UNIX bioinformatics exercises

A) In class

Week 1:

Open a terminal emulator (puTTY in the PC clients).

Log into the remote server,

ssh -X [yourusername]@biocluster.igb.uiuc.edu

Initialize a session on a cluster node:

qsub -I

Set up a subdirectory of your home directory called gene_1

Initialize an editing session using nano or nedit, for a file called gene_1.fna

Using the NCBI website from a browser on your PC (www.ncbi.nlm.nih.gov), recover the nucleic acid sequence for the mRNA encoding accession number AF159587, in FASTA format

Paste this sequence from the browser into your terminal window with the text editor open and save the text file to your home directory. Close the editor.

Then recover the protein sequence corresponding to this mRNA, and save this also as a different text file.

Use the program blastn to search the Arabidopsis genome database for hits to the nucleic acid sequence. Is this sequence from this organism?

Arabidopsis genome database:
/home/classroom/cpsc565/Athaliana_167.fa

What can you deduce about the intron / exon structure of this gene? Try a search using fasta36 with and without the -A option.

Use the program blastp to search the protein sequence against the Arabidopsis protein database, and to create a text file from the results.

Arabidopsis protein database:

/home/classroom/cpsc565/Athaliana_167_protein.fa

What kind of protein is this? Is it related to any other classes of protein? Do they all have the same biological function? Hint, you may need to use the nr database. Note that this will take a while. (/home/mirrors/NCBI/BLAST_DBS/latest).

find out:

1) Which organisms have this family of genes
2) What is the generic name for them

B) On your own or at home

You should be able to complete the following tasks on your own over the course of the **next two weeks** (no need to do them all by next week, we will cover some more of the necessary material then). You should work on these in your own time, after you finish the classwork, or in the tutorial session on Wednesday. Write up a brief description of how you solved each problem.

1) Downloading data

Download DNA sequence, protein sequence and BLAST databases of your own choice from either NCBI or your organism-specific site of interest. How can you do this remotely, directly into the server machine?

2) Using data

Run a BLAST search against your downloaded database with a gene of interest. Do a FASTA search against a downloaded sequence file. Redirect output to a file. Make a BLAST database from a FASTA protein or DNA sequence file. Run a BLAST search against your own database.

3) Downloading programs

On the remote server, download the HMMER ***source code*** from:

http://selab.janelia.org/software/hmmer3/3.0/hmmer-3.0.tar.gz

unzip and untar the downloaded archive. The archive contains source code – text files of programs written in the C programming language. To turn this into an ***executable, binary*** file you need to ***compile*** the source code. This process allows

you to run a program on a computer with any type of processor (Mac, Intel PC, AMD, Alpha, SPARC, P690, etc) as long as it is running the UNIX operating system.

4) Installing programs

Make a ~/bin directory. Install the HMMER binaries you made into it. Put it in your path using .bashrc, and use it as you would use any program already installed on the system.

Unix Exercises – Week 2.

1)
Use grep to find 1) the name of the input sequence and 2) the top hit from a BLAST output file without viewing it directly.

2)
Archive the contents of your home directory using tar. Compress the tar file with gzip. Now uncompress and unarchive the .tar.gz file using tar on one command line.

3)
Use ps, w and top to show all processes that are executing.

4)
Combine ps -fae with grep to show all processes that you are executing

5)
Use locate to find all filenames that contain the word blast. Can you combine this with grep to avoid displaying all filenames containing the word debug?

6)
Use the HMMER package to search the exampleprotein.txt sequence against the Pfam_fs database in /usr/BLAST/db. Redirect output to a file. Use the tail –f command to look at the file while the process is running. What happens?

7)
Using clustalw on the command line, align the protein sequences in exampleproteins.txt. Are these related? How closely?

8)

Using the HMMER package you compiled last week, create a Hidden Markov Model from your alignment. Use this to search the exampleproteins.txt file with hmmsearch to see examples of positive matches. Then if time try searching the file /usr/BLAST/db/testproteins.fasta (will take a while).