# Program Assignment

due day: 1 March 2020

This assignment requires you to implement an algorithm that find the best alignment between the prefixes of two sequences S1 and S2. (In other words, we forgive the ending spaces for both S1 and S2.)

You need to implement three programs: the standard dynamic programming, the banded dynamic programming and the X-drop algorithm.

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.

For each of your program, you need to output the alignment score, the alignment and the number of entries you filled in.

The input sequences are in FASTA format. For example,

```
>seq1
tgacaatccc
>seq2
tgaggatggt
```

The score matrix, the alphabet, the gap penalty score, the bandwidth, the threshold X are specified in the parameter file.

You need to output the optimal score, the number of entries you filled in 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
tga-caat
>seq2
tgagca-t
```

## Detail of the programming task

You are required to write three programs:

1. **A program for DP**
   ```
   java align_DP parameter.txt input.txt output.txt
   ```
2. **A program for banded DP**
   ```
   java align_band parameter.txt input.txt output.txt
   ```
3. **A program for X-drop**
   ```
   java align_drop parameter.txt input.txt output.txt
   ```

## Testing data

You are given four sets of testing data.

The first testing dataset is without affline gap penalty. They are contained in parameter1.txt and input1.txt. The sample output files are output1_DP.txt, output1_band.txt, and output1_Xdrop.txt.

For the rest of the three testing datasets, only parameter and input files are provided.

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.

## Written question

You are required to submit a report.

In the report, you need to describe your algorithm for handling affine gap penalty.

Also, you need to perform analysis on the running time and the number of entries to fill in for your three programs.

For your analysis, you need to use some homology sequences. For example, you can get some homology sequences from http://eggnogdb.embl.de/#/app/downloads.

## Submission

Please email your three programs, a README on how to run your program and your report and analysis to ksung@comp.nus.edu.sg.