CAP5510 Fall 2019, HW#1

Question 1:

The given two sequences are:

x = TTCATA

y = TGCTCGTA

Given Scoring Function:

w(match)=+5

w(mismatch)=-2

w(indel)=-6

Here, length of x and y are consecutively 6 and 8. Let's consider a 2D matrix S with 7 rows and 9 columns for calculating the alignment scores. Also, we need to consider another 2D matrix PTR of the same size which will be used for tracing back the aligned sequence. If match then $c(x_i,y_j)=+5$, else $c(x_i,y_j)=-2$.

Initialization of S:

6	**A**	-36								
5	**T**	-30								
4	**A**	-24								
3	**C**	-18								
2	**T**	-12								
1	**T**	-6								
0		0	-6	-12	-18	-24	-30	-36	-42	-48
i			**T**	**G**	**C**	**T**	**C**	**G**	**T**	**A**
	j	0	1	2	3	4	5	6	7	8

Calculation of S Matrix:

1ST Column:

$$\mathbf{S(1,1)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(0,0) + c(T,T)} = \mathbf{0 + 5} = +\mathbf{5} \\ S(0,1) + c(T,-) = -6 - 6 = -12 \\ S(1,0) + c(-,T) = -6 - 6 = -12 \end{cases}$$

$$\mathbf{S(2,1)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(1,0) + c(T,T)} = \mathbf{-6 + 5} = -\mathbf{1} \\ S(1,1) + c(T,-) = +5 - 6 = -1 \\ S(2,0) + c(-,T) = -12 - 6 = -18 \end{cases}$$

$$\mathbf{S(3,1)} = \boldsymbol{max}\begin{cases} S(2,0) + c(C,T) = -12 - 2 = -14 \\ \boldsymbol{S(2,1) + c(C,-)} = \mathbf{-1 - 6} = -\mathbf{7} \\ S(3,0) + c(-,T) = -18 - 6 = -24 \end{cases}$$

$$\mathbf{S(4,1)} = \boldsymbol{max}\begin{cases} S(3,0) + c(A,T) = -18 - 2 = -20 \\ \boldsymbol{S(3,1) + c(A,-)} = \mathbf{-7 - 6} = -\mathbf{13} \\ S(4,0) + c(-,T) = -24 - 6 = -30 \end{cases}$$

$$\mathbf{S(5,1)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(4,0) + c(T,T)} = \mathbf{-24 + 5} = -\mathbf{19} \\ S(4,1) + c(T,-) = -13 - 6 = -19 \\ S(5,0) + c(-,T) = -30 - 6 = -36 \end{cases}$$

$$\mathbf{S(6,1)} = \boldsymbol{max}\begin{cases} S(5,0) + c(A,T) = -30 - 2 = -32 \\ \boldsymbol{S(5,1) + c(A,-)} = \mathbf{-19 - 6} = -\mathbf{25} \\ S(6,0) + c(-,T) = -36 - 6 = -42 \end{cases}$$

2nd Column:

$$\mathbf{S(1,2)} = \boldsymbol{max}\begin{cases} S(0,1) + c(T,G) = -6 - 2 = -8 \\ S(0,2) + c(T,-) = -12 - 6 = -18 \\ \boldsymbol{S(1,1) + c(-,G)} = \mathbf{+5 - 6} = -\mathbf{1} \end{cases}$$

$$\mathbf{S(2,2)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(1,1) + c(T,G)} = \mathbf{+5 - 2} = +\mathbf{3} \\ S(1,2) + c(T,-) = -1 - 6 = -7 \\ S(2,1) + c(-,G) = -1 - 6 = -7 \end{cases}$$

$$\mathbf{S(3,2)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(2,1) + c(C,G)} = \mathbf{-1 - 2} = -\mathbf{3} \\ S(2,2) + c(C,-) = +3 - 6 = -3 \\ S(3,1) + c(-,G) = -7 - 6 = -13 \end{cases}$$

$$\mathbf{S(4,2)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(3,1) + c(A,G)} = \mathbf{-7 - 2} = -\mathbf{9} \\ S(3,2) + c(A,-) = -3 - 6 = -9 \\ S(4,1) + c(-,G) = -13 - 6 = -19 \end{cases}$$

$$\mathbf{S(5,2)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(4,1) + c(T,G) = -13 - 2 - 15} \\ S(4,2) + c(T,-) = -9 - 6 = -15 \\ S(5,1) + c(-,G) = -19 - 6 = -25 \end{cases}$$

$$\mathbf{S(6,2)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(5,1) + c(A,G) = -19 - 2 = -21} \\ S(5,2) + c(A,-) = -15 - 6 = -21 \\ S(6,1) + c(-,G) = -25 - 6 = -31 \end{cases}$$

