4096 Aug 20 04:10 secondary_endosymbiont_of_Ctenarytaina_eucalypti_uid172737
dr-xr-xr-x    2 ftp      anonymous       4096 Nov  2 16:10 secondary_endosymbiont_of_Heteropsylla_cubana_Thao2000_uid172738
dr-xr-xr-x    2 ftp      anonymous       4096 Jan 28  2011 uncultured_Termite_group_1_bacterium_phylotype_Rs_D17_uid59059
lftp ftp.ncbi.nih.gov:/genomes/Bacteria> ▮
```

cd ne

Then press tab key to auto-complete or list

```
lftp ftp.ncbi.nih.gov:/genomes/Bacteria> cd Ne
Neisseria_gonorrhoeae_FA_1090_uid57611/    Neisseria_meningitidis_M01_240355_uid162075/
Neisseria_gonorrhoeae_NCCP11945_uid59191/  Neisseria_meningitidis_M04_240196_uid162081/
Neisseria_gonorrhoeae_TCDC_NG08107_uid161097/  Neisseria_meningitidis_MC58_uid57817/
Neisseria_lactamica_020_06_uid60851/       Neisseria_meningitidis_NZ_05_33_uid162077/
Neisseria_meningitidis_053442_uid58587/    Neisseria_meningitidis_WUE_2594_uid162093/
Neisseria_meningitidis_8013_uid161967/     Neisseria_meningitidis_Z2491_uid57819/
Neisseria_meningitidis_FAM18_uid57825/     Neisseria_meningitidis_alpha14_uid61649/
Neisseria_meningitidis_G2136_uid162085/    Neisseria_meningitidis_alpha710_uid161971/
Neisseria_meningitidis_H44_76_uid162083/   Neorickettsia_risticii_Illinois_uid58889/
Neisseria_meningitidis_M01_240149_uid162079/  Neorickettsia_sennetsu_Miyayama_uid57965/
lftp ftp.ncbi.nih.gov:/genomes/Bacteria> cd Ne
```

How to transfer file between a linux and a windows machine?
