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.

Some commands written on multiple lines should be entered as a one-line commands.

See `http://jura.wi.mit.edu/bio/education/bioinfo2006/unix-perl/` for course page

#	**To do / To answer , Commands , and Comments**
0	**Mac OS X**: Open X11 (In the dock, or Applications >> Utilities >> X11) or the terminal on your computer. If you're running Mac OS X, you're in Unix now. **Windows**: Open Cygwin (preferred) or SSH See the course page for links to more information about these applications.

1	Open your home account on hebrides.
	`ssh -l username hebrides.wi.mit.edu`
	
	replacing 'username' with yours. The switch '-l' is the 12th letter (not "one"). You will be prompted for your password. If it's the first time for connecting, you may get a message about a "RSA key fingerprint" and then asking,
	"Are you sure you want to continue connecting?" You can answer "yes".
	
	Also if it's the first time connecting, change your password.
	
	To change the password whenever you want:
	`passwd`

2	What is the full path to your home directory? `pwd` "print working directory" show the full path to your current location
3	What files are in your home directory? `ls` "list" all files (except hidden files)

4	What files (including hidden files) are in your home directory? How big are they? `ls –a` `ls –al` (the letter l) The -a option will also show files starting with a dot; the -l option shows a long listing
5	What's in these files? `more myfile` “myfile” is replaced by any 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`

9	Change to the ‘dbs’ directory: `cd dbs`
10	(Optional) To get a preview of what you’ll be downloading, open a web browser to view the NCBI’s FTP site. Point it to the rat RefSeq sequence “database”: ftp://ftp.ncbi.nih.gov/refseq/R_norvegicus/ You can view the “README” file and note the “mRNA_Prot” directory. Note that this isn’t really a database but rather a set of multiple sequence files.

11	Download the README file, renaming it as refseq_README:
	
	`wget ftp://ftp.ncbi.nih.gov/refseq/R_norvegicus/README –O refseq_README`
	
	This is a one-line command. The –O refseq_README (the “O” is the letter, not a zero), which is optional, lets you the rename the downloaded file. In any cases of poor network connections to NCBI, `README` can be copied from `/home/george/unix/dbs/`:
	
	`cp /home/george/unix/dbs/README ./refseq_README`

12	Download the “rat.rna.fna.gz” file and rename it to rat.fna.gz. This is a fasta-format file of all rat RefSeq gene sequences: `wget ftp://ftp.ncbi.nih.gov/refseq/R_norvegicus/mRNA_Prot/`

	`rat.rna.fna.gz –O rat.fna.gz`
	
	This is a one-line command. In any cases of poor network connections to NCBI, `rat.fna.gz` can be copied from `/home/george/unix/dbs/:`
	
	`cp /home/george/unix/dbs/rat.fna.gz .`

13	Check to make sure you downloaded what you wanted to get: `ls` You should have 'refseq_README' and 'rat.fna.gz'
14	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'

15	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.

16	How big is the sequence file? `ls -l`
17	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)?
18	Look at rat.fna to check that it's in fasta format `more rat.fna` Fasta format requires a one-line header (starting with ">") followed by sequence
19	What are the arguments to use for "grep"? `grep` "general regular expression parser" – very useful!
20	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-format sequence
21	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"

22	How many sequences are in the sequence file? `grep ">" rat.fna | wc –l` wc ("word count"), with the –l option (for "lines"), prints the number of lines.
23	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. The "-o T" option indexes the sequences in such a way that, given accession(s) you can extract sequence(s) from the big file using 'fastacmd'.

24	What files have been created? What does the log file say?
	`ls`
	`more formatdb.log`
	
	formatdb.log will show any formatdb errors. You can safely ignore any errors that look like
	CoreLib [002.003] FileOpen(".formatdbrc","r") failed
	How many sequences were indexed?

25	Change to the 'rat-ests' directory `cd ../rat-ests` Remember that '**..**' means up one level in the directory tree
26	Get a file of ESTs from /home/george/rat-ests and place it into the directory "rat-ests" `cp /home/george/rat-ests/* .` Note the final "." which indicates that you want to copy the files into the current directory.

27	Extract the first sequence and place it into a file by itself.
	`head ests.fa`
	`head -8 ests.fa > est1.fa`

28	Run BLAST on the single sequence, using an expect cutoff of 0.05, printing text output (only the best 5 hits): `blastall -i est1.fa -d ../dbs/rat.fna -p blastn -e 0.05 -v 5 -b 5 -o est1_blast.txt` This is a one-line command. The command "blastall" shows and describes all options. What do the options mean in the command above?
29	Remembering that you use the ↑ to get back to the previous command, 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 9` This is similar to the previous command, but with "-m 9" added. Running BLAST on est1.fa with "–m 8" instead of "-m 9" performs the exact same search but creates an output file without the header line.
30	Open est1_blast.txt in pico (a simple text editor), using ^X (control-x) to exit. `pico est1_blast.txt`
31	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 when indexing rat.fna with formatdb

32	Make a file with the five sequence IDs from est1_blast.txt, and extract these sequences from rat.fna
	
	`tail -5 est1_blast_tab.txt | cut -f2 > list1.txt`
	`fastacmd –i list1.txt –d ../dbs/rat.fna > myseqs1.fa`
	
	The command 'cut' can cut out a field (by number, after the '-f') from a tab-delimited file. Accessions (ex: NM_133594) or GIs (ex: 19424297) can be used (or both).

33	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.txt` This is a one-line command. Note that BLAST is very fast when searching a database smaller than the default "nt" database
34	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 9 -o est_blast_tab.txt`
35	Any questions?
36	Logout from hebrides and return to your desktop terminal: `logout` Make sure the "command prompt" no longer shows something like "username@hebrides".
37	Copy the BLAST output files from hebrides to your laptop using 'scp' (secure copy): `scp username@hebrides.wi.mit.edu:/home/username/unix_class/rat-ests/* .`

	replacing username with yours. This is a one-line command.
38	Look at the files you downloaded with Unix commands or with Mac/Windows software. Note that the file(s) will be in the directory where you ran the last command, probably C:/cygwin/home/student (Windows) or /Users/student (Mac)
39	Delete any of your files from the laptop. Thanks!

Notes:

1 – If, when trying to use the 'pico' text editor, you get a message that hebrides or barra doesn't know anything about your terminal, use the command
`setenv TERM vt100`
and try again with `pico`.