Sci Bio Assi

1. Applications of multiple alignments; In Protein sequences, it is used in:

   - DNA encoding
   - SNP Identification
   - Inferring evolutionary relationships
   - Genome sequence assemble- shotgun sequencing
   - Developing primers and probes

   In DNA sequence, it is used in:

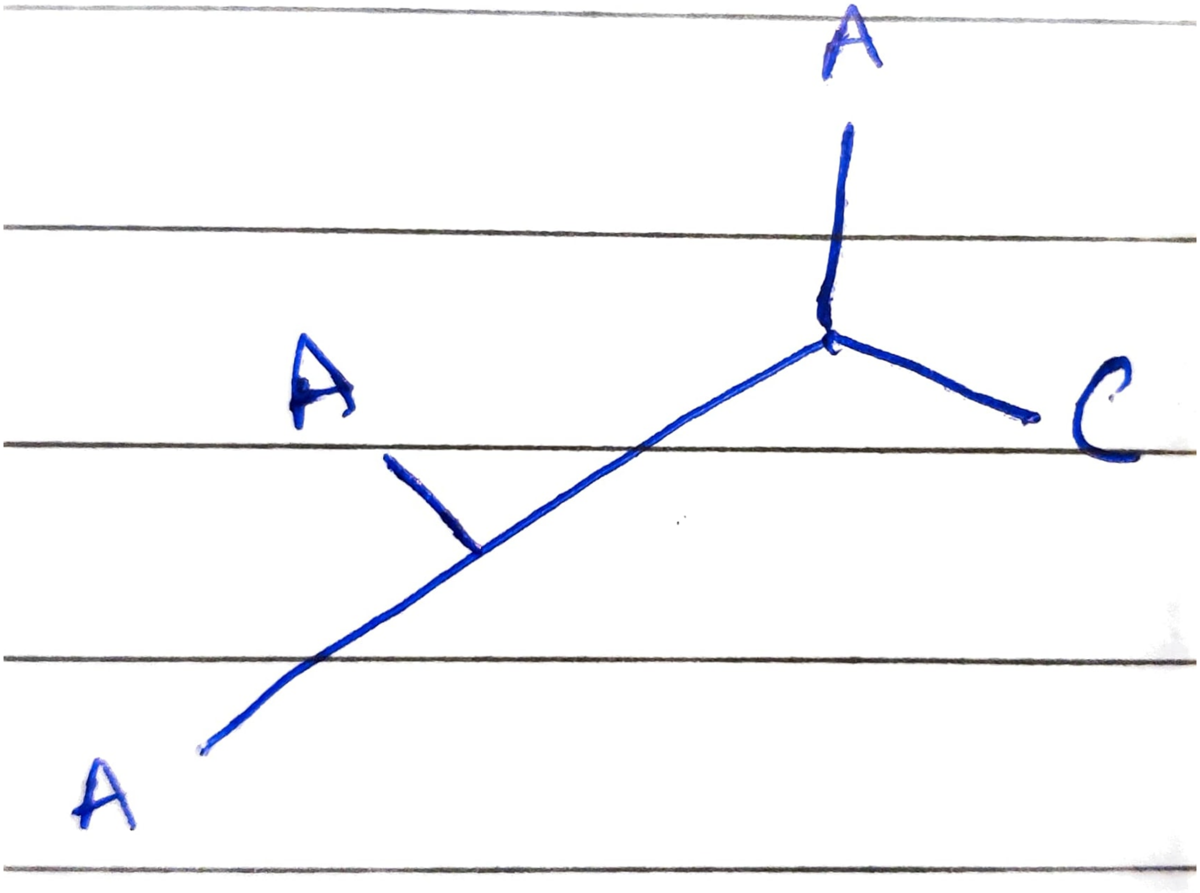
   - Building phylogenetic tree
   - Homology modeling of proteins
   - Predicting secondary and tertiary structures of new sequences
   - Identifying conserved patterns, motifs, and blocks in protein sequences
   - Identify related proteins in database searches
   - Constructing scoring matrices like PAM, BLOSUM

   More biological information is revealed by multiple alignments than by pairwise alignments. It enables the detection of conserved sequence patterns and motifs over the whole sequence family, which are difficult to identify when comparing only two sequences. A protein's multiple alignments can reveal several conserved and functionally important amino acids. It also allows for the visual representation of various types of amino acid residues using various colors, which aids in the analysis of events such as point mutations and indels.

2. For evaluating a given multiple sequence alignment, the sum of pairs is a scoring scheme that evaluates the cost column by column. The column cat is given by $\Sigma_{i<j} D(S_i, S_j)$. 1. Eg. Assuming mismatch cost as 1, match cost as 0, and indel cost as 1, the column cost of the following column is 26.

$$\begin{bmatrix} L \\ L \\ A \\ P \\ G \\ S \\ - \\ G \end{bmatrix} = 6+6+5+4+3+2 = 26$$

The sum of the pairs scoring scheme tends to overweight the contributions of differences from many very similar sequences. For example, if we consider the column $\begin{bmatrix} A \\ A \\ A \\ C \end{bmatrix}$, the sum of pairs score for the column is $3-3=0$. However, evolutionary speaking, the relationship between the sequences is



Then, a single $A \rightarrow C$ mutation can explain the data and thus SP tends to overcount mutations.

