DATABASES & MULTIPLE SEQUENCE ALIGNMENT

PREPARED BY AYESHA

There are four structural types of database management systems:

- **Hierarchical databases.**
- **Network databases.**
- **Relational databases.**
- **Object-oriented databases.**

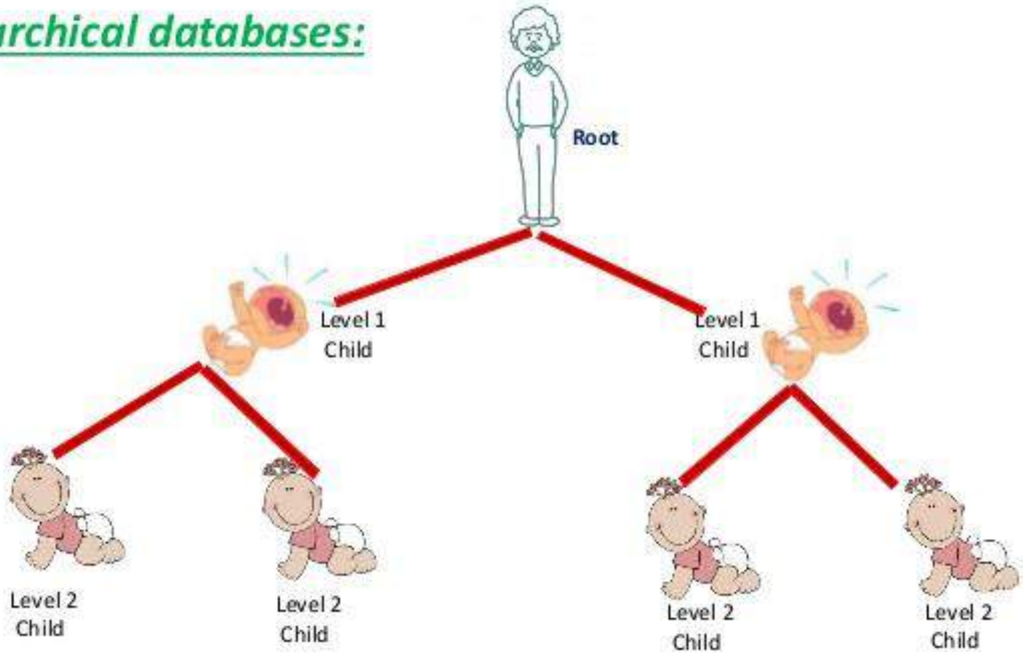
Slide:5

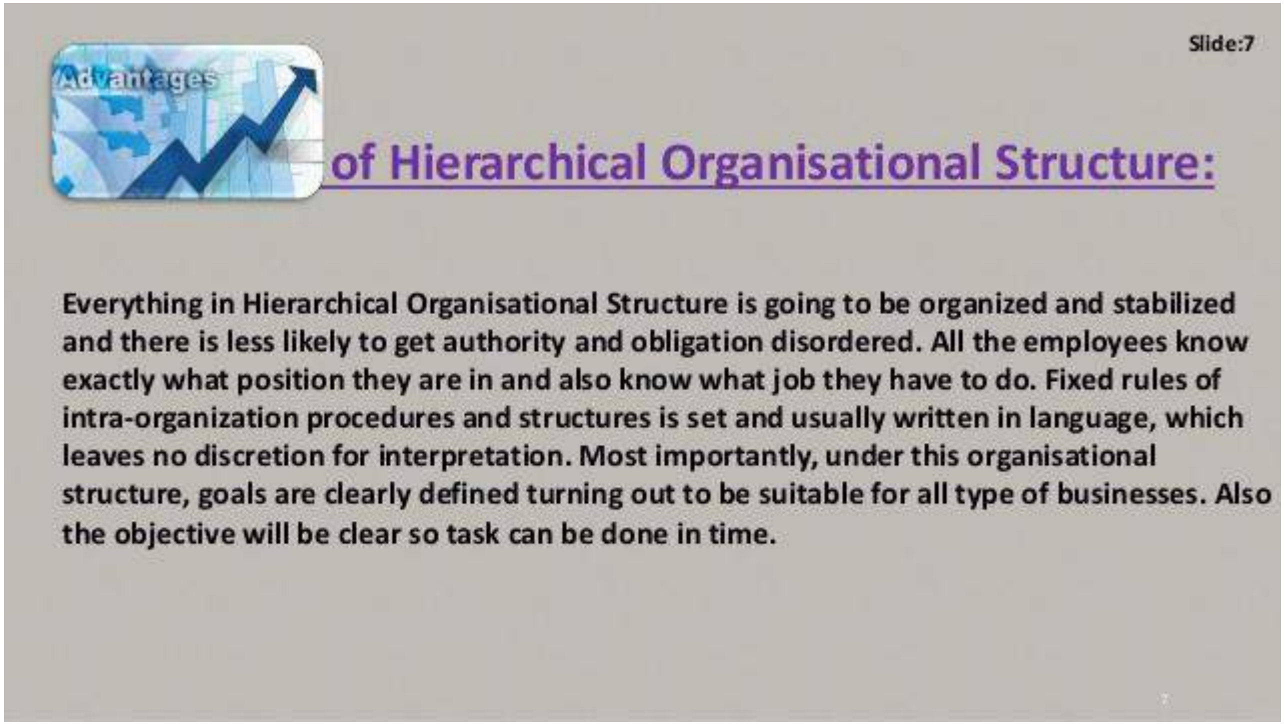
Hierarchical databases

A hierarchical database model is a data model in which the data is organized into a tree-like structure. The data is stored as records which are connected to one another through links. A record is a collection of fields, with each field containing only one value.

Slide:6

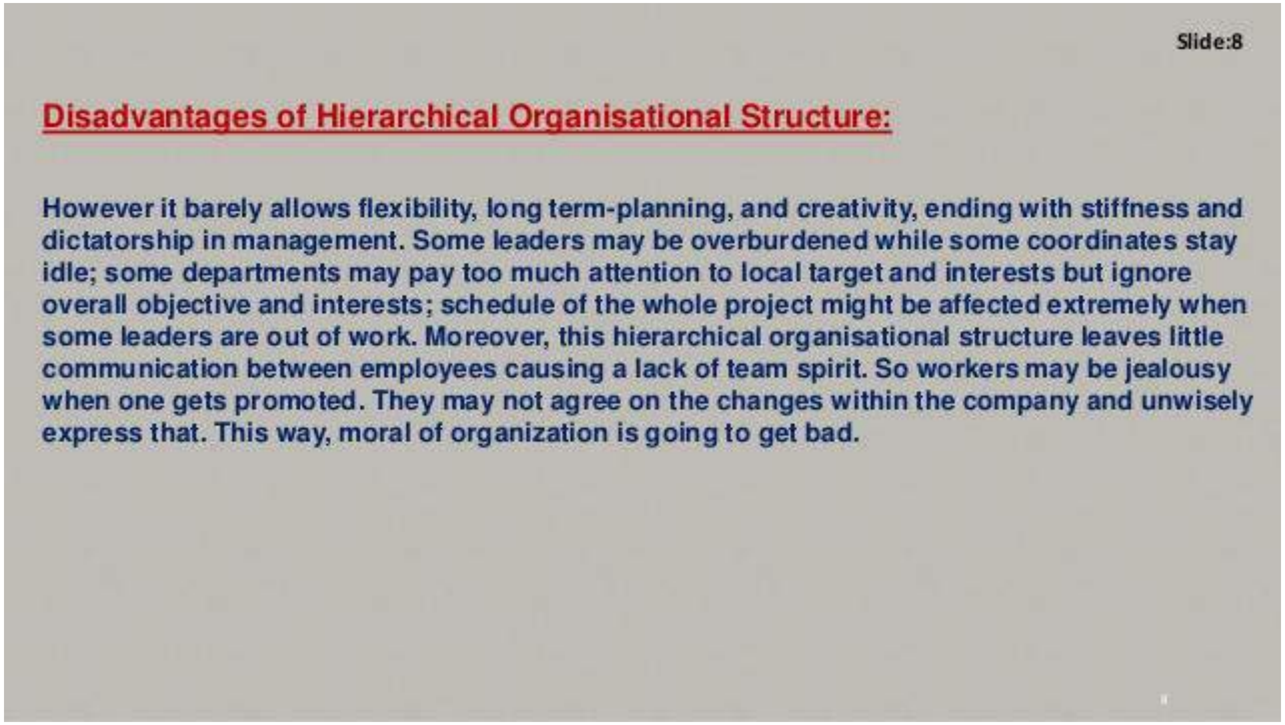
Hierarchical databases:





Advantages of Hierarchical Organisational Structure:

Everything in Hierarchical Organisational Structure is going to be organized and stabilized and there is less likely to get authority and obligation disordered. All the employees know exactly what position they are in and also know what job they have to do. Fixed rules of intra-organization procedures and structures is set and usually written in language, which leaves no discretion for interpretation. Most importantly, under this organisational structure, goals are clearly defined turning out to be suitable for all type of businesses. Also the objective will be clear so task can be done in time.



