

Sequence Comparison: Dynamic Programming

Genome 373

Genomic Informatics

Elhanan Borenstein

GAATC

CATAC

Mission:

**Find the best alignment
between two sequences.**

A “search” algorithm for
finding the alignment
with the best score



- Dynamic programming

A method for
scoring
alignments



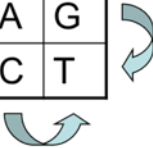
- Substitution matrix
- Gap penalties

Scoring Aligned Bases

- **Substitution matrix:**

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Purine	A	G
Pyrimidine	C	T



- **Gap penalty:**

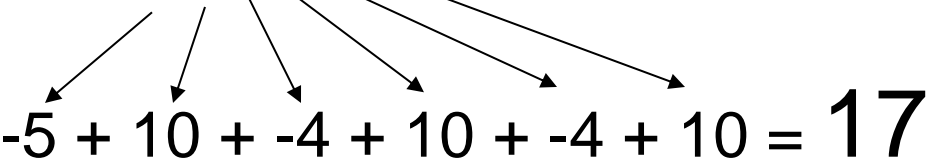
- **Linear** gap penalty
- **Affine** gap penalty



GAAT-C

d = -4

CA-TAC



$-5 + 10 + -4 + 10 + -4 + 10 = 17$

Exhaustive search

- *Align the two sequences: GAATC and CATAC*

GAATC	GAAT-C	-GAAT-C
CATAC	C-ATAC	C-A-TAC
GAATC-	GAAT-C	GA-ATC
CA-TAC	CA-TAC	CATA-C

Simple (exhaustive search) algorithm

- 1) *Construct all possible alignments*
- 2) *Use the substitution matrix and gap penalty to score each alignment*
- 3) *Pick the alignment with the best score*

How many possibilities?

GAATC

GAAT-C

-GAAT-C

CATAC

C-ATAC

C-A-TAC

GAATC-

GAAT-C

GA-ATC

CA-TAC

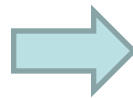
CA-TAC

CATA-C

- How many different possible alignments of two sequences of length n exist?

$$\binom{2n}{n} = \frac{(2n)!}{(n!)^2}$$

2n choose n
the binomial coefficient



5	2.5×10^2
10	1.8×10^5
20	1.4×10^{11}
30	1.2×10^{17}
40	1.1×10^{23}

The Needleman–Wunsch Algorithm

- An algorithm for **global alignment** on two sequences
- A **Dynamic Programming (DP)** approach
 - Yes, it's a weird name.
 - DP is closely related to recursion and to mathematical induction
- We can prove that the resulting score is optimal.

DP matrix

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

$j \rightarrow$ 0 1 2 3 etc.

$i \downarrow$		G	A	A	T	C
0						
1	C					
2	A					
3	T					
4	A					
5	C					

DP matrix

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

j → 0 1 2 3 etc.

i ↓

0

1

2

3

4

5

G

A

A

T

C

C

A

T

A

C

initial row and column

Best alignment
of GA to CA

DP matrix

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d=-4$

j → 0 1 2 3 etc.

i ↓		G	A	A	T	C
0						
1	C					
2	A		5			
3	T					
4	A					
5	C					

The value at (i, j) is the score of the best alignment of the first i characters of one sequence versus the first j characters of the other sequence.

