Sequence Alignment

– Objectives

1. Understand the purpose of aligning sequences.
2. Understand the difference between global and local alignment algorithms in terms of specific purpose and usefulness, and alignment matrix.
3. Understanding the difference between pairwise and multiple sequence alignment.

– Learning Outcomes

1. The students will be able to explain the major steps in pairwise and multiple sequence alignment, explain the principle for, and execute sequence alignment using ClustalX.

- The overall learning objectives for today's topic are as follow:

Sequence Alignment

– Why Aligning Sequences?

1. Identifying similarity between two sequences

2. Sequence specifies the function of a protein

3. Similarity in sequence can imply similarity in function.
   - Assign function to uncharacterized sequences based on characterized sequences

4. Sequence from different species can be compared to estimate the evolutionary relationships between species
   - We will come back to this in Phylogenetic trees

– Similarity VS Identity

Term 1 – Similarity

1. Statistical assessment on how well database and query sequences match that can infer homology.
2. This term usually being used in amino acid sequence similarity searching.

Term 2 – Identity

1. Sequence identity is the amount of characters which match exactly between two different sequences.
2. Hereby, gaps are not counted and the measurement is relational to the shorter of the two sequences.

– Global VS Local

```
Global FTFTALILLAVAV
       F--TAL-LLA-AV
```

Type 1 – Global Alignment *(The Needleman-Wunsch algorithm)*

1. A global alignment performs an end-to-end alignment of the query sequence with the reference sequence.
2. Ideally, this alignment technique is most suitable for closely related sequences of similar lengths.

```
Local  FTFTALILL-AVAV
       --FTAL-LLAAV--
```

Type 2 – Local Alignment *(The Smith–Waterman algorithm)*

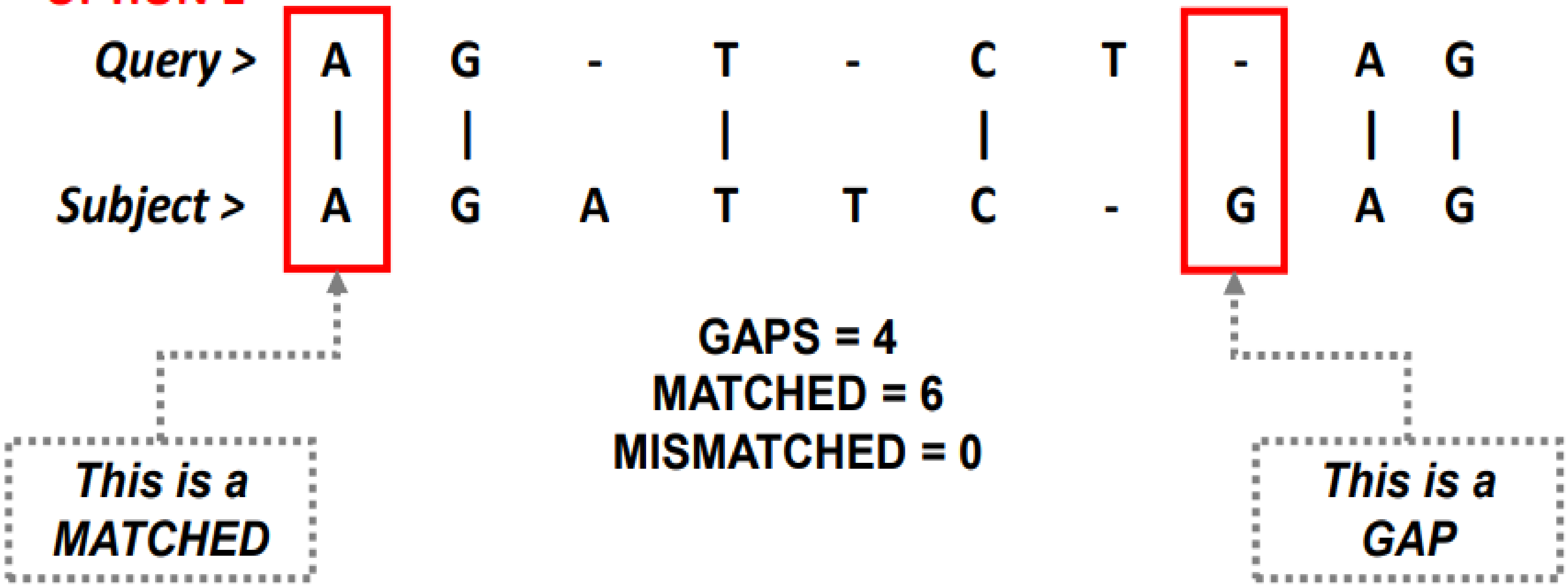
1. A local sequence alignment matches a contiguous sub-section of one sequence with a contiguous sub-section of another.

