CS 482/682, 2022, Assignment 1.

Due date: Feb. 6, 11:59pm.

Assignment 1. Fit Alignment

Fit alignment: Given two sequences S and T, the fit alignment asks to find a substring T' of T, such that the global sequence alignment score between S and T' is maximized.

Part 1. Write a program to "fit align" two sequences. The program reads from a FASTA file that contains exactly two DNA sequences; and prints the alignment score and the alignment to an output file. We will call your program by the following command line:

assn1-part1 in.fasta out.txt

Note that the file names may be different when we call.

The output file should consist of three lines:

- The first line is the alignment score;
- The second and third lines correspond to the first and second sequences in the input, with dash inserted to form the alignment.

If there are multiple alignments that can give the same optimal score, your program only needs to output any one of them.

We use the following score scheme and linear gap penalty: match = 1, mismatch = -1, indel=-1.

For example, the following input:

```
>seq1
AACCCCTAG
>seq2
TTAATCCCCAGGGTCGTTT
```

Should produce the following output (or possibly another solution with equal score to the following):

```
6
AA-CCCCTAG
AATCCCC-AG
```

Part 2. Consider the scenario where you already have a library (in a FASTA file) of known SARS-Cov2 virus genomes. Now you obtained a new partial sequence from the virus detected from a patient. You need to find which variant it is likely from. You can use the code you developed in part 1 as a subroutine for this purpose.

In this part of the assignment, your program is given two FASTA files. The first file contains only one sequence. The second file contains many sequences. It then fit-align the first sequence to each of the

sequences in the second file. Finally, it outputs the best alignment it can achieve to a new file. We will call your program with the following command line:

assn1-part2 query.fasta library.fasta out.txt

The output file format will look like the following. The first line is the same header line of the best sequence that maximizes the fit alignment score in the library fasta file. The second line is the fit alignment score. The third and fourth lines are the aligned sequences from the query sequence and the library sequence, respectively.

```
> variant xxxx
6
AA-CCCCTAG
AATCCCC-AG
```

Your program needs to be reasonably efficient (quadratic time complexity and at most quadratic space complexity) to get full marks.

What to submit:

Your program and a readme.pdf file. Your program can contain multiple source files. For this program, only standard libraries for the programming language are allowed (such as java.util or similar libraries for other programming languages). The readme.pdf file contains your information (names, student number, etc.) and any additional information you would like the TA to know about your program. Keep the readme.pdf file short. Further details about the submission procedure will be announced later.

Programming language:

Java or Python.