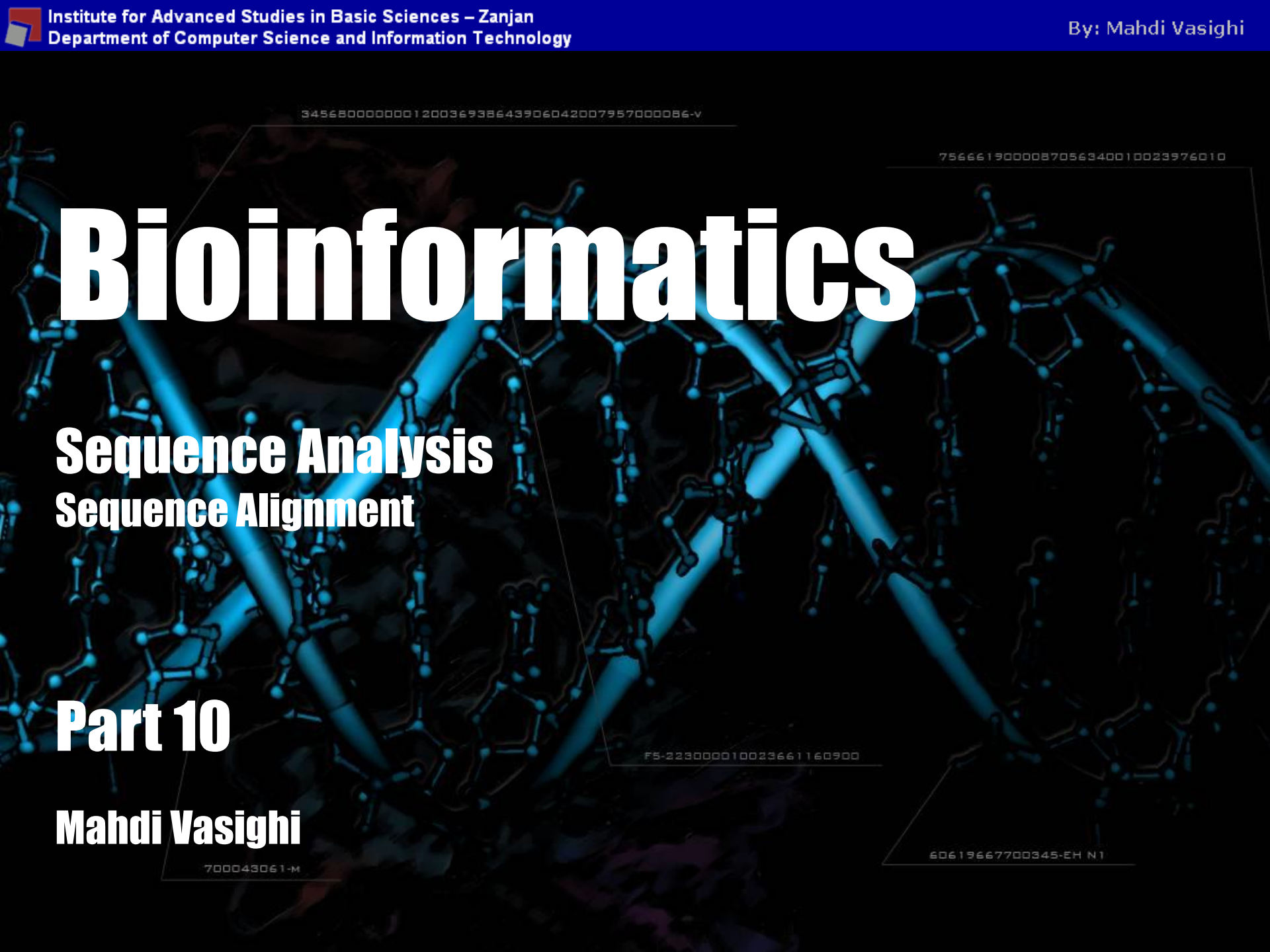


Bioinformatics

Sequence Analysis
Sequence Alignment

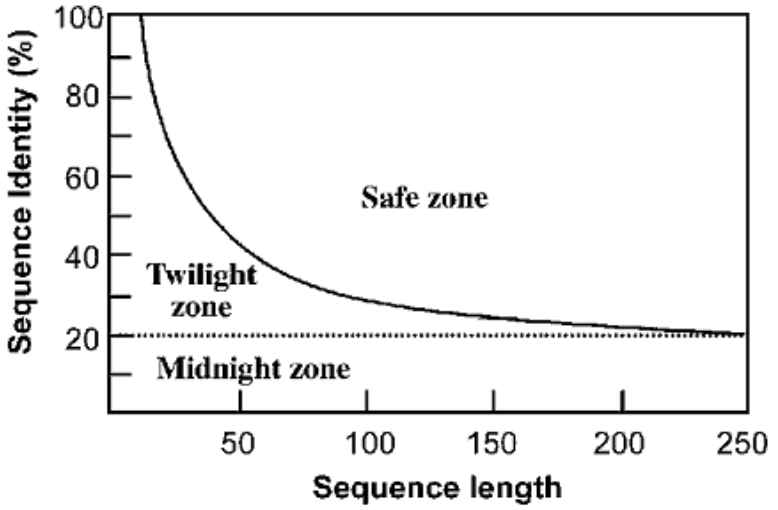
Part 10

Mahdi Vasighi

Pairwise Sequence Alignment

- An important concept in sequence analysis is sequence homology
  - A conclusion about a common ancestral relationship drawn from sequence similarity comparison
- A related but different term is sequence similarity, which is the percentage of aligned residues
  - A direct result of observation from the sequence alignment.
  - Can be quantified using percentages
  - similarity level → evolutionary relationship can be inferred

Sequence length is also a crucial factor. The shorter the sequence, the higher the chance that some alignment is attributable to random chance.



Pairwise Sequence Alignment

- Sequence similarity vs. Sequence identity
  - Synonymous for nucleotide sequences.
  - In a protein sequence alignment, sequence identity refers to the percentage of matches of the same amino acid residues.
  - Sequence similarity refers to the percentage of aligned residues that have similar physicochemical characteristics.
- There are two ways to calculate the sequence similarity/identity:

$$S = [(L_s \times 2)/(L_a + L_b)] \times 100$$

$$I = [(L_i \times 2)/(L_a + L_b)] \times 100$$

Pairwise Sequence Alignment

Dynamic programming

Dynamic programming is a computational method that can be applied only to problems exhibiting the properties of overlapping subproblems. Examples include:

- Travelling salesman problem
- finding the best chess move

The method is very important for sequence analysis because it provides the **very best** or **optimal alignment** between sequences.

Idea: Build up an optimal alignment using previous solutions for optimal alignments of smaller subsequences.

Pairwise Sequence Alignment

Dynamic programming

Global Alignment : Needleman-Wunsch technique

The following is an example of global sequence alignment using **Needleman-Wunsch** technique:

```
Seq.1:  G A A T T C A G T T A
Seq.2:  G G A T C G A
```

A simple scoring scheme is assumed:

Status	Score
Match	+1
Mismatch	0
Indel (gap)	0

Pairwise Sequence Alignment

Dynamic programming

Global Alignment : Needleman-Wunsch technique

Needleman-Wunsch technique has **three steps**:

1. Initialization
2. Matrix fill (scoring)
3. Traceback (alignment)

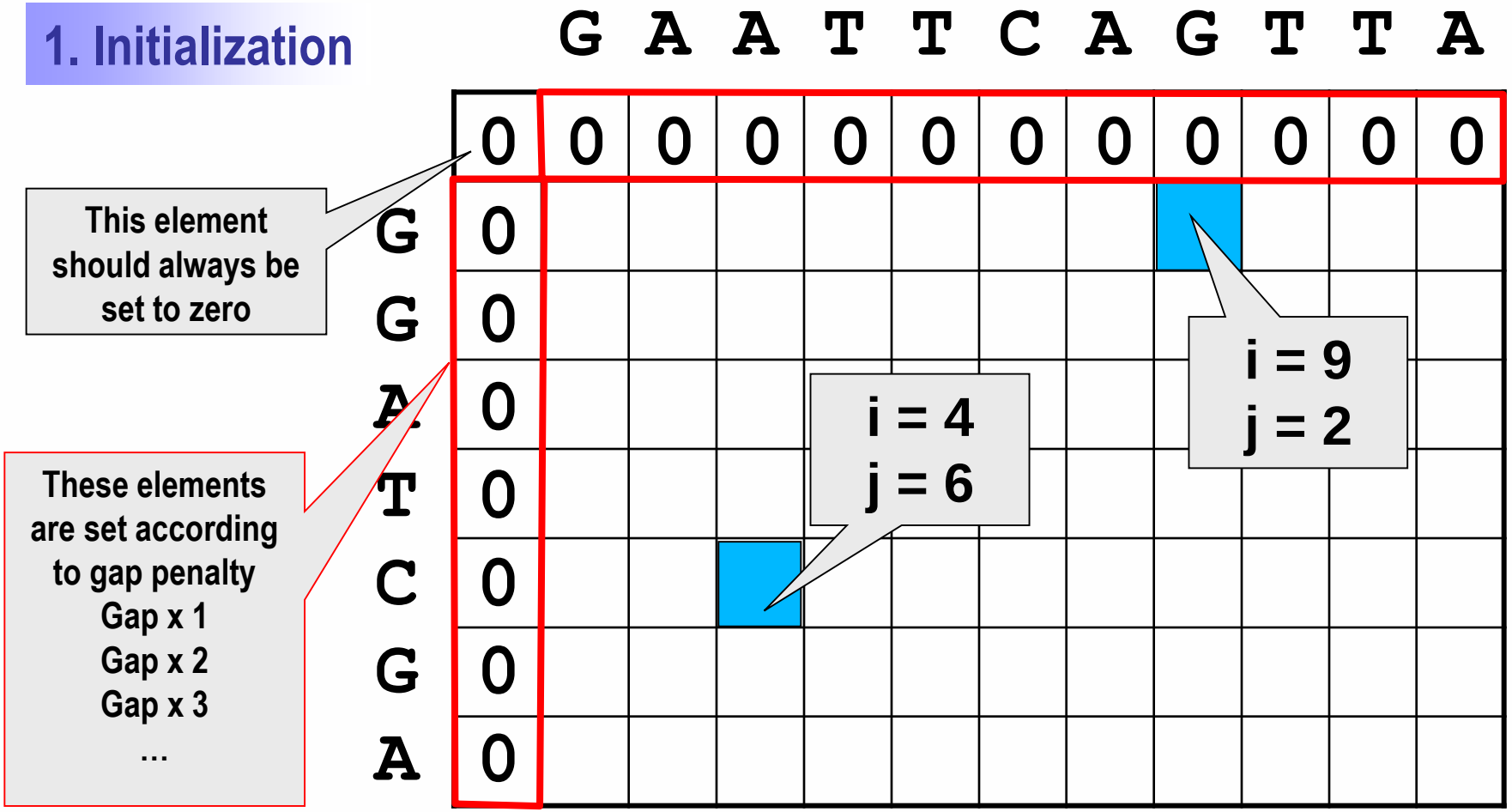
The first step in the global alignment dynamic programming approach is to create a matrix with M + 1 columns and N + 1 rows where M and N correspond to the size of the sequences to be aligned.

Pairwise Sequence Alignment

Dynamic programming

Global Alignment : Needleman-Wunsch technique

1. Initialization



Pairwise Sequence Alignment

$M_{i,j} = MAXIMUM [M_{i-1,j-1} + S_{i,j} , M_{i,j-1} + w , M_{i-1,j} + w]$

Dynamic programming

Global Alignment : Needleman-Wunsch technique

2. Matrix fill (scoring)

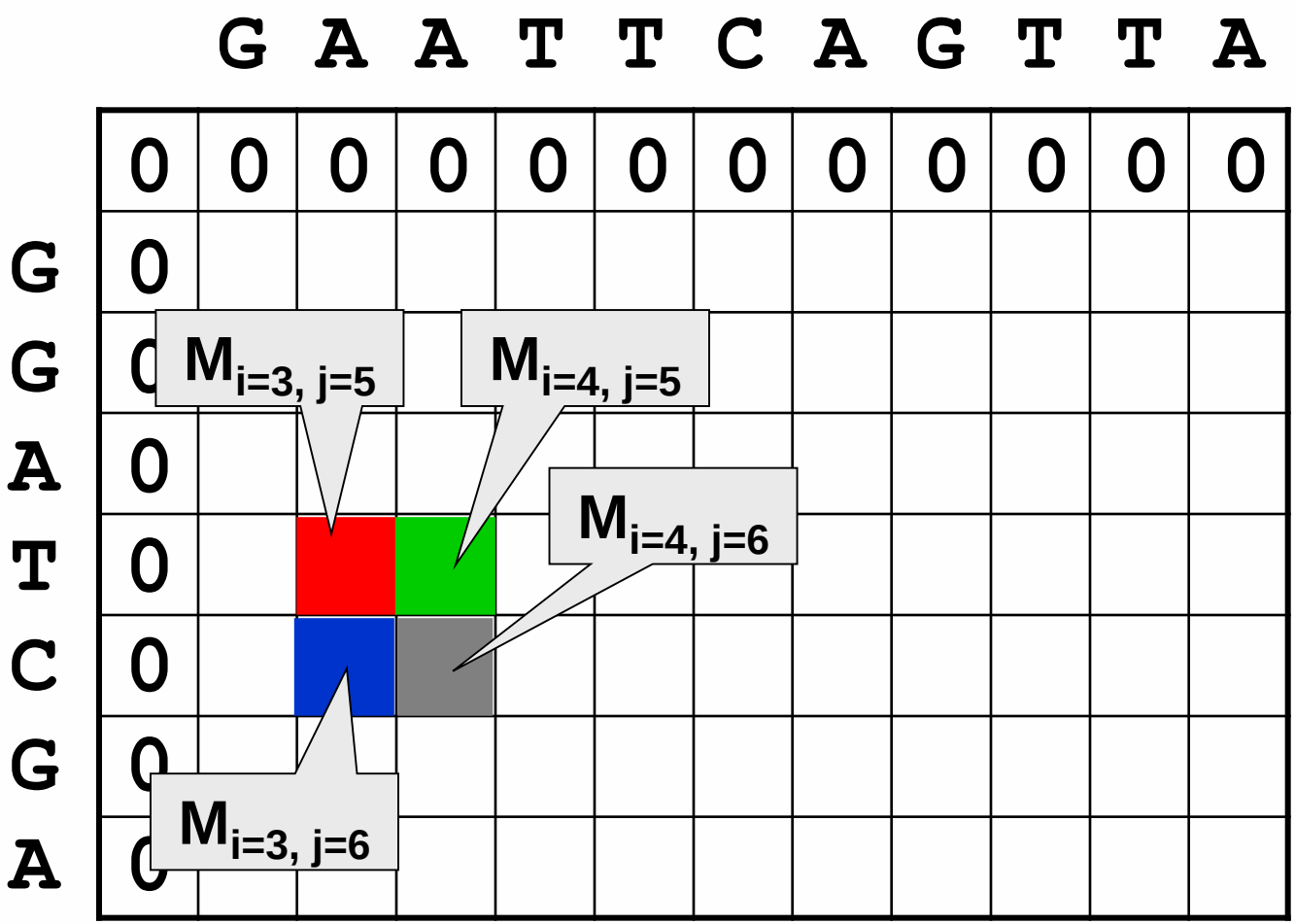
One possible solution of the matrix fill step finds the maximum global alignment score by starting in the upper left hand corner in the matrix and finding the maximal score $M_{i,j}$ for each position in the matrix.

$$M_{i,j} = MAXIMUM [M_{i-1,\,j-1} + S_{i,j} , M_{i,j-1} + w , M_{i-1,j} + w]$$

- $M_{i-1,\,j-1} + S_{i,j}$: match/mismatch in the diagonal
- $M_{i,j-1} + w$: gap in sequence #1
- $M_{i-1,j} + w$: gap in sequence #2

Global Alignment : Needleman-Wunsch technique

$$M_{i,j} = \text{MAXIMUM} [M_{i-1, j-1} + S_{i,j} , M_{i,j-1} + w , M_{i-1,j} + w]$$



