Sequence Alignment

SBS BIOHACKATHON || PYTHON

Scope

1. Global alignment with Backtracking
2. Local alignment with Backtracking

Dynamic Programming

Dynamic Programming is a technique in computer programming that helps to efficiently solve a class of problems that have overlapping subproblems and optimal substructure property.

Global Alignment

A global alignment aligns two sequences from beginning to end. Each letter in the sequence is aligned only once. A general global alignment technique used is the Needleman–Wunsch algorithm, which is based on dynamic programming.

Global alignment

Step 1: Prepare the scoring matrix

Remember to leave a gap in the front. This gap represents nucleotides in front that does not match.

	gap	T	C	G
gap				
A				
T				
C				
G				

Global alignment

Step 2: Perform scoring.

~ Match → +1

~ Mismatch → -1

~ Gap → -2

~ Gap to gap → 0

The value for each box can come from:

~ Left (+ gap)

~ Bottom (+gap)

~ Diagonal (consider match or mismatch)

First, fill in the values for the first row and column

	gap	T	C	G
gap				
A				
T				
C				
G				

Global alignment

Step 2: Perform scoring.

~ Match → +1

~ Mismatch → -1

~ Gap → -2

~ Gap to gap → 0

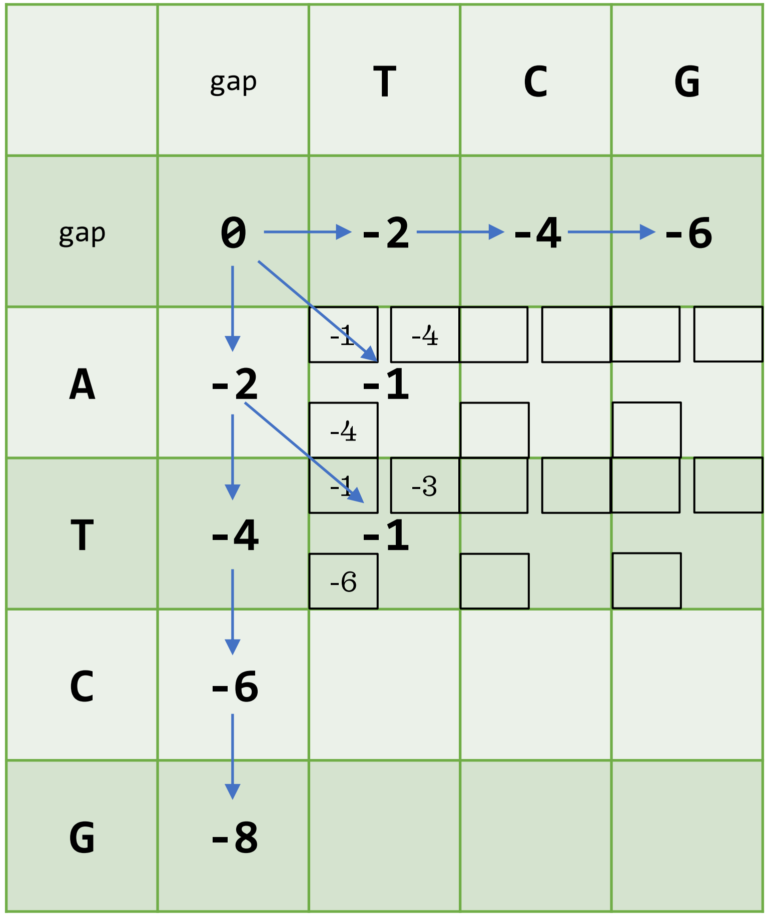
The value for each box can come from:

~ Left (+ gap)

~ Bottom (+gap)

~ Diagonal (consider match or mismatch)

First, fill in the values for the first row and column



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

Global alignment

Step 2: Perform scoring.

