Unix, Perl and BioPerl

Session 1: Introduction to Unix for Bioinformatics

Exercise 1: BLASTing ESTs against a RefSeq database

Goal: Learn the most common Unix commands while manipulating sequence files and "identifying" some rat ESTs by searching RefSeq, an annotated database, with BLAST.

Note: Each command written on multiple lines should be entered as a one-line command, except for actual multiple-line commands, which are delimited with semicolons.

See `http://jura.wi.mit.edu/bio/education/bioinfo-mini/unix-perl/` for course page

#	To do / To answer	Command	Comments
0	Mac OS X: Open the terminal on your computer Mac OS 9: open MacSSH Windows: Open SSHSecureShellClient (See http://jura.wi.mit.edu/bio/education/docs/ssh-sftp.html for more information about these applications.)		If you're running Mac OS X, you're in Unix now.

1	Open your home account on hebrides.	`ssh -l username hebrides.wi.mit.edu`
		
		1st time users:
		`passwd`
			Username is replaced by your's. You will be prompted for your password.
			If it's the first time for connecting, change your password.

2	What is the full path to your home directory?	`pwd`	"print working directory"
3	What files are in your home directory?	`ls`	"list" – won't show hidden files
4	What files (including hidden files) are in your home directory? How big are they?	`ls -a;` `ls -al`	-a option will also show files starting with a dot
5	What's in these files?	`more myfile`	myfile is replaced by a name you got with the "ls" command
6	Create a directory called "unix_class"	`mkdir unix_class`	"make directory"
7	Go to the "unix_class" directory	`cd unix_class`	"change directory"

8	Make directories called "rat-ests" and "dbs", and go to the "dbs" directory.	`mkdir rat-ests;`
		`mkdir dbs;`
		`cd dbs`	

9	Access the NCBI by FTP to get the rat RefSeq sequence "database" (which isn't really a database but rather a multiple sequence file)	`ftp ftp.ncbi.nih.gov`
		
		`[Follow instructions from the FTP site]`
			Note: Sometimes FTP requires a specific user name and password, and other times FTP access can be anonymous (in which case you use "anonymous" as the username and your e-mail as the password)

10	Go to the "refseq" directory	`cd refseq`	
11	List the refseq files. Get the README file, renaming it as refseq_README	`ls -F;` `get README refseq_README;`	-F makes it easier to tell directories from other files. README is often a very helpful file on FTP sites.
12	Go to ./R_norvegicus/ mRNA_Prot	`cd R_norvegicus;` `cd mRNA_Prot`	
13	Get the file "rat.fna.gz" containing the RefSeq set of rat cDNA sequences	`get rat.fna.gz`	The opposite of "get" is "put"
14	Disconnect from the FTP site	`quit`	or "bye"
15	Check to make sure you downloaded what you wanted to get	`ls`	You should have 'refseq_README' and 'rat.fna.gz'
16	Look at the README one screenful at a time	`more refseq_README`	Hit the space bar to advance to the next screenful or 'q' to quit 'more'
17	Unzip the sequence file. What's the file called now?	`gunzip rat.fna.gz;` `ls`	It's generally assumed that a file ending in .gz needs to be unzipped; the opposite is gzip
18	How big is the sequence file?	`ls -l`	
19	How would you list files in order of modification time? (Consult the man pages for ls, using the space bar to advance and "q" to quit)	`man ls`	Extra credit: How would you list in reverse order of modification time (from oldest to newest)?
20	Look at rat.fna to check that it's in FASTA format	`more rat.fna`	FASTA format is a one-line header followed by sequence
21	What are the arguments to use for "grep"?	`grep`	"general regular expression parser" – very useful!

22	Use grep to print all the header lines into a file called rat.headers	`grep ">" rat.fna > rat.headers`	">" marks the beginning of a fasta file
23	Check out the new file to be sure it looks okay at the beginning and the end	`head rat.headers; tail rat.headers`	Add the option –n to print n lines with "head" or "tail"
24	How many sequences are in the sequence file?	`grep ">" rat.fna | wc -l`	wc –l ("word count") actually prints the number of lines
25	Make a BLAST database of the rat.fna sequence file using the "formatdb" command.	`formatdb -i rat.fna -p F -o T`	"formatdb –" prints all options. The options used here are the minimal/usual ones.
26	What files have been created? What does the log file say?	`ls; more formatdb.log`	formatdb.log will show any formatdb errors.
27	Change to the 'rat-ests' directory	`cd ../rat-ests`	Remember that '..' means up one level in the directory tree
28	Get a file of ESTs from /home/george/rat-ests and place it into the directory "rat-ests"	`cp /home/george/rat-ests/* .`	
29	Extract the first sequence and place it into a file by itself.	`head ests.fa; head -8 ests.fa > est1.fa`	Can you think of another way to do this?
30	Run BLAST on the single sequence, using an expect cutoff of 0.5, printing text output (only best 5 hits)	`blastall -i est1.fa -d ../dbs/rat.fna -p blastn -e 0.05 -v 5 -b 5 -o est1_blast.txt`	"blastall" command shows and describes all options. What do these options mean?
31	Run a similar BLAST search but with a default expect value cutoff and generate tab-delimited output	`blastall -i est1.fa -d ../dbs/rat.fna -p blastn -v 5 -b 5 -o est1_blast_tab.txt –m 8`	Similar to the command above, but with "-m 8" added. Did you use the ↑ command to get back to the previous one?
32	Open est1_blast.txt in pico (a simple text editor), using ^X (control-x) to exit.	`pico est1_blast.txt`	Try running BLAST on est1.fa with "–m 9" instead of "-m 8" to find what each field is showing.
33	Extract a sequence (ex: NM_199463) from the BLAST database	`fastacmd –s NM_199463 –d ../dbs/rat.fna`	This only works if you had used the "–o T" option with formatdb

34	Make a file with the five sequence IDs from est1_blast.txt, and extract these sequences from rat.fna	`fastacmd -i list.txt -d ../dbs/rat.fna > myseqs1.fa`	Use pico to make the list. Accessions (ex: NM_133594) or GIs (ex: 19424297) can be used.
35	BLAST the set of ESTs with standard text output	`blastall -i ests.fa -d ../dbs/rat.fna -p blastn -e 0.05 -v 5 -b 5 -o est_blast_1.txt`	Note that BLAST is very fast when searching a database smaller than the default "nt" database
36	BLAST the set of ESTs with tab-delimited output	`blastall -i ests.fa -d ../dbs/rat.fna -p blastn -e 0.05 -v 5 -b 5 -T F -m 8 -o est_blast_tab.txt`	
37	Any questions?		
38	Logout from hebrides and return to your desktop terminal.	`logout`	Make sure the "command prompt" no longer shows something like "username@hebrides".
39 a	If you're using Mac OS X, download the tab-delimited output using sftp and view in it Excel.	`sftp username@hebrides.wi.mit.edu`	Replace "username" with yours. Use cd, ls, and get as before for ftp.
39 b	If you're using Mac OS 9 or Windows, use an application to download the tab-delimited output using sftp and view in it Excel.	`[Follow instructions for Fetch or Mac SFTP.]`	
40	Delete any of your files from the laptop		Thanks!

Notes: In any cases of poor FTP connections to NCBI,
`rat.fna.gz` and the associated `README`
can be copied from
`/home/george/unix/dbs/`