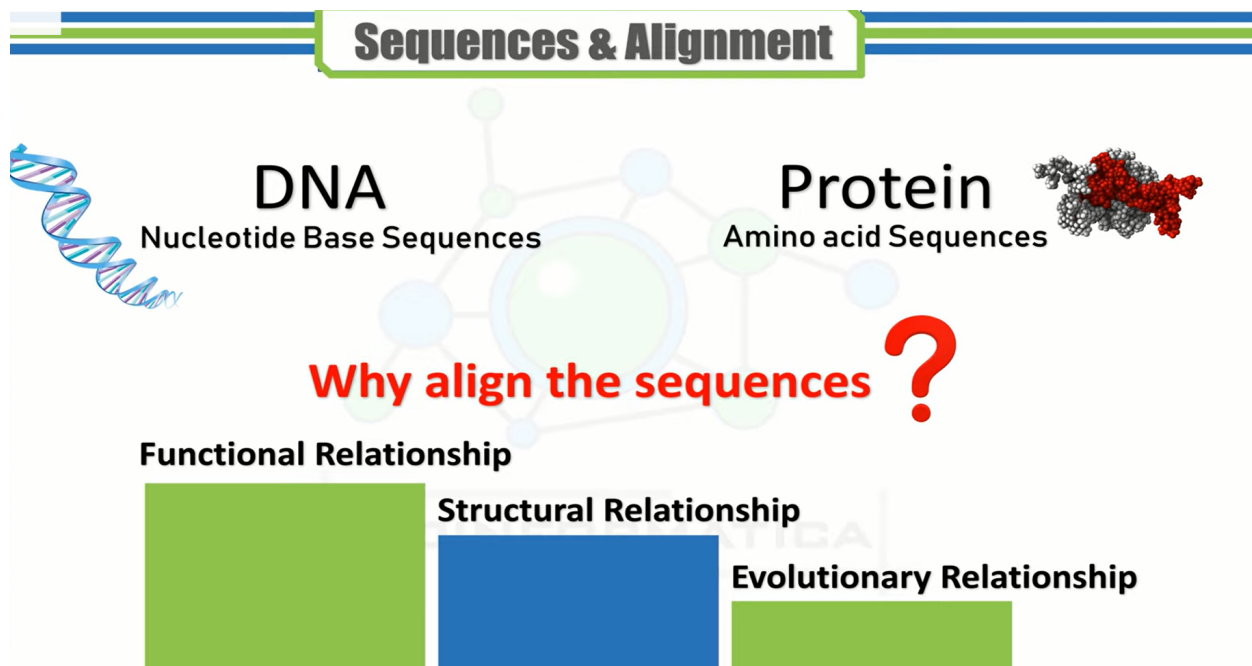## Session 2

Sequence Alignment

● **What is it**

○ **It is the procedure of comparing two *Pairwise Alignment* or more multiple sequences by searching for a series of individual characters or patterns that are in the same order in the sequences.**

○ **It is also a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences.**

Identity: For this column, the nucleotides from two sequences are identical

A T G A C T G G A — Alignment column
A T A A — T G — A

Probable insertion

Alignment cannot tell us molecular origin of a gap

Probable deletion

A T A A T G A

Gaps

Gaps in a DNA sequence alignment represent evolutionary deletions and insertions

Similarity: For this column, nucleotides from two sequences are not identical, but similar, both are purine

---

Identity: <-(*Alignment*)
The two nucleotides are identical from the two sequences.

Similarity: <-(*Mismatch*)
The two nucleotides aren't identical but from the same nitrogenous base division (both are Purines or both are Pyrimidines).

Gap: <-(*Mismatch*)
They tell us that there is an evolutionary deletions and insertion of nucleotides but it can't tell us the molecular origin of a gap
Gaps are marked by (—).

○ **Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix.**

○ **Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns.**

○ **It is used to find out degrees of similarity between two *Pairwise Alignment* or more nucleic acid sequences of DNA or RNA and Amino-acid sequences of proteins.**

○ **Sequence alignments are also used for non-biological sequences, such as calculating the distance cost between strings in a natural language or in financial data.**

○ **There are two types: Global & Local Alignment**

○ **Global Alignment:**
        Is End to End through all of the Sequence.
        Suitable for Divergent Sequences.

○ **Local Alignment:**
        Is Part of the Sequence that has the highest Similarity.
        Suitable for Closely related Sequences.

Alignment

Global Alignment

Target Sequence
```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
    ||||||||||      |||||||   ||||||||||||| ||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```
Query Sequence

Tries to align entire sequence

Align all letters from query and target

Suitable for closely related sequences

A general Global Alignment Method is
NEEDLEMAN - WUNSCH

GLOBAL VS. LOCAL SEQUENCE ALIGNMENT

Local Alignment

Target Sequence
```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
                  ||||  |||||| ||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Align regions having highest similarities

Align substring of target with substring of query

Suitable for more divergent sequences

A general Local Alignment Method is
SMITH - WATERMAN

o

Local Alignment

Target Sequence
```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
                  ||||  |||||| ||||||||||||||||
Query Sequence 5' TACTCACGGATGAGGTACTTTAGAGGC 3'
```

