2.4 Problem 4

Step 1

Locating Insertions

In this problem you are given 6 different tests for the proposed problem in 6 different tabs. You should submit only answers in the plain text format. Depending on the quality of the answer you will be given some fraction of the points on the corresponding test (the exact formula is described in the problem statement). Maximum points for each test are described in the corresponding tab.

The tests for this problem were generated based on some manipulations on real-life genomes as well as by artificial simulation. For each test we give a short background of what to expect in that test.

The points are distributed as follows: for tests 1-2 we give 200 points, for tests 3-4 we give 250 points and for tests 5-6 we give 300 points.

You can download all tests via this link: https://stepik.org/media/attachments/lesson/40285/tests.zip

Step 2

Locating Insertions

One day several of John Doe's friends asked him to analyze genomes of some organisms with strange properties they found. He sequenced their genomes and tries to understand what is happening there. His gut feeling says that something got inserted into each of these genomes and that is what causes these strange properties. He wants to know where these insertion are located inside the genomes. Please, help him.

Input format

You are given a DNA sequence of a genome in FASTA file format.

Output format

The output should contain two integers separated by space — the start position of the inserted sequence and the end position of the inserted sequence.

Scoring

Suppose your answer contains two integers L and R, while the correct answer is $\overline{L}$ and $\overline{R}$. Then the percentage of points you get for each test is $100\% \times \max(0, 1 - \frac{20}{|S|}(\max(0, |L - \overline{L}| - 10) + \max(0, |R - \overline{R}| - 10))$, where S is the length of the input sequence. Maximum number of points is different for different tests: you can get up to 200 points for tests 1 and 2, up to 250 points for tests 3 and 4, and up to 300 points for tests 5 and 6.

Step 3

This test (01.fasta) contains a real plasmid with a random nucleotide sequence inserted in a random place.

Submit two integers separated by a space specifying the left and right borders of the inserted sequence.

To solve this problem please visit https://stepik.org/lesson/40285/step/3

Step 4

This test (02.fasta) contains a real bacterial genome with a lot of base-changes/indel errors introduced. Of course, we also inserted a random nucleotide sequence in a random place.

Submit two integers separated by a space specifying the left and right borders of the inserted sequence.

To solve this problem please visit https://stepik.org/lesson/40285/step/4

Step 5

This test (03.fasta) contains a short fragment of an unknown organism genome where a sequence from a drastically different unknown organism genome was inserted.

Submit two integers separated by a space specifying the left and right borders of the inserted sequence.

To solve this problem please visit https://stepik.org/lesson/40285/step/5

Step 6

This test (04.fasta) contains a genome of an unknown organism where a sequence from a drastically different unknown organism was inserted.

Submit two integers separated by a space specifying the left and right borders of the inserted sequence.

To solve this problem please visit https://stepik.org/lesson/40285/step/6

Step 7

This test (05.fasta) contains a genome of an unknown bacteria found in a patient's gut. Find out what might be the cause of his disease.

Submit two integers separated by a space specifying the left and right borders of the inserted sequence.

To solve this problem please visit https://stepik.org/lesson/40285/step/7

Step 8

This test (06.fasta) contains a genome of an unknown bacteria. It is conjectured that this genome contains some kind of a transposable element.

Submit two integers separated by a space specifying the left and right borders of the inserted sequence.

To solve this problem please visit https://stepik.org/lesson/40285/step/8