3rd Column:

$$\mathbf{S(1,3)} = \boldsymbol{max}\begin{cases} S(0,2) + c(T,C) = -12 - 2 = -14 \\ S(0,3) + c(T,-) = -18 - 6 = -24 \\ \boldsymbol{S(1,2) + c(-,C) = -1 - 6 = -7} \end{cases}$$

$$\mathbf{S(2,3)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(1,2) + c(T,C) = -1 - 2 = -3} \\ S(1,3) + c(T,-) = -7 - 6 = -13 \\ S(2,2) + c(-,C) = +3 - 6 = -3 \end{cases}$$

$$\mathbf{S(3,3)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(2,2) + c(C,C) = +3 + 5 = +8} \\ S(2,3) + c(C,-) = -3 - 6 = -9 \\ S(3,2) + c(-,C) = -3 - 6 = -9 \end{cases}$$

$$\mathbf{S(4,3)} = \boldsymbol{max}\begin{cases} S(3,2) + c(A,C) = -3 - 2 = -5 \\ \boldsymbol{S(3,3) + c(A,-) = +8 - 6 = +2} \\ S(4,2) + c(-,C) = -9 - 6 = -15 \end{cases}$$

$$\mathbf{S(5,3)} = \boldsymbol{max}\begin{cases} S(4,2) + c(T,C) = -9 - 2 = -11 \\ \boldsymbol{S(4,3) + c(T,-) = +2 - 6 = -4} \\ S(5,2) + c(-,C) = -15 - 6 = -21 \end{cases}$$

$$\mathbf{S(6,3)} = \boldsymbol{max}\begin{cases} S(5,2) + c(A,C) = -15 - 2 = -17 \\ \boldsymbol{S(5,3) + c(A,-) = -4 - 6 = -10} \\ S(6,2) + c(-,C) = -21 - 6 = -27 \end{cases}$$

4th Column:

$$\mathbf{S(1,4)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(0,3) + c(T,T) = -18 + 5 = -13} \\ S(0,4) + c(T,-) = -24 - 6 = -30 \\ S(1,3) + c(-,T) = -7 - 6 = -13 \end{cases}$$

$$\mathbf{S(2,4)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(1,3) + c(T,T) = -7 + 5 = -2} \\ S(1,4) + c(T,-) = -13 - 6 = -19 \\ S(2,3) + c(-,T) = -3 - 6 = -9 \end{cases}$$

$$\mathbf{S(3,4)} = \boldsymbol{max}\begin{cases} S(2,3) + c(C,T) = -3 - 2 = -5 \\ S(2,4) + c(C,-) = -2 - 6 = -8 \\ \boldsymbol{S(3,3) + c(-,T) = +8 - 6 = +2} \end{cases}$$

$$\mathbf{S(4,4)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(3,3) + c(A,T) = +8 - 2 = +6} \\ S(3,4) + c(A,-) = +2 - 6 = -4 \\ S(4,3) + c(-,T) = +2 - 6 = -4 \end{cases}$$

$$\mathbf{S(5,4)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(4,3) + c(T,T) = +2 + 5 = +7} \\ S(4,4) + c(T,-) = +6 - 6 = 0 \\ S(5,3) + c(-,T) = -4 - 6 = -10 \end{cases}$$

$$\mathbf{S(6,4)} = \boldsymbol{max}\begin{cases} S(5,3) + c(A,T) = -4 - 2 = -6 \\ \boldsymbol{S(5,4) + c(A,-) = +7 - 6 = +1} \\ S(6,3) + c(-,T) = -10 - 6 = -16 \end{cases}$$

