Comparison of sequences

INDRAPRASTHA INSTITUTE *of* INFORMATION TECHNOLOGY **DELHI**

Dr. Jaspreet Kaur Dhanjal

Assistant Professor, Center for Computational Biology

Email ID: jaspreet@iiitd.ac.in

August 19, 2025

Global Alignment

Sequence 1: ATTAC

Sequence 2: ATGC

Scoring scheme: Match: +2 Mismatch: -1 Gap: -1

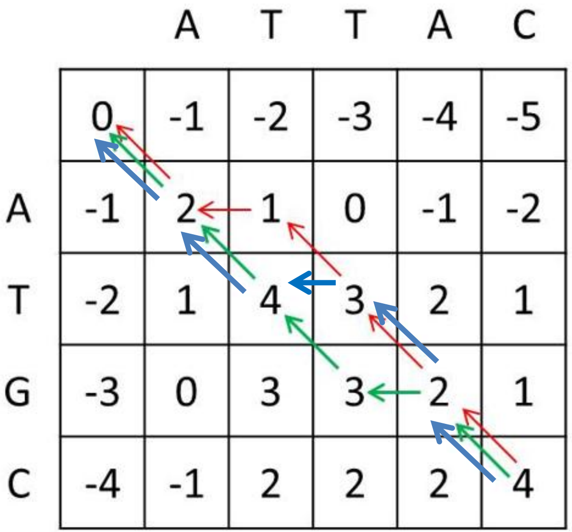
		A	T	T	A	C
	0	-1	-2	-3	-4	-5
A	-1	2	1	0	-1	-2
T	-2	1	4	3	2	1
G	-3	0	3	3	2	1
C	-4	-1	2	2	2	4

Global Alignment

- **Mismatching base pairs**
- **Two possibilities of optimal alignment**

Sequence 1: ATTAC
Sequence 2: ATGC

Scoring scheme:
Match: +2
Mismatch: -1
Gap: -1



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

Optimal alignments:

```
ATTAC      ATTAC      ATTAC
| | |      ||  |      ||  |
A-TGC      ATG-C      AT-GC
```

Score: 4

Score: 4

Score:4

Global Alignment

Sequence 1: ATTAC
Sequence 2: ATGC

	Empty string	A	T	T	A	C
Empty string		A -	AT --	ATT ---	ATTA ----	ATTAC -----
A	- A	Best alignment between A and A	Best alignment between A and AT AT / AT A- / -A	Best alignment between A and ATT	Best alignment between A and ATTA	Best alignment between A and ATTAC
T	-- AT	Best alignment between AT and A	Best alignment between AT and AT	Best alignment between AT and ATT	Best alignment between AT and ATTA	Best alignment between AT and ATTAC
G	--- ATG	Best alignment between ATG and A	Best alignment between ATG and AT	Best alignment between ATG and ATT	Best alignment between ATG and ATTA	Best alignment between ATG and ATTAC
C	---- ATGC	Best alignment between ATGC and A	Best alignment between ATGC and AT	Best alignment between ATGC and ATT	Best alignment between ATGC and ATTA	Best alignment between ATGC and ATTAC

