AIN411: Introduction to Bioinformatics

(Fall 2023)

Assignment 1

Hüseyin Eren DOĞAN

2210765009

Question 1)

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

The central dogma describes the unidirectional flow of the genetic information from DNA to RNA then protein. Replication, transcription and translation are the main components of the central dogma. This unidirectional flow of information ensures the transfer of genetic instructions from DNA to proteins, forming the basis of cellular functions.

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

Homology in bioinformatics refers to the similarity in the biological genetic sequences -such as DNA, RNA, or protein structures- or functions between different species with common evolutionary origin. Significance lies in the fact that homologous sequences often share functional or structural characteristics, allowing researchers to infer information about the function of a gene or protein by studying its homologs in other organisms.

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

Multiple sequence alignment (MSA) is a bioinformatics technique used to align three or more biological sequences simultaneously. The purpose of MSA is to identify regions of similarity and difference among the sequences, providing insights into their evolutionary relationships and functional similarities. It aids in analyzing the structure and function of genes or proteins and helps identify conserved motifs.

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?

BLAST (Basic Local Alignment Search Tool) achieves faster execution than optimal sequence alignment algorithms by utilizing heuristics and shortcuts. Instead of exhaustively comparing all possible sequence pairs, BLAST rapidly identifies local regions of similarity and then extends these regions. While this approach sacrifices optimality, it significantly improves speed. BLAST may not always produce the optimal alignment, but it is a trade-off for efficiency, and its results are generally accurate for many practical applications.

Question 2)

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).

I used python (3.10.0) to write my code. You can run it with the command:

python3 <python_file.py> <input_file.txt> <name of blosum matrix> <gap_open> <gap_extend>

as example:

python3 sequence_alignment.py input.txt blosum62 -10 -5

b) Align the sequences of Protein A and Protein B, 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?

```
MDQLEEQIAEFKEAFSLFDKDG-NTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMK--DSEEEIREAFRVFDKDGNG--SAAELRHVMTNLGEKLTDEEVDEMIIGMEVVEESDVLSPELEEMEVYVRD
|     |  |  ||||||||||  ||||||||||    ||||||| ||| |||| || |||| || ||||||||||  ||||||||||||||||||  |||||||||| ||| ||| |||| |         |   |   |
M--AKAQ-PEWFEAFSLFDKDGDGTITTKELGTV----GQNPTEA-LQD-INEV-AD-NGTI-FP-FLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMT-LGE-LTD-EVDE-IREADIDGDGQVNYEEFVQM-MTAKQ

Alignment score: 282.0
Identity value: 98/149 (65.8%)
```

I chose a global sequence alignment algorithm, Needleman-Wunsch, because i want to look at overall similarity instead of finding specific regions of proteins or motifs. Since my goal is to compare the overall similarity and finding conserved regions, it can be counted as successful, i got 98 matches in the 149 length alignment and got a score with 282.

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.

Result for BLOSUM45:

```
MDQLEEQIAEFKEAFSLFDKDG-NTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMK--DSEEEIREAFRVFDKDGNG--SAAELRHVMTNLGEKLTDEEVDEMIIGMEVVEESDVLSPELEEMEVYVRD
|     |  |  ||||||||||  ||||||||||    ||||||| ||| |||| || |||| || ||||||||||  ||||||||||||||||||  |||||||||| ||| ||| |||| |         |   |   |
M--AKAQ-PEWFEAFSLFDKDGDGTITTKELGTV----GQNPTEA-LQD-INEV-AD-NGTI-FP-FLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMT-LGE-LTD-EVDE-IREADIDGDGQVNYEEFVQM-MTAKQ

Alignment score: 375.0
Identity value: 98/149 (65.8%)
```

Result for BLOSUM80:

```
MDQLEEQIAEFKEAFSLFDKDG-NTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMK--DSEEEIREAFRVFDKDGNG--SAAELRHVMTNLGEKLTDEEVDEMIIGMEVVEESDVLSPELEEMEVYVRD
|     |  |  ||||||||||  ||||||||||    ||||||| ||| |||| || |||| || ||||||||||  ||||||||||||||||||  |||||||||| ||| ||| |||| |         |   |   |
M--AKAQ-PEWFEAFSLFDKDGDGTITTKELGTV----GQNPTEA-LQD-INEV-AD-NGTI-FP-FLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMT-LGE-LTD-EVDE-IREADIDGDGQVNYEEFVQM-MTAKQ

Alignment score: 311.0
Identity value: 98/149 (65.8%)
```

