Bio 4342 Lab practice

Unix problems. Do each problem by yourself or with a neighbor. Proceed at your own pace and ask for help if you need it. Some of the problems will NOT be the easiest way to accomplish the task at hand but will be done in a way to practice various techniques. Please try to follow the examples. I have added room to write notes. These problems are designed to give you an introduction into several techniques that you will be using in the second half of the semester. Question 6 is an optional question that you may work on if you want to try and qualify as a "unix bioinformatics geek".

1. Connecting to and setting up goose:
    a. Start by logging into your goose account
    b. For security step one should be to change your password. What is the command to change your password on a unix based system? Use it to set a new password for your account.
    c. Also to set up your accounts enter the following two commands (you only need to do this only once):

       cp ~chris/.tcshrc ~/

       source ~/.tcshrc

2. Finding and retrieving sequences from online databases
    a. Start up the mozilla browser on goose and do the following:
    b. Go to NCBI (national center for Biotechnology information at www.ncbi.nih.gov this is a good place to start for many Bio-

informatic databases), go to locus link (listed in the right) and see if you can follow the right links to get to the DNA sequence file for the mRNA for Myoglobin in Cow.

c. Save the file to goose in fasta format.

3. Blast searching using web pages

   a. Use explorer on you mac and search the swissprot database for proteins similar to the cow myoglobin, use the goose web page (goose.wustl.edu/blast/blast.html) to do the search. For this type of search I recommend using explorer, as the safari browser will often "time out" and fail to load the results if search take too long.

      i. You will need to view the protein sequence you recovered above (cow myoglobin) somehow and copy it into your clipboard. How do you see the contents of a file?

      ii. use explorer on your mac and go to the blast page (goose.wustl.edu/blast/blast.html)

      iii. Paste your protein sequence into the sequence box

iv. Select the correct blast program and database to search with your sequence

v. Hit the search button to start the search

vi. From the results what protein in the database is the most similar to cow myoglobin that is NOT a myoglobin from another species.

b. Go to NCBI (www.ncbi.nih.gov) and find the sequence of the protein based on the information you see for this sequence

i. Find the sequence to this protein in fasta format and save a copy of the sequence to your PB desktop

c. Do everything necessary to transfer the sequence to your home directory on goose.

d. Look at the contents of the fasta file on goose, and show the results to chris.

4. Using RepeatMasker

a. Attached is a copy of the mini-help page for RepeatMasker. This is what you see if you type RepeatMasker with no arguments (a not uncommon tactic with many Unix programs

that have lots of arguments and/or require input files). RepeatMasker is on goose and is used to analyze and "remove" repetitive DNA sequences from genomic DNA prior to many analysis programs.

b. to start, login to goose and create a directory to hold some sequence files
c. copy the three fasta files from the tostudents directory in chris' home directory into the directory you created above. These three files are genomic sections from Human (hs.fasta), Drosophila (dm.fasta) and nematodes (ce.fasta). We will run RepeatMasker on each of these sections and examine the results.
d. The default repeat library used is a database of repeats found in primates. To run RepeatMasker on a primate sequence simply type `Repeatmasker filename.fasta`.
    i. Run RepeatMasker on the human genomic clone you saved in step 1 above.
    ii. How many files were created.
    iii. Take a look at the contents of each file created.

iv. What fraction of the Human DNA was removed as being repetitive?

e. What command line argument is necessary to have RepeatMasker search for melanogaster or C elegans repeats?

f. Rerun RepeatMasker on the melanogaster and elegans sequences.

g. What fraction of DNA was repetitive in the melanogaster clone?

h. What fraction of DNA was repetitive in the elegans clone?

5. The genome browser at University of California Santa Cruz.

   a. Start your web browser on your PB and go to the URL: genome.ucsc.edu.

   b. click on “Genome Browser” on the left side panel.

   c. Lets say we are interested in looking at the genomic region around the gene for the human homolog of the histone methyltransferase Su(var)3-9 from drosophila. Just like with LocusLink or NCBI, searches here are broad and only work sporadically. Make sure “Human” is the selected organism and

lets try “histone methyltransferase”. Even though there are some genes proposed these are not the right gene. Go “back” and enter the gene name SUV39H1..This should result in a list that includes this gene. To view the genomic region around this gene, follow the link.

d. You can adjust the width of the image I have found a width of 1000 to work well on the PB. Type 1000 in the width box and click the “Submit” button. You can click on any entry to get a view of this genomic region around your gene of interest.
e. Each of the horizontal lines represents the location of some kind of analysis mapped onto this genomic region. You can get help on any line by clicking on the grey box along the left edge.
f. There are many ways to control the output:
    i. The buttons above the picture allow you to move to the left or right and zoom in or out. The menus below the picture allow you control the level detail given for any of the information in the picture.
    ii. Practice controlling the picture by
        1. get a view of several genes in the region,
        2. get more information on the repeats in the region

3. get more details on the SNP's in the region.
4. how many spliced EST's have been sequenced for your gene of interest?

iii. What happens when you click the title for a given data line? Try clicking on "Repeating Elements by Repeatmasker" several times.

iv. What happens when you click on a given box?

1. Use the menu's at the bottom to set "Repeatmasker" to dense and click the "refresh" button. Click on a box under "Repeating Elements by Repeatmasker"
2. Now click on a box under "Repeating Elements by Repeatmasker". What was the result? Notice that in "dense" mode clicking on a box set the display to "full" mode, but clicking on a box in "full" mode links to an information page.
3.

6. Command line blast searches.

a. Set your current working directory to /db2

i. List out the directories here. The nt and nr directories are the databases you will search when you use the command line blast program. “nt” in the nucleotide database while “nr:” is the non-redundant protein databases.

b. The program on goose to run blast searches is called blastall because it runs all 5 version of BLAST

c. You can use any browser you choose and go here: `goose.wustl.edu/~chris/blastall.html`

d. You might want to bookmark this page for future use. It describes the command line structure for blastall. Most things will work exactly s described. The main difference is that you will need to give the full path to the database you will to search.

Eg `-d /db2/nt/nt` will search the set of files that make up the non-redundant nucleotide database

-d `/db2/nr/nr` will search the non-redundant protein database

We expect that Drs. Brent and/or Buhler will discuss the various parameters than blast will take and how they affect your search but for now you may use the default values.

e. Use your cDNA sequence for cow myoglobin to search the non-redundant nt database. Remember that your myoglobin is a protein sequence and non-redundant is DNA, which program would you use in this case? If you see this warning you can ignore it:

"[blastall] WARNING: [000.000] RNA: Could not find taxdb.bti"

f. Repeat one of the searches but force the output to a file in HTML format. Copy the output file into the directory called "public_html" in your home directory. You can use your browser to open the file and view it, to do this point your browser to http://goose.wustl.edu/~name/filename. See the example of a file in my public_html above in section c