Dept. of Computer Science and Engineering
University of Rajshahi
www.ru.ac.bd

Dr. Shamim Ahmad

Local alignment

Local alignment methods find related regions ***within*** sequences - they can consist of a **subset** of the characters within each sequence.

For example, positions 20-40 of *sequence A* might be aligned with positions 50-70 of *sequence B*.

This is a more flexible technique than *global alignment* and has the advantage that *related regions* which appear in a different order in the two proteins (which is known as **domain shuffling**) can be identified as being related.

This is **not** possible with *global alignment* methods.

Similarity Matrix

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

The Smith Waterman algorithm

The **Smith-Waterman algorithm** (1981) is for determining similar regions between two nucleotide or protein sequences.

Smith-Waterman is also a dynamic programming algorithm and improves on Needleman-Wunsch. As such, it has the desirable property that it is guaranteed to find the **optimal local alignment** with respect to the scoring system being used (which includes the substitution matrix and the gap-scoring scheme).

However, the Smith-Waterman algorithm is **demanding of time and memory** resources: in order to align two sequences of lengths m and n, O(mn) time and space are required.

As a result, it has largely been replaced in practical use by the **BLAST algorithm**; although not guaranteed to find optimal alignments, BLAST is much more efficient.

Smith–Waterman Algorithm

	Smith–Waterman algorithm	Needleman–Wunsch algorithm
Initialization	First row and first column are set to 0	First row and first column are subject to gap penalty
Scoring	Negative score is set to 0	Score can be negative
Traceback	Begin with the highest score, end when 0 is encountered	Begin with the cell at the lower right of the matrix, end at top left cell

The Smith-Waterman algorithm

1. Create a table of size (*m*+1)x(*n*+1) for sequences **s** and **t** of lengths *m* and *n*,
2. Fill table entries (1,1:*m*+1) and (1:n+1,1) with zeros.
3. Starting from the top left, compute each entry using the recursive relation:

$$M_{i,j} = \max \begin{Bmatrix} M_{i-1,j-1} + \sigma(\mathbf{s}_i, \mathbf{t}_j) \\ M_{i-1,j} + \sigma(\mathbf{s}_i, -) \\ M_{i,j-1} + \sigma(-, \mathbf{t}_j) \\ 0 \end{Bmatrix}$$

4. Perform the trace-back procedure from the maximum element in the table to the first zero element on the trace-back path.

3. Fill the scoring matrix using the equation below.

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ \max_{k\geq 1}\{H_{i-k,j} - W_k\}, \\ \max_{l\geq 1}\{H_{i,j-l} - W_l\}, \\ 0 \end{cases} \qquad (1 \leq i \leq n, 1 \leq j \leq m)$$

where

$H_{i-1,j-1} + s(a_i, b_j)$ is the score of aligning a_i and b_j,

$H_{i-k,j} - W_k$ is the score if a_i is at the end of a gap of length k,

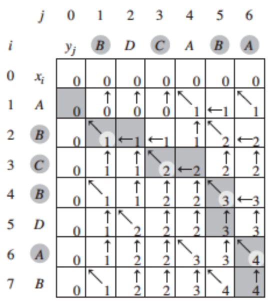
$H_{i,j-l} - W_l$ is the score if b_j is at the end of a gap of length l,

0 means there is no similarity up to a_i and b_j.

Step 3: Computing the length of a LCS

```
LCS-LENGTH(X, Y)
m = X.length
n = Y.length
let b[1..m, 1..n] and c[0..m, 0..n] be new tables
for i = 1 to m
    c[i, 0] = 0
for j = 0 to n
    c[0, j] = 0
for i = 1 to m
    for j = 1 to n
        if x_i == y_j
            c[i, j] = c[i - 1, j - 1] + 1
            b[i, j] = "↖"
        elseif c[i - 1, j] ≥ c[i, j - 1]
            c[i, j] = c[i - 1, j]
            b[i, j] = "↑"
        else c[i, j] = c[i, j - 1]
            b[i, j] = "←"
return c and b
```

