Pairwise alignment

Optimal alignment by dynamic programming

Why align?

Given the following sequences of letters:

1. martin
2. marvin
3. martina
4. marina

Determine which pair(s) of sequences are:

a. Most similar to each other

b. Most dissimilar to each other

Alignment (gap insertion) required:

```
          m  mm  gap                    m  mm  gap
martin                      martin-
|||x||    5   1    0        ||||||      6   0    1
marvin                      martina


martin-                     marvin-
||| ||    5   0    2        |||x||      5   1    1
mar-ina                     martina


marvin-                     martina
||| ||    5   0    2        ||| |||     6   0    1
mar-ina                     mar-ina
```

m=no. of matches, mm=no of mismatches, gap=no. of gaps

A scoring system

$$S \leftarrow (N_m \times P_m) + (N_{mm} \times P_{mm}) + (N_{gap} \times P_{gap})$$

S	Alignment score
N_m	Number of matches
P_m	Match penalty
N_{mm}	Number of mismatches
P_{mm}	Mismatch penalty
N_{gap}	Number of gaps
P_{gap}	Gap penalty

- *Mismatches are equivalent to substitutions*
- *Gaps are equivalent to insertions and deletions*

Global alignment scores

```
          m  mm  gap  score
martin
|||x||    5   1    0     16
marvin



martin-
||| ||    5   0    2      4
mar-ina



marvin-
||| ||    5   0    2      4
mar-ina
```

```
          m  mm  gap  score
martin-
||||||    6   0    1     16
martina



marvin-
|||x||    5   1    1     12
martina



martina
||| |||   6   0    1     16
mar-ina
```

$$S \leftarrow (N_m \times P_m) + (N_{mm} \times P_{mm}) + (N_{gap} \times P_{gap})$$

$P_m \leftarrow 4$ $P_{mm} \leftarrow -4$ $P_{gap} \leftarrow -8$

Local alignment scores

```
          m  mm  gap  score
martin
|||x||    5   1    0     16
marvin
```

```
          m  mm  gap  score
martin
||||||    6   0    0     24
martina
```

```
martin
||| ||    5   0    1     12
mar-ina
```

```
marvin
|||x||    5   1    0     16
martina
```

```
marvin
||| ||    5   0    1     12
mar-ina
```

```
martina
||| |||   6   0    1     16
mar-ina
```

$$S \leftarrow (N_m \times P_m) + (N_{mm} \times P_{mm}) + (N_{gap} \times P_{gap})$$

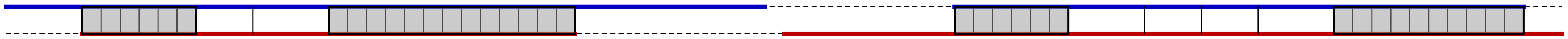
$P_m \leftarrow 4$ $P_{mm} \leftarrow -4$ $P_{gap} \leftarrow -8$

Global and Local alignment

- Global
    - Includes all residues in both sequences
    - Must include gaps if unequal length
- Local
    - Highest scoring sub-alignment
- Others
    - Differential treatment of internal and terminal gaps
    - May ignore residues at ends of sequences

Why align?

To determine the similarity of two sequences

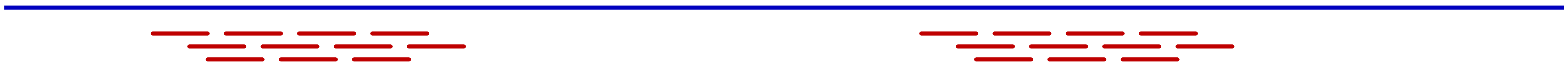


To identify similar (conserved) regions. Likely to have function.

- Functional Analyses
  - Evolutionary relationships (homology / paralogy)
  - Important regions
  - Inferring structure
- Phylogenetic analyses
  - Evolutionary divergence
  - Basis for tree building

Why align?

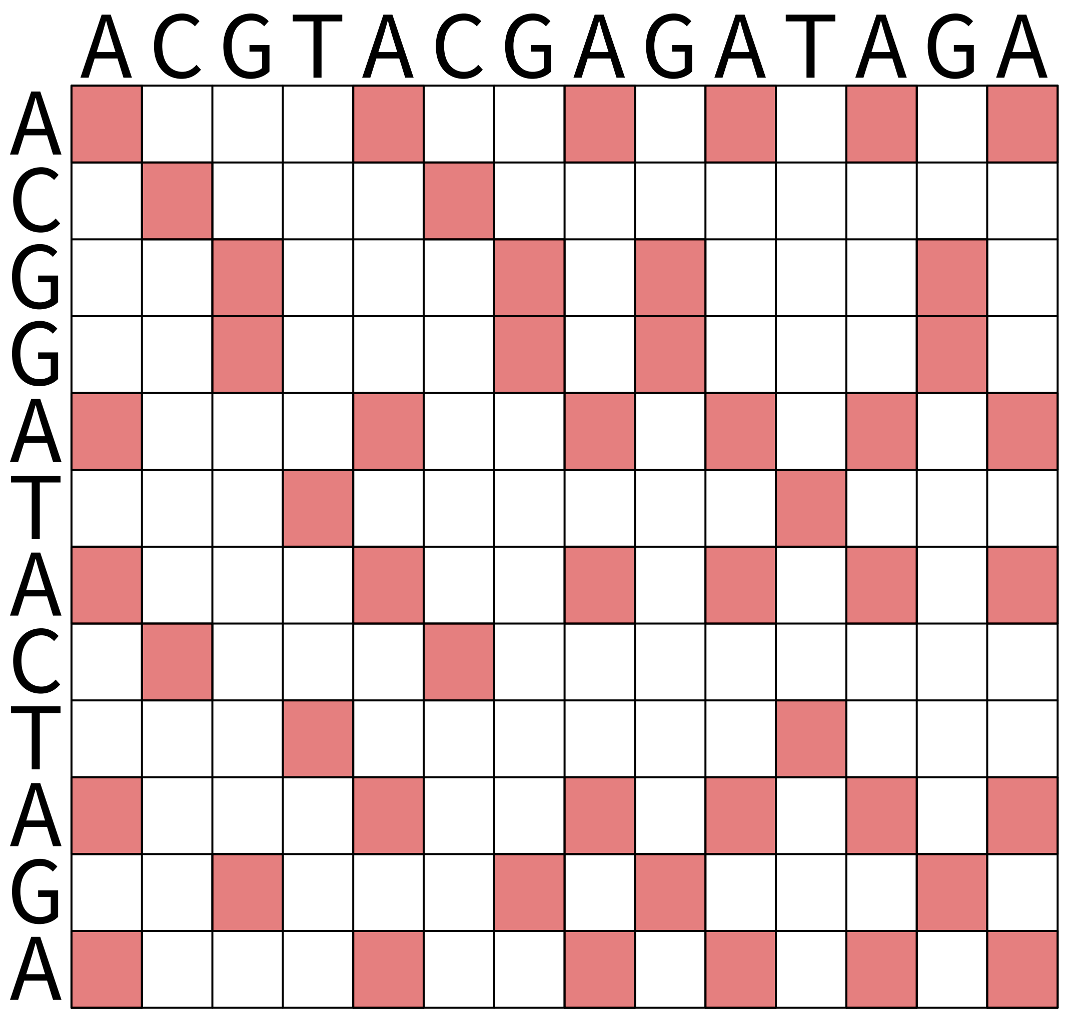
Map locations of short sequences to genomes



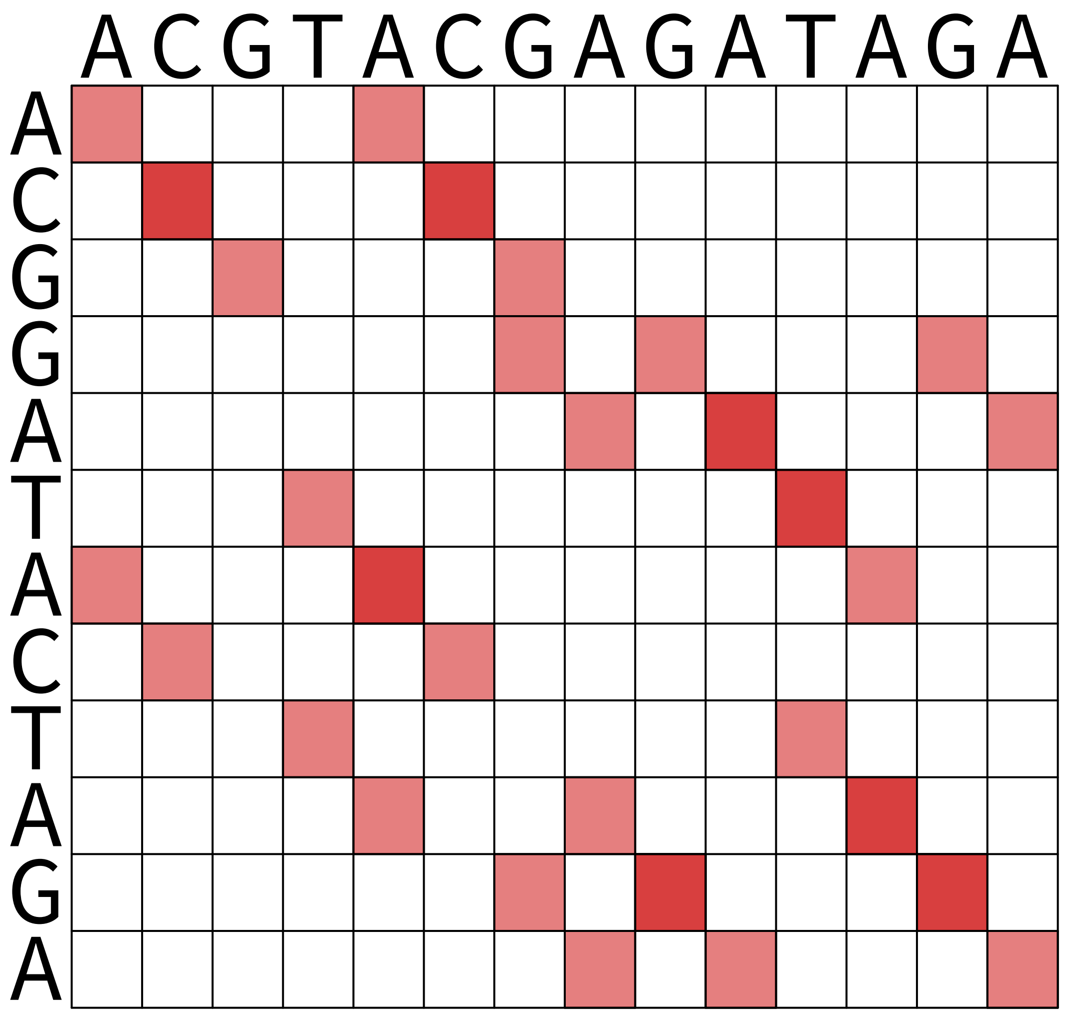
- Identify expressed regions
- Estimate expression from RNA-seq
- Identify sequence variants
- Identify amplified regions
- Evaluate primers and probes
- ...