Disadvantages of Hierarchical Organisational Structure:

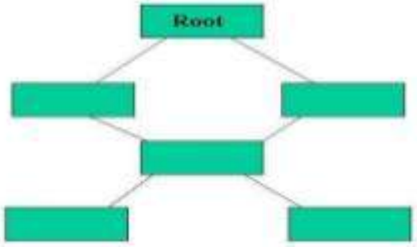
However it barely allows flexibility, long term-planning, and creativity, ending with stiffness and dictatorship in management. Some leaders may be overburdened while some coordinates stay idle; some departments may pay too much attention to local target and interests but ignore overall objective and interests; schedule of the whole project might be affected extremely when some leaders are out of work. Moreover, this hierarchical organisational structure leaves little communication between employees causing a lack of team spirit. So workers may be jealousy when one gets promoted. They may not agree on the changes within the company and unwisely express that. This way, moral of organization is going to get bad.

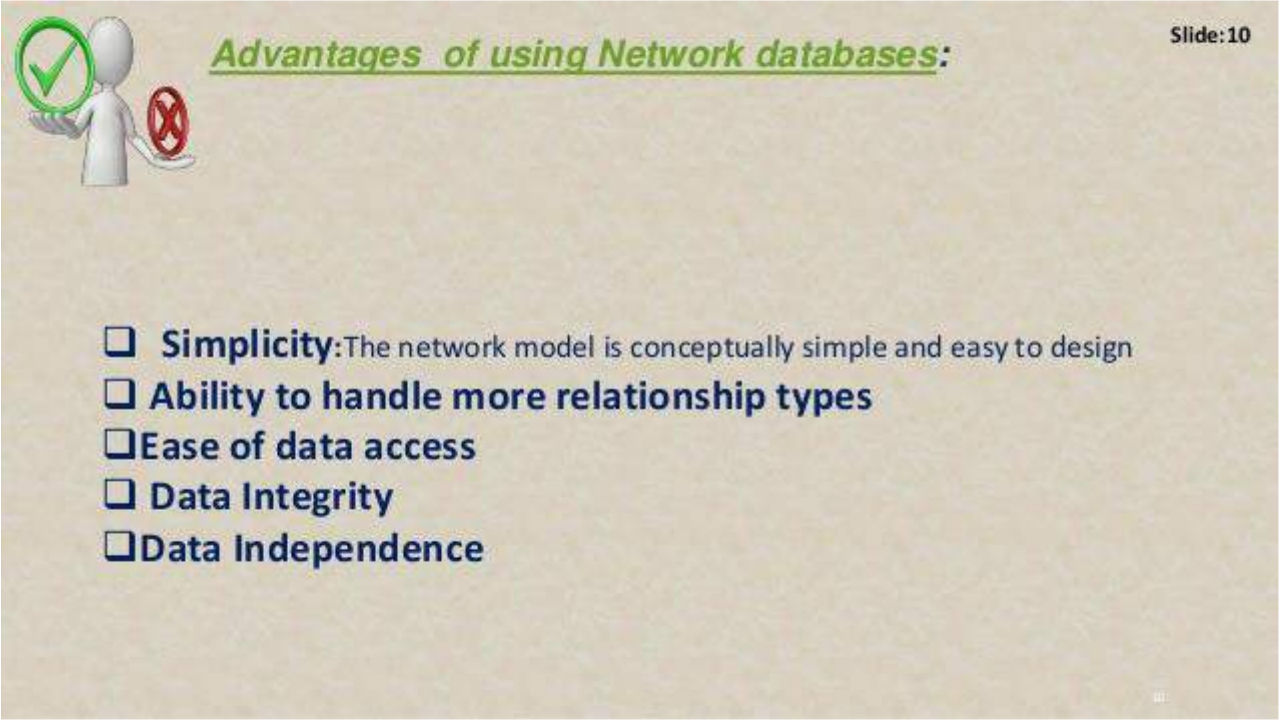
Network databases

Slide:9

The **network** model is a **database** model conceived as a flexible way of representing objects and their relationships. Its distinguishing feature is that the schema, viewed as a graph in which object types are nodes and relationship types are arcs, is not restricted to being a hierarchy or lattice.

Network Database Model





Advantages of using Network databases:

- ❑ **Simplicity**:The network model is conceptually simple and easy to design
- ❑ **Ability to handle more relationship types**
- ❑**Ease of data access**
- ❑ **Data Integrity**
- ❑**Data Independence**

Disdvantages of using Network databases

System complexity:

In a network model, data are accessed one record at a time. This males it essential for thedatabase designers, administrators, and programmers to be familiar with the internal datastructures to gain access to the data. Therefore, a user friendly database management systemcannot be created using the network model

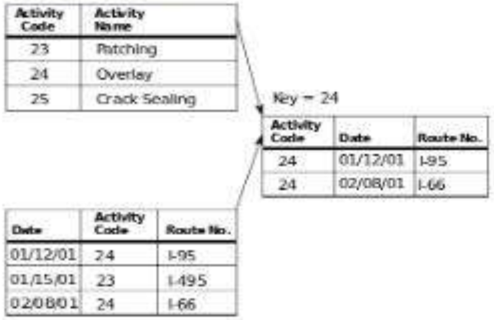
Lack of Structural independence:

Making structural modifications to the database is very difficult in the network database modelas the data access method is navigational. Any changes made to the database structure require theapplication programs to be modified before they can access data. Though the network modelachieves data independence, it still fails to achieve structural independence

Relational databases

A relational database is a digital database whose organization is based on the relational model of data, as proposed by E. F. Codd in 1970. The various software systems used to maintain relational databasesare known as a relational database management system (RDBMS).

Relational Model



Activity Code	Activity Name
23	Patching
24	Overlay
25	Crack Sealing

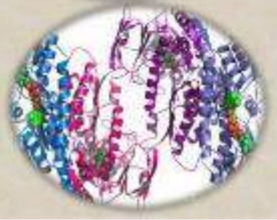
Key = 24

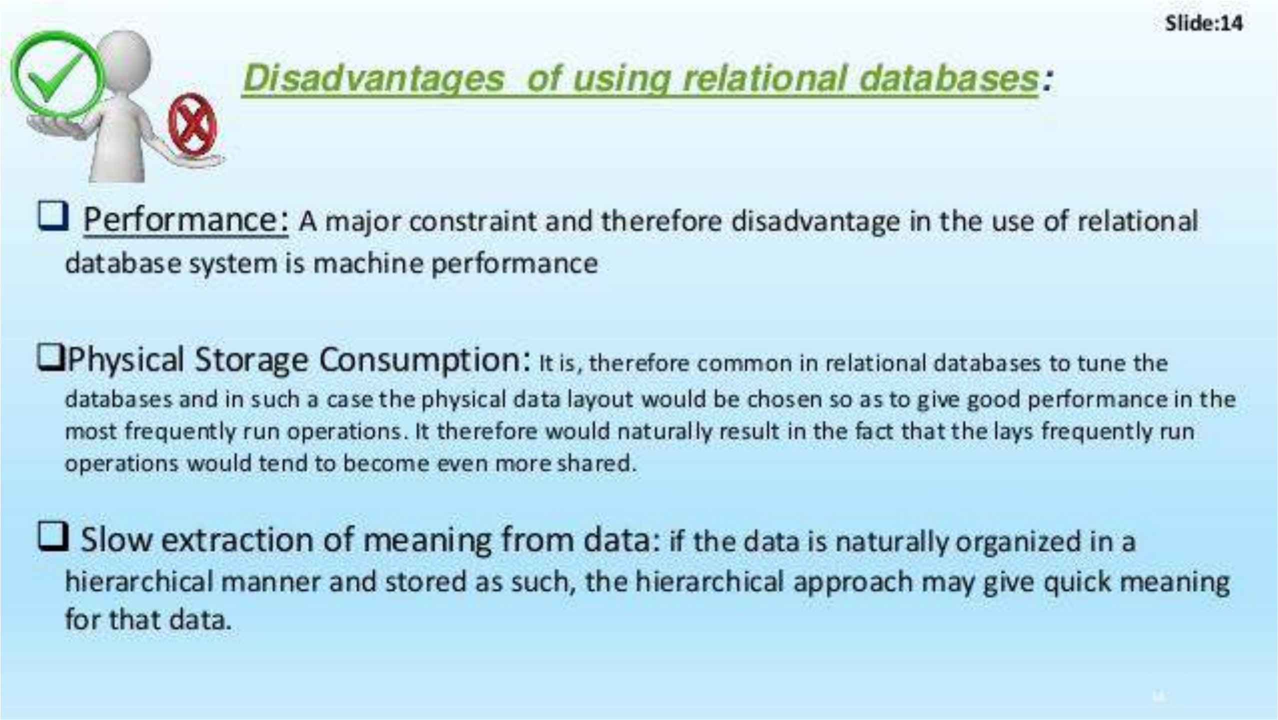
Activity Code	Date	Route No.
24	01/12/01	I-95
24	02/08/01	I-66

Date	Activity Code	Route No.
01/12/01	24	I-95
01/15/01	23	I-495
02/08/01	24	I-66

Advantages of using relational databases:

- ❑ Ease of use
- ❑ Flexibility
- ❑ Precision
- ❑ Security
- ❑ Data Independence
- ❑ Data Manipulation Language





Disadvantages of using relational databases:

□ Performance: A major constraint and therefore disadvantage in the use of relational database system is machine performance

□Physical Storage Consumption: It is, therefore common in relational databases to tune the databases and in such a case the physical data layout would be chosen so as to give good performance in the most frequently run operations. It therefore would naturally result in the fact that the lays frequently run operations would tend to become even more shared.

