

Project: Alignment problem

This project requires you to implement the global alignment and the local alignment methods.

The programs should be flexible, i.e., it should be possible to:

- Align sequences over any alphabet. The alphabet can be {A, C, G, T} for DNA or 20 letters alphabet for amino acids, or any alphabet. The alphabet is specified in the parameter file.
- Use any score matrix. The score matrix is specified in the parameter file.
- Handle **affline** gap penalty.
- Output the optimal alignment score and the optimal alignment.

The input sequences are in FASTA format. For example,

```
>seq1
Ctacgaaaggtcgtgtcacgatgtccgcaaggat
ggcattgcatagaggaattgattgcaac
```

```
>seq2
cttaatgtcccgcgtacaaggatagcatgtggca
tagaggaatagaatagcagc
```

The score matrix, the alphabet, and the gap penalty score are specified in the parameter file.

You need to output the optimal score and an optimal alignment. The computed optimal alignment should be output in FASTA alignment format (also called Pearson format after the creator of the FASTA alignment program). In FASTA alignment format, the two aligned sequences are printed above each other with gaps inserted as described by the computed alignment. For example,

```
>seq1
ctacgaaaggtcgtgtcacg-atgtcc-----gc
aagggatggcat---tgcatagaggaattgat-tg
caac
```

```
>seq2
ctt-----aatgtcccgcgta-c
aagggatagcatgtg-gcatagaggaatagaata-
gcagc
```

Detail of the assignment

You are required to write two programs:

1. A program for global alignment

```
java global_align parameter.txt input.txt output.txt
```

2. A program for local alignment

```
java local_align parameter.txt input.txt output.txt
```

Testing data

You are given three sets of testing data.

The first testing dataset is without affine gap penalty. They are contained in parameter1.txt and input1.txt. The sample output files are output1_global.txt, and output1_local.txt.

The second testing datasets are with affine gap penalty. The detail is specified in parameter2.txt, input2.txt, output2_global.txt, and output2_local.txt.

For the last testing dataset, I only give you two files parameter3.txt and input3.txt. You are required to generate output3_global.txt and output3_local.txt and send them to me.

Note that everything after ';' in the dataset is comment. You must make sure your programs can read the parameter files and the input files. Also, your programs must follow the output format as stated in the sample output files.

Submission

Please submit the two programs, output3_global.txt, and output3_local.txt.

In addition, please handin a sheet to summarize your system. The sheet should tell me the algorithms you have implemented, their time complexity, and their space complexity. Also, please also include the output for the third testing dataset.

Please make sure your programs are efficient enough. I will time your program.