How to determine an alignment

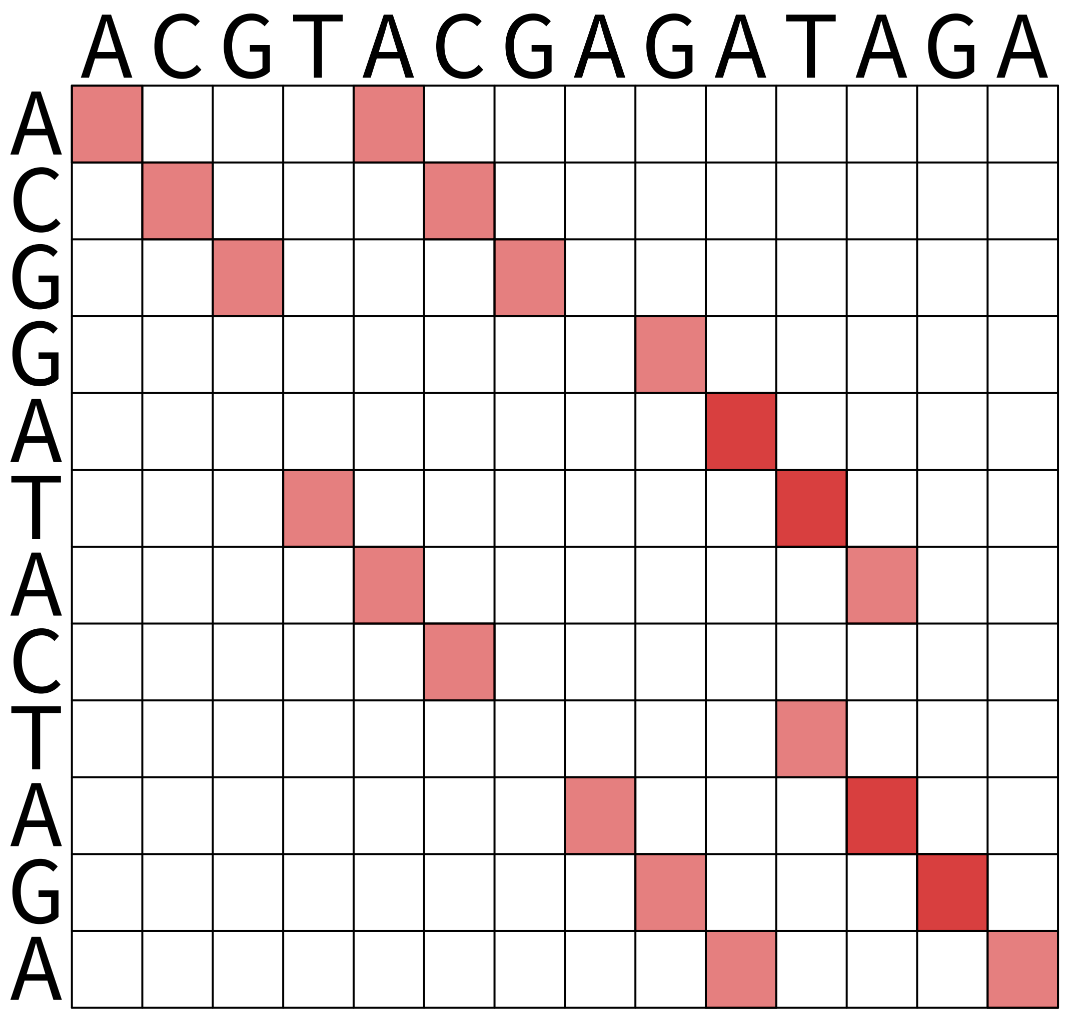
Visualising the alignment space (dot plot)



window=1



window=2

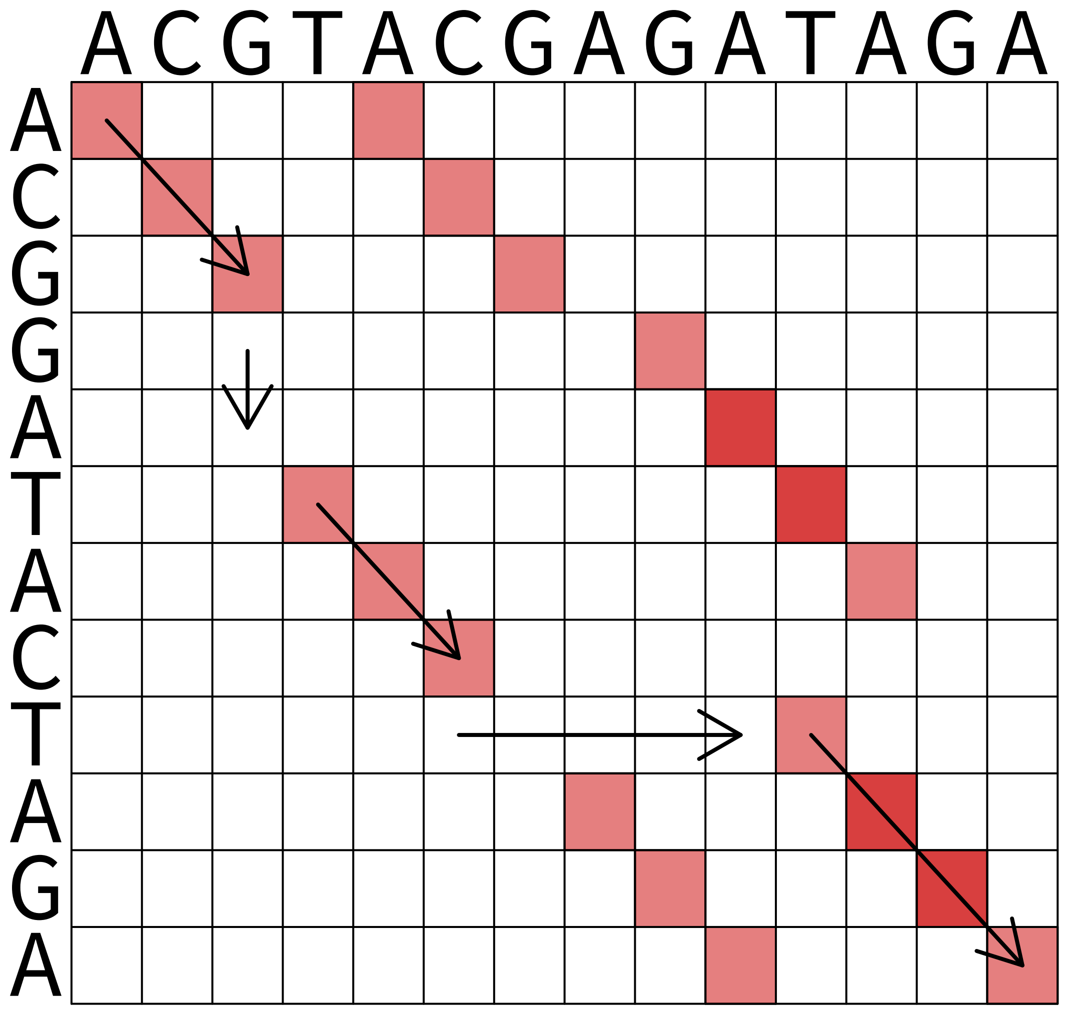


window=3

window: the number of nucleotides used for each comparison

How to determine an alignment

How to choose a path through the matrix



window=3

```
ACG--TACGAGATAGA
|||  |||    ||||
ACGGATAC----TAGA
```

→ Gap inserted into vertical sequence

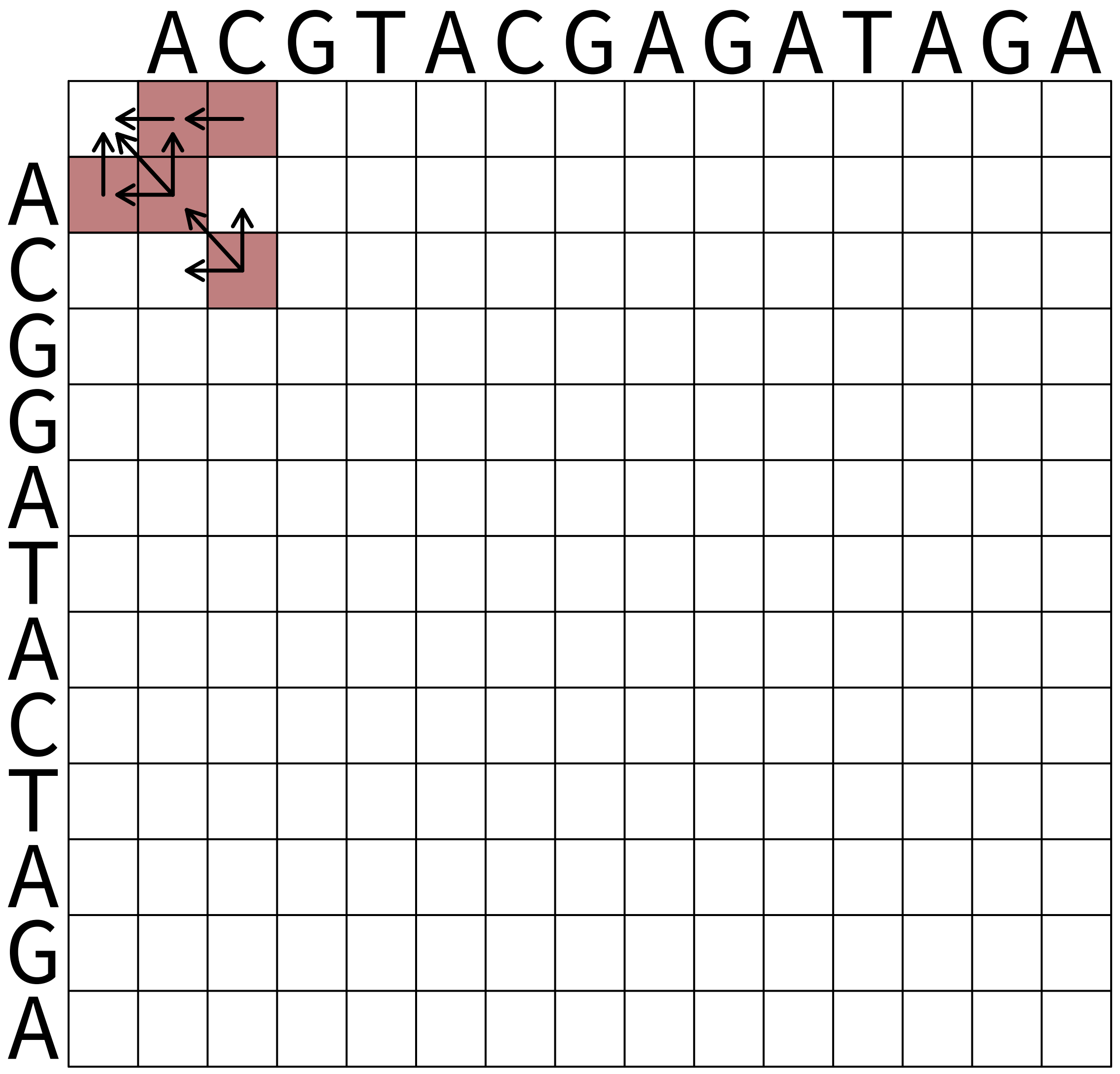
↓ Gap inserted into horizontal sequence

↘ character added from both sequences

How to find the optimal alignment?

How to determine an alignment

How to choose a path through the matrix



What is the optimal alignment of the first pair of nucleotides?

```
-    A    A        -A    A-
A    A    -        A-    -A
```

What is the optimal alignment to position 0,2 in the matrix?

