Sequence comparison: Dynamic programming

Genome 559: Introduction to Statistical and Computational Genomics

Prof. James H. Thomas

Sequence comparison overview

- Problem: Find the “best” alignment between a query sequence and a target sequence.
- To solve this problem, we need
  - a method for scoring alignments, and
  - an algorithm for finding the alignment with the best score.
- The alignment score is calculated using
  - a substitution matrix
  - gap penalties.
- The algorithm for finding the best alignment is dynamic programming.

A simple alignment problem.

- Problem: find the best pairwise alignment of `GAATC` and `CATAC`.
- Use a linear gap penalty of -4.
- Use the following 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

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 alignments of two sequences of length ***n*** exist?

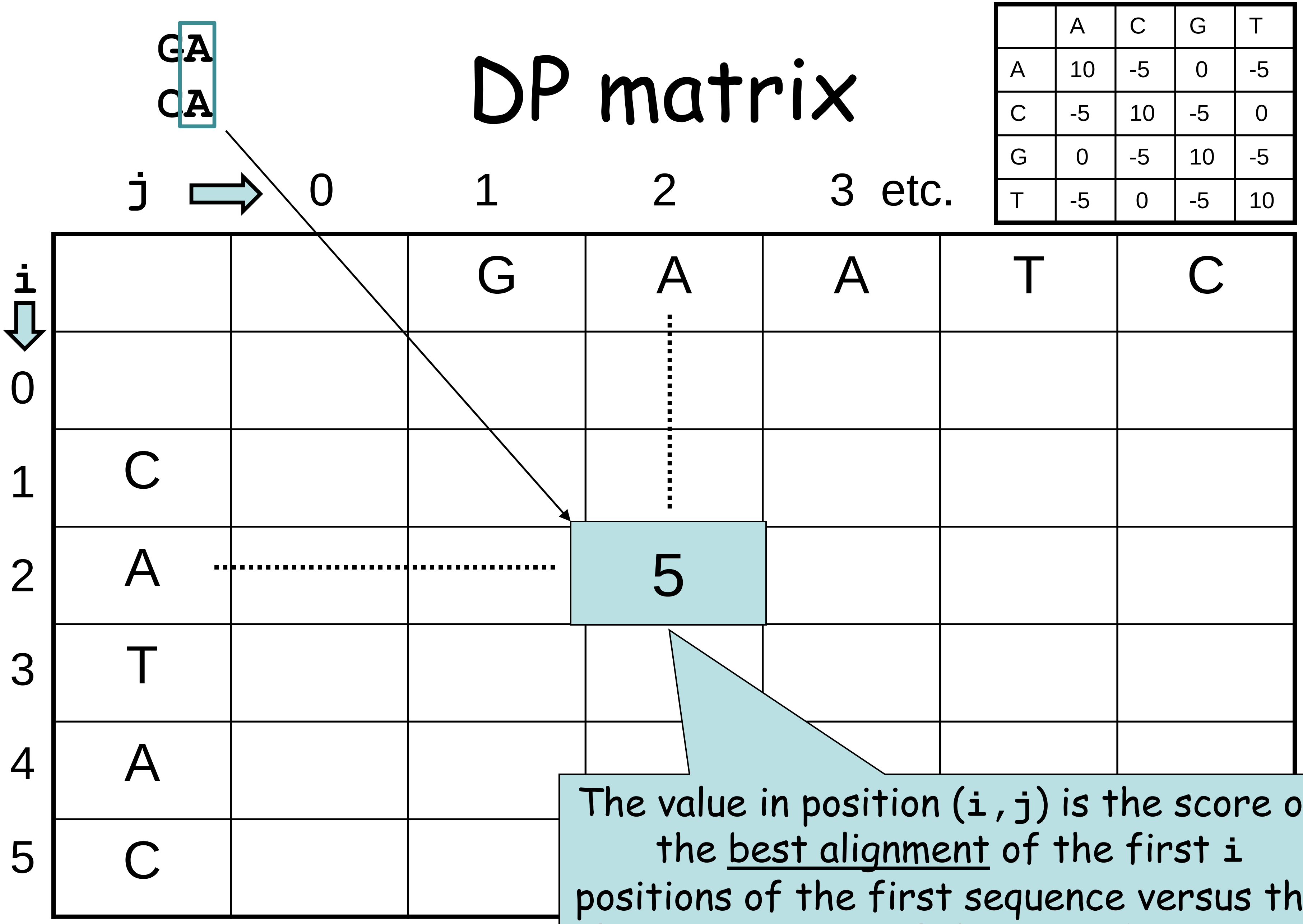
5	9.2×10^{2}
10	1.8×10^{5}
20	1.4×10^{11}
30	1.2×10^{17}
40	1.1×10^{23}

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

A really really huge number

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



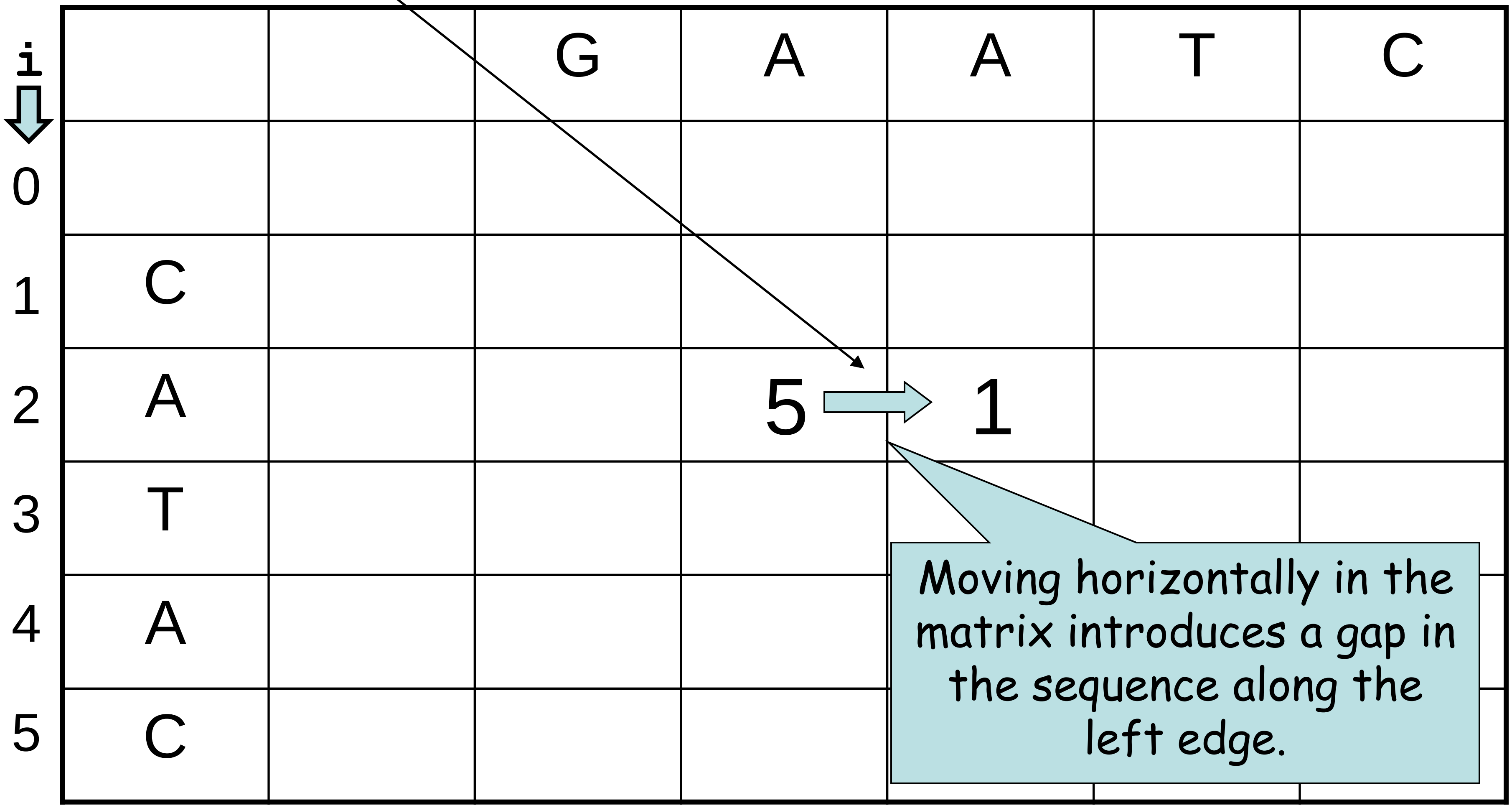
		0	1	2	3 etc.		
			G	A	A	T	C
0							
1	C						
2	A			5			
3	T						
4	A						
5	C						

The value in position (i,j) is the score of the best alignment of the first i positions of the first sequence versus the first j positions of the second sequence.