Global Alignment

Target Sequence
```
5' ACTACTAGATTACTTACGGATCAGGTACTTTAGAGGCTTGCAACCA 3'
    ||||||||||      |||||||   ||||||||||||| ||||||
5' ACTACTAGATT----ACGGATC--GTACTTTAGAGGCTAGCAACCA 3'
```
Query Sequence

o

3

ATCG
*TCG

→ Match   +1
→ Mis match −1
→ Gap   −2

[→/↑] + Gap

[↗] + Match/Mismatch

		Gap	T	C	G
G	−8	−10 / −5	−7 / −2	−4 / +1	+1
		−7 (−5)	−4 (−2)	(+1) −4	
C	−6	−8 / −3	−5 / 0	0 → (−2)	−2
		−5 (−3)	(0) −4	−3 −6	
T	−4	−6 / −1	−3 / −2	−2 → (−4)	−4
		(−1) −3	(−2) −5	(−4) −7	
A	−2	−4 / −1	−1 → (0)	−3 → (−5)	−5
		(−1) −4	(−3) −6	(−5) −8	
Gap	Zero	−2	−4	−6	
	Gap	T	C	G	

★ Get the minimum every time
★ When you get the maximum → Trace Back

Rewards and Penalties

Match	:	1
Mismatch	:	-1
GAP	:	-2

Sequences

Seq 1 = $\dfrac{\text{A T G C T}}{M}$

Seq 2 = $\dfrac{\text{A G C T}}{N}$

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

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

Seq-1   **A T G C T**

Seq-2   **A - G C T**

5

Differences between Global & Local Alignment

BASIS OF COMPARISON	GLOBAL SEQUENCE ALIGNMENT	LOCAL SEQUENCE ALIGNMENT
Description	In global alignment, an attempt is made to align the entire sequence (end to end sequence alignment).	Finds local regions with the highest level of similarity between the two sequences.
Examples Of Tools	-EMBOSS Needle -Needleman-Wunsch Global Align Nucleotide Sequences (Specialized BLAST)	-BLAST -EMBOSS Water -LALIGN
Function	A global alignment contains all letters from both the query and target sequences.	A local alignment aligns a substring of the query sequence to a substring of the target sequences.
Two Sequences	If two sequences have approximately the same length and are quite similar, they are suitable for global alignment.	Any two sequences can be locally aligned as local alignment finds stretches of sequences with high levels of matches without considering the alignment of the rest of the sequence regions.
Suitability	Suitable for aligning two closely related sequences.	Suitable for aligning more divergent sequences or distantly related sequences.
Use	Global alignments are usually done for comparing homologous genes like comparing two genes with same function (in human vs. mouse) or comparing two proteins with similar function.	Used for finding out conserved patterns in DNA sequences or conserved domains or motifs in two proteins.
General Technique	Needleman–Wunsch algorithm.	Smith–Waterman algorithm.

○ **Global Alignment**

- **Global alignment is used to compare sequences in cases where we have reason to believe that the sequences are related along their entire length. If for example, sequences s and t are two independent sequencing runs of the same PCR product, then they should differ only in positions where there are sequencing errors. In order to find those sequencing errors, we align all of sequence s with all of sequence t. Other applications of global alignment include finding mutations in closely related gene or protein sequences and identification of single nucleotide polymorphisms (SNPs)**

```
  1   2   3   4   5   6   7   8   9  10  11

  C   G   T   C   C   G   A   A   G   T   G
          |   .   |   |   |   |
  ⋆   ⋆   T   A   C   G   A   A   ⋆   ⋆   ⋆
```

(a) Global alignment

- **Examples of Global alignment tools include:**

  - **EMBOSS Needle**

  - **Needleman-Wunsch Global Align Nucleotide Sequences (Specialized BLAST)**

- **In global alignment, an attempt is made to align the entire sequence (end to end alignment).**

- **Contains all letters from both the query and target sequences.**

- **Suitable If two sequences have approximately the same length and are quite similar.**

- **Suitable for aligning two closely related sequences.**

- **Global alignments are usually done for comparing homologous genes like comparing two genes with same function (in human vs. mouse) or comparing two proteins with similar function.**

- **A general Global alignment technique:**

  is Needleman–Wunsch algorithm.

○ Local Alignment

- Local alignment addresses cases where we only expect to find isolated regions of similarity. One example is alignment of genomic DNA upstream from two co-expressed genes to find conserved regions that may correspond to transcription factor binding sites. Another application is identification of conserved domains1 in two amino acid sequences that encode proteins that share one or more domains, but are unrelated.

```
3   4   5   6   7   8
T   C   C   G   A   A
|   .   |   |   |   |
T   A   C   G   A   A
```

(b) Local alignment

- Examples of Local alignment tools include:

  - BLAST

  - EMBOSS Water

  - LALIGN

- Finds local regions with the highest level of similarity between the two sequences.
- A local alignment aligns a substring of the query sequence to a substring of the target sequence.
- Any two sequences can be locally aligned as local alignment finds stretches of sequences with high levels of matches without considering the alignment of the rest of the sequence regions.
- Suitable for aligning more divergent or distantly related Seq.
- Used for finding out conserved patterns in DNA sequences or conserved domains or motifs in two proteins.

- A general Local alignment method:

    is Smith–Waterman algorithm.