Science-2 Assignment 3

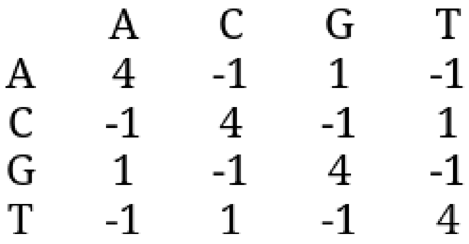
Question 1

You are given 2 nucleotide sequences:

`GGCTGCAACTAGCTC`

`GGGTAAGCTTGC`

Given gap penalty: -3 and the transition-transversion scoring matrix (expressed in similarity):



	A	C	G	T
A	4	-1	1	-1
C	-1	4	-1	1
G	1	-1	4	-1
T	-1	1	-1	4

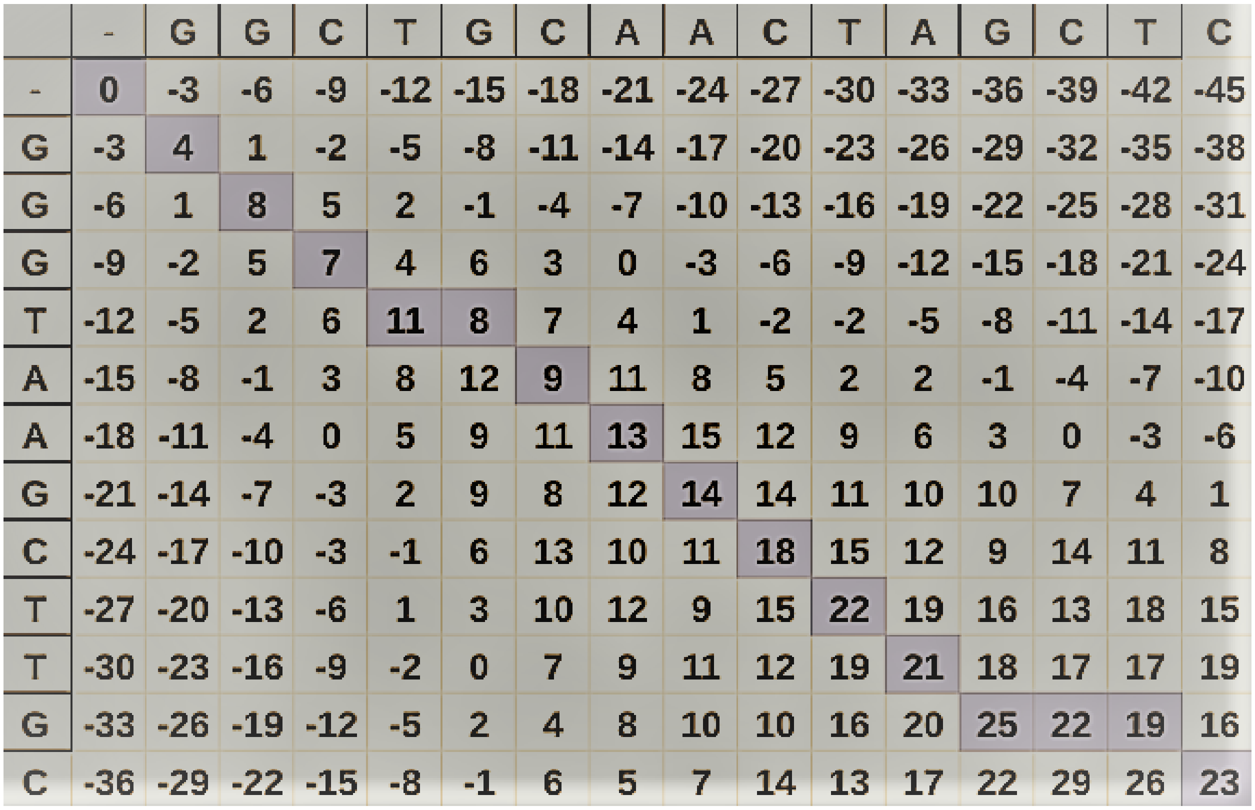
Carry out the global and local alignment (dynamic programming algorithm), and indicate the final similarity score and the best alignment.

