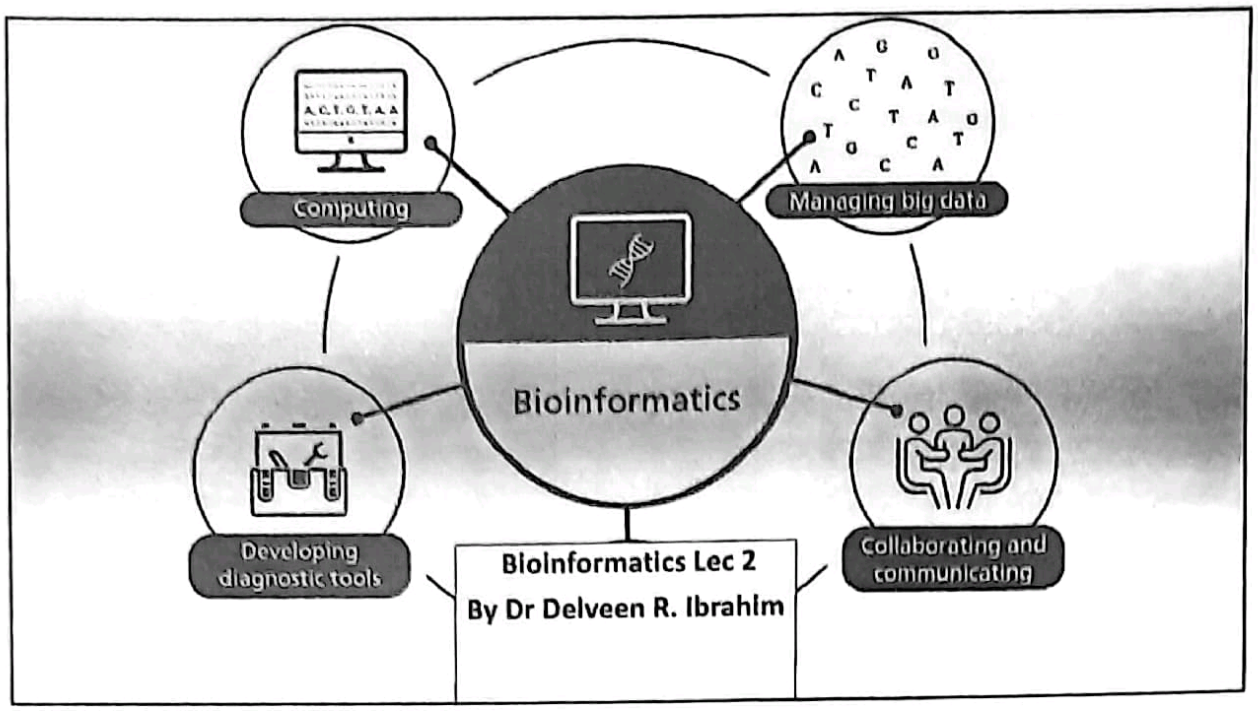
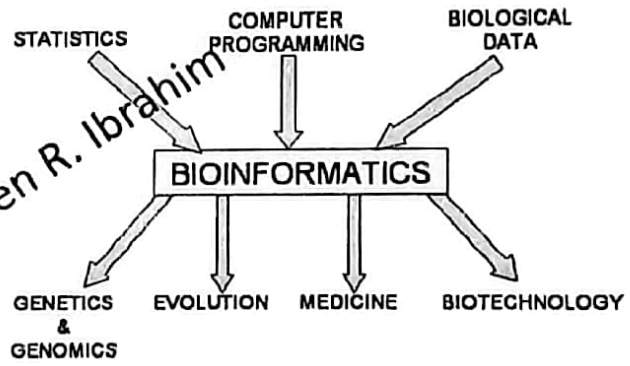
(2)



1

What is Bioinformatics??

- Bioinformatics, is a scientific subdiscipline that involves using computer technology to collect, store, analyze and disseminate biological data and information, such as DNA and amino acid sequences.



