Problem Set #1 CH391L Bioinformatics Marcotte due 9/16/2004

Getting started on LINUX

For these problems, you’ll be using your computer accounts on the chemistry department computer *argon* (those of you with your own accounts are welcome to use those). You can log into argon using programs running SSH, which is available for free download from the web link given on the course web page. You should be able to log in to argon by entering the name of the computer (argon.cm.utexas.edu) into ssh, along with the username & password you got in class.

Unlike a PC or Mac, if you log into argon from another computer, you’ll be typing commands in one at a time (i.e. using a command line interface), instead of using a mouse to select operations from menus and folders. However, LINUX is basically organized like a PC/Mac, in that files are stored in folders (called *directories* in LINUX), which can in turn be in other folders, etc. The list of folders that identify where a file is stored is called the *path*, and a typical path might look like /usr/bin/perl. This path means that the file named perl (the name of the perl interpreter program that usually runs your own perl scripts) is stored in the directory named “bin” which is stored in the directory named “usr” which is stored in a special directory named “/”. The top directory in LINUX is always named “/”—all of the other slashes in the path are only there to indicate where one directory name stops and the name of the next directory starts. For those of you who haven’t used LINUX before, here’s a quick summary of the most common commands.

Command	**Function**
ls	list the files stored in the current directory
pwd	print out the name of the current directory
cd /usr/bin	change directories so that you are now in /usr/bin
cd ..	move to the next higher directory
mkdir xxxx	create a subdirectory named xxxx inside your current directory
cd xxxx	move into the subdirectory named xxxx
rmdir xxxx	delete the subdirectory named xxxx **(***USE CAUTION !***)**
ee myfile	edit (or create) the file named myfile with the ee text editor

ee will let you edit/create your programs. The commands are listed at the top of the screen (“^” is the control key, so “^[“ means hold down the control key & type “[“).

more myfile	print the file myfile to the screen. Use the space bar to advance one page or the return key to advance one line.
chmod +x myfile	give permission to run the file named myfile
chmod –x myfile	remove permission to run the file named myfile
man command	read a description (unfortunately, usually a difficult to understand description) of the LINUX command named command
rm myfile	delete the file named myfile **(***USE CAUTION !*** LINUX will delete virtually any file with no questions asked!)**
mv myfile myfile.pl	rename the file myfile as myfile.pl
mv myfile.pl xxxx/myfile.pl	move the file myfile.pl into the subdirectory xxxx
perl myfile.pl	execute the perl program named myfile
./myfile.pl	execute the perl program named myfile (this requires the first line of myfile.pl to be “#!/usr/bin/perl –w”

Detailed instructions for ee can be found on the web, such as at
http://www.ucl.ac.uk/is/unix/ee.htm
There are also other text editors you can use to write your program (such as vi), but ee is by far the easiest for non-specialists.

Log into your LINUX account & play around a bit, creating files & subdirectories, to get used to it. Now, on to some problems. For each, turn in all of the products of your analysis (e.g., for # 1, turn in the frequencies, for # 2, turn in a printout of the program and the frequencies, and so on.)

A Couple of Perl Problems

1. Write a short Perl program to calculate the frequencies of nucleotides in a DNA sequence. This can be the program from the Perl primer or a program of your own construction. Run it on the *E. coli* genome and the *T. volcanium* genome (you can get the nucleotide sequence files from the CH391L web page and transfer it to your argon account using the file transfer program in SSH). Turn in the nucleotide frequencies of the two genomes (& your program, if it was your own design).

2. Write a short Perl program to calculate the frequencies of all dinucleotides in a DNA sequence. Again, this can be based on the program example in the Perl primer or of your own construction. Run it on the *E. coli* genome. Turn in the program and the dinucleotide frequencies of the *E. coli* genome that are output when you run the program.

3. Run your Perl program from problem #2 on the genome of *T. volcanium*. This can also be downloaded from the CH391L web page. Turn in the observed dinucleotide frequencies.

4. Calculate the dinucleotide frequencies that you would expect for *E. coli*, based upon the frequencies of the single nucleotides in *E. coli*. Are the observed dinucleotide frequencies of the *E. coli* genome that you found in problem 2 consistent with what you expected for the dinucleotide frequencies? If not, speculate what might account for the difference. Turn in the expected dinucleotide frequencies & your speculations.

5. Run your Perl program from problem #2 on the 3 mystery gene DNA sequences that you can download from the CH391L web page. Print out the dinucleotide frequencies of each. Based on the observed dinucleotide frequencies, guess which genome each of the 3 genes is taken from. Turn in the genes' dinucleotide frequencies & your guesses.

Scale of biological data

Just to get a feeling for the scale of some of this data, use the LINUX command:
wc *dnasequence*

where *dnasequence* is the name of one of your DNA sequence files, to measure the size of the *E. coli* genome and the *T. volcanium.* The output of the *wc* command (which stands for "word count") consists of three numbers. The first number is the number of lines in the file, the second is the number of words, and the third is the number

of characters. So, in this case, the third number corresponds to the number of nucleotides in the DNA sequence + the number of carriage returns. The first number will tell you the number of carriage returns.
6. Report the total number of nucleotides in both genomes.

Now, try the LINUX command:
ls –l
to measure how much space these files occupy on your hard disk. This command will list every file in your directory, along with some details about it, such as the permissions for reading/writing/executing and the size of the file. The size will be written in the 5th column, in units of bytes.

7. Report the amount of hard disk space occupied by each genome. How does this compare to the number of nucleotides in each file?

8. Given the answers to 6 & 7, how many bytes does it take to store a nucleotide in the format we have them in? On a 20 Gb (20 gigabyte, or $20x10^9$ bytes) hard drive, how many times would the human genome fit? Could I store my entire genome on a 128 Mb memory stick (such as the ones you can carry on your keychain)? There is a notion that within the next 20 years, it will be technologically & economically possible to sequence a complete human genome for ~$1000. How much space would it take to store the genomes of everyone in Austin (assuming we were foolish enough to store the data without compressing it somehow)?

9. How many times larger is the human genome sequence than the *E. coli* genome? The *E. coli* genome contains ~4,500 known genes, and the current estimate for the human genome is ~35,000-45,000 genes, or around 10 times more genes than in *E. coli*. Making the assumption that the genes of *E. coli* & human are about the same size (which isn't really quite true), calculate the density of genes in the two genomes

Lastly, Some Substitution Matrix Problems
Some exercises to familiarize you with the properties of protein sequences. Consult the BLOSUM50 substitution matrix in the 9/7/2004 class handout to answer these.

9. Which amino acid is most likely not to be substituted for by another?

10. Which amino acids are most easily substituted for by others?

11. What are the most disfavored amino acid substitutions?

12. Exercise 2.1 from Durbin *et al.*:
Amino acids D, E, and K are charged; V, I, and L are hydrophobic (greasy).
What is the average BLOSUM50 score within the group of 3 charged amino acids?
Within the 3 hydrophobic amino acids?
Between the 2 groups?
Suggest reasons for the pattern observed.