Global Alignment : Needleman-Wunsch technique

$$M_{i,j} = \text{MAXIMUM} [M_{i-1, j-1} + S_{i,j} , M_{i,j-1} + w , M_{i-1,j} + w]$$

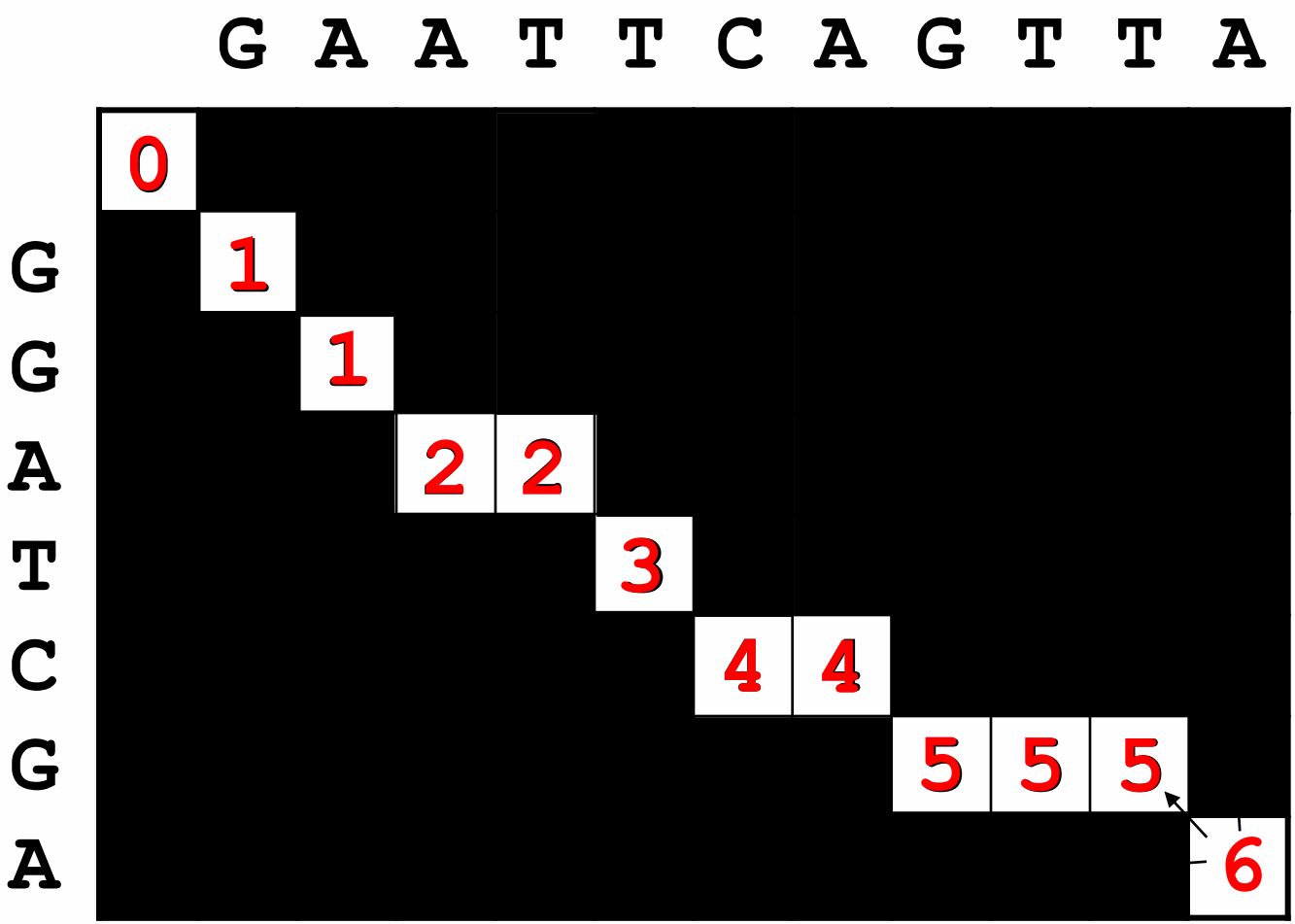
		G	A	A	T	T	C	A	G	T	T	A
	0	0	0	0	0	0	0	0	0	0	0	0
G	0	1	1	1	1	1	1	1	1	1	1	1
G	0	1	1	1								
A	0	1	2	2								
T	0	1	2	2								
C	0	1	2	2								
G	0	1	2	2								
A	0	1	2	3								

Global Alignment : Needleman-Wunsch technique

the maximum alignment score for the two test sequences is 6. The traceback step determines the actual alignment(s) that result in the maximum score.

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

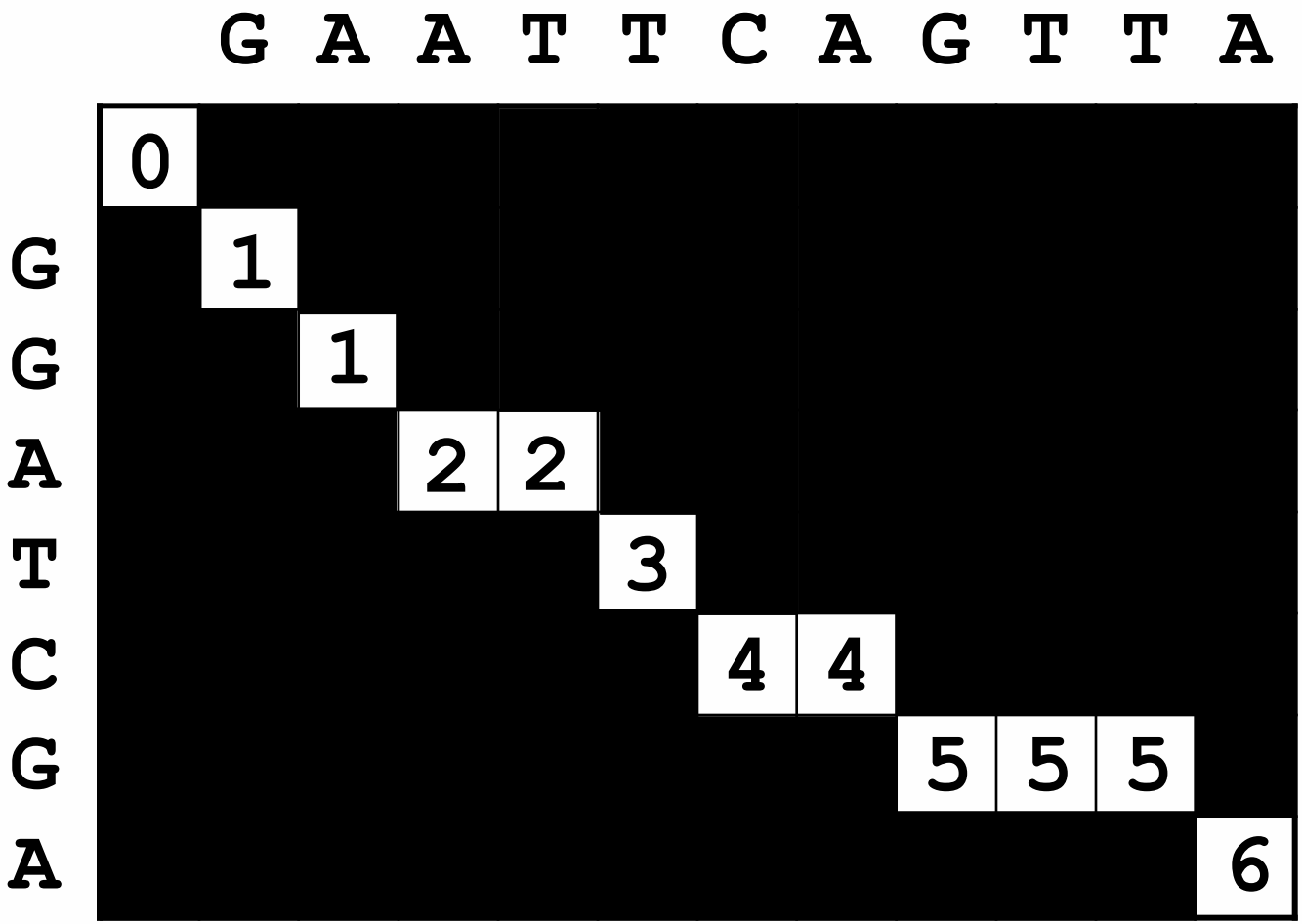
Global Alignment : Needleman-Wunsch technique