2

Aims of bioinformatics

The field of bioinformatics has three main objectives:

1. To organize vast reams of molecular biology data in an efficient manner
2. To develop tools that aid in the analysis of such data
3. To interpret the results accurately and meaningfully

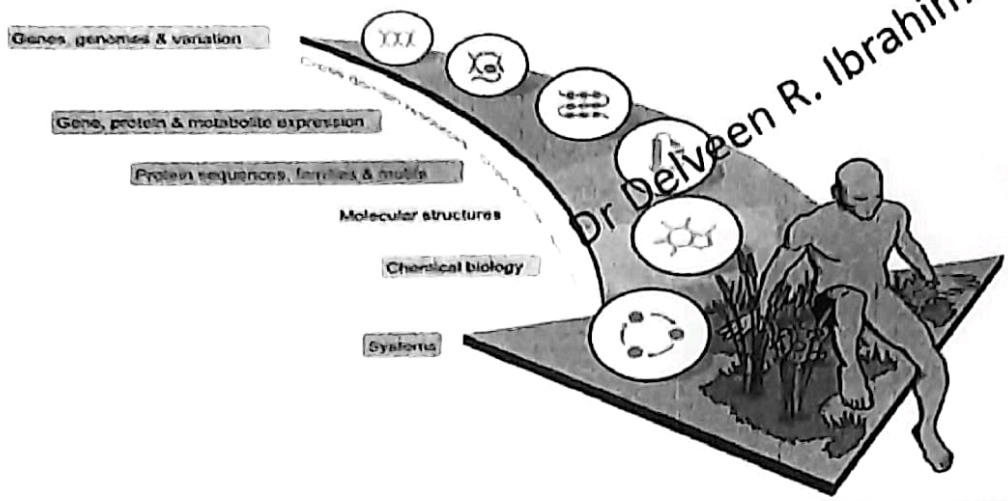
3

Since the foundation of molecular biology, there has been a need to collect/manage, analyze and present biological information in an organized fashion. **Biological information is not just sequences of DNA, it can be:**

- Genomic DNA, SNPs (single nucleotide polymorphism), genetic markers, mobile elements
- ESTs (Expressed sequence tags (ESTs), short sequences of random cDNAs derived from the mRNA of particular species, tissues, etc.
- Protein sequences
- Protein structures
- Protein interactions, macromolecule interactions
- Gene expression (microarray data)
- Phenotypes such as mutants and QTLs (A quantitative trait locus)
- Molecular functions of genes, proteins
- Metabolic (biochemical) pathways
- Journal articles

4

A broad overview of the different types of data that fall within the scope of bioinformatics



5

Applications of bioinformatics:

- Manipulation of DNA and protein sequence,
- Maintenance of sequence and other databases in **databases** designed to allow their easy manipulation and inspection,
- Analytical methods, particularly for the comparison of DNA and protein sequences **(sequence similarity searching, sequence alignment)**
- Analysis of **phylogenetic** relationships by **multiple sequence alignment**, most often of protein or rDNA sequences.
- Statistical analysis of genomic and proteomic data.

6

History of the bioinformatics

Paulien Hogeweg

Term bioinformatics was invented by Paulien Hogeweg and Ben Hesper in 1970 as "the study of informatic processes in biotic system

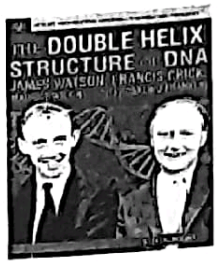
While the first informatician is Margaret Dayhoff (1925–1983) , was an american physical chemist who pioneered the application of computational methods to the field of biochemistry

Ben Hesper

7

History of Bioinformatics

1952 Sanger sequenced bovine insulin



1953 Watson and Crick published DNA structure

Margaret Dayhoff

An American physical chemist and pioneer in the field of bioinformatics, she pioneered the use of mathematics and computational methods in the field of biochemistry.

1965 Margert Dayhoffs published protein sequence Atlas

1970 Paulien Hogeweg and Ben Hesper invented the term of Bioinformatics

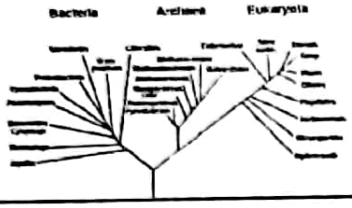
1977 DNA sequencing (Sanger) and software to analyze it (Staden)

SANGER SEQUENCING

8

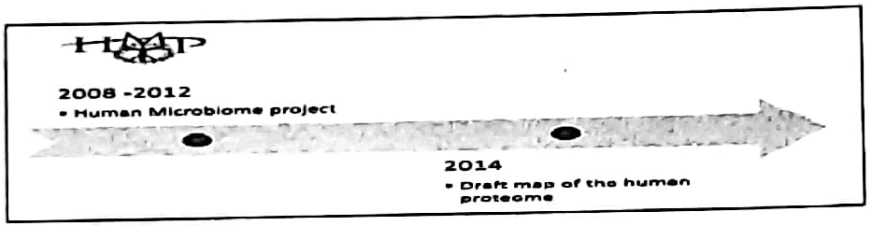
History of Bioinformatics

- **1993** Sanger center for sequencing, UK
- **1995** *Hemophilus influnzae* genome Sequenced and Methanococcus genome sequenced , which confirmed the existence of third major branch of life on earth



- **1997** *E. coli* Genome sequenced completely
- 1998 worm genome and fly genome sequenced
- In the 1990–2000s, use of the Internet, coupled with next-generation sequencing (NGS) , led to an exponential entry of data and a rapid increase of bioinformatics tools.
- 2001-2003 Human genome sequenced complete

9



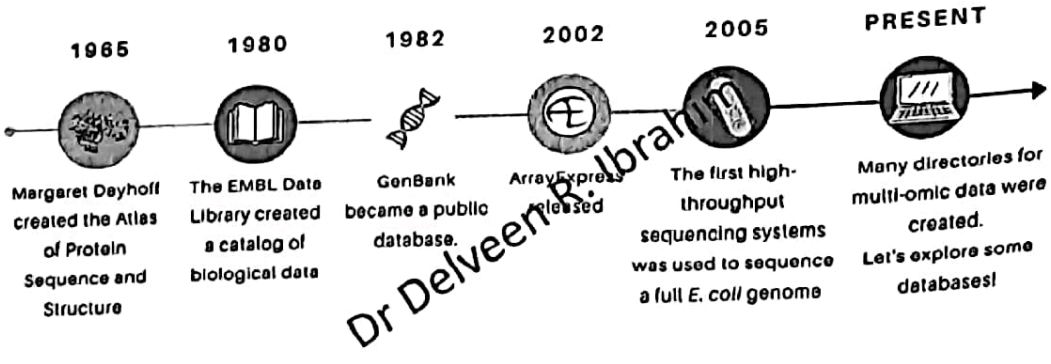
- Today, bioinformatics faces multiple challenges, such as handling Big Data, ensuring the reproducibility of results and a proper integration into academic curriculums.