Result for BLOSUM90:

```
MDQLEEQIAEFKEAFSLFDKDG-NTITTKELGTVMRSLGQNPTEAELQDMINEVDADGNGTIDFPEFLTMMARKMK--DSEEEIREAFRVFDKDGNG--SAAELRHVMTNLGEKLTDEEVDEMIIGMEVVEESDVLSPELEEMEVYVRD
|     |  |  ||||||||||  ||||||||||    ||||||| ||| |||| || |||| || ||||||||||  ||||||||||||||||||  |||||||||| ||| ||| |||| |         |   |   |
M--AKAQ-PEWFEAFSLFDKDGDGTITTKELGTV----GQNPTEA-LQD-INEV-AD-NGTI-FP-FLTMMARKMKDTDSEEEIREAFRVFDKDGNGYISAAELRHVMT-LGE-LTD-EVDE-IREADIDGDGQVNYEEFVQM-MTAKQ

Alignment score: 359.0
Identity value: 98/149 (65.8%)
```

There is no change on the matches or alignments while blosum matrix changes. Since blosum matrices is just scoring matrices, they just affects the alignment score.

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?

In this situation, it may be correct to use a local sequence alignment algorithm, so i'd choose Smith-Waterman algorithm. Because it focuses on specific regions instead of looking for similarity in whole sequence, so it makes it suitable for using this algorithm for such a problem.

Question 3)

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	>S2:mouse	>S3:monkey	>S4:frog	>S5:bacteria
ATCGATCGA	ATCGATCGT	ATCGATCGAT	ATCATCGTAA	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).

<u>Pairwise Alignments</u>

S1 vs S2

	-	A	T	C	G	A	T	C	G	A
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7
T	-2	0	2	1	0	-1	-2	-3	-4	-5
C	-3	-1	1	3	2	1	0	-1	-2	-3
G	-4	-2	0	2	4	3	2	1	0	-1
A	-5	-3	-1	1	3	5	4	3	2	1
T	-6	-4	-2	0	2	4	6	5	4	3
C	-7	-5	-3	-1	1	3	5	7	6	5
G	-8	-6	-4	2	0	2	4	6	8	7
T	-9	-7	-5	-3	-1	1	3	5	7	7

Alignment:

A T C G A T C G A

A T C G A T C G T

Similarity = 8/9

S1 vs S3

	-	A	T	C	G	A	T	C	G	A
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7
T	-2	0	2	1	0	-1	-2	-3	-4	-5
C	-3	-1	1	3	2	1	0	-1	-2	-3
G	-4	-2	0	2	4	3	2	1	0	-1
A	-5	-3	-1	1	3	5	4	3	2	1
T	-6	-4	-2	0	2	4	6	5	4	3
C	-7	-5	-3	-1	1	3	5	7	6	5
G	-8	-6	-4	2	0	2	4	6	8	7
A	-9	-7	-5	-3	-1	1	3	5	7	9
T	-10	-8	-6	-4	-2	0	2	4	6	8

Alignment:

A T C G A T C G A -

A T C G A T C G A T

Similarity = 9/10

S1

vs

S4

	-	A	T	C	G	A	T	C	G	A
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7
T	-2	0	2	1	0	-1	-2	-3	-4	-5
C	-3	-1	1	3	2	1	0	-1	-2	-3
A	-4	-2	0	2	2	3	2	1	0	-1
T	-5	-3	-1	1	1	2	4	3	2	1
C	-6	-4	-2	0	0	1	3	5	4	3
G	-7	-5	-3	-1	1	0	2	4	6	5
T	-8	-6	-4	-2	0	0	1	3	5	5
A	-9	-7	-5	-3	-1	1	0	2	4	6
A	-10	-8	-6	-4	-2	0	0	1	3	5

Alignment:

A T C G A T C G – A -

A T C – A T C G T A A

Similarity = 8/11

S1

vs

S5

	-	A	T	C	G	A	T	C	G	A
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7
C	-2	0	0	1	0	-1	-2	-3	-4	-5
C	-3	-1	-1	1	0	-1	-2	-1	-2	-3
G	-4	-2	-2	0	2	1	0	-1	0	-1
G	-5	-3	-3	-1	1	1	0	-1	0	-1
T	-6	-4	-2	-2	0	0	2	1	0	-1
A	-7	-5	-3	-3	-1	1	1	1	0	1
T	-8	-6	-4	-4	-2	0	2	1	0	0
G	-9	-7	-5	-5	-3	-1	1	1	2	1