→

1. **Global alignment**
    a. Alignment score: 23
    b. Best Alignment: →
    c. Algorithm Used: Needleman-Wunsch

```
GGCTGCAACTAGCTC
GGGT-AAGCTTG--C
```

d. DP Table:

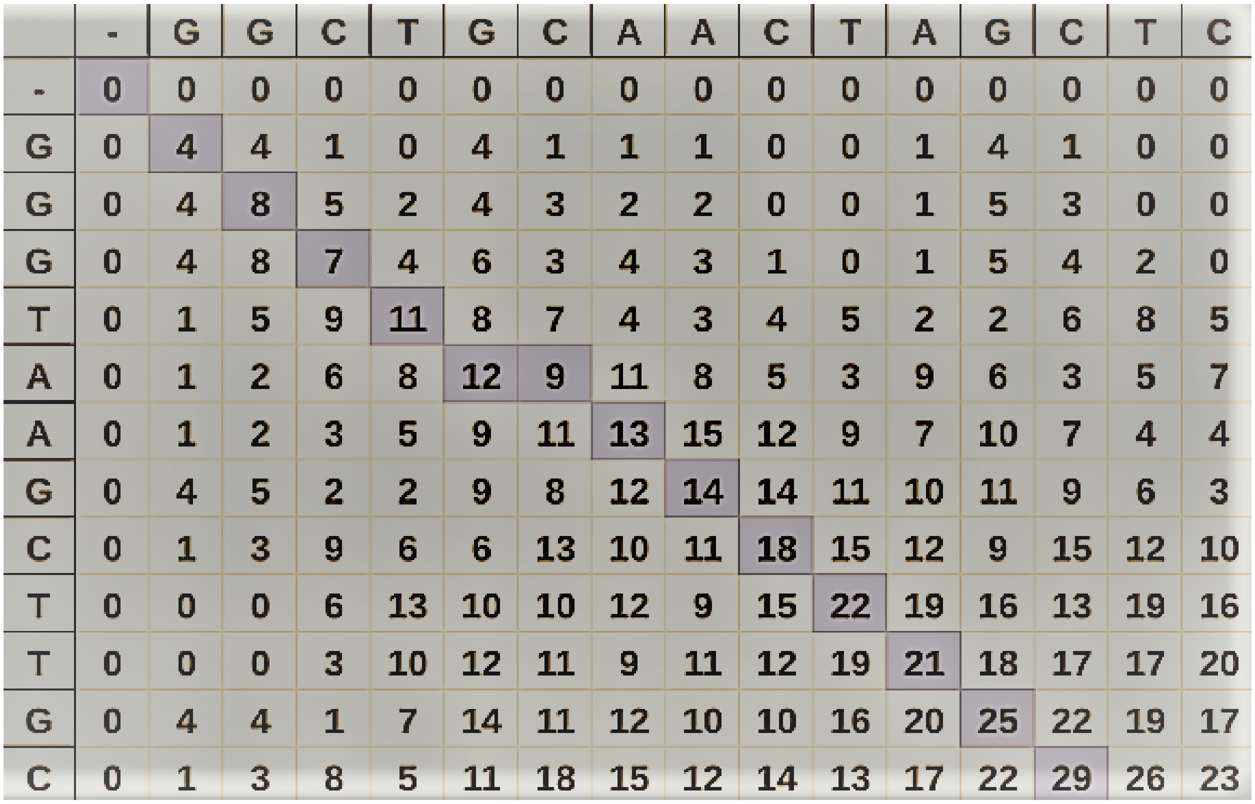


	-	G	G	C	T	G	C	A	A	C	T	A	G	C	T	C
-	0	-3	-6	-9	-12	-15	-18	-21	-24	-27	-30	-33	-36	-39	-42	-45
G	-3	4	1	-2	-5	-8	-11	-14	-17	-20	-23	-26	-29	-32	-35	-38
G	-6	1	8	5	2	-1	-4	-7	-10	-13	-16	-19	-22	-25	-28	-31
G	-9	-2	5	7	4	6	3	0	-3	-6	-9	-12	-15	-18	-21	-24
T	-12	-5	2	6	11	8	7	4	1	-2	-2	-5	-8	-11	-14	-17
A	-15	-8	-1	3	8	12	9	11	8	5	2	2	-1	-4	-7	-10
A	-18	-11	-4	0	5	9	11	13	15	12	9	6	3	0	-3	-6
G	-21	-14	-7	-3	2	9	8	12	14	14	11	10	10	7	4	1
C	-24	-17	-10	-3	-1	6	13	10	11	18	15	12	9	14	11	8
T	-27	-20	-13	-6	1	3	10	12	9	15	22	19	16	13	18	15
T	-30	-23	-16	-9	-2	0	7	9	11	12	19	21	18	17	17	19
G	-33	-26	-19	-12	-5	2	4	8	10	10	16	20	25	22	19	16
C	-36	-29	-22	-15	-8	-1	6	5	7	14	13	17	22	29	26	23

2. **Local alignment**
    a. Alignment Score: 29
    b. Best Alignment: →
    c. Algorithm Used: Smith-Waterman

```
GGCTGCAACTAGCTC
GGGTA-AGCTTGC-
```

d. DP Table:



	-	G	G	C	T	G	C	A	A	C	T	A	G	C	T	C
-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
G	0	4	4	1	0	4	1	1	1	0	0	1	4	1	0	0
G	0	4	8	5	2	4	3	2	2	0	0	1	5	3	0	0
G	0	4	8	7	4	6	3	4	3	1	0	1	5	4	2	0
T	0	1	5	9	11	8	7	4	3	4	5	2	2	6	8	5
A	0	1	2	6	8	12	9	11	8	5	3	9	6	3	5	7
A	0	1	2	3	5	9	11	13	15	12	9	7	10	7	4	4
G	0	4	5	2	2	9	8	12	14	14	11	10	11	9	6	3
C	0	1	3	9	6	6	13	10	11	18	15	12	9	15	12	10
T	0	0	0	6	13	10	10	12	9	15	22	19	16	13	19	16
T	0	0	0	3	10	12	11	9	11	12	19	21	18	17	17	20
G	0	4	4	1	7	14	11	12	10	10	16	20	25	22	19	17
C	0	1	3	8	5	11	18	15	12	14	13	17	22	29	26	23

Question 2

Identify the dinucleotide CA repeat region and the score in the following sequence: `TGGCACACTCACACCACACAGACAGTTA`

