

Unix, Perl, and Python

Session 1: Introduction to Unix and LSF

Exercise 1

Goal: Given 100 assembled mouse transcripts, I would like to know which of these represent known proteins, so I can BLAST them against known human proteins in a database such as NCBI's Refseq. To help characterize these transcripts, I may also want to BLAST them against protein databases of other species.

Note: Commands are in the `Courier` font and sometimes appear over multiple lines. Each command should be executed as a single line on tak.

1	<p>The X Windows System provides a graphical user interface (GUI) for computers. Mac comes with the program X11, an X windows program. If you are using Mac, open terminal with Go->Utilities->Terminal, and login to tak with <code>ssh -Y username@tak</code> -Y option will allow you to use X windows. For the PC, we use two software products: Xming and PuTTY. Xming is an X Window Server, which can display tak GUI programs on your local desktop. PuTTY is a free SSH client, which allows you to connect to tak. X server (Xming) must be started before setting up a SSH connection with PuTTY. Refer to the website on how to connect your PC to tak: http://jura.wi.mit.edu/bio/education/docs/ssh-sftp.php</p>
2	<p>What is the full path of your current directory? <code>pwd</code></p>
3	<p>Make a new directory called unix-exercise to store your files today. Enter this directory. <code>mkdir unix-exercise</code> <code>cd unix-exercise</code></p>
4	<p>Copy seq.fa file from our website to your working directory. Check if seq.fa is in the folder with ls. Make ls command write out file sizes in human friendly mode. <code>wget</code> http://jura.wi.mit.edu/bio/education/hot_topics/Unix_Perl_Python/ex/seq.fa <code>ls -lh</code></p>
5	<p>seq.fa is in fasta format. Look at the sequence file and see how the sequences are separated? What types of sequences are in this file - genes or peptides? <code>more seq.fa</code> When you're finished with <i>more</i>, type q to exit from the command.</p>

6	<p>How many sequences are in the fasta file?</p> <pre>grep \> seq.fa wc -l</pre> <p>Note: > has a special meaning in the Unix shell; it means redirect the output of a command to a file. We can add \ in front of > to make the <i>grep</i> command look for this character.</p>
7	<p>We are going to BLAST the sequences against human, mouse and rat RefSeq proteins with the <i>blastall</i> command. The <i>blastall</i> command, however, is set up differently from most commands on tak. When you type <i>blastall</i> (and hit <return>),</p> <pre>blastall</pre> <p>tak treats it as the command <i>bsub blastall</i>. This is because <i>blastall</i> can use lots of memory on tak, so the command is sent directly to the cluster. As with any command sent to the cluster, you will receive an email when the command finishes. In this case, the email should contain details of how to use the <i>blastall</i> command. <i>blastall</i> can perform several types of BLAST as described here:</p> <p>http://www.ncbi.nlm.nih.gov/BLAST/blast_program.shtml.</p> <p>BaRC has downloaded and formatted many databases (see http://iona.wi.mit.edu/bio/databases/list.php) that are commonly used for BLAST searches.</p> <p>For easily parsing the BLAST result, we are going to use -m 8 option to have the output in tabular format. We also use an expectation value of 1e-10 to increase the stringency.</p> <pre>blastall -p blastx -e 1e-10 -m 8 -i seq.fa -d hs.faa -o hs_blast.txt</pre> <pre>blastall -p blastx -e 1e-10 -m 8 -i seq.fa -d rn.faa -o rn_blast.txt</pre> <pre>blastall -p blastx -e 1e-10 -m 8 -i seq.fa -d mouse.faa -o mm_blast.txt</pre>
8	<p>After the jobs are submitted to the cluster, how do you check the status of your jobs?</p> <pre>bjobs</pre> <p>When a command (“job”) finishes on the cluster, you will receive an email telling how the job performed and if any errors or warnings have occurred. If your job was done correctly, the email should include "Successfully completed".</p>
9	<p>List the files in your directory to see if the BLAST output files exist.</p> <pre>ls</pre> <p>Look at the top 10 lines of <i>hs_blast.txt</i>.</p> <pre>head hs_blast.txt</pre> <p>You can see that the BLAST result file does not have header line, so we include the column descriptions here:</p> <ol style="list-style-type: none"> 1. query id 2. subject id 3. percent id 4. alignment length 5. number of mismatches

	<p>6. number of gap openings 7. query start 8. query end 9. subject start 10. subject end 11. expect value 12. bit score</p>
10	<p>Each line in the BLAST output represents a single BLAST alignment. If a query sequence has multiple BLAST hits, there would be several lines with the same query sequence (first column)</p> <p>What are the names of the query sequences mapped to human RefSeq proteins? Please save the unique gene names to a file. How many are there? Please run the same analysis for the BLAST results from rat and mouse.</p> <pre>cut -f1 hs_blast.txt sort -u > query_has_hs.txt cut -f1 rn_blast.txt sort -u > query_has_rn.txt cut -f1 mm_blast.txt sort -u > query_has_mm.txt wc -l query_has_hs.txt wc -l query_has_rn.txt wc -l query_has_mm.txt</pre>
11	<p>The transcript sequences that had BLAST hits in all three species encode proteins that are probably more conserved than the others. Can you find the names of these query sequences?</p> <pre>cat query_has_*.txt sort uniq -c sort -k1,1nr</pre>