<u>5th Column:</u>

$$\mathbf{S(1,5)} = \boldsymbol{max}\begin{cases} S(0,4) + c(T,C) = -24 - 2 = -26 \\ S(0,5) + c(T,-) = -30 - 6 = -36 \\ \boldsymbol{S(1,4) + c(-,C) = -13 - 6 = -19} \end{cases}$$

$$\mathbf{S(2,5)} = \boldsymbol{max}\begin{cases} S(1,4) + c(T,C) = -13 - 2 = -15 \\ S(1,5) + c(T,-) = -19 - 6 = -25 \\ \boldsymbol{S(2,4) + c(-,C) = -2 - 6 = -8} \end{cases}$$

$$\mathbf{S(3,5)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(2,4) + c(C,C) = -2 + 5 = +3} \\ S(2,5) + c(C,-) = -8 - 6 = -14 \\ S(3,4) + c(-,C) = +2 - 6 = -4 \end{cases}$$

$$\mathbf{S(4,5)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(3,4) + c(A,C) = +2 - 2 = 0} \\ S(3,5) + c(A,-) = +3 - 6 = -3 \\ S(4,4) + c(-,C) = +6 - 6 = 0 \end{cases}$$

$$\mathbf{S(5,5)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(4,4) + c(T,C) = +6 - 2 = +4} \\ S(4,5) + c(T,-) = 0 - 6 = -6 \\ S(5,4) + c(-,C) = +7 - 6 = +1 \end{cases}$$

$$\mathbf{S(6,5)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(5,4) + c(A,C) = +7 - 2 = +5} \\ S(5,5) + c(A,-) = +4 - 6 = -2 \\ S(6,4) + c(-,C) = +1 - 6 = -5 \end{cases}$$

6th Column:

$$\mathbf{S(1,6)} = \boldsymbol{max}\begin{cases} S(0,5) + c(T,G) = -30 - 2 = -32 \\ S(0,6) + c(T,-) = -36 - 6 = -42 \\ \boldsymbol{S(1,5) + c(-,G) = -19 - 6 = -25} \end{cases}$$

$$\mathbf{S(2,6)} = \boldsymbol{max}\begin{cases} S(1,5) + c(T,G) = -19 - 2 = -21 \\ S(1,6) + c(T,-) = -25 - 6 = -31 \\ \boldsymbol{S(2,5) + c(-,G) = -8 - 6 = -14} \end{cases}$$

$$\mathbf{S(3,6)} = \boldsymbol{max}\begin{cases} S(2,5) + c(C,G) = -8 - 2 = -10 \\ S(2,6) + c(C,-) = -14 - 6 = -20 \\ \boldsymbol{S(3,5) + c(-,C) = +3 - 6 = -3} \end{cases}$$

$$\mathbf{S(4,6)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(3,5) + c(A,G) = +3 - 2 = +1} \\ S(3,6) + c(A,-) = -3 - 6 = -9 \\ S(4,5) + c(-,G) = 0 - 6 = -6 \end{cases}$$

$$\mathbf{S(5,6)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(4,5) + c(T,G) = 0 - 2 = -2} \\ S(4,6) + c(T,-) = +1 - 6 = -5 \\ S(5,5) + c(-,G) = +4 - 6 = -2 \end{cases}$$

$$\mathbf{S(6,6)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(5,5) + c(A,G) = +4 - 2 = +2} \\ S(5,6) + c(A,-) = -2 - 6 = -8 \\ S(6,5) + c(-,G) = +5 - 6 = -1 \end{cases}$$

7th Column:

$$\mathbf{S(1,7)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(0,6) + c(T,T) = -36 + 5 = -31} \\ S(0,7) + c(T,-) = -42 - 6 = -48 \\ S(1,6) + c(-,T) = -25 - 6 = -31 \end{cases}$$

$$\mathbf{S(2,7)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(1,6) + c(T,T) = -25 + 5 = -20} \\ S(1,7) + c(T,-) = -31 - 6 = -37 \\ S(2,6) + c(-,G) = -14 - 6 = -20 \end{cases}$$

$$\mathbf{S(3,7)} = \boldsymbol{max}\begin{cases} S(2,6) + c(C,T) = -14 - 2 = & -16 \\ S(2,7) + c(C,-) = -20 - 6 = & -26 \\ \boldsymbol{S(3,6) + c(-,T) = -3 - 6 =} & \boldsymbol{-9} \end{cases}$$

$$\mathbf{S(4,7)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(3,6) + c(A,T) = -3 - 2 =} & \boldsymbol{-5} \\ S(3,7) + c(A,-) = -9 - 6 = & -15 \\ S(4,6) + c(-,T) = 1 - 6 = & -5 \end{cases}$$

$$\mathbf{S(5,7)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(4,6) + c(T,T) = +1 + 5 =} & \boldsymbol{+6} \\ S(4,7) + c(T,-) = -5 - 6 = & -11 \\ S(5,6) + c(-,T) = -2 - 6 = & -8 \end{cases}$$

$$\mathbf{S(6,7)} = \boldsymbol{max}\begin{cases} S(5,6) + c(A,T) = -2 - 2 = & -4 \\ \boldsymbol{S(5,7) + c(A,-) = +6 - 6 =} & \boldsymbol{0} \\ S(6,6) + c(-,T) = +2 - 6 = & -4 \end{cases}$$

