BBM 411 REPORT

Fundamentals of (Introduction to) Bioinformatics

Assignment 1

name, surname: Duygu Sena Öztürk
number: 2200356056

* How to use my code:

```
python duygu_script.py input.txt global BLOSUM62.txt -10 -5
```

- duygu_script.py: script name
- input.txt: input file that has the 2 sequences seperated with new line chor.
- global: either global or local (all lowercase)
- BLOSUM62.txt: file name that has the matrix
- -10: opening gap penalty
- -5: gap extension penalty

* (to run using command line)

Question 1

a) Gene is the both physical and the functional unit of heredity, it's a section of a DNA that produces a functional product.
A genome is an organism's full set of DNA which consists genes, chromosomes. Chromosomes are coiled structures that keeps the DNA strands safe.

b) Gene expression is the process by which the information encoded in a gene is used to either make RNA molecules that code for proteins or to make non-coding RNA molecules that serve other functions. Alternative splicing is a way to get more than 1 protein product out of the same gene.

c) Aligning biomolecular sequences are useful for identifying sequence similarity, producing phylogenetic trees and developing homology models of protein structures The similarities in genes.

* This has several advantages: (i) it allows a high sequence flexibility of exonic regulatory sequences that puts no constrains on coding requirements, (ii) the protein interaction can be influenced by small changes in the concentration of regulatory proteins which allows the alternative usage of exons *

d) Scoring matrices are used to determine the relative score of two charachters matching in a sequence alignment.

Scoring metrics (statistical versus biological): When evaluating a sequence alignment, one would like to know how meaningful it is. This requires a scoring matrix, or a table of values that describes the probability of a biologically meaningful amino-acid or nucleotide residue-pair occurring in an alignment. Scores for each position are obtained frequencies of substitutions in blocks of local alignments of protein sequences

BLOSUM r: the matrix built from blocks with less than r% of similarity – E.g., BLOSUM62 is the matrix built using sequences with less than 62% similarity (sequences with ≥ 62% identity were clustered) – Note: BLOSUM 62 is the default matrix for protein BLAST. Experimentation has shown that the BLOSUM-62 matrix is among the best for detecting most w eak protein similarities

e) BLAST is expected to have a faster run time than FASTA. But FASTA is more sensitive than BLAST. FASTA uses local sequence alignment first, then extends to global alignment while BLAST uses local sequence alignment.

Question 2

b)

Global alignment

```
(base) duygusenaozturk@Duygu-MacBook-Air pythonProject8 %  python duygu_script.py input.txt global BLOSUM62.txt -10 -5
MTAIIKEIVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIYNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCEDLDQWLSEDDNHVAAIHCKAGKGRTGVMICAYLLLSLAHRGKFLKAQEALDGDIFYGEVRTRDKKGVT
PSQRRYVYYYSYLLKNHLDYRPVALLFHKMMFETIPMFSGGTCNPQFVVCFTNKQLKVKIYSSNSIRIEFTGPTRREDKFMYFEFDRMFMFPQPLPVCGDIKVEFFHKQNKDKMFHFWVNTFFIPGPEETSEKVENGSLCDQEIDSICSIERADNDKEYLVLTLTKNDLDKANK
KANRYFSPNFKVKLYFTKTVEEPSNPEASSSTSVTPDVSDNEPDHYRYSDTTDSDPENEPFDEDQHTQITKV

                                        |                            |                          |  |       |       |        |     |   |              |
              |                                        |||||  |  |      ||  |              |      |  ||||| |         |  |          ||           |
      |       ||         |           |            |   | |
-----------------------------------------G-------------------------------NTI--------------G------------N-DEYISF---E---I---------G---------ALS-----L----A----------------G--
-----------------D-----------------------------------------IRIEF---TNKQD------DR--M--------------F-------M--FWVNTSF----------V--G---NW---GF-SI---------IIITF----IVRGI-
----MY-P-------LTK-----A--QYTSMAKMRM-L---QP---KI-QAMR---ERLG-D-D---------
Alignment score: -1613.0
Identity value: 42/423 9.929078014184398%
```

* alignment score was really low with an identity value close to only 10 %

Local alignment

```
(base) duygusenaozturk@Duygu-MacBook-Air pythonProject8 %  python duygu_script.py input.txt local BLOSUM62.txt -10 -5
DIKVEFFHKQNKDKMFHFWVNTFFI
||  ||  ||  | || ||||| |
DIRIEFTNKQD-DRMFMFWVNTSFV
Alignment score: 82.0
Identity value: 15/25 60.0%
```

* alignment score was higher also identity value was even greater than 50 %

c) Smith-Waterman local alignment algorithm was able to locate a 25 aminoacids-long functional region with more than 50 % identical between them. The local alignment function performs better when we are looking for regions. The global algorithm tries to match them as whole which can make these regions hard to identify.

