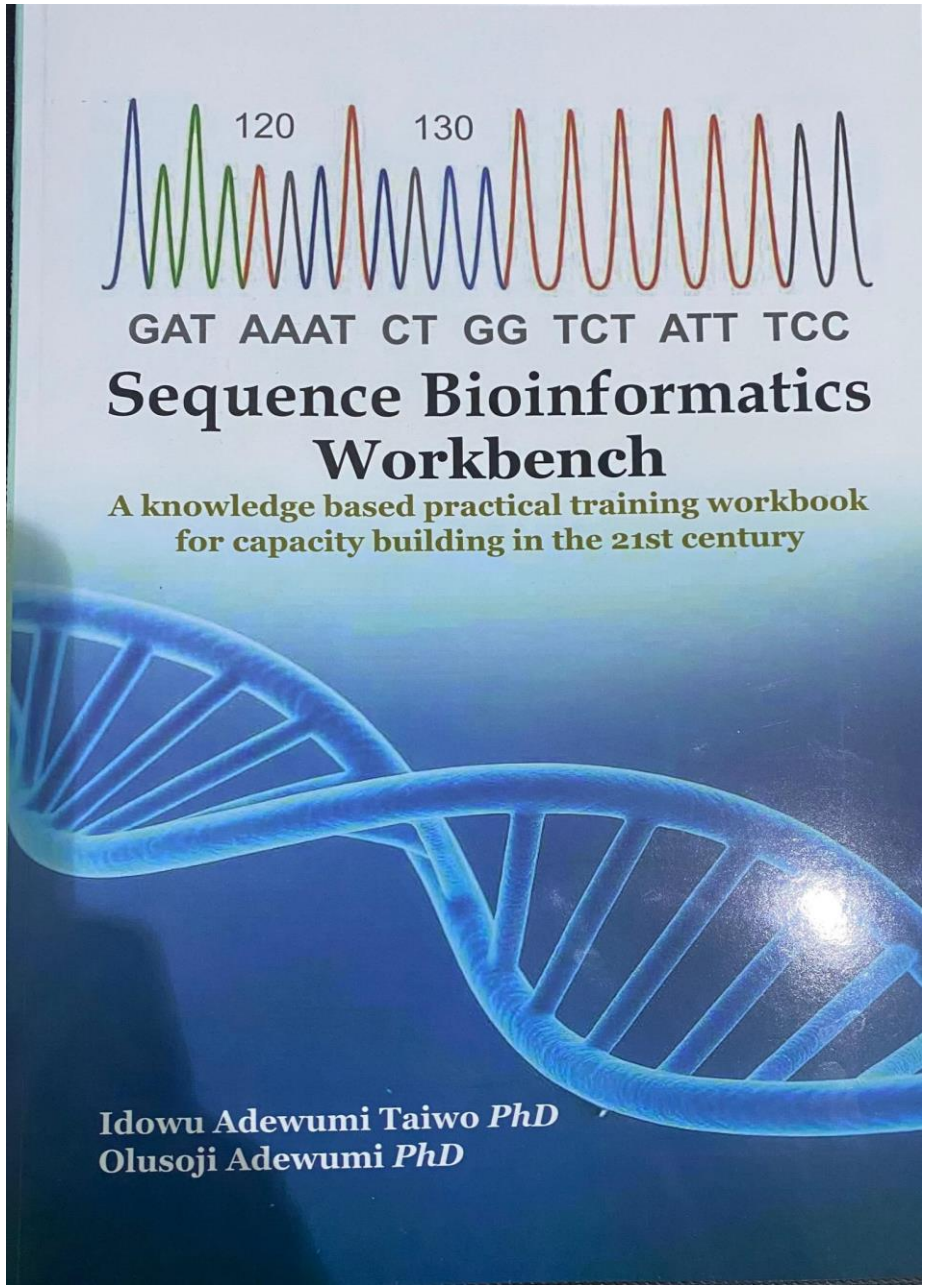
CGNPH Summer School

Data Mining – Data Retrieval

I. A. Taiwo PhD
(Professor of Human Genetics and Bioinformatics)

Outline of Presentation

- Introductory Lecture
- Historical Perspective
- Applications of Bioinformatics
- Databases
- File Formats



Content

Section One – Data Retrieval and Mining

1. Literature Databases
2. Fundamentals of Sequence Analysis
3. Protein Sequence Databases
4. Nucleotide Sequence Databases

