Pairwise Sequence Alignment

Outline

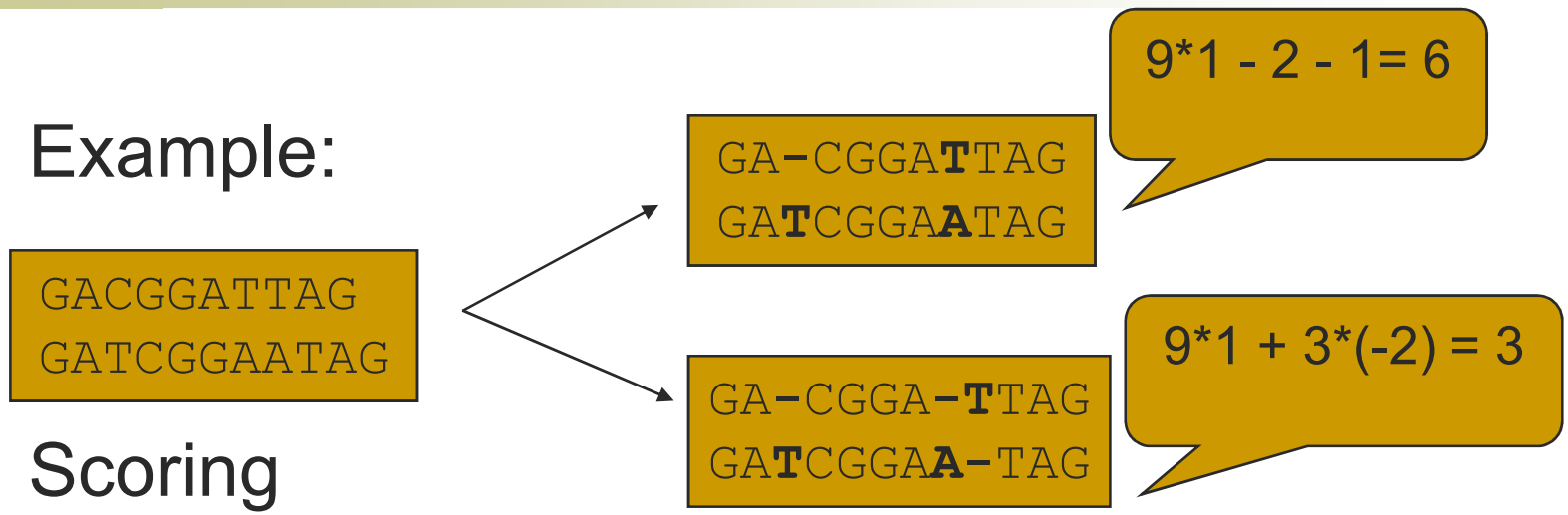
- Introduction
- Global Alignment
  - The Basic Algorithm
- Local Alignment
- Semiglobal Alignment

Introduction

- Sequence similarity is an indicator of homology
- There are other uses for sequence similarity
  - Database queries
  - Comparative genomics
  - ...

Introduction

- Example:



- Scoring
  - Match +1
  - mismatch -1
  - Gap penalty -2

Introduction

- The sequences may have different sizes.
- We define an alignment as the insertion of spaces in arbitrary locations along the sequences so that they end up with the same size.
- In general , there may be many alignments with maximum score.

Different alignment

- Global Alignment

- Local Alignment

- Semiglobal Alignment

Global Alignment

- Using Dynamic Programming
  - Reuses the results of previous computations
- Example (two sequence x , y):