For related sequences, it ignores the assumption of a shared ancestor. As a result, the SP approach for evaluating MSAs is inherently problematic from an evolutionary standpoint.

We can score using a free-form technique. The total lengths of the tree branches may be determined using the substitutions in the MSA column since the MSA software finds a phylogenetic tree representing the connections among the sequences. Alternatively, a simpler tree can be utilized, with one of the sequences serving as the ancestor of all the others.

3. Steps involved in progressive alignment approach:

    a. Align each sequence to every other pair-wise

    b. Compute distances between each aligned pair

    c. Construct a phylogenetic tree

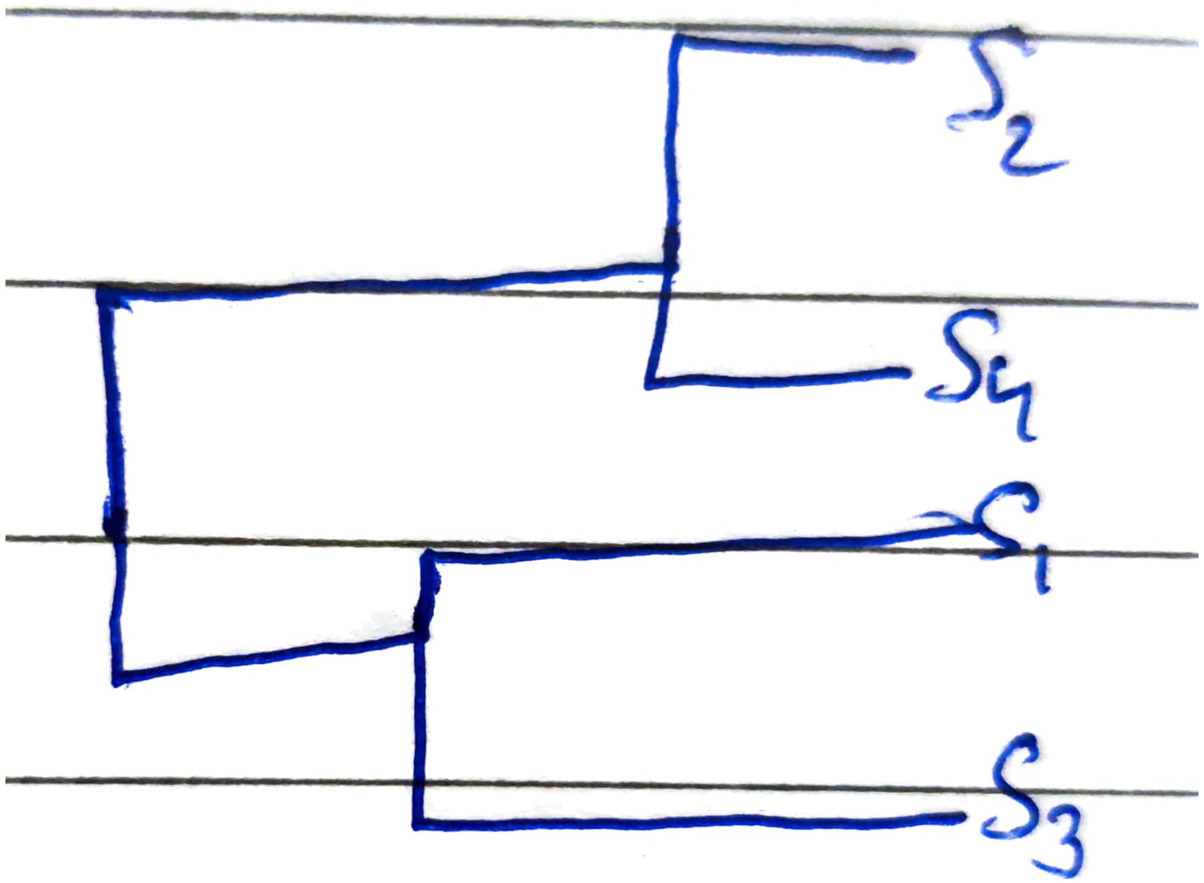
    d. Cluster closely related sequences

    e. Align closely related sequences first

    f. Gaps inserted in closely related sequences are propagated throughout

    For example, 4 sequences S_1, S_2, S_3, S_4.

    First, we perform 6 pairwise comparisons and then cluster analysis to construct the phylogenetic tree.



Then, we align the most similar pair

S_2 _____________________ _________

S_4 ____ __ ________________________

Then, we align the most similar pair

S_2 _____ ____________ ____________

S_3 ____ ____________ ______________

Then we align the above two alignments

S_2 __________________ _____ _______

S_4 __ __ ________________ _______

S_1 _________ ___________ _________

S_3 ____ ____________ ____________

The main drawback of progressive alignment is that the final MSA is dependent on the initial pair-wise sequence alignments. On the sequence tree, the first sequences to be aligned are the most closely related. If these sequences align perfectly, there will be a few faults; the more errors there are, the more errors are created, and these errors are propagated to the MSA. Another disadvantage is the selection of appropriate scoring matrices and gap penalties for the sequence, as the final MSA is dependent on initial pairwise alignment, which is sensitive to scoring schemes. We employ Bayesian approaches like Hidden Markov Models to solve the challenge of aligning more distantly linked sequences. It can also be fixed by continuously realigning subgroups of sequences and then aligning these subgroups into a global alignment of all sequences to address the problem. It enhances the overall alignment score. Groups can be chosen based on the phylogenetic tree's ordering or at random. To optimize as much as feasible, we can manually alter the alignment by eye examination.