10

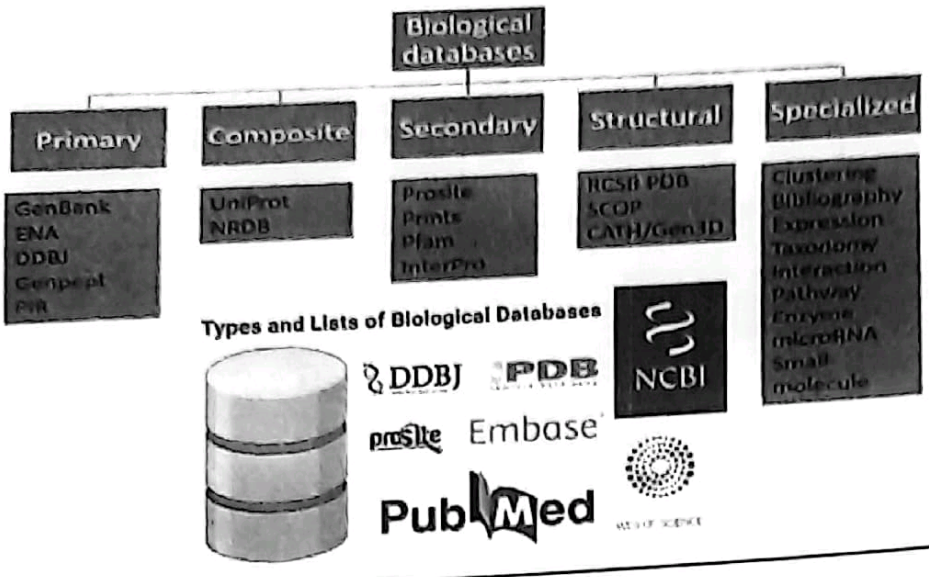
Biological Databases

11

A brief history of biological databases

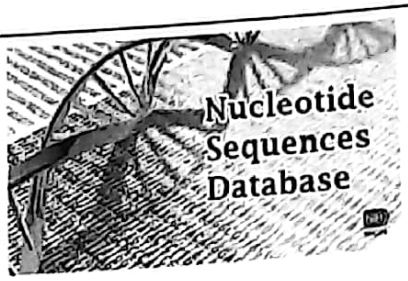


12



13

Biological databases



- Nucleotide and protein sequences databases:
- **Primary databases**
- **Secondary databases**

14

Primary databases

- It can also be called an archival database since it archives the experimental results submitted by the scientists. The primary database is populated with experimentally derived data like genome sequence, macromolecular structure, etc.
- It obtains unique data obtained from the laboratory and these data are made accessible to normal users without any change.
- The data are given accession numbers when they are entered into the database. The same data can later be retrieved using the accession number. Accession number identifies each data uniquely and it never changes.
- **Examples –**
- **Nucleic Acid Databases:** GenBank, EMBL and DDBJ
- **Protein Databases** are: PDB, Swiss-Prot, PIR,TrEMBL etc.

15

Secondary databases

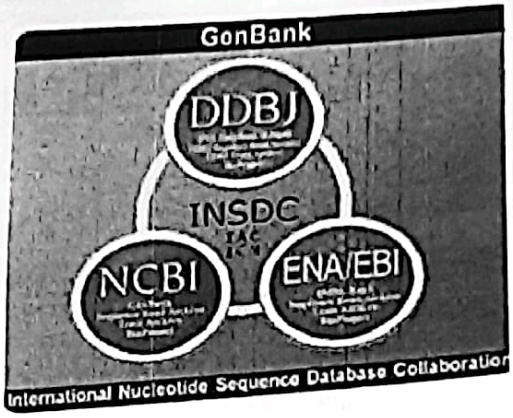
- The data stored in these types of databases are the analyzed result of the primary database.
- The data here are highly curated(processing the data before it is presented in the database). A secondary database is better and contains more valuable knowledge compared to the primary database.
- **Examples –**
- DNA databses: Refseq **(comprehensive, integrated, well-annotated set of sequences, including genomic DNA)**
- SNP- disease databases
- Protein Databases:
- InterPro (protein families, motifs, and domains)
- UniProt/KB (sequence and functional information on proteins)

