Alexandria University
Faculty of Engineering
Computer and Systems Engineering Department

Bioinformatics

Master of Computer Sc.

Fall 2020

Assignment 1

Yousef Mohamed Zook

Alexandria University
Faculty of Engineering
Computer and Systems Engineering Department

Problem 1. Global alignment

You are given two sequences AGATT and AGTT. Assume a **match score of 1**, a **gap penalty of 3,** and a **substitution score of -1**. Using these scores, obtain the global alignment of these two sequences in the following two steps:

(a) Fill in the entries of the F matrix by applying the recurrence relationship for global alignment to these sequences. Please show the back pointers to the matrix entry/entries that give you the maximal score for any entry.

Solution

	-	A	G	A	T	T
-	0	-3	-6	-9	-12	-15
A	-3	1	-2	-5	-8	-11
G	-6	-2	2	-1	-4	-7
T	-9	-5	-1	1	0	-3
T	-12	-8	-4	-2	2	1

Red + Green arrows, show the entry/entries that give the maximal score for each entry. Green arrows show the best alignment score path.

(b) Apply the traceback procedure to obtain an optimal alignment. If there are multiple possible alignments, please show all of them along with their traceback paths.

Solution

A G A T T
A G - T T

Alexandria University
Faculty of Engineering
Computer and Systems Engineering Department

Problem 2. Local alignment

Using the same scores above perform a local alignment for the sequences, GAAGAG and AAGC in the following two steps:

(a) Fill in the entries of the F matrix using the recurrence relation for the local alignment of these sequences. Show back pointers to matrix entry/entries that give you the maximal score.

Solution

	-	A	A	G	C
-	0	0	0	0	0
G	0	0	0	1	0
A	0	1	1	0	0
A	0	1	2	0	0
G	0	0	0	3	0
A	0	1	1	0	0
G	0	0	0	2	0

(b) Apply the traceback procedure to generate a local alignment.

Solution

- A A G C - / C - - / - C / - - C / - C -
G A A G AG / - A G / AG / AG- / A - G

Alexandria University
Faculty of Engineering
Computer and Systems Engineering Department

Problem 3. (a)Analyzing a DNA Sequence

Using high-throughput methods, scientists are now able to sequence entire genomes in a very short period of time. Sequencing a genome is quite an accomplishment in itself, but it is really only the beginning of the study of an organism. Further study can be done both at the wet lab bench and on the computer. In this problem, you will use a computer to help you identify an open reading frame, determine the protein that it will express, and find the bacterial source for that protein. Here is the DNA sequence:

TACGCAATGCGTATCATTCTGCTGGGCGCTCCGGGCGCAGGTAAAGGTACTCAGGCTCAATTCATCATGGAGAAAT
ACGGCATTCCGCAAATCTCTACTGGTGACATGTTGCGCGCCGCTGTAAAAGCAGGTTCTGAGTTAGGTCTGAAAGC
AAAAGAAATTATGGATGCGGGCAAGTTGGTGACTGATGAGTTAGTTATCGCATTACTCAAAGAACGTATCACACA
GGAAGATTGCCGCGATGGTTTTCTGTTAGACGGGTTCCCGCGTACCATTCCTCAGGCAGATGCCATGAAAGAAGCC
GGTATCAAAGTTGATTATGTGCTGGAGTTTGATGTTCCAGACGAGCTGATTGTTGAGCGCATTGTCGGCCGTCGGG
TACATGCTGCTTCAGGCCGTGTTTATCACGTTAAATTCAACCCACCTAAAGTTGAAGATAAAGATGATGTTACCGG
TGAAGAGCTGACTATTCGTAAAGATGATCAGGAAGCGACTGTCCGTAAGCGTCTTATCGAATATCATCAACAAACT
GCACCATTGGTTTCTTACTATCATAAAGAAGCGGATGCAGGTAATACGCAATATTTTAAACTGGACGGAACCCGTA
ATGTAGCAGAAGTCAGTGCTGAACTGGCGACTATTCTCGGTTAATTCTGGATGGCCTTATAGCTAAGGCGGTTTAA
GGCCGCCTTAGCTATTTCAAGTAAGAAGGGCGTAGTACCTACAAAAGGAGATTTGGCATGATGCAAAGCAAACCC
GGCGTATTAATGGTTAATTTGGGGACACCAGATGCTCCAACGTCGAAAGCTATCAAGCGTTATTTAGCTGAGTTTT
TGAGTGACCGCCGGGTAGTTGATACTTCCCCATTGCTATGGTGGCCATTGCTGCATGGTGTTATTTTACCGCTTCGG
TCACCACGTGTAGCAAAACTTTATCAATCCGTTTGGATGGAAGAGGGCTCTCCTTTATTGGTTTATAGCCGCCGCC
AGCAGAAAGCACTGGCAGCAAGAATGCCTGATATTCCTGTAGAATTAGGCATGAGCTATGGTTCAC