– Simple Alignment

Alignment

Query >	A	G	T	C	T	A	G			*7nt*
Subject >	A	G	A	T	T	C	G	A	G	*9nt*

OPTION 1



```
Query >    A  G  -  T  -  C  T  -  A  G
           |  |     |     |        |  |
Subject >  A  G  A  T  T  C  -  G  A  G
```

GAPS = 4
MATCHED = 6
MISMATCHED = 0

This is a MATCHED

This is a GAP

- A simple sequence alignment can generate several types of alignment profiles.
- This is because the sequence alignment algorithm will try to find the optimal matching profiles of all the regions in both sequences.
- These algorithms are LOCAL and GLOBAL alignment algorithms which we have discussed in the previous lecture.
- Regardless of these algorithms, the optimal alignment profile is determined by the percentages of matched nucleotides, mismatches and regions of gaps.
- Let's take a look at other types of alignment profiles then.

Simple Alignment

Alignment

```
Query >    A   G   T   C   T   A   G          7nt
Subject >  A   G   A   T   T   C   G   A   G  9nt
```

OPTION 3

```
Query >    A   G   T   C   T   -   -   A   G
           |   |   X   X   |           |   |
Subject >  A   G   A   T   T   C   G   A   G
```

GAPS = 2
MATCHED = 5
MISMATCHED = 2

- Meanwhile in the third alignment profile you have 2 gaps, 5 matched and 2 mismatched regions. Did you noticed the difference of alignment profiles with the first profile, the second and the third?

Alignment

```
Query >    A   G   T   C   T   A   G           7nt
Subject >  A   G   A   T   T   C   G   A   G   9nt
```

OPTION 1

GAPS = 4
MATCHED = 6
MISMATCHED = 0

```
Query >   A G - T - C T - A G
          | |   |   |     | |
Subject > A G A T T C - G A G
```

OPTION 2

GAPS = 2
MATCHED = 6
MISMATCHED = 1

```
Query >   A G - - T C T A G
          | |     | | X | |
Subject > A G A T T C G A G
```

OPTION 3

GAPS = 2
MATCHED = 5
MISMATCHED = 2

```
Query >   A G T C T - - A G
          | | X X |     | |
Subject > A G A T T C G A G
```

OPTION 4

GAPS = 4
MATCHED = 6
MISMATCHED = 0

```
Query >   A G - T C T - - A G
          | |   |   |     | |
Subject > A G A T - T C G A G
```

Which one has the best optimal alignment?

