Mustafa Ozan Alpay - 2309615

CENG 465
Spring 2020-2021

Due Date: April 14, Wedneday 23:59 (via ODTU-Class)

Assignment #1

Problem 1 (60 Points): Fill in the dynamic programming score table using the Needleman-Wunsch algorithm to globally align the two sequences given below. Use the BLOSUM62 matrix given below as the scoring matrix. Use the linear gap model with a gap penalty of -4. Show the best alignment of these two sequences. Also, show the alignment path on the partial scores table.

	-	**M**	**C**	**G**	**M**	**G**	**C**	**M**	**E**	**L**
-	**0**	**-4**	**-8**	**-12**	**-16**	**-20**	**-24**	**-28**	**-32**	**-36**
G	**-4**	-3	-7	-2	-6	-10	-14	-18	-22	-26
M	**-8**	1	-3	-6	3	-1	-5	-9	-13	-17
C	**-12**	-3	10	6	2	0	8	4	0	-4
M	**-16**	-7	6	7	11	7	4	13	9	5
E	**-20**	-11	2	4	7	9	5	9	18	14
D	**-24**	-15	-2	1	3	6	6	5	14	14
K	**-28**	-19	-6	-3	0	2	3	5	10	12

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	**9**	-1	-1	-3	0	-3	-3	-3	-4	-3	-3	-3	-3	-1	-1	-1	-1	-2	-2	-2
S	-1	**4**	1	-1	1	0	1	0	0	0	-1	-1	0	-1	-2	-2	-2	-2	-2	-3
T	-1	1	**4**	1	-1	1	0	1	0	0	0	-1	0	-1	-2	-2	-2	-2	-2	-3
P	-3	-1	1	**7**	-1	-2	-1	-1	-1	-1	-2	-2	-1	-2	-3	-3	-2	-4	-3	-4
A	0	1	-1	-1	**4**	0	-1	-2	-1	-1	-2	-1	-1	-1	-1	-1	-2	-2	-2	-3
G	-3	0	1	-2	0	**6**	-2	-1	-2	-2	-2	-2	-2	-3	-4	-4	0	-3	-3	-2
N	-3	1	0	-2	-2	0	**6**	1	0	0	-1	0	0	-2	-3	-3	-3	-3	-2	-4
D	-3	0	1	-1	-2	-1	1	**6**	2	0	-1	-2	-1	-3	-3	-4	-3	-3	-3	-4
E	-4	0	0	-1	-1	-2	0	2	**5**	2	0	0	1	-2	-3	-3	-3	-3	-2	-3
Q	-3	0	0	-1	-1	-2	0	0	2	**5**	0	1	1	0	-3	-2	-2	-3	-1	-2
H	-3	-1	0	-2	-2	-2	1	1	0	0	**8**	0	-1	-2	-3	-3	-2	-1	2	-2
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	**5**	2	-1	-3	-2	-3	-3	-2	-3
K	-3	0	0	-1	-1	-2	0	-1	1	1	-1	2	**5**	-1	-3	-2	-3	-3	-2	-3
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	**5**	1	2	-2	0	-1	-1
I	-1	-2	-2	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	**4**	2	1	0	-1	-3
L	-1	-2	-2	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	**4**	3	0	-1	-2
V	-1	-2	-2	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	**4**	-1	-1	-3
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	**6**	3	1
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	**7**	2
W	-2	-3	-3	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	**11**

BLOSUM62 Substitution Matrix

Show the alignment path on the partial scores table above. Show the alignment below:

Best alignment:
```
- - G M - C M E D K
M C G M G C M E - L
```

	-	M	C	G	M	G	C	M	E	L
-	0	-4	-8	-12	-16	-20	-24	-28	-32	-36
G	-4	-3	-7	-2	-6	-10	-14	-18	-22	-26
M	-8	1	-3	-6	3	-1	-5	-9	-13	-17
C	-12	-3	10	6	2	0	8	4	0	-4
M	-16	-7	6	7	11	7	4	13	9	5
E	-20	-11	2	4	7	9	5	9	18	14
D	-24	-15	-2	1	3	6	6	5	14	14
K	-28	-19	-6	-3	0	2	3	5	10	12

Problem 2 (40 Points): What is the maximum length of the DNA sequence that is expected to occur at least once in its entirety in another DNA sequence of length 1000? In other words given two DNA sequences, ***A*** and ***B***, where *length*(***A***)=1000, what should be the maximum length of ***B***, so that it is expected to observe a perfect local alignment (all matches, no mismatch or gap) between ***A*** and ***B***? Show the steps of your calculation.

To find the possibility of a match, we can use the following formula:

$$E(x) = p^{length} (m - length + 1) (n - length + 1)$$

where

- `p` is the possibility for the sequence (1/4 for DNA, 1/20 for proteins)
- `m` is the length of the first sequence
- `n` is the length of the second sequence
- `length` is the expected number of consequent matches

Since we are interested in finding the length of B such that it matches perfectly (i.e. the E(x) value is 1), the length of B should be equal to the expected number of consequent matches, therefore `length = n` must hold. To simplify the calculations, we can omit the “ `- length + 1`” parts from the formula. By doing that and plugging in the values, we get

$$E(x) = (0.25)^{b} * 1000 * b = 1$$

If we solve the equation above, we would get `b = 6.31193` and `b = 0.001` Since we are interested in maximum length of the DNA sequence, with having a DNA sequence with length 6 we are expected to observe a perfect local alignment without any mismatches or gaps.