BCB 444 Fall 2015 Project 5

Forming Data Pipeline and Formatting with Perl

Part II

Due Oct. 7 by 1 p.m.

1. Introduction

In the previous Project 4 description we have explained that Polymerase Chain Reaction (PCR) is an important molecular biology technique that allows the amplification of targeted DNA sequences. To use PCR to amplify a target gene, researchers must design a pair of *primers* based on the target gene sequence. Primers are short DNA oligonucleotides that can anneal (i.e., forming hydrogen bonds) to either ends of the amplification site on the target gene in order to uniquely control the amplification process. The previous project makes use of one of the best-known PCR primer design software — Primer3. If you have done Project 4 correctly, you should have obtained the correctly formatted "primer3_input.txt" file. When the file is given to Primer3 as input, "primer3_output.txt" will be produced. The output file is provided in this project, so you don't really need to run Primer3 to obtain it. However, it is fun to run "primer3_input.txt" through the "primer3_core" command to see if you can obtain the same output as in "primer3_output.txt". To run it, type the following command on your class account (i.e., gdcb-bcbteach.gdcb.iastate.edu):

```
primer3_core primer3_input.txt > my_primer3_output.txt
```

2. Extract designed primers from Primer3 output

In this project, we wish to extract the primers designed by Primer3 to a clean FASTA formatted file for downstream analysis or for primer synthesis ordering. The content of the extracted primer file should look like the following:

```
>egl-2_L0
gctaacacttgacgcggcgc
>egl-2_R0
actcgccggcttcaagctcg
>egl-2_L1
acacgccgtcaacgatgcca
>egl-2_R1
Gcgccgcgtcaagtgttagc
...
```

The primer header line is made up of the target gene name (e.g., egl-2), the direction of the primer in a pair (e.g., L or R) and the primer pair number (e.g., 0, 1 ... 9) as reported in the example Primer3 output file "primer3_output.txt". The actual primer sequence is then given below this header line just like in any FASTA file. Because it is often necessary to analyze the primers based on their direction, and sometimes the amplified gene fragment will also need to be sequenced to confirm their identity, your Primer3 output extraction Perl pipeline code should accept the following three options:

```
  -left: extract left primers only
  -right: extract right primers only
  -product: extract amplicon sequences only
```

Only one option above should be given to your program at a time, and when no option is given, your program should take the default action to output both primers in a pair like the example above. The provided "project_5_skeleton.pl" file contains the required code of getting command line options using the Perl Getopt::Long package. In addition, the "select.pl" example code also contains an example of using the Perl Getopt::Long package to retrieve command line options. If you do not know how to process command line options in order to combine the functionality prescribed above in one single program, you are allowed to create three different Perl programs, one each for extracting the left and right primers, and a third one for extracting the amplified PCR product sequences. All necessary data for generating the proper output in this project are contained in the Primer3 output file "primer3_output.txt", thus you do not and should not need to read in other input files.

3. Forming a data processing pipeline

As mentioned earlier, it is a common practice (and sometimes a necessity) to link up available bioinformatic software tools in a data processing pipeline to accomplish tasks. The data pipeline converts input data from one format to the next, and extract relevant output information from the last processing step. Although in this project you are only required to implement the front end and the back end of a pipeline that links up with just one bioinformatic tool, the Primer3 program, you can see the value of Perl in these types of bioinformatic applications already. After you implement both ends of the data pipeline to using Primer3 in the two projects, you can accomplish an automatic, closed form solution for PCR primer design similar to the following:

```
cat target_genes.seq | call_primer3.pl - | primer3_core | extract_primers.pl -
> primers.seq
```

The symbol '-' is required in the two Perl commands above because these two programs both expect command line options and the '-' symbol tell them to read from the standard input instead. Compare the individual commands linked by the Linux *pipe* operator `|` above to the figure on page 2 of Project 4 description, and you can see that they exactly match the boxes of the pipeline as depicted earlier. This combined command pipeline can prepare target gene sequences for Primer3 design, call Primer3, extract its output and finally save the designed primers in the output file all at once. A custom data processing pipeline like this is frequently needed in bioinformatics to provide powerful solutions that are otherwise unavailable or inefficiency to come by.

4. Project submission and grading

A skeletal implementation “project_5_skeletion.pl” has been provided to help you get started with this project. As said earlier, if you cannot understand how to obtain command line options and act separately on each option given in the same program, say, “project_5.pl”, you are allowed to create four separate programs “project_5_both_primers.pl”, “project_5_left_primers.pl”, “project_5_right_primers.pl” and “project_5_amplicons.pl” to submit for this project. Otherwise, roll your functionality into the one “project_5.pl” program and submit that for this project.

To submit your project, archive all your Perl programs(s) to a ZIP file or any compressed formats, then upload the archive to Blackboard Learn project submission area. To facilitate grading, please add Perl comments in your code to identify yourself (e.g., your full name and University ID) and to explain your implementation. It is not necessary but it may be helpful to include a README.txt file in your submission to provide more information or any special consideration about your Project implementation to the TA.

The TA will grade your project after the deadline, so you may submit multiple times before the deadline. If your project works on the sample data provided, your score should be no less then 15 (out of 20 total for the project). However, We will certainly test your code on other input data not made available to you to see if your code is general enough and is not making any assumption on some fixed characteristics of the input. Keep in mind that it is THIS project description that officially defines how your project code should behave and will be graded, not the sample input and output files provided. Therefore, after successfully testing your project on the provided sample data files, you may wish to create your own additional test files by changing the “primer3_output.txt” input data file to test for different input data. Better yet, now you have the whole pipeline working, try to use your Projects 4 and 5 programs to design new primers for any FASTA gene sequences you can download from the Internet. If you have any question about this project, post your project related questions to the Project discussion forum on Blackboard Learn.