Alignment:

A T C G - - A T C G A

A C C G G T A T - G -

Similarity = 6/11

S2

vs

S3

	-	A	T	C	G	A	T	C	G	T
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7
T	-2	0	2	1	0	-1	-2	-3	-4	-5
C	-3	-1	1	3	2	1	0	-1	-2	-3
G	-4	-2	0	2	4	3	2	1	0	-1
A	-5	-3	-1	1	3	5	4	3	2	1
T	-6	-4	-2	0	2	4	6	5	4	3
C	-7	-5	-3	-1	1	3	5	7	6	5
G	-8	-6	-4	2	0	2	4	6	8	7
A	-9	-7	-5	-3	-1	1	3	5	7	7
T	-10	-8	-6	-4	-2	0	2	4	6	8

Alignment:

A T C G A T C G - T

A T C G A T C G A T

Similarity = 9/10

S2

vs

S4

	-	A	T	C	G	A	T	C	G	T
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7
T	-2	0	2	1	0	-1	-2	-3	-4	-5
C	-3	-1	1	3	2	1	0	-1	-2	-3
A	-4	-2	0	2	2	3	2	1	0	-1
T	-5	-3	-1	1	1	2	4	3	2	1
C	-6	-4	-2	0	0	1	3	5	4	3
G	-7	-5	-3	-1	1	0	2	4	6	5
T	-8	-6	-4	-2	0	0	1	3	5	7
A	-9	-7	-5	-3	-1	1	0	2	4	6
A	-10	-8	-6	-4	-2	0	0	1	3	5

Alignment:

A T C G A T C G T - -

A T C – A T C G T A A

Similarity = 8/11

S2

vs

S5

	-	A	T	C	G	A	T	C	G	T
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7
C	-2	0	0	1	0	-1	-2	-3	-4	-5
C	-3	-1	-1	1	0	-1	-2	-1	-2	-3
G	-4	-2	-2	0	2	1	0	-1	0	-1
G	-5	-3	-3	-1	1	1	0	-1	0	-1
T	-6	-4	-2	-2	0	0	2	1	0	1
A	-7	-5	-3	-3	-1	1	1	1	0	0
T	-8	-6	-4	-4	-2	0	2	1	0	1
G	-9	-7	-5	-5	-3	-1	1	1	2	1

Alignment:

A T C G - - A T C G T

A C C G G T A T C G -

Similarity = 6/11

S3

vs

S4

	-	A	T	C	G	A	T	C	G	A	T
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
T	-2	0	2	1	0	-1	-2	-3	-4	-5	-6
C	-3	-1	1	3	2	1	0	-1	-2	-3	-4
A	-4	-2	0	2	2	3	2	1	0	-1	-2
T	-5	-3	-1	1	1	2	4	3	2	1	0
C	-6	-4	-2	0	0	1	3	5	4	3	2
G	-7	-5	-3	-1	1	0	2	4	6	5	4
T	-8	-6	-4	-2	0	0	1	3	5	5	6
A	-9	-7	-5	-3	-1	1	0	2	4	6	5
A	-10	-8	-6	-4	-2	0	0	1	3	5	5

Alignment:

A T C G A T C G – A T

A T C – A T C G T A A

Similarity = 8/11

S3

Vs

S5

	-	A	T	C	G	A	T	C	G	A	T
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
C	-2	0	0	1	0	-1	-2	-3	-4	-5	-6
C	-3	-1	-1	1	0	-1	-2	-1	-2	-3	-4
G	-4	-2	-2	0	2	1	0	-1	0	-1	-2
G	-5	-3	-3	-1	1	1	0	-1	0	-1	-2
T	-6	-4	-2	-2	0	0	2	1	0	-1	0
A	-7	-5	-3	-3	-1	1	1	1	0	1	0
T	-8	-6	-4	-4	-2	0	2	1	0	0	2
G	-9	-7	-5	-5	-3	-1	1	1	2	1	1

Alignment:

A T C G A T C G A T -

A C C G G T - - A T G

Similarity = 6/11

S4

Vs

S5

	-	A	T	C	G	A	T	C	G	A	T
-	0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-10
A	-1	1	0	-1	-2	-3	-4	-5	-6	-7	-8
C	-2	0	0	1	0	-1	-2	-3	-4	-5	-6
C	-3	-1	-1	1	0	-1	0	-1	-2	-3	-4
G	-4	-2	-2	0	0	-1	-1	1	0	-1	-2
G	-5	-3	-3	-1	-1	-1	-2	0	0	-1	-2
T	-6	-4	-2	-2	-2	0	-1	-1	1	0	-1
A	-7	-5	-3	-3	-1	-1	-1	-2	0	2	1
T	-8	-6	-4	-4	-2	0	-1	-2	-1	1	1
G	-9	-7	-5	-5	-3	-1	-1	0	-1	0	0

Alignment:

A T C A T C G – T A A -

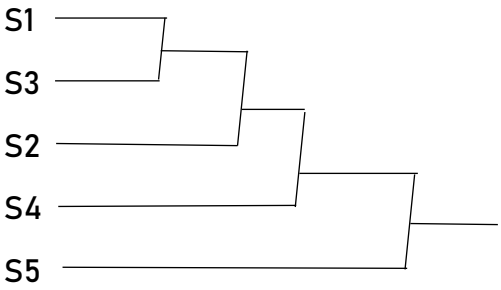
A – C - - C G G T A T G

Similarity = 6/12

Similarity Matrix:

	S1	S2	S3	S4	S5
S1	1	8/9	9/10	8/11	6/11
S2	8/9	1	9/10	8/11	6/11
S3	9/10	9/10	1	8/11	6/11
S4	8/11	8/11	8/11	1	1/2
S5	6/11	6/11	6/11	1/2	1

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.



Step 1: Aligning S1 and S3

ATCGATCGA-

ATCGATCGAT

Profile 1: ATCGATCGAT

Step 2: Aligning Profile 1 and S2

ATCGATCGA-

ATCGATCGAT

ATCGATCG-T

Profile 2: ATCGATCGAT

Step 3: Aligning Profile 2 and S4

ATCGATCGA--

ATCGATCGAT-

ATCGATCG-T-

ATC-ATCG-TA

Profile 3: ATCGATCGAT

Step 4: Aligning Profile 3 and S5

ATCGATCGA--

ATCGATCGAT-

ATCGATCG-T-

ATC-ATCG-TA

ACCGGT--ATG

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.

1st column: S(A,A) x (4+3+2+1) = 1 x 10 = 10

2nd column: S(T,T) x (3+2+1) + S(T,C) x 4 = 1 x 6 + (-1) x 4 = 6 – 4 = 2

3rd column: S(C,C) x (4+3+2+1) = 1 x 10 = 10

4th column: S(G,G) x (3+2+1) + S(G, -) x 4 = 1 x 6 + (-1) x 4 = 6 – 4 = 2

5th column: S(A,A) x (3+2+1) + S(A,G) x 4 = 1 x 6 + (-1) x 4 = 6 – 4 = 2

6th column: S(T,T) x (4+3+2+1) = 1 x 10 = 10

7th column: S(C,C) x (3+2+1) + S(C, -) x 4 = 1 x 6 + (-1) x 4 = 6 – 4 = 2

8th column: S(G,G) x (3+2+1) + S(G, -) x 4 = 1 x 6 + (-1) x 4 = 6 – 4 = 2

9th column: S(A,A) x (2+1) + S(A,-) x (2 x 3) + S(-,-) = 1 x 3 + (-1) x 6 + 0 = -3

10th column: S(T,T) x (3+2+1) + S(-,T) x 4 = 1 x 6 + (-1) x 4 = 6 – 4 = 2

11th column: S(-,-) x (2+1) + S(-,A) x 3 + S(-, G) x 3 + S(A,G) = 0 – 3 – 3 – 1 = -7

Sum of Pairs: 10 + 2 + 10 + 2 + 2 + 10 + 2 + 2 – 3 + 2 – 7 = 32

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?

If we look at Gene X of these four organisms, we can say that bacteria is the most distantly related organism to human, we can take a look at similarity matrix, the lowest similarity value is between human and bacteria. And we got the lowest score in alignments when we aligned s1(human) and s5(bacteria).

There is probability that we can get a different result when we used another gene instead of gene X. Different genes can evolve at different rates and under different selective pressures. Analyzing a different gene might reveal variations in genetic relatedness among organisms. Additionally, some genes may be subject to horizontal gene transfer, leading to discrepancies in the evolutionary history inferred from different genes. Therefore, using another gene could indeed yield different results in terms of the perceived relationships between organisms.