Programming for Biologists (MMG1002H)

Assignment 5
10 Points

Assignment goals:

Practice interfacing external programs and processing their output

Part I: Practice running an external program (7 marks)

Primer3 (https://sourceforge.net/projects/primer3/) is a widely used program for designing PCR primers. It is possible to design primers one set at a time over the web using e.g.
http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi
For this lab, you will write a script to automate primer design for many sequences. Imagine you want to do design primers for all genes in a chromosome.
To do this, you will need the primer3 program to run as a command line tool on your local machine. Normally, primer3 must be compiled from source code in the C programming language, but we have supplied the executables for you (see 'primer3_core.exe' for Windows and 'primer3_core' for Mac OS X). If these don't work for you, you will need to follow the primer3 instructions to compile them, or see the TA for help. It doesn't matter which version of primer3 you use for this assignment.

Write a script that will design primers to 'detect a sequence' for all genes in yeastChromosome16Genes-DNA.fasta. This file contains the DNA of all of the genes on yeast chromosome 16 and was downloaded from http://yeastmine.yeastgenome.org.

There is some detailed documentation inside the primer3-1.1.4 folder, but you really only have to look at the "example" file, where an example input file for primer3 is located. Most important are the "PRIMER_SEQUENCE_ID" and "SEQUENCE" tags.

You will need to create the required input files for primer3, call the program (named primer3_core) with that input file (you may need to specify the full path to the program), read the results and process it.

Print out the best primer pairs for all the sequences in the input file (yeastChromosome16Genes-DNA.fasta). Note, for the purposes of this assignment, the best primer pair is the first one output by primer3. Note also that primer3 tries to find the 5 best primer pairs by default.

Print out the melting temperature (TM), GC% and length of all left primers in a file (all_primers_yChr16.txt) as a table of the format:

```
         TM   GC% Length
Gene1    .    .    .
Gene2    .    .    .
Etc…
```

Part II: Practice basic data I/O and plotting in R (3 marks)

Read in the melting temperatures, GC% and length of all primers you computed in part I.

Calculate the mean TM, length and GC%. Plot the TM versus the GC%. What is the correlation coefficient between TM and GC content? Is it significant?
Make a histogram for the TMs
Make a boxplot of TM's for primers with a GC < 50% and GC >50% (two categories). Is the difference between the two statistically significant? (Hint: you can use a single R function to compute this)
Print out the three plots with labeled axes as PDF files.

Submitting your assignment
When you're finished, gather the scripts that you modified or wrote and any requested output. Place them in a folder, and create a single archive file called assignment5.tar.gz

```
tar –czf assignment5.tar.gz yourFolder
```

You can also submit a zip archive (on a mac, right click and "compress").

When you're finished upload `assignment5.tar.gz` or `assignment5.zip` to the course website