  x : AGC

  y : AAAC
- Scoring function:
  - Match +1
  - mismatch -1
  - Gap penalty -2

Global Alignment

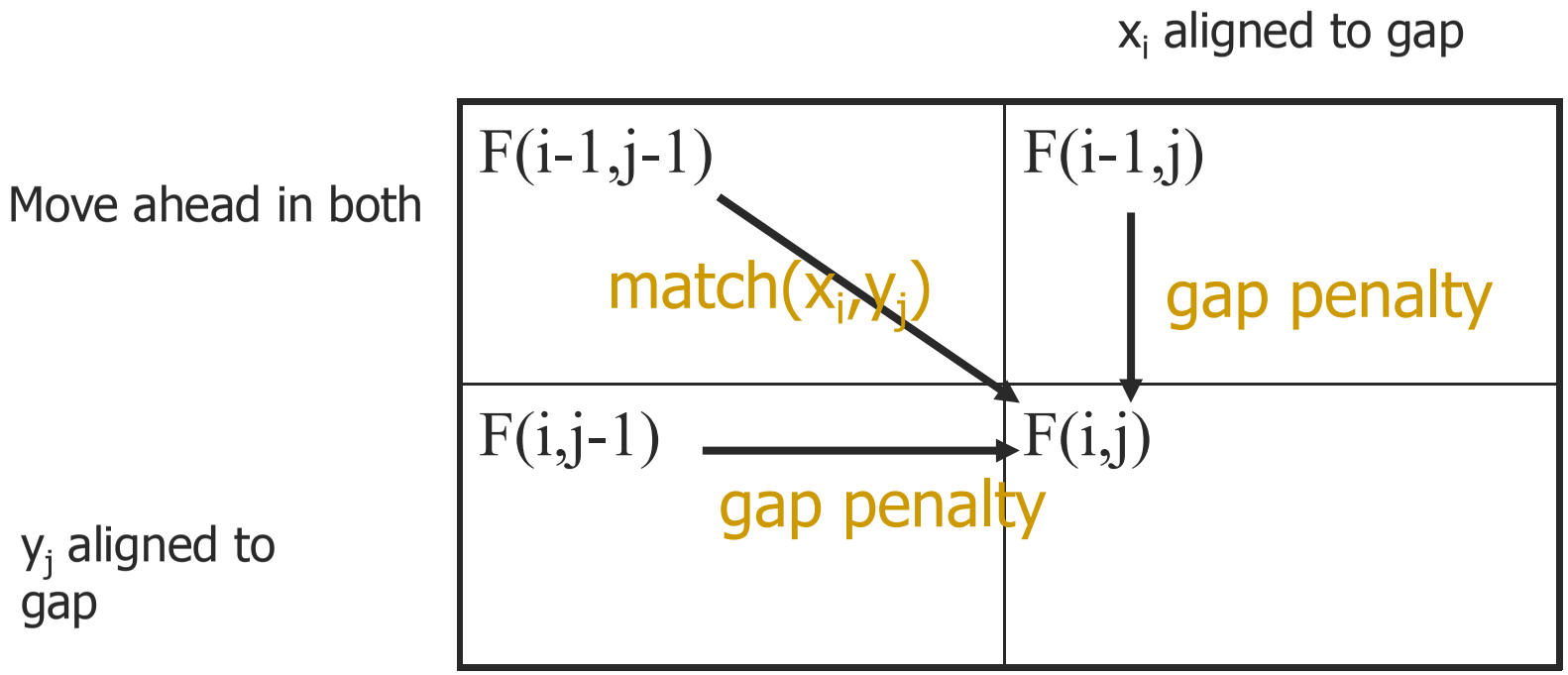
Step 1 : forming a matrix F(i,j)

y_j →

x_i ↓

	–	A	A	A	C
–	0	-2	-4	-6	-8
A	-2				
G	-4				
C	-6				

Global Alignment



While building the table, keep track of where optimal score came from, reverse arrows

Global Alignment

- $F(i, j) = \max \text{ of } \begin{cases} F(i-1,j-1) + \text{match}(x_i, y_j) \\ F(i-1,j) + \text{gap penalty} \\ F(i,j-1) + \text{gap penalty} \end{cases}$