~ Match → +1

~ Mismatch → -1

~ Gap → -2

~ Gap to gap → 0

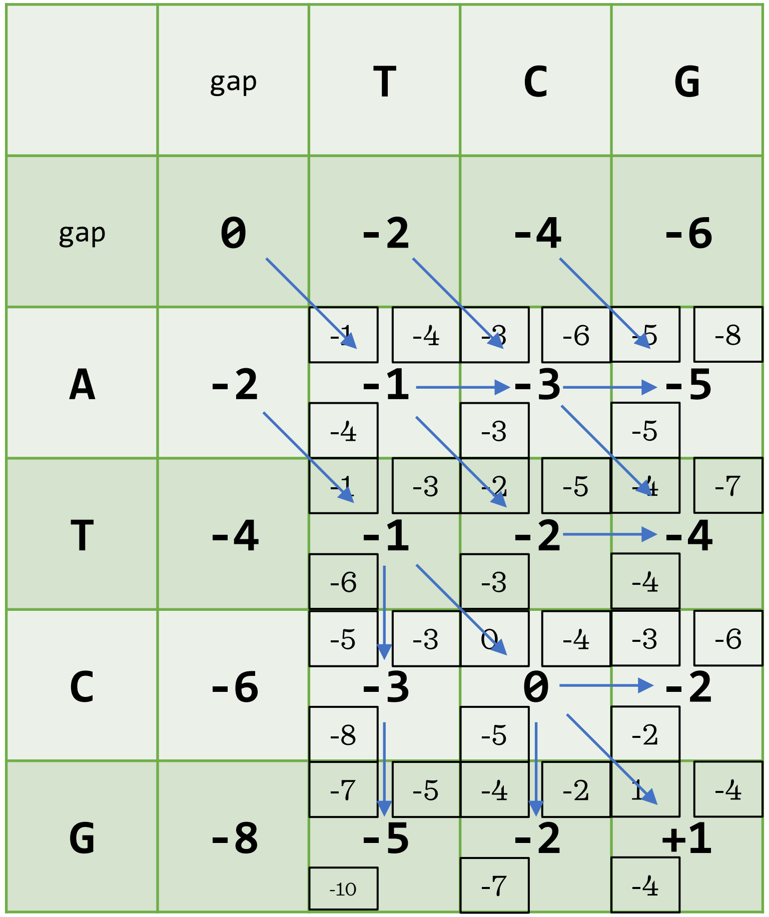
The value for each box can come from:

~ Left (+ gap)

~ Bottom (+gap)

~ Diagonal (consider match or mismatch)

First, fill in the values for the first row and column



	gap	T	C	G
gap	0	-2	-4	-6
A	-2	-1 -4 -1 -4	-3 -6 -3 -3	-5 -8 -5 -5
T	-4	-1 -3 -1 -6	-2 -5 -2 -3	-4 -7 -4 -4
C	-6	-5 -3 -3 -8	0 -4 0 -5	-3 -6 -2 -2
G	-8	-7 -5 -5 -10	-4 -2 -2 -7	1 -4 +1 -4

Global alignment

Step 2: Perform scoring.

~ Match → +1

~ Mismatch → -1

~ Gap → -2

~ Gap to gap → 0

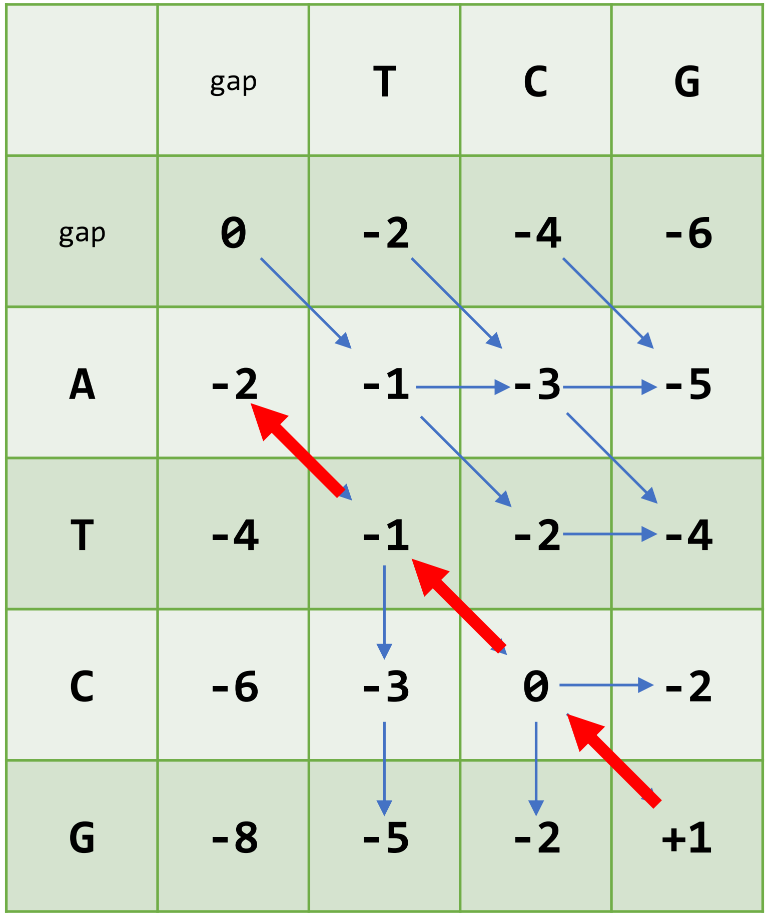
The value for each box can come from:

~ Left (+ gap)

~ Bottom (+gap)

~ Diagonal (consider match or mismatch)

First, fill in the values for the first row and column



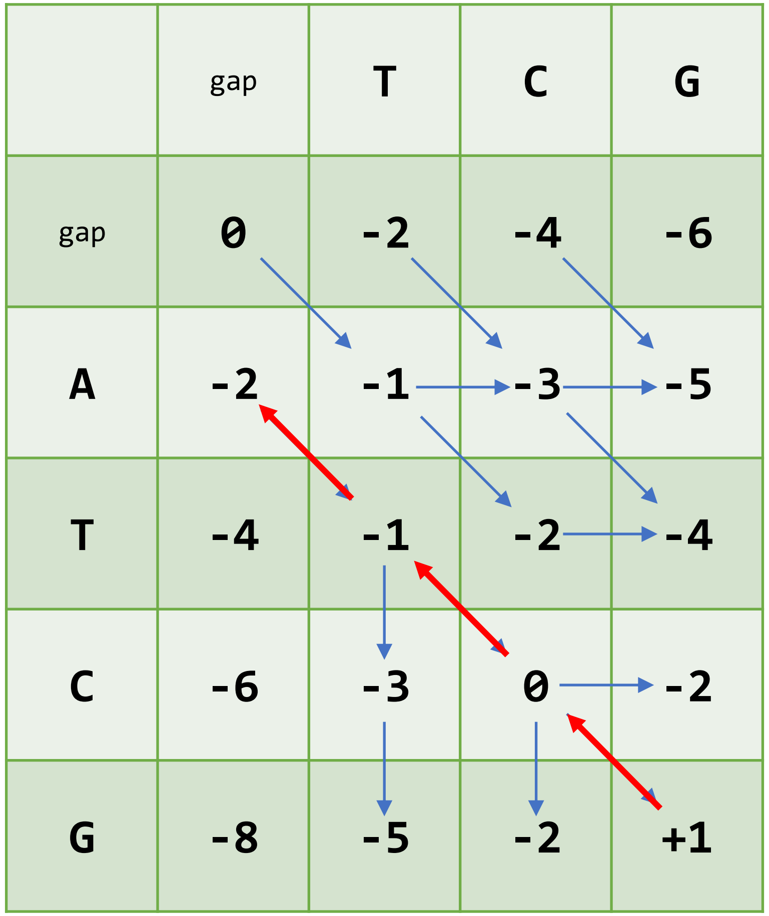
	gap	T	C	G
gap	0	-2	-4	-6
A	-2	-1	-3	-5
T	-4	-1	-2	-4
C	-6	-3	0	-2
G	-8	-5	-2	+1