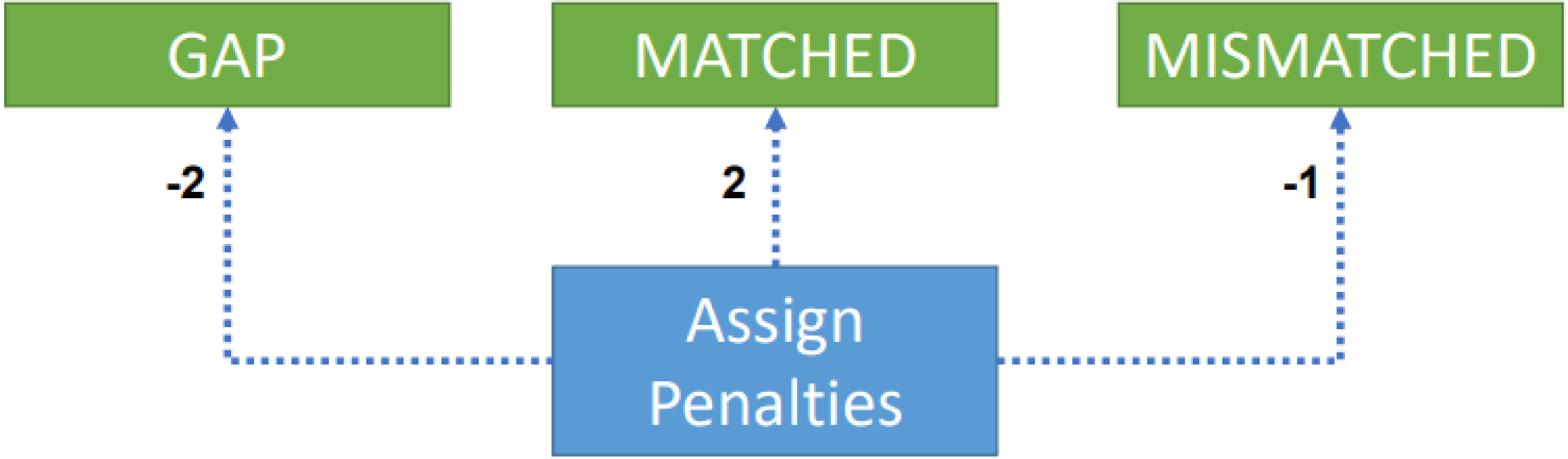
How do you chose the best alignment?

- Everything seems CORRECT! Don't you think so?
- However, in sequence alignment algorithm, it requires only one arrangement that gives the HIGHEST MATCHED SCORE.
- How do I know which one?
- How do you chose which one is the best alignment?

– *Simple Alignment*

Alignment

Query >	**A**	**G**	**T**	**C**	**T**	**A**	**G**			*7nt*
Subject >	**A**	**G**	**A**	**T**	**T**	**C**	**G**	**A**	**G**	*9nt*



GAP

MATCHED

MISMATCHED

-2

2

-1

Assign Penalties

- The fundamental way to determine which alignment profile has the best optimal sequence alignment is by assigning a different VALUE to GAP, MATCHED and MISMATCHED.

- These values are what we call as PENALTIES, that will be used to assessed the sequence similarities between these sequences.

- You may assign any integers to these penalties but generally we will be using

- GAP penalty as -2
- MATCHED penalty as 2
- and MISMATCHED penalty as -1

Scoring Matrix & Dynamic Programming Algorithm

OPTION 2

```
Query >    A   G   -   -   T   C   T   A   G      7nt
           |   |           |   |   X   |   |
Subject >  A   G   A   T   T   C   G   A   G      9nt
           2   2  -2  -2   2   2  -1   2   2       7
```

What is Scoring Matrix?

1. The aim of a sequence alignment is to match "the most similar elements" of at least two sequences (pairwise alignment). This similarity must be evaluated somehow.

2. The previous few slides gave you an overview of how to find the best sequence alignment – manually & laborious. Is there any other way?

3. The answer is yes – we use Scoring Matrix method to align longer and much more complex sequences.

4. What type of information can I get from this matrix?

Score: A number used to assess the biological relevance of a finding.

In the context of sequence alignments, a score is a numerical value that describes the overall quality of an alignment. Higher numbers correspond to higher similarity. The score scale depends on the scoring system used.

Alignment

Query >	A	G	T	C	T	A	G			*7nt*
Subject >	A	G	A	T	T	C	G	A	G	*9nt*

How do I do it?

