CSCI 451/551 Fall 2023 Homework 4
Due: October 31, at the start of class

(Group submissions: please make sure all group members are listed)

Problems 1 (30 pts) *Programming project:* Write a program to compute the FM-index for a given text string S and search for a given query string Q.

The text string should be read in from a FASTA file (like HW3). You can assume that the file just contains a single DNA sequence, e.g.

```
> seq1
ACAATGGAT
```

Remember to add a `'$'` to the end of the string. The query string should be provided as a command line parameter.

Output:

Part 1: First provide the computed FM-index for the string. You do not need to compress the occ table.

```
BW = TC$GAAGTAA

C[A] = 1
C[C] = 5
..

OCC[$,1] = 0
..
OCC[T,10] = 2
```

(it is fine to 0-based indexing if you want)

Part 2: Implement the FM-index based range search algorithm and output the range of Q found (recall this is the range of Q in the suffix array; if you keep the suffix array around you can check that it worked), e.g.

```
range[S, 'AT'] = [4,5]
```

Demonstrate your program on a couple sample input strings and queries strings.

Bonus question (1 pt): Implement "suffix array sampling". Just store the suffix array locations for every kth (e.g. k=10) suffix in S. Then use the procedure discussed in class in order to find entries that are not stored. Demonstrate this works with a longer string.