G	A	A	T	T	C	A	G	T	T	A	Seq(1)
G	G	A	-	T	C	-	G	-	-	A	Seq(2)

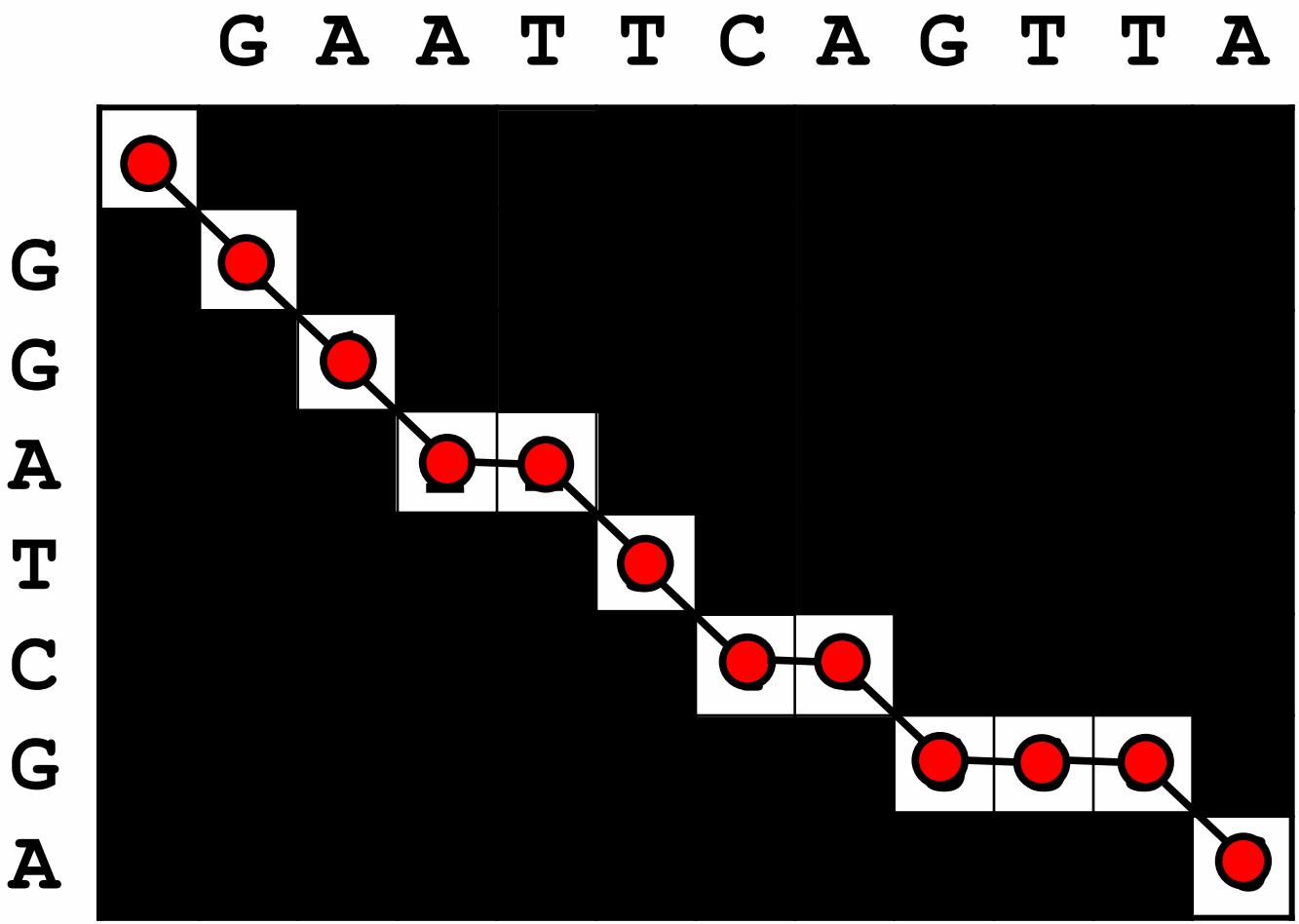
Global Alignment : Needleman-Wunsch technique

There are more alternative solutions each resulting in a maximal global alignment score of 6.



G	-	A	A	T	T	C	A	G	T	T	A	Seq(1)
G	G	-	A	-	T	C	-	G	-	-	A	Seq(2)

Global Alignment : Needleman-Wunsch technique



G	-	A	A	T	T	C	A	G	T	T	A	Seq(1)
G	G	-	A	-	T	C	-	G	-	-	A	Seq(2)

Global Alignment : Needleman-Wunsch technique

Here is another example of global sequence alignment by Needleman-Wunsch technique using a different scoring scheme:

```
Seq.1:  G A A T T C A G T T A
Seq.2:  G G A T C G A
```

The scoring scheme is:

Status	Score
Match	+2
Mismatch	-1
Indel (gap)	-2

Global Alignment : Needleman-Wunsch technique

1. Initialization

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
G	-2											
G	-4											
A	-6											
T	-8											
C	-10											
G	-12											
A	-14											

$$M_{i,j} = \text{MAXIMUM} [M_{i-1,\, j-1} + S_{i,j} , M_{i,j-1} + w , M_{i-1,j} + w]$$

Global Alignment : Needleman-Wunsch technique

2. Matrix filling

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
G	-2	2										
G	-4											
A	-6											
T	-8											
C	-10											
G	-12											
A	-14											

$$M_{i-1,\,j-1} + S_{i,j} \;,\; M_{i,j-1} + w \;,\; M_{i-1,j} + w$$

$$M_{1,1} = \text{MAXIMUM} [\; 0 \quad +2 \;,\; -2 - 2 \;,\; -2 - 2 \;]$$

Global Alignment : Needleman-Wunsch technique

2. Matrix filling

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
G	-2	2										
G	-4	0										
A	-6											
T	-8											
C	-10											
G	-12											
A	-14											

$$M_{i-1,\,j-1} + S_{i,j} \;,\; M_{i,j-1} + w \;,\; M_{i-1,j} + w$$

$$M_{1,2} = \text{MAXIMUM} [\; -2 + 2 \;,\; 2 - 2 \;,\; -4 - 2 \;]$$

Global Alignment : Needleman-Wunsch technique

2. Matrix filling

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
G	-2	2										
G	-4	0										
A	-6	-2										
T	-8											
C	-10											
G	-12											
A	-14											