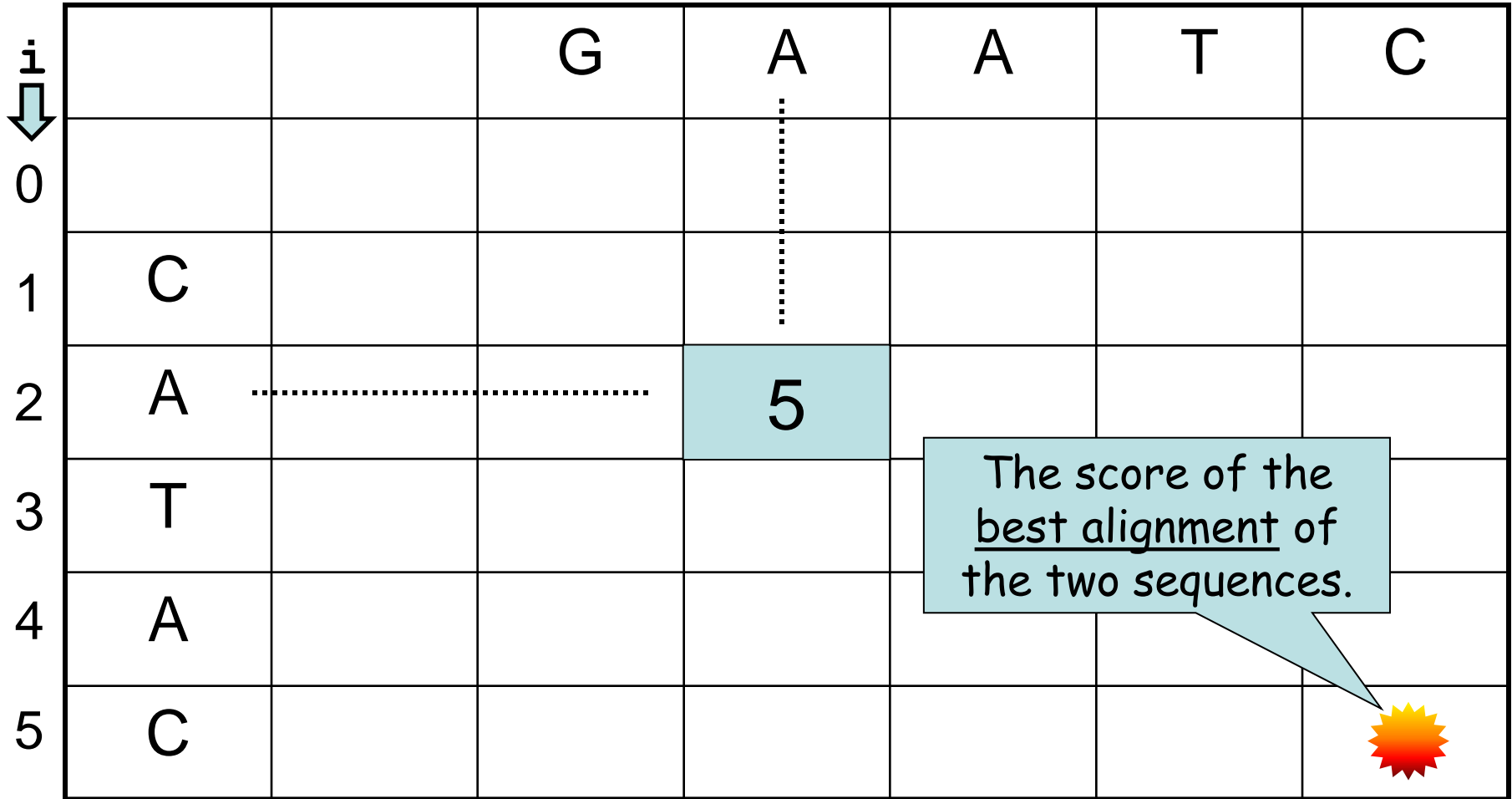
DP matrix

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

$j \rightarrow$ 0 1 2 3 etc.



Moving in the DP matrix

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
C						
A			5			
T						
A						
C						

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

DP matrix

GAA
CA-

		G	A	A	T	C
C						
A			5	1		
T						
A						
C						

Moving horizontally in the matrix introduces a gap in the sequence along the left edge.

GA-
CAT

DP matrix

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
C						
A						
T						
A						
C						

Moving vertically in the matrix introduces a gap in the sequence along the top edge.

5
↓
1

GAA
CAT

DP matrix

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
C						
A			5			
T					0	
A						
C						

Moving diagonally in the matrix aligns two residues

Start at top left and
move progressively

Initialization

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
	0					
C						
A						
T						
A						
C						

G
-

Introducing a gap

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
		0 → -4				
C						
A						
T						
A						
C						

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

Introducing a gap

-
C

		G	A	A	T	C
		0 → -4				
C		-4				
A						
T						
A						
C						

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

Complete first row and column

CATAC

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4					
A	-8					
T	-12					
A	-16					
C	-20					

What about $i=1, j=1$

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d=-4$

		$j \rightarrow$	0	1	2	3	etc.	
$i \downarrow$				G	A	A	T	C
	0		0	-4	-8	-12	-16	-20
	1	C	-4	?				
	2	A	-8					
	3	T	-12					
	4	A	-16					
	5	C	-20					