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

```
GAA
CA-
```



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

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

```
GA-
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

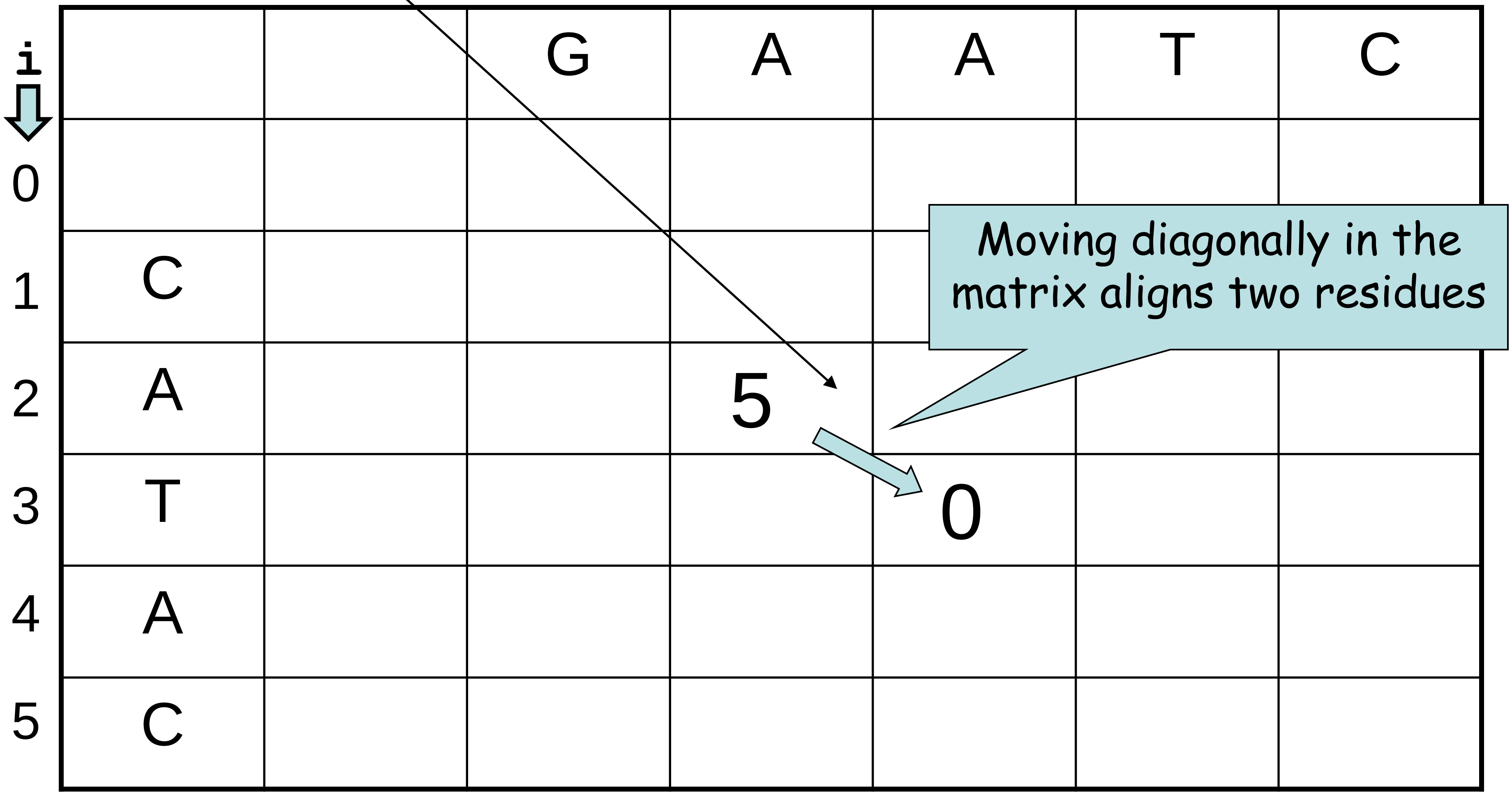
i			G	A	A	T	C
0							
1	C						
2	A			5			
3	T			1			
4	A						
5	C						

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

DP matrix

```
GAA
CAT
```

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



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

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

i ⇩			G	A	A	T	C
0		0 →	-4				
1	C						
2	A						
3	T						
4	A						
5	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

-
C

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

```
G-
-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

i ⇩			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

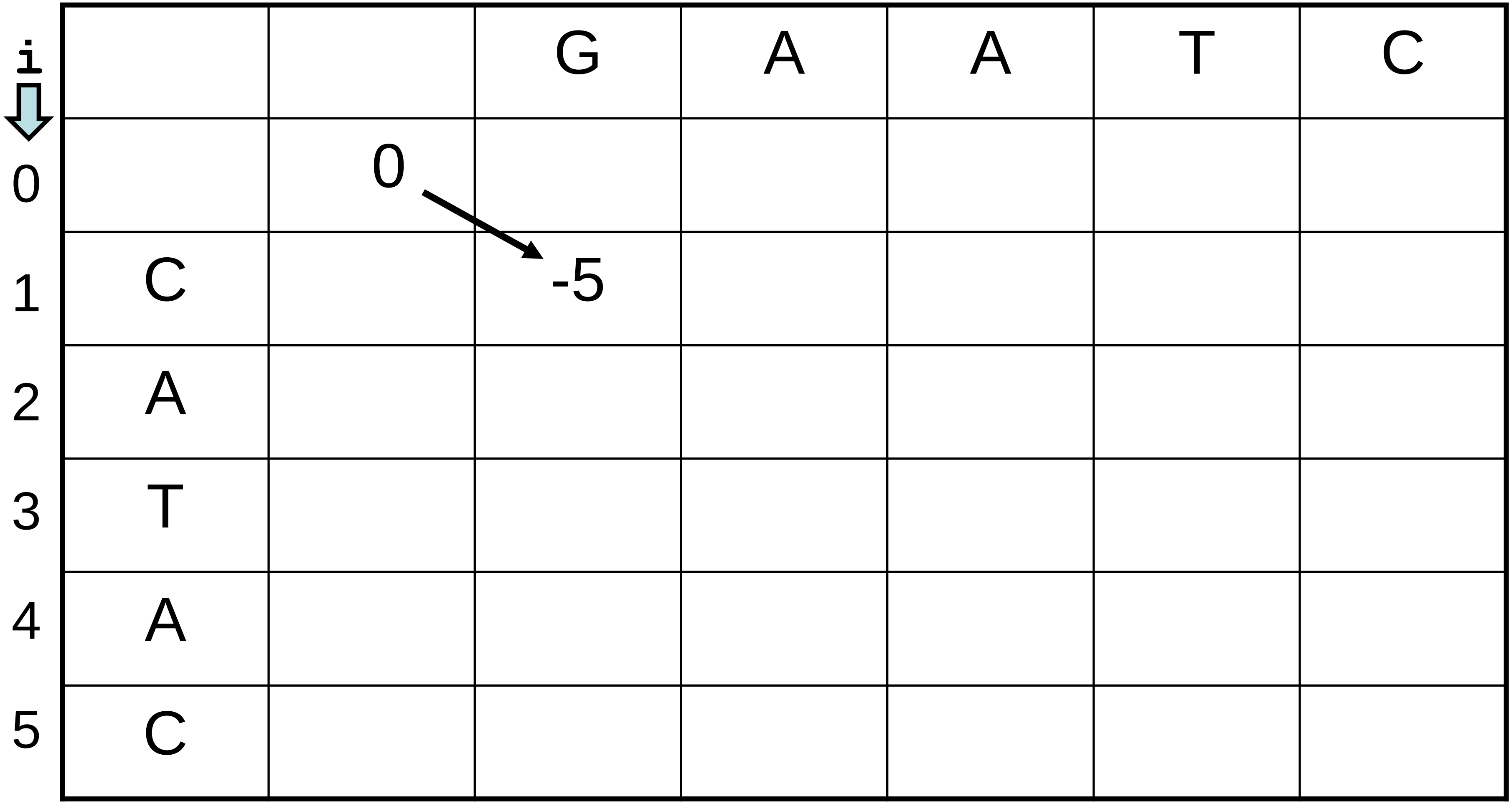
	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10

