BBM411/AIN411: Fundamentals of (Introduction to) Bioinformatics (Fall 2022)

Assignment 1

Due date: November 19, 2022, 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 document, and also, in the pdf and script filename) in the given time frame (tuncadogan@gmail.com). Please enter "BBM411 – Fall 2022 – Assignment 1" to the email subject.

Please note that, although sharing of 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 a gene? What is a protein? What is the biological relation between genes and proteins? What is the difference between a chromosome and genome?

b) What is gene expression? What is the name of the process that allows the production of multiple versions of a protein from the same gene, and why is this important?

c) Why do we align biomolecular sequences? What do the detected similarities tell us?

d) What is the use of scoring matrices in the process of alignment? What is the meaning of the numerical values in a scoring matrix and how are they calculated? Explain and discuss over the example of BLOSUM62 matrix. What is the meaning of "62" in the name of this matrix?

e) Which one of the following algorithms, BLAST or FASTA, is expected to have a faster run time, given the same search database and query sequence? Which one do you expect to be more accurate? Explain both over the algorithmic approaches and complexities of these algorithms.

Question 2 (60 points)

Implement the pairwise sequence alignment of amino acid (protein) sequences via dynamic programming. 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 arguments:

- alignment algorithm: local (Smith-Waterman) or global (Needleman-Wunch),
- scoring matrix (should be able to take any scoring scheme in the format of square a matrix),
- gap opening penalty (a negative integer),
- gap extension penalty (a negative integer),

an output 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 output should be the raw alignment score. The third 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 (in local alignment, terminal gaps are not included in the calculation). You may use any programming language (Python is preferred), but your script should run on a basic Unix/Linux shell (such as bash). It is okay to use basic data manipulation libraries such as NumPy; however, it is not okay to use specialized libraries especially ones related to bioinformatics.

a) Submit your script file via email, for testing. Explain dependencies and show the run command over an example.

b) Align the sequences of Protein A and Protein B given below, once with local and once with global alignment, paste the alignment output and percent identities below, discuss the difference in the alignments (parameters: BLOSUM62, gap open= -10, gap extend= -5). Observe the change in matching positions and the change in the alignment score. Terminal gaps in the global alignment will be treated the same as internal gaps.
(download BLOSUM62 matrix from: https://www.ncbi.nlm.nih.gov/Class/FieldGuide/BLOSUM62.txt)

c) We are looking for a region of functional importance that is similar between these two sequences. This region should be at least 25 amino acids long and should be more than 50% identical between two proteins. Which algorithm was able to locate it and what is the reason behind? Why the other algorithm could not identify this region?

d) Play around with the parameter values for gap open and gap extend, and switch your scoring matrix between different PAM and BLOSUM versions, until you get different alignments. Which parameters gave you different alignments compared to the local and global alignments from the previous item and why? Discuss your results.

> **Protein_A**
MTAIIKEIVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIYNLCAERHYDTAKFNCRVA
QYPFEDHNPPQLELIKPFCEDLDQWLSEDDNHVAAIHCKAGKGRTGVMICAYLLLSLAHRGKFLKAQEALDGDIFYGEVRTRDKKG
VTIPSQRRYVYYYSYLLKNHLDYRPVALLFHKMMFETIPMFSGGTCNPQFVVCFTNKQLKVKIYSSNSIRIEFTGPTRREDKFMYF
EFDRMFMFPQPLPVCGDIKVEFFHKQNKDKMFHFWVNTFFIPGPEETSEKVENGSLCDQEIDSICSIERADNDKEYLVLTLTKNDL
DKANKDKANRYFSPNFKVKLYFTKTVEEPSNPEASSSTSVTPDVSDNEPDHYRYSDTTDSDPENEPFDEDQHTQITKV
> **Protein_B**
GNTIGNDEYISFEIGALSLAGDIRIEFTNKQDDRMFMFWVNTSFVGNWGFSIIIITFIVRGIMYPLTKAQYTSMAKMRMLQPKIQA
MRERLGDD

Question 3 (30 points)

Please construct a multiple sequence alignment (MSA) using progressive alignment (ClustalW) for sequence fragments of *Gene X* of 4 different organisms given below. Steps:

1) Construct pairwise alignments (pairwise alignment parameters: match=1, mismatch=-1, gap open/extend/terminal=-1)
2) Build Guide Tree (by Neighbor Joining method)
3) Progressive Alignment (guided by the tree) – remember, once a gap always a gap! (multiple alignment parameters: match=1, mismatch=-1, gap open/extend/terminal=-1)

>S_1: Homo sapiens (human)
GTGCCTATGG
>S_2: Mus musculus (mouse)
GTGCCTTCG
>S_3: Capra hircus (goat)
GAGCCTAGGA
>S_4: Cyprinus carpio (fish)
TACGCGGTT

a) Show all pairwise alignments over partial scores tables (together with arrows and scores) and fill the similarity matrix below (Similarity = # of exact matches / alignment length).

	s_1	s_2	s_3	s_4
s_1				
s_2				
s_3				
s_4				

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.

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 3 organisms is the most closely related organism to human and why? Would it be possible to find a different result if we used another gene instead of *Gene X* ?