Hacettepe University

BBM 411: Fundamentals of Bioinformatics - 2022 FALL

Assignment 1

TUNCA DOĞAN

Student name:
Ataberk ASAR

Student Number:
2210356135

1 Question 1

(a) **Gene:** Physical and functional unit structure of DNA
Protein: Complex molecules made up of amino acids
Genes are used to synthesize proteins
Set of all chromosomes is called genome

(b) **Gene Expression:** Process of producing RNA and Proteins
Alternative splicing is the process of producing multiple versions of a protein from the same gene. It is important since it increases diversity

(c) We align biomolecular sequences to detect similarities between these sequences. These similarities used to identify evolutional, functional, or structural relations between them.

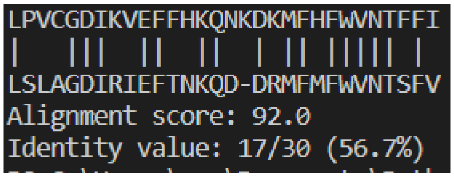
(d) Scoring matrices allow us to consider chemical, physical, and evolutionary relationships between different pair of nucleotides or amino acids. Scores obtained by studying occuring frequencies. Numbers of BLOSUM matrices indicates how similar protein sequences used to obtain frequencies (e.g. BLOSUM62 obtained by using protein sequences with proteins with less than 62% similarity).

(e) BLAST looks for known significant patterns in the sequences using local sequence alignment, while FASTA uses both local and global alignment consecutively. Thus BLAST works faster than FASTA.

2 Question 2

(a) **Run command:**
python q2.py seq_file alignment_algorithm scoring_matrix_file gap_open gap_extend output_file
Note that output file is not necessary, and used to see partial scores table
Aligned sequences will be printed onto command prompt

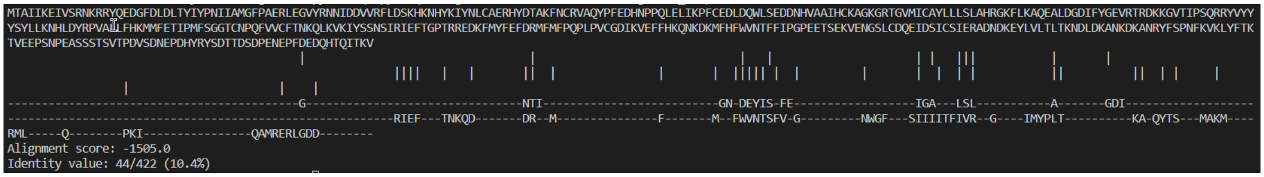
(b) **Local output:**



```
LPVCGDIKVEFFHKQNKDKMFHFWVNTFFI
|   |||  ||  ||  | || ||||| |
LSLAGDIRIEFTNKQD-DRMFMFWVNTSFV
Alignment score: 92.0
Identity value: 17/30 (56.7%)
```

Figure 1: Local Alignment

Global output:



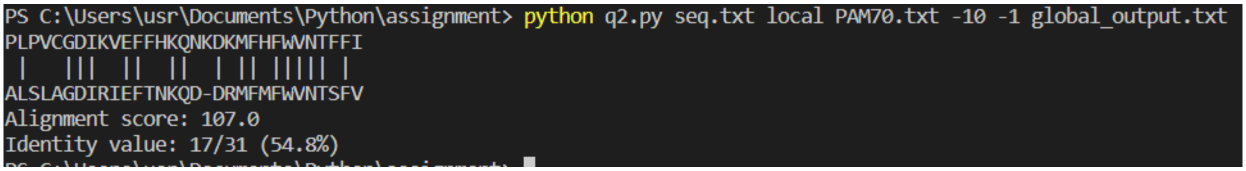
```
MTAIIKEIVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIYNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCEDLDQWLSEDDNHVAAIHCKAGKGRTGVMICAYLLLSLAHRGKFLKAQEALDGDIFYGEVRTRDKKGVTIPSQRRYVYY
YSYLLKNHLDYRPVALLFHKMMFETIPMFSGGTCNPQFVVCFTNKQLKVKIYSSNSIRIEFTGPTRREDKFMYFEFDRMFMFPQPLPVCGDIKVEFFHKQNKDKMFHFWVNTFFIPGPEETSEKVENGSLCDQEIDSICSIERADNDKEYLVLTLTKNDLDKANKDKANRYFSPNFKVKLYFTK
TVEEPSNPEASSSTSVTPDVSDNEPDHYRYSDTTDSDPENEPFDEDQHTQITKV
                                           |                                 |                              |   |                     | |   |||           |       |
                                                         ||||   |   |       ||  |               |       |  ||||| |  |         |       |  |  | |           ||          ||  | |     |
                 |                      |    |
-------------------------------------------G--------------------------------NTI--------------------------GN-DEYIS-FE------------------IGA---LSL-----------A-------GDI-------------------
---------------------------------------------------------RIEF---TNKQD-------DR--M---------------F-------M--FWVNTSFV-G---------NWGF---SIIIITFIVR--G----IMYPLT----------KA-QYTS---MAKM----
RML-----Q--------PKI----------------QAMRERLGDD--------
Alignment score: -1505.0
Identity value: 44/422 (10.4%)
```