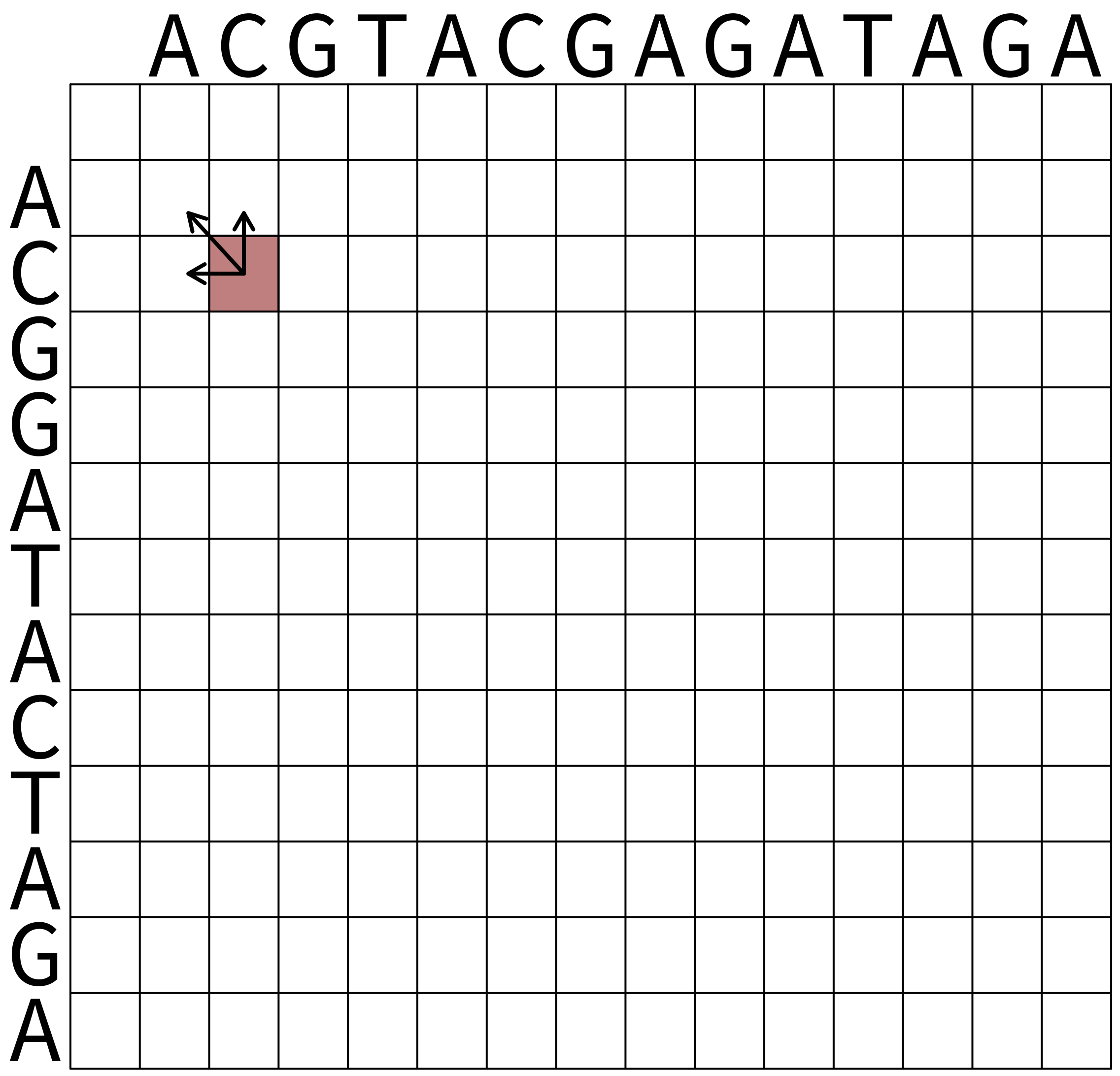
```
AC
--
```

What is the optimal alignment to position 2,2 in the matrix?

depends on scores in (1,1), (1,2), (2,1)

How to determine an alignment

How to choose a path through the matrix



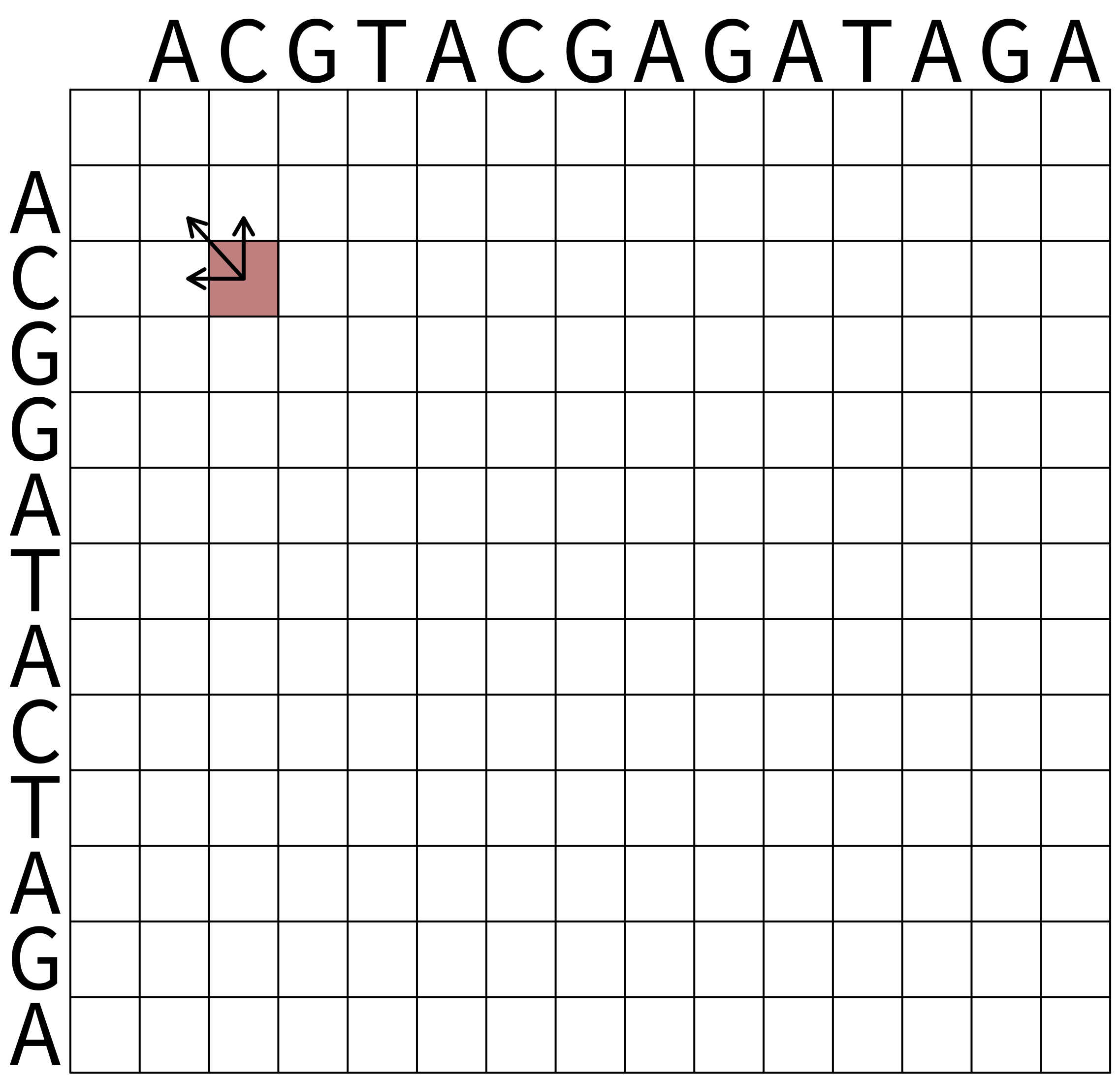
What is the optimal alignment to position 2,2 in the matrix?

The optimal alignment to position (2,2) must include one of:

1. optimal alignment to (1,2)
2. optimal alignment to (1,1)
3. optimal alignment to (2,1)

How to determine an alignment

How to choose a path through the matrix



What is the score of the optimal alignment to position 2,2 in the matrix?

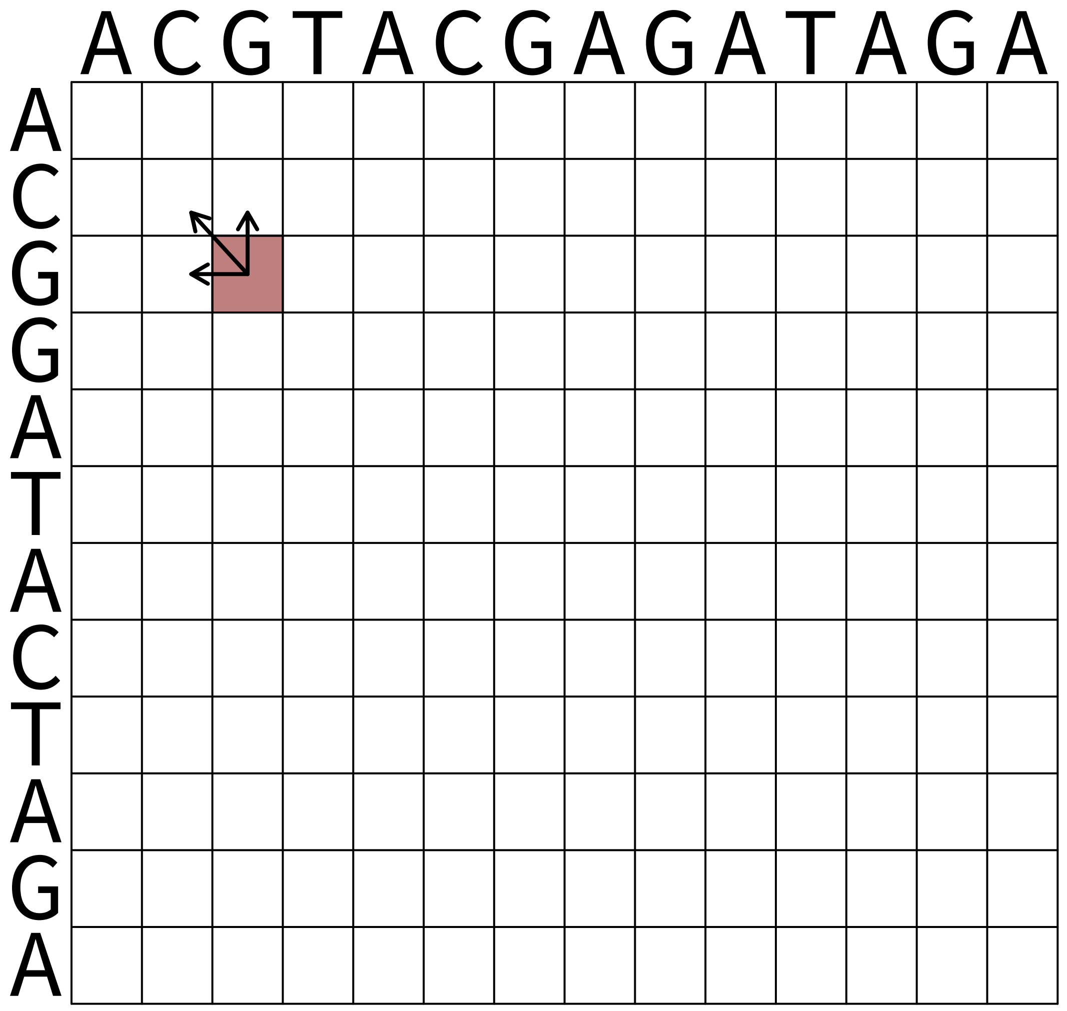
The optimal score at position (2,2) must be one of:

1. optimal score at (1,2) + ?
2. optimal score at (1,1) + ?
3. optimal score at (2,1) + ?

