REGULATIONS

- **Due date:** 15 November 2021, Monday, 23:59 (Not subject to postpone)
- Electronically. You will be submitting your program source code through a text file which you will name as the1.py and the report nameSurname.pdf by means of the AYBUZEM system.
- **Team:** There is **no** teaming up. The homework has to be done and turned in individually.
- **Cheating:** Source(s) and Receiver(s) will receive zero and be subject to disciplinary action.

INTRODUCTION

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) causes a disease called COVID-19. It is very contagious and has rapidly spread globally. Due to various symptomatic and asymptomatic cases and the possibility of asymptomatic transmission, there is a pressing need for a fast and sensitive detection protocol to diagnose asymptomatic people. Various SARS-CoV-2 diagnostic kits are already available from many companies and national health agencies. However, publicly available information on these diagnostic kits is lacking. Accurate designing of polymerase chain reaction (PCR) depends on a number factors related to the probe design quality. You may basically consider that the probe is a fragment of the genome. In this homework, you are expected to describe a set of probes targeting conserved regions identified from SARS-CoV-2 genomes from the Global Initiative on Sharing All Influenza Data (GISAID). In this way, you will help the production of the PCR test kit. Remember that the most important process of designing PCR test kit is probe design, and in this homework, you will identify probes.

BACKGROUND

SARS-CoV-2 genome sequences:

- They include four nucleotides: adenine (A), guanine (G), cytosine (C), and thymine (T).
- They are called as base pairs (bps).

In this homework, you will use SARS-CoV-2 genome sequences whose lengths are exactly 27000 bps as the input. That is, you may basically consider a SARS-CoV-2 sequence as a string whose length is 27000.

PROBLEM

- You are expected to write the following *probes* function which takes two input parameters. The first parameter is the file name and the second argument is length of the probe. This function reads the genome sequences from the input file and returns a list of probes that is the maximum repeated substring with the given length.

def probes(fileName, probeLen)

SPECIFICATION

Input Specification:

- The input includes SARS-CoV-2 sequences which are sequences of four nucleotides, A, C, T, and G. If there is a sequence whose nucleotides are different from A, C, T, and G in the first 27000 bps, please ignore this sequence.
- SARS-CoV-2 sequences must belong to different people. This means that all sequences should be different from the others and you must remove the same sequences from the input file.
- Each line in the input file represents a SARS-CoV-2 genome sequence belonging to one person.
- Length of the sequences must be 27000. There may be some sequences grater than 27000 in length and you take the first 27000 bps. If there is a sequence whose length is less than 27000 bps in the input file, please ignore it.

Output Specification:

- Your *probes* function must return a list of probes which are the maximum occurring subsequences with the given length.
- Results must be written in an output file,
  "nameSurname.txt"
- Content of the output file is explain in the following:
  <The number of valid sequences> <The maximum number sequences that the probes are found>
  <length of the returning list of the probes function>
  <the first element of the returning list of the probes>
  <the second element of the returning list of the probes>
  ....
  .....

- A valid sequence is the sequence whose length is greater or equal than 27000 and only includes A, C, T, and G nucleotides in the first 27000 bps.
- The output file does NOT include additional information strings.

EXAMPLE

- In this example, the length of the genome sequences is 10 bps.
- There are 6 genome sequences.
- Probe length is 5.
- Suppose you are given an input file "example.txt"with the following content:
  TTCGATCTCT
  ATGCTTAGTG
  ACTTTCGATC
  ACCAACCAAC
  TAAAGGTTTA
  TTTCGANATT

- OutputFile, for instance, "nameSurname.txt"
  5 2
  2
  TTCGA
  TCGAT

Submission

1. the1.py
2. nameSurname.pdf (report)
   - The report contains pseudocode of the algorithm and
   - a table summarizing the results of your algorithm and the following table is given as an example.

The number of valid sequences	The number of sequences probe is found	Probe length	Time in seconds
5	2	2	1.2
10	?	90	?
50	?	90	?
100	?	90	?
250	?	90	?
500	?	90	?
1000	?	90	?
10	?	100	?
50	?	100	?
100	?	100	?
250	?	100	?
500	?	100	?
1000	?	100	?