Local Alignment

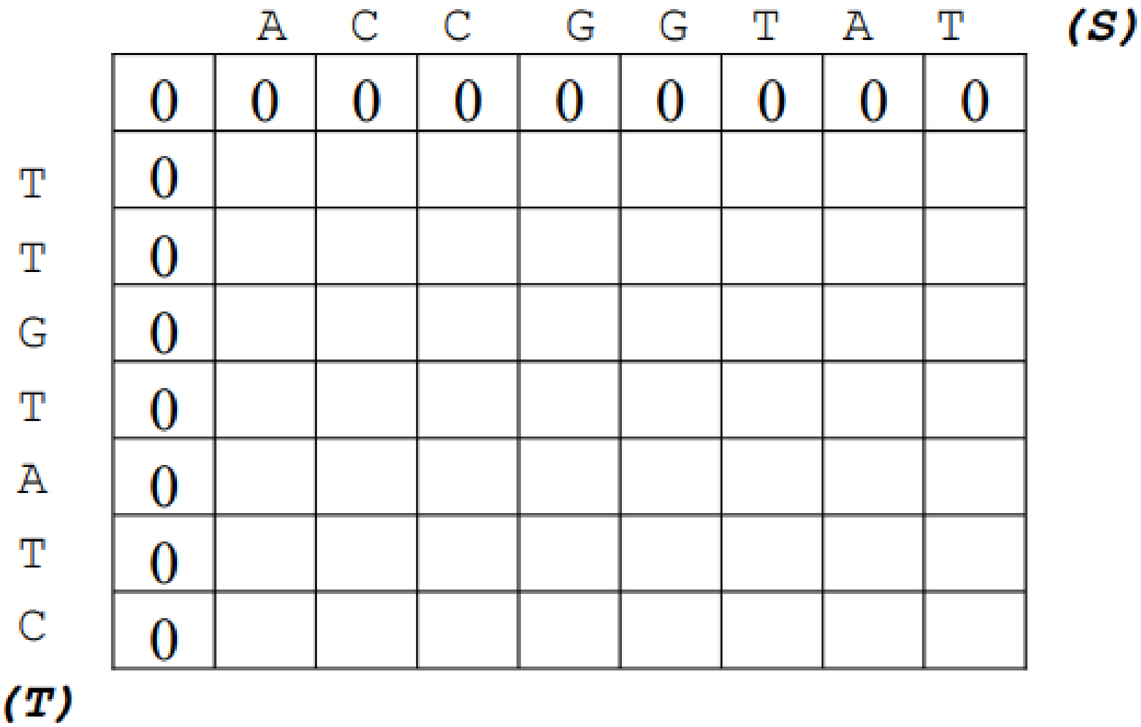
- Between two sequences, find the best two subsequences and their score
- We want to ignore gaps in the matched sequences

```
ATGCGCTACCGTATCCTAGGAC
       |||||||| ||
-------ACCGTATC-TA----
```

- Use the same types of substitution matrix and gap penalties
  - If S*i* matches T*j* then σ(S*i*,T*j*) >=0
  - If they do not match or represent a gap then <=0
- Use a modification of the previous dynamic programming approach
  - Initialize row0/col0 with 0
  - Lowest allowable value of any cell is 0
  - Find the cell with the highest value (i,j) and extend
  - The alignment back to the first zero value
  - The score of the alignment is the value in that cell

Local Alignment

min value of any cell is 0



		A	C	C	G	G	T	A	T	*(S)*
	0	0	0	0	0	0	0	0	0	
T	0									
T	0									
G	0									
T	0									
A	0									
T	0									
C	0									
(T)										

Local Alignment

Find biggest cell and map alignment from there

(S)

		A	C	C	G	G	T	A	T
	0	0	0	0	0	0	0	0	0
T	0	0	0	0	0	0	2	1	2
T	0	0	0	0	0	0	2	1	3
G	0	0	0	0	2	2	1	1	2
T	0	0	0	0	1	1	4	3	3
A	0	2	1	0	0	0	3	6	5
T	0	1	1	0	0	0	2	5	8
C	0	0	3	3	2	1	1	4	7

(T)

$$V(i,0) = 0,\ V(0,j) = 0$$

$$V(i,j) = \max \begin{cases} 0 \\ V(i-1,j-1) + \sigma(S_i,T_j) \\ V(i-1,j) + \sigma(S_i,-) \\ V(i,j-1) + \sigma(-,T_j) \end{cases}$$

Local Alignment

Find biggest cell and map alignment from there

		A	C	C	G	G	T	A	T	*(S)*
	0	0	0	0	0	0	0	0	0	
T	0	0	0	0	0	0	2	1	2	
T	0	0	0	0	**0**	0	2	1	3	
G	0	0	0	0	2	**2**	1	1	2	
T	0	0	0	0	1	1	**4**	3	3	
A	0	2	1	0	0	0	3	**6**	5	
T	0	1	1	0	0	0	2	5	**8**	
C	0	0	3	3	2	1	1	4	7	
(T)										

```
GTAT (S)
||||
GTAT (T)
```

This is also known as Smith–Waterman algorithm.