→

a. The dinucleotide `CA` tandem repeats from position 10 and the length of this sequence is 10.

b. The region is `CACACCACAC`.

c. This gives us a total score of **20**.

Question 3

When would you encounter a situation for using DP for overlap regions? How are the boundary conditions and recursive relations different from that for global alignment?

→

a. Consider situations where there are overlapping regions in our two sequences or one sequence is the sub-sequence of another, in such situations we use DP.

b. The recursive relation for the DP is same as that for global alignment but the boundary conditions are different.

$$\text{Recurence relation: } F(i,j) = max \begin{cases} F(i-1,j-1) + s(i,j) \\ F(i-1,j) - d \\ F(i,j-1) - d \end{cases}$$
$$\text{Boundary conditions}$$
$$F(0,0) = 0$$
$$F(i,0) = 0$$
$$F(0,j) = 0$$

Question 4

What is the advantage of using affine gap scores?

→ An affine score suggests that the deletions/insertions are the consequence of a single mutation rather than numerous mutations. As a result, large gaps are penalised less. It does so because a large shift may occur as a result of a single mutation that replaced a large section. We can't tell how many mutations occurred to modify the sequence in the gap if we utilised several gap penalties. As a result, we can argue that employing affine gap scores improves the sensitivity of our matching and allows us to find more homologs of a sequence. Further benefit is that during the algorithm's traceback, we simply have to verify if a new gap is forming or not, rather than searching through all of the gap lengths.

Question 5

Give the time and space complexity of DP. Under what conditions is time an issue and under what conditions would space be a problem?

