Programming for Bioinformatics
BIOL 7200
October 24th, 2016

This week we will begin the noble study of Perl. Perl concepts and commands for this week:

- scalars
- arrays
- `shift`, `unshift`, `pop`, `push`
- `split`, `join`
- hashes
- `strict`
- `if`, `elsif` and `else`
- `while`, `for` and `foreach`
- streams
- `open`, `close`
- `chomp`
- `@ARGV`

Instructions for submission:

- You should submit **four** perl scripts named `week8_1.pl`, `week8_2.pl`, `week8_3.pl` and `week8_4.pl` for exercises 1, 2, 3 and 4 above.
- All scripts should write the output to STDOUT.
- You should NOT use perl modules other than 'use strict' and 'use warnings'

Exercises

1) Write a script that will get numbers from the user using STDIN and add them to an array. If the numbers are positive, add them to the back of the array. If the numbers are negative, add them to the front of the array. Stop if the user enters 0. Print the array in the end, with the values separated by dots. Also print the sum of the numbers entered.

2) Download the RepeatMasker table for the hg19 version of the human genome using the Table Browser on the UCSC Genome Browser. The RepeatMasker table is under the 'Repeats' group. There are three columns in the RepeatMasker table which classify the repeat: repClass, repFamily and repName. Using Perl, count the occurrences of every repName, repFamily and repClass in the output file. Print the results in some sort of pretty table, *i.e.*, format the results in some visually pleasing way.

3) Write a script to summarize an input BED file (download the UCSC gene table from the Table Browser in BED format as you did for Week 5; it should include at least the strand information along with the chr, start and stop), the name of which you will take from the command line. The summary should include:

1. The total number of entries in the file
2. The total length of the entries in the file
3. The number of entries on each strand
4. The longest entry
5. The shortest entry
6. The average and standard deviation of the gene lengths

Put this script somewhere on your PATH!

4) A real research example with hashes

Navigate to ftp://hgdownload.cse.ucsc.edu/goldenPath/, go into hg19 -> database. This is the place where you'll see a bunch of flat files that make up the databases for the UCSC Genome and Table Browsers. We are going to use two of the flat files from here (knownGene and kgXref from Week 3) to find the coordinates of some genes of interest.

With this assignment on T-Square, you'll find a file named InfectiousDisease-GeneSets.txt

Your objective is to write a script that reads in these three files in a specific logical order and returns the coordinates for the genes in the InfectiousDisease-GeneSets.txt file.

The filenames will be taken as command line arguments in the order: knownGene kgXref InfectiousDiesease-GeneSets

NOTE: It can happen that certain genes are absent in the kgXref table, which is ok. This inconsistency is due to discordance in the update dates of the table and GeneSets file, but there shouldn't be a lot of these cases.

Submission Instructions

- You should submit **four** perl scripts named `week8_1.pl`, `week8_2.pl`, `week8_3.pl` and `week8_4.pl` for exercises 1, 2, 3 and 4 above.
- All scripts should write the output to STDOUT.
- You should NOT use perl modules other than 'use strict' and 'use warnings'