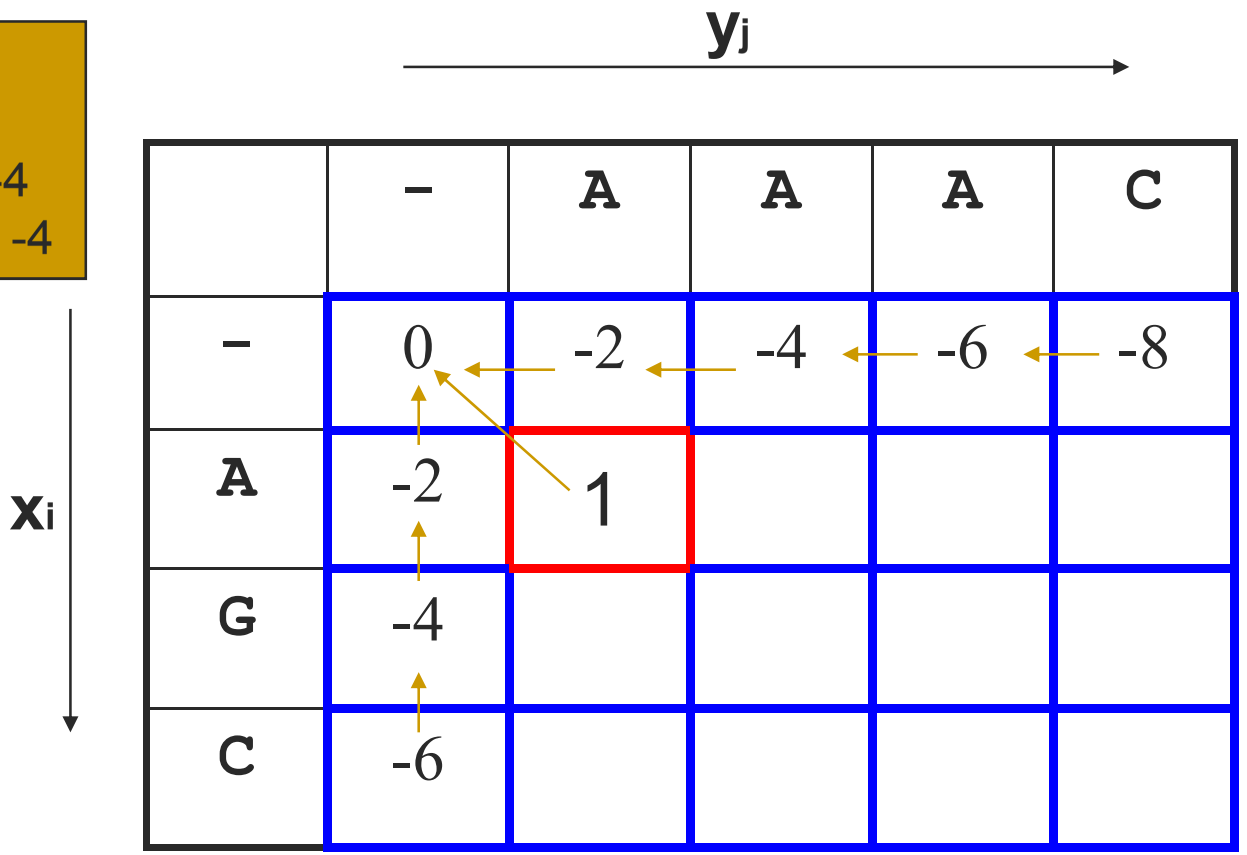
Global Alignment

- Seq1:AAAC
- Seq2:AGC

	–	A	A	A	C
–	0	-2	-4	-6	-8
A	-2				
G	-4				
C	-6				

