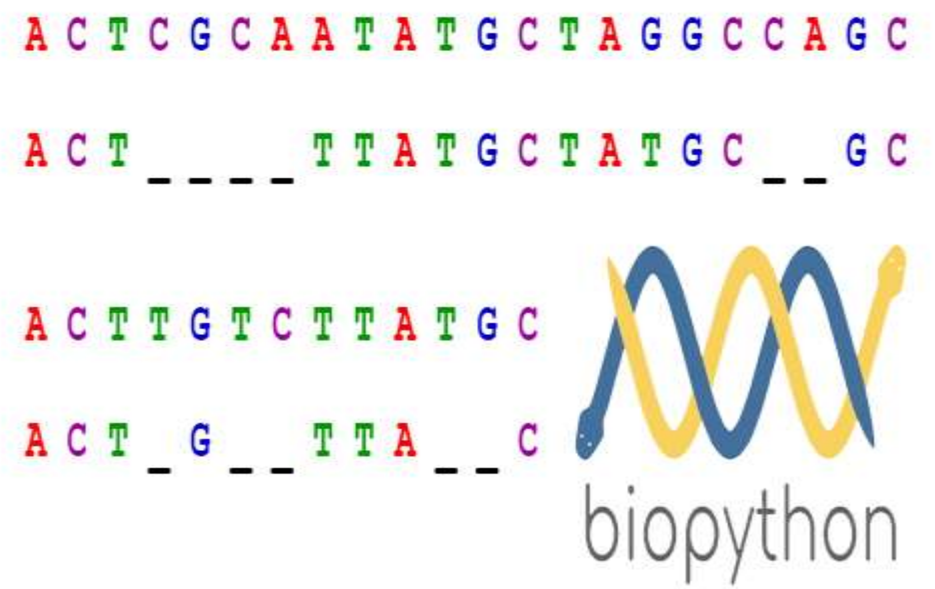
# Sequence Alignment

Lecture – 5

Department of CSE, DIU



# CONTENTS

1. Sequence Alignment
   - Why align sequences

2. Sequence Alignment Methods
   - Pairwise Alignment
   - Multiple Sequence Alignment

3. Pairwise Sequence Alignment Methods
   - Global Alignment (Needleman-Wunsch)
   - Local Alignment (Smith-Waterman)

# 1. Sequence Alignment

Why and how align sequences

# Sequence Alignment

A way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences

```
CTGTCG-CTGCACG
-TGC-CG-TG----
```

# Why align sequences?

- Useful for discovering
    - Functional
    - Structural and
    - Evolutionary relationship
  - For example
    - To find whether two (or more) genes or proteins are evolutionarily related to each other
    - Two proteins with similar sequences will probably be structurally or functionally similar

# 2. Sequence Alignment Methods

Pairwise and Multiple

# Pairwise Sequence Alignment

- A pair of sequences as input

- Align them in such a way that, for that particular alignment the assumed region of similarity produces higher score than all the other alignments

- Methods
  - Global Alignment (Needleman-Wunsch)
  - Local Alignment (Smith-Waterman)

```
CTGTCGCTGCACG--
-------TGC-CGTG
```

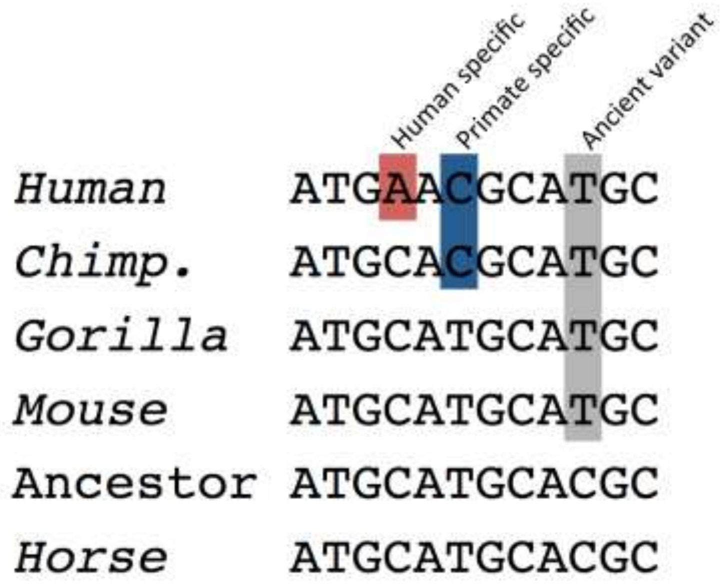
# Pairwise Sequence Alignment

- Idea:

  Display one sequence above another with spaces inserted in both to reveal similarity

```
A:  C A T - T C A - C
    |   |     | |   |
B:  C - T C G C A G C
```

# Multiple Sequence Alignment



```
Human     ATGAACGCATGC
Chimp.    ATGCACGCATGC
Gorilla   ATGCATGCATGC
Mouse     ATGCATGCATGC
Ancestor  ATGCATGCACGC
Horse     ATGCATGCACGC
```

- Three or more than three sequences as input

- Align all the sequences altogether in such a manner that the alignment produces highest score

# 3. Pairwise Sequence Alignment

Global and Local methods

# Global Alignment (Needleman-Wunsch)

3 Major Steps

- Create 2D Matrix
- Trace back
- Final Alignment

Create 2D Matrix

- Row x Col 2D matrix draw (Row , Col size of seq1 and seq2 respectively)
- Place 2 seqs as Row and Column Header
- Cell (0,0) = 0
- Cell (0,1) to Cell (0,Column) and Cell (1,0) to Cell (Row,0) value = delete gap value from previous cell value
- For other cell values, follow equation in (1)

Trace back

- Start from Cell (Row, Col)
- Go back up to Cell (0,0)

Final Alignment

- Start from Cell (Row, Col)
- If ↖ then, place character in both seq
- If ← or ↑ then character in start seq & gap in end seq

# Global Alignment (Needleman-Wunsch) - Example

Input

- seq1 = TTGT
- seq2 = ATTTGCT

Scoring Scheme

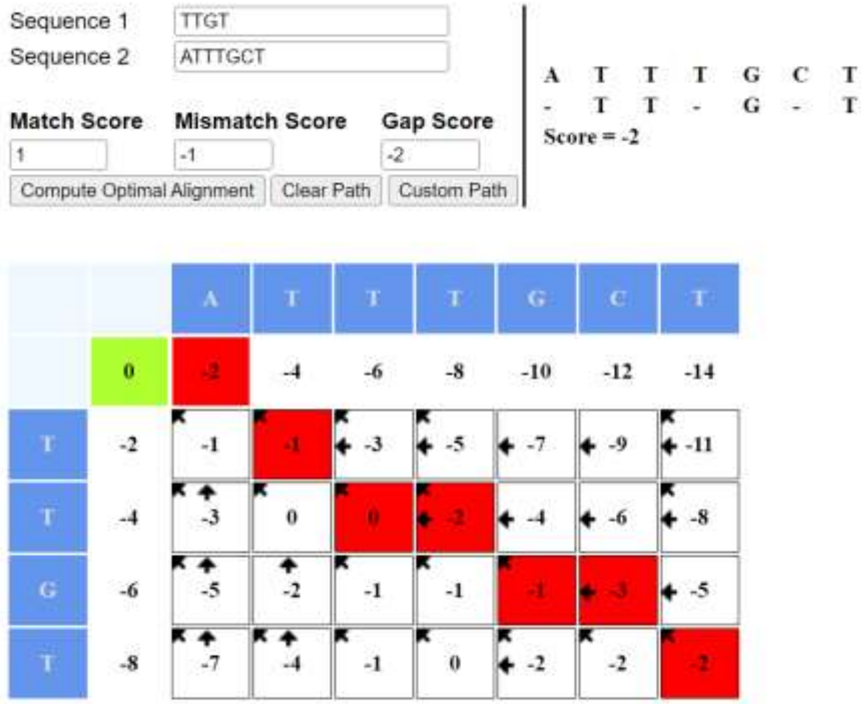
$\delta(x, x) = 1$ (Match)

$\delta(x,-) = -2$ (Gap)

$\delta(x, y) = -1$ (Mis match)

$$V_{i,j} = \max \begin{cases} V_{i-1,j} + \delta(s_i, -) \\ V_{i,j-1} + \delta(-, t_j) \\ V_{i-1,j-1} + \delta(s_i, t_j) \end{cases}$$