4. $L = 50$ residues long

The alignment of N sequences fakes $= (2L)^{N-2}$ = 10^{2N-4}

$$\text{Time} = 5 \text{ billion years} = 5 \times 10^9 \text{ years} \\ = 5 \times 10^9 \times 365 \text{ days} = 5 \times 10^9 \times 365 \times 86400 \text{ s} = 15768 \times 15768 \times 10^13 \text{ s}$$

$$10^{2N-4} = 15768 \times 10^{13} \\ 2N - 4 = 13 + \log_{10} 15768 = 17.19 \\ 2N = 10.545$$

This implies our computer can align 10 sequences in 5 billion years.

6. MSA using a progressive approach:

S1: GATTCA

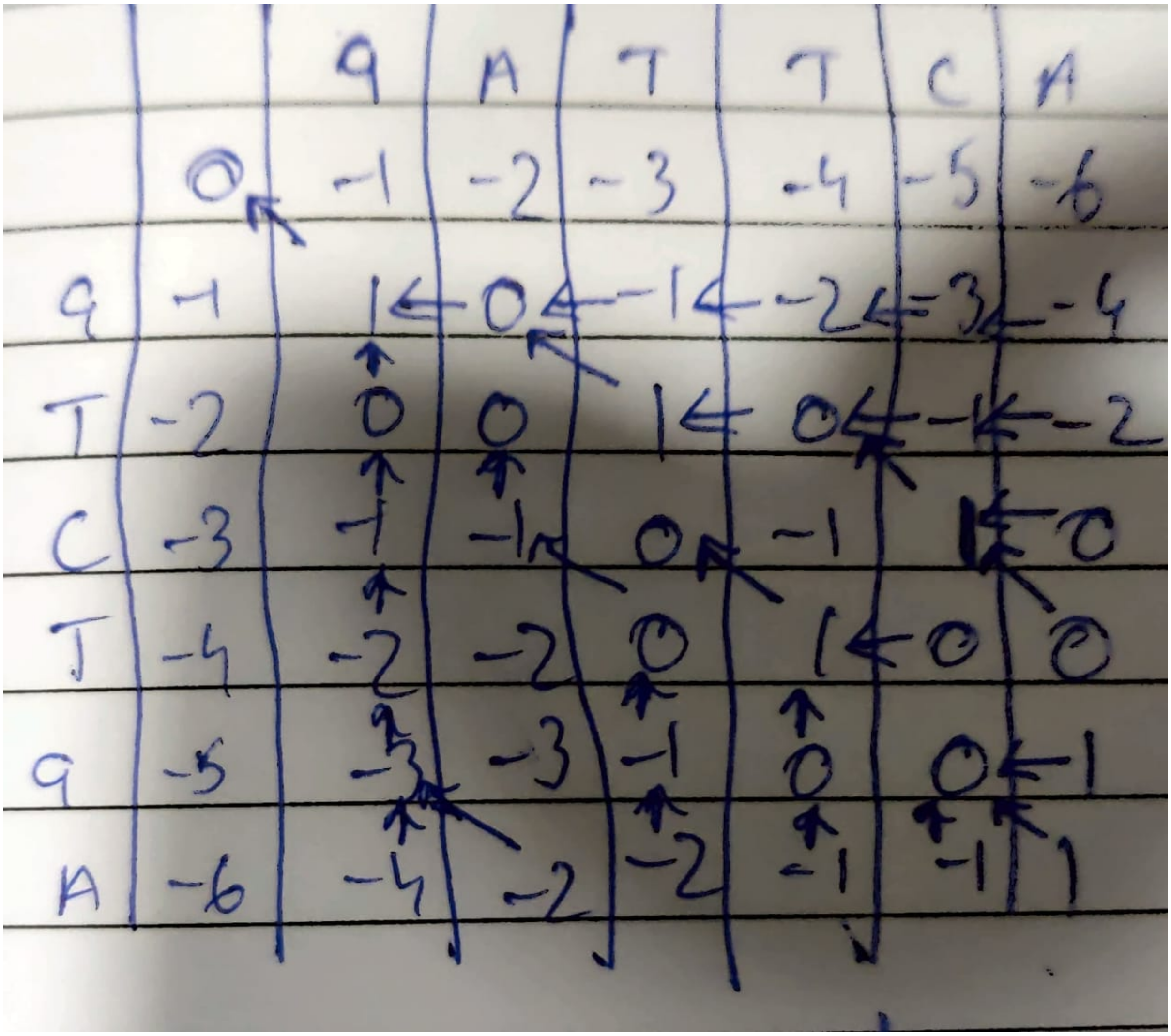
S2: GTCTGA

S3: GATATT

S4: GTCAGC

First, we have to ${}_4C_2 = 6$ pairwise alignments using Needleman Wunsch Algorithm. We take the distance between 2 sequences as the no. of mismatches.