Local Alignment

- For finding local alignments, the Needleman-Wunsch algorithm is slightly modified to start over and find a new local alignment whenever the existing alignment score goes negative.

- Since a local alignment can start anywhere, the first row and column in the matrix is initialized to zeros.

- The iteration step is modified to include a zero to include the possibility that starting a new alignment would be cheaper than having many mismatches.

- Furthermore, since the alignment can end anywhere, the entire matrix is traversed to find the optimal alignment score (not only in the bottom right corner). The rest of the algorithm, including traceback, remains unchanged, with traceback indicating an end at a zero, indicating the start of the optimal alignment.

End Free Alignment

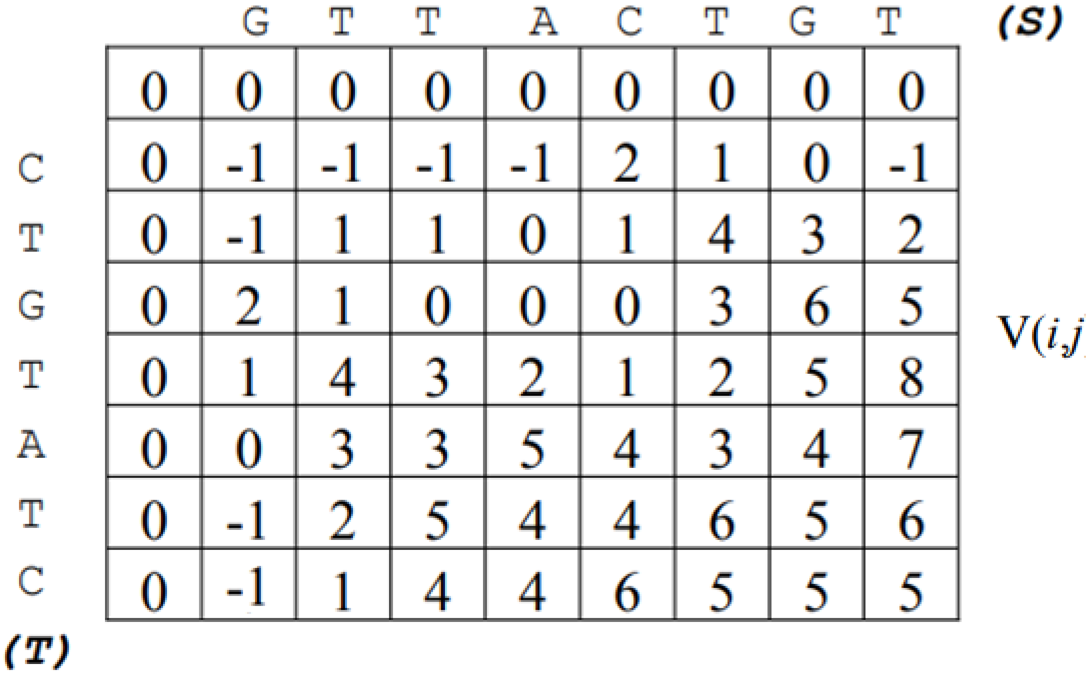
- To find overlap between two sequences, i.e. alignment between the start of one and the end of the other sequence

```
CGCTACC   TCCTAGGAC
    |||   ||||           ⟶   CGCTACCGTATCCTAGGAC
    ACCGTATCCT
```

- Essential to DNA sequencing strategies
  – Building genome fragments out of shorter sequencing data
- Another variant of the Global Alignment Problem
  - Set the initial conditions to zero weight, this allow indels/gaps at the ends without penalty
  - Fill the array/table using the same recursion model used in global/local alignment
  - Find the best alignment that ends in one row or column
  - Trace this back

