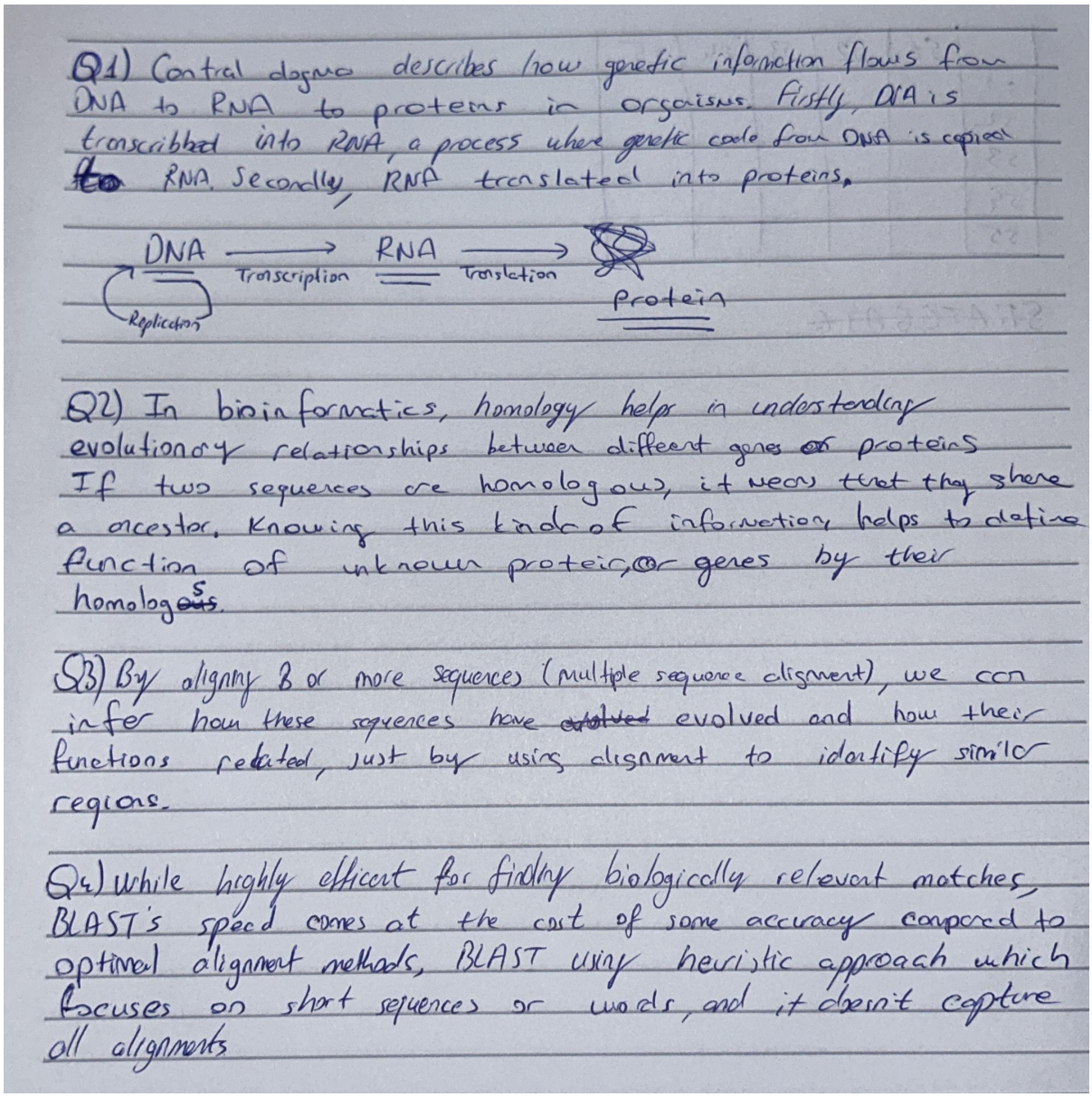
Çağrı Çakıroğlu

November,2023

BBM 411-ASSIGNMENT1



PART 2

HOW MY CODE WORK

-Ensuring “blosum” package is installed.