Eq. 1: Cell Value



Sequence 1 TTGT

Sequence 2 ATTTGCT

Match Score 1

Mismatch Score -1

Gap Score -2

```
A  T  T  T  G  C  T
-  T  T  -  G  -  T
Score = -2
```

| | | A | T | T | T | G | C | T |
|---|---|---|---|---|---|---|---|---|
| | 0 | -2 | -4 | -6 | -8 | -10 | -12 | -14 |
| T | -2 | ↖ -1 | ↖ -1 | ↖ ← -3 | ↖ ← -5 | ← -7 | ← -9 | ↖ ← -11 |
| T | -4 | ↖ ↑ -3 | ↖ 0 | ↖ 0 | ↖ ← -2 | ← -4 | ← -6 | ↖ ← -8 |
| G | -6 | ↖ ↑ -5 | ↑ -2 | ↖ -1 | ↖ -1 | ↖ -1 | ← -3 | ← -5 |
| T | -8 | ↖ ↑ -7 | ↖ ↑ -4 | ↖ -1 | ↖ 0 | ↖ ← -2 | ↖ -2 | ↖ -2 |

# Local Alignment (Smith-Waterman)

3 Major Steps

- -Create 2D Matrix
- -Trace back
- -Final Alignment

Create 2D Matrix

- Row x Col 2D matrix draw (Row , Col size of seq1 and seq2 respectively)
- Place 2 seqs as Row and Column Header
- First Row, First Column all value = 0
- For other cell values, follow equation in (2)

Trace back

- Start from each Cell which has the maximum value in the entire matrix
- Go back up to the Cell where first time 0 occurs

Final Alignment

- Start from each Cell with max value
- If ↖ then, place character in both seq
- If ← or ↑ then character in start seq & gap in end seq

# Local Alignment (Smith-Waterman) - Example

Input

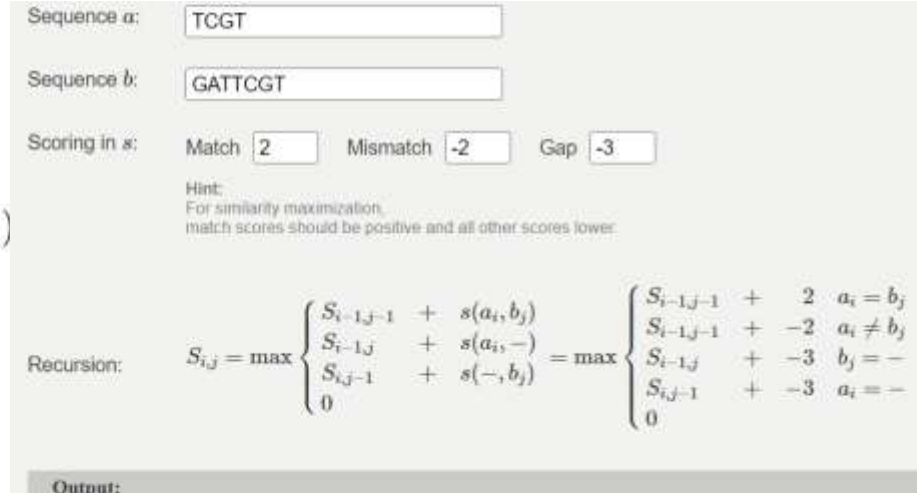
- seq1 =  TCGT
- seq2 = GATTCGT

Scoring Scheme

$\delta(x, x) = 2$ (Match)
$\delta(x,-) = -3$ (Gap)
$\delta(x, y) = -2$ (Mis match)

$$A[i,j] = \max \begin{cases} A[i, j-1] + \text{gap} \\ A[i-1, j] + \text{gap} \\ A[i-1, j-1] + \text{match}(i,j) \\ 0 \end{cases}$$

Eq. 2: Cell Value



Sequence $a$: TCGT

Sequence $b$: GATTCGT

Scoring in $s$: Match 2 Mismatch -2 Gap -3

Hint:
For similarity maximization,
match scores should be positive and all other scores lower.

Recursion:

$$S_{i,j} = \max \begin{cases} S_{i-1,j-1} & + & s(a_i, b_j) \\ S_{i-1,j} & + & s(a_i, -) \\ S_{i,j-1} & + & s(-, b_j) \\ 0 \end{cases} = \max \begin{cases} S_{i-1,j-1} & + & 2 & a_i = b_j \\ S_{i-1,j-1} & + & -2 & a_i \neq b_j \\ S_{i-1,j} & + & -3 & b_j = - \\ S_{i,j-1} & + & -3 & a_i = - \\ 0 \end{cases}$$

Output: