CBB752b17 Homework Assignment 1

DUE DATE: February 20th (Monday) 2017, 11:59pm

Choose to do either MCDB&MBB or CBB&CS homework, depending on your academic affiliation. No late submissions will be accepted.

MCDB & MBB 752/753

1. Multiple sequence alignments (MSA) cannot be efficiently handled using purely dynamic programming. Choose one existing MSA software and describe how it implements MSA. (for example Muscle, clustalW, Kalign, MView, T-coffee. . .)
2. Align the following two sequences using the Smith-Waterman algorithm (local alignment), with the following scores: Match: 2; Mismatch: 0; Gap: -1. In addition to filling out the alignment matrix, indicate the traceback and write out the final alignment.

		A	A	A	A	C	G	C	T	T
	0	0	0	0	0	0	0	0	0	0
T	0									
T	0									
T	0									
A	0									
A	0									
T	0									
C	0									
G	0									
C	0									

3. In class and "The Game of Genomes" the difficulty of properly identifying structural variation in a genome was discussed. Explain how the Smith-Waterman and Needleman-Wuncsh algorithms would align a read containing a large deletion relative to the reference genome. How would changing the gap penalty impact the alignment?
4. Machine learning approaches are becoming extremely useful in the analysis of genome-scale data, as reviewed in the following paper: Yip, KY, Cheng, C, Gerstein, M (2013). Machine learning and genome annotation: a match meant to be?. Genome Biol., 14, 5:205. Choose one article that describes the application of supervised machine learning to genomics and answer the following:
   - What are the researchers trying to predict/infer?
   - What information in being used for the prediction? What is the logic behind using these data?
   - What preprocessing steps are used to prepare the data for machine learning?
   - What is the model the researchers use, and why did they select their particular method?
   - How do the researchers evaluate their predictions? Were they effective? What biological insight was gained?

CBB & CPSC 752

Scripting must be done from scratch, without the use of any preexisting packages.

The programming task is to implement the Smith-Waterman local alignment algorithm for protein sequences.

Gap penalties: opening gap -2, extension gap = -1

Requirements:

The program should automatically read in the similarity matrix file called “blosum62.txt” and input sequences in “input.txt”, where each line is a sequence.

The output should contain a human-readable alignment such as the following:

```
T  C  W  A

   |     |

S  C  -  A
```

where | represents amino acid identity and - represents a sequence gap.

For each sequence pair, the output must include the completed scoring matrix (including the sequences themselves) in tab-delimited format (akin to the hand-drawn DP scoring matrix), best-scoring local alignment(s) and the score. (Just to be precise, the completed scoring matrix contains the best score in the alignment up to this point.) These will constitute 90% of your grade, with the remaining 10% coming from your programming style (e.g. clear comments).

Programs that do not compile will get an immediate 0. To receive partial credit, please make sure your program is well-commented.

If you use Python, please edit the code template “hw1.py”.

If you use R, please edit the following code block:

```r
## Specifying author and email
p <- c(person("First name", "Last name", role = "aut", email = "your email"))

## Define the main function
MyOwn_Smith_Waterman <- function(input = "input.txt", output = "score.txt") {

}

## Run the main function and generate results
MyOwn_Smith_Waterman(input = "input.txt", output = "score.txt")
```