d) Pam is based on global alignments of closely related proteins while Blosum is based on local alignments. BLOSUM performed better and showed better results.
I've tried PAM250, PAM350 and BLOSUM90.
Even though results were close PAM300 performed worse in both global and local alignment.

according to identical percentage resul

Global Alignments

BLOSUM90

```
(base) duygusenaozturk@Duygu-MacBook-Air pythonProject8 %  python main.py input.txt global blosum90.txt -10 -5
MTAIIKEIVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIYNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCEDLDQWLSEDDNHVAAIHCKAGKGRTGVMICAYLL
LSLAHRGKFLKAQEALDGDIFYGEVRTRDKKGVTIPSQRRYVYYYSYLLKNHLDYRPVALLFHKMMFETIPMFSGGTCNPQFVVCFTNKQLKVKIYSSNSIRIEFTGPTRREDKFMYFEFDRMFMFPQPLPVCGDIKVEF
FHKQNKDKMFHFWVNTFFIPGPEETSEKVENGSLCDQEIDSICSIERADNDKEYLVLTLTKNDLDKANKDKANRYFSPNFKVKLYFTKTVEEPSNPEASSSTSVTPDVSDNEPDHYRYSDTTDSDPENEPFDEDQHTQIT
KV
                                        |                  |  |                        |             |  |                  | |
  |     |   |      |    |    |                                                 |||||                    |      ||  |
|      |  ||||| |  |        |      |  | |          ||         ||  | |    |                    |        |        |   | |

-----------------------------------------G------------------NT--I-----------------G-------N--------------DEYISFE-------------------IGA---
--LS----L----A------G----D----I---------------------------R------I--------E----FTNKQ----------------------D-------DR--M---------------
F-------M--FWVNTSFV-G---------NWGF---SIIIITFIVRG------IMYPLT----------KA-QYTS---MAKM---RM-------------LQPKI-----QAMR---------ERLG-D-D------
---
Alignment score: -1562.0
Identity value: 49/423 11.58392434988179 6%
```

PAM250

```
(base) duygusenaozturk@Duygu-MacBook-Air pythonProject8 %  python main.py input.txt global pam250.txt -10 -5
MTAIIKEIVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIYNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCEDLDQWLSE
LSLAHRGKFLKAQEALDGDIFYGEVRTRDKKGVTIPSQRRYVYYYSYLLKNHLDYRPVALLFHKMMFETIPMFSGGTCNPQFVVCFTNKQLKVKIYSSNSIRIEFTGPTRREDK
FHKQNKDKMFHFWVNTFFIPGPEETSEKVENGSLCDQEIDSICSIERADNDKEYLVLTLTKNDLDKANKDKANRYFSPNFKVKLYFTKTVEEPSNPEASSSTSVTPDVSDNEPD
KV

                  |  |         |            |  |    |              |                      |||||  |  |
|      |  ||||| |         |  |           |         |      |     ||       |                      |

-----------------------------------------------------------------------------------------------------------------
--IGNDE-Y----------ISF-E----------IGA----------L-S-L-----A---------------GD-----------------------IRIEF---TNKQD--
F-------M--FWVNTSF----------V--GNW------GFSIIII-----TFIVRGI-MYPLT-----KAQ--YTSMAKMRM-------------------LQPKI------
---
Alignment score: -1537.0
Identity value: 38/423 8.983451536643026%
```

PAM300

```
(base) duygusenaozturk@Duygu-MacBook-Air pythonProject8 %  python main.py input.txt global pam300.txt -10 -5
MTAIIKEIVSRNKRRYQEDGFDLDLTYIYPNIIAMGFPAERLEGVYRNNIDDVVRFLDSKHKNHYKIYNLCAERHYDTAKFNCRVAQYPFEDHNPPQLELIKPFCEDLDQWLSEDDN
LSLAHRGKFLKAQEALDGDIFYGEVRTRDKKGVTIPSQRRYVYYYSYLLKNHLDYRPVALLFHKMMFETIPMFSGGTCNPQFVVCFTNKQLKVKIYSSNSIRIEFTGPTRREDKFMY
FHKQNKDKMFHFWVNTFFIPGPEETSEKVENGSLCDQEIDSICSIERADNDKEYLVLTLTKNDLDKANKDKANRYFSPNFKVKLYFTKTVEEPSNPEASSSTSVTPDVSDNEPDHYR
KV

                  |             |               |       |              |                       |||||  |
|      |  ||||| |         |  |          ||         |      |     |        |                            |

-----------------------------------------------------------------------------------------------------------------G
--------Y----------ISF------E-----IGA--------------L-----S-L------------AGD-----------------------IRIEF---T--NKQ---
F-------M--FWVNTSF----------V--GNW------GF-SII-II---TFIVRGI-MYPLT-----K-AQ-YTSMAKMRM-----------------------L---QP---K
---
Alignment score: -1540.0
Identity value: 35/423 8.274231678486997%
```

local alignments:

BLOSUM90

```
(base) duygusenaozturk@Duygu-MacBook-Air pythonProject8 %  python duygu_script.py input.txt local BLOSUM62.txt -10 -5
DIKVEFFHKQNKDKMFHFWVNTFFI
||  ||  ||  | || ||||| |
DIRIEFTNKQD-DRMFMFWVNTSFV
Alignment score: 82.0
Identity value: 15/25 60.0%
```

PAM300

```
(base) duygusenaozturk@Duygu-MacBook-Air pythonProject8 %  python main.py input.txt local pam300.txt -10 -5
LPVCGDIKVEFFHKQNKDKMFHFWVNTFFIPG
|   |||  ||  ||  | || ||||| |
LSLAGDIRIEFTNKQD-DRMFMFWVNTSFVGN
Alignment score: 122.0
Identity value: 17/32 53.125%
```

PAM250

```
(base) duygusenaozturk@Duygu-MacBook-Air pythonProject8 %  python main.py input.txt local pam250.txt -10 -5
GDIKVEFFHKQNKDKMFHFWVNTFFI
|||  ||  ||  | || ||||| |
GDIRIEFTNKQD-DRMFMFWVNTSFV
Alignment score: 99.0
Identity value: 16/26 61.53846153846154%
```

* : the alignments that I used are highlighted in yellow

Question 3: Pairwise sequence alignments

1) allign

a) >S1 >S2
```
GTGCCTATGG
GTGCCT-TCG
```

b) >S1 >S3
```
GTGCCTATGG-
GAGCCTA-GGA
```

c) >S1 >S4
```
GTGC-CTATGG
-TACGCGGT-T
```
different one
```
GT--GC-CTATGG
-TACGCGGT-T--
```

** had more alignments than 1 (both c and d and f)

e) **
```
GTGC-C--TTCG
-TACGCGGTT--
```
different one

d) >S2 >S3
```
GTGCCTTCG-
GAGCCTAGGA
```

e) >S2 >S4 *
```
GT--GC-CTTCG
-TACGCGGTT--
```

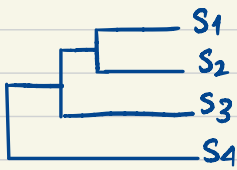
f) >S3 >S4
```
GA-GCCTAGG-A
TACGC---GGTT
```
different one *
```
GAGC-CTAGGA-
TA-CGC--GGTT
```

a) 8/10	80 %	d) 6/10	60 %
b) 8/11	72.7 %	e) 5/12	41.6 %
c) 9/11	36.3 %	f) 5/12	41.6 %

	S_1	S_2	S_3	S_4
S_1	1	0.80	0.73	0.36
S_2	1 0.80	1	0.60	0.42
S_3	2 0.73	3 0.60	1	0.42
S_4	0.36	0.42	0.42	1

b) Guide tree

Phylogenetic tree



$(((S_1, S_2), S_3), S_4)$

since S_1 and S_2 are already aligned from a, I can continue with aligning S_3 with them

*

GTGCCTATGG — Seq 1

GTGCCT-TCG — Seq 2

GAGCCTAGGA *1

Seq3

GAGCCTAGGA

either 1 or 2

*2 GAGCCTA-GGA

Not chosen because it'll also mean 2 gaps for now.

*

GTGC-CTATGG--

GTGC-CT-TCG--

GAGC-CTA-GGA-

-TACGC---GGTT

* * *

GAGCCTA-GGA

seq 4

TACGCGGTT

gaps to match the C's and GG's

if they are frozen (not just the gaps created) then it'd look like this

GTGCCTATGG

GTGCCT-TCG

GAGCCTAGGA

-TACGCGGTT

*

Sum of Pairs Scoring

```
GTGC-CTATGG--
GTGC-CT-TCG--
GAGC-CTA-GGA-
-TACGC---GGTT
0 0 0 6 -3 6 0 -3 -3 0 6 -5 -3
```

d) the column 4, 6, 11 were conserved in all 4 of them.

c)

	1	2	3	4	5	6	
	12	13	14	23	24	34	
1	GG,	GG,	G-,	GG,	G-,	G-	= 1+1-1+1-1-1 = 0
2	TT,	TA,	TT,	TA,	TT,	AT	= 0
3	GG,						= 0
4	CC,						= 6
5	--						= -3
6	CC						= 6
7	TT						= 0
8	A-						= -4+1 = -3
9	TT						= -3
10	GC						= 0
11	GG						= 6
12	--						= 0-5 = -5
(?) 13	--						= -3

e) The mouse is the organism that is most closely related to human. (according to gene x similarities) It had the highest similarity amongst the given fragments of gene x. Yes we could find a different result if we use a different gene than gene X.