End Free Alignment

min value row0 & col0 is 0



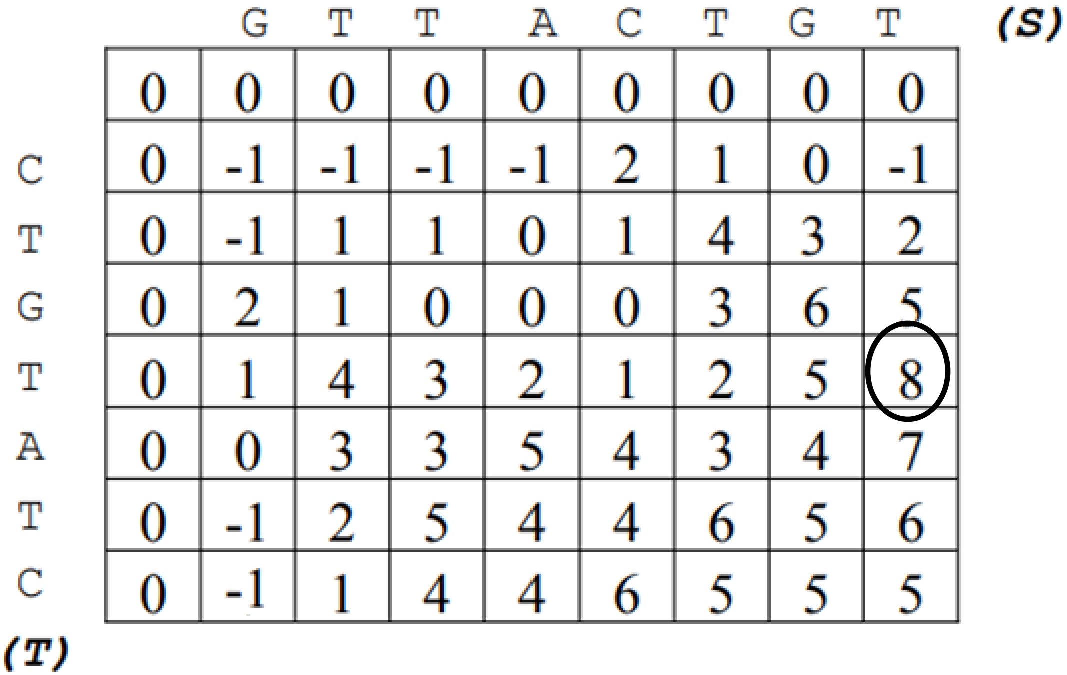
		G	T	T	A	C	T	G	T	(S)
	0	0	0	0	0	0	0	0	0	
C	0	-1	-1	-1	-1	2	1	0	-1	
T	0	-1	1	1	0	1	4	3	2	
G	0	2	1	0	0	0	3	6	5	
T	0	1	4	3	2	1	2	5	8	
A	0	0	3	3	5	4	3	4	7	
T	0	-1	2	5	4	4	6	5	6	
C	0	-1	1	4	4	6	5	5	5	
(T)										

$$V(i,0) = 0,\ V(0,j) = 0$$

$$V(i,j) = \max \begin{cases} V(i-1,j-1) + \sigma(S_i,T_j) \\ V(i-1,j) + \sigma(S_i,-) \\ V(i,j-1) + \sigma(-,T_j) \end{cases}$$

