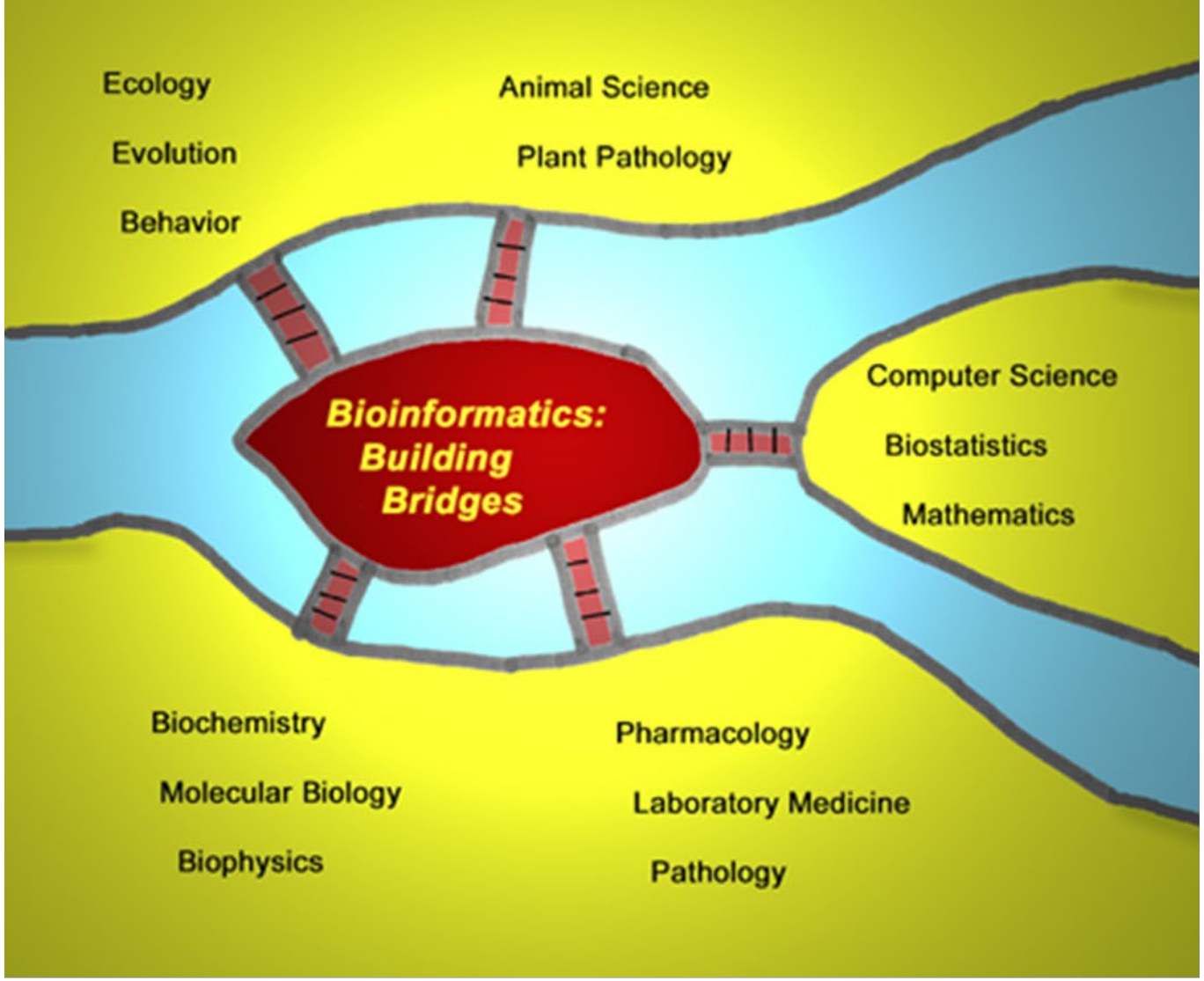