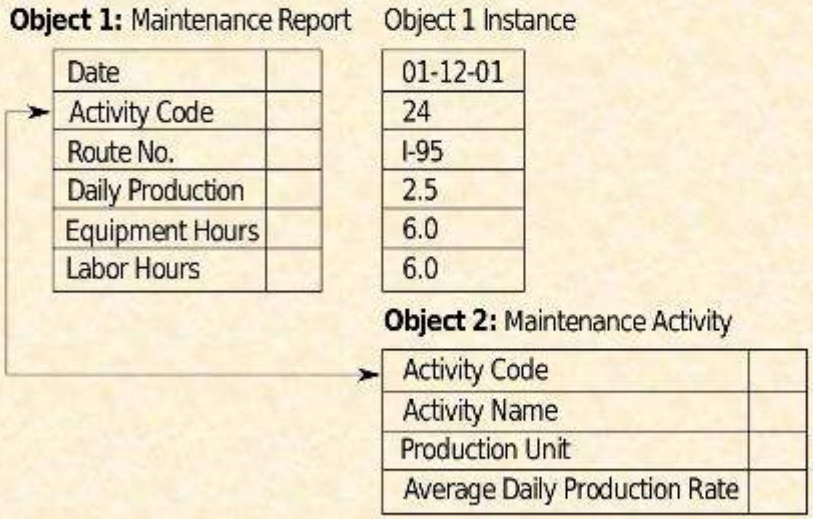
□ Slow extraction of meaning from data: if the data is naturally organized in a hierarchical manner and stored as such, the hierarchical approach may give quick meaning for that data.

Slide:15

Object-oriented databases

An object database (also object-oriented databasemanagement system, OODBMS) is a databasemanagement system in which information is represented in the form of objects as used in object-oriented programming. Object databases are different from relational databases which are table-oriented.

Object-Oriented Model



Object 1: Maintenance Report

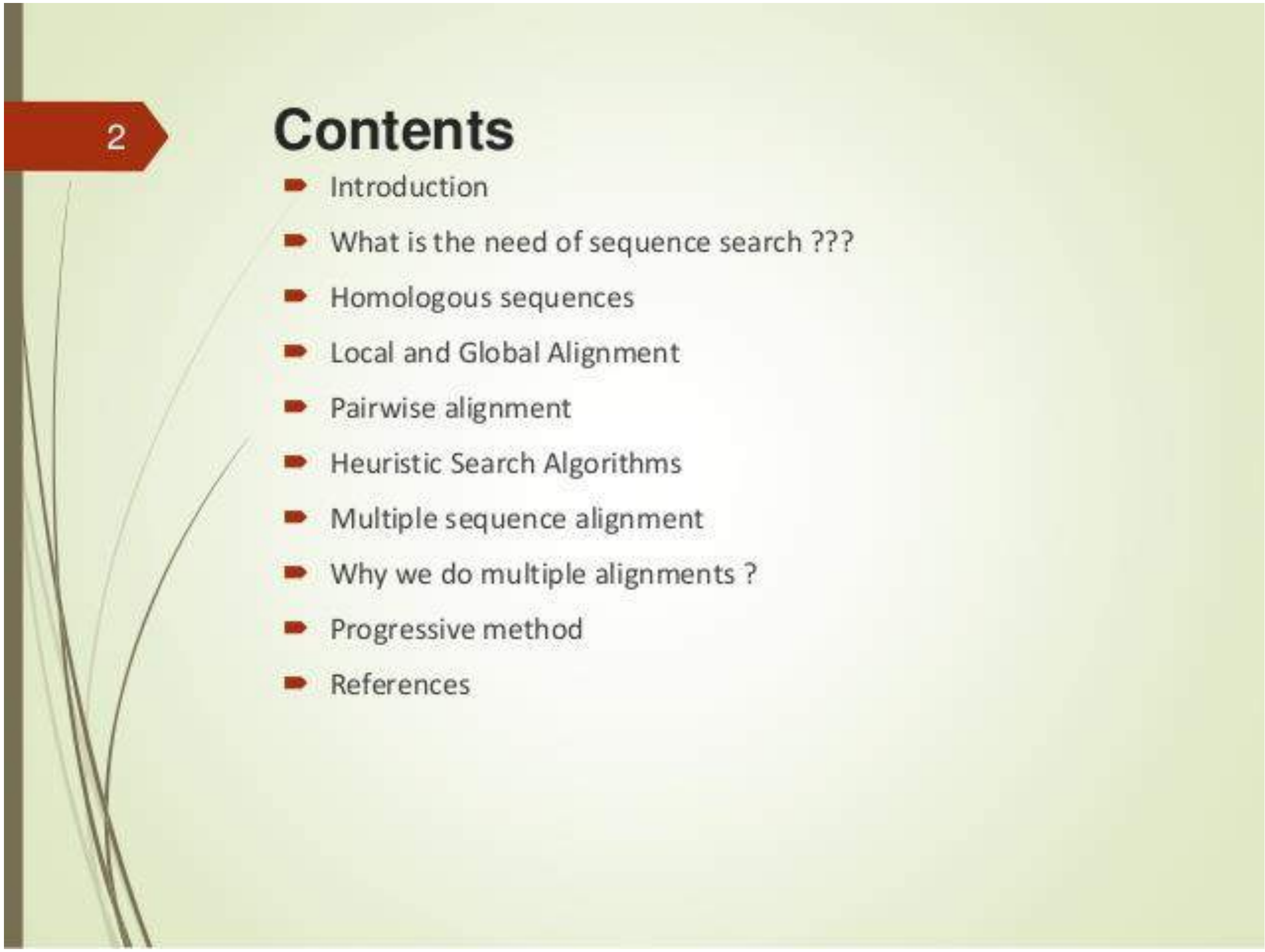
Date	
Activity Code	
Route No.	
Daily Production	
Equipment Hours	
Labor Hours	

Object 1 Instance

01-12-01
24
I-95
2.5
6.0
6.0

Object 2: Maintenance Activity

Activity Code	
Activity Name	
Production Unit	
Average Daily Production Rate	

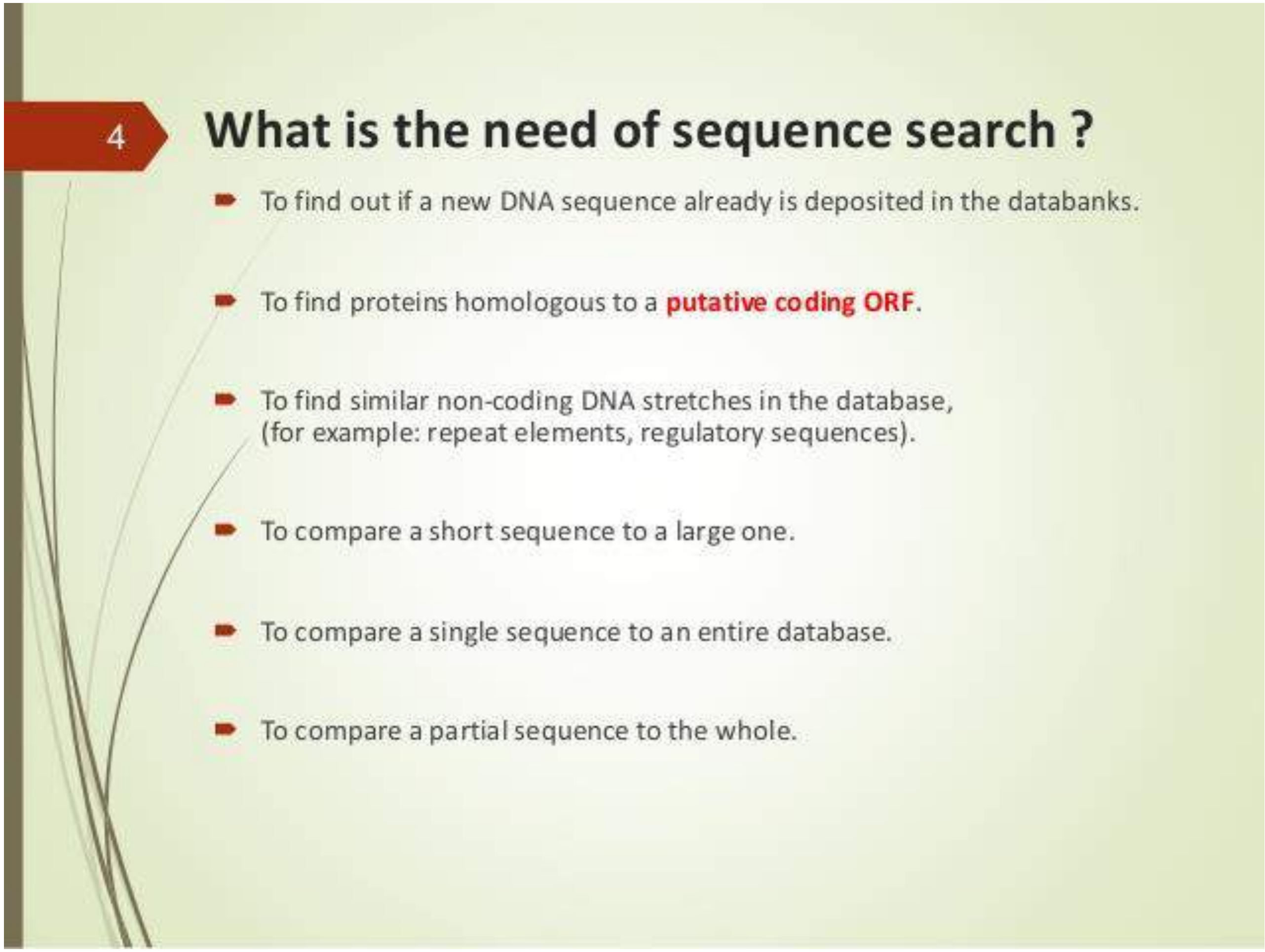


Contents

- Introduction
- What is the need of sequence search ???
- Homologous sequences
- Local and Global Alignment
- Pairwise alignment
- Heuristic Search Algorithms
- Multiple sequence alignment
- Why we do multiple alignments ?
- Progressive method
- References

Introduction

- Sequence similarity searching to identify homologous sequences is one of the first, and most informative steps in any analysis for newly determined sequences.
- Modern protein sequence databases are very comprehensive, so that more than 80% of metagenomic sequence samples typically share significant similarity with proteins in sequence databases.
- Widely used similarity searching programs, like BLAST , PSI-BLAST , SSEARCH and the HMMER3 programs produce accurate statistical estimates, ensuring protein sequences that share significant similarity also have similar structures.