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

G
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



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

```
-----
CATAC
```

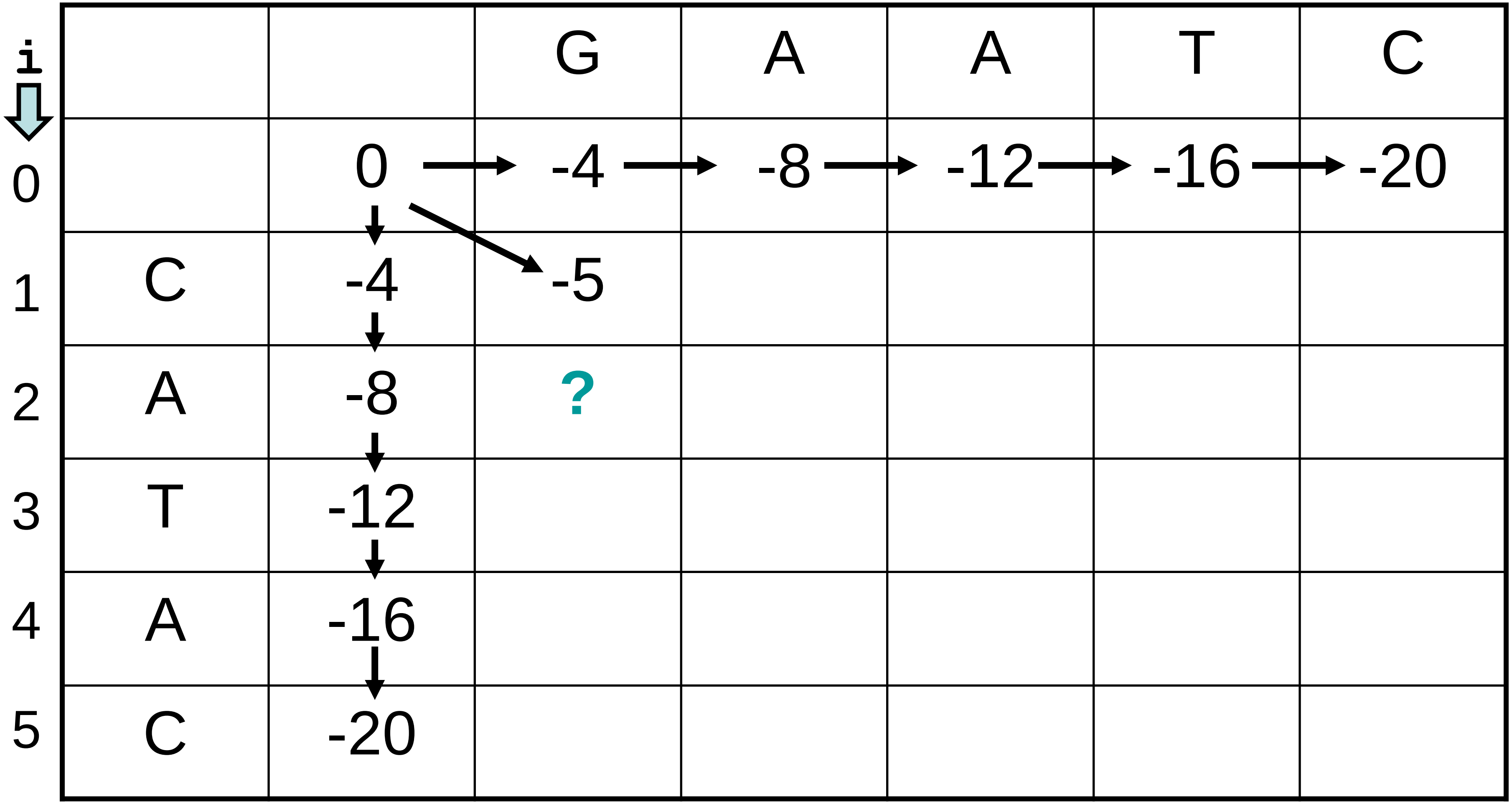
DP matrix

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

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					

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

DP matrix

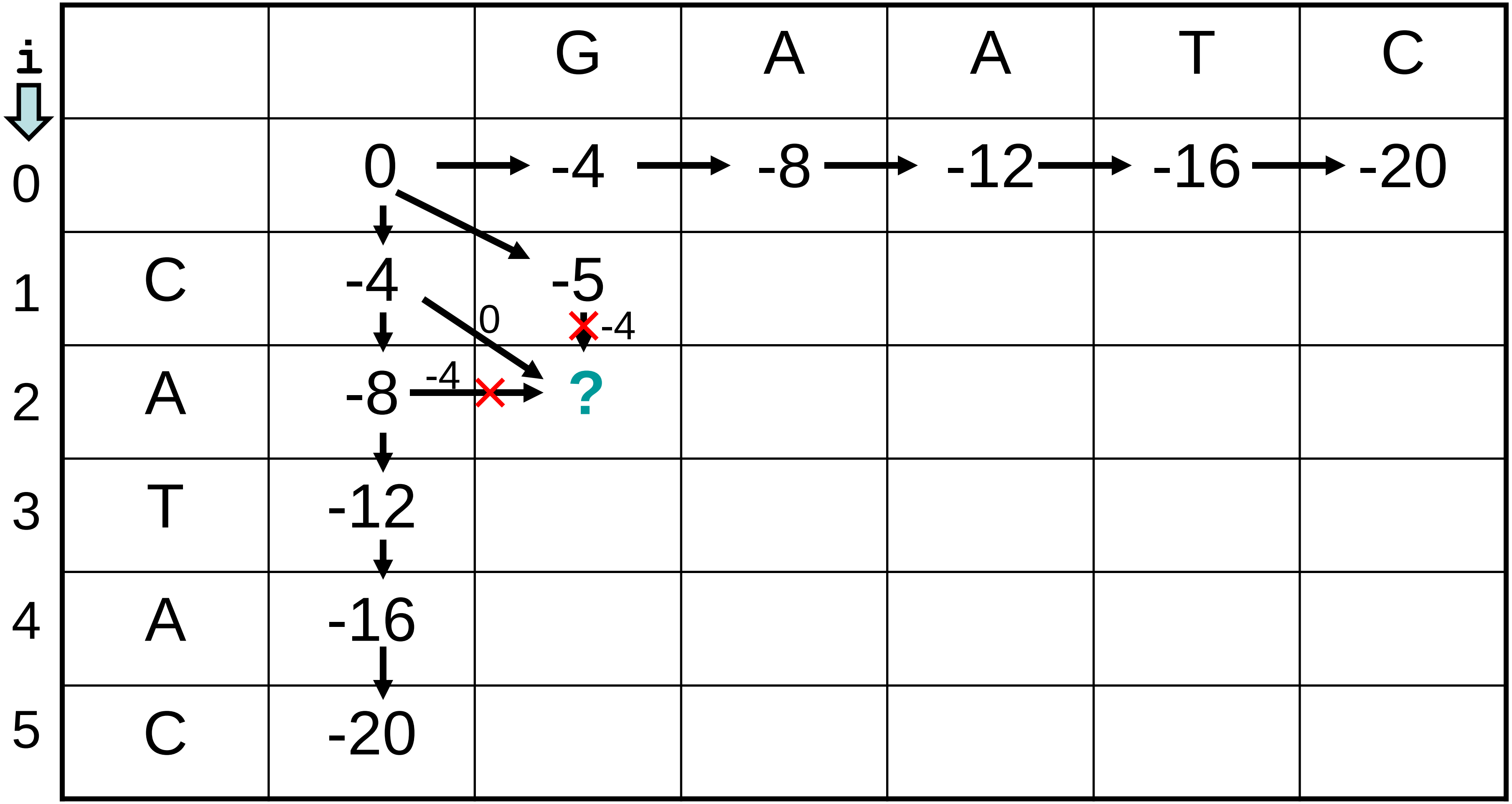


i			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-5				
2	A	-8	?				
3	T	-12					
4	A	-16					
5	C	-20					