Global alignment

Step 3: Traceback

Follow the arrow where the value came from

If the value comes from two directions, choose the one with the higher value



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

Global alignment

Step 4: Alignment

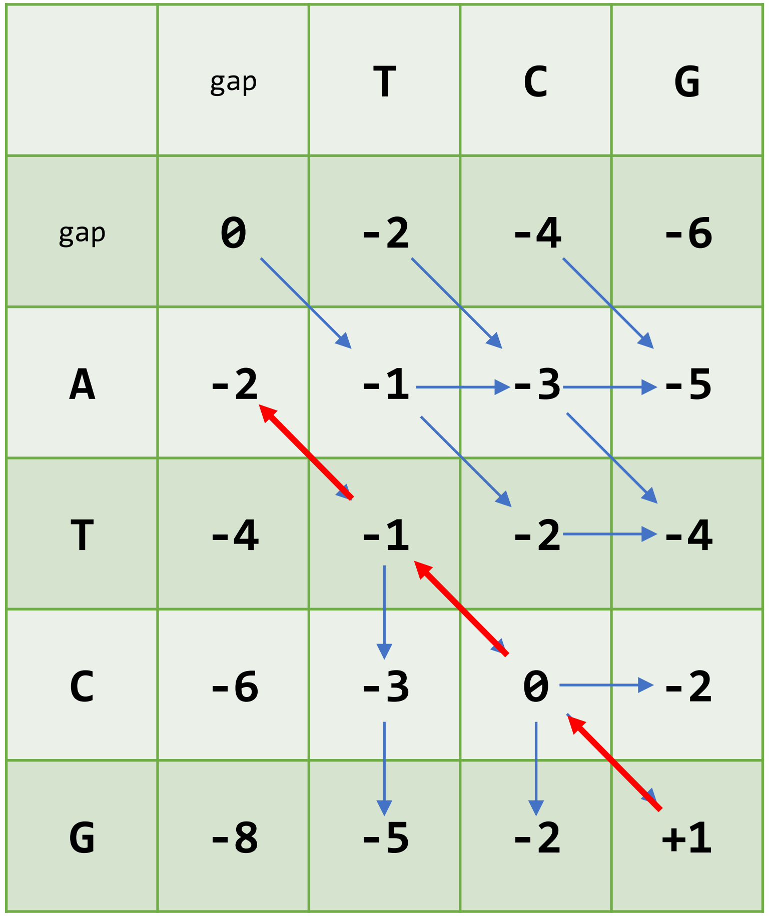
Look at traceback arrow and direction

Up and down arrows, gap: Gap

Diagonal: Character

ATCG

_TCG



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

Local Alignment

Local alignments are used for dissimilar sequences that may contain regions of similarity. The is a general local algorithm used is the Smith-Waterman algorithm which is also based on the dynamic programming but with additional choices to start and end at any place.