8th Column:

$$\mathbf{S(1,8)} = \boldsymbol{max}\begin{cases} S(0,7) + c(T,A) = -42 - 2 = & -44 \\ S(0,8) + c(T,-) = -48 - 6 = & -54 \\ \boldsymbol{S(1,7) + c(-,A) = -31 - 6 =} & \boldsymbol{-37} \end{cases}$$

$$\mathbf{S(2,8)} = \boldsymbol{max}\begin{cases} S(1,7) + c(T,A) = -31 - 2 = & -33 \\ S(1,8) + c(T,-) = -37 - 6 = & -43 \\ \boldsymbol{S(2,7) + c(-,A) = -20 - 6 =} & \boldsymbol{-26} \end{cases}$$

$$\mathbf{S(3,8)} = \boldsymbol{max}\begin{cases} S(2,7) + c(C,A) = -20 - 2 = & -22 \\ S(2,8) + c(C,-) = -26 - 6 = & -32 \\ \boldsymbol{S(3,7) + c(-,A) = -9 - 6 =} & \boldsymbol{-15} \end{cases}$$

$$\mathbf{S(4,8)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(3,7) + c(A,A) = -9 + 5 =} & \boldsymbol{-4} \\ S(3,8) + c(A,-) = -15 - 6 = & -21 \\ S(4,7) + c(-,A) = -5 - 6 = & -11 \end{cases}$$

$$\mathbf{S(5,8)} = \boldsymbol{max}\begin{cases} S(4,7) + c(T,A) = -5 - 2 = & -7 \\ S(4,8) + c(T,-) = -4 - 6 = & -10 \\ \boldsymbol{S(5,7) + c(-,A) = +6 - 6 =} & \boldsymbol{0} \end{cases}$$

$$\mathbf{S(6,8)} = \boldsymbol{max}\begin{cases} \boldsymbol{S(5,7) + c(A,A) = +6 + 5 =} & \boldsymbol{+11} \\ S(5,8) + c(A,-) = 0 - 6 = & -6 \\ S(6,7) + c(-,A) = 0 - 6 = & -6 \end{cases}$$

6	**A**	-36	-25	-21	-10	1	5	2	0	11
5	**T**	-30	-19	-15	-4	7	4	-2	6	0
4	**A**	-24	-13	-9	2	6	0	1	-5	-4
3	**C**	-18	-7	-3	8	2	3	-3	-9	-15
2	**T**	-12	-1	3	-3	-2	-8	-14	-20	-26
1	**T**	-6	5	-1	-7	-13	-19	-25	-31	-37
0		0	-6	-12	-18	-24	-30	-36	-42	-48
i			**T**	**G**	**C**	**T**	**C**	**G**	**T**	**A**
	j	0	1	2	3	4	5	6	7	8

PTR matrix for tracing back:

6	**A**	-36	D	Di	D	D	Di	Di	D	Di
5	**T**	-30	Di	Di	D	Di	Di	Di	Di	L
4	**A**	-24	D	Di	D	Di	Di	Di	Di	Di
3	**C**	-18	D	Di	Di	L	Di	L	L	L
2	**T**	-12	Di	Di	Di	Di	L	L	Di	L
1	**T**	-6	Di	L	L	Di	L	L	Di	L
0		0	-6	-12	-18	-24	-30	-36	-42	-48
i			**T**	**G**	**C**	**T**	**C**	**G**	**T**	**A**
	j	0	1	2	3	4	5	6	7	8

Trace Back Path:

6	A	-36	-25	-21	-10	1	5	2	0	11
5	T	-30	-19	-15	-4	7	4	-2	6	0
4	A	-24	-13	-9	2	6	0	1	-5	-4
3	C	-18	-7	-3	8	2	3	-3	-9	-15
2	T	-12	-1	3	-3	-2	-8	-14	-20	-26
1	T	-6	5	-1	-7	-13	-19	-25	-31	-37
0		0	-6	-12	-18	-24	-30	-36	-42	-48
i			T	G	C	T	C	G	T	A
	j	0	1	2	3	4	5	6	7	8

So, Final Global Alignment:

T	-	-	T	C	A	T	A
T	G	C	T	C	G	T	A

Question 2:

i) Using Protein BLAST:

Query Sequence: pannexin 1 [Bos taurus]
Accession: ADH10263.1
Link: https://www.ncbi.nlm.nih.gov/protein/ADH10263.1
Search Domain: Human (taxid : 9606)
Search results: 15 sequences selected

Description:

Among the 15 selected sequences, the Expect Value (E-value) of first 6 sequences (Accession AAH16931.1, NP_056183.2, AAC61779.1, CAR31475.1, AAK91713.1, AAK73361.1) is 0.0 which means they are more related to the query sequence. Expect value parameter points towards the possible numbers of random sequences with similar score. Lower E-value represents more significant match. Among those six, Pannexin 1 [Homo sapiens] (Accession AAH16931.1) is the most identical one. It has 100% query cover which means the whole query sequence is included in the aligned segments. Also it has the highest maximum score and total score (both 679). Though score depends on the length of the sequence (two not so similar sequences can have high max score due to long length), but here the length of the sequences are quite similar (422 aa to 426 aa for the first five and 357 aa for the 6^{th}). So, Pannexin 1 [Homo sapiens] (Accession AAH16931.1) can be considered to be the most related one to the given query sequence. The other 9 sequences also have Expect-value < 1e-04, so all of them are considered homologues to the query sequence with an error rate of less than 0.01%.

Alignments:

For pannexin-1 [Homo sapiens] (both Sequence ID: AAH16931.1 and NP_056183.2), length of aligned sequence is 427 with a Identities of 82%, Positivists of 89% (7% similar proteins) and they have only 4 Gaps (0%). The other four sequences (Accession AAC61779.1, CAR31475.1, AAK91713.1, AAK73361.1) with E-value 0 also have high percentage of Identities, Positives and very low percentage of Gaps. This means protein sequence Pannexin 1 of cattle and human is highly related. Also, 9 other protein sequeunces of Homo Sapiens (pannexin-1 isoform X1, pannexin-3, pannexin-1, pannexin-2 isoform 2, pannexin 2 isoform1, pannexin 2 isoform CRA_a, pannexin-2 isoform 1, dJ402G11.9 of chromosome 22, pannexin 2 isoform 2) have low E-value means they are also considered to be related to the given query sequence.

ii) Using Nucleotide BLAST:

Query Sequence: Human 2-5A-dependent RNase gene, complete cds
Accession: L10381
Link: https://www.ncbi.nlm.nih.gov/nuccore/L10381.1
Search Domain: Mouse genomic + transcript
Search results: 9 sequences selected

Description:

In total 9 sequences (having significant alignment to the query sequence) are selected by the Nucleotide BLAST program. If the selected sequences are sorted according to E-value then

selected top 8 sequences are of ‘Mus musculus ribonuclease L (2', 5'-oligoisoadenylate synthetase-dependent) Rnasel mRNA (different transcript variants). All of them had E-value of 0.0 with max score of 974 (total score is same as max score) and percent identity of 75.56%. Also, 70% of the query sequence is covered in the aligned sequence. The 9th sequence selected is also ‘Mus musculus ribonuclease L (2', 5'-oligoisoadenylate synthetase-dependent) Rnasel , transcript variant X8 mRNA with an E-value of 0.0. But the max score, total score and query cover is a bit less for this one compared to other ones. The 0.0 E-value and all these other properties of the selected sequences indicate that the RNase RNA found in human body is highly identical to ribonuclease L Rnasel of mouse.

Alignment:

From the graphical representation, it can be seen that the alignment score is pretty high (>=200) within the range 189-2253 of the query sequence for the first 8 selected sequences (for the 9th sequence, the range is 189-2133). The identities of all the selected sequences are same (76%). The first 8 sequences have only 3% gap in the aligned portion while the 9th sequence have a gap of 2%. Still the E-value is higher for the first 8 sequences as for the 9th sequence, the length of the aligned sequence is less (length 1972) compared to the first 8 sequences (length 2095).