$$M_{i-1,\,j-1} + S_{i,j} \;,\; M_{i,j-1} + w \;,\; M_{i-1,j} + w$$

$$M_{1,3} = \text{MAXIMUM} [\; -4 \quad -1 \;,\; 0 \quad -2 \;,\; -6 -2 \;]$$

Global Alignment : Needleman-Wunsch technique

2. Matrix filling

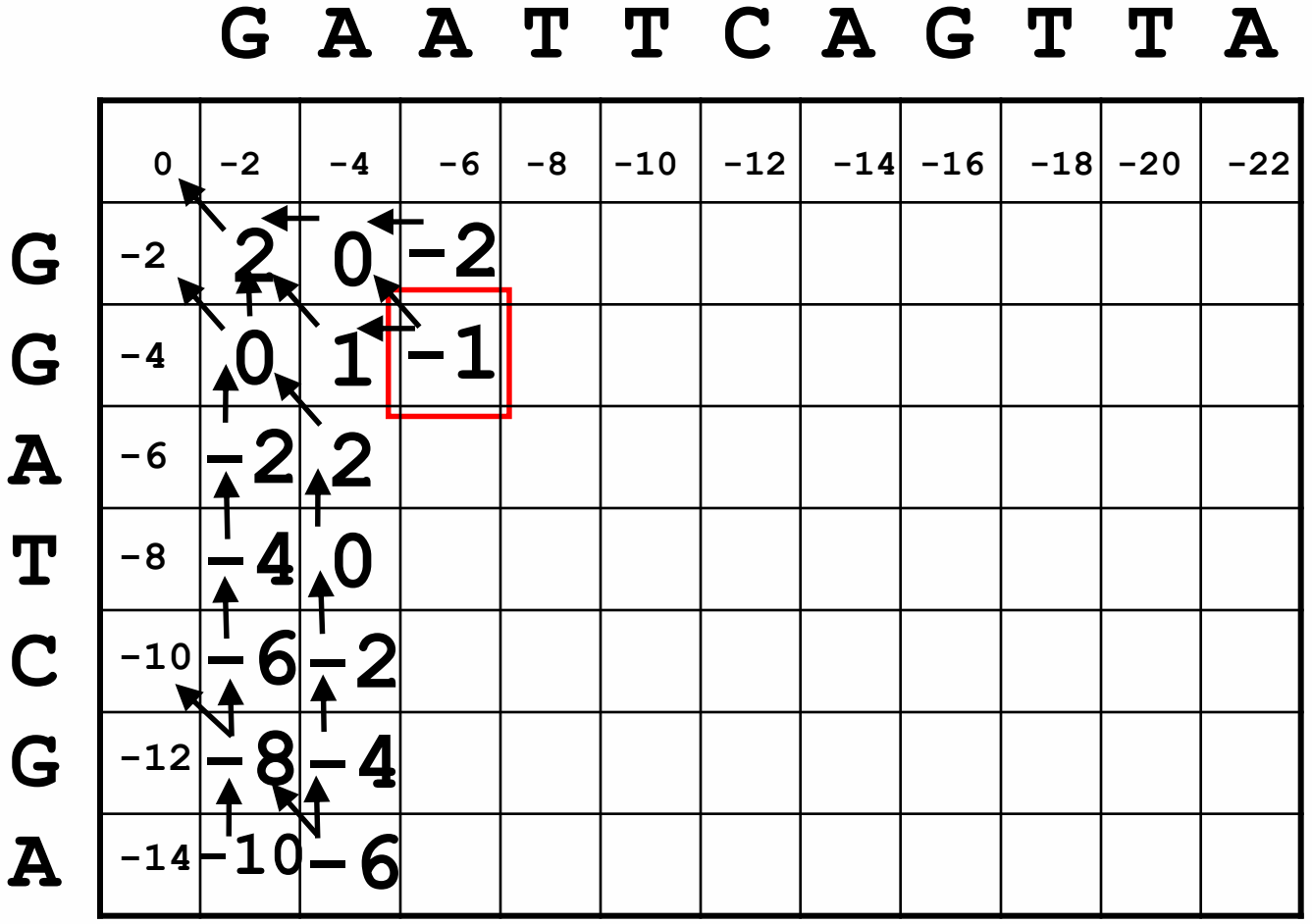
		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
G	-2	-2										
G	-4	0										
A	-6	-2										
T	-8	-4										
C	-10											
G	-12											
A	-14											

$$M_{i-1,\,j-1} + S_{i,j} \;,\; M_{i,j-1} + w \;,\; M_{i-1,j} + w$$

$$M_{1,4} = \text{MAXIMUM} [\; -6 - 1 \;,\; -2 - 2 \;,\; -8 - 2 \;]$$

Global Alignment : Needleman-Wunsch technique

2. Matrix filling



$M_{i-1,\,j-1} + S_{i,j}$, $M_{i,j-1} + w$, $M_{i-1,j} + w$

$M_{3,2}$ = MAXIMUM [0 -1 , −2 -2 , 1 -2]

Global Alignment : Needleman-Wunsch technique

2. Matrix filling

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
G	-2	2	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	-4	0	1	-1	-3	-5	-7	-9	-8	-10	-12	-14
A	-6	-2	2	3	1	-1	-3	-5	-7	-9	-11	-10
T	-8	-4	0	1	5	3	1	-1	-3	-5	-7	-9
C	-10	-6	-2	-1	3	4	5	3	1	-1	-3	-5
G	-12	-8	-4	-3	1	2	3	4	5	3	1	-1
A	-14	-10	-6	-2	1	0	1	5	3	4	2	3

After the matrix fill step, the maximum global alignment score for the two sequences is 3.

Global Alignment : Needleman-Wunsch technique

3. Traceback

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
G	-2	2	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	-4	0	1	-1	-3	-5	-7	-9	-8	-10	-12	-14
A	-6	-2	2	3	1	-1	-3	-5	-7	-9	-11	-10
T	-8	-4	0	1	5	3	1	-1	-3	-5	-7	-9
C	-10	-6	-2	-1	3	4	5	3	1	-1	-3	-5
G	-12	-8	-4	-3	1	2	3	4	5	3	1	-1
A	-14	-10	-6	-2	1	0	1	5	3	4	2	3

Global Alignment : Needleman-Wunsch technique

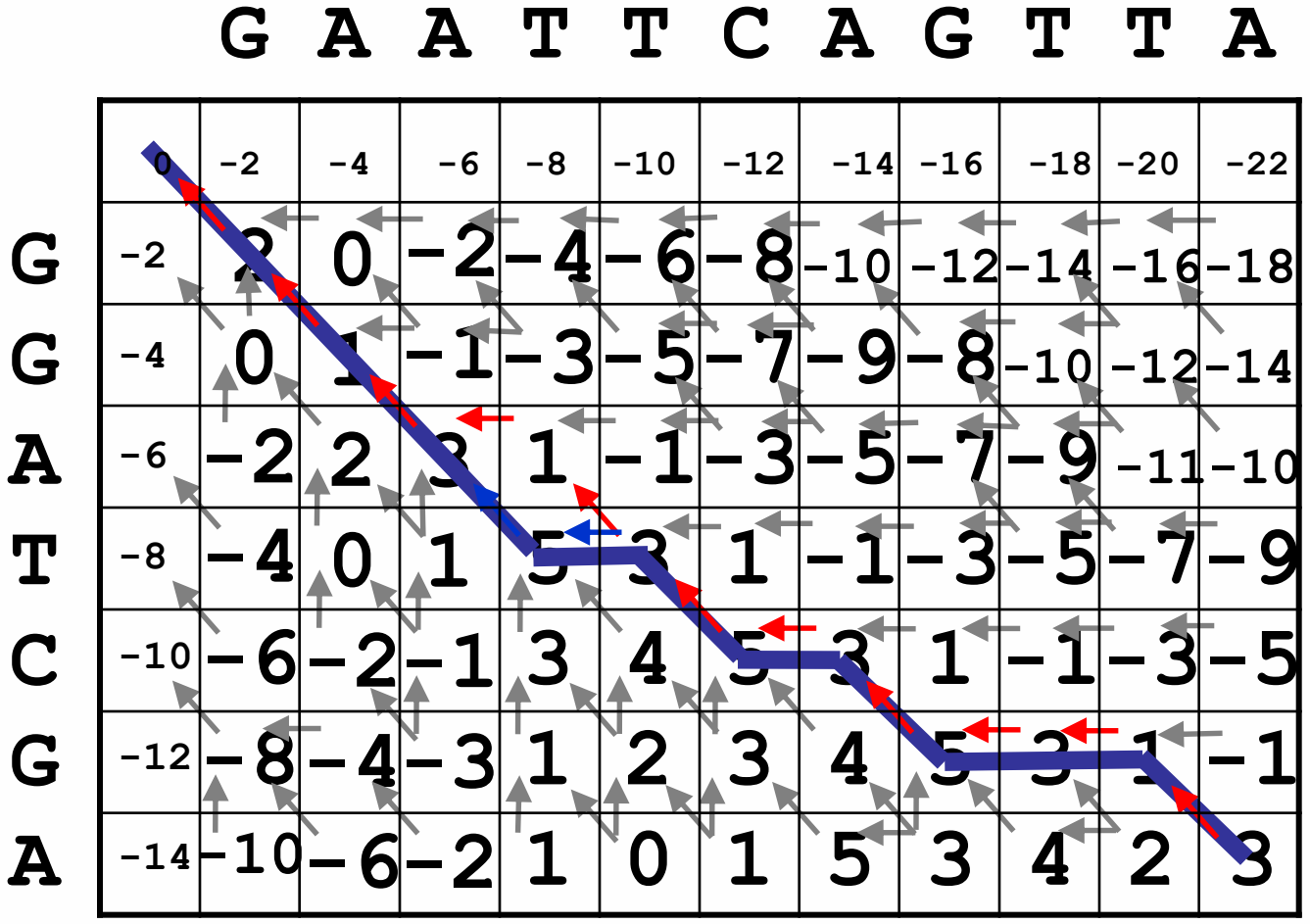
3. Traceback

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
G	-2	2	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	-4	0	1	-1	-3	-5	-7	-9	-8	-10	-12	-14
A	-6	-2	2	3	1	-1	-3	-5	-7	-9	-11	-10
T	-8	-4	0	1	5	3	1	-1	-3	-5	-7	-9
C	-10	-6	-2	-1	3	4	5	3	1	-1	-3	-5
G	-12	-8	-4	-3	1	2	3	4	5	3	1	-1
A	-14	-10	-6	-2	1	0	1	5	3	4	2	3