$S_1 S_2$:



		G	A	T	T	C	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
T	-2	0	0	1	0	-1	-2
C	-3	-1	-1	0	-1	1	0
T	-4	-2	-2	0	1	0	0
G	-5	-3	-3	-1	0	0	1
A	-6	-4	-2	-2	-1	-1	1

Alignment: GAT_TCA

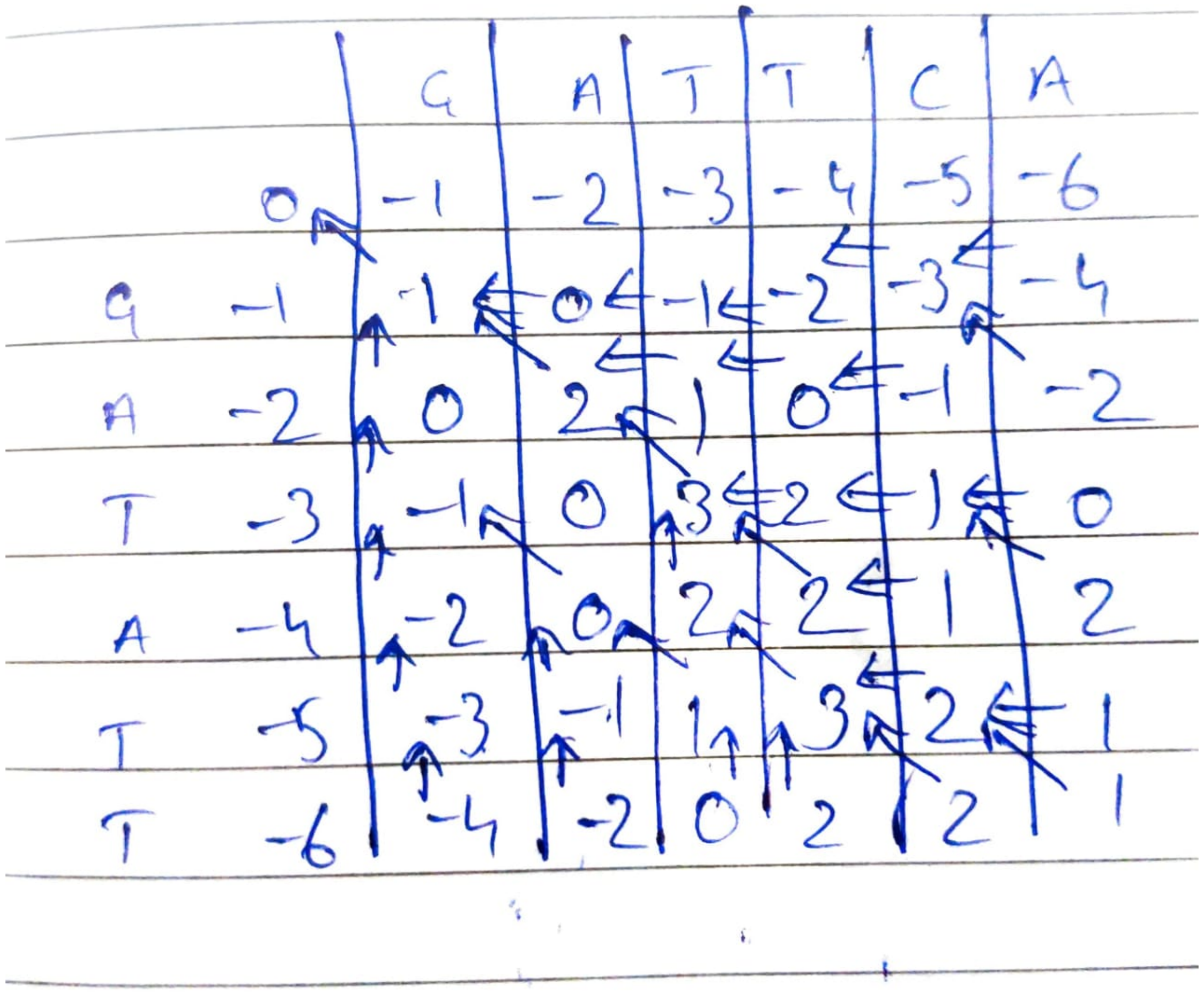
G_TCTGA

Alignment: GAT_TCA
G_TCTGA

Score: 1

Distance: 3

S_1S_3 :



		G	A	T	T	C	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
A	-2	0	2	1	0	-1	-2
T	-3	-1	0	3	2	1	0
A	-4	-2	0	2	2	1	2
T	-5	-3	-1	1	3	2	1
T	-6	-4	-2	0	2	2	1