DP matrix

-G	~~G-~~	~~--G~~
CA	~~CA~~	~~CA-~~
-4	-9	-12

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



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					

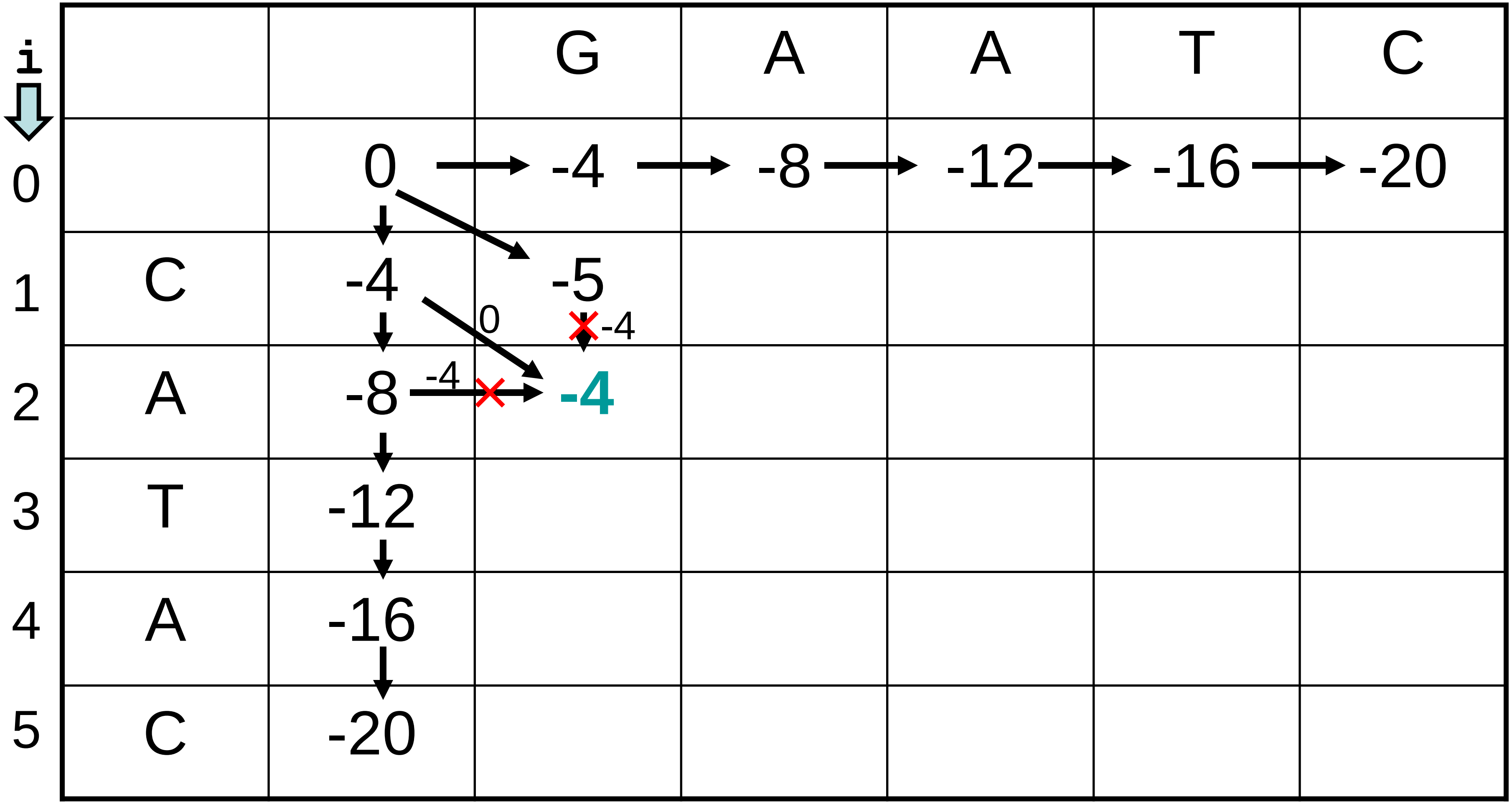
DP matrix

```
-G
CA
  -4
```

```
~~G-~~
~~CA~~
 -9
```

```
~~--G~~
~~CA-~~
  -12
```

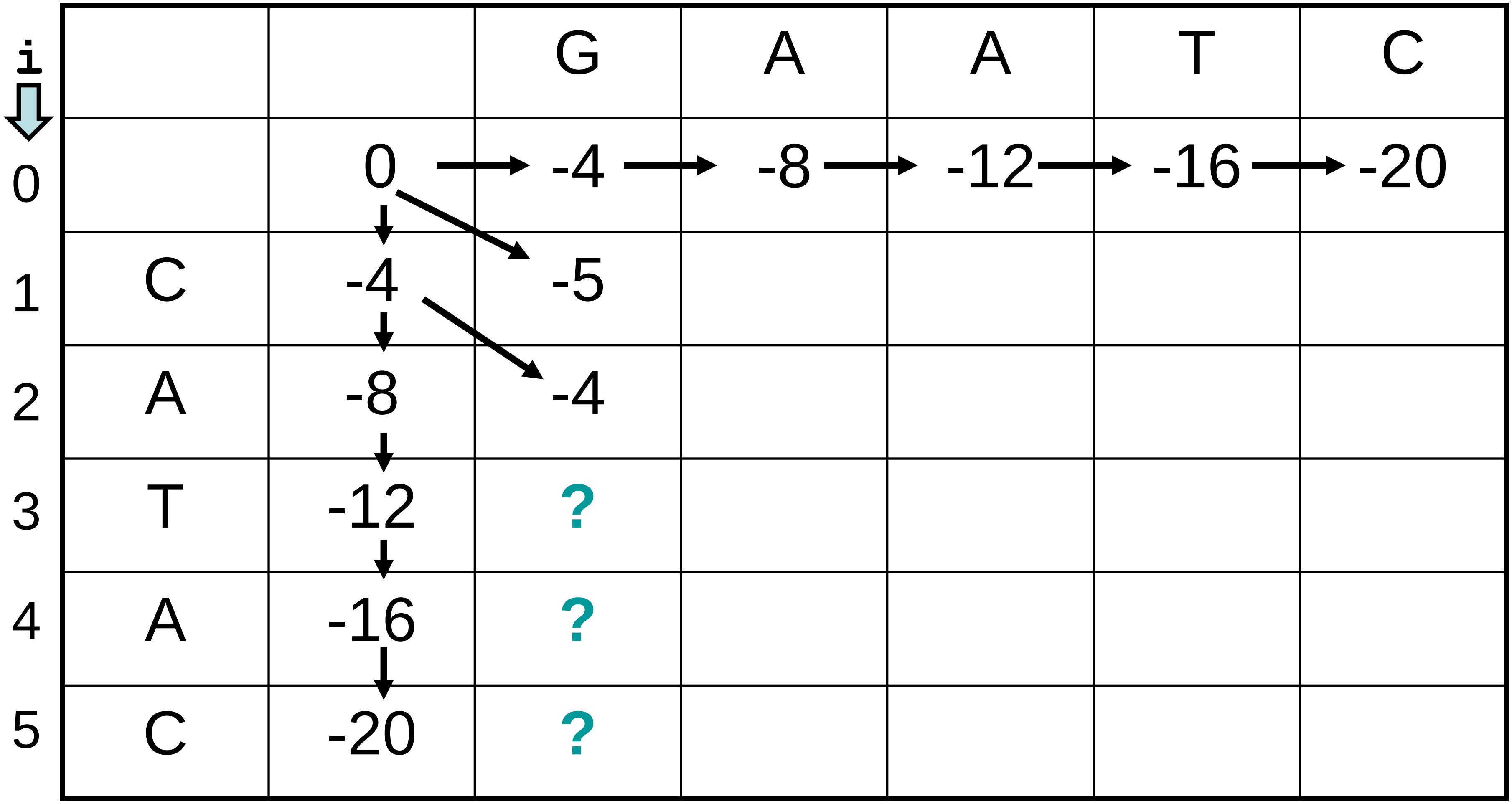
	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10



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

DP matrix

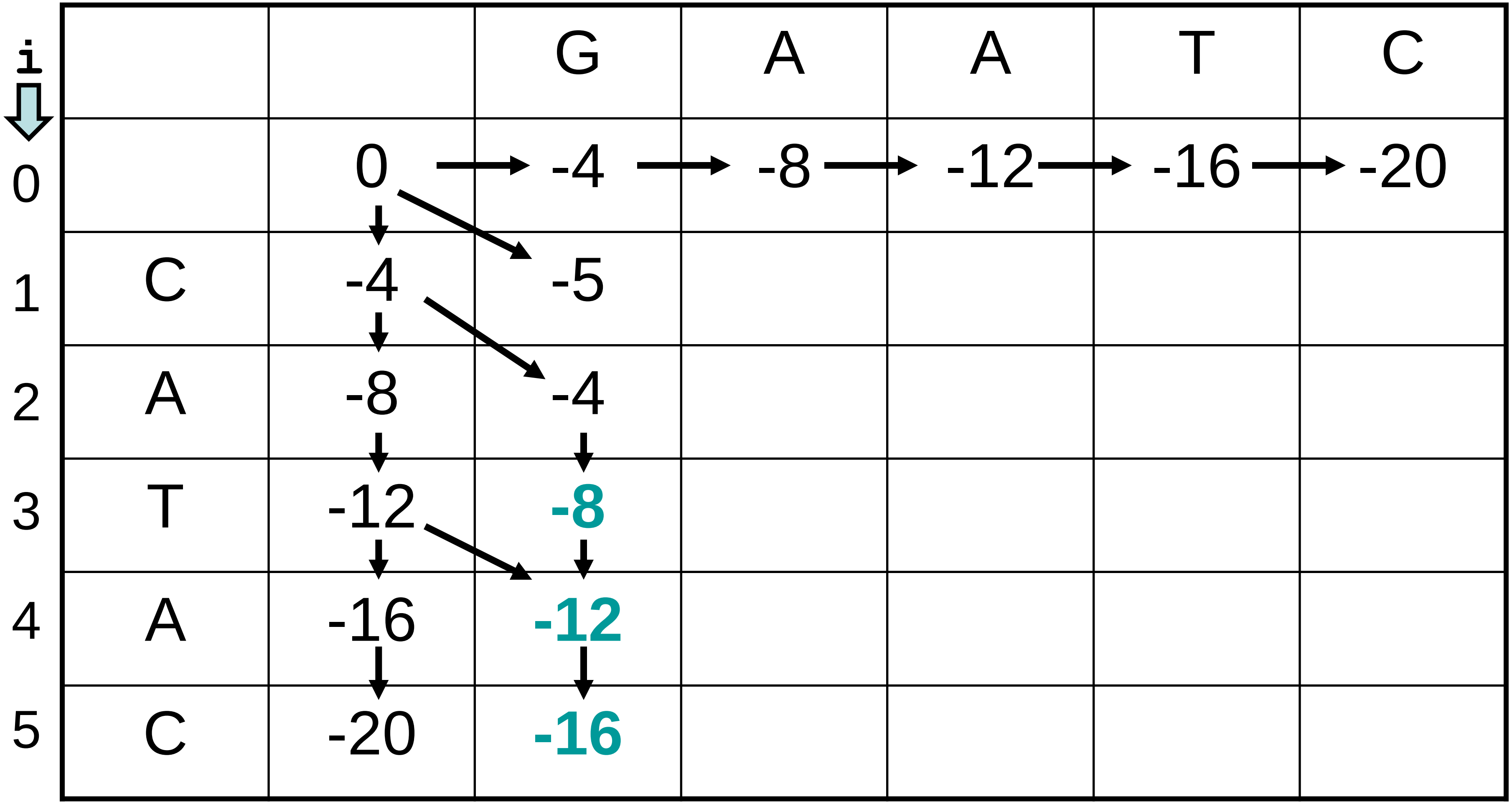
	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10