What is the need of sequence search ?

- To find out if a new DNA sequence already is deposited in the databanks.

- To find proteins homologous to a **putative coding ORF**.

- To find similar non-coding DNA stretches in the database, (for example: repeat elements, regulatory sequences).

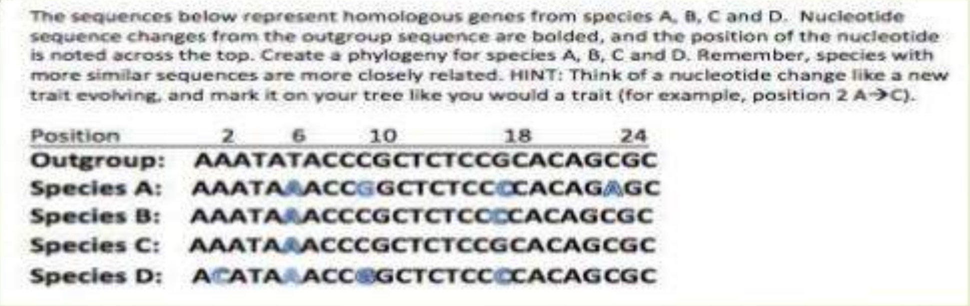
- To compare a short sequence to a large one.

- To compare a single sequence to an entire database.

- To compare a partial sequence to the whole.

Homologous sequences

- A homologous sequence, in molecular biology, means that the sequence is similar to another sequence. The similarity is derived from **common ancestry**.
- Homology among DNA, RNA, or proteins is typically inferred from their nucleotide or amino acid **sequence similarity**.



The sequences below represent homologous genes from species A, B, C and D. Nucleotide sequence changes from the outgroup sequence are bolded, and the position of the nucleotide is noted across the top. Create a phylogeny for species A, B, C and D. Remember, species with more similar sequences are more closely related. HINT: Think of a nucleotide change like a new trait evolving, and mark it on your tree like you would a trait (for example, position 2 A→C).

Position	2 6 10 18 24
Outgroup:	**AAATATACCCGCTCTCCGCACAGCGC**
Species A:	**AAATAAACCGGCTCTCCCCACAGAGC**
Species B:	**AAATAAACCCGCTCTCCCCACAGCGC**
Species C:	**AAATAAACCCGCTCTCCGCACAGCGC**
Species D:	**ACATAAACCGGCTCTCCCCACAGCGC**

Fig.1 Homologous sequences
Source : https://en.wikipedia.org/wiki/Sequence_homology

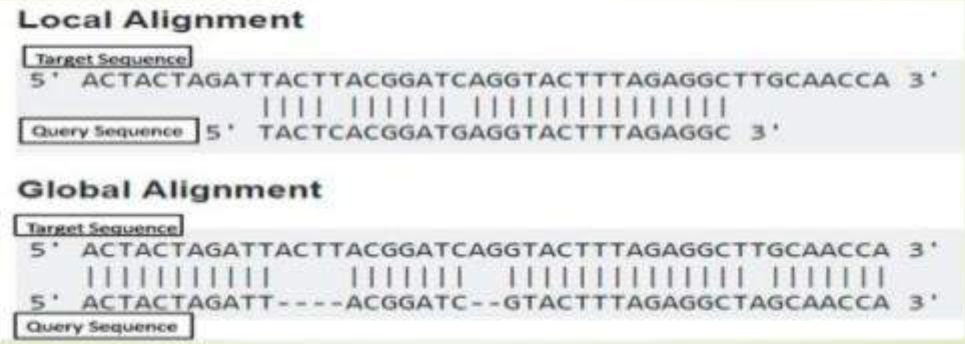
Local and Global Alignment

Local Alignment :

- Stretches of sequences with **highest density** of matches are aligned.
- Suitable for **partially similar different length** and **conserved region** containing sequences.

Global Alignment :

- Attempts to align the **maximum of the entire sequence**.
- Suitable for similar and **equal length sequences**.



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
```

Fig. 2 local and global alignment
Source : https://www.majordifferences.com/2016/05/difference-between-global-and-local.html#.XAc6W-LhXIU

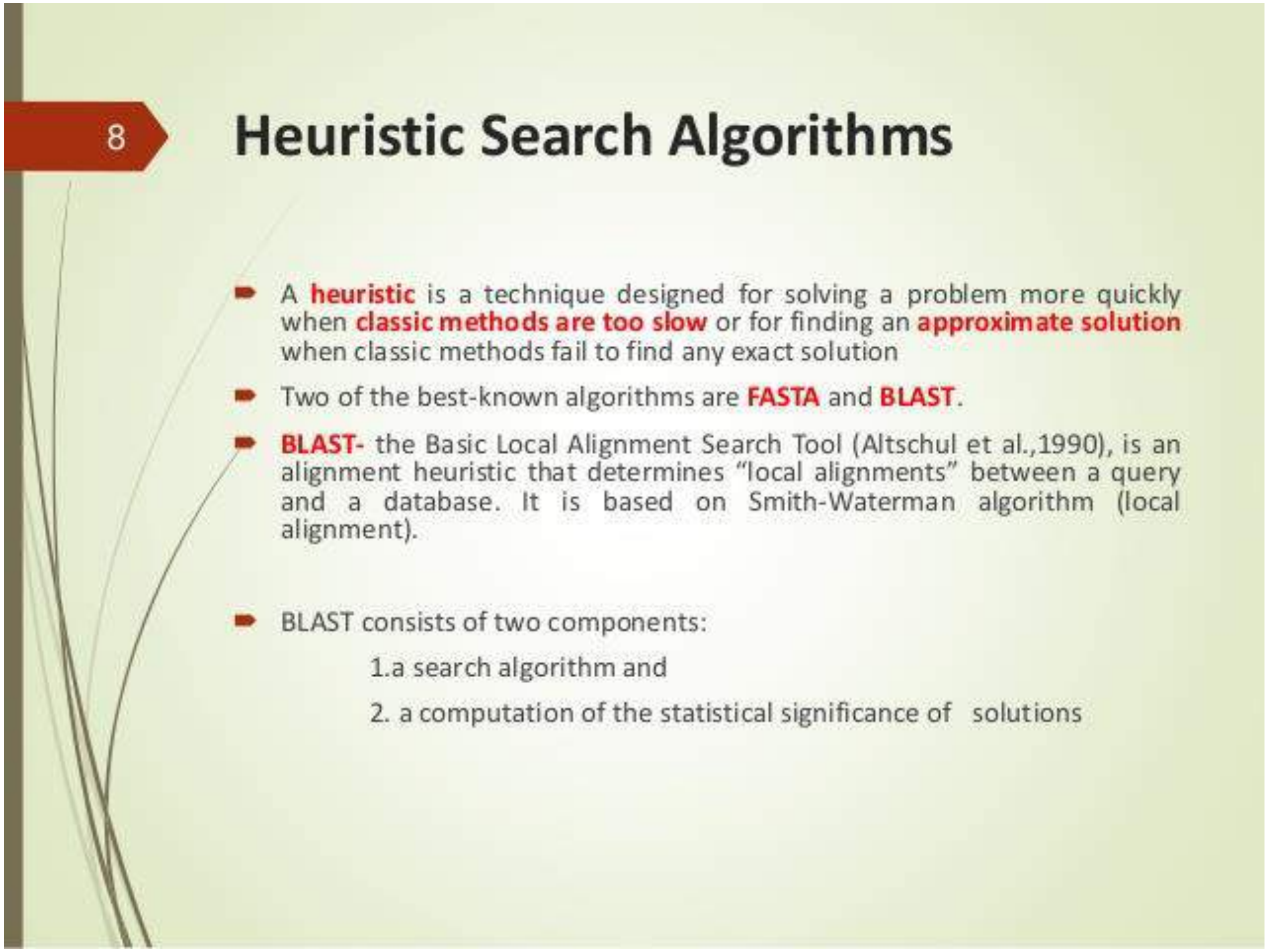
Pair wise Alignment

- **Pair wise Sequence Alignment** is used to identify regions of similarity that may indicate **functional, structural and/or evolutionary** relationships between two biological sequences (protein or nucleic acid).

- It is used to decide if two proteins (or genes) are related structurally or functionally.

- **Needleman Wunsch Algorithm → Global Alignment**

- **Smith waterman Algorithm → Local Alignment**

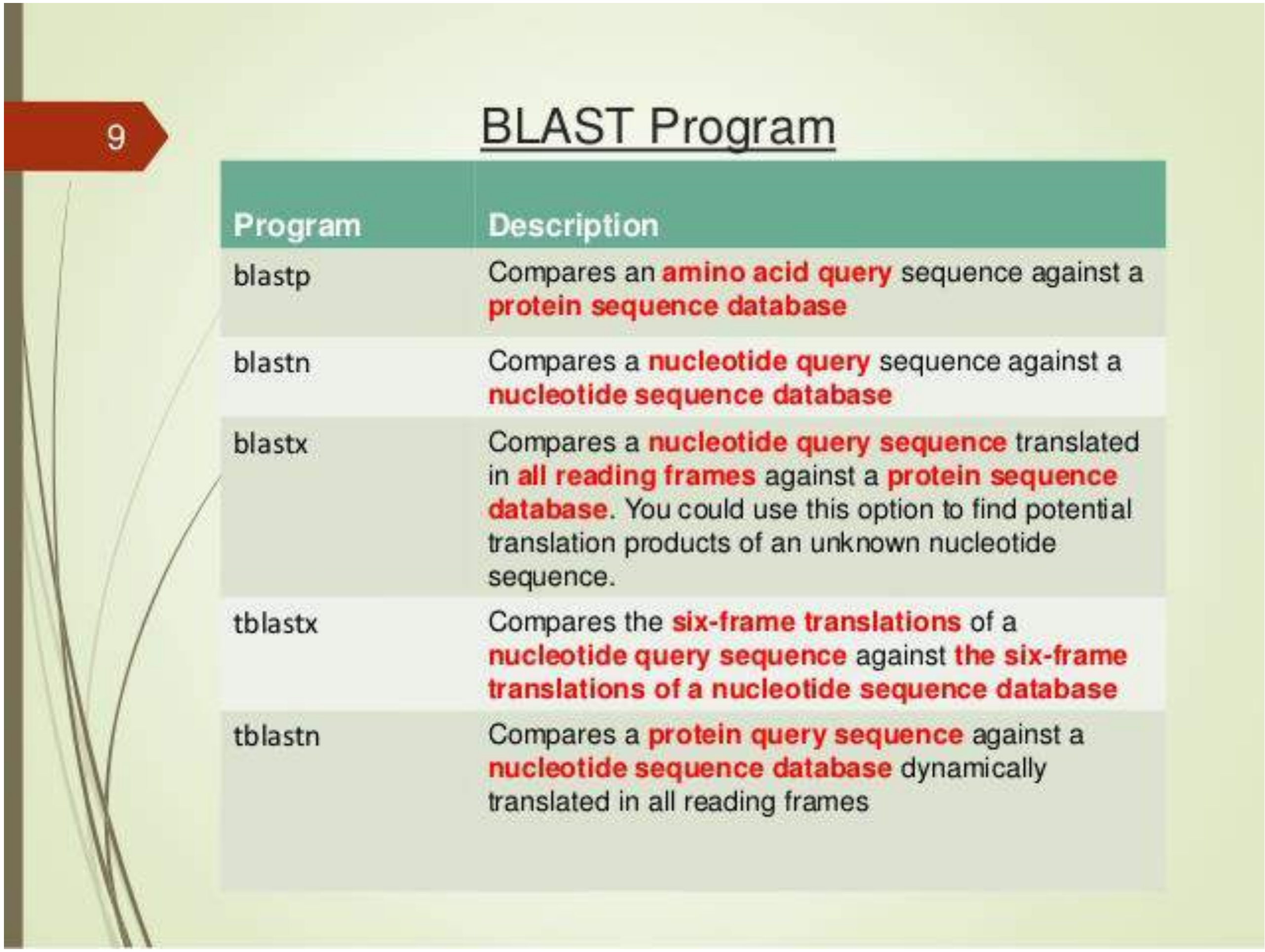


Heuristic Search Algorithms

- A **heuristic** is a technique designed for solving a problem more quickly when **classic methods are too slow** or for finding an **approximate solution** when classic methods fail to find any exact solution
- Two of the best-known algorithms are **FASTA** and **BLAST**.
- **BLAST-** the Basic Local Alignment Search Tool (Altschul et al.,1990), is an alignment heuristic that determines "local alignments" between a query and a database. It is based on Smith-Waterman algorithm (local alignment).
- BLAST consists of two components:

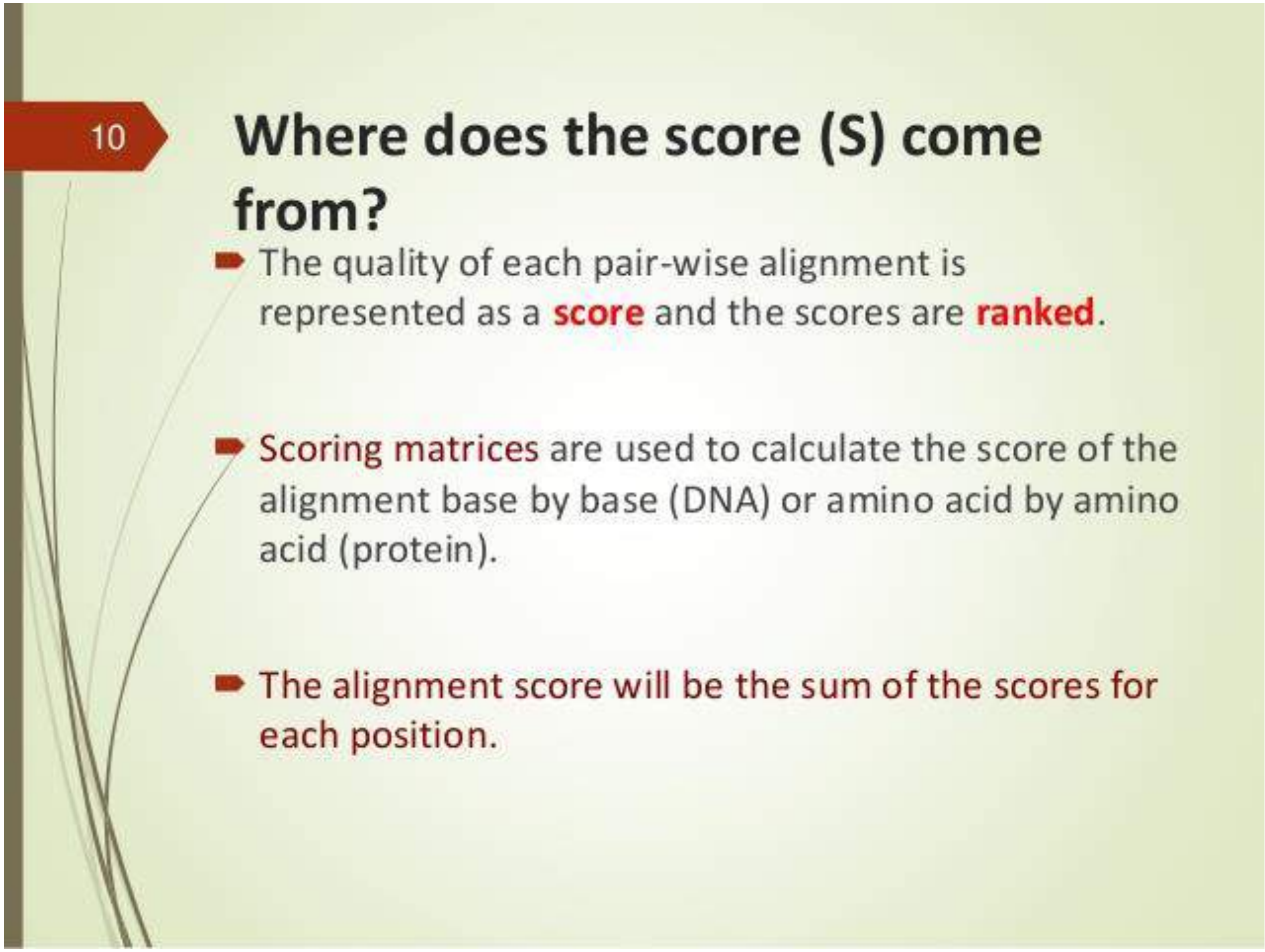
  1.a search algorithm and

  2. a computation of the statistical significance of solutions



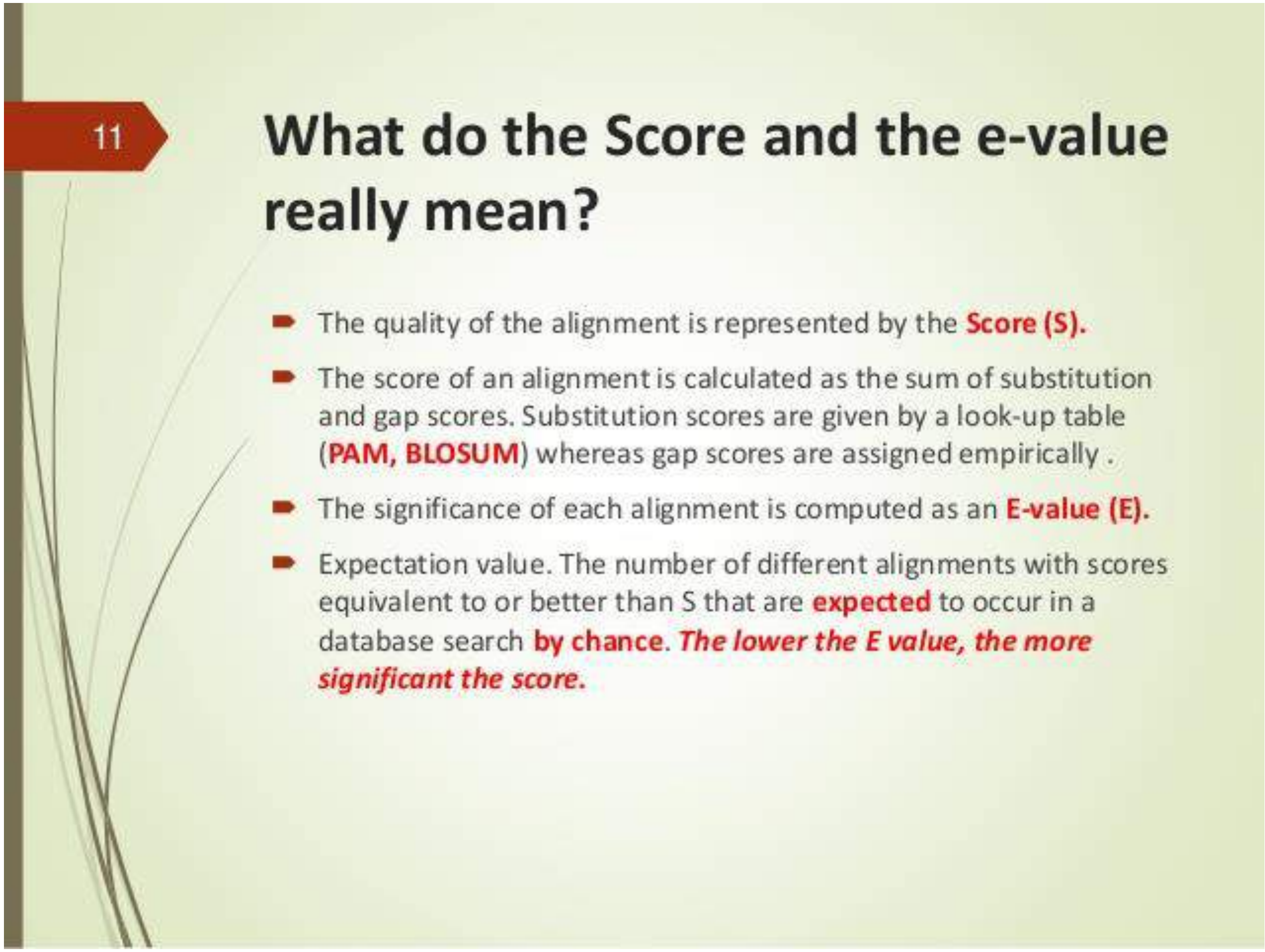
BLAST Program

Program	Description
blastp	Compares an **amino acid query** sequence against a **protein sequence database**
blastn	Compares a **nucleotide query** sequence against a **nucleotide sequence database**
blastx	Compares a **nucleotide query sequence** translated in **all reading frames** against a **protein sequence database**. You could use this option to find potential translation products of an unknown nucleotide sequence.
tblastx	Compares the **six-frame translations** of a **nucleotide query sequence** against **the six-frame translations of a nucleotide sequence database**
tblastn	Compares a **protein query sequence** against a **nucleotide sequence database** dynamically translated in all reading frames



Where does the score (S) come from?

- The quality of each pair-wise alignment is represented as a **score** and the scores are **ranked**.
- Scoring matrices are used to calculate the score of the alignment base by base (DNA) or amino acid by amino acid (protein).
- The alignment score will be the sum of the scores for each position.



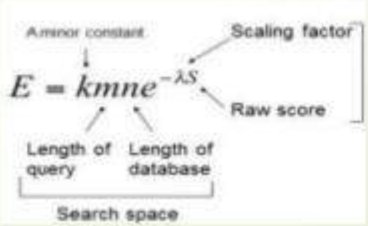
What do the Score and the e-value really mean?

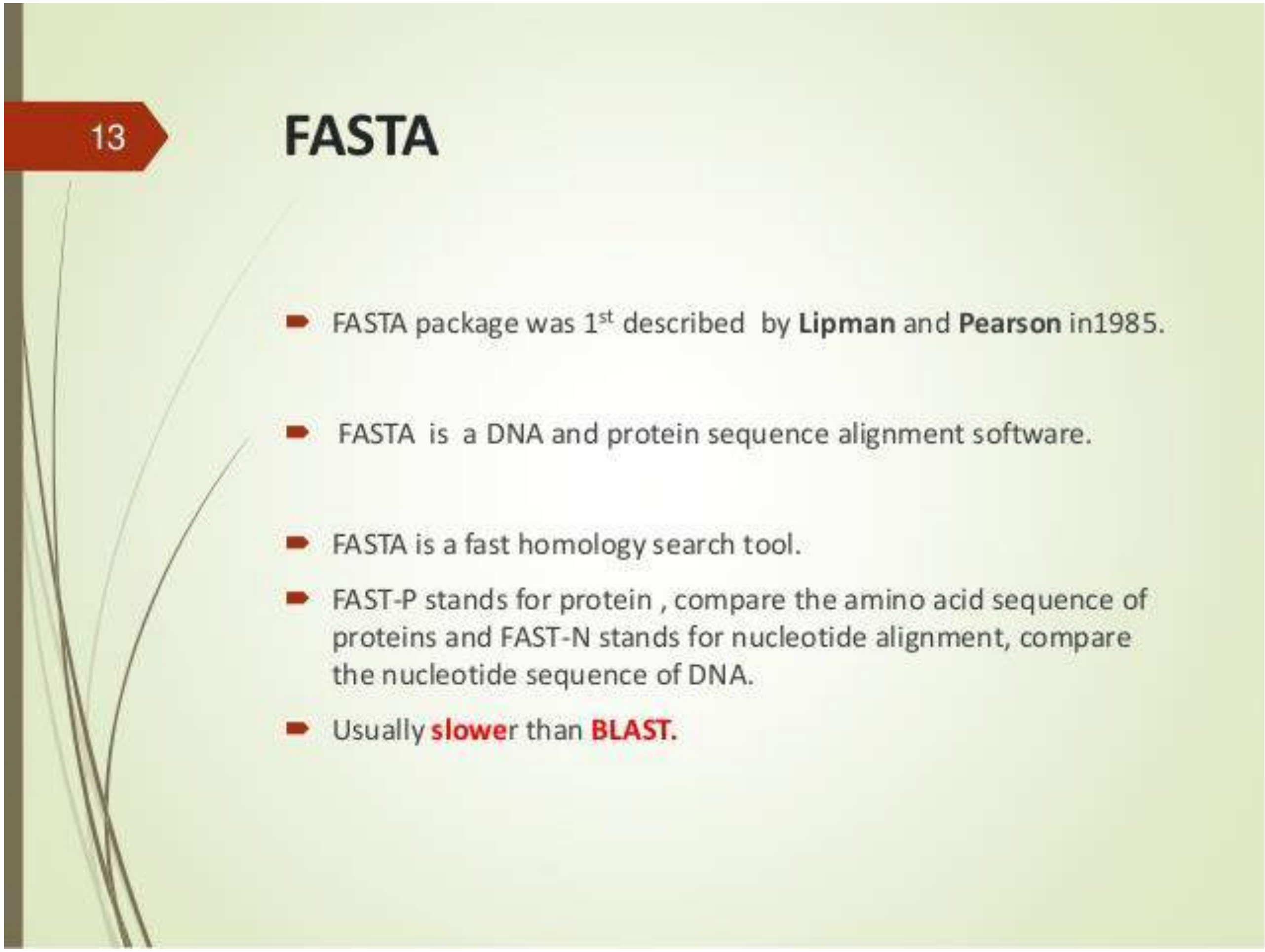
- The quality of the alignment is represented by the **Score (S).**
- The score of an alignment is calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (**PAM, BLOSUM**) whereas gap scores are assigned empirically .
- The significance of each alignment is computed as an **E-value (E).**
- Expectation value. The number of different alignments with scores equivalent to or better than S that are **expected** to occur in a database search **by chance**. ***The lower the E value, the more significant the score.***

- *Low E-values suggest that sequences are homologous*

  - Statistical significance depends on both the size of the alignments and the size of the sequence database.

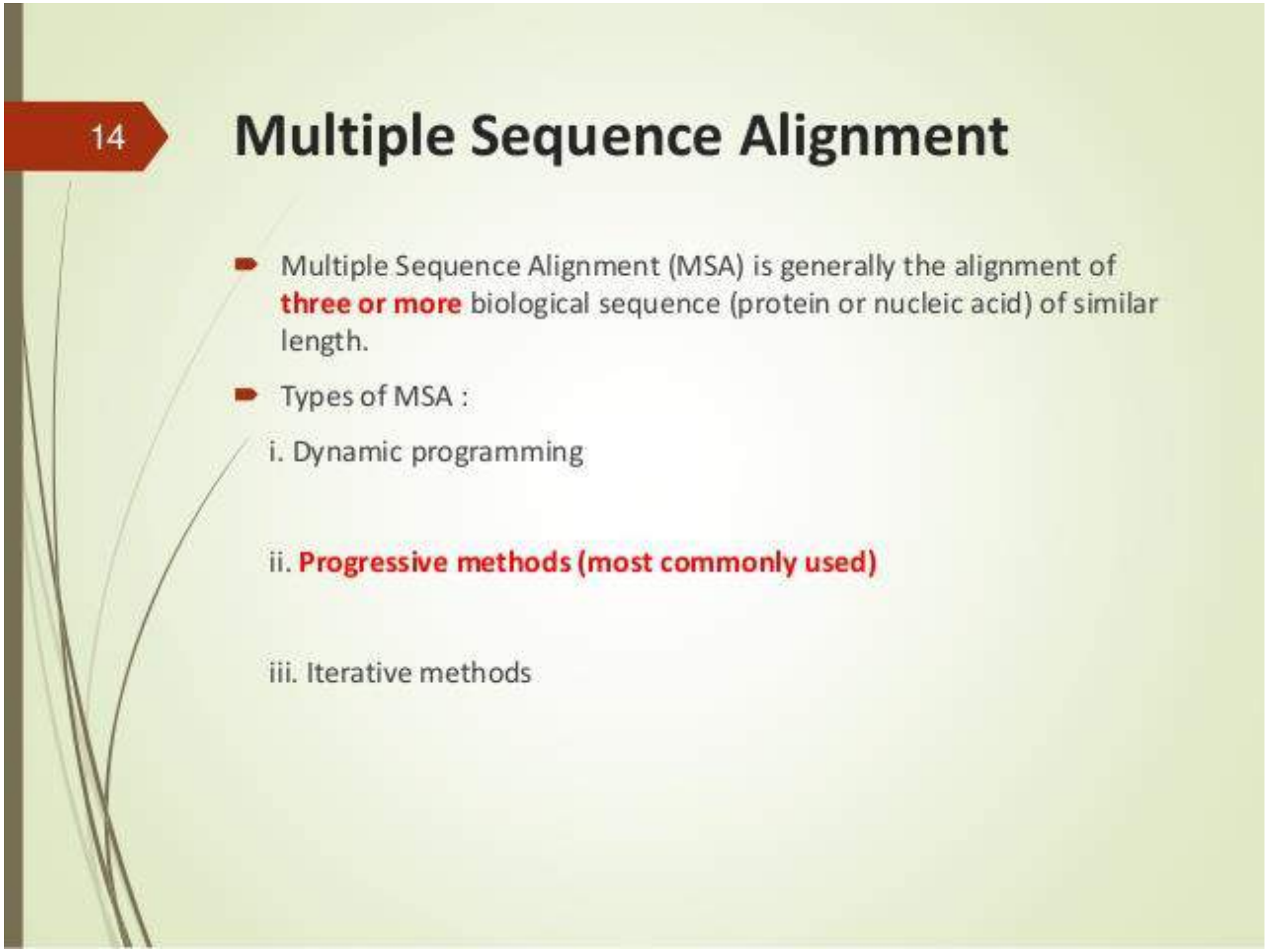