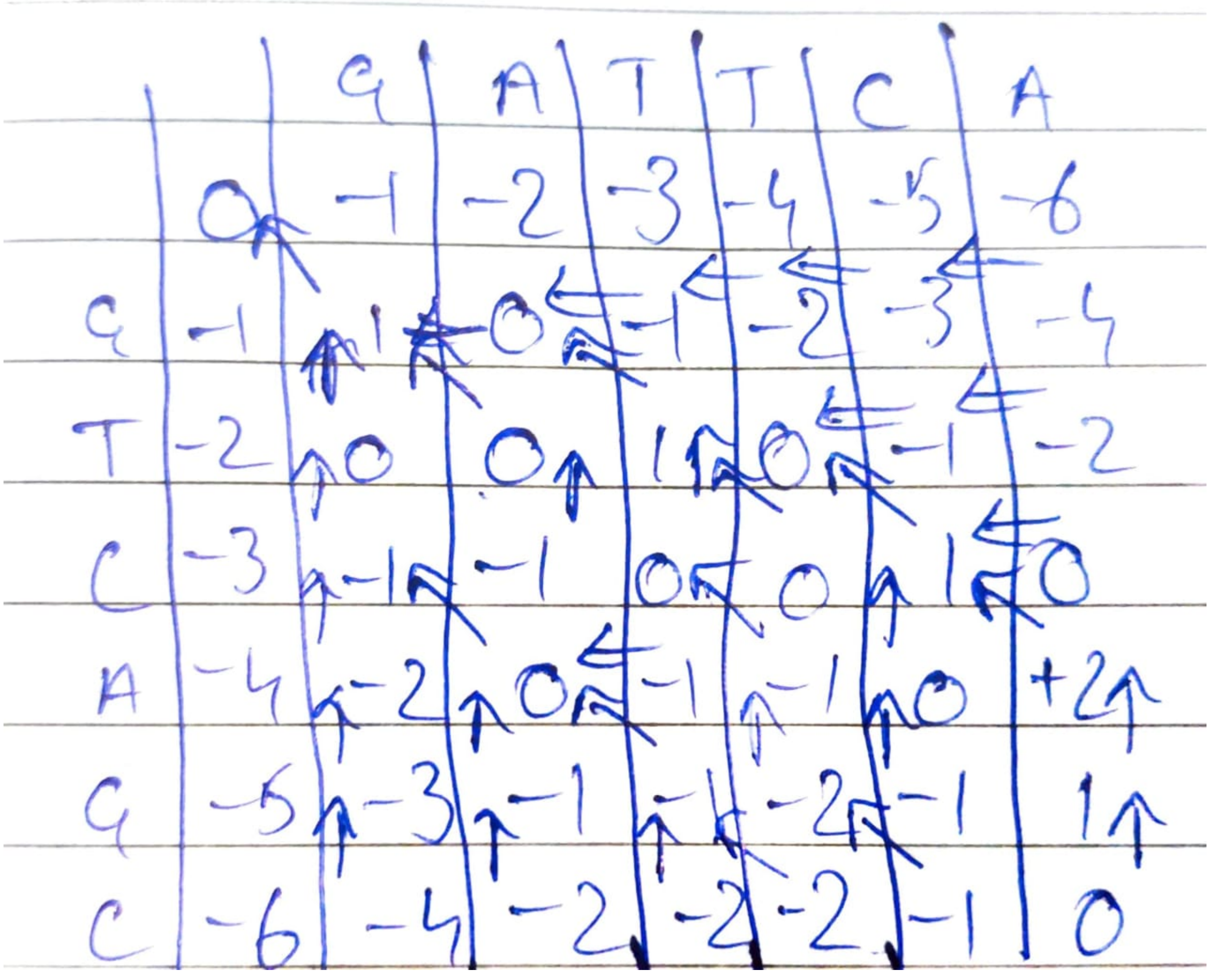
Alignment: GATTCA_ _

G_T_CAGC

Score: 1

Distance: 3

S_1S_4 :



		G	A	T	T	C	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
T	-2	0	0	1	0	-1	-2
C	-3	-1	-1	0	0	1	0
A	-4	-2	0	-1	-1	0	+2
G	-5	-3	-1	-1	-2	-1	1
C	-6	-4	-2	-2	-2	-1	0