a. First, try to find an open reading frame in this segment of DNA. What is an open reading frame (ORF)? You can find the answer in your textbook or online with a simple Internet search (http://www.google.com).

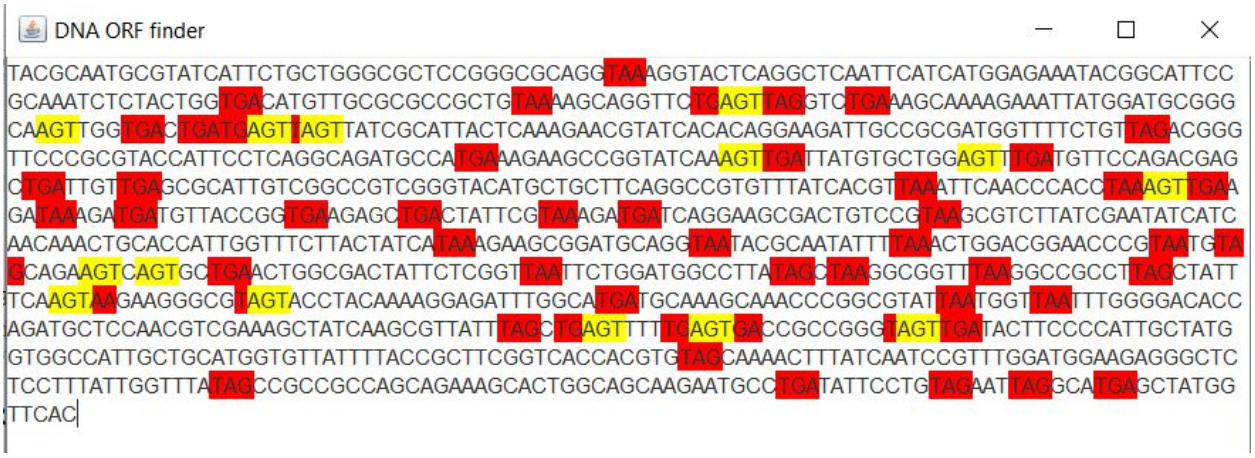
You may also wish to try the bookshelf at PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books). In bacteria, an open reading frame on a piece of mRNA almost always begins with AUG, which corresponds to ATG in the DNA segment that codes for the mRNA. According to the standard genetic code, there are three Stop codons on mRNA: UAA, UAG, and UGA, which correspond to TAA, TAG, and TGA in the parent DNA segment. Here are the rules for finding an open reading frame in this piece of bacterial DNA:

1. It must start with ATG. In this exercise, the first ATG is the Start codon. In a real gene search, you would not have this information.
2. It must end with TAA, TAG, or TGA.
3. It must be at least 300 nucleotides long (coding for 100 amino acids).
4. The ATG Start codon and the Stop codon must be in frame. This means that the total number of bases in the sequence from the Start to the Stop codon must be evenly divisible by 3.