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d=-4$

Three ways to get
to $i=1, j=1$

G-

-C

j \Rightarrow 0 1 2 3 etc.

i		G	A	A	T	C	
0		0	-4	-8	-12	-16	-20
1	C	-4	-8				
2	A	-8					
3	T	-12					
4	A	-16					
5	C	-20					

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d=-4$

Three ways to get
to $i=1, j=1$

-G

C-

j



0

1

2

3 etc.

i

0

1

2

3

4

5

G

A

A

T

C

0

-4

-8

-12

-16

-20

-4

-8

-8

-12

-16

-20

C

A

T

A

C

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

Three ways to get
to $i=1, j=1$

G
C

j → 0 1 2 3 etc.

i ↓		G	A	A	T	C	
0		0	-4	-8	-12	-16	-20
1	C	-4	-5				
2	A	-8					
3	T	-12					
4	A	-16					
5	C	-20					

Accept the highest scoring
of the three

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5				
A	-8					
T	-12					
A	-16					
C	-20					

Then simply repeat the
same rule progressively
across the matrix

DP matrix

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C	
		0	-4	-8	-12	-16	-20
C		-4	-5				
A		-8	?				
T		-12					
A		-16					
C		-20					

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

DP matrix

~~G-~~
~~CA~~

-G
CA

~~--G~~
~~CA-~~

$-5 + -4 = -9$

$-4 + 0 = -4$

$-8 + -4 = -12$

		G	A	A	T	C	
		0	-4	-8	-12	-16	-20
C	-4	-5					
A	-8	?					
T	-12						
A	-16						
C	-20						

0 → -4 → -8 → -12 → -16 → -20

-4 ↓ -5
0

-8 → -4 × ?

?

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

DP matrix

~~G-~~
~~CA~~

-G
CA

~~--G~~
~~CA-~~

$-5 + -4 = -9$

$-4 + 0 = -4$

$-8 + -4 = -12$

		G	A	A	T	C	
		0	-4	-8	-12	-16	-20
C	-4	-5					
A	-8	-4					
T	-12						
A	-16						
C	-20						

Arrows from (0,0) to (0,1), (0,1) to (0,2), (0,2) to (0,3), (0,3) to (0,4), (0,4) to (0,5).
 Arrows from (0,0) to (1,1), (0,1) to (2,2).
 Red 'X' marks on (1,2) and (2,1).
 Red 'X' with '-4' on (1,2).
 Red 'X' with '-4' on (2,1).
 Teal '-4' on (2,2).

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

DP matrix

		G	A	A	T	C	
		0	-4	-8	-12	-16	-20
C		-4	-5				
A		-8	-4				
T		-12	?				
A		-16	?				
C		-20	?				

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

DP matrix

		G	A	A	T	C	
		0	-4	-8	-12	-16	-20
C		-4	-5				
A		-8	-4				
T		-12	-8				
A		-16	-12				
C		-20	-16				

Arrows indicate the path of the optimal alignment from the top-left cell (0) to the bottom-right cell (-16):

- 0 → -4 (right)
- 0 → -4 (down)
- 4 → -5 (down-right)
- 4 → -8 (down)
- 8 → -4 (down-right)
- 8 → -12 (down)
- 12 → -8 (down-right)
- 12 → -16 (down)
- 16 → -20 (down)
- 16 → -12 (down-right)
- 16 → -16 (down)

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d=-4$

DP matrix

		G	A	A	T	C	
		0	-4	-8	-12	-16	-20
C	-4	-5	?				
A	-8	-4	?				
T	-12	-8	?				
A	-16	-12	?				
C	-20	-16	?				

Traceback

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C	
		0	-4	-8	-12	-16	-20
C	-4	-5	-9				
A	-8	-4	5				
T	-12	-8	1				
A	-16	-12	2				
C	-20	-16	-2				

What is the alignment associated with this entry?

Just follow the arrows back - this is called the **traceback**

-G-A
CATA

Full Alignment

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9			
A	-8	-4	5			
T	-12	-8	1			
A	-16	-12	2			
C	-20	-16	-2			?

Continue and find the optimal global alignment, and its score.