Right and left moves introduce gaps
Diagonal moves align residues to each other

How to determine an alignment

How to choose a path through the matrix



What is the score of the optimal alignment to position 2,2 in the matrix?

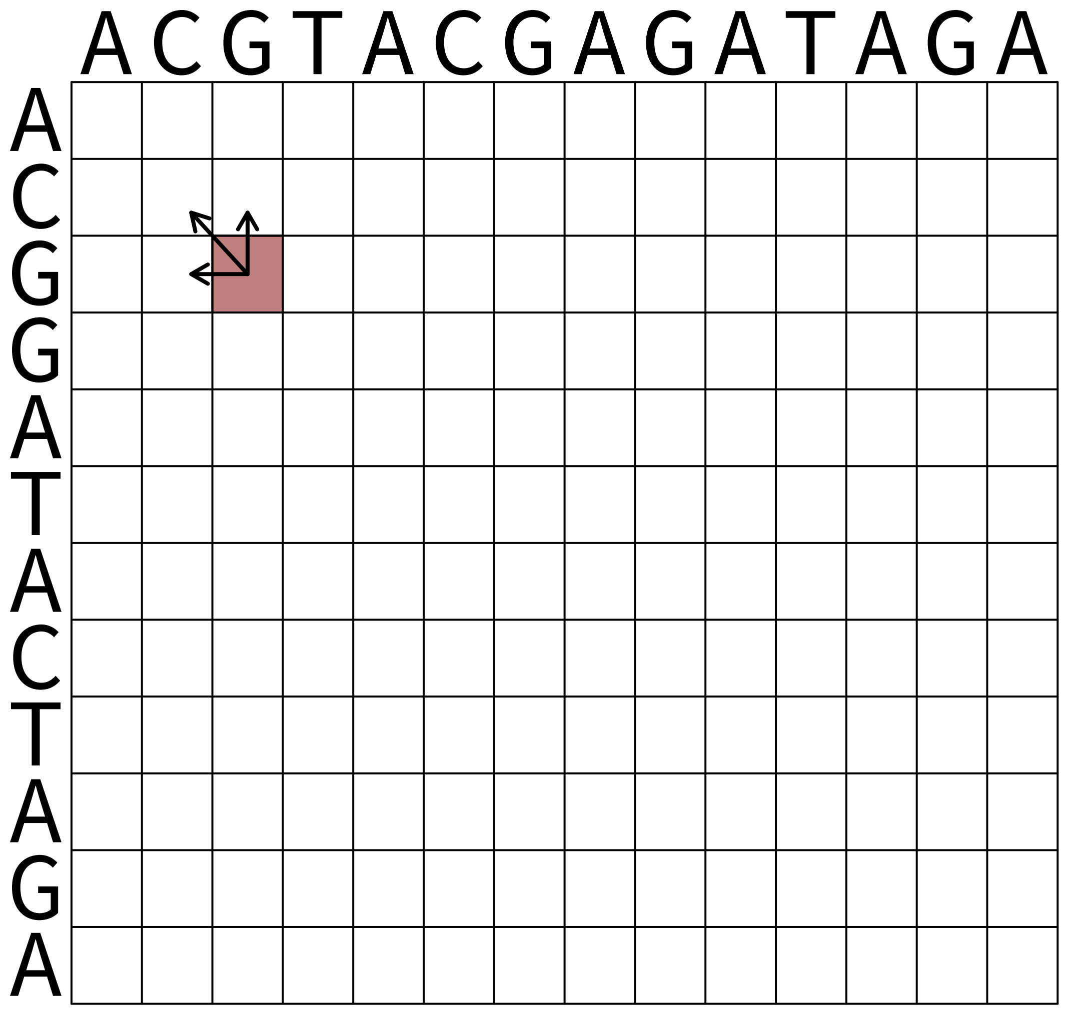
The optimal score at position (2,2) is the maximum of:

1. optimal score at (1,2) + gap
2. optimal score at (1,1) + match/mismatch
3. optimal score at (2,1) + gap

This is the basis for sequence alignment by dynamic programming

How to determine an alignment

How to choose a path through the matrix



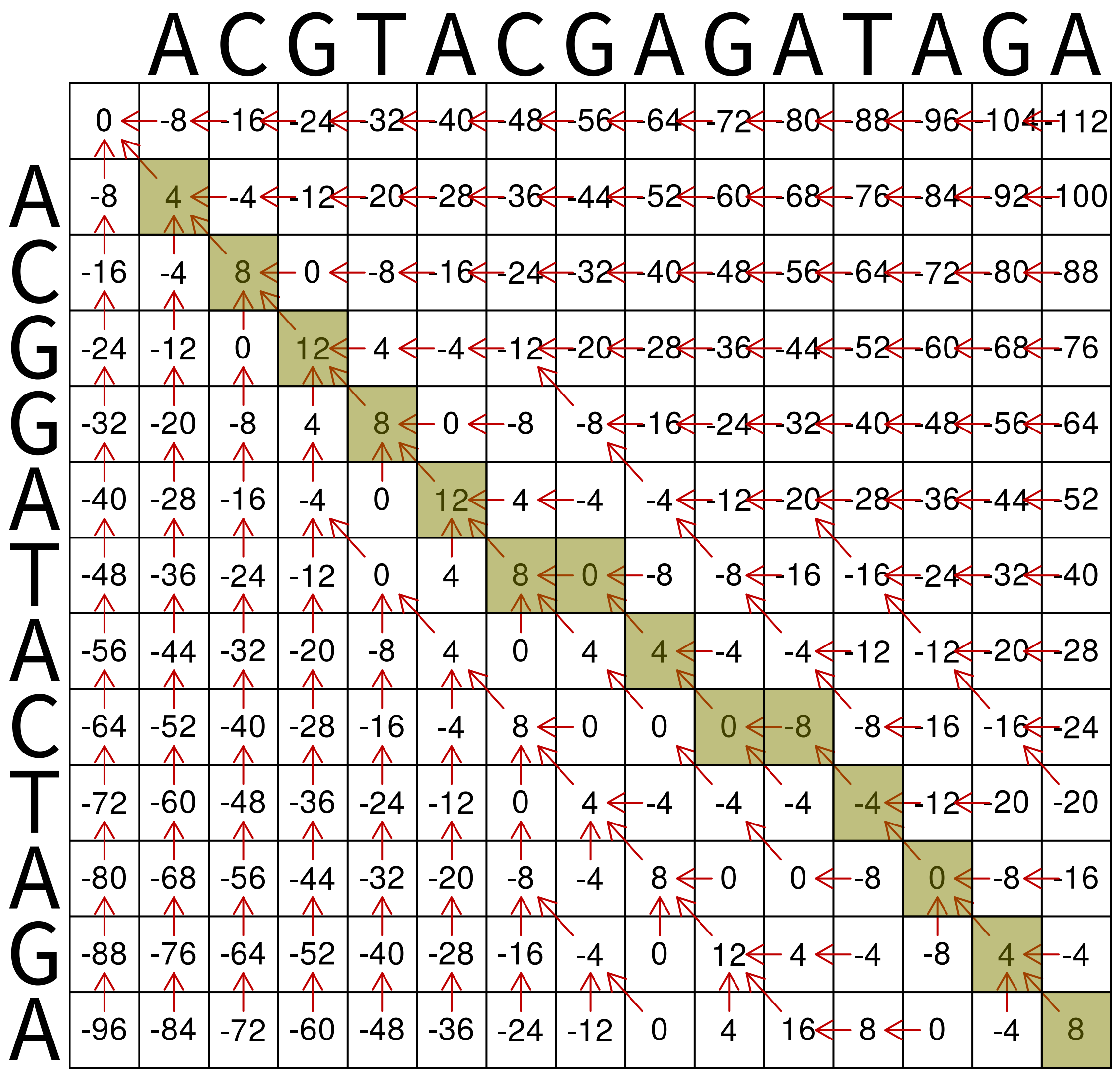
What is the score of the optimal alignment to position i,j in the matrix?

The optimal score at position (i,j) is the maximum of:

1. optimal score at (i-1,j) + gap
2. optimal score at (i-1,j-1) + match/mismatch
3. optimal score at (i,j-1) + gap

This is the Needleman-Wunsch equation

The Needleman-Wunsch algorithm



		A	C	G	T	A	C	G	A	G	A	T	A	G	A
	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80	-88	-96	-104	-112
A	-8	4	-4	-12	-20	-28	-36	-44	-52	-60	-68	-76	-84	-92	-100
C	-16	-4	8	0	-8	-16	-24	-32	-40	-48	-56	-64	-72	-80	-88
G	-24	-12	0	12	4	-4	-12	-20	-28	-36	-44	-52	-60	-68	-76
G	-32	-20	-8	4	8	0	-8	-8	-16	-24	-32	-40	-48	-56	-64
A	-40	-28	-16	-4	0	12	4	-4	-4	-12	-20	-28	-36	-44	-52
T	-48	-36	-24	-12	0	4	8	0	-8	-8	-16	-16	-24	-32	-40
A	-56	-44	-32	-20	-8	4	0	4	4	-4	-4	-12	-12	-20	-28
C	-64	-52	-40	-28	-16	-4	8	0	0	0	-8	-8	-16	-16	-24
T	-72	-60	-48	-36	-24	-12	0	4	-4	-4	-4	-4	-12	-20	-20
A	-80	-68	-56	-44	-32	-20	-8	-4	8	0	0	-8	0	-8	-16
G	-88	-76	-64	-52	-40	-28	-16	-4	0	12	4	-4	-8	4	-4
A	-96	-84	-72	-60	-48	-36	-24	-12	0	4	16	8	0	-4	8

```
A C G T A C G A G A T A G A
| | |   |     |     | | | |
A C G G A T – A C – T A G A
```