Hints: Try this search by pasting the DNA sequence into a word processing program, then searching for the Start and Stop codons. Once you have found a pair, highlight the text of the

proposed ORF and use the program's Word Count function to count the number of characters between (or including) the Start and Stop codons. This number must be evenly divisible by 3. You can also use a fixed-width font such as Courier, enlarge the size of the text, and adjust the margins so that each line holds just three characters (one codon). Once you find the first ATG, delete the characters that precede it. Then search for a Stop codon that fits all on one line (is in the same reading frame as the Start codon).

Solution



b. Admittedly, Part (a) is a tedious approach. Here is an easier one: Highlight the entire DNA sequence again and copy it. Then go to the Translate tool on the ExPASy server (http://www.expasy.org/tools/dna.html). Paste the sequence into the box entitled “Please enter a DNA or RNA sequence in the box below (numbers and blanks are ignored).” Then select “Verbose (“Met”, “Stop”, spaces between residues)” as the Output format and click on “Translate Sequence.” The “Results of Translation” page that appears contains six different reading frames. What is a reading frame and why are there six? Identify the reading frame that contains a protein (more than 100 continuous amino acids with no interruptions by a Stop codon) and note its name. Now go back to the Translate tool page, leave the DNA sequence in the sequence box, but select “Compact (“M”, “-”, no spaces)” as the Output format. Go to the same reading frame as before and copy the protein sequence (by one-letter abbreviations) starting with “M” for methionine and ending in “-” for the Stop codon. Save this sequence to a separate text file.

Solution

using https://www.ncbi.nlm.nih.gov/orffinder/ instead:

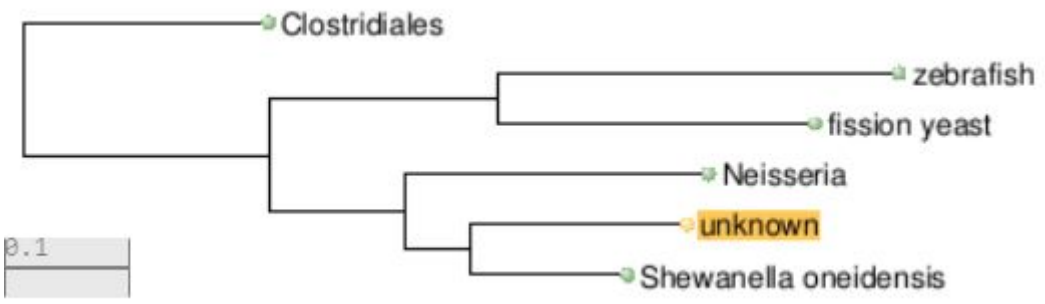
Mark subset... Marked: 0 Download marked set as Protein FASTA

Label	Strand	Frame	Start	Stop	Length (nt | aa)
ORF1	+	1	7	651	645 | 214
ORF2	**+**	**2**	**383**	**466**	**84 | 27**
ORF3	+	2	479	613	135 | 44
ORF4	+	2	872	979	108 | 35
ORF5	+	3	741	>1052	312 | 103
ORF6	-	1	555	391	165 | 54
ORF7	-	1	384	253	132 | 43
ORF8	-	2	275	57	219 | 72

c. Now you will identify the protein and the bacterial source. Go to the NCBI BLAST page (http://www.ncbi.nlm.nih.gov/BLAST/). What does BLAST stand for? You will do a simple BLAST search using your protein sequence, but you can do much more with BLAST. You are encouraged to work the Tutorials on the BLAST home page to learn more. On the BLAST page, select "Protein-protein BLAST." Enter your protein sequence in the "Search" box. Use the default values for the rest of the page and click on the "BLAST!" button. You will be taken to the "formatting BLAST" page. Click on the "Format!" button. You may have to wait for the results.

Solution

The most similar is: Shewanella oneidensis



Alexandria University
Faculty of Engineering
Computer and Systems Engineering Department

Problem 3. (b) Sequence Homology

I can’t find the BLAST platform steps mentioned in the pdf.