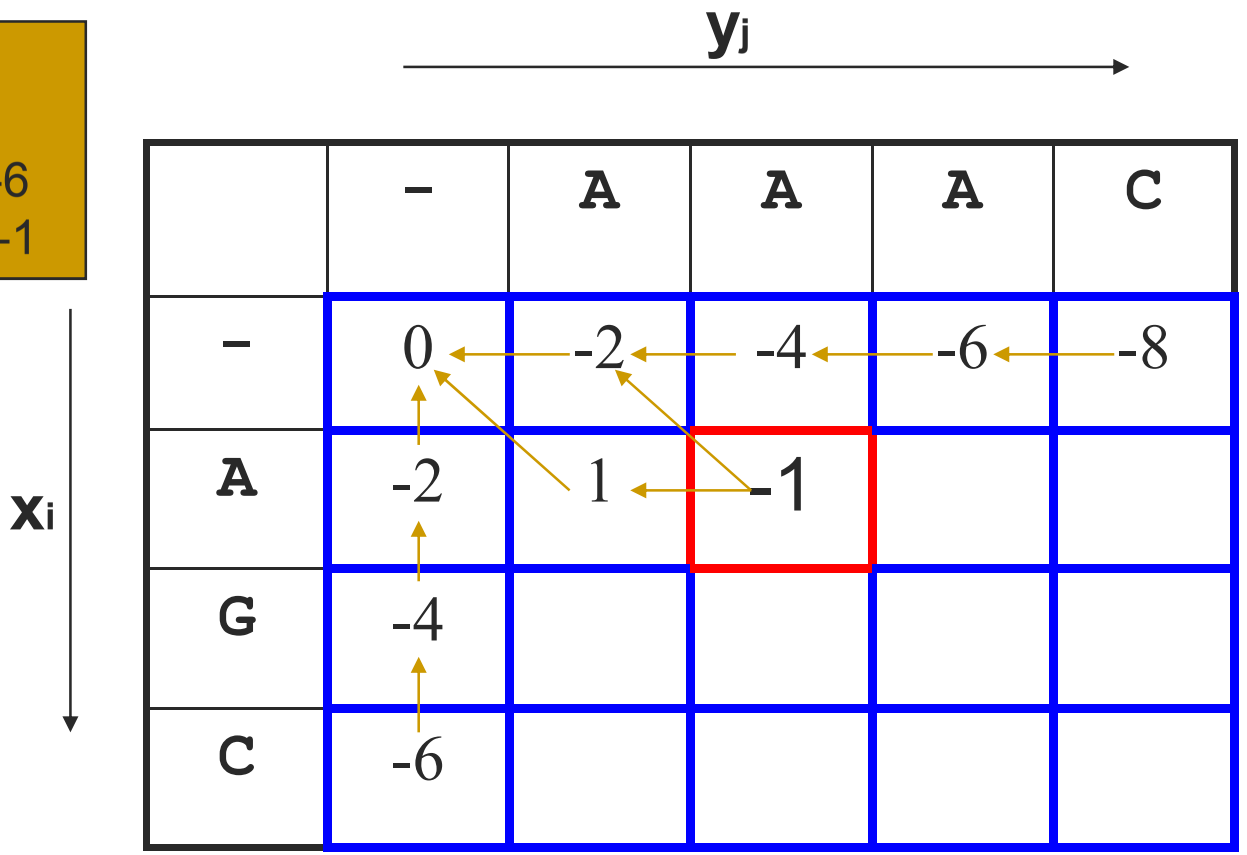
Global Alignment

F(1,1) :
↘ : 0 + 1 = 1
↓ : -2 + (-2) = -4
→ : -2 + (-2) = -4

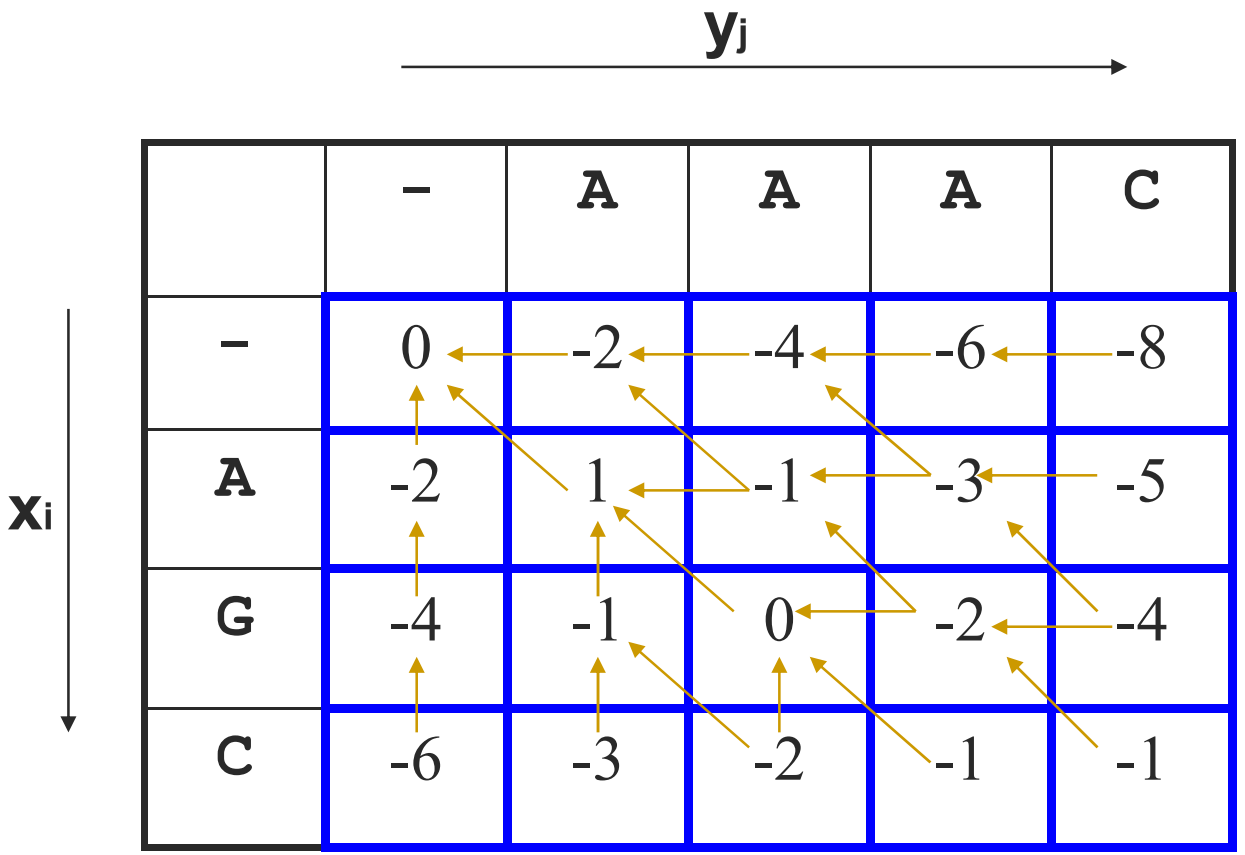