Use SSH secure file transfer client
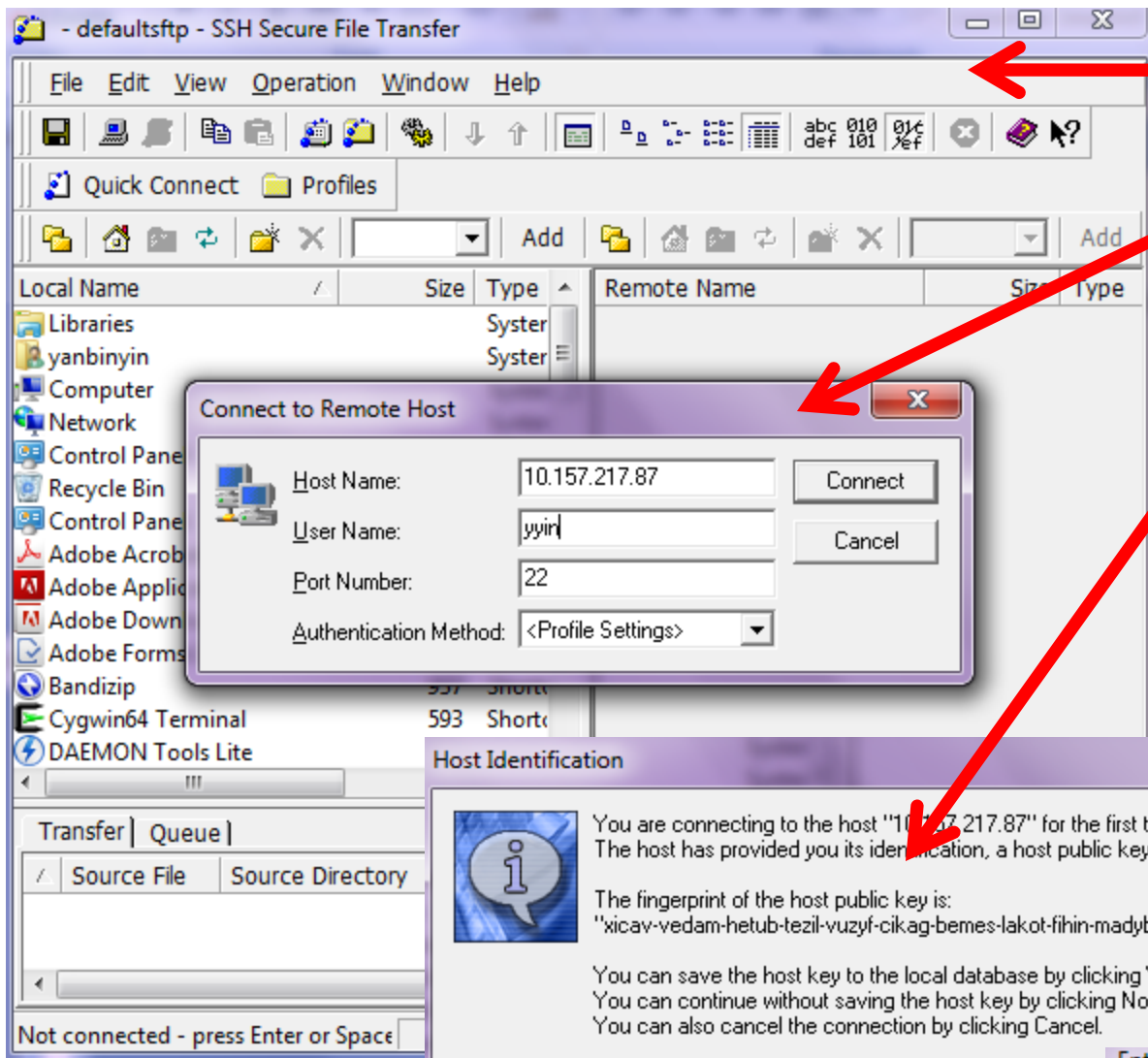
Open the software
Hit enter

Put IP address (10.157.217.25)
Put user name
Hit connect

Choose yes

Put password
Hit ok

If transfer from local to remote: locate your file and drag to the right
If transfer from remote to local: locate your file and drag to the left
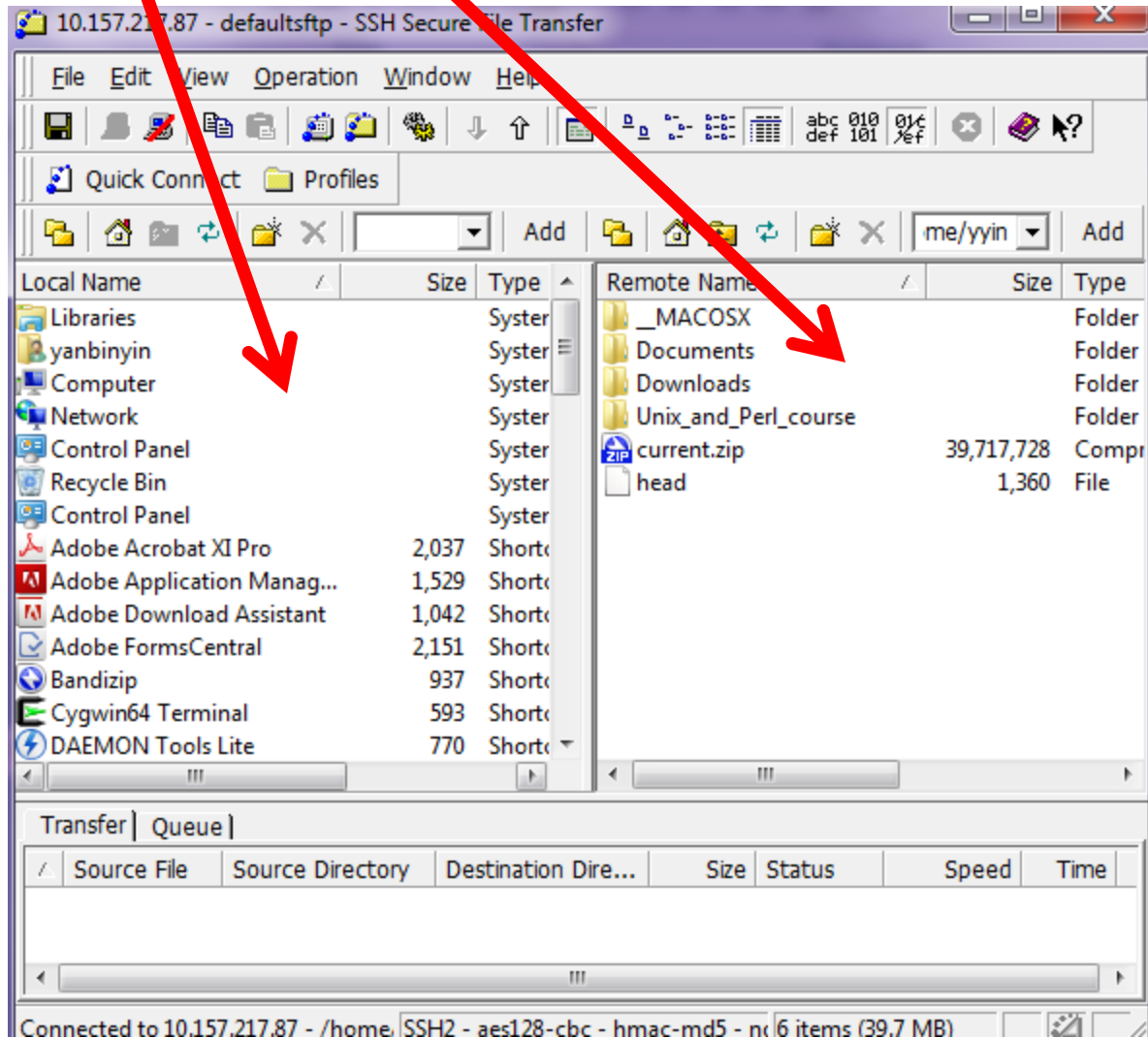
Transfer files between two Linux machines (or mac and linux)
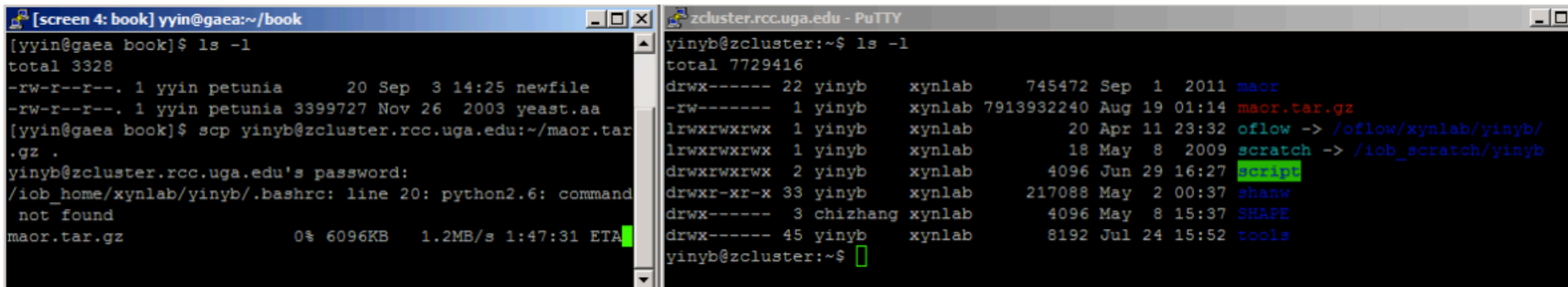
scp: secure copy files/folders between hosts on a network

You are at a Linux or Mac machine, e.g. your laptop with Ubuntu installed and you want to copy some file from ser

Open a terminal in your machine

