Unix, Perl and BioPerl

Session 2: Sequence analysis with Perl (introduction)

Exercise 2: Retrieving a set of genes encoding growth factors and extracting their proximal promoters

Goal: Learn the some common Perl commands and file handling while parsing genome mapping data, generating sequence files, extracting some promoters, and analyzing them.

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

To do:

1	Log onto hebrides and enter the `unix_class` and `perl_1` directories
2	Copy starting script and data file from /home/george/perl_1 script = `analyzeGFs.pl` data file = `mapped_growth_factors.txt`
3	Check permissions of script and change to 744 if necessary. `chmod 744 analyzeGFs.pl`
4	Look at the data variables at the top of the script and, if you want, change the length of the promoter you'll be extracting.

5	Look at the format of the input data (`mapped_growth_factors.txt`), which is as follows:
	
	`Input file contains the following fields:`
	`0: RefSeq accession`
	`1: gene name`
	`2: date of assembly used for mapping data`
	`3: chromosome (ex: chr10)`
	`4: strand (+ or -)`
	`5: transcript end 1 (start if strand = +; end otherwise)`
	`6: transcript end 2 (end if strand = +; start otherwise)`
	`7: gene symbol`
	
	This file was created from Gene Ontology annotation, together with UCSC mapping data (using an SQL query of relational databases).

6	Run the script, check the output, and debug.
7	Debugging option: To speed up script execution time, you may wish to create a file of only the first few lines of data and process that file until your script is debugged.
8	Under the line "# 0": Use fastacmd to extract the sequences from a BLAST-formatted database. Syntax for fastacmd is `fastacmd –d database –s accession`

	where database in this case is "nt", the default GenBank DNA database. Redirect output to appropriate file.
9	The code for extracting a promoter is placed in a subroutine called `getUpstreamSeq` The line `getUpstreamSeq();` calls the subroutine, which appears below in the script. Skip down to the subroutine.
10	Note: The two cDNA ends given in the input file are listed from smaller to larger nucleotide along the appropriate chromosome, regardless of strand on which the gene appears. As a result, the ends of a gene on the "+" strand are listed as the 5'- 3' direction, whereas the ends of a gene on the "–" strand are listed in the 3'-5' direction. Thus to extract "upstream sequence", one must extend the cDNA chromosome location at the appropriate end, based on the strand.
11	Under line "# 1": Extract the "promoter" sequence (defined here as the 1000 nt upstream of the cDNA) using the "nibFrag" command. nibFrag can rapidly extract a region of a chromosome if that chromosome file has first been indexed as a "nib" file for the BLAT (not BLAST) tool – the tool that was initially used to generate the mapping data for the entire genome. See http://iona.wi.mit.edu/bio/bioinfo/docs/blat_barra.html for more information. The nibFrag syntax is `nibFrag chrFile toExtractStart toExtractEnd strand outputFile`
12	Under line "# 2": Using EMBOSS's "descseq" command, change the header of the promoter sequence to show the gene with which it's associated. You could change the $newHeader variable if you'd like. Changing a sequence header could be done using a variety of Perl or Unix commands too, so you can do it another way if you'd prefer. The "descseq" syntax is `descseq -seq seqFile -out newSeqFile -name newHeader` If `newHeader` is more than one word, you may need to enclose your variable in double quotes, like `"newHeader"`
13	Under line "# 3": Add this promoter to `$promoterFile`, a big file of all promoters.
14	Under line "# 4": After gathering the set of cDNAs and promoters for all growth factors, select an analysis that you think would be interesting for one of these multiple sequence files. Going to EMBOSS's page listing applications by groups `http://emboss.sourceforge.net/apps/groups.html` should lead to an interesting application in the EMBOSS package. Execute the command and look at the output file.
15	Under line "# 5": Complete the script by adding command(s) to delete any temporary files you've generated during the execution of the script.