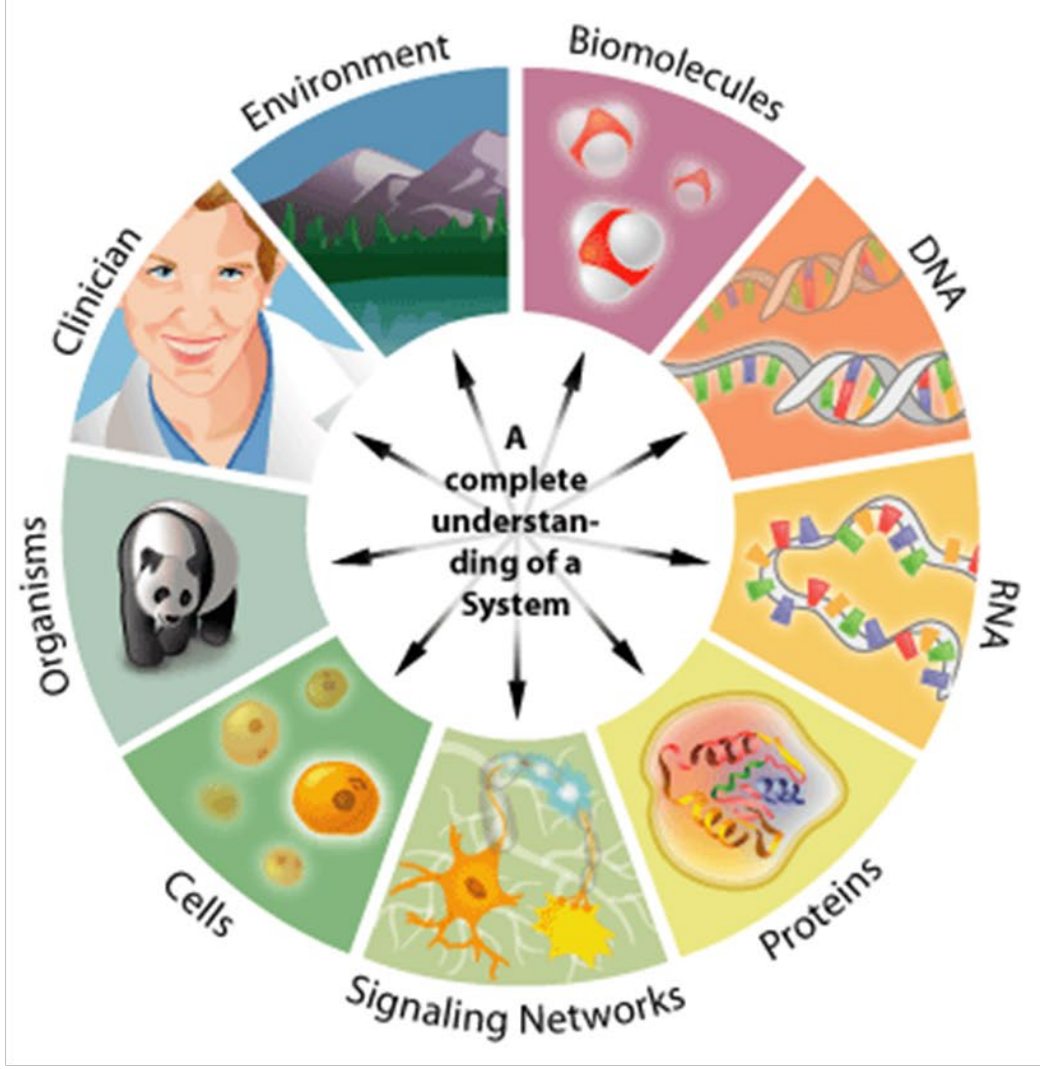
MSc III Sem – Life Sciences

Course – Bioinformatics

Introduction to Bioinformatics and Historical Background

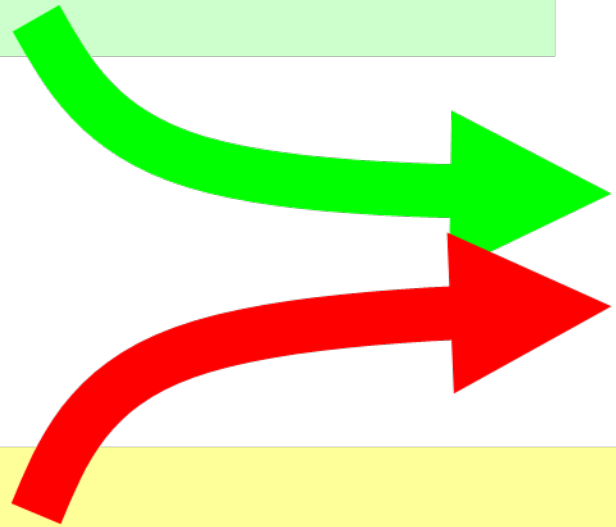




What is Bioinformatics?

The field of science in which **biology**, **computer science** and **information technology** merge into a single discipline

Biologists collect molecular data: DNA & Protein sequences, gene expression, etc.



Bioinformaticians
Study biological questions by analyzing molecular data

Computer scientists
(+Mathematicians, Statisticians, etc.)
Develop tools, softwares, algorithms to store and analyze the data.

Definition

Large databases that can be accessed and analyzed with sophisticated tools have become central to biological research and education. The information content in the genomes of organisms, in the molecular dynamics of proteins, and in population dynamics, to name but a few areas, is enormous. Biologists are increasingly finding that the management of complex data sets is becoming a bottleneck for scientific advances. Therefore, **bioinformatics** is rapidly become a key technology in all fields of biology.

Bioinformatics (Oxford English Dictionary):

The branch of science concerned with information and information flow in biological systems, esp. the use of computational methods in genetics and genomics.

- **Bioinformatics** is an interdisciplinary field mainly involving molecular biology and genetics, computer science, mathematics, and statistics. Data intensive, large-scale biological problems are addressed from a computational point of view.
- A bioinformatic analysis usually involves the following steps: (i) collects statistics from biological data, (ii) builds a computational model, (iii) solves a computational modeling problem, and (iv) test and evaluate a computational algorithm.
- Bioinformatics mainly resolves problems by organizing the types of data sources.
- Sequence analysis is the analysis of DNA and protein sequences for clues regarding function and includes subproblems such as identification of homologs, multiple sequence alignment, searching sequence patterns, and evolutionary analyses.
- Protein structures are three-dimensional data and the associated problems are structure prediction (secondary and tertiary), analysis of protein structures for clues regarding function, and structural alignment.
- Gene expression data is usually represented as matrices and analysis of microarray data mostly involves statistics analysis, classification, and clustering approaches.
- Biological networks such as gene regulatory networks, metabolic pathways, and protein-protein interaction networks are usually modeled as graphs and graph theoretic approaches are used to solve associated problems such as construction and analysis of large-scale networks.

Molecular Bioinformatics

Molecular Bioinformatics involves the use of computational tools to discover new information in complex data sets (from the one-dimensional information of DNA through the two-dimensional information of RNA and the three-dimensional information of proteins, to the four-dimensional information of evolving living systems).

What is Database

- **General:**
- A database is any collection of related data.
- A Computerized archive used to store and organize data in such a way that information can be retrieved easily.
- A database is a collection of interrelated data store together without harmful and unnecessary redundancy (duplicate data) to serve multiple applications
- Retrieving is called firing a query.

DATABASE SYSTEM

Database System is an integrated collection of related files along with the detail about their definition, interpretation, manipulation and maintenance

A database system is based on the data. Also a database system can be run or executed by using software called DBMS (Database Management System).

A database system controls the data from unauthorized access.

A database management system (DBMS) is a collection of programs that enables users to create and maintain a database.

What Does a DBMS Do?

Database management systems provide several functions in addition to simple file management:

- allow concurrency
- control security
- maintain data integrity
- provide for backup and recovery
- control redundancy
- allow data independence
- provide non-procedural query language
- perform automatic query optimization

What is a relational database?

- a database that treats all of its data as a collection of relations

Origin of bioinformatics and biological databases:

The first protein sequence reported was that of bovine insulin in **1956**, consisting of 51 residues.

Nearly a decade later, the first nucleic acid sequence was reported, that of yeast $tRNA^{alanine}$ with 77 bases.

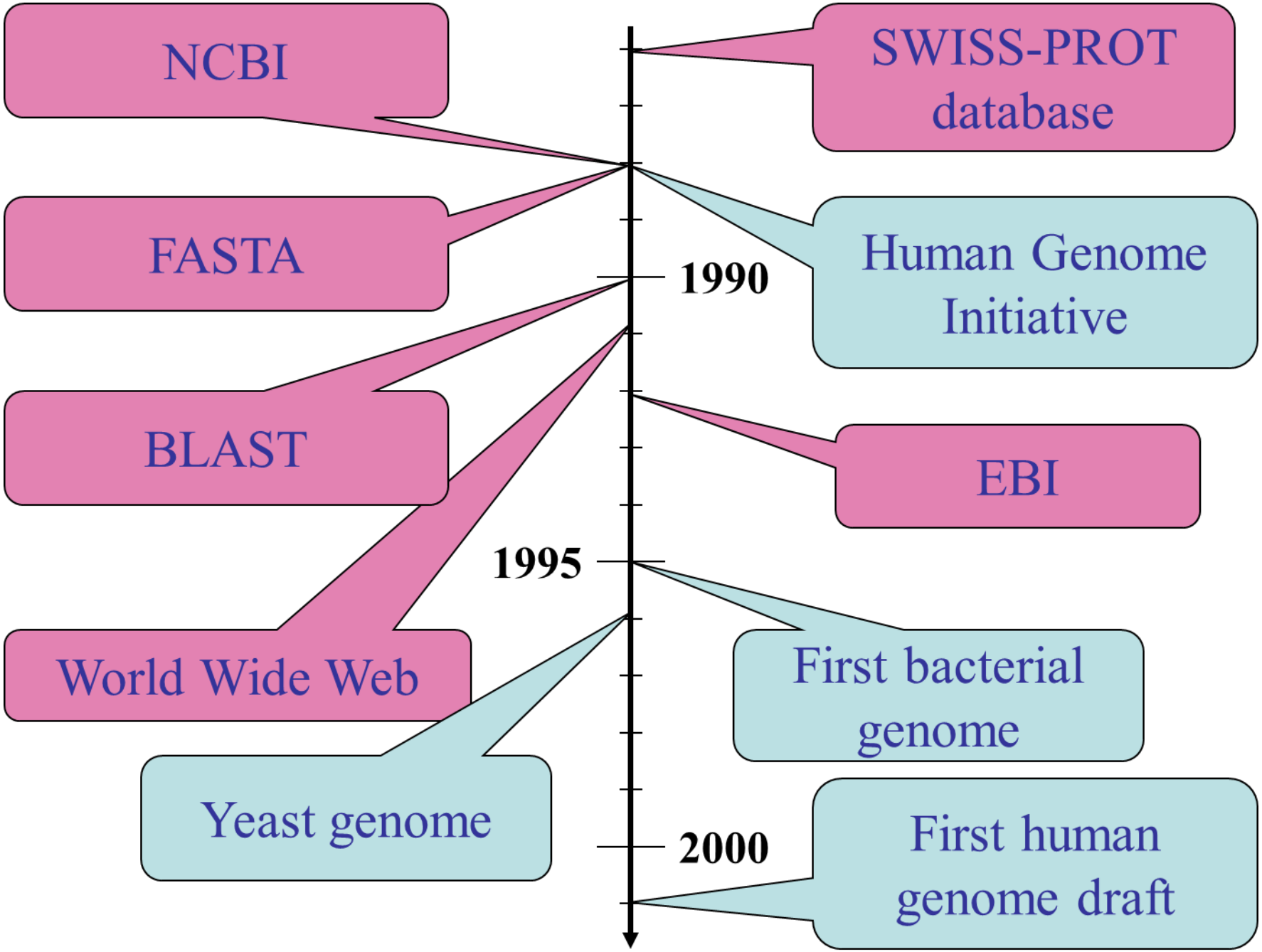
In 1965, **Dayhoff** gathered all the available sequence data to create the first bioinformatic database (*Atlas of Protein Sequence and Structure*).