```
scp yyin@10.157.217.87:/home/yyin/Unix_and_Perl_course/Data/Arabidopsis/At_genes.gff .
```

```
scp username@IP:/path .
```

You will be asked for password on ser

wget

wget is a program useful for downloading files from both FTP and HTTP sites.

wget is non-interactive: you simply enter the necessary options and arguments on the command line and the file is downloaded for you.

You must identify the links first: browse a http webpage or a ftp site and locate the remote files/folders you want to download and then go to the terminal and type

```
wget -q ftp.ncbi.nih.gov/blast/db/FASTA/yeast.aa.gz

wget -r -q
ftp://ftp.ncbi.nih.gov/genomes/archive/old_refseq/Bacteria/Esc
herichia_coli_K_12_substr__MG1655_uid57779

wget -q ftp.ncbi.nih.gov:/blast/executables/LATEST/ncbi-blast-
2.2.27+-x64-linux.tar.gz

wget ftp://emboss.open-bio.org/pub/EMBOSS/emboss-latest.tar.gz
```

It take time to download
Put **&** at the end of command line to put the job to the background

-q quiet
-r recursive (for folders)

Archive and compress files/folders

To save disk space, we can compress large files if we do not intend to use them for a while. A lot of files downloaded from the web are compressed and need to be uncompressed before any processing can take place.

Common compressed formats:

zless to view zipped files

- gzip (gz)

    gzip my_file *(compresses file* my_file*, producing its compressed version,* my_file.gz*)*

    gzip –d my_file.gz *(decompress* my_file.gz*, producing its original version* my_file*)*

- bzip2

    bzip2 my_file *(compresses file* my_file*, producing its compressed version,* my_file.bz2*)*

    bunzip2 my_file.bz2 *(decompress* my_file.bz2*, producing its original version* my_file*)*

Common compressed formats (continued):

- zip

      zip my_file.zip my_file1 my_file2 my_file3 *(create a compressed archive*
            called my_files.zip*, containing three files:* my_file1, my_file2,
my_file3*)*

      zip -r my_file.zip my_file1 my_dir *(if* my_dir *is a directory, create an*
            archive my_file.zip *containing the file* my_file1 *and the directory*
my_dir

            with all its content)

      zip –l my_file.zip *(list contents of the zip archive* my_file.zip*)*

      unzip my_files.zip *(decompress the archive into the constituent files and*
            directories

- tar

      tar -cvf my_file.tar my_file1 my_file2 my_dir *(create a compressed*
            archive called my_files.tar*, containing files* my_file1, my_file2
and the

            directory my_dir *with all its content)*

      tar –tvf my_file.tar *(list contents of the tar archive* my_file.tar*)*

      tar - xvf my_files.tar *(decompress the archive into the constituent files*
            and directories)

            Use man tar to learn more

Common compressed formats (continued):
• tgz (also, tar.gz – essentially a combo of "tar" and "gzip")

tar -czvf my_file.tgz my_file1 my_file2 my_dir *(create a compressed*
archive called my_files.tgz*, containing files* my_file1, my_file2
and the

directory my_dir *with all its content)*
tar –tzvf my_file.tgz *(list contents of the tar archive* my_file.tar*)*
tar -xzvf my_files.tgz *(decompress the archive into the constituent files*
and directories)

Wget the book materials of **Unix and Perl Primer for Biologists**
http://korflab.ucdavis.edu/Unix_and_Perl/

```
mkdir book


cd book
wget http://korflab.ucdavis.edu/Unix_and_Perl/current.zip


unzip current.zip
```


Unpackage the emboss package

```
cd
mkdir tools
cd tools
mv ../emboss-latest.tar.gz tools
tar -zxf emboss-latest.tar.gz &
```

Check disk usage

Disk space is a limited resource, and you want to frequently monitor how much disk space you have used. To check the disk space usage for a folder, use the *du* (disk usage) command

```
yyin@ser:~$ du -hs .
318M    .
yyin@ser:~$ du -hs Unix_and_Perl_course/
131M    Unix_and_Perl_course/
```

To check how much space left on the entire storage file system, use the *df* command

```
yyin@ser:~$ df -h
Filesystem      Size  Used Avail Use% Mounted on
/dev/sdb1       412G  8.1G  383G   3% /
udev             24G  4.0K   24G   1% /dev
tmpfs           4.8G  900K  4.8G   1% /run
none            5.0M     0  5.0M   0% /run/lock
none             24G  156K   24G   1% /run/shm
/dev/sde        2.7T  202M  2.6T   1% /disk4
/dev/sda        2.7T  202M  2.6T   1% /disk1
/dev/sdc        2.7T  202M  2.6T   1% /disk2
/dev/sdd        2.7T  202M  2.6T   1% /disk3
none             24G  144M   24G   1% /tmp/guest-4ynGWF
```

\- Save history of your commands:
```
history > hist1
less 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)