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

Global Alignment : Needleman-Wunsch technique

3. Traceback



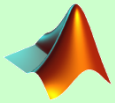
G	A	A	T	T	C	A	G	T	T	A
G	G	A	T	-	C	-	G	-	-	A

Global Alignment : Needleman-Wunsch technique

		G	A	A	T	T	C	A	G	T	T	A
	0	-2	-4	-6	-8	-10	-12	-14	-16	-18	-20	-22
G	-2	2	0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	-4	0	1	-1	-3	-5	-7	-9	-8	-10	-12	-14
A	-6	-2	2	3	1	-1	-3	-5	-7	-9	-11	-10
T	-8	-4	0	1	5	3	1	-1	-3	-5	-7	-9
C	-10	-6	-2	-1	3	4	5	3	1	-1	-3	-5
G	-12	-8	-4	-3	1	2	3	4	5	3	1	-1
A	-14	-10	-6	-2	1	0	1	5	3	4	2	3

Global Alignment : Needleman-Wunsch technique

nwalign



Purpose: Globally align two sequences using the Needleman-Wunsch algorithm

Syntax:

```
[Score, Alignment] = nwalign(Seq1, Seq2, 'PropertyName', PropertyValue)
```

Values are 'PAM40', 'PAM250', 'BLOSUM30' 'BLOSUM62', or 'BLOSUM100'.

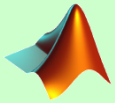
```
nwalign(...,'ScoringMatrix', ScoringMatrixValue)
nwalign(...,'Alphabet', AlphabetVlaue)
```

Property to select the type of sequence. Value is either'AA' or 'NT'. The default value is 'AA'.

```
nwalign(...,'Showscore', ShowscoreValue)
```

Global Alignment : Needleman-Wunsch technique

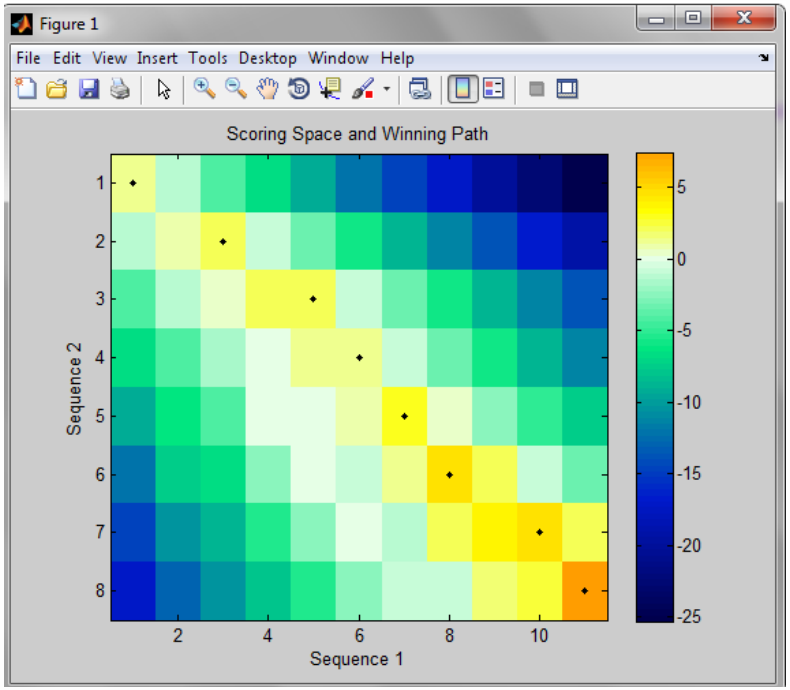
nwalign



Example

```
[Score, Alignment] = nwalign('IGRHRYHIGG','SRYIGRG',...
                             'scoringmatrix','pam250')

[Score, Alignment] = nwalign('VSPAGMASGYD','IPGKASYD','Showscore',1)
```

```
Score =

    7.3333


Alignment =

VSPAGMASGYD
:  |  |  ||  ||
I-P-GKAS-YD
```



Pairwise Sequence Alignment

Dynamic programming

Local Alignment : Smith-Waterman technique

The following is an example of local sequence alignment using **Smith-Waterman** technique:

```
Seq.1:  A T C A G A G T C
Seq.2:  G T C A G T C A
```

A simple scoring scheme is assumed:

Status	Score
Match	+2
Mismatch	-1
Indel (gap)	-2

Local Alignment : Smith-Waterman technique

1. Initialization

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

Local Alignment : Smith-Waterman technique

2. Matrix fill (scoring)

The main difference to the Needleman-Wunsch algorithm is that negative scoring matrix cells are set to zero, which renders the (thus positively scoring) local alignments visible.

$$M_{i,j} = \text{MAXIMUM}\left[M_{i-1,\,j-1} + S_{i,j},\ M_{i,j-1} + w,\ M_{i-1,j} + w,\ 0 \right]$$

- $M_{i-1,\,j-1} + S_{i,j}$: match/mismatch in the diagonal
- $M_{i,j-1} + w$: gap in sequence #1
- $M_{i-1,j} + w$: gap in sequence #2

$$H_{i,j} = \text{MAXIMUM}\left[\underbrace{M_{i-1,j-1} + S_{i,j}}_{0} , \underbrace{M_{i,j-1} + w}_{-1} , \underbrace{M_{i-1,j} + w}_{0} , \underbrace{0}_{-2} \right]$$

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

$$H_{i,j} = \text{MAXIMUM}\,[\; M_{i-1,\,j-1} + S_{i,j} \;,\; M_{i,j-1} + w \;,\; M_{i-1,j} + w \;]$$

$$H_{i,j} = \text{MAXIMUM}\,[\; 0 + 2 \;,\; 0 - 2 \;,\; 0 - 2 \;,\; 0 \;]$$

		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	0	0							
T	0	0	2							
C	0	0								
A	0	2								
G	0	0								
T	0	0								
C	0	0								
A	0	2								

$$H_{i,j} = \text{MAXIMUM}\,[\; 0 \quad M_{i-1,\,j-1} + S_{i,j} \quad M_{i,j-1} + w \quad M_{i-1,j} + w \;]$$

$$H_{i,j} = \text{MAXIMUM}\,[\; 0 \quad -1 \quad 2 \quad -2 \quad 0 \quad -2 \quad 0 \;]$$

		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	0	0							
T	0	0	2							
C	0	0	0							
A	0	2	0							
G	0	0								
T	0	0								
C	0	0								
A	0	2								

$$H_{i,j} = \text{MAXIMUM}\,[\; M_{i-1,\,j-1} + S_{i,j} \;,\; M_{i,j-1} + w \;,\; M_{i-1,j} + w \;,\; 0 \;]$$

$$H_{i,j} = \text{MAXIMUM}\,[\; 2 \;,\; -1 \;,\; 0 \;,\; -2 \;,\; 0 \;,\; -2 \;,\; 0 \;]$$

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

To obtain the optimum local alignment, we start with the highest value in the matrix (i,j). Then, we go backwards to one of positions (i-1,j), (i,j-1), and (i-1,j-1) depending on the direction of movement used to construct the matrix. We keep the process until we reach a matrix cell with zero value