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

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

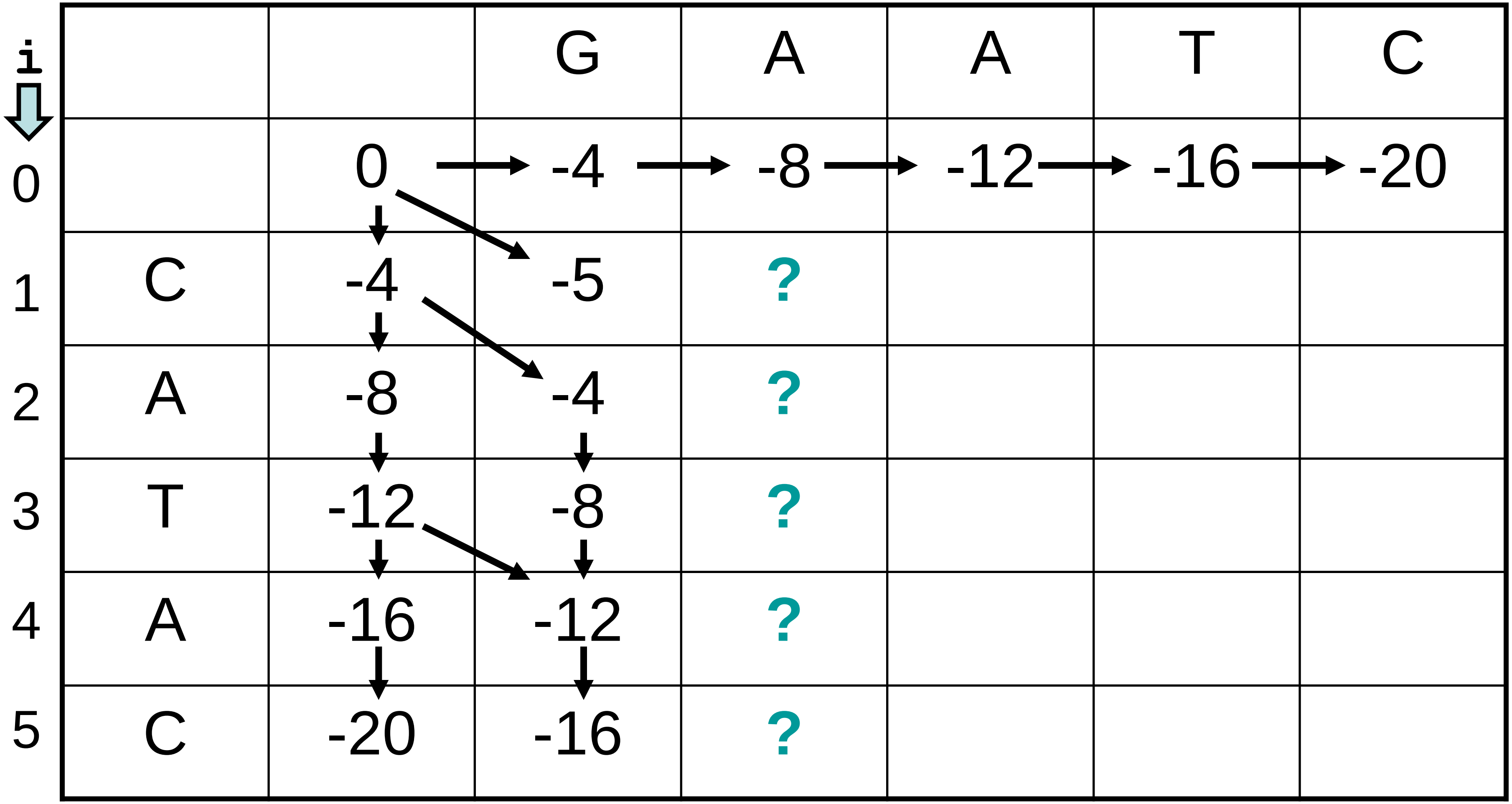
DP matrix



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

DP matrix

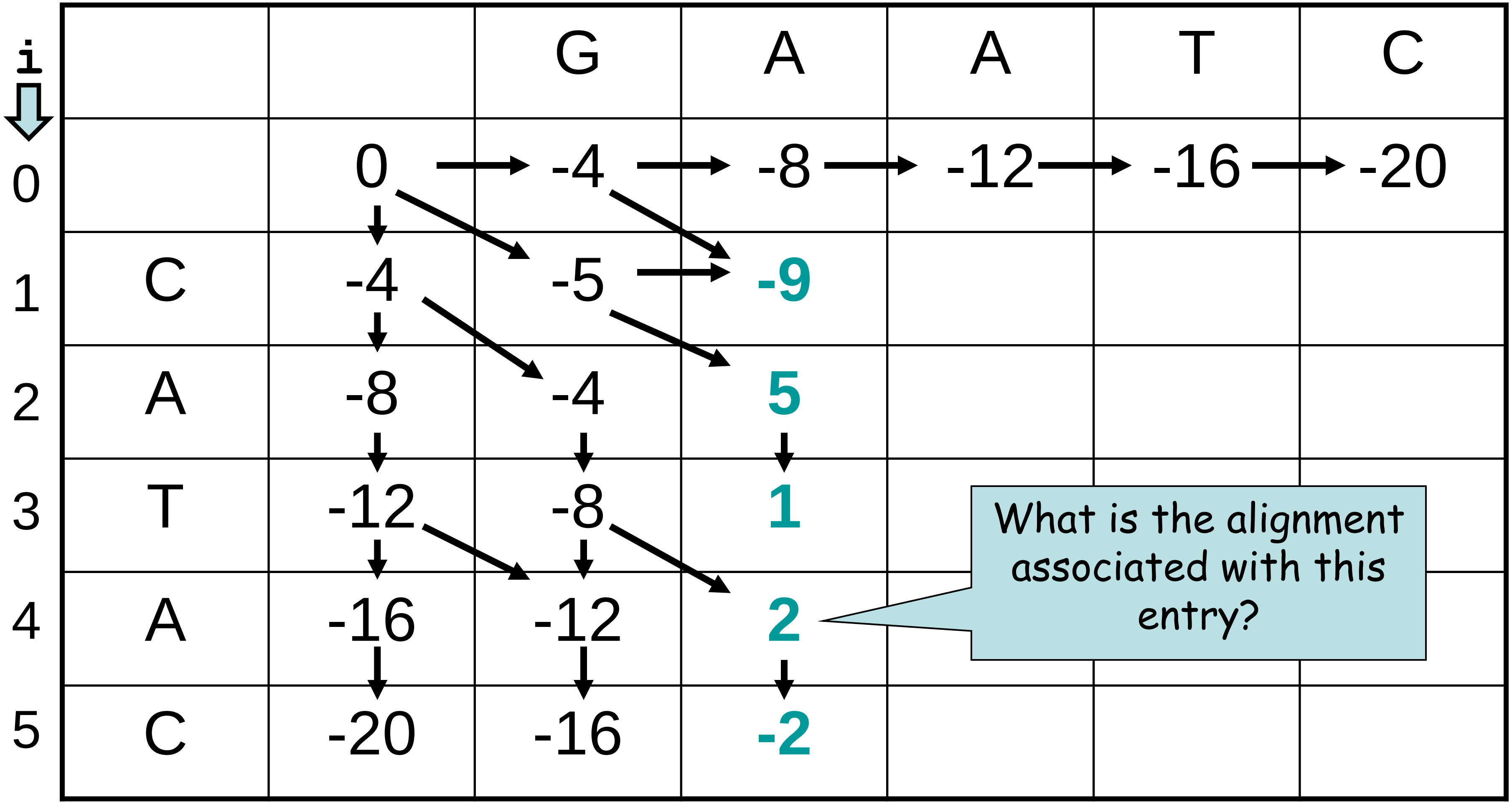
	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10