Global Alignment

F(1,2) :
↘ : -2 + 1 = -1
↓ : -4 + (-2) = -6
→ : 1 + (-2) = -1



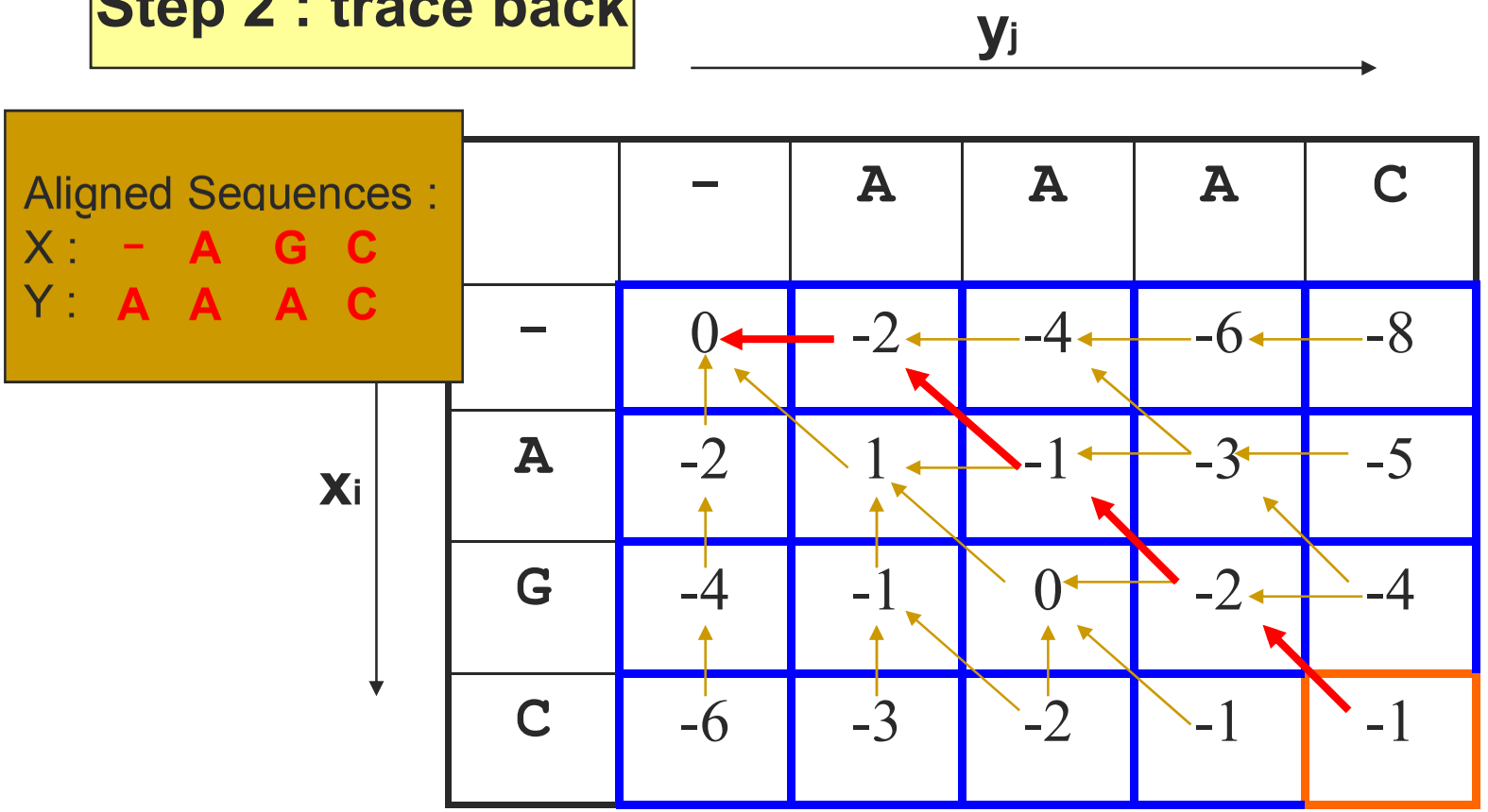