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

A T C A G A G T C

G T C A G T C A

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

A T C A G A G T C

G T C A G T C A

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

A T C A G A G T C

G T C A G - - T C A

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

Finding the sub-optimal alignments

1. fix optimal alignment positions to zero

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

Finding the sub-optimal alignments

2. Recalculate the matrix elements again

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

Finding the sub-optimal alignments

2. Recalculate the matrix elements again

		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	2	0	2	0	0
T	0	0	0	0	0	0	1	0	4	2
C	0	0	0	0	0	0	0	0	2	6
A	0	2	0	0	0	0	2	0	0	4
G	0	0	1	0	0	0	0	0	0	2
T	0	0	2	0	0	0	0	0	0	0
C	0	0	0	4	2	0	0	0	0	0
A	0	2	0	2	6	4	2	0	0	0

Finding the sub-optimal alignments

2. Recalculate the matrix elements again

		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	2	0	2	0	0
T	0	0	0	0	0	0	1	0	4	2
C	0	0	0	0	0	0	0	0	2	6
A	0	2	0	0	0	0	2	0	0	4
G	0	0	1	0	0	0	0	0	0	2
T	0	0	2	0	0	0	0	0	0	0
C	0	0	0	4	2	0	0	0	0	0
A	0	2	0	2	6	4	2	0	0	0

→ New values

A T C A G A G T C

G T C A G T C A

		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	2	0	2	0	0
T	0	0	0	0	0	0	1	0	4	2
C	0	0	0	0	0	0	0	0	2	6
A	0	2	0	0	0	0	2	0	0	4
G	0	0	1	0	0	0	0	0	0	2
T	0	0	2	0	0	0	0	0	0	0
C	0	0	0	4	2	0	0	0	0	0
A	0	2	0	2	6	4	2	0	0	0

A T C A G A G T C

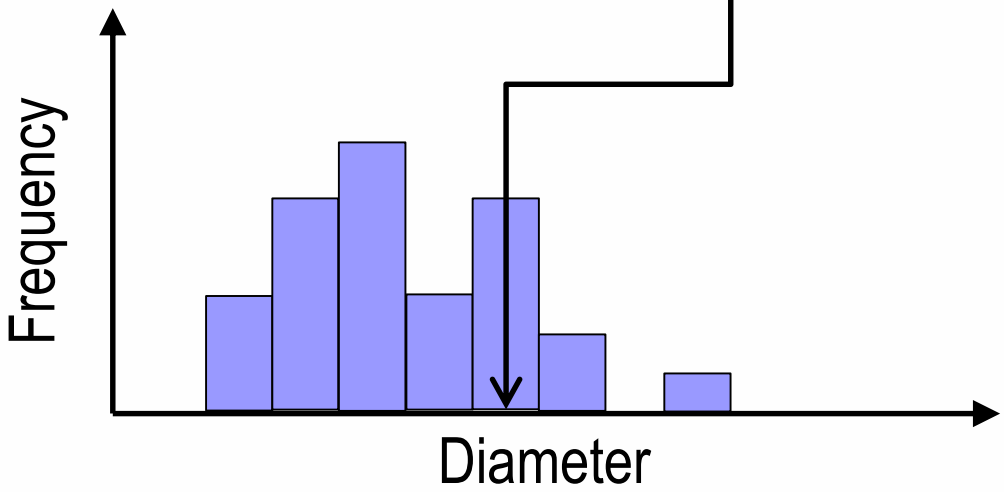
G T C A G T C A

		A	T	C	A	G	A	G	T	C
	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	2	0	2	0	0
T	0	0	0	0	0	0	1	0	4	2
C	0	0	0	0	0	0	0	0	2	6
A	0	2	0	0	0	0	2	0	0	4
G	0	0	1	0	0	0	0	0	0	2
T	0	0	2	0	0	0	0	0	0	0
C	0	0	0	4	2	0	0	0	0	0
A	0	2	0	2	6	4	2	0	0	0

Significance of Sequence Alignments

Assessing significance requires a *distribution*

I have an apple of diameter 10 cm. Is that unusual?



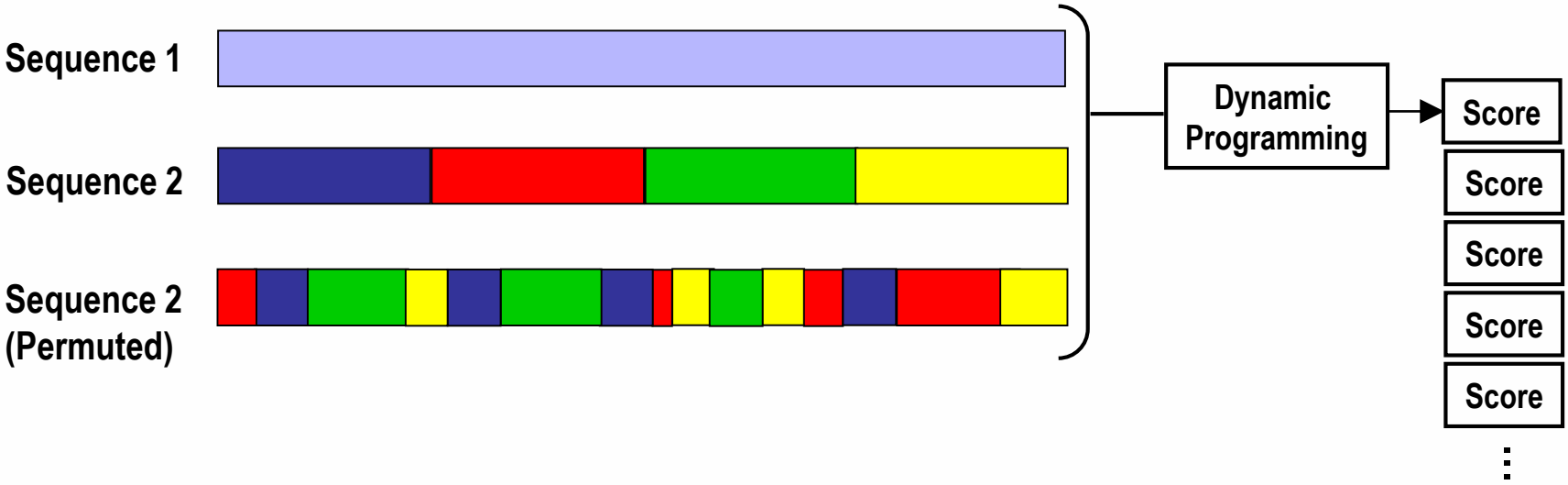
Significance of Sequence Alignments

- When given a sequence alignment showing a certain degree of similarity, it is often important to ask whether the observed sequence alignment can occur by random chance or the alignment is indeed statistically sound.
- The truly statistically significant sequence alignment will be able to provide evidence of homology between the sequences involved.
- Solving this problem requires a distribution of the alignment scores of two unrelated sequences of the same length.

Significance of Sequence Alignments

A practical approach to the problem is as follows:

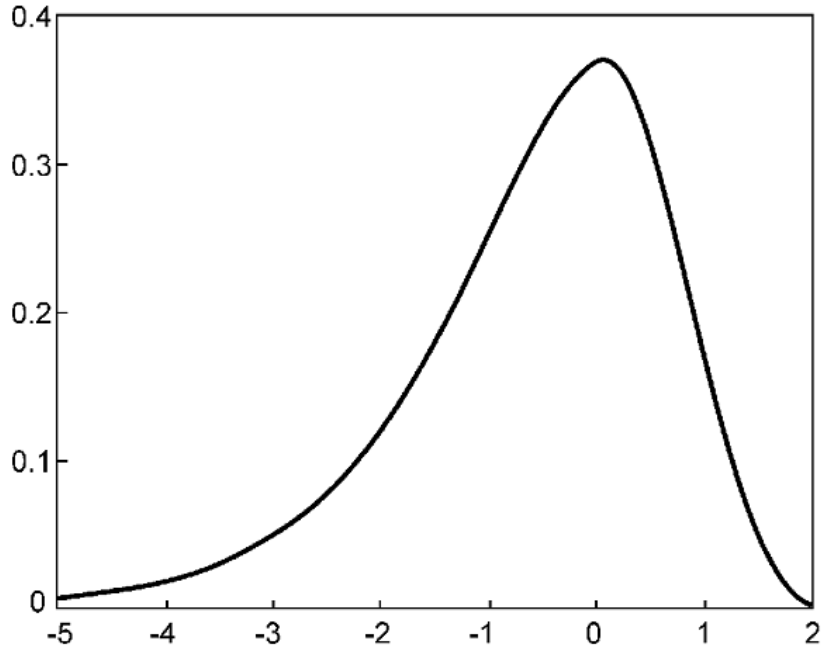
We may randomize one of the sequences, many times, re-align each result to the second sequence (held fixed), and collect the distribution of resulting scores.