1. Set up a score matrix with an additional row and column

2. Set the score at (1,1) to 0

3. Fill the first row and column with gap penalties

4. Cell by cell:

   a) Determine maximum score

   b) Record score

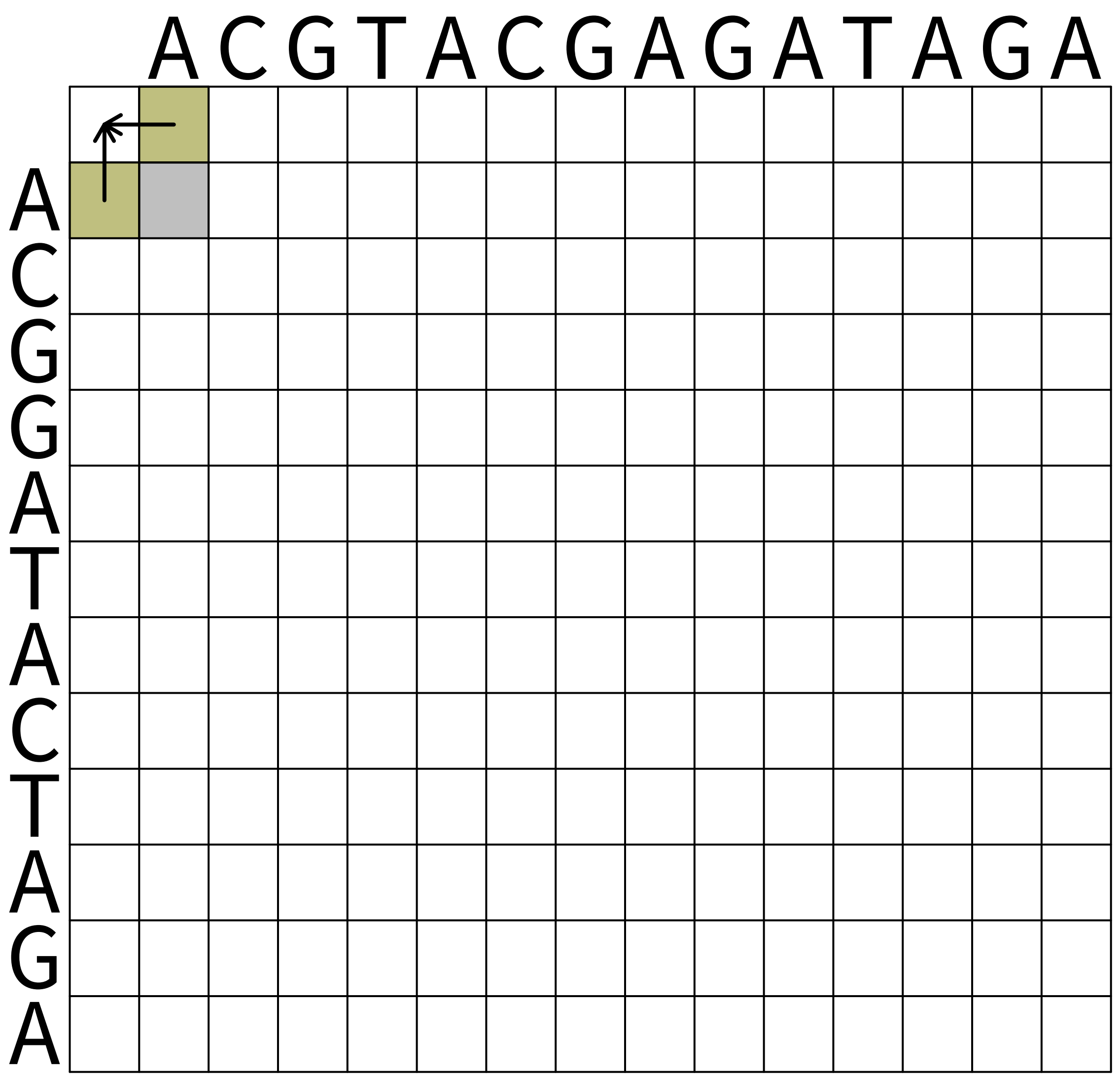
   c) Record the cell from which the alignment was extended

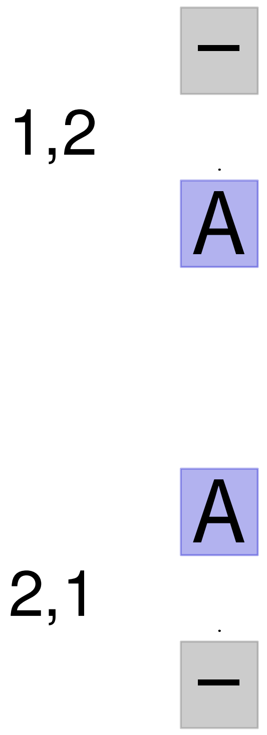
5. Trace alignment from bottom right to top left

The Needleman-Wunsch algorithm

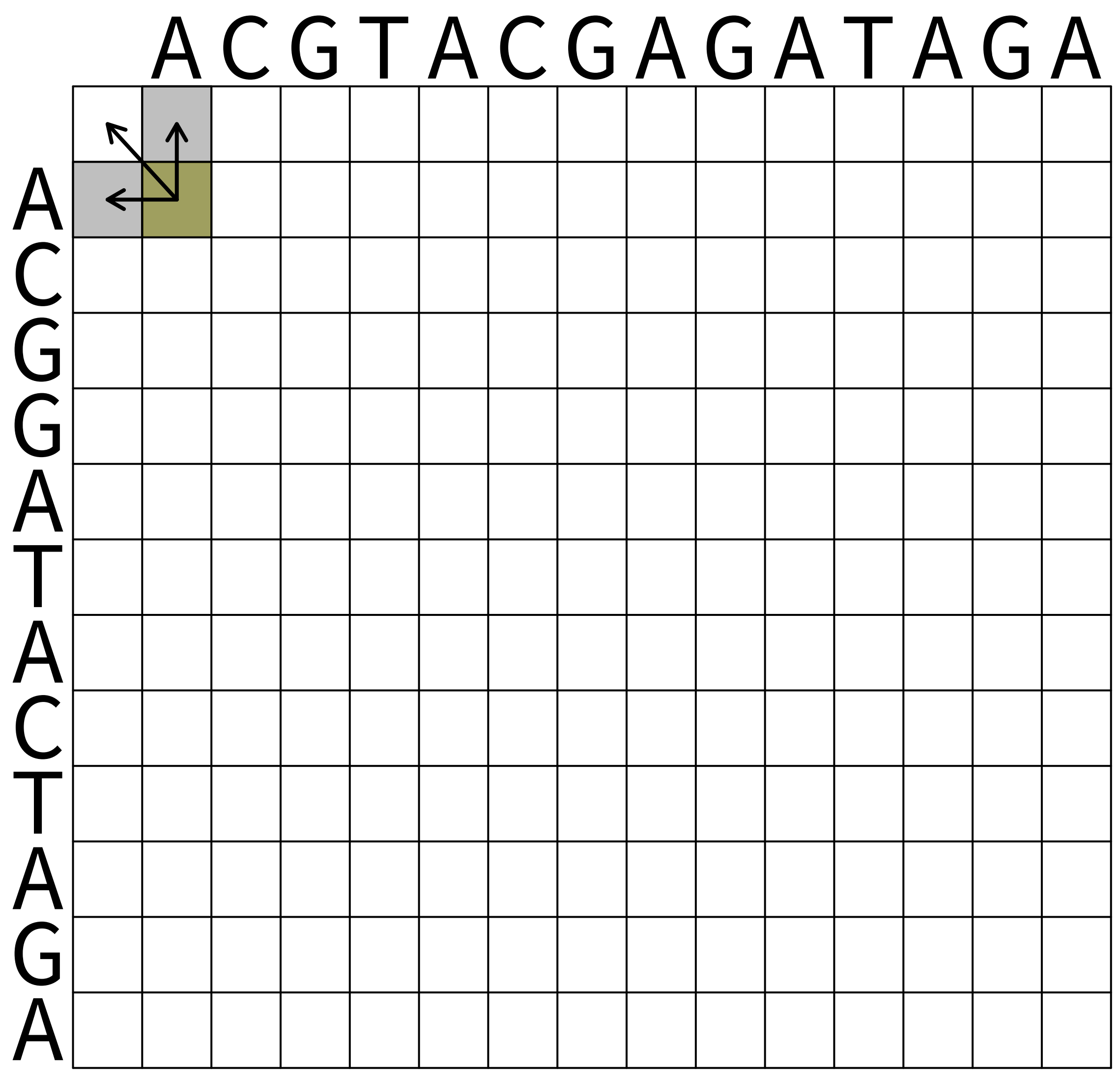
		A	C	G	T	A	C	G	A	G	A	T	A	G	A
A		▨													
C															
G															
G															
A															
T															
A															
C															
T															
A															
G															
A															