Figure 2: Global Alignment

Note that global alignment with blosum 62 given as a seperate file for easier grading

(c) As seen from part b, local alignment algorithm should be used to find these. Reason for this is that local alignment algorithm proposed to find subsequences with highest alignment scores between two sequences. On the other hand, global alignment algorithm could not find it, since it tries to align all of the sequences.

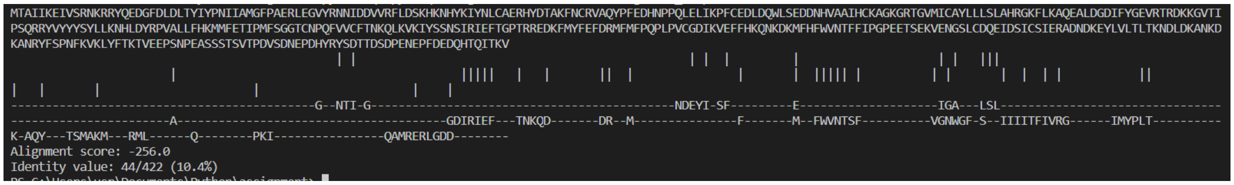
(d) **PAM70 Local Alignment Open Penalty = -10 Extend Penalty = -1:**



```
PS C:\Users\usr\Documents\Python\assignment> python q2.py seq.txt local PAM70.txt -10 -1 global_output.txt
PLPVCGDIKVEFFHKQNKDKMFHFWVNTFFI
 |   |||  ||  ||  | || ||||| |
ALSLAGDIRIEFTNKQD-DRMFMFWVNTSFV
Alignment score: 107.0
Identity value: 17/31 (54.8%)
PS C:\Users\usr\Documents\Python\assignment>
```

Figure 3: PAM70 Local Alignment

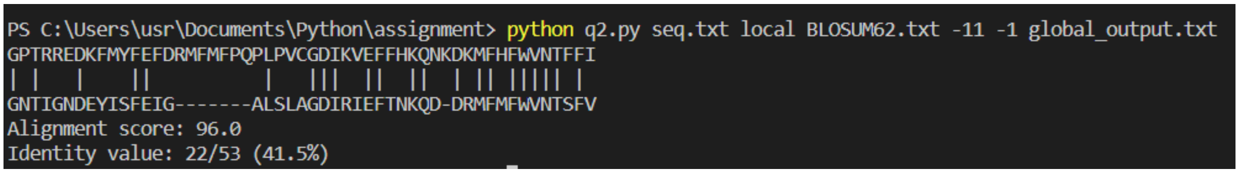
PAM70 Global Alignment Open Penalty = -10 Extend Penalty = -1:



```
MTAIIKEIVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIYNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCEDLDQWLSEDDNHVAAIHCKAGKGRTGVMICAYLLLSLAHRGKFLKAQEALDGDIFYGEVRTRDKKGVTI
PSQRRYVYYYSYLLKNHLDYRPVALLFHKMMFETIPMFSGGTCNPQFVVCFTNKQLKVKIYSSNSIRIEFTGPTRREDKFMYFEFDRMFMFPQPLPVCGDIKVEFFHKQNKDKMFHFWVNTFFIPGPEETSEKVENGSLCDQEIDSICSIERADNDKEYLVLTLTKNDLDKANKD
KANRYFSPNFKVKLYFTKTVEEPSNPEASSSTSVTPDVSDNEPDHYRYSDTTDSDPENEPFDEDQHTQITKV
                                               | |                                                | |  |         |                    | |   |||
                       |                                         |||||   |   |       ||  |               |       |  ||||| |          | |       |  |  | |           ||
|   |       |                      |                      |    |
--------------------------------------------G--NTI-G--------------------------------------------NDEYI-SF---------E--------------------IGA---LSL--------------------------------
-----------------------A---------------------------------------GDIRIEF---TNKQD-------DR--M---------------F-------M--FWVNTSF----------VGNWGF-S--IIIITFIVRG------IMYPLT----------
K-AQY---TSMAKM---RML------Q--------PKI----------------QAMRERLGDD--------
Alignment score: -256.0
Identity value: 44/422 (10.4%)
PS C:\Users\usr\Documents\Python\assignment>
```