Significance of Sequence Alignments

Many studies have demonstrated that the distribution of similarity scores assumes a peculiar shape that resembles a highly skewed normal distribution with a long tail on one side

Gumble extreme value distribution



The distribution can be expressed as:

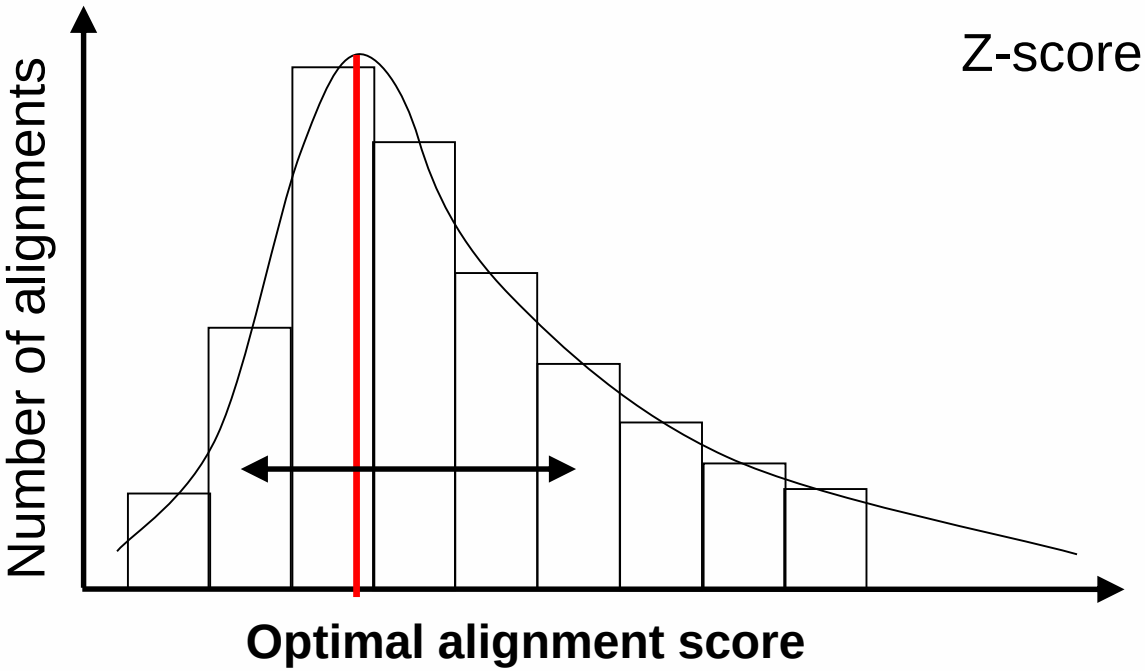
$$P = 1 - e^{-Kmn e^{-\lambda x}}$$

Where **m** and **n** are the sequence lengths, **λ** is a scaling factor for the scoring matrix used, and **K** is a constant that depends on the scoring matrix and gap penalty combination that is used.

A **P-value** is given to indicate the probability that the original alignment is due to random chance.

Significance of Sequence Alignments

A population of alignment score will be obtained and z-score can be calculated as follow:



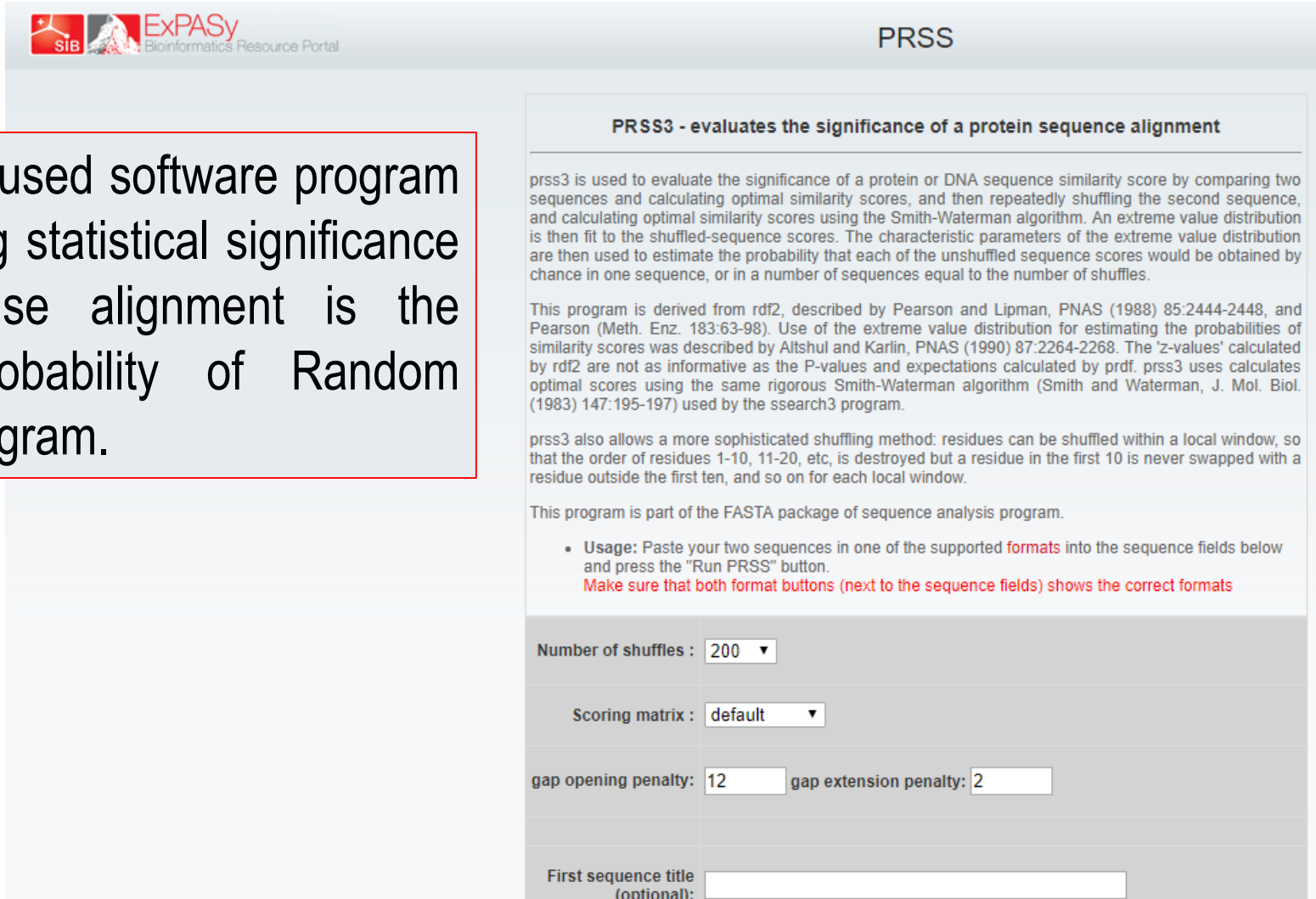
$$\text{Z-score} = \frac{\text{Score - mean}}{\text{Standard deviation}}$$

- **The Z-score reflects the extent to which the original result is an outlier from the population.**
- **Experience suggests that Z-scores ≥ 5 are significant.**

Significance of Sequence Alignments

http://www.ch.embnet.org/software/PRSS_form.html

A frequently used software program for assessing statistical significance of a pairwise alignment is the **PRSS3** (Probability of Random Shuffles) program.



1. Globally align the **AK2H_ECOLI** and **AK3_ECOLI** using the toolbox and calculate similarity and identity..

2. How could the **divide and conquer** method be used to solve the global sequence alignment more efficiently?