CSci 3003/5465: Introduction to Computing in Biology

Lab Assignment #2

15 points
Assigned: 09/12/19
Due: 09/19/19 (Thursday), before 11:55pm

Goals of this lab:

- 5pts - Become familiar with using public databases for browsing gene information and obtaining sequence data.
- 5pts - Get practice using the Linux environment for writing and running Python scripts.
- 5pts - Begin understanding Python syntax, debugging programs using the python IDE Spyder

Part I: Finding a human disease gene for sequence analysis

(1) Go to the Online Inheritance in Man (OMIM) database: http://omim.org. Read about the purpose and the history of the OMIM database by looking in the “About” and “FAQ” (inside ‘Help’) pages.

(2) You will use OMIM to pick a gene of your choice to serve as the focus of various bioinformatics analyses throughout the next few lab assignments. There are several different ways to browse the OMIM database, but you can start by simply searching for a specific disease (e.g. lung cancer).

To find a gene specific to the disease (and its sequence), click the “Advanced Search: OMIM” link (inside Options) and select the “* gene with known sequence” option under “MIM Number Prefix”. Also, under the “Only Records with:” box, select “Allelic Variants” so that we restrict our search to genes with known variations in the human population.

Once you have found a gene, read about its function and its association with the disease you chose. Answer the following questions and turn in with your lab assignment. (To get functional and orthology information about your gene, you might also find GeneCards useful: http://www.genecards.org/).

a. Name the gene you chose and briefly describe its function and its association with the disease.

b. How many allelic variants are associated with your gene? List a few examples, and describe the mutations associated with each variant (e.g. substitution vs. insertion/deletion).

c. On which chromosome is your gene located in the human genome?

d. Does your gene have an ortholog in any other species? If so, name 2-3 other species and report the sequence similarity to the human gene.

(3) Download the sequence for your gene (or one of its variants) in FASTA format (https://en.wikipedia.org/wiki/FASTA_format). Create a folder named "*Lab2*" in your CSCI3003 directory (if you have not already) and save a text file in that folder with this sequence (Copy and Paste into a text editor will work fine).

HINT: Click "DNA->NCBI RefSeq" on the right menu and then "FASTA" for one of the variants to get a sequence.

Part II: Editing and running Python scripts to process sequence data

Your goal is to edit and run a Python script that we have mostly written for you. Once you modify it correctly, this script will read in your sequence from your file and perform some simple processing. For the rest of this course, we will be using a tool known as an integrated development environment (IDE) called Spyder. It is a combination of a text editor and python interpreter that lets you easily develop and debug scripts all in a single program.

1. Download the *process_sequence.py* script from the "Lab 2 Materials" folder on the course Moodle website

2. Open Spyder by entering the following commands into the shell (or open it using Graphical User Interface on your personal machine):

   ```
   spyder &
   ```

3. Open the script called *process_sequence.py* using Spyder. Study the code, and guess what it will do to your gene sequence. Modify it so that it will load your sequence file (**HINT**: notice the file_name variable). Run it, and include the output in your lab write-up.

   TIP: In Spyder, you can run the script loaded in the editor by clicking the green "Play" arrow in the toolbar or by hitting F5. You can also run individual lines or selections of code using F9.

Part III: Debugging Python scripts

The final part of this lab assignment is to debug and fix a Python script that we've written (badly) for you. One of the best ways to fix bugs is to follow an iterative process: start at the top and "Edit, Run, and Revise". Follow the error tracebacks provided by the Python interpreter to understand what the bugs are.

1. Download the *buggy_script.py* script from the course Moodle website

2. Fix any bugs in the script until the Python interpreter runs it without any warnings.
   TIP: Clear the interpreter window by hitting CTRL-L (or type `clear all`).
   When it's fixed, you should get the following output:

```
This seems to be 'ok'

I am trying to write a multiline statement but I am too lazy
to write multiple lines separately using multiple print statements, and don't
feel like using new lines '\n'.
Something isn't quite right, though!
Can you fix me? .....
"Do. Or do not. There is no try."

6
AAAATTTT
AAAATTTTTTTT
pythonMaster@umn.edu
32
Adenine: A
Cytosine: C
Guanine: G
Thymine: T
Start Codon: AUG
Stop Codon: UAG
Length of Coding Region: 36
```

When you're finished with the lab, make a report of any questions you answered plus any requested output, and gather the scripts that you modified.

Submit to Canvas

Submit your homework file using Canvas, include the following documents in your archive:

1. Fasta File for your gene of interest
2. Answered questions about your gene of interest in .txt format
3. Output of *process_sequence.py* run on your gene of interest. (can be included in above questions .txt)
4. Copy of *process_sequence.py* with any additional comments you added
5. Corrected *buggy_script.py* with any additional comments you added

Congrats!! You're done with Lab 2!