Local alignment

Step 1: Prepare scoring matrix

	gap	C	A	T	D	O	G	F	I	S	H
gap											
D											
O											
G											

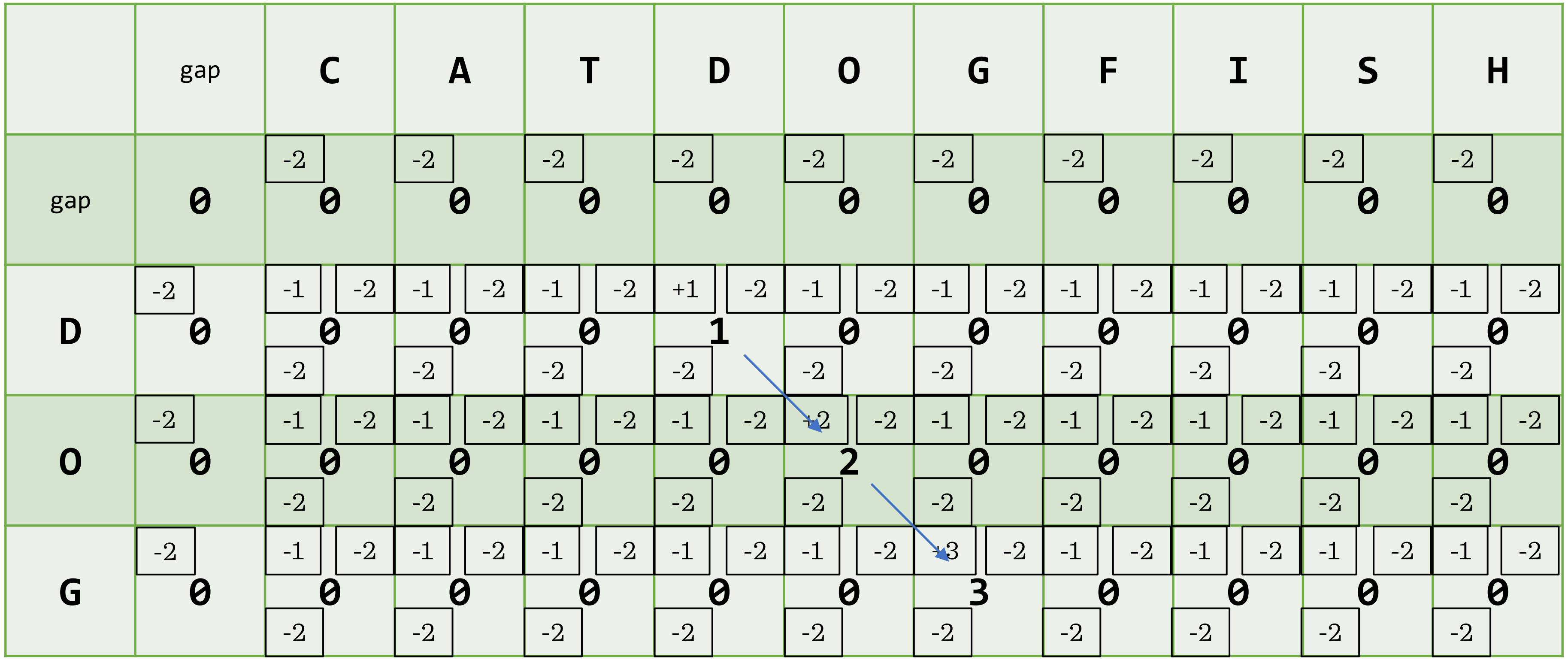
Local alignment

Step 2: Scoring (gap = -7, match = 10, mismatch = -5)

	gap	C	A	T	D	O	G	F	I	S	H
gap	0	□	□	□	□	□	□	□	□	□	□
D	□	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □
O	□	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □
G	□	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □	□ □ □

Local alignment

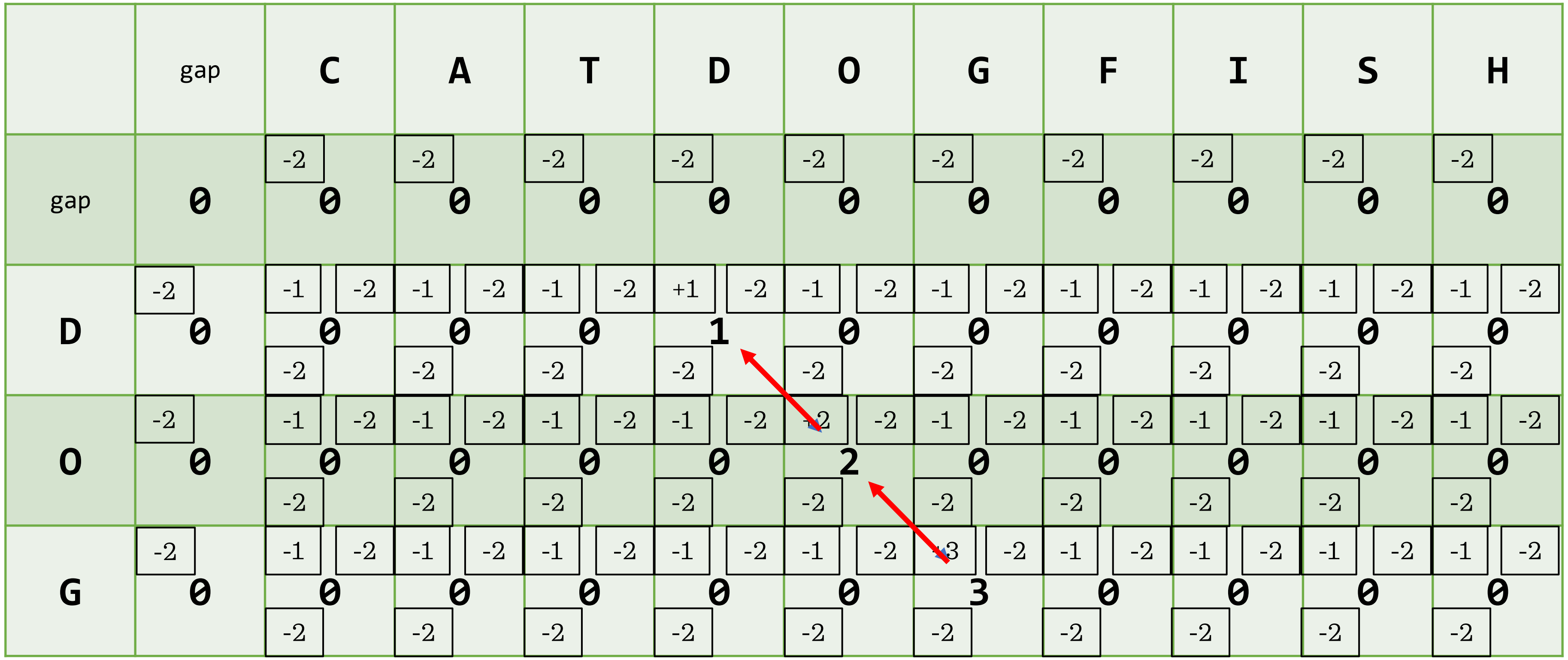
Step 2: Scoring (gap = -2, match = 1, mismatch = -1)



	gap	C	A	T	D	O	G	F	I	S	H
gap	0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0
D	-2 0	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	+1 -2 1 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2
O	-2 0	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	+2 -2 2 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2
G	-2 0	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	+3 -2 3 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2