FIRST:

Assign a value for MATCHED, MISMATCHED and GAP. They could be random values but common practise is ‘ 2 ’ for MATCHED penalty, ‘ -2 ’ for GAP penalty and ‘ -1 ’ for MISMATCHED penalty. These values will be used to calculate the probability of highest IDENTITY (how similar they are) between two sequences.

SECOND:

Do the maths! Take a look at the next slide.

Global Alignment

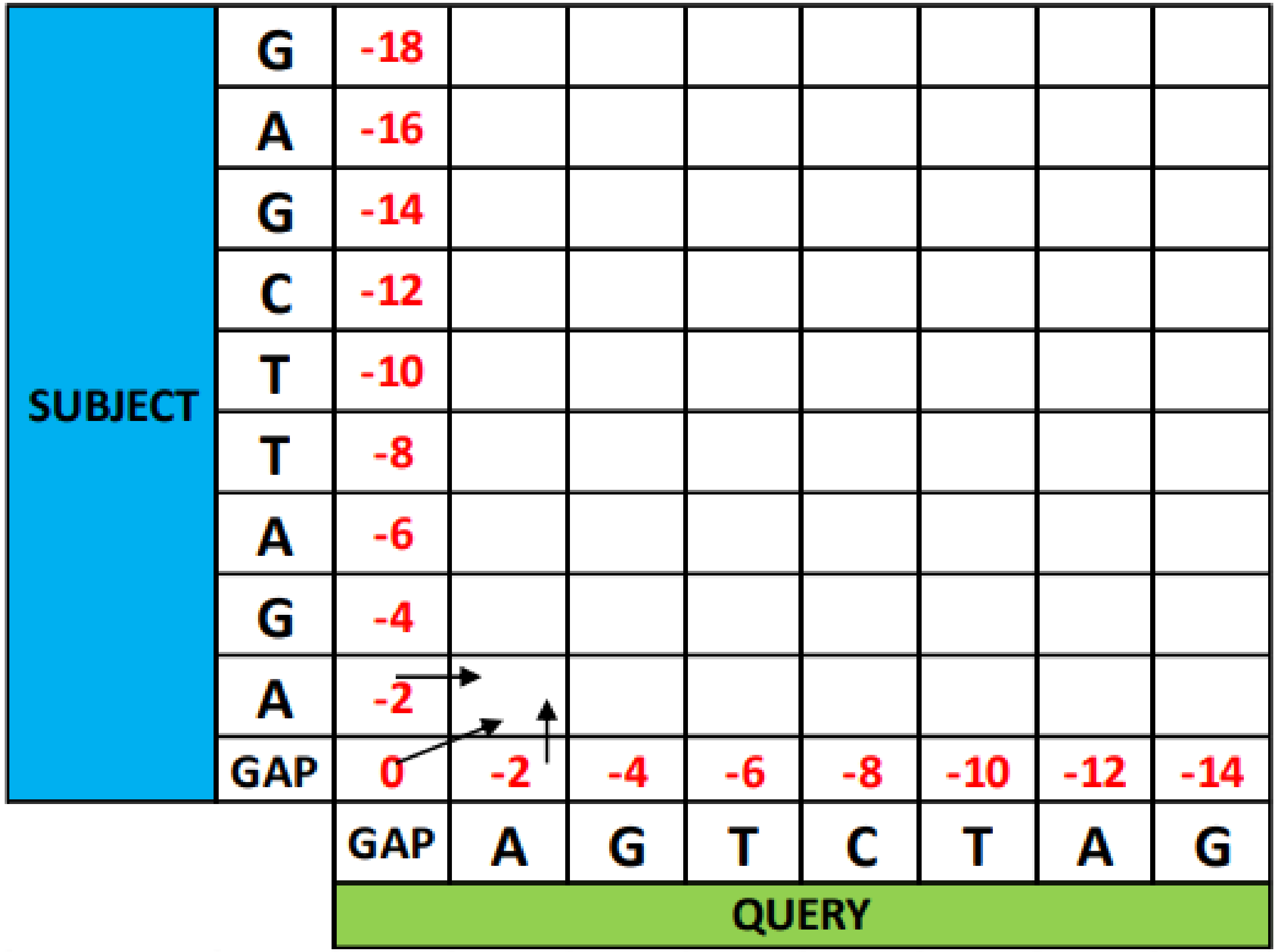
SUBJECT	G	-18							
	A	-16							
	G	-14							
	C	-12							
	T	-10							
	T	-8							
	A	-6							
	G	-4							
	A	-2							
	GAP	0	-2	-4	-6	-8	-10	-12	-14
		GAP	A	G	T	C	T	A	G
		QUERY							