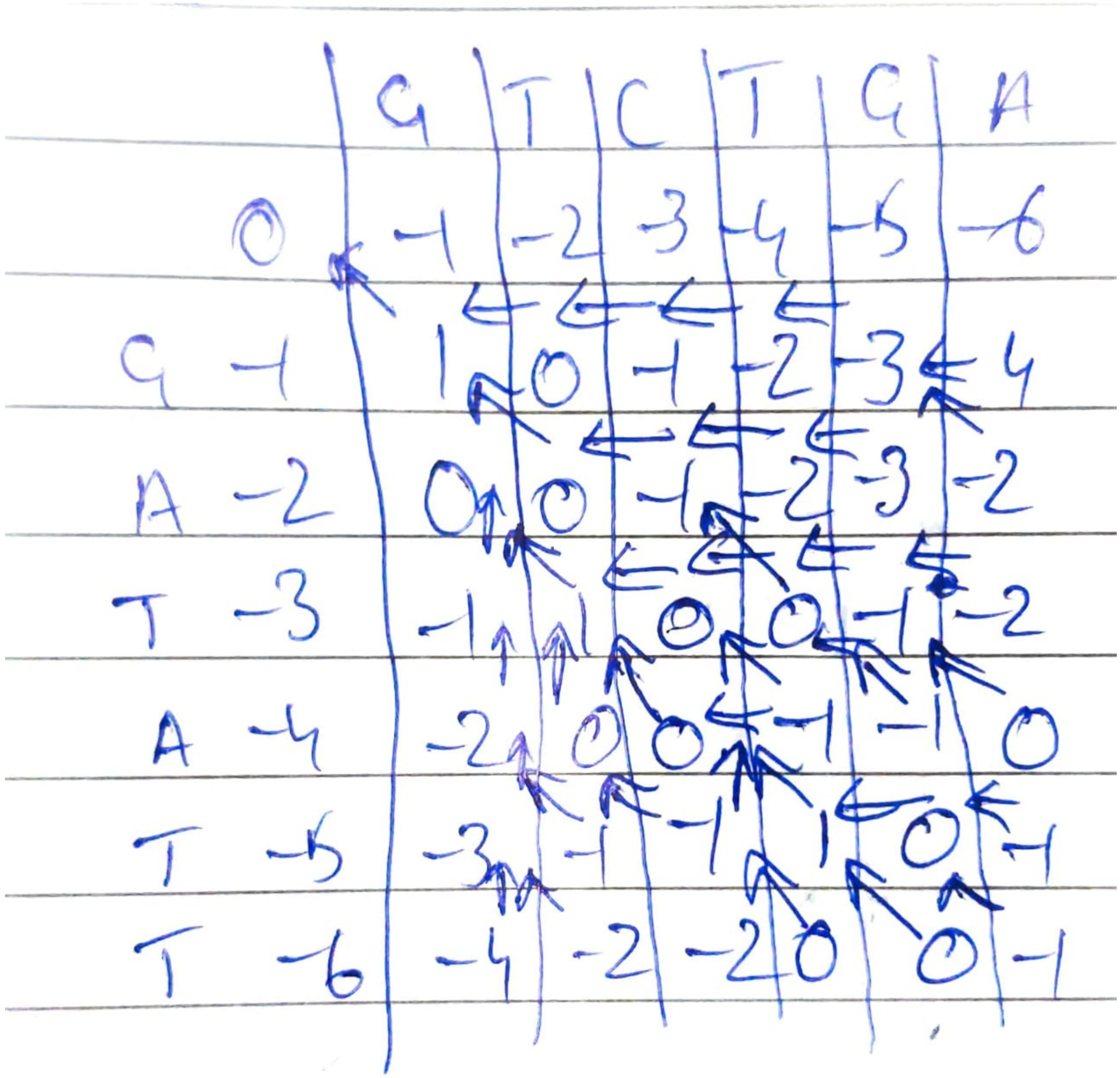
Alignment: GATTCA_ _

G_T_CAGC

Score: 0

Distance: 4

$S_2 S_3$:



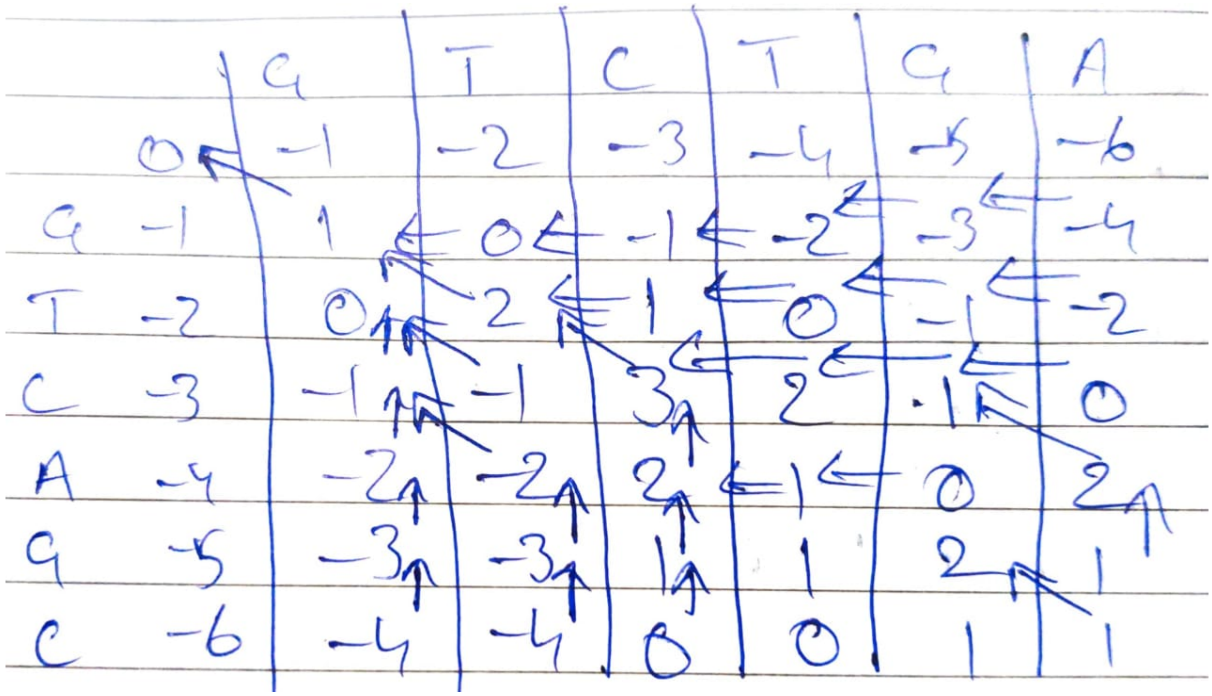
		G	T	C	T	G	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	4
A	-2	0	0	-1	-2	-3	-2
T	-3	-1	1	0	0	-1	-2
A	-4	-2	0	0	-1	-1	0
T	-5	-3	-1	-1	1	0	-1
T	-6	-4	-2	-2	0	0	-1