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

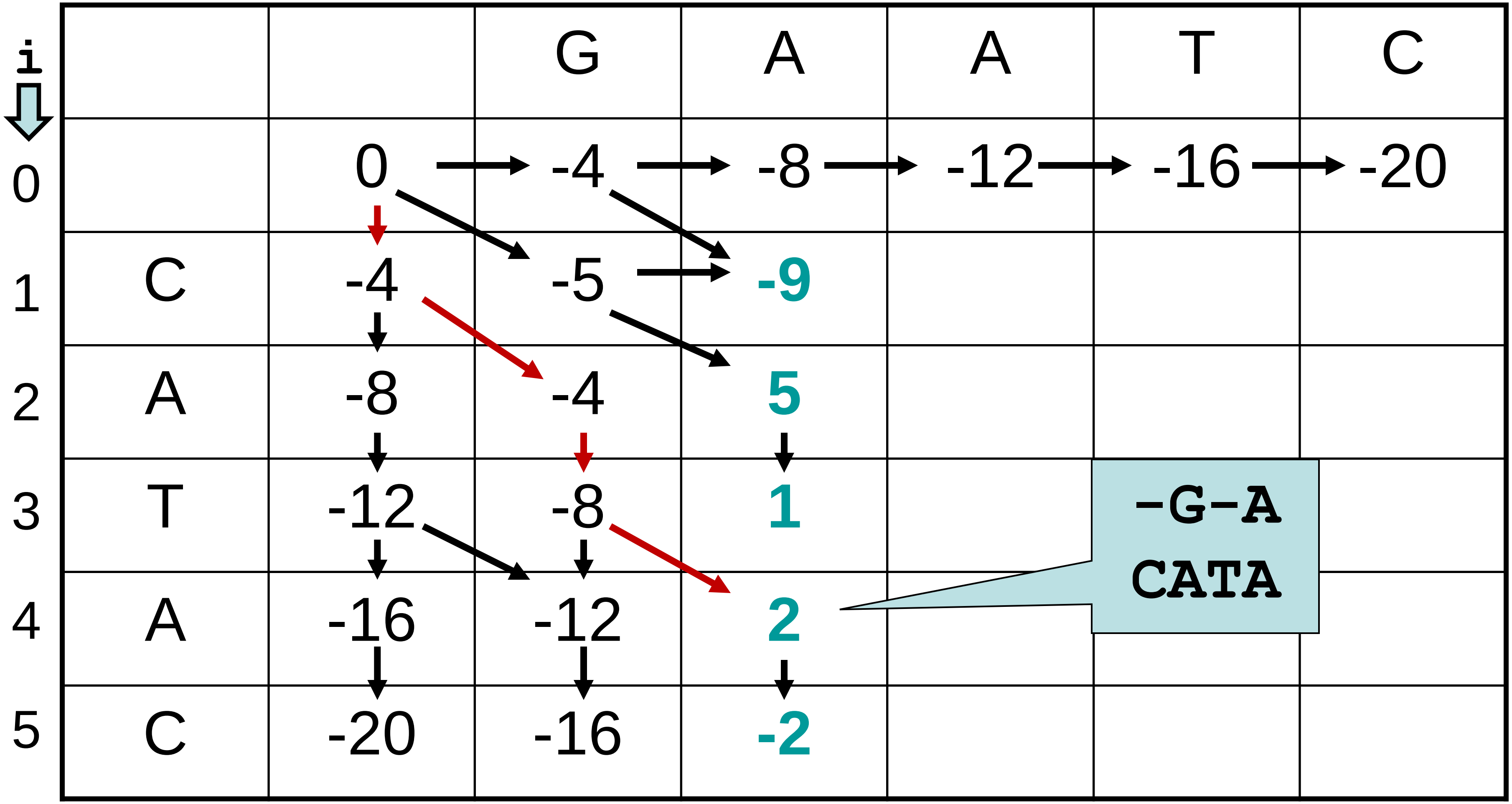


i			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-5	-9			
2	A	-8	-4	5			
3	T	-12	-8	1			
4	A	-16	-12	2			
5	C	-20	-16	-2			

What is the alignment associated with this entry?

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

DP matrix

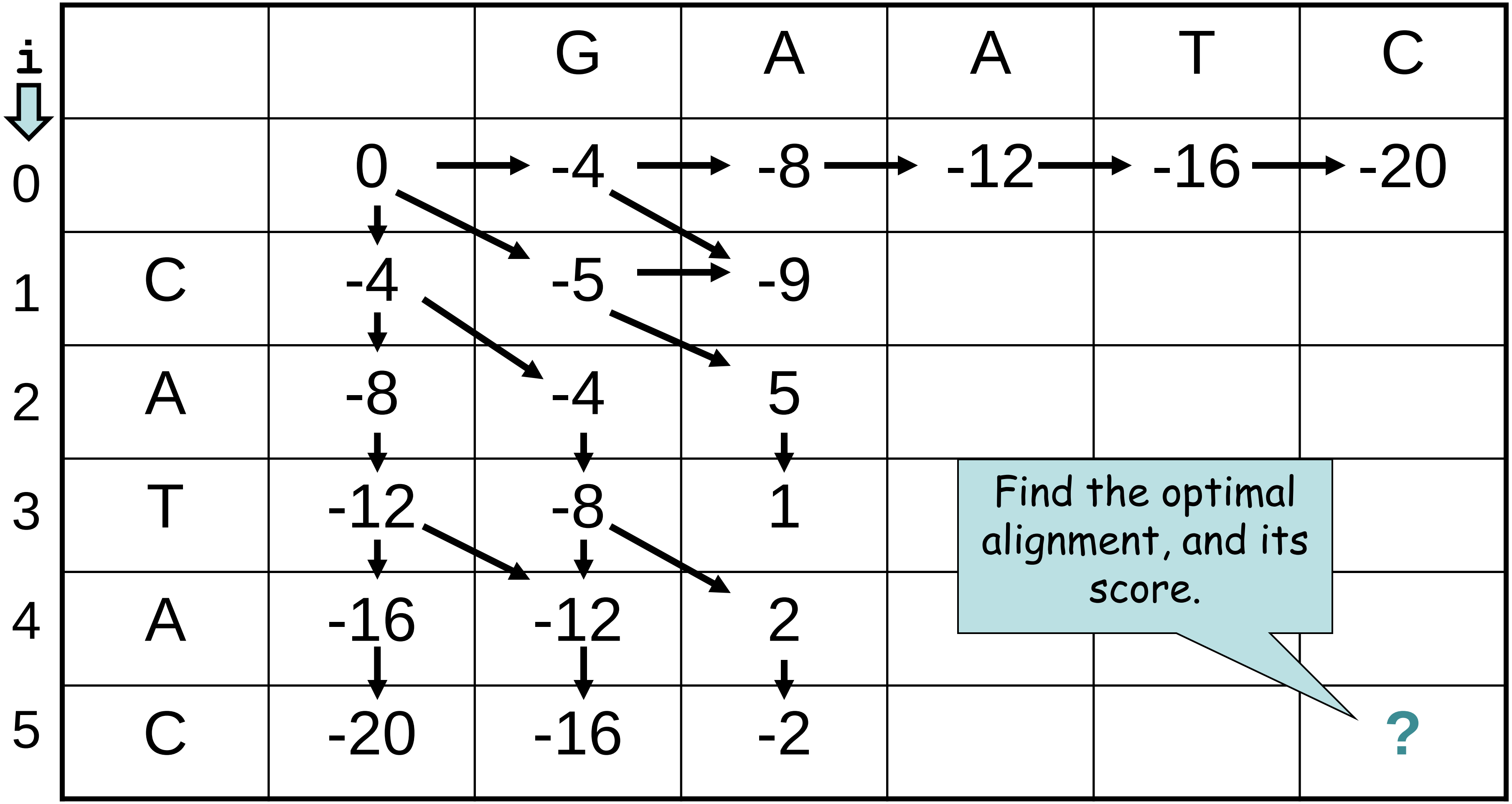


i			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-5	-9			
2	A	-8	-4	5			
3	T	-12	-8	1			
4	A	-16	-12	2			
5	C	-20	-16	-2			

```
-G-A
CATA
```