Full Alignment

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Full Alignment

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	2	-6	-7	-7	-6
A	-8	-7	-7	-7	-7	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Best alignment starts at bottom right and follows traceback arrows to top left

GA-ATC
CATA-C

One best traceback

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C	
		0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6	
A	-8	-4	5	1	-3	-7	
T	-12	-8	1	0	11	7	
A	-16	-12	2	11	7	6	
C	-20	-16	-2	7	11	17	

GAAT-C
-CATAC

Another best traceback

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

GAAT-C
-CATAAC

GA-ATC
CATA-C

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5	-9	-13	-12	-6
A	-8	-4	5	1	-3	-7
T	-12	-8	1	0	11	7
A	-16	-12	2	11	7	6
C	-20	-16	-2	7	11	17

Multiple solutions

GA-ATC
CATA-C

GAAT-C
CA-TAC

GAAT-C
C-ATAC

GAAT-C
-CATAC

- When a program returns a single sequence alignment, it may not be the only best alignment but it is guaranteed to be one of them.
- In our example, all of the alignments at the left have equal scores.

Practice problem:

Find a best pairwise alignment of **GAATC** and **AATTC**

Substitution matrix

	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

Gap penalty: $d = -4$

		G	A	A	T	C
		0				
A						
A						
T						
T						
C						

DP in equation form

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5				
A	-8	-5	-4			
T	-12					
A	-16					
C	-20					

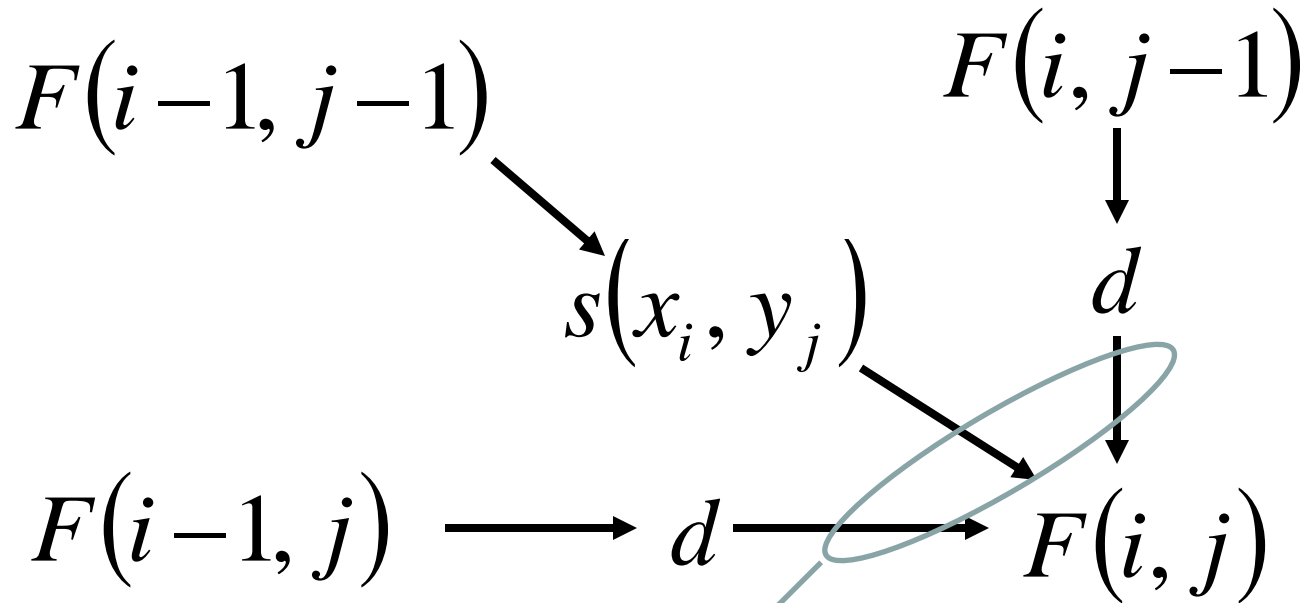
- Align sequence **x** and **y**.
- **F** is the DP matrix; **s** is the substitution matrix; **d** is the linear gap penalty.

$$F(0,0) = 0$$

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \end{cases}$$

DP equation graphically

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4	-5				
A	-8	-4	?			
T	-12					
A	-16					
C	-20					



take the max
of these three

