BBM411: Fundamentals of Bioinformatics (Fall 2019)

Assignment 1

Due date: November 25, 2019, time: 23:59

Burak Emre Özer -> 21527266 \
(burakemreozer@gmail.com)

Question 1 (16 points)

Please carefully explain each biological phenomenon given below, in 2-3 sentences.

a) Define homology in terms of biomolecular sequence similarities.

Calculating the similarities between sequences (e.g. BLAST) increases our chances of making inferences about common ancestors, neighborhoods and evolution between these structures. Therefore, generally similar DNA, RNA or protein sequences can be called homologous. However, this does not necessarily mean that similar sequences are always homologous.

b) What is the biological relation between genes and proteins?

Genes encode proteins by specifying the sequence of amino acids, the basic building blocks of proteins. Thus, genes actually convert the encoded genetic information into functions through proteins.

c) What is gene expression? Explain one of the two steps of gene expression.

Gene expression is the process of building a gene product such as protein based on the genetic sequence data on the gene. **Transcription**, which is one of the two steps, can be explained as follows: Transferring the genetic information in the nucleus (DNA) to the RNA by mRNA.

d) Give one example way to extract biological data (a.k.a. transforming a biological sample into data) and briefly explain how it is done.

We can extract biological data using the **Sanger's Dideoxy Method**. This method can be summarized as follows: Fragments are generated by replicating DNA labeled with colors. Afterwards, the fragments are separated according to their size by electrophoresis. A machine scans fragments and records each color according to its position. Finally, the software predict the base for each position.

Question 2 (20 points)

a) Please write the correct regular expression below.

< [Q] - X - [KH] - X - E - [DE] - L - [PTSAG] - [LI] >

b) Search the sequence below to locate the hits to your regular expression. How many hits can you find? Give the positions of the hits on the given sequence.

	Expression	Index
1.	QRKVEDLSI	2-10
2.	QEHIEDLTL	13-21
3.	QDKIEELGI	39-47
4.	QDHVEELTL	51-59
5.	QEHIEELTL	129-137
6.	QEHCEELTL	160-168
7.	QEKIEDLSL	172-180
8.	QRHVEELGI	192-200
9.	QDKVEDLPI	205-213
10.	QRKIEDLTL	227-235

Question 3 (64 points)

a) Submit your script file via email, for testing.

The script file submitted from **burakemreozer@gmail.com**

b) Aligned the sequences of Gene A and Gene 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: match=5, mismatch=-4, 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.

Global Alignment	Local Alignment
ACTACTAGATTACTGACGGATAAGGTACTTTAGAGGCTTGCAACCA ||||| || ||||| |||||||||||| | | || ACTAC-----T-C--ACGGATCAGGTACTTTAGA-G---G----CA Percent identity: 63.04347826086956 Total Alignment Score: 46	ACTACTAGATTACTGACGGATAAGGTACTTTAGAGGC | |||| ||||| ||||||||||||||| --------ACTACTCACGGATCAGGTACTTTAGAGGC Percent identity: 89.65517241379311 Total Alignment Score: 118

Global alignment attempts to align sequences from end-to-end. Therefore, too many gaps were formed as a result of alignment. Due to the high gap penalty and the different lengths of the sequences, the total alignment score was 46. Then, the local alignment algorithm was used. Local alignment tries to align the given sequences between the subsequences and to find smaller motifs. As expected, local alignment achieved a higher success and found the total alignment score 118. Local alignment appears to be more appropriate if there are cases where the sequences are partially similar.

c) Play around with the parameter values for match, mismatch, gap open and gap extend 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.

I observed that the alignment score increased as a result of reducing the gap penalty while performing global alignment. It would almost reach the score of the local alignment algorithm. In general, I think that the fact that gap penalties are less in global alignment will create more effective alignments.

Algorithm :global Match score :5 Miss match :-4 Gap opening :-2 Gap Extension :-1 **Total Alignment Score: 122**	**Algorithm :local** Match score :5 Miss match :-4 Gap opening :-2 Gap Extension :-1 **Total Alignment Score: 126**

When I reduced the miss match penalty, I found that the local alignment algorithm tried to align the sequences end-to-end.

Algorithm :global Match score :5 Miss match :-1 Gap opening :-1 Gap Extension :-1 **Total Alignment Score: 128**	**Algorithm :local** Match score :5 Miss match :-1 Gap opening :-1 Gap Extension :-1 **Total Alignment Score: 132**

```
Algorithm       :local
Match score     :5
Miss match      :-1
Gap opening     :-1
Gap Extension   :-1

INFO:root:Matrix initialized...

INFO:root:Sequence alignment printing...

ACTACTAGATTACTGACGGATAAGGTACTTTAGAGGCTTGCA
|||||     | |  |||||| |||||||||||| |   |||
ACTAC-----T-C--ACGGATCAGGTACTTTAGA-G---GCA

Percent identity: 67.44186046511628
Total Alignment Score: 132
```