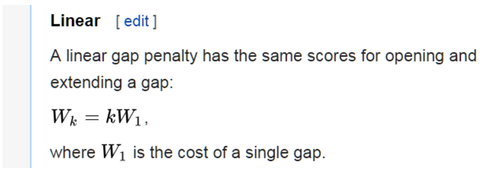
BCBA ⇦ AB C BDAB ⇦
BDCAB A



	j	0	1	2	3	4	5	6
i		y_j	B	D	C	A	B	A
0	x_i	0	0	0	0	0	0	0
1	A	0	↑0	↑0	↑0	↖1	←1	↖1
2	B	0	↖1	←1	←1	↑1	↖2	←2
3	C	0	↑1	↑1	↖2	←2	↑2	↑2
4	B	0	↖1	↑1	↑2	↑2	↖3	←3
5	D	0	↑1	↖2	↑2	↑2	↑3	↑3
6	A	0	↑1	↑2	↑2	↖3	↑3	↖4
7	B	0	↖1	↑2	↑2	↑3	↖4	↑4

Simplified Smith–Waterman algorithm

When linear gap penalty function is used
A linear gap penalty has the same scores for opening and extending a gap:



Linear [edit]

A linear gap penalty has the same scores for opening and extending a gap:

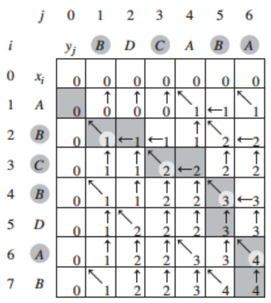
$$W_k = kW_1,$$

where W_1 is the cost of a single gap.

Step 4: Constructing a LCS (Backtracking)

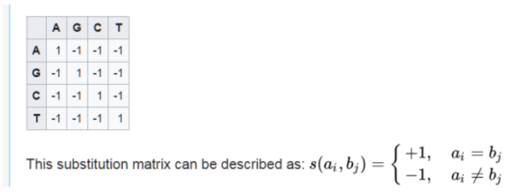
BCBA ⇦ AB C BDAB ⇦
BDCAB A

```
PRINT-LCS(b, X, i, j)
if i == 0 or j == 0
    return
if b[i, j] == "↖"
    PRINT-LCS(b, X, i - 1, j - 1)
    print x_i
elseif b[i, j] == "↑"
    PRINT-LCS(b, X, i - 1, j)
else PRINT-LCS(b, X, i, j - 1)
```



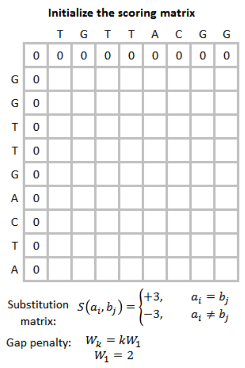
	j	0	1	2	3	4	5	6
i		y_j	B	D	C	A	B	A
0	x_i	0	0	0	0	0	0	0
1	A	0	↑0	↑0	↑0	↖1	←1	↖1
2	B	0	↖1	←1	←1	↑1	↖2	←2
3	C	0	↑1	↑1	↖2	←2	↑2	↑2
4	B	0	↖1	↑1	↑2	↑2	↖3	←3
5	D	0	↑1	↖2	↑2	↑2	↑3	↑3
6	A	0	↑1	↑2	↑2	↖3	↑3	↖4
7	B	0	↖1	↑2	↑2	↑3	↖4	↑4

Similarity Matrix



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

This substitution matrix can be described as: $s(a_i, b_j) = \begin{cases} +1, & a_i = b_j \\ -1, & a_i \neq b_j \end{cases}$

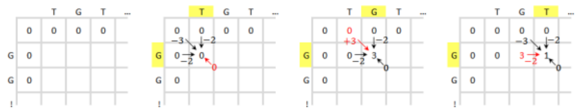


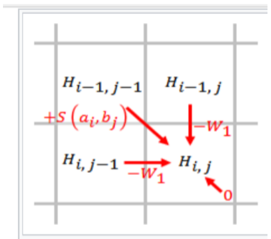
Initialize the scoring matrix

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0								
G	0								
T	0								
T	0								
G	0								
A	0								
C	0								
T	0								
A	0								