    - Important consideration for comparing results across different searches.
    - ***E-value increases as database gets bigger***
    - ***E-value decreases as alignments get longer***





FASTA

- FASTA package was 1st described by **Lipman** and **Pearson** in1985.
- FASTA is a DNA and protein sequence alignment software.
- FASTA is a fast homology search tool.
- FAST-P stands for protein , compare the amino acid sequence of proteins and FAST-N stands for nucleotide alignment, compare the nucleotide sequence of DNA.
- Usually **slowe**r than **BLAST.**



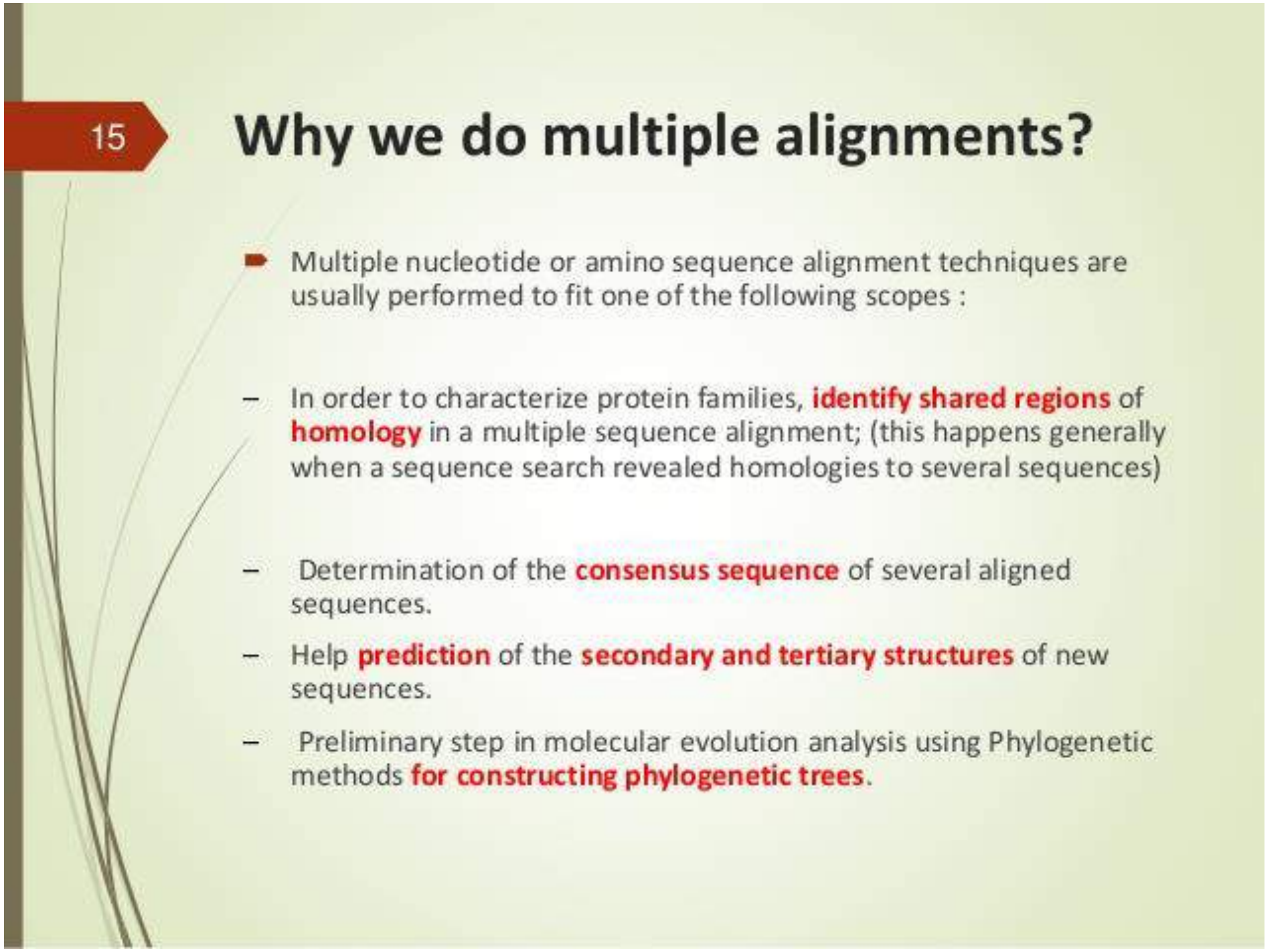
Multiple Sequence Alignment

- Multiple Sequence Alignment (MSA) is generally the alignment of **three or more** biological sequence (protein or nucleic acid) of similar length.
- Types of MSA :

i. Dynamic programming

ii. **Progressive methods (most commonly used)**

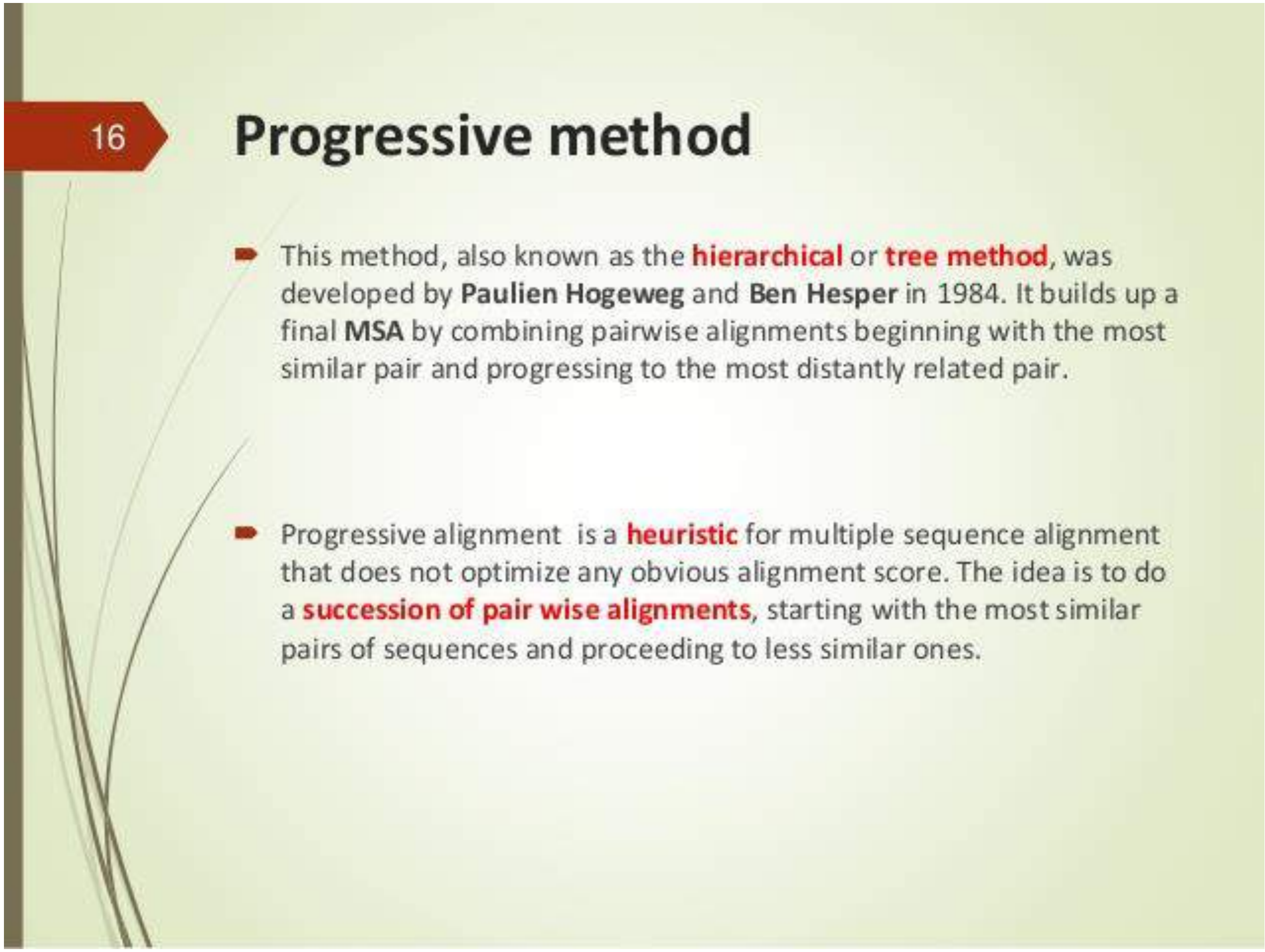
iii. Iterative methods



Why we do multiple alignments?

- Multiple nucleotide or amino sequence alignment techniques are usually performed to fit one of the following scopes :

- In order to characterize protein families, **identify shared regions** of **homology** in a multiple sequence alignment; (this happens generally when a sequence search revealed homologies to several sequences)
- Determination of the **consensus sequence** of several aligned sequences.
- Help **prediction** of the **secondary and tertiary structures** of new sequences.
- Preliminary step in molecular evolution analysis using Phylogenetic methods **for constructing phylogenetic trees**.



Progressive method