-pip install blosum

-My python file is “CagriCakiroglu411_A1.py” it needs “ input.txt” as :

MDQLEEQIAEFKEAFSLFDKDGNTITTKELGTVMRSLGQNPTEAELQDMINEVDA
DGNGTIDFPEFLTMMARKMKDSEEEIREAFRVFDKDGNGSAAELRHVMTNLGEK
LTDEEVDEMIIGMEVVEESDVLSPELEEMEVYVRD

-

MAKAQPEWFEAFSLFDKDGDGTITTKELGTVGQNPTEALQDINEVADNGTIFPFL
TMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMTLGELTDEVDEIRE
ADIDGDGQVNYEEFVQMMTAKQ

-And runs as :

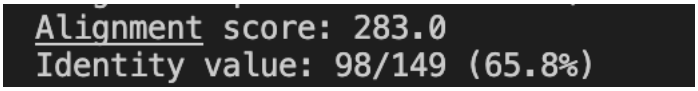
python CagriCakiroglu411_A1.py input.txt -10 -5 62 45

The provided Python script implements the Needleman-Wunsch algorithm for global sequence alignment, a fundamental technique in bioinformatics for aligning protein or nucleotide sequences. The script is designed to read two sequences from an input file with each sequence on a distinct line. It facilitates user customization by allowing the specification of gap opening and extension penalties, as well as parameters for two different BLOSUM (BLOcks SUbstitution Matrix) scoring matrices via command line arguments. These BLOSUM matrices are essential for scoring alignments based on amino acid substitutions, which is particularly important in protein sequence alignment. For each pair of BLOSUM parameters, the script computes and displays the aligned sequences, a match representation highlighting sequence matches, the final alignment score, and the identity percentage, reflecting the proportion of identical matches in the alignment. This utility serves as a valuable tool in bioinformatics for analyzing sequence similarity and functional relationships.

B)

FOR BLOOSUM 62

(Gap open: -10 , Gap extend:-5)



We used global alignment instead of local alignment because we wanted to compare the entire length of our two protein sequences. Global alignment is best when you expect the sequences to be similar all the way through, not just in certain parts. It tries to align every part of the sequence, from start to finish, which is different from local alignment that only looks for the most similar sections, no matter where they are. So, if our proteins are related in their whole structure, global alignment gives us a complete picture of their similarities and differences.Yes, the global alignment was successful because it aligned the entire sequences, showing us how similar they were from beginning to end.

C)

BLOSUM45

- When using the BLOSUM45 matrix, which is designed for sequences with lower similarity (approximately 45%), the alignment score increased from 283 to 377.
- This increase in alignment score indicates that the BLOSUM45 matrix is more permissive of mismatches and conservative substitutions between the sequences.
- The higher score suggests that the sequences may be more divergent, with many mismatches that are less penalized by BLOSUM45.
- The aligned sequences themselves might contain more mismatches with BLOSUM45 compared to BLOSUM62, as BLOSUM45 allows for more flexibility in scoring substitutions.

D)

The Smith-Waterman algorithm would be the preferred choice for identifying a region of functional importance that encompasses the entire shorter sequence while being a subset of the longer one. This algorithm excels at local alignment, allowing it to precisely pinpoint and score the similarity of the specific region in question. In contrast, global alignment algorithms like Needleman-Wunsch are better suited for aligning entire sequences, while other methods like overlap alignment are not tailored for this specific scenario. Therefore, Smith-Waterman's ability to focus on local similarities makes it the most appropriate choice for identifying the functional region in question.

