CAP5510 - Bioinformatics Fall 2017, Homework 1

Due date: 11 / 1 / 2017
Turn in hard copy of report in class.
Upload source code and executable using e-learning.

October 18, 2017

This is a programming homework. The purpose is to gain experience on sequence alignment. In this homework, you will implement

1. Global alignment
2. Local alignment
3. End space free (dove-tail) alignment

according to Needleman-Wunsch and Smith-Waterman methods. The methods are available in my course slides at `www.cise.ufl.edu/~tamer/teaching/fall2017/lectures` (Slides 2).

Your program will get the following as input:

1. an integer that denotes which alignment method will be used. Use 1, 2, and 3 for global, local, and dove-tail alignment respectively.
2. name of the file that contains query sequences in Fasta format
3. name of the file that contains database sequences in Fasta format
4. name of the file that contains the alphabet
5. name of the file that contains the scoring matrix for the given alphabet
6. a positive integer k that indicates the number of nearest neighbors

7. a negative integer m that indicates the gap penalty

Sample input files: Sample query, database, alphabet, and scoring matrix files are on the e-learning website.

Output: Your program will output the top k alignments over all query/database sequence pairs along with their score. Each alignment in the output should contain the score, ids of the sequences, and the starting positions of the alignment. It will look like the following:

```
Score = 47
id1 36 AQT..KNGQGWVPSNYITPV
id2 39 ARLNDKEGYVPRNLLGLYP.
```

Here id1 and id2 are ids, 36 and 39 are starting positions.

Return: You will return a soft copy of the following four items. If you are not an EDGE student, also return a hard copy of the first three items.

1. A graph which shows the score of the top 100 results for each alignment strategy using the given files. You can put the plots into a single graph where y-axis is the score and x-axis runs from 1 to k.
2. Split the given sample query file into multiple query files each containing one sequence. Align each query with all the sequences in the given sample database. Return a graph which shows the running time of each method for each query. Here, x-axis will be the query sequence length and y-axis will be the running time.
3. A brief discussion of the two plots.
4. Source code and executable. Executable file's name will be "*hw1*".

Final details:

- You can use C, C++, or Java.
- Make sure that the program runs on CISE linux machine (such as thunder or storm)
- Do NOT forget to use the name "hw1" as the executable of your code. For example, a command line to find top 10 results using global alignment should look like

  ```
  > hw1 1 queryfile datafile alphabet scorematrix 10 -3
  ```

  to run global alignment and return top 10 results with a gap penalty of -3.