COSC 1437, Professor Clark

Assignment 9 Genome Analysis: 24 points

Write a Program: COVID-19 Genome Analysis

Your goal is to do analysis on the COVID-19 virus (*technical name: severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)*).

In order to understand the terminology within these requirements, read the GenomeDefinitions.pdf file located in the Canvas Assignment where you found this document.

Program Requirements:

You will write a program that will allow the user to choose a reading frame for the genome. Based upon that choice, you will:

1. Output a report file with the gene analysis for that reading frame. The report will contain information on each gene located in that frame. That information needs to include:
    a. The number of amino acids in the gene
    b. The start and stop positions of the gene (by nucleotide number) This would look something like 1344..1883
    c. The amino acid string for the gene
    d. For this assignment, use an ORF of 50. Only include genes in the report that have at least 50 amino acids.
2. Provide the user with the ability to do codon bias analysis for that reading frame. The user will have two options:
    a. Generate a complete report file of codon bias for every amino acid in the genome (ensure that you do this for each reading frame so that you may turn in the report with your submission). If the user chooses this option, only write the data to a file. Do not display that output to the screen. Simply tell the user the name of the file you generated.
    b. Do a codon bias analysis for a single amino acid. The user will choose which amino acid to analyze (by entering the single letter representation of the amino acid), and the program will display to the screen the results of that analysis. The output will contain:
        i. The full name of the amino acid along with its one letter code
        ii. All codons that can code for that amino acid
        iii. The percentage for which each codon is found for that amino acid within the genome

The user should be allowed to continue analyzing single amino acids if they wish. When they are finished, write out the report from (a) to the file. They should not leave the codon bias analysis portion of the program without a report generated. Note, the single analysis screen output will not be written to a file.

Design Requirements and Suggestions:

You are provided with the following files to begin (located in GenomeLabNeededFiles.zip):

- aminoAcidTable.csv – a file of all amino acids with their names, abbreviations, and all codons that can create that amino acid
- covidSequenceRF1.csv – all codons of the COVID-19 genome for reading frame 1
- covidSequenceRF2.csv – all codons of the COVID-19 genome for reading frame 2
- covidSequenceRF3.csv – all codons of the COVID-19 genome for reading frame 3

You're going to need a minimum of 3 classes but may choose to create as many as you wish. You'll need one for Amino Acids, one for Genes, and a Main class. If you go this design route, your Main class will be huge. This is fine, as long as there are many well documented methods in it. Or you may find the design to be cleaner with more classes. It's completely up to you.

Design this before you start trying to code it! With multiple team members, communicating who is completing which parts is crucial, and determining how those parts will interact with each other ahead of time will make the merging of your code pieces much smoother. I've provided a (hopefully) helpful file called ProjectDesign.txt to help with your team project management for the components.

Minimally, you should consider holding the following information for your data type classes design:

- Amino Acid
    - Full name
    - Single letter abbreviation
    - All codons capable of creating this amino acid
- Gene
    - The string of amino acids that make up the gene
    - The start nucleotide position
    - The end nucleotide position
    - Gene length (number of amino acids)

You'll be handling a lot and will likely have many arrays/arraylists/variables/methods/etc. being managed in your Main class (or additional classes you choose to create). Be sure to document everything well (Javadoc comments, comments within code as needed).

Let me be clear, this project will not be easy. It will take your whole group working together to figure out how you want to tackle this.

Submission Requirements:

Upload a .zip file for your group on Canvas containing the following items:

- A folder containing all of your code for the project
- Codon bias reports for reading frames 1, 2, and 3
- Gene analysis reports for reading frames 1, 2, and 3
  Note that this will be 6 files plus the code folder

Example Output:

I will not provide example output for the COVID-19 genome you are analyzing. However, I have provided a .zip file on Canvas (in our Chapter 9 module) with the requested generated reports on a Measles genome. You may look at those reports to see the format etc. of what you need to provide.

A note: you may wish to write your project first against the Measles genome (I have provided files for all reading frames in that .zip as well). If you can get that to work and get my output, then it will only require very minor updates to your code to get it to work for COVID-19.