BBM411/AIN411: Fundamentals of (Introduction to) Bioinformatics
(Fall 2023)

Assignment 1

Due date: November 20, 2023, time: 23:59 **(**10 points reduction for each day late**)**

Please submit your assignment as a single PDF file + your script file over e-mail (include your name both inside the document and also in the PDF and script filenames) in the given time frame (tuncadogan@gmail.com). Please enter "BBM411 – Fall 2023 – Assignment 1" to the email subject.

Please note that although sharing ideas and discussions is encouraged, solutions/results, codes and text should only belong to you. In the case of copy/cheat, serious point deductions will be applied.

Question 1 (10 points)

Please carefully explain each question below (in a total of 2-3 sentences for each item)

a) What is central dogma? What does it say about the information flow at the molecular level in living organisms?

b) What is the significance of homology in bioinformatics??

c) Briefly describe the purpose of a multiple sequence alignment in bioinformatics.

d) Explain how BLAST manages to run faster than the optimal sequence alignment. algorithm Does BLAST perform the same as the optimal alignment regarding accuracy?

Question 2 (55 points)

Implement the pairwise sequence alignment of amino acid (protein) sequences via dynamic programming (you should select the correct alignment algorithm for the sequences given below and implement only that algorithm, either local / Smith-Waterman or global / Needleman-Wunch). Your implementation should take 2 sequences of any length (written in 2 different lines of the same text file) as input (text file should be accepted as a command line argument), include the following additional input arguments:

- a scoring matrix (should be able to take any scoring scheme in the format of square a matrix),
- a gap opening penalty value (a negative integer), and
- a gap extension penalty value (a negative integer).

In the first part of the output, include the aligned sequences in the classical 3-line notation (first line for the first sequence, second line for the '|' characters for the positions where there is a match between 2 sequences and space characters when there is no match, and third line for the second sequence) including '–' character for gaps in the aligned sequences. This should either be printed on screen or written in an output text file. The second part of the output should be the raw alignment score. The third part of the output should be the percent identity between the two aligned sequences, which can be calculated by multiplying the number of matches in the pair by 100 and dividing by the length of the aligned region, including gaps. You may use any programming language (Python is preferred), but your script should run on a basic Unix/Linux shell (such as bash) without any external dependencies. It is not okay to use specialized libraries such as the ones related to bioinformatics. Sample input-output is given below:

Global sequence alignment (blosum62, gap open: -10, gap extend: -5):

input:

```
MVSPADKTNVKAAWGVG
MVLSGEDKSNIKWGKV
```

output:

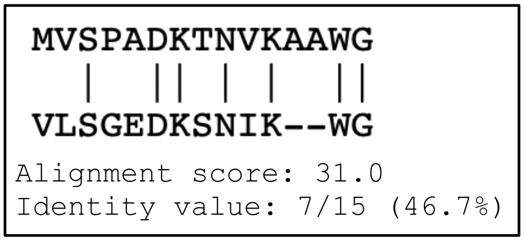
```
MV-SPADKTNVKAAWGVG
|| |  || | |  ||
MVLSGEDKSNIK--WGKV
Alignment score: 23.0
Identity value: 9/18 (50.0%)
```

Local sequence alignment (blosum62, gap open: -10, gap extend: -5):

input:

```
MVSPADKTNVKAAWGVG
MVLSGEDKSNIKWGKV
```

output:



```
MVSPADKTNVKAAWG
  |  || | |  ||
VLSGEDKSNIK--WG
Alignment score: 31.0
Identity value: 7/15 (46.7%)
```

a) Explain how your code runs and show the run command over an example (submit your script file as part of your assignment submission for testing).

b) Align the sequences of Protein A and Protein B given below, with the alignment algorithm of your choice (using your own implementation), paste the alignment output and percent identities in your answer sheet (parameters: BLOSUM62, gap open= -10, gap extend= -5). Discuss why did you chose this particular algorithm? Was the algorithm of your choice successful in the end?
(you can obtain BLOSUM62 matrix via the module at https://pypi.org/project/blosum/)

c) Investigate the impact of changing the scoring matrix on the alignment. Choose a different scoring matrix (e.g., BLOSUM45, BLOSUM90, etc.) and discuss how they affect the alignment result. Include specific changes in the alignment score and any variations in the aligned sequences. Discuss your results (you can obtain BLOSUM matrices via the module at https://pypi.org/project/blosum/).

d) Suppose we are looking for a region of functional importance that is similar between these two sequences. This region spans the whole of the shorter sequence but a subset of the longer one. Which algorithm would you choose, and what is the reason behind it? Why could the other algorithms not correctly identify this region?

```
> Protein_A
MDQLEEQIAEFKEAFSLFDKDGNTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMM
ARKMKDSEEEIREAFRVFDKDGNGSAAELRHVMTNLGEKLTDEEVDEMIIGMEVVEESDVLSPELEEMEV
YVRD
> Protein_B
MAKAQPEWFEAFSLFDKDGDGTITTKELGTVGQNPTEALQDINEVADNGTIFPFLTMMARKMKDTDSEEE
IREAFRVFDKDGNGYISAAELRHVMTLGELTDEVDEIREADIDGDGQVNYEEFVQMMTAKQ
```

Question 3 (35 points)

No implementation is required for this question; you can just apply the algorithms manually by hand. Please construct a multiple sequence alignment (MSA) using progressive alignment (ClustalW) for sequence fragments of *Gene X* of 5 different organisms given below. Steps: ***(1)*** construct global pairwise alignments (pairwise alignment parameters: match=1, mismatch=-1, gap open/extend/terminal=-1), ***(2)*** build the guide tree, and ***(3)*** progressive alignment (guided by the tree) – remember, once a gap always a gap!

>S1:human

ATCGATCGA

>S2:mouse

ATCGATCGT

>S3:monkey

ATCGATCGAT

>S4:frog

ATCATCGTAA

>S5:bacteria

ACCGGTATG

a) Show all pairwise global alignments including its output and partial scores tables and fill the similarity matrix below (Similarity = # of exact matches / alignment length).

	S1	S2	S3	S4	S5
S1					
S2					
S3					
S4					
S5					

b) Draw the guide tree and construct the final MSA using the guide tree. Show the guide tree, each step of your multiple alignments, and the finalized MSA output (multiple alignment parameters: match=1, mismatch=-1, gap open/extend/terminal=-1).

c) Score your MSA with Sum of Pairs (SP) Scoring. Calculate the scores column by column using the following scoring scheme: S(X,X) = 1, S(X,Y) = -1, S(X,-) = -1, S(-,X) = -1 and S(-,-) = 0. Show your calculation.

d) Please show the conserved residues and patterns on your MSA.

e) According to similarities in terms of *Gene X*, which one of these 4 organisms is the most distantly related organism to human and why? Would it be possible to find a different result if we used another gene instead of *Gene X* ?