Substitution matrix: $s(a_i, b_j) = \begin{cases} +3, & a_i = b_j \\ -3, & a_i \neq b_j \end{cases}$

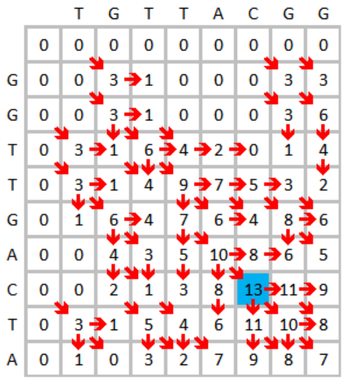
Gap penalty: $W_k = kW_1$, $W_1 = 2$





Simplified Smith–Waterman algorithm when linear gap penalty function is used

$$H_{ij} = \max \begin{cases} H_{i-1,j-1} + s(a_i, b_j), \\ H_{i-1,j} - W_1, \\ H_{i,j-1} - W_1, \\ 0 \end{cases}$$



		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

		T	G	T	T	A	C	G	G
	0	0	0	0	0	0	0	0	0
G	0	0	3	1	0	0	0	3	3
G	0	0	3	1	0	0	0	3	6
T	0	3	1	6	4	2	0	1	4
T	0	3	1	4	9	7	5	3	2
G	0	1	6	4	7	6	4	8	6
A	0	0	4	3	5	10	8	6	5
C	0	0	2	1	3	8	13	11	9
T	0	3	1	5	4	6	11	10	8
A	0	1	0	3	2	7	9	8	7

3	6	9	7	10	13
G	T	T	-	A	C
I	I	I		I	I
G	T	T	G	A	C

Dynamic Programming

	GAP	M	N	A	L	S	D	R	T
GAP	0	0	0	0	0	0	0	0	0
M	0	6	0	0	4	0	0	0	0
G	0	0	6	1	0	5	1	0	0
S	0	0	1	7	0	2	5	1	1
D	0	0	2	1	3	0	6	4	1
R	0	0	0	0	0	3	0	12	3
T	0	0	0	1	0	1	3	0	15
T	0	0	0	1	0	1	1	2	3
E	0	0	1	0	0	0	4	0	2
T	0	0	0	2	0	1	0	3	3

```
SDRT

SDRT
```

Substitution Score

Substitution matrix (BLOSUM 50 matrix)

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

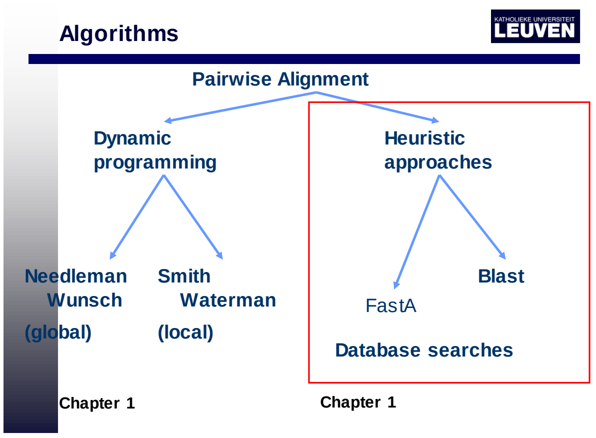
Log odds score can be positive (identities, conservative replacements) and negative

Aligning locally using BLOSUM 62

		A	A	E	E	K	K	L	A	A	A
	0	0	0	0	0	0	0	0	0	0	0
A	0	4	4	0	0	0	0	0	4	4	4
A	0	4	8	3	0	0	0	0	4	8	8
R	0	0	3	8	3	2	2	0	0	3	7
R	0	0	0	3	8	5	4	0	0	0	2
I	0	0	0	0	0	5	2	6	0	0	0
A	0	4	4	0	0	0	4	1	10	4	4

```
KKLA
RRIA
```

Score: 10



Algorithms

KATHOLIEKE UNIVERSITEIT LEUVEN

Pairwise Alignment

Dynamic programming

Heuristic approaches

Needleman Wunsch

(global)

Smith Waterman

(local)

FastA

Blast

Database searches

Chapter 1

Chapter 1