Sequence Alignment:-

Is a way of arranging of two or more sequences of DNA, RNA or protein to identify and highlight regions of similarity. the similarity may indicate the functional, structural and evolutionary signifane of the sequence, it's made between a known sequence and unknown sequence or between two unknown sequences

i- the known sequence is called reference sequence.
ii- the unknown sequence is called Query sequence.

TYPES of ALIGNMENT

According to the number of sequence.
i- pairwise alignment (Global or local sequence alignment)
ii- multiple alignment.

PAIR-wise Sequence Alignment.

- it only compaires two sequence at a time
- it is concerned with finding the best-matching of prot amino acid or DNA (Nucleic acid) sequences.
- typically, the purpose of this is to find homologues (creative) of a gene in a database of knowns.
- this information is useful for answering a variety of biological questions.
  i- the study of molecular evolution.
- It consist of a series of paired bases, one base from each sequence

TYPES OF PAIRS:

i- **Matches**:- the same nucleotide appears in both sequences.

ii- **Mismatches**:- different nucleotides are found in the two sequence, it's also alignment correspond to **mutation**.

iii- **Gaps**:- A base in one sequence and a null base in the other. Gaps correspond to **Insertions** or **deletions**.

Difference between pairwise and multiple sequence Alignment

Pair-wise sequence Alignment:- Algorithm used are:-

- Needlemen-wunsch algorithm & smith-waterman algorithm
- Blast (Basic local alignment search tool).
- it determine similar regions between two nucleotide or protein sequences; optimality is based on score.
- The BLAST algorithm is much more efficient, much faster ease of use, statistical rigor than other approaches.

Multiple sequence Alignment: (MSA): is a sequence alignment of three or more biological sequences, generally protein, DNA or RNA. It try to align all of the sequences in a specified set. the most popular multiple alignment tools is; **clustalW, MUSCLE, PROBCONS, MAFFT**, and **T-Coffee**.

TYPE OF Sequence Alig

According to sequence coverage:

i- Global sequence alignment

ii- Local sequence alignment.

Global Sequence Alignment: is a matching the residues of two sequences accross their entire length. that is alignment is carried out from begining to end of both sequences to find the best possible alignment accross the entire length between the two sequences.

Local Sequence Alignment: is a matching of two sequence from regions which have more similarity with each other, It does not assume that the two sequence in question have similarity over the entire length. It only find local regions with the highest level of similarity between the two sequences and aligns these regions without regard for the alignment of the rest of the sequence regions.

DATABASE SIMILARITY SEARCHING:

A main application of pair wise alignment is retrieving biological sequences in databases based on similarity; this process involves submission of a query sequences and performing a pairwise comparison of the query with all individual sequence in a database.

(Mycobacterium tuberculosis DNA, complete genome strain H37Rv)

There are two major algorithms for performing database searches: BLAST and FASTA.

BLAST (Basic Local Alignment Search Tool) is the most common local tool developed by Altschul et al 1990, it is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins or the nucleotides of DNA sequences, A BLAST search enables a researcher to compare a query sequence with a library or database of sequence, and identify library sequences that resemble the query sequence above a certain threshold.

Example the question that BLAST ask:

"Which bacterial species have a protein that is related in lineage to a certain protein with known amino-acid sequence"?

The BLAST algorithm and the computer programs that implements it were developed by Stephen Altschul, Warren Gish and David Lipman at the US national center for biotechnology Information (NCBI)

FASAT FORMAT

is the standard format in the field of bioinformatics to represent either nucleotide sequences or peptide sequences. this format is single-letter code and it allows sequence names and description at the begining followed by sequence data in multiple lines. the length of the each chunk (line) of the sequence must not exceeds 80 characters, sequence identifiers are defined by a standard called NCBI