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

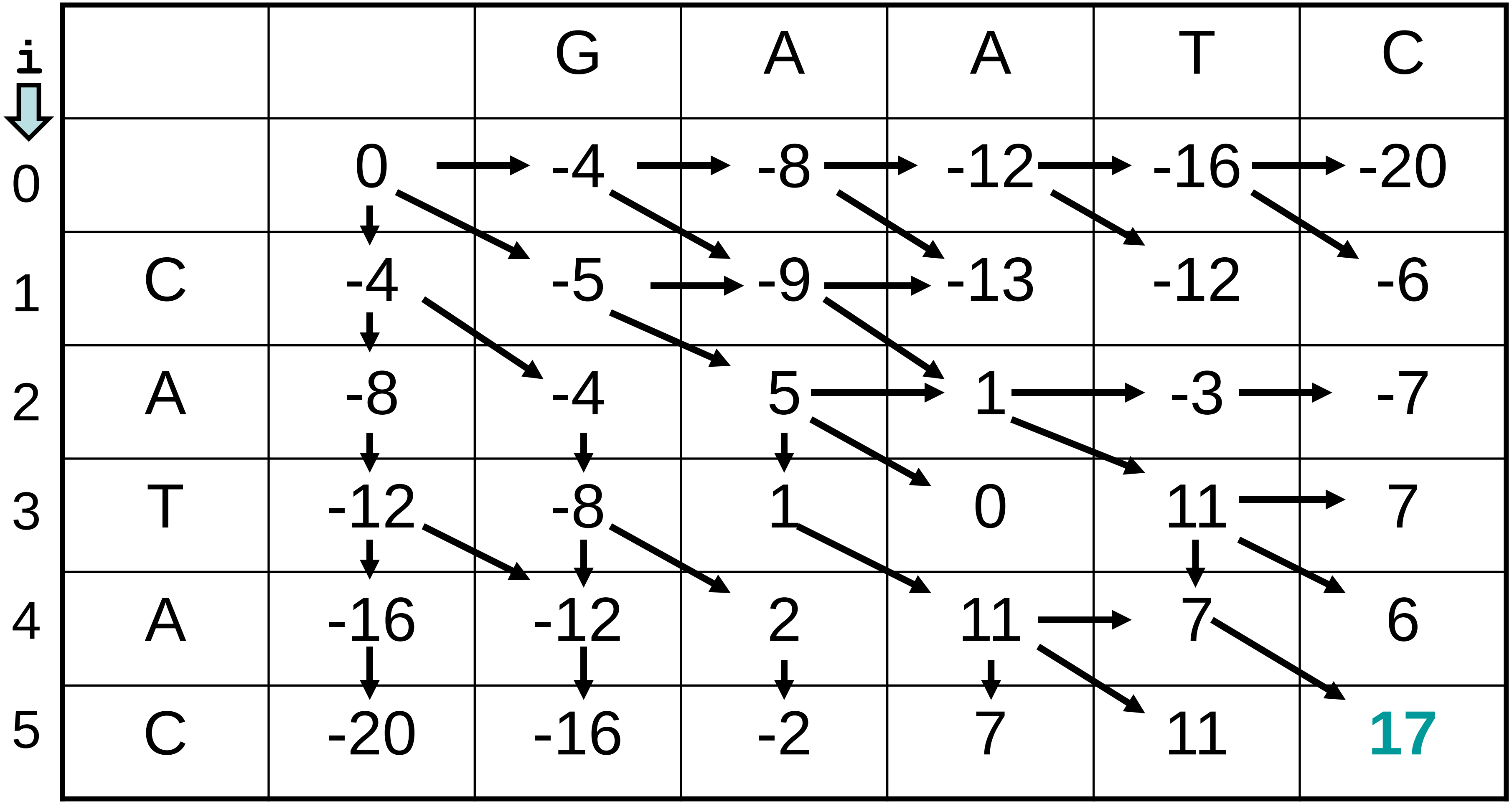


i			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-5	-9			
2	A	-8	-4	5			
3	T	-12	-8	1			
4	A	-16	-12	2			
5	C	-20	-16	-2			?

Find the optimal alignment, and its score.

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

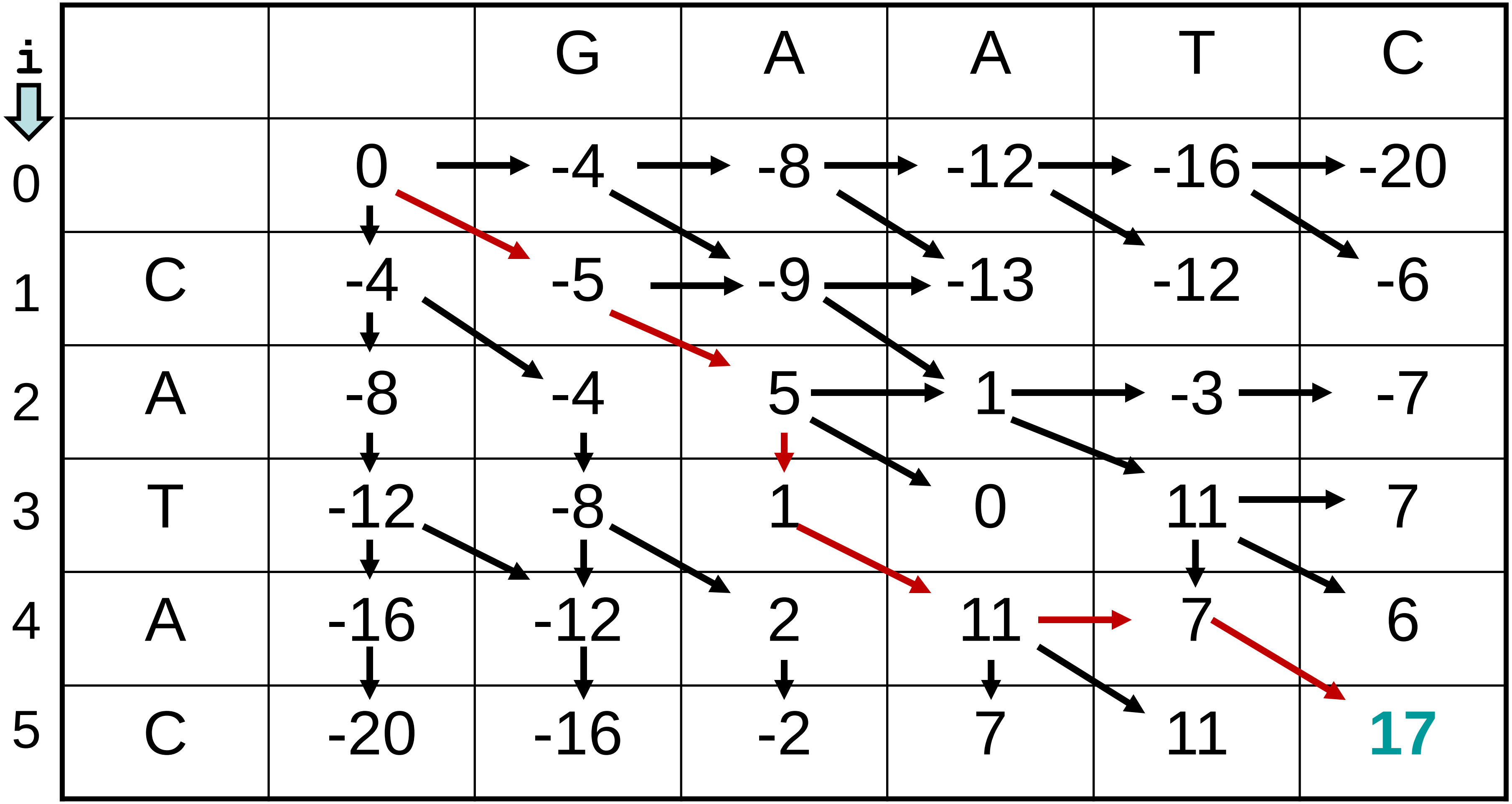


i			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-5	-9	-13	-12	-6
2	A	-8	-4	5	1	-3	-7
3	T	-12	-8	1	0	11	7
4	A	-16	-12	2	11	7	6
5	C	-20	-16	-2	7	11	**17**

```
GA-ATC
CATA-C
```