- This method, also known as the **hierarchical** or **tree method**, was developed by **Paulien Hogeweg** and **Ben Hesper** in 1984. It builds up a final **MSA** by combining pairwise alignments beginning with the most similar pair and progressing to the most distantly related pair.

- Progressive alignment is a **heuristic** for multiple sequence alignment that does not optimize any obvious alignment score. The idea is to do a **succession of pair wise alignments**, starting with the most similar pairs of sequences and proceeding to less similar ones.

- The steps are summarized as follows:

- Compare all sequences pairwise.

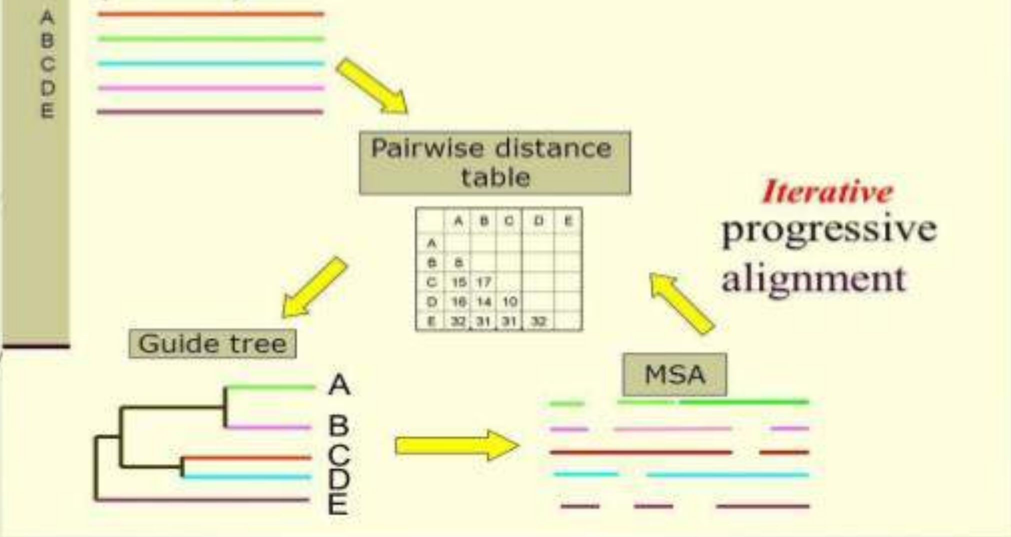
- Perform cluster analysis on the pairwise data to generate a hierarchy for alignment. This may be in the form of a **binary tree** or a simple ordering.

- Build the multiple alignment by first aligning the most similar pair of sequences, then the next most similar pair and so on.

- Once an alignment of two sequences has been made, then this is fixed.

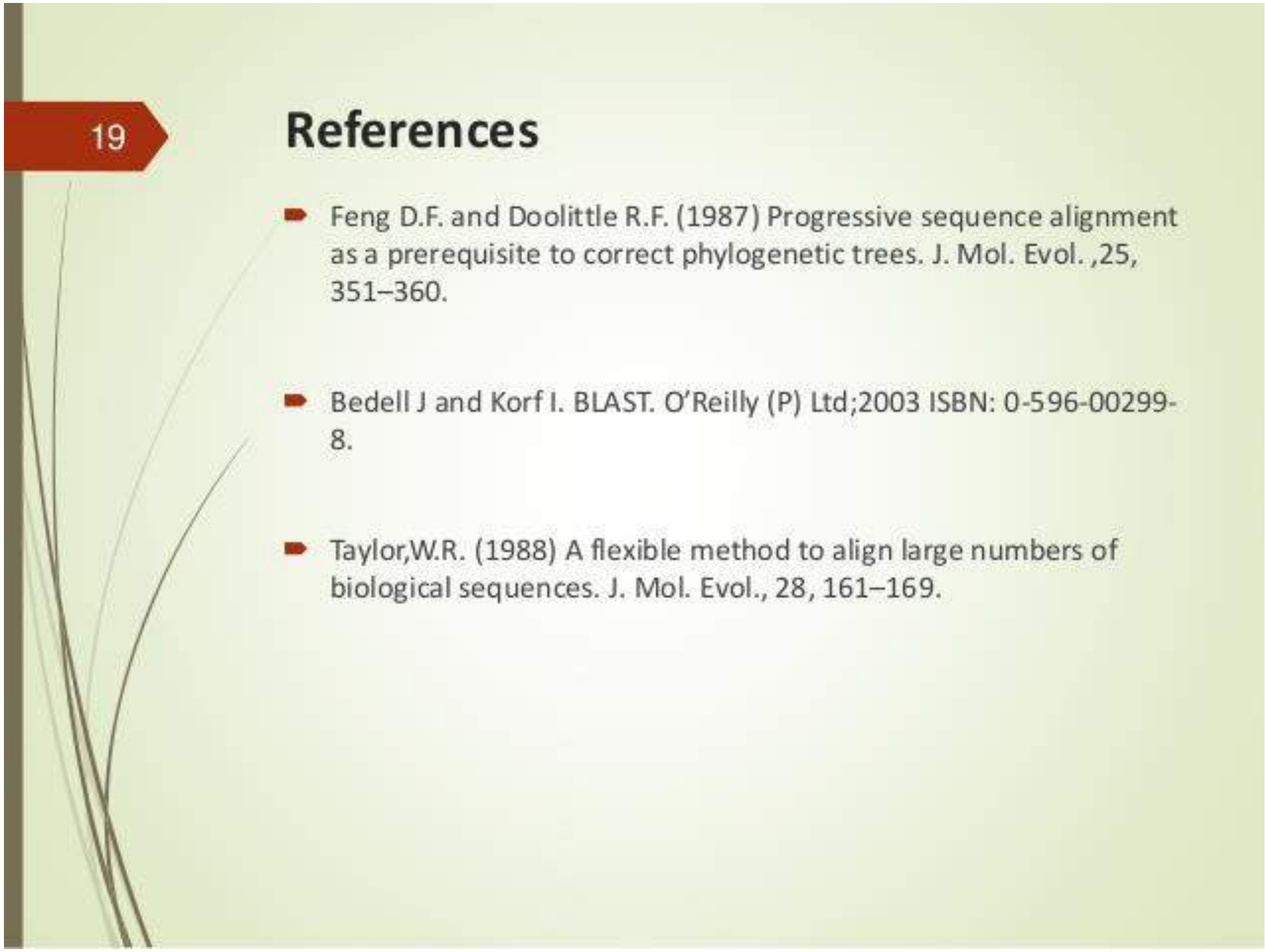
- Thus for a set of sequences A, B, C, D having aligned A with C and B with D the alignment of A, B, C, D is obtained by comparing the alignments of A and C with that of B and D using averaged scores at each aligned position.

Multiple sequence alignment (MSA)



	A	B	C	D	E
A					
B	8				
C	15	17			
D	16	14	10		
E	32	31	31	32	

Fig. 3 Steps of MSA
Source : https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html



References

- Feng D.F. and Doolittle R.F. (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. ,25, 351–360.
- Bedell J and Korf I. BLAST. O'Reilly (P) Ltd;2003 ISBN: 0-596-00299-8.
- Taylor,W.R. (1988) A flexible method to align large numbers of biological sequences. J. Mol. Evol., 28, 161–169.

thank you!
pptbackgrounds.org