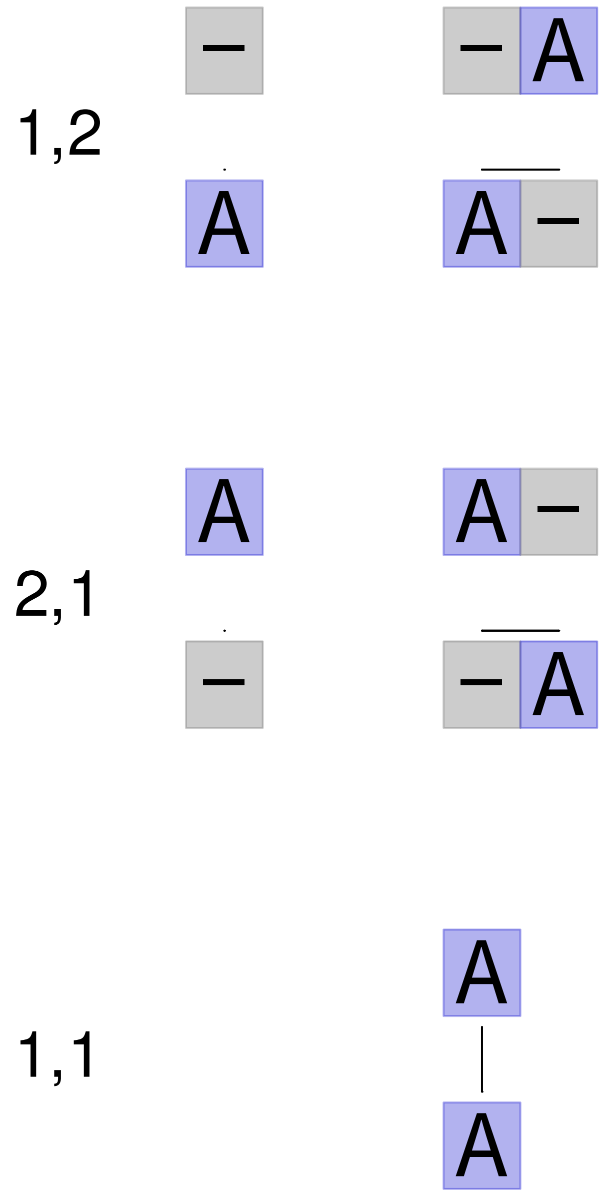
The Needleman-Wunsch algorithm



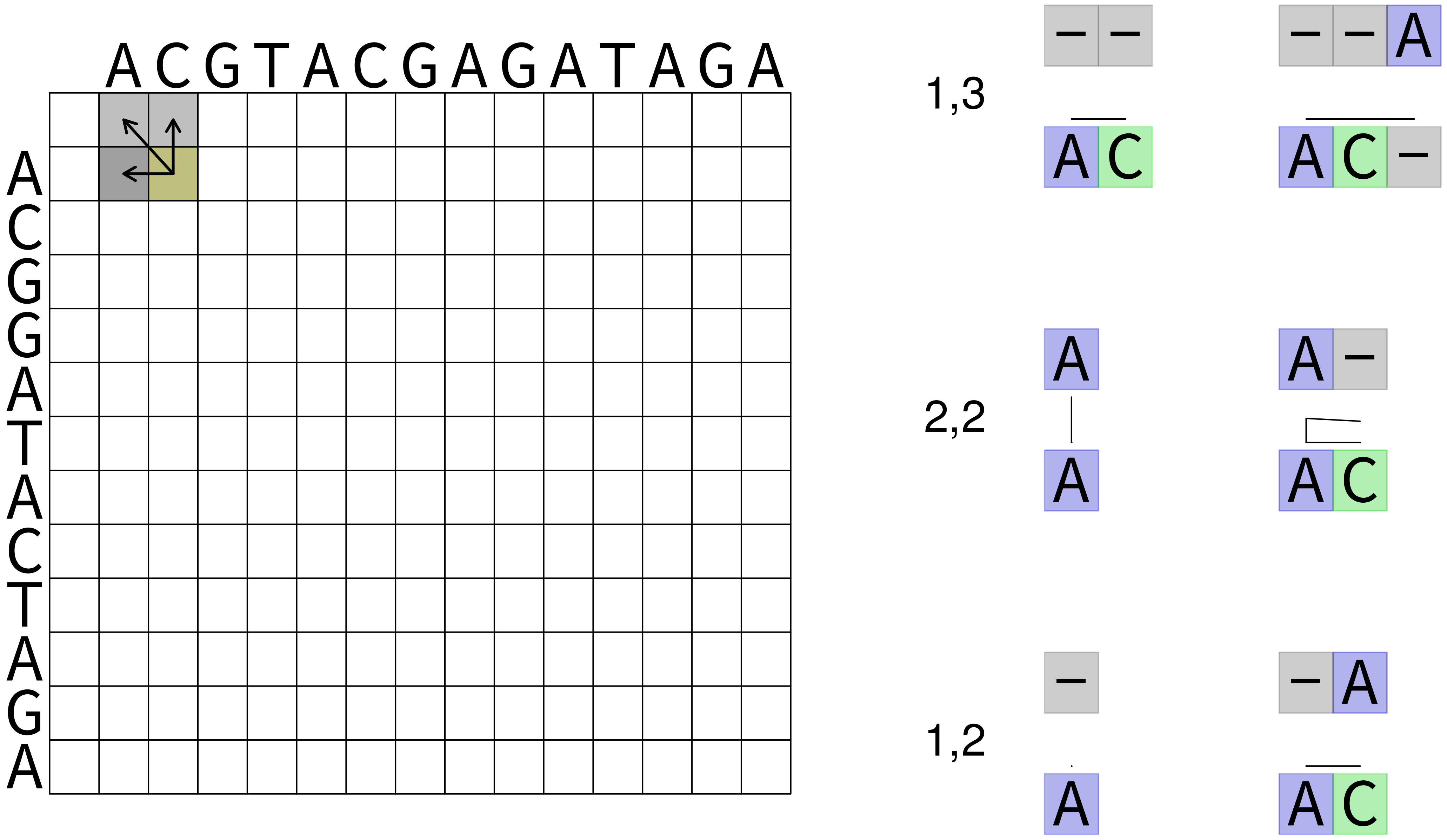


The Needleman-Wunsch algorithm

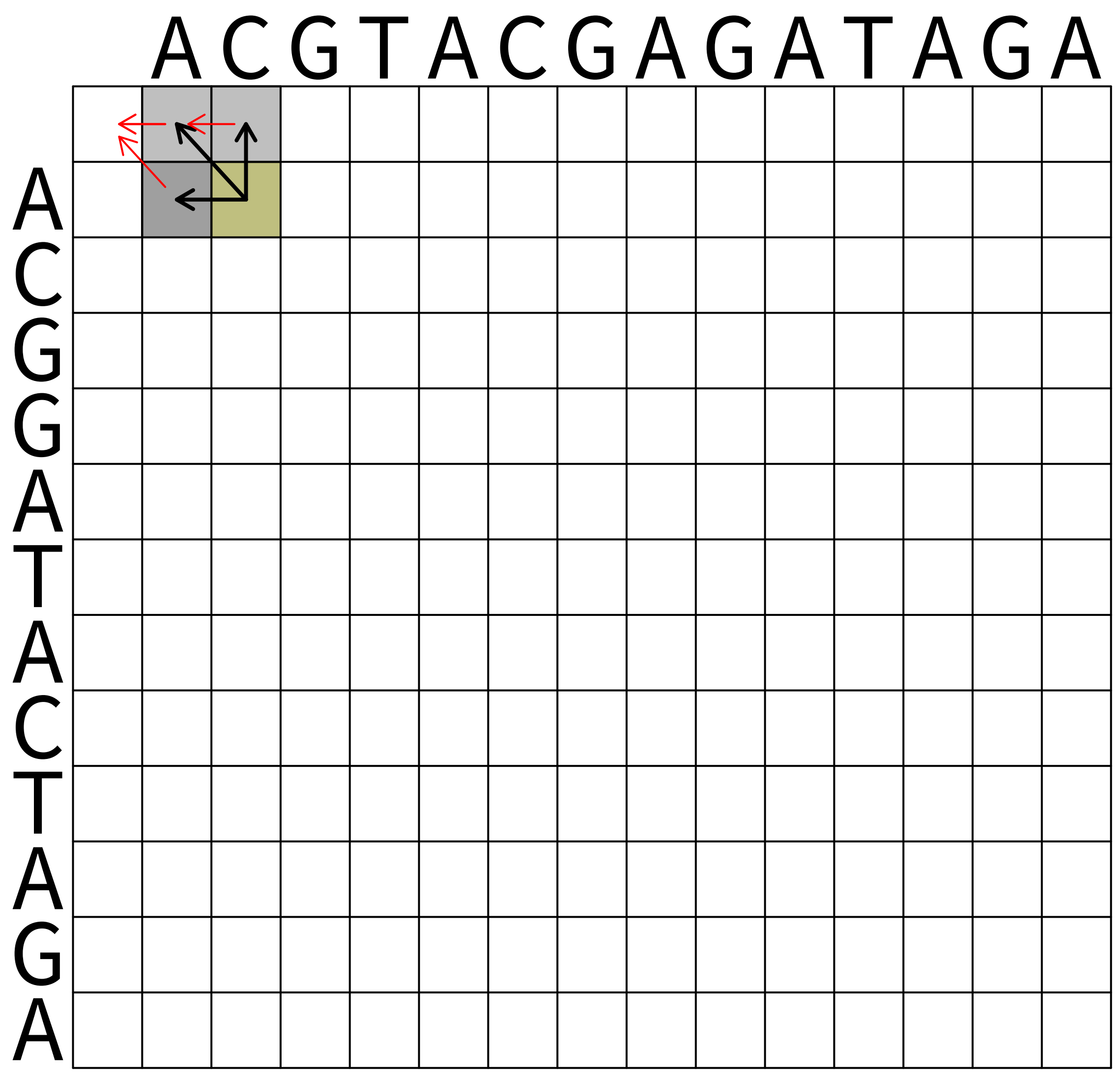


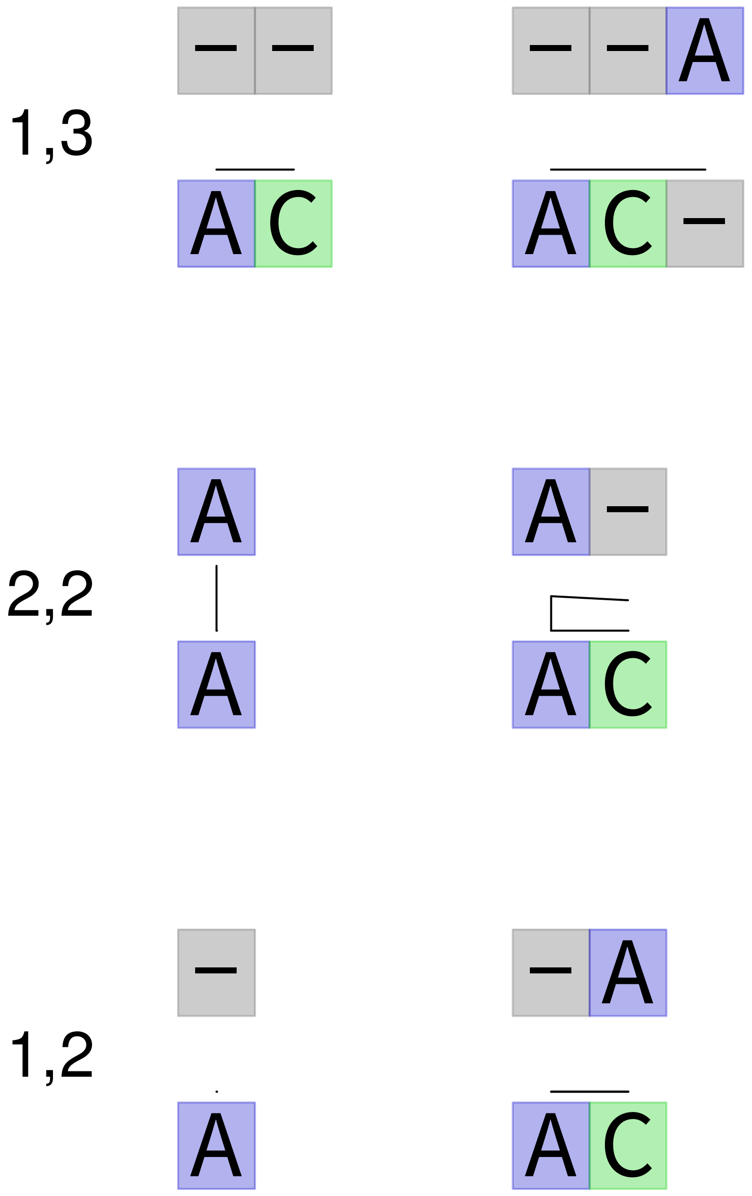


The Needleman-Wunsch algorithm

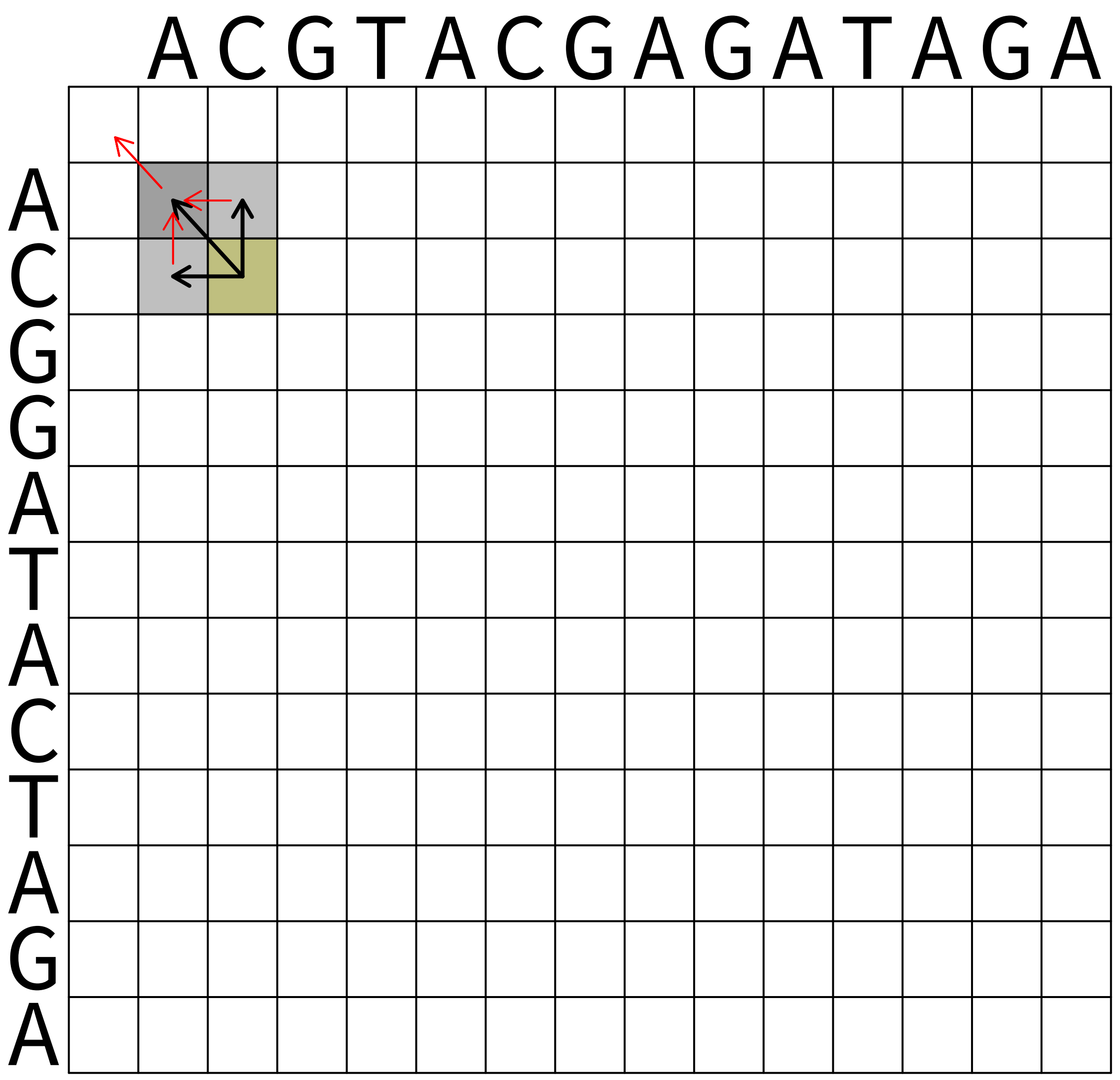


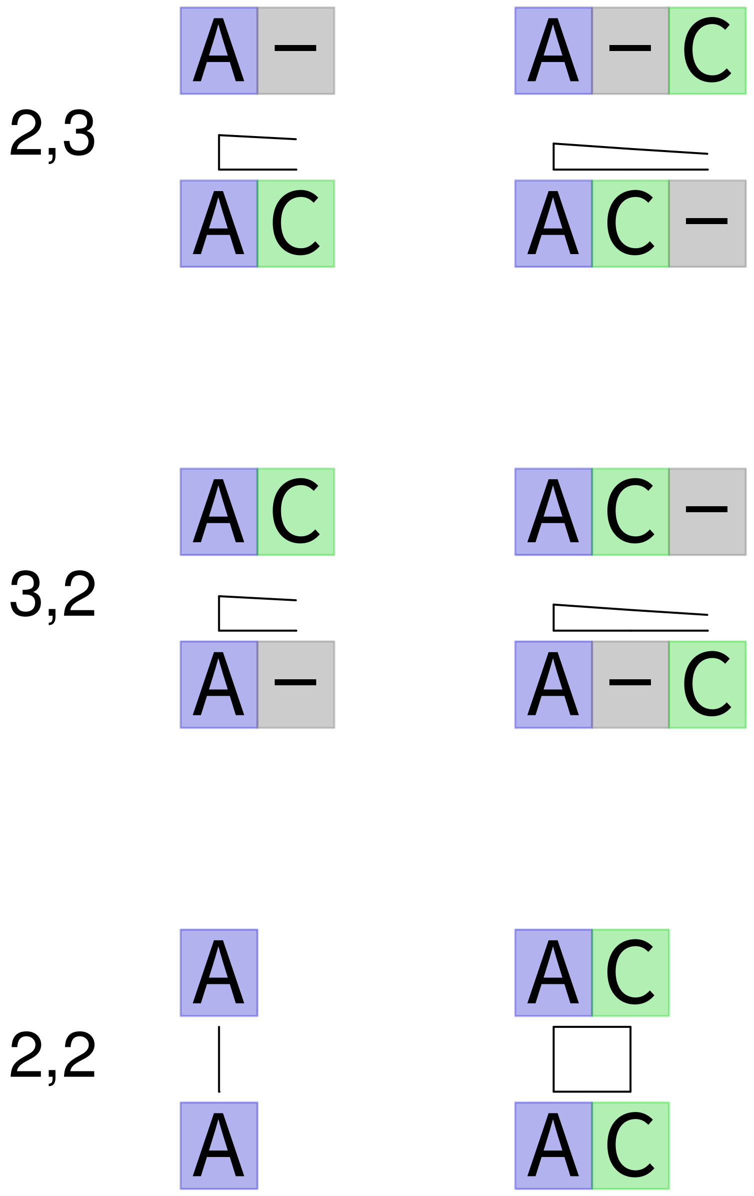
The Needleman-Wunsch algorithm



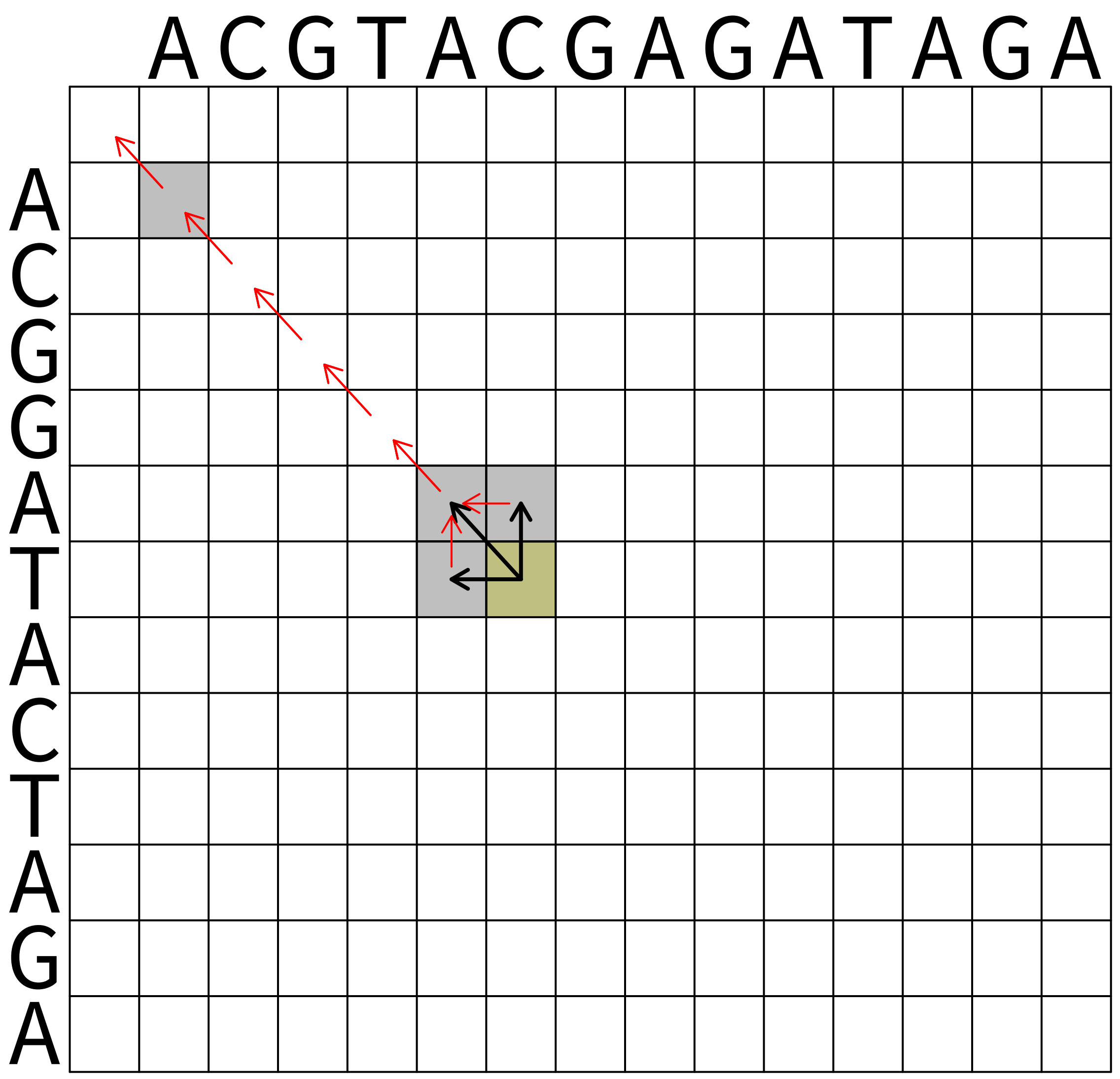


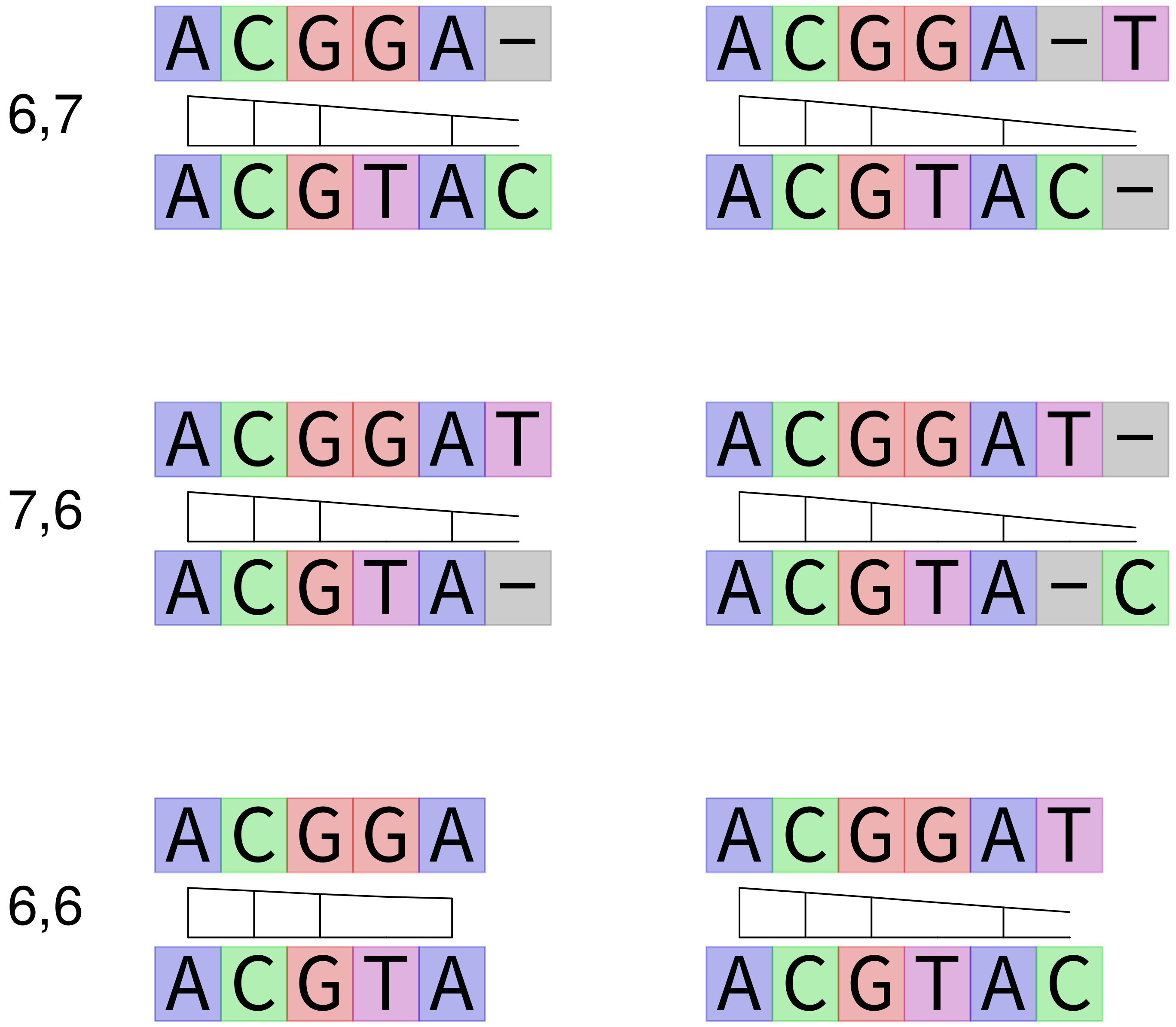
The Needleman-Wunsch algorithm





The Needleman-Wunsch algorithm





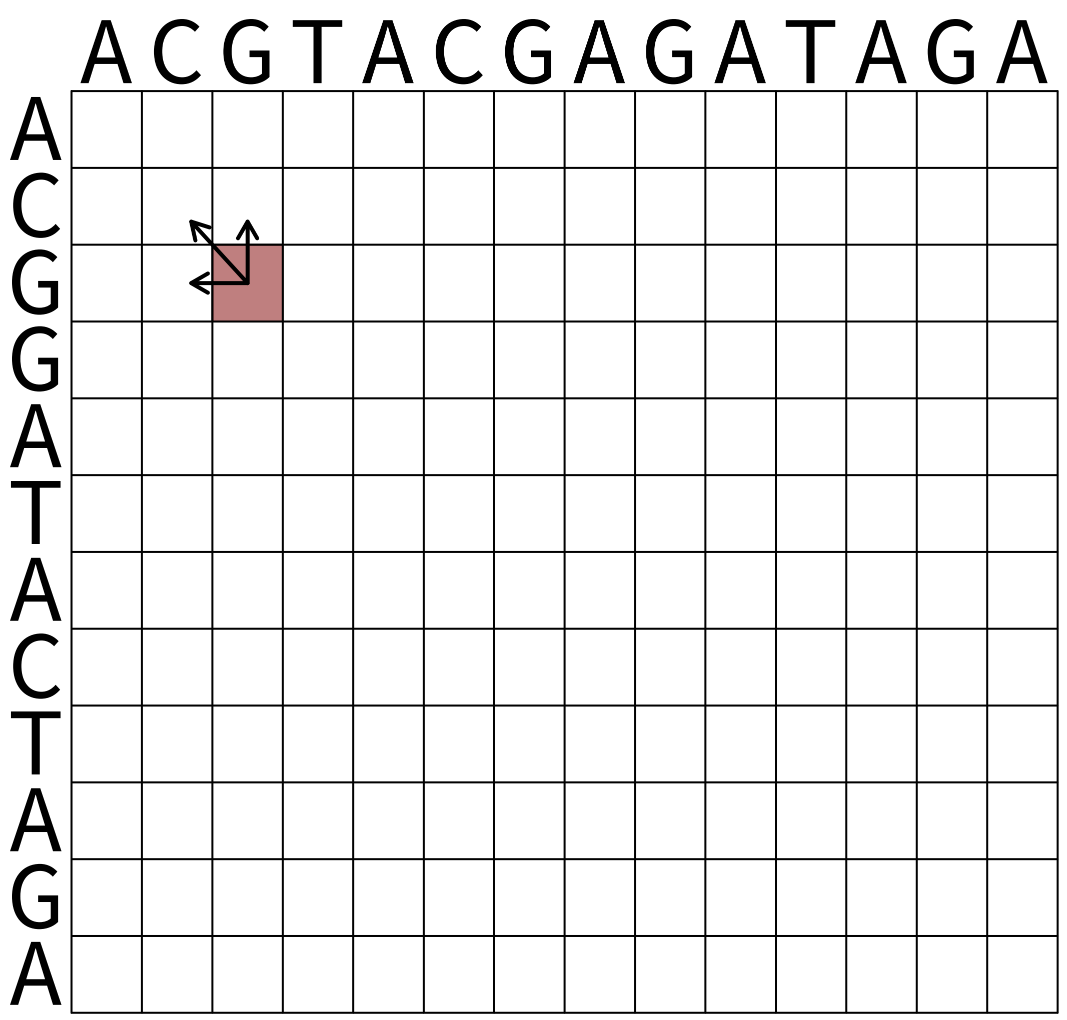
Global vs Local Alignment

- Needleman-Wunsch: Global alignment
- Smith-Waterman: Local alignment

Smith-Waterman is a modification of Needleman-Wunsch

Local alignment

How to choose a sub-path in the matrix



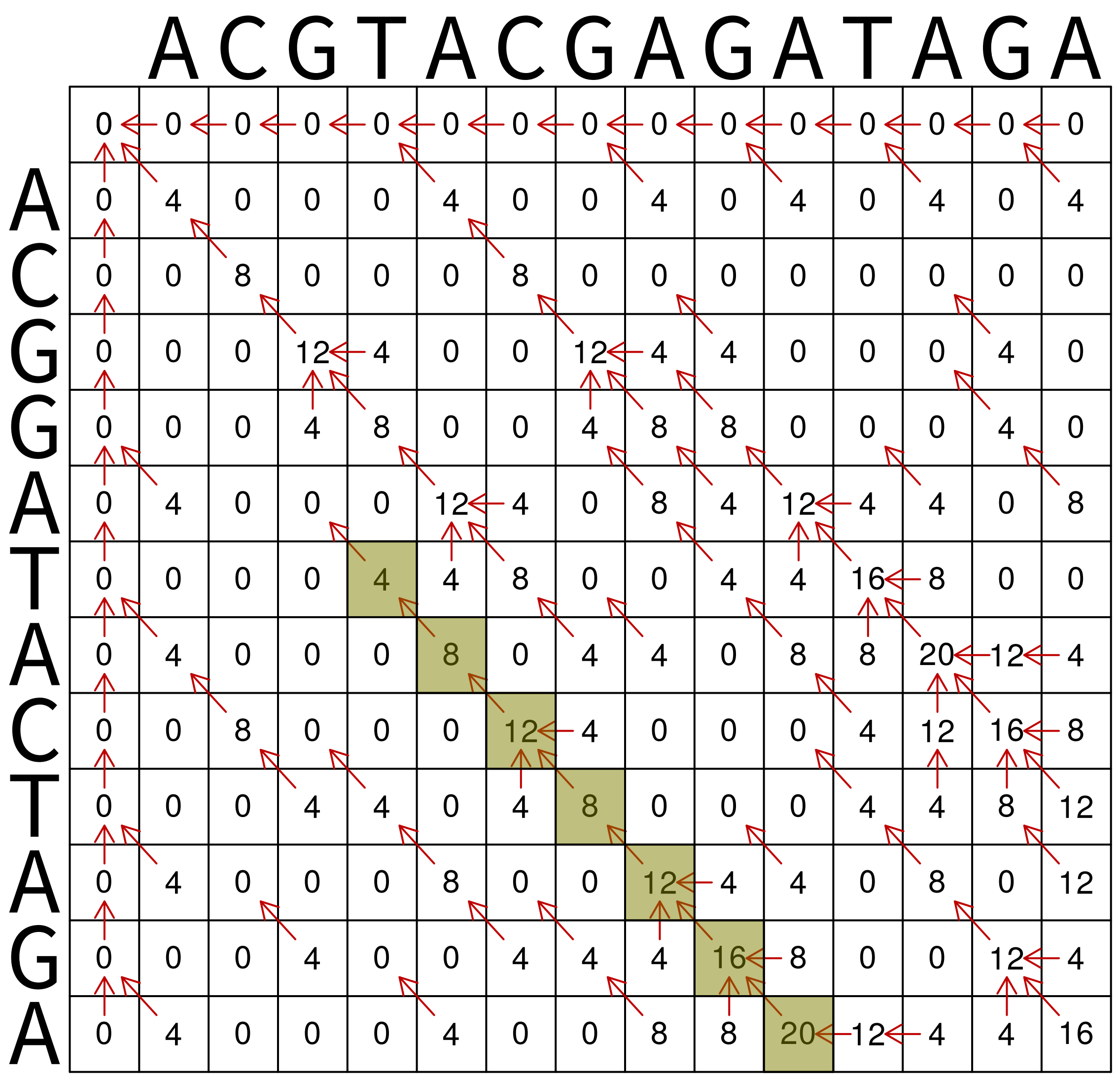
What is the score of the optimal alignment to position i,j in the matrix?

The optimal score at position (i,j) is the maximum of:

1. optimal score at (i-1,j) + gap
2. optimal score at (i-1,j-1) + match/mismatch
3. optimal score at (i,j-1) + gap
4. 0

This is the Smith-Waterman equation

The Smith-Waterman algorithm



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

```
T A C G A G A
| | |   | | |
T A C T A G A
```

1. Set up a score matrix with an additional row and column

2. Set the score at (1,1) to 0

3. Fill the first row and column according to the equation (all 0)

4. Cell by cell:

   a) Determine maximum score

   b) Record score

   c) Record the cell from which the alignment was extended

5. Trace alignment from the cell with the maximum score

Affine gap penalties