Global Alignment



Global Alignment

Step 2 : trace back

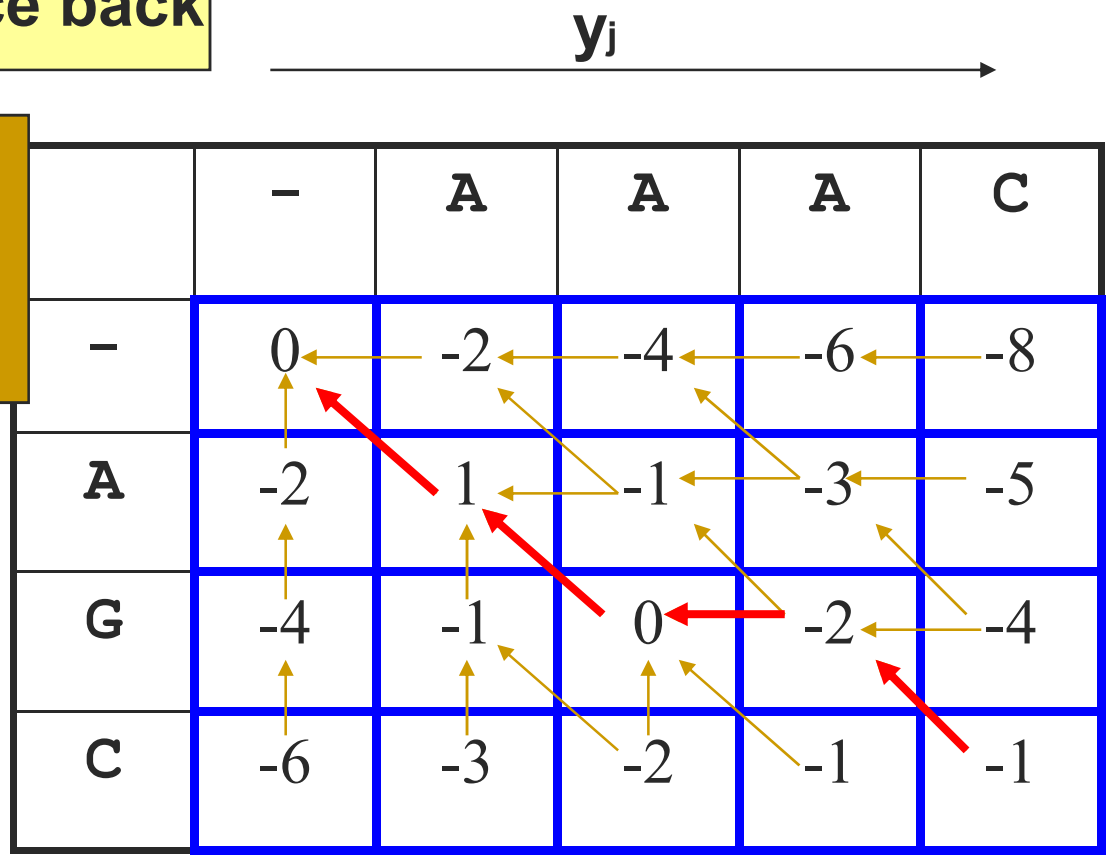
Aligned Sequences :
X : – A G C
Y : A A A C



