Systems Programming - CSI 223 Assignment 1

Issue Date: 19 February, 2020 **Due Date:** 23 March, 2020

Total Marks: 47

Instructions: Work in team of 3 members max. Check Moodle for submission link and for any further instructions.

Background

DNA can be thought of as a sequence of nucleotides. Each nucleotide is Adenine, Cytosine, Guanine, or Thymine. These are abbreviated as A, C, G, and T. A nucleotide is also called a nucleotide base, nitrogenous base, nucleobase, or just a base.

DNA data can be stored in a .fasta file format as shown below: One fasta file can contain multiple sequences.

The .fasta format

```
>JX297515.1 Bovine viral diarrhea virus type 1b isolate Corona polyprotein gene, complete cds
ATGGAGTTGATCACAAATGAACTTTTATATAAAACATACAAACAGAAACCCACTGGAGTGGAGGAACCAG
TATATGACCAAGCTGGCAACCCTTTGTTTGGAGAAAGAGGAGAGATTCATCCGCAATCAACGCTAAAACT
GCCACATAAAAGAGGGGAGCGCGAAGTCCCCACCAATTTGGCTTCTTTACCAAGAAAAGGTGACTGCAGG
TCGGGTAATAGCAAGGGGCCTGTGAGTGGAATCTATTTAAAACCAGGGCCGTTATTCTACCAGGATTACA
AAGGACCCGTCTATCATAGGGCCCCATTGGAGTTTTTTGAGGAGGTGCCTATGTGTGAGATAACTAAAAG
AATTGGGAGAGTAACTGGTAGTGACAGCAAATTATACCACATTTATGTGTGTATTGATGGATGCATAATA
GTTAAAAGCGCTACAAAAGATCGTCAAAAAGTACTCAAATGGGTCCACAACAAGCTAAACTGCCCCATAT
GGATATCAAGCTGCTCCGACACAAAGGATGAAGGGGCGGTAAGAAAGAGGCAACAAAAGCCAGATAGGTT
TGAAAAGGGGAGAATGAAGATAACACCTAAGGAGTCAGAGAAAGACAGTAAGACCAAGCCACCAGATGCC
ACGATAGTGGTAGATGGAGTCAAATATCAGGTAAAGAAAAAAGGAAAAGTCAAGAGTAAGAATACCCAGG
ATGGCTTATACCACAACAAAAATAAACCTCAAGAGTCACGCAAGAAACTAGAGAAAGCCCTACTGGCCTG
GGCAATAATAGCCTTGGTTTTGATCCAGGTCACGGAGGGAGAGAATATAACACAGTGGAACTTACAAGAT
AATGGAACTGAAGGAATACAACGAGCCATGTTTCAAAGAGGTGTGAATAGAAGTCTACATGGGATATGGC
CAGAAAAAATCTGTACAGGTGTTCCTTCCCATCTGGCCACTGACACAGAACTGAAGGCAATTCACGGTAT
GATGGATGCAAGTGAGAAGACAAATTATACGTGCTGCAGACTCCAACGCCATGAGTGGAACAAACATGGT
TGGTGCAATTGGTATAATATTGAACCTTGGGTCCTTCTTATGAATAAAACTCAAGCCAACCTTACTGAGG
GTCAGCCGCTAAAGGAGTGTGCCGTTACATGCCGGTACGATCGAGATAGTGACCTGAATGTAGTAACACA
AGCTAGGGATAGCCCCACACCATTGACAGGCTGCAAGAAAGGCAAGAACTTTTCCTTTGCAGGCATATTG
GTACAAGGGCCTTGCAACTTTGAAATAGCCGCAAGTGATGTGCTGTTCAAAGAGCATGATTGCACTGGTG
TGTTTCAAGACACAGCTCACTACCTCGTAGATGGGATGACCAACACCTTAGAGAGTGCTAGACAAGGGAC
CGCAAAACTAACGACTTGGCTGGGCAGGCAGCTTGGGATACTAGGAAAGAAACTGGAAAACAAGAGTAAG
```

Header line with sequence name, starting with ">"

Sequence, typically 60 characters per row

For this assignment, you will be interacting with the DNA data set for the Corona virus, i.e. **Corona.fasta** file.

Within the .fasta file, a **string** is simply an ordered collection of symbols selected from some **alphabet** and formed into a **word**; the **length** of a string is the number of symbols that it contains.

An example of a length **21 DNA string** (whose alphabet contains the symbols 'A', 'C', 'G', and 'T') is "**ATGCTTCAGAAAGGTCTTACG**."

Question 1 – Establishing size of the DNA.

i. Which Unix/Linux command(s) would you execute to extract the sequence in the 13th line of the Corona.fasta file? **[2 mark]**

ii. Which Unix/Linux commands would you execute to count the number of sequences (number of lines) in the Corona.fasta file? **[5 marks]**

Note: since every line containing a sequence name starts with '>' character

iii. Using Bash, count nucleotides (characters) only on the Corona.fasta file. **[5 marks]**

Note: ignore lines with sequence names and counts all the characters

iv. Now, count only **T's** in the Corona.fata file. **[5 marks]**

v. Write a Bash Script that will read an input sequence e.g. **`ATGGAGTTGAT`** and output four (4) integers, (separated by spaces) counting the respective number of times that the symbols **'A', 'C', 'G',** and **'T'** occur.

`Sample Output: 20 12 17 21` **[10 marks]**

Question 2 - Transcribe DNA to RNA

An **RNA** string is a string formed from the alphabet containing **'A', 'C', 'G',** and **'U'**.

Given a DNA string **tt** corresponding to a coding strand, its transcribed RNA string **uu** is formed by replacing all occurrences of '**T**' in **tt** with **'U'** in **uu**.

Given a DNA string **`GATGGAACTTGACTACGTAAATT`**

Write a Bash Script that will return the transcribed RNA string.

`Output: GAUGGAACUUGACUACGUAAAUU` **[10 marks]**

Question 3 - Complementing a Strand of DNA

In DNA strings, symbols **'A' and 'T'** are complements of each other, as are **'C' and 'G'**.

The reverse complement of a DNA string **S** is the string **S^C** formed by reversing the symbols of **S**, then taking the complement of each symbol (e.g., the reverse complement of "**GTCA**" is "**TGAC**").

Given a DNA string **`AAAACCCGGT`**

Execute a Bash Script that accept input **S** = **`AAAACCCGGT`** and return the reverse complement **S^C** of **S**.

`Sample Output: ACCGGGTTTT` **[10 marks]**