16

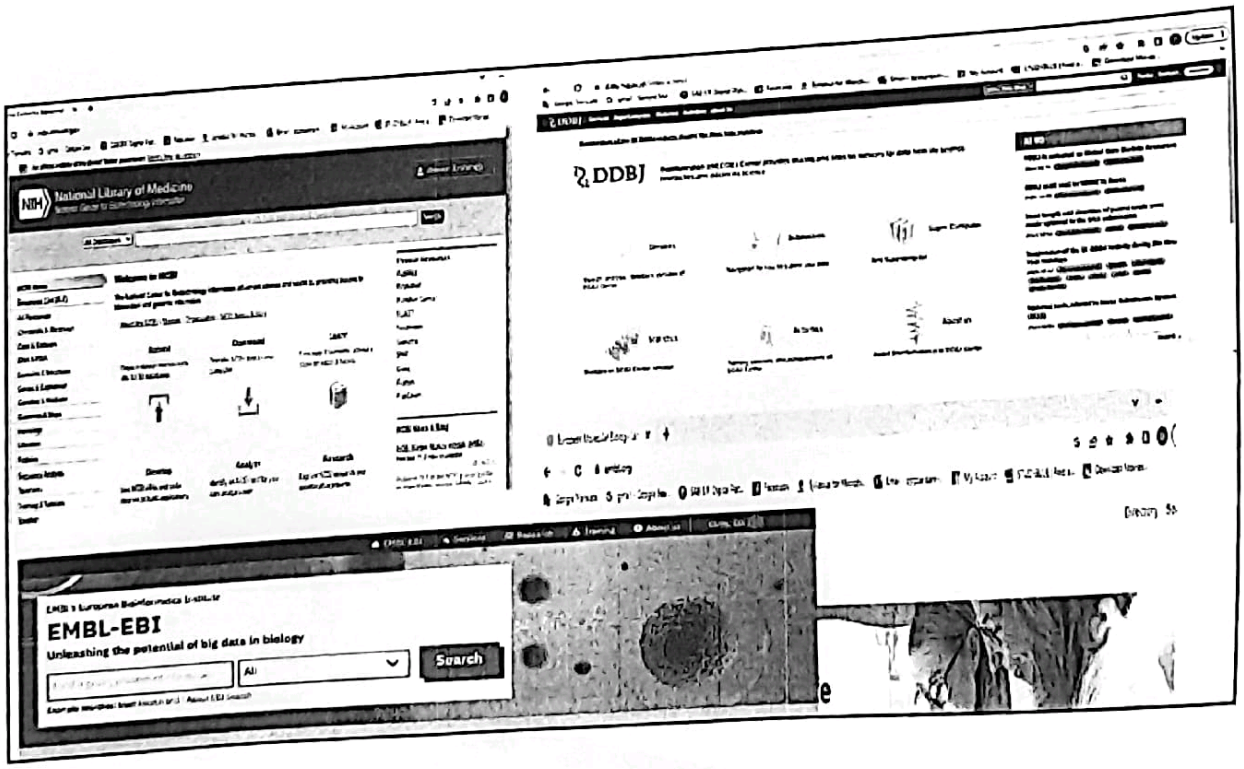
Nucleic acid databases

- GenBank (NCBI), USA.
- DDBJ (DNA Data Bank of Japan).
- EMBL-EBI (European Bioinformatics Institute).

- Three databases are the sources for Nucleotide sequence data from all organisms.
- All three Databases accept nucleotide sequence submission.
- They exchange new and updated data on daily basis to achieve optimal synchronization between them.



17



18

Abbreviation

- NCBI: National Center for Biotechnology information
- EMBL: European Molecular Biology Library
- DDBJ: DNA Databank of Japan
- UniProtKB: Universal Protein Knowledgebase
- PIR: Protein Information Resources
- PDB: Protein Data Bank
- NRDB: Non-Redundant Databases
- PubMed: Public/Publisher MEDLINE
- NLM: National library of medicine
- PMC: PubMed Central
- BLAST: The Basic Local Alignment Search Tool

19

- Lets Navigate EBI at https://www.ebi.ac.uk/
- DDBJ at https://www.ddbj.nig.ac.jp/
- PMC at https://www.ncbi.nlm.nih.gov/pmc/
- Pubmed https://pubmed.ncbi.nlm.nih.gov/

20