→ Given two sequences of length n and m, to solve sequence alignment of the two given sequences, the DP algorithm has the following complexities.

$$\text{Time complexity: } O(nm)$$
$$\text{Space complexity: } O(nm)$$

Time complexity: Consider sequences of length $0.1 - 1\ Mbp$, time complexity might be a concern since each pair comparison would take a long time and would be impossible to accomplish in a database.

Space complexity: Consider sequences of length $\geq 1\ Mbp$, space complexity might be a concern because it would need roughly 10^{12} bytes in memory, which is nearly

$1\ TB$ of RAM.

Question 6

Describe the construction of Nucleic acid PAM scoring matrices.

→ PAM or **P**oint/Percent **A**ccepted **M**utations indicate the degree of sensitivity between two sequences, which varies depending on how far apart they are evolutionarily.

$$S(a,b \mid t) = \log(\frac{P(a \mid b,t)}{q_a q_b})$$

Here, $P(a \mid b,t)$ is the probability of an amino acid b being replaced by an amino acid a in time t and q_a, q_b denotes the frequency with which they occur, respectively.

The matrix is normalised by selecting a constant with an average change of around 1% across all places. We now have the PAM-1 matrix. By extrapolating from PAM-1, we can now utilise this matrix to create a PAM matrix for any distance.

Question 7

Take any gene sequence and its corresponding protein sequence and perform databases searches with both these sequences. Which of these two searches identifies more significant matches? Give reasons.

→ The protein sequence matches are more significant than matches generated by a gene sequence despite the fact that the gene sequence gives more matches. This is because the protein sequence matches are much more accurate. The following reasons are why:

1. A gene sequence has only four bases, but a protein sequence contains twenty amino acids. As a consequence, there's a higher chance of two unrelated genomic sequences matching, but a lower chance of the same happening with proteins.
2. Protein matchers also employ more sensitive matrices like `BLOSUM` or `PAM` to determine how similar two sequences are, but there is no such ‘sensitive’ matrix for genomic sequence comparison.

Question 8

What is the difference in the working of PSI-BLAST and BLAST programs?

PSI-BLAST

a. PSI-BLAST stands for **P**osition-**S**pecific **I**terative **B**asic **L**ocal **A**lignment **S**earch **T**ool

b. PSI-BLAST uses position-specific scoring matrices (PSSMs) to score matches between query and database sequences

c. While both tools search multiple databases for similar proteins, PSI-BLAST is much better at searching and finding remote homologs, i.e., sequences that are distantly related to your query sequence

BLAST

a. BLAST stands for **B**asic **L**ocal **A**lignment **S**earch **T**ool

b. BLAST which uses pre-defined scoring matrices such as BLOSUM

c. While both tools search multiple databases for similar proteins, BLAST tends to miss out on remote homologs

Question 9

i. **In BLAST database search algorithm, the match/mismatch ratio for comparing nucleotide sequences is chosen to be large for highly conserved sequences, while it is small for divergent sequences. Give reasons, why?**

ii. **Give the BLAST nucleotide substitution matrix for comparing sequences that are 95% conserved.**

→

i. a. Identical base matches are rewarded in the BLAST database search method, whereas mismatches are penalised.

b. In conserved sequences, there are more matches, while in small divergent sequences, there are less matches.

c. As a result, for conserved sequences, the match/mismatch ratio is kept high, but for divergent sequences, it is kept low.

ii. A ratio of 0.5 is ideal for sequences that are 95 percent preserved; the scoring matrix for this would be:

	A	G	T	C
A	1	-2	-2	-2
G	-2	1	-2	-2
T	-2	-2	1	-2

C	-2	-2	-2	1

Question 10

In BLOSUM62 matrix, a conserved Tryptophan position has score S(W,W) = 11, but a conserved Leucine position has score S(L,L) = 4. Give at least one reason why these values differ.

→

a. A protein sequence's amino acid distribution is not uniform. Some amino acids have a higher frequency of occurrence, whereas others have a lower frequency of occurrence.

b. Furthermore, certain amino acids mutate at a faster rate than others. Amino acids like Leucine have been shown to have a greater frequency than Tryptophan.

c. It's also difficult to substitute Tryptophan with Leucine since Tryptophan is more stable. As a result, Tryptophan has a higher identity score than Leucine.