PART 3

A)

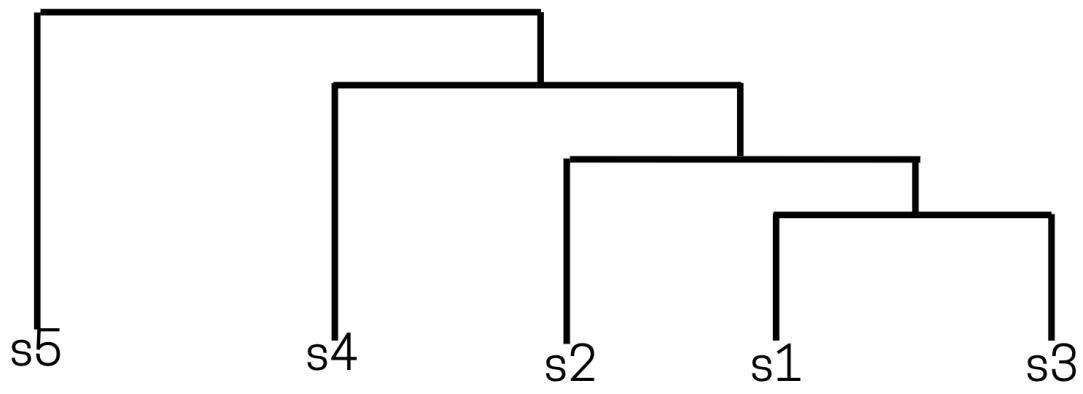
Pairs	Alignment	Score
S1-S2	ATCGATCGA ATCGATCGT	**7**
S1-S3	ATCGATCGA- ATCGATCGAT	8
S1-S4	ATCGATCG--A ATC-ATCGTAA	5
S1-S5	ATCG-A-TCGA ACCGGTAT-G-	1
S2-S3	ATCGATCG-T ATCGATCGAT	8
S2-S4	ATCGATCGT-- ATC-ATCGTAA	5
S2-S5	ATC-G-ATCGT ACCGGTAT-G-	1
S3-S4	ATCGATCG-AT ATC-ATCGTAA	5
S3-S5	ATCGATCGAT- ACCGGT--ATG	1
S4-S5	ATCATC-GTA-A A-C--CGGTATG	0

Table : Global Alignment Scores

	s1	s2	s3	s4	s5
s1	-				
s2	0.89	-			
s3	0.90	0.90	-		
s4	0.727	0.727	0.727	-	
s5	0.545	0.545	0.545	0.5	-

Similarity Matrix

B) Guide Tree



S1-S3 alignment:

A T C G A T C G A T
A T C G A T C G A -

Profile and s2 alignment:

A T C G A T C G - T
A T C G A T C G A T

Profile and s4 alignment:

A T C - A T C G T A A
A T C G A T C G A T A

Profile and s5 alignment:

A C C G G T - - A T - G
A T C G A T C G A T A G

Final MSA

s5	A	C	C	G	G	-	T	-	-	A	T	G
s4	A	T	C	-	A	T	C	G	T	A	A	-
s3	A	T	C	G	A	T	C	G	-	T	-	-
s2	A	T	C	G	A	T	C	G	A	T	-	-
s1	A	T	C	G	A	T	C	G	A	-	-	-

C)
SUM OF PAIRS

s5	A	C	C	G	G	-	T	-	-	A	T	G
s4	A	T	C	-	A	T	C	G	T	A	A	-
s3	A	T	C	G	A	T	C	G	-	T	-	-
s2	A	T	C	G	A	T	C	G	A	T	-	-
s1	A	T	C	G	A	T	C	G	A	-	-	-
SoP	10	2	10	2	2	2	2	2	-8	-6	-10	-10

D)

s5	A	C	C	G	G	-	T	-	-	A	T	G
s4	A	T	C	-	A	T	C	G	T	A	A	-
s3	A	T	C	G	A	T	C	G	-	T	-	-
s2	A	T	C	G	A	T	C	G	A	T	-	-
s1	A	T	C	G	A	T	C	G	A	-	-	-

residues **pattern**

E)

The most distantly related organism to humans, based on the similarity matrix for Gene X, appears to be the one represented by s5, as it branches out first in the UPGMA guide tree, indicating the greatest distance from humans. However, if a different gene were analyzed, the results could indeed vary because different genes can have different evolutionary histories and rates of mutation. Thus, the perceived evolutionary distance between species can change depending on the particular gene being studied.