Section Two–Sequence Comparison and Homology Search

5. Pairwise Sequence Alignment
6. Multiple Sequence Alignment
7. Basic Local Alignment Search Tool (BLAST)

Section Three – Applications of Sequence Bioinformatics

8. In Silico PCR
9. Restriction Analysis
10. Phylogenetic Analysis

Introductory Lecture

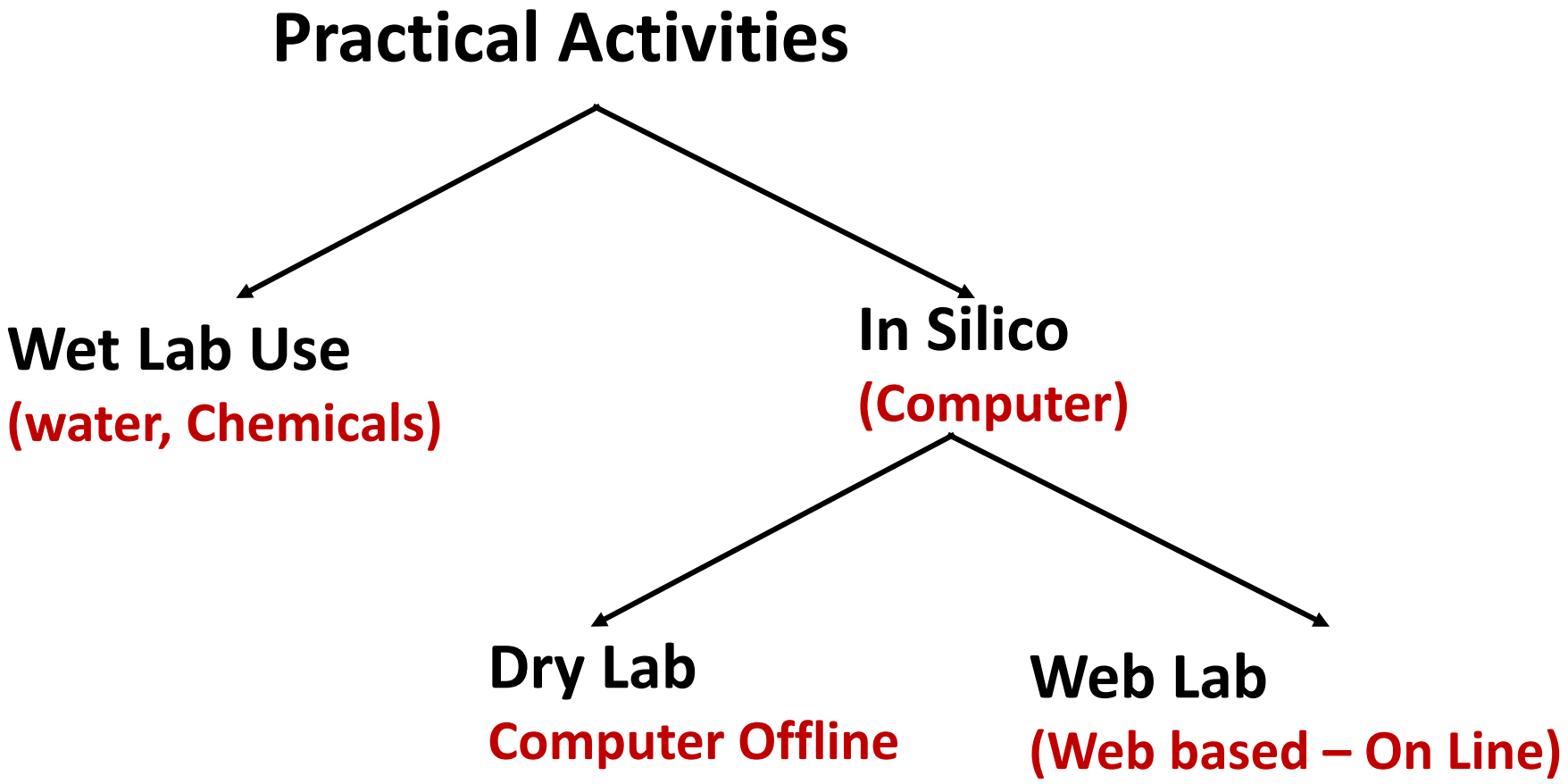
What is Bioinformatics?

The use of computer to manage and analyze biological data contained in databases.