Global Alignment

Step 2 : trace back

Aligned Sequences :
X : A G – C
Y : A A A C



Global Alignment

- Arrow preference

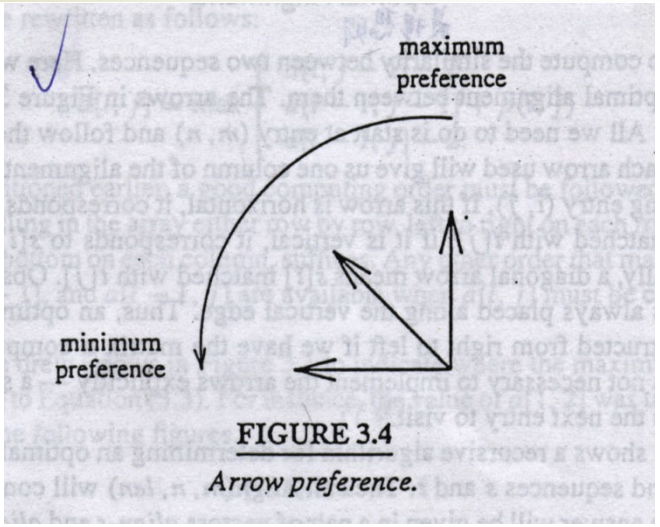


FIGURE 3.4
Arrow preference.

- For instance, when aligning x=ATAT with y=TATA , we get

```
- A T A T        rather than        A T A T -
T A T A -                          - T A T A
```

Global Alignment

- Summary
  - Uses recursion to fill in intermediate results table
  - Uses O(nm) space and time
    - $O(n^2)$ algorithm
    - Feasible for moderate sized sequences, but not for aligning whole genomes.

Local Alignment

- A local alignment between *x* and *y* is an alignment between a substring of *x* and a substring of *y*.
- $F(i, j) = \max \text{ of } \begin{cases} 0 \\ F(i-1,j-1) + \text{match}(x_i, y_j) \\ F(i-1,j) + \text{gap penalty} \\ F(i,j-1) + \text{gap penalty} \end{cases}$

Local Alignment

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

Local Alignment

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

Local Alignment

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

Local Alignment

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

Local Alignment

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

Local Alignment

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

Local Alignment

Local alignment :
G A A G T
G A C G T

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

Local Alignment

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

Semiglobal Alignment

- we score alignments ignoring some of the *end gaps* in the sequences.
  - Example:

```
CAGCA-CTTGGATTCTCGG
---CAGCCTGG--------
```

Observe that this is not the best alignment. Below we present another alignment with a higher score(-12 against -19).

```
CAGCACTTGGATTCTCGG
CAGC-----G-T----GG
```

Semiglobal Alignment

	-	C	G	A	A	G	T	T	G
-	0	0	0	0	0	0	0	0	0
G	−2								
A	−4								
C	−6								
G	−8								

Semiglobal Alignment

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

Semiglobal Alignment

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

Semiglobal Alignment

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

Semiglobal Alignment

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

Semiglobal Alignment

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

CGAAGTTG
-GACG---

Summary

- Global Alignment

- Local Alignment

- Semiglobal Alignment