End Free Alignment

Find the best ‘end’ point in an end col or row



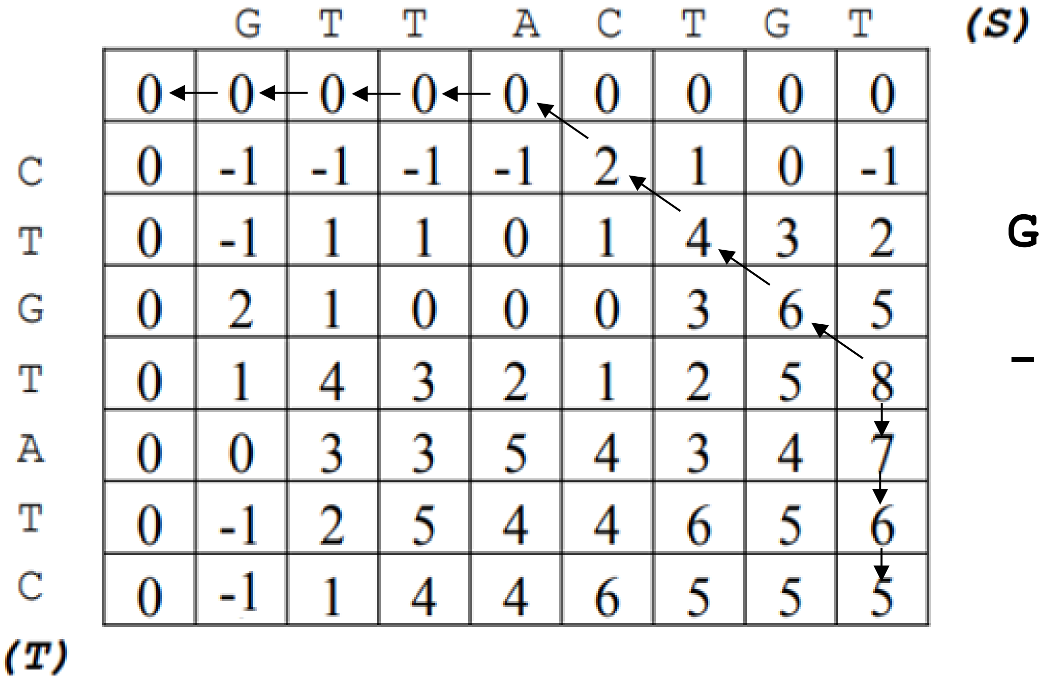
		G	T	T	A	C	T	G	T
	0	0	0	0	0	0	0	0	0
C	0	-1	-1	-1	-1	2	1	0	-1
T	0	-1	1	1	0	1	4	3	2
G	0	2	1	0	0	0	3	6	5
T	0	1	4	3	2	1	2	5	8
A	0	0	3	3	5	4	3	4	7
T	0	-1	2	5	4	4	6	5	6
C	0	-1	1	4	4	6	5	5	5

(S)

(T)

End Free Alignment

Trace the best route from there to the origin and end



		G	T	T	A	C	T	G	T	(S)
	0	0	0	0	0	0	0	0	0	
C	0	-1	-1	-1	-1	2	1	0	-1	
T	0	-1	1	1	0	1	4	3	2	
G	0	2	1	0	0	0	3	6	5	
T	0	1	4	3	2	1	2	5	8	
A	0	0	3	3	5	4	3	4	7	
T	0	-1	2	5	4	4	6	5	6	
C	0	-1	1	4	4	6	5	5	5	
(T)										

```
GTTACTGT---  (S)
    ||
----CTGTATC  (T)
```

Dynamic programming for pairwise alignment

- Dynamic programming algorithms can solve global, local and ends-free alignment
- They give the optimum score and alignment using the parameters given
- When searching multiple genomes, the sizes still get too big!
- Choice of GAP penalty and substitution matrix are critically important

Practice question

Use dynamic programming and the provided scoring scheme to globally align the following DNA sequences

Seq 1: GCTTAGC
Seq 2: GCATTGC

Scoring scheme:
Match = +3
Mismatch = -2
Gap = -3

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

Practice question

Use dynamic programming and the provided scoring scheme to globally align the following DNA sequences

```
Seq 1:  GCTTAGC
Seq 2:  GCATTGC
```

Scoring scheme:
Match = +3
Mismatch = -2
Gap = -3

		G	C	T	T	A	G	C
	0	-3	-6	-9	-12	-15	-18	-21
G	-3	3	0	-3	-6	-9	-12	-15
C	-6	0	6	3	0	-3	-6	-9
A	-9	-3	3	4	1	3	0	-3
T	-12	-6	0	6	7	4	1	-2
T	-15	-9	-3	3	9	6	3	0
G	-18	-12	-6	0	6	7	9	6
C	-21	-15	-9	-3	3	4	6	12

Optimal alignment:

```
GC-TTAGC
|| || ||
GCATT-GC
```