CSE4094 – Assignment I

Due To: 13/03/2018 23:55

In this project, you will work on a particular region in a genome, so you will only take some part of one genome as input. This input can be defined as one line of text in a file where the text contains only A, T, G or C and nothing else (even whitespace). The total number of characters may be very high, but in this assignment, we will assume that we only have 500 characters/bases.

Your objective is to find all possible **k-mers** appering at least **x** times. Assume the value of k will be at most 9 and x will be at least 2.

After you find all possible k-mers, search for the reverse complement of each k-mer and if you find any, give it as output.

Inputs: integer **k**, **x**. An input file.

Output: All possible k-mers in the file appearing at least x times. Reverse complement of eack k-mer if found any.

Example:

Assume there are 32 characters in the file, so the following line shows the contents of a file.

ACAAATTTGCATAATTTCGGGAAATTTCCTTT

Inputs: k=3, x=4

Outputs:

3-mer: TTT

Reverse complement: AAA appearing 2 times

Inputs: k=4, x=3

Outputs:

4-mer: AATT, ATTT

Reverse complement: AATT appearing 3 times, AAAT appearing 2 times.

You can use any programming language. You can work in groups of 2 or 3.