Local alignment

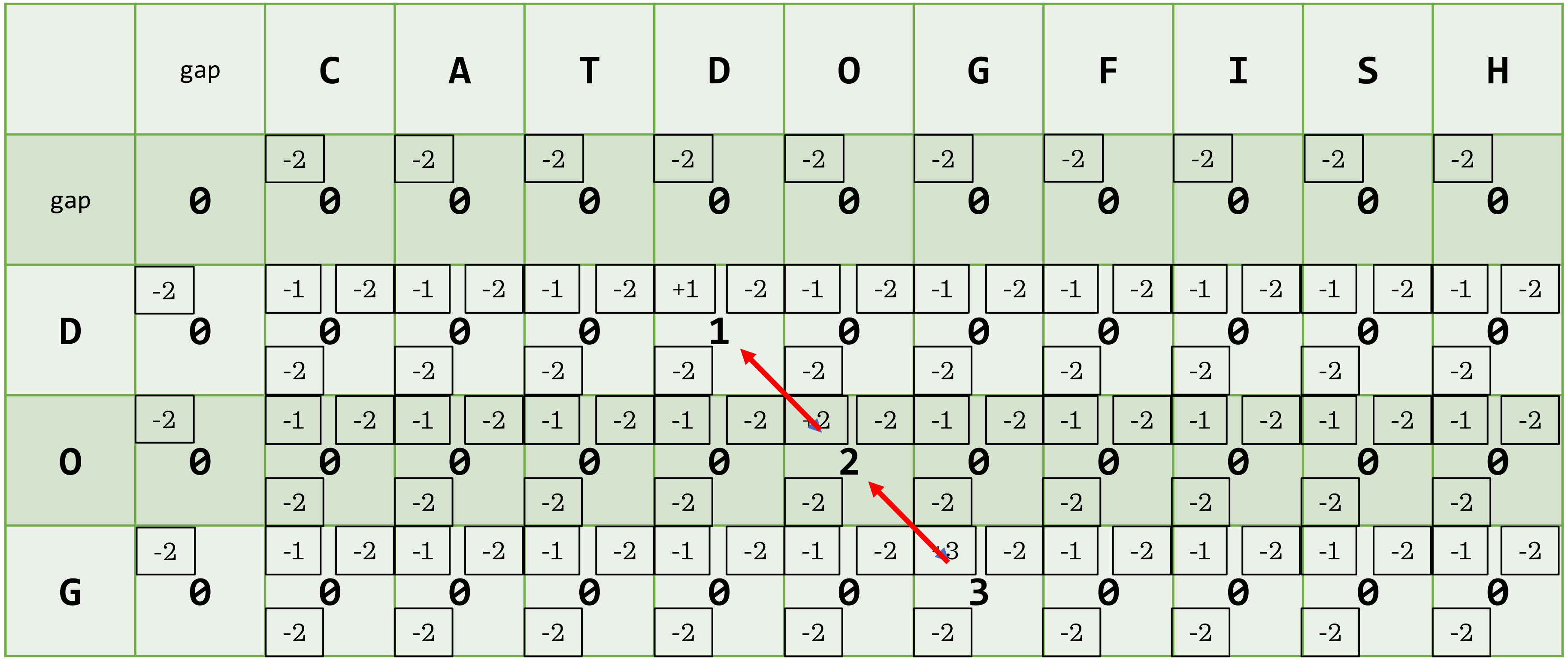
Step 3: Traceback



	gap	C	A	T	D	O	G	F	I	S	H
gap	0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0
D	-2 0	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	+1 -2 1 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2
O	-2 0	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	+2 -2 2 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2
G	-2 0	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	+3 -2 3 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2

Local alignment

Step 4: Alignment (Result: DOG)



	gap	C	A	T	D	O	G	F	I	S	H
gap	0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0	-2 0
D	-2 0	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	+1 -2 1 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2
O	-2 0	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	+2 -2 2 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2
G	-2 0	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	+3 -2 3 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2	-1 -2 0 -2