The Protein DataBank (PDB) followed in 1972 with a collection of ten X-ray crystallographic protein structures.

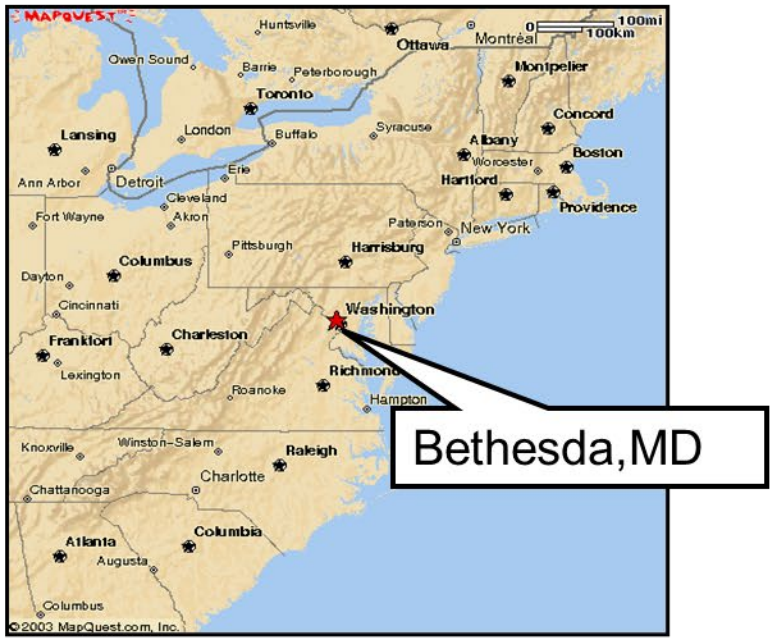
The SWISSPROT protein sequence database began in 1987.

Categories of databases for Life Sciences

- Sequences (DNA, protein)
- Genomics
- Mutation/polymorphism
- Protein domain/family
- Proteomics (2D gel, Mass Spectrometry)
- 3D structure
- Metabolic networks
- Regulatory networks
- Bibliography
- Expression (Microarrays,…)
- Specialized


NCBI
SWISS-PROT database
FASTA
1990
Human Genome Initiative
BLAST
EBI
1995
World Wide Web
First bacterial genome
Yeast genome
2000
First human genome draft

The National Center for Biotechnology Information



Created in 1988 as a part of the National Library of Medicine at NIH

- Establish public databases
- Research in computational biology
- Develop software tools for sequence analysis
- Disseminate biomedical information

NCBI

- Very comprehensive biological database
- GENBANK: The nucleotide sequence database
- Provides 42 different resource
- Provides a simple and easy to use web interface
- Sequence submission: done using Bankit or Sequin
- Search Engine for data retrieval: Entrez
- Retrieves information across all the resources under NCBI

  Example: PubMed, taxonomy, SNP, PubChem etc.

http://www.ncbi.nlm.nih.gov/

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: https://www.ncbi.nlm.nih.gov/sars-cov-2/.

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog
Hiding sequences in an alignment now available in the MSA Viewer!
04 Sep 2020
Do you ever wish there was a quick way to hide partial or poor quality sequences

GenBank release 239 is available
04 Sep 2020
GenBank release 239.0 (8/18/2020) is now available on the NCBI FTP site. This release has 9.89 trillion bases and 2.12

Coronavirus host gene regulatory elements now annotated by RefSeq Functional Elements

Contents from Open Access Online Resources