What is Scoring Matrix?

1. Scoring matrices are used to determine the relative score made by matching two characters in a sequence alignment.

2. These are usually the likelihood of two characters being derived from a common ancestral character.

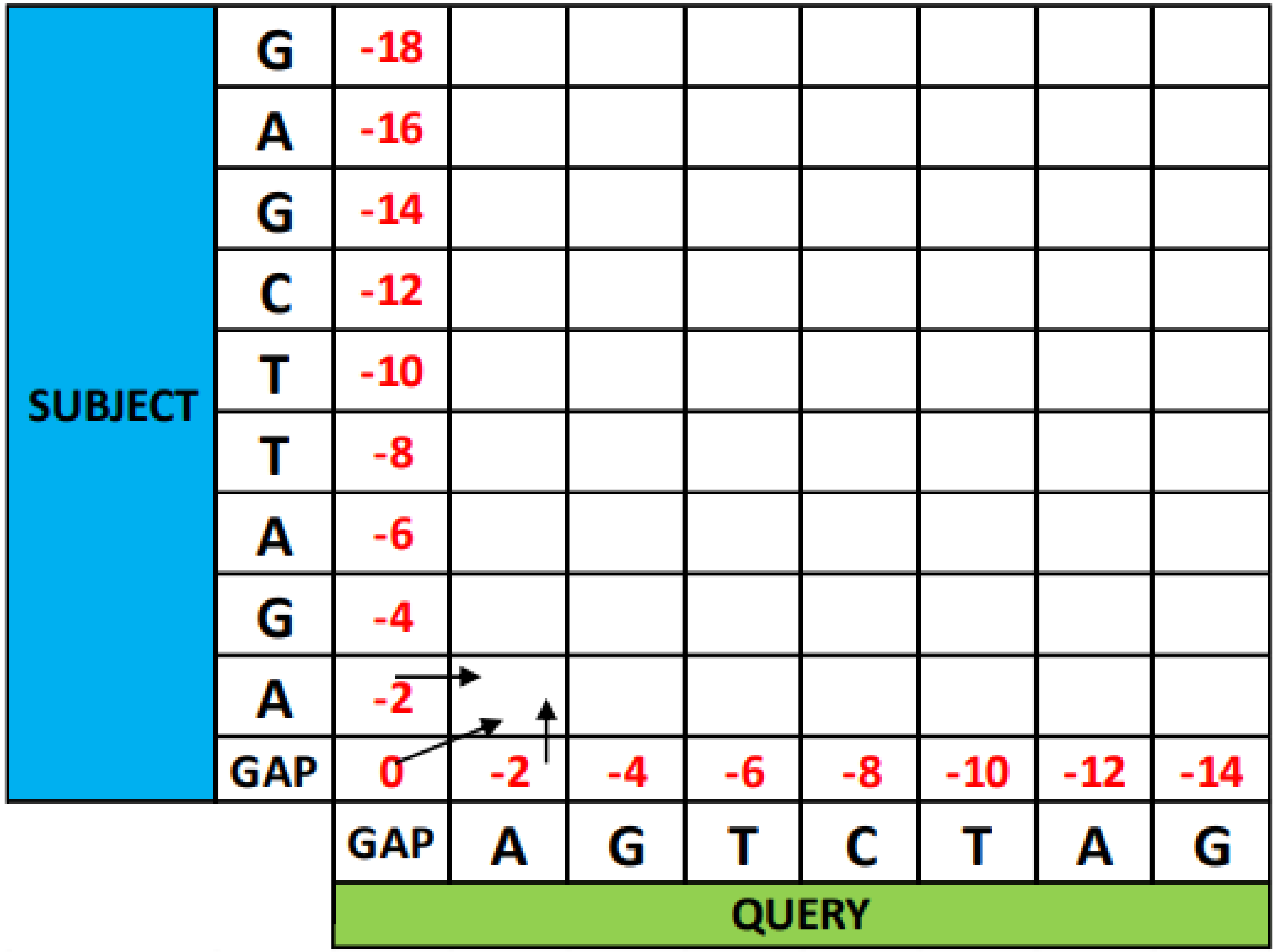
Global Alignment



SUBJECT	G	-18							
	A	-16							
	G	-14							
	C	-12							
	T	-10							
	T	-8							
	A	-6							
	G	-4							
	A	-2							
	GAP	0	-2	-4	-6	-8	-10	-12	-14
		GAP	A	G	T	C	T	A	G
		QUERY							

RULES

1. To fill in the empty boxes, you must use the values of GAP, MATCHED and MISMATCHED.
2. The values of any box can come from bottom box, left box, and bottom-left box.



SUBJECT	G	-18							
	A	-16							
	G	-14							
	C	-12							
	T	-10							
	T	-8							
	A	-6							
	G	-4							
	A	-2							
	GAP	0	-2	-4	-6	-8	-10	-12	-14
		GAP	A	G	T	C	T	A	G
		QUERY							

RULES

1. To fill in the empty boxes, you must calculate the value from each direction.

2. The highest value will become the new value of that empty box.

A (from Subject) and A (from Query) is a match. Therefore:

Value from 1 (horizontal) = - 2 + (- 2) *eq* -4

Value from 2 (vertical) = - 2 + (- 2) *eq* -4

Value from 3 (diagonal) = 0 + 2 *eq* 2

NEEDLEMAN WUNSCH ALGORITHM

- MATCH : 1
- MISMATCH : -1
- GAP : -2

		T	A	T	G	A
	0	-2	-4	-6	-8	-10
T	-2					
A	-4					
C	-6					
G	-8					
A	-10					

0	-2
-2	

TOP VALUE :

LEFT VALUE :

DIAGONAL VALUE :