One best path

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

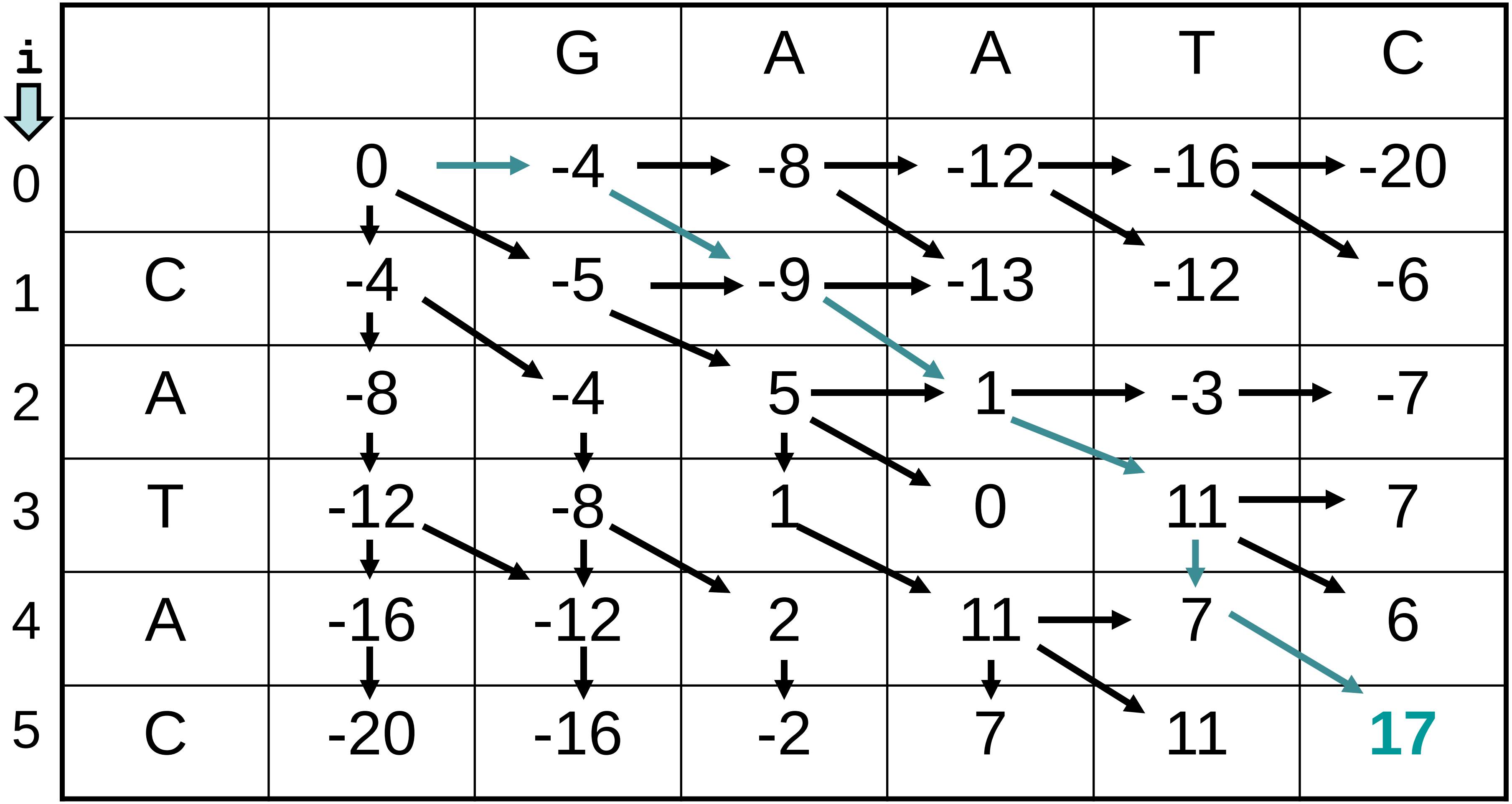


i			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-5	-9	-13	-12	-6
2	A	-8	-4	5	1	-3	-7
3	T	-12	-8	1	0	11	7
4	A	-16	-12	2	11	7	6
5	C	-20	-16	-2	7	11	17

```
GAAT-C
-CATAC
```

Another best path

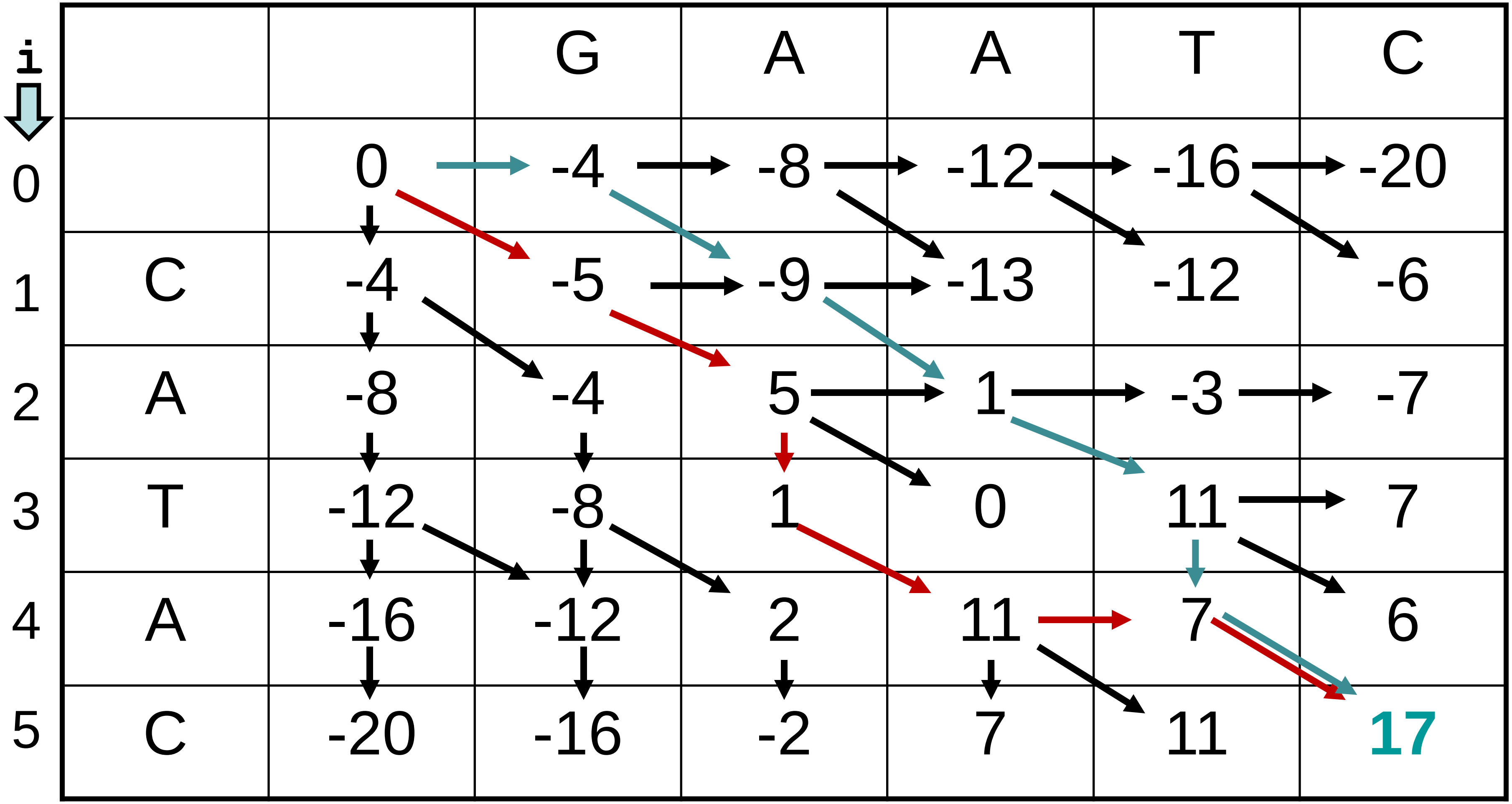
	A	C	G	T
A	10	-5	0	-5
C	-5	10	-5	0
G	0	-5	10	-5
T	-5	0	-5	10



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

```
GAAT-C        GA-ATC
-CATAC        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



i			G	A	A	T	C
0		0	-4	-8	-12	-16	-20
1	C	-4	-5	-9	-13	-12	-6
2	A	-8	-4	5	1	-3	-7
3	T	-12	-8	1	0	11	7
4	A	-16	-12	2	11	7	6
5	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 sequence alignment, it may not be the **only** best alignment.

DP in equation form

- 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

$$\begin{tikzcd}
F(i-1, j-1) \arrow[dr] & & F(i, j-1) \arrow[d] \\
& s(x_i, y_j) \arrow[dr] & d \arrow[d] \\
F(i-1, j) \arrow[r] & d \arrow[r] & F(i, j)
\end{tikzcd}$$

take the best of these three

Dynamic programming

- 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.

Summary

- Scoring a pairwise alignment requires a substitution matrix and gap penalties.
- Dynamic programming is an efficient algorithm for finding an optimal alignment.
- Entry (`i,j`) in the DP matrix stores the score of the best-scoring alignment up to that position.
- DP iteratively fills in the matrix using a simple mathematical rule.

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						