Alignment: G_TCTGA
GATAT_T

Score: -1

Distance: 4

S_2S_4 :



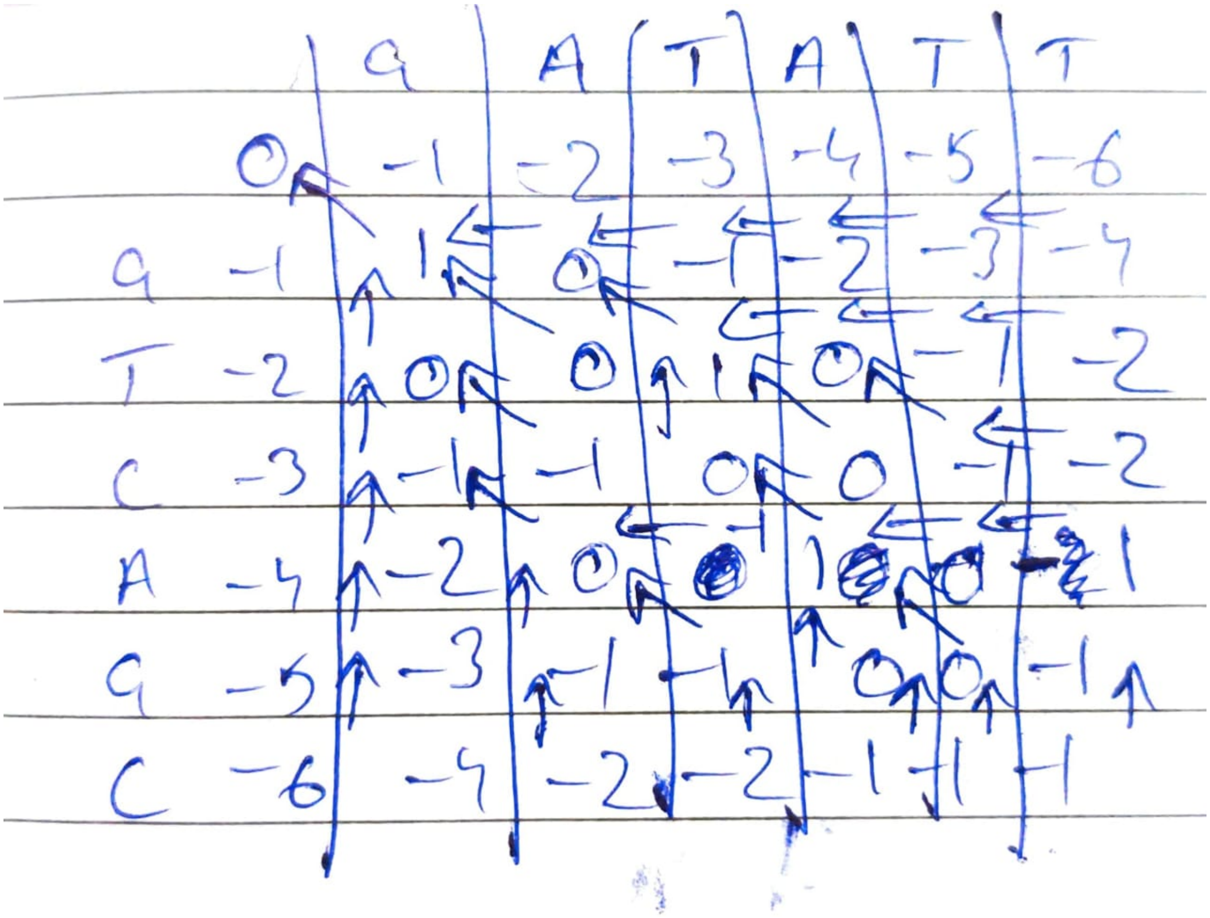
		G	T	C	T	G	A
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
T	-2	0	2	1	0	-1	-2
C	-3	-1	-1	3	2	1	0
A	-4	-2	-2	2	1	0	2
G	-5	-3	-3	1	1	2	1
C	-6	-4	-4	0	0	1	1

Alignment: GTC_TGA
GTCA_GC

Score: 1

Distance: 3

S_3S_4 :



		G	A	T	A	T	T
	0	-1	-2	-3	-4	-5	-6
G	-1	1	0	-1	-2	-3	-4
T	-2	0	0	1	0	-1	-2
C	-3	-1	-1	0	0	-1	-2
A	-4	-2	0	-1	1	0	-1
G	-5	-3	-1	-1	0	0	-1
C	-6	-4	-2	-2	-1	-1	-1

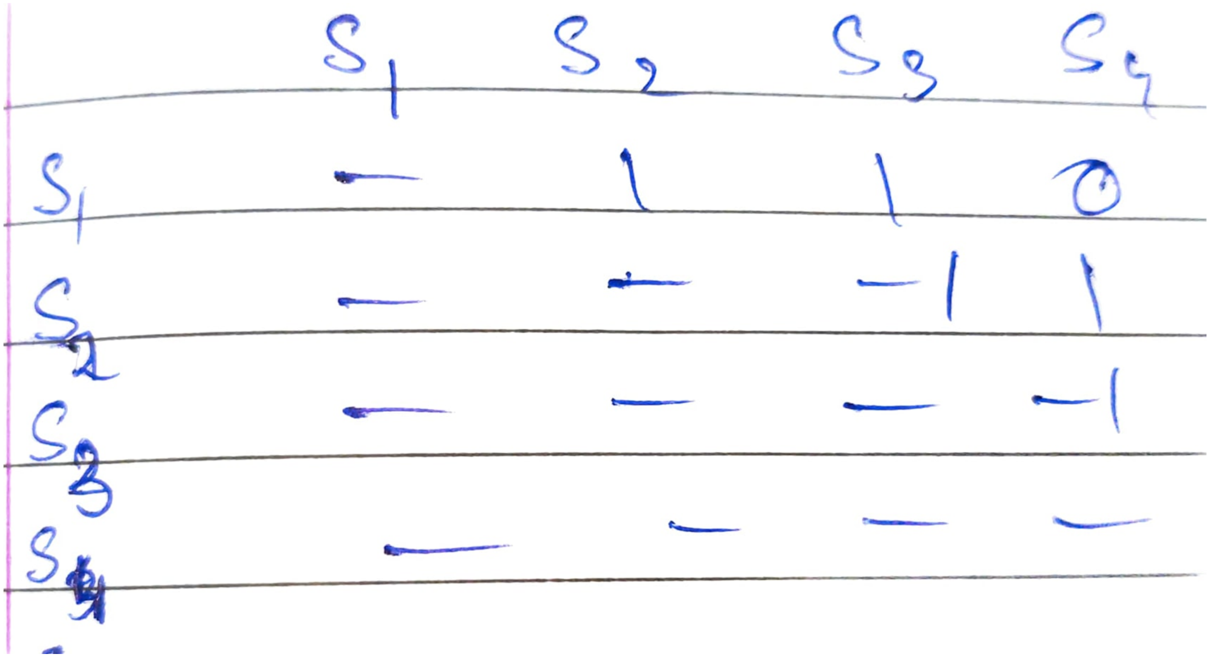
Alignment: GAT_ATT

G_TCAGC

Score: -1

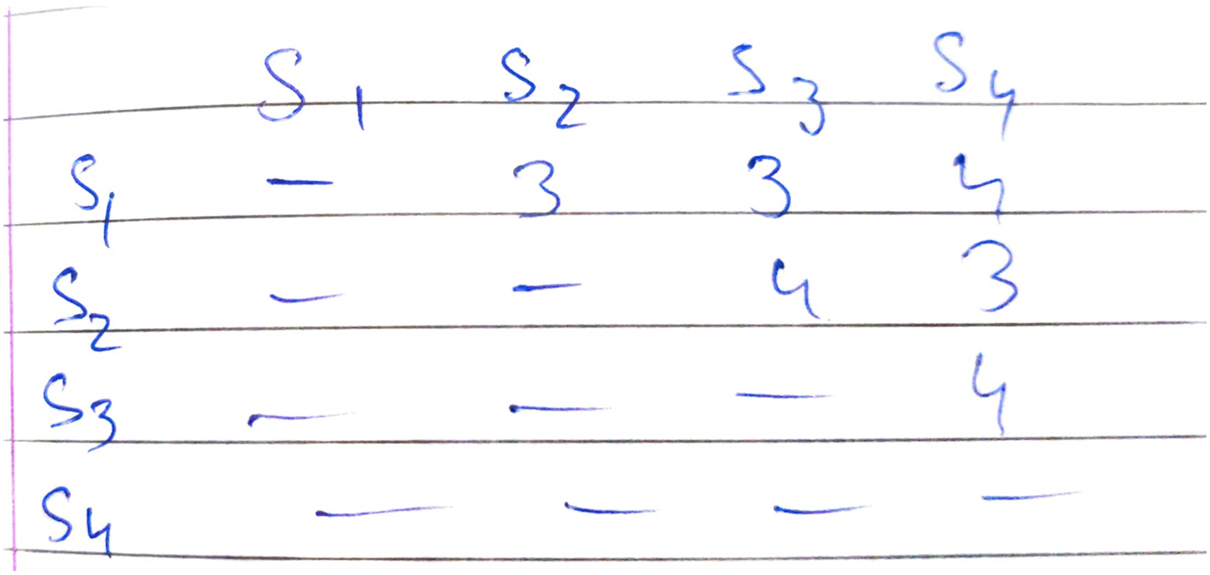
Distance: 4

Scoring pairwise matrix:



	S_1	S_2	S_3	S_4
S_1	—	1	1	0
S_2	—	—	—1	1
S_3	—	—	—	—1
S_4	—	—	—	—

Distance Matrix:



	S_1	S_2	S_3	S_4
S_1	—	3	3	4
S_2	—	—	4	3
S_3	—	—	—	4
S_4	—	—	—	—