NEEDLEMAN WUNSCH ALGORITHM

- MATCH : 1
- MISMATCH : −1
- GAP : −2

		T	A	T	G	A
	0	−2	−4	−6	−8	−10
T	−2	1				
A	−4					
C	−6					
G	−8					
A	−10					

0	−2
−2	1

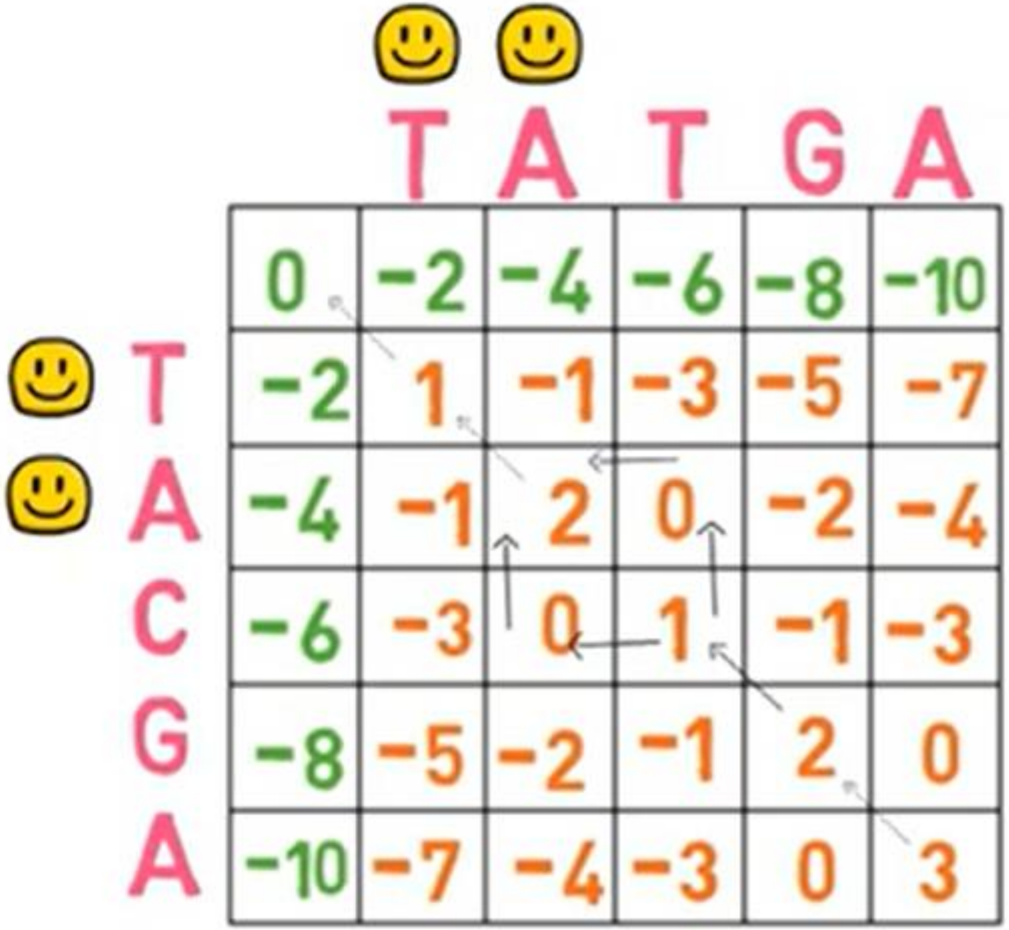
TOP VALUE : (−2) + (−2) = −4

LEFT VALUE : (−2) + (−2) = −4

DIAGONAL VALUE : 0 + 1 = 1

1

NEEDLEMAN WUNSCH ALGORITHM



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

T A T G A
T A C G A

Finally, we have the result of the alignment. Here is the result of the Needleman-Wunsch alignment.

Because it is a global alignment, the full sequence is included and begins and ends on the first and last positions (even if a gap is at the first/last position).

```
CA-CATA
|| | ||
CAGC-TA
```

Scoring Matrix

.	.	C	A	G	C	T	A
.	0	-1	-2	-3	-4	-5	-6
C	-1	1	0	-1	-2	-3	-4
A	-2	0	2	1	0	-1	-2
C	-3	-1	1	1	2	1	0
A	-4	-2	0	0	1	1	2
T	-5	-3	-1	-1	0	2	1
A	-6	-4	-2	-2	-1	1	3

Local Alignment

Query >	**A**	**G**	**T**	**C**	**T**	**A**	**G**			*7nt*
Subject >	**A**	**G**	**A**	**T**	**T**	**C**	**G**	**A**	**G**	*9nt*

How do I perform a local alignment using the same sequences?

FIRST:

Assign a value for MATCHED, MISMATCHED and GAP. They could be random values but common practise is **‘ 1 ’ for MATCHED penalty, ‘ -1 ’ for GAP penalty and ‘ -1 ’ for MISMATCHED penalty**. These values will be used to calculate the probability of highest IDENTITY (how similar they are) between two sequences.

SECOND:

Do the maths! Take a look at the next slide.