Figure 4: PAM70 Global Alignment

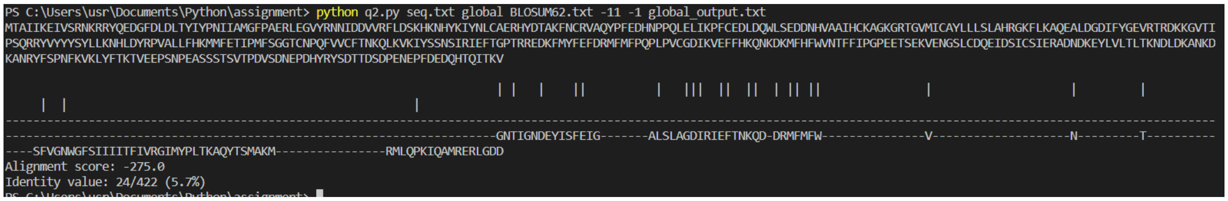
BLOSUM62 Local Alignment Open Penalty = -11 Extend Penalty = -1:



```
PS C:\Users\usr\Documents\Python\assignment> python q2.py seq.txt local BLOSUM62.txt -11 -1 global_output.txt
GPTRREDKFMYFEFDRMFMFPQPLPVCGDIKVEFFHKQNKDKMFHFWVNTFFI
| |   |    ||          |   |||  ||  ||  | || ||||| |
GNTIGNDEYISFEIG-------ALSLAGDIRIEFTNKQD-DRMFMFWVNTSFV
Alignment score: 96.0
Identity value: 22/53 (41.5%)
```

Figure 5: BLOSUM62 Local Alignment

BLOSUM62 Global Alignment Open Penalty = -11 Extend Penalty = -1:



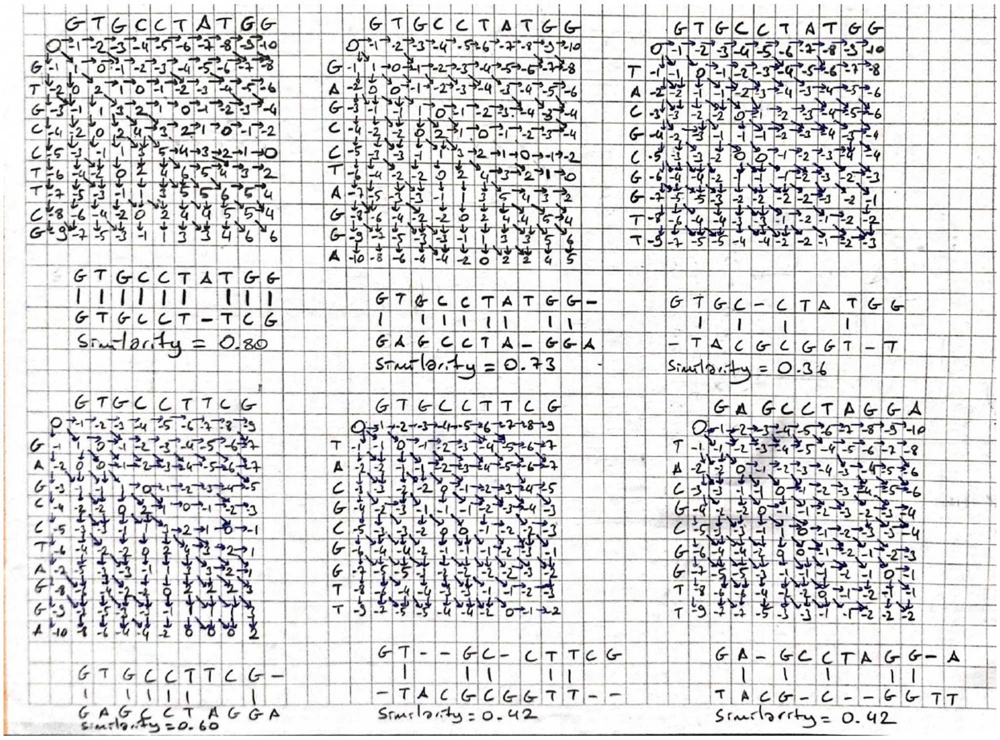
```
PS C:\Users\usr\Documents\Python\assignment> python q2.py seq.txt global BLOSUM62.txt -11 -1 global_output.txt
MTAIIKEIVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIYNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCEDLDQWLSEDDNHVAAIHCKAGKGRTGVMICAYLLLSLAHRGKFLKAQEALDGDIFYGEVRTRDKKGVTI
PSQRRYVYYYSYLLKNHLDYRPVALLFHKMMFETIPMFSGGTCNPQFVVCFTNKQLKVKIYSSNSIRIEFTGPTRREDKFMYFEFDRMFMFPQPLPVCGDIKVEFFHKQNKDKMFHFWVNTFFIPGPEETSEKVENGSLCDQEIDSICSIERADNDKEYLVLTLTKNDLDKANKD
KANRYFSPNFKVKLYFTKTVEEPSNPEASSSTSVTPDVSDNEPDHYRYSDTTDSPENEPFDEDQHTQITKV

                                                                        | |   |    ||          |   |||  ||  ||  | || ||               |                    |         |
     |  |                                                 |
-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------GNTIGNDEYISFEIG-------ALSLAGDIRIEFTNKQD-DRMFMFW---------------V--------------------N---------T---------
----SFVGNWGFSIIIITFIVRGIMYPLTKAQYTSMAKM---------------RMLQPKIQAMRERLGDD
Alignment score: -275.0
Identity value: 24/422 (5.7%)
PS C:\Users\usr\Documents\Python\assignment>
```

Figure 6: BLOSUM62 Global Alignment

As one can see, using different scoring matrix and different gap penalties will give us different results. Thus choosing right scoring matrix and gap penalties is important to determine relations between sequences.

3 Question 3

(a)



```
GTGCCTATGG
|||||| |||
GTGCCT-TCG
Similarity = 0.80
```

```
GTGCCTATGG-
| ||||| ||
GAGCCTA-GGA
Similarity = 0.73
```

```
GTGC-CTATGG
 | |  |  |
-TACGCGGT-T
Similarity = 0.36
```

```
GTGCCTTCG-
| ||||   |
GAGCCTAGGA
similarity = 0.60
```

```
GT--GC-CTTCG
 |   ||  ||
-TACGCGGTT--
Similarity = 0.42
```

```
GA-GCCTAGG-A
 |  |  |  ||
TACG-C--GGTT
Similarity = 0.42
```

Figure 7: Partial Scores Table

(b)

	s1	s2	s3	s4
s1	-	.8	.73	.36
s2		-	.6	.42
s3			-	.42
s4				-

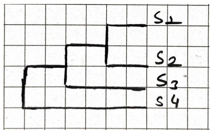


Figure 8: Guide Tree

```
GTGCCTATGG-
GTGCCT-TCG-
GAGCCTA-GGA
TACGCG--GTT
```

(c) $0 + 0 + 0 + 6 + 0 + 0 - 4 - 3 + 0 + 0 - 5 = -6$

(d) :::*::   .::

(e) Mouse is the most similar one to human, since it has the highest similarity score to human. Using another gene can change the results.

References

PAM70 Scoring Matrix