

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

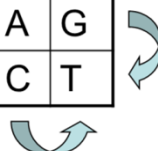
- Substitution matrix
- Gap penalties

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



- **Substitution matrix:**

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



GAAT-C

d=-4

CA-TAC

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

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?

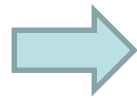
How many possibilities?

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}

FYI for two sequences of length m and n , possible alignments number:

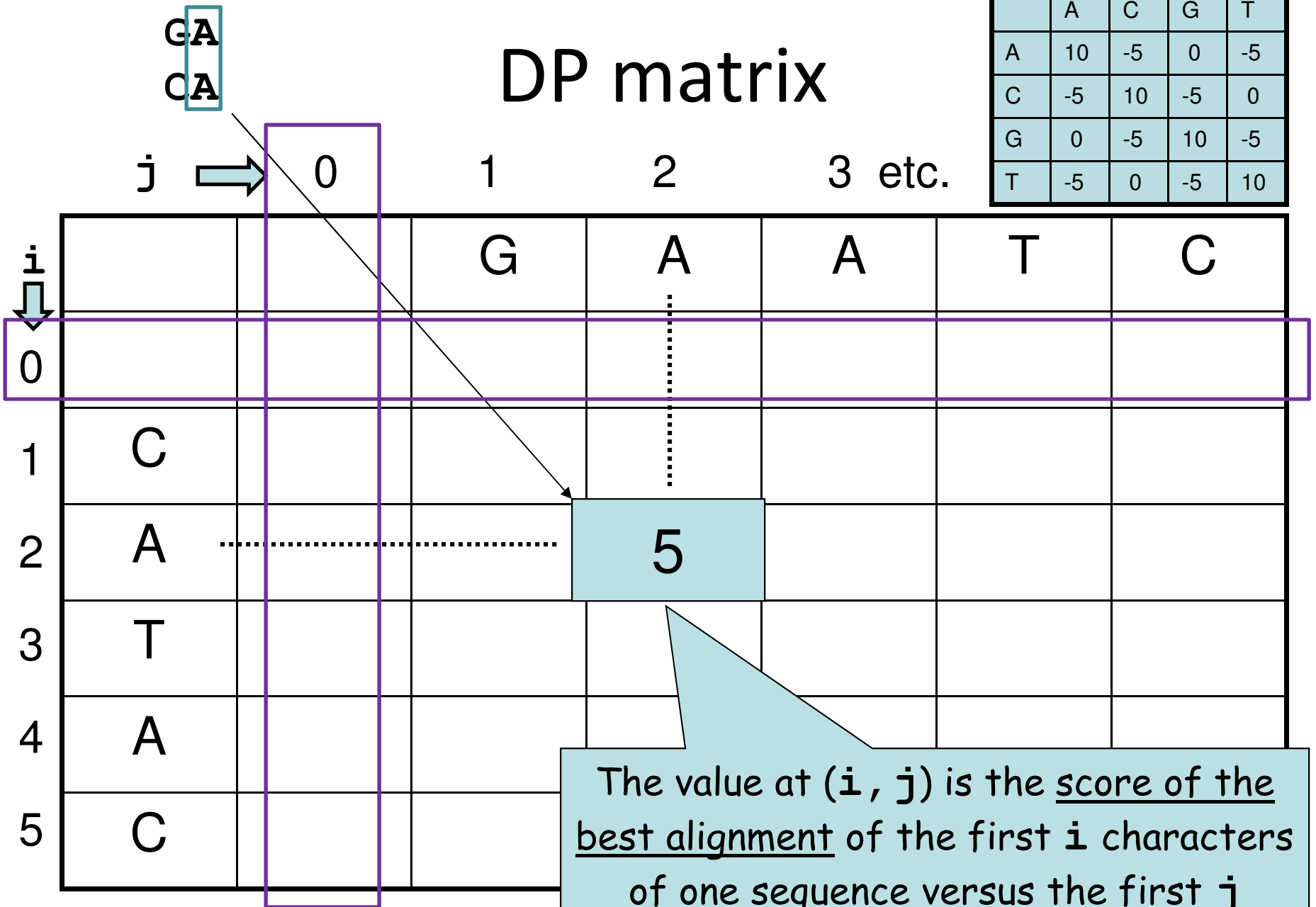
$$\binom{mn}{\min(m,n)} = \frac{(mn)!}{(\min(m,n)!)^2}$$

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

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



initial row and column

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.

GAA
CA-

DP 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

		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-
CA.T

DP 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

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

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

GAA
CAT

DP 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

		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

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

		G	A	A	T	C
	0					
C						
A						
T						
A						
C						

G
|
-

Introducing a gap

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

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

-
C

Introducing a gap

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

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

Complete first row and column

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

CATAC

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

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

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

$G-$
 $-C$

$j \rightarrow$ 0 1 2 3 etc.

$i \downarrow$		G	A	A	T	C
0	0	-4				
1	C	-8				
2	A					
3	T					
4	A					
5	C					

-G

C-

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

j → 0 1 2 3 etc.

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

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

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

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

$j \rightarrow$

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

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

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

	j → 0	1	2	3 etc.		
i ↓		G	A	A	T	C
0		0				
1	C	-4	-8			
2	A					
3	T					
4	A					
5	C					

	j → 0	1	2	3 etc.		
i ↓		G	A	A	T	C
0		0	-4			
1	C		-8			
2	A					
3	T					
4	A					
5	C					

	j → 0	1	2	3 etc.		
i ↓		G	A	A	T	C
0		0				
1	C		-5			
2	A					
3	T					
4	A					
5	C					

Which of these three ways should we use?

Accept the highest scoring of the three

Accept the highest scoring
of the three

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

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

DP 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

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

~~G-~~
~~CA~~

-G
CA

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

$-5 + -4 = -9$

$-4 + 0 = -4$

$-8 + -4 = -12$

DP 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

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

Diagram illustrating the DP matrix calculation for sequence alignment. The matrix shows scores for alignments of prefixes of 'CAGATC' (rows) and 'GATCA' (columns). The diagonal path (0, -4, -8, -12, -16, -20) represents the alignment of 'CAGATC' with 'GATCA'. The cell (A, G) contains a question mark, indicating the value to be calculated. Red 'X' marks and arrows indicate the path taken to reach the cell (A, G): from (0,0) to (0,1) to (1,1) to (2,1). A red 'X' is also shown on the path from (0,1) to (1,2) to (2,2).

~~G-~~
~~CA~~

-G
CA

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

$-5 + -4 = -9$

$-4 + 0 = -4$

$-8 + -4 = -12$

DP 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

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

Diagram illustrating the DP matrix calculation for sequence alignment. The matrix shows scores for alignments between the sequence CAGATC (rows) and GAACTC (columns). The diagonal path (0, -4, -8, -12, -16, -20) represents the alignment of CAGATC with GAACTC. The value -4 in the cell (A, G) is highlighted in blue, indicating the optimal alignment score for that cell. Red 'X' marks are placed over the values -5 and -8 in the cells (C, G) and (A, A) respectively, indicating that these alignments are not optimal. Arrows show the path from the top-left cell (0) to the bottom-right cell (-20) through the diagonal cells.

DP 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

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

DP 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

		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				

DP 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

		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

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

		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

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

		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

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

		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

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

		G	A	A	T	C
	0	-4	-8	-12	-16	-20
C	-4					-6
A	-8					-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

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

		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

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

		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

GA-ATC
CATA-C

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

		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

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

$$d = -4$$

		G	A	A	T	C
	0					
A						
A						
T						
T						
C						