-
  ```
  ATTACTTAGGATTATAGA
  ||   | | || | |  |
  AT---T-A-GA-T-T--A
  ```

-
  ```
  ATTACTTAGGATTATAGA
  ||||     |||||
  ATTA-----GATTA----
  ```

Different alignments, but same scores

- Alignment 1 looks bad

- Alignment 2 looks better

Affine gap penalties

```
● ATTACTTAGGATTATAGA
  ||   | | || | |  |
  AT---T-A-GA-T-T--A

● ATTACTTAGGATTATAGA
  ||||     |||||
  ATTA-----GATTA----
```

		S		*Alignment score*
9	9	N_m		*Number of matches*
		P_m	4	*Match penalty*
0	0	N_{mm}		*Number of mismatches*
		P_{mm}	-4	*Mismatch penalty*
6	2	N_{gap_o}		*Number of gap openings*
		P_{gap_o}	-8	*Gap opening penalty*
3	7	N_{gap_e}		*Number of gap extensions*
		P_{gap_e}	-1	*Gap extension penalty*

Modify scoring system:

$$S \leftarrow (N_m \times P_m) + (N_{mm} \times P_{mm}) + (N_{gap\_o} \times P_{gap\_o}) + (N_{gap\_e} \times P_{gap\_e})$$

Substitution matrices

- Mismatch penalties should reflect the likelihood of nucleotide substitutions
- All substitutions are not equally likely to happen during evolution
  - Transitions (purine <-> purine and pyrimidine <-> pyrimidine)
  - Transversions (purine <-> pyrimidine)
  - C -> T more likely at CG positions
- But we used simple mismatch penalty; How to improve?

Separate penalties for different substitutions

	A	C	T	G
A	1	-2	-2	-1
C	-2	1	-1	-2
T	-2	-1	1	-2
G	-1	-2	-2	1

Protein alignments

- Use substitution matrices instead of match / mismatch
  - Mutation distances of codons (Fitch substitution model)
  - Chemical properties of amino acids. Dissimilar pairs of amino acids have larger penalties
  - Observed frequencies in alignments of homologous sequences
    - PAM: Percentage of acceptable point mutations
    - BLOSUM: Blocks substitution matrix
- BLOSUM (and PAM) most commonly used matrices

BLOSUM 62

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	J	Z	X	X.
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	-1	-1	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	-2	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	4	-3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	-3	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-1	-3	-1	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	-2	4	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	-3	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-4	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	-3	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	3	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	-3	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	2	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	0	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-3	-1	-1	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	-2	0	-1	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	-1	-1	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-2	-2	-1	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-1	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	2	-2	-1	-4
B	-2	-1	4	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	-3	0	-1	-4
J	-1	-2	-3	-3	-1	-2	-3	-4	-3	3	3	-3	2	0	-3	-2	-1	-2	-1	2	-3	3	-3	-1	-4
Z	-1	0	0	1	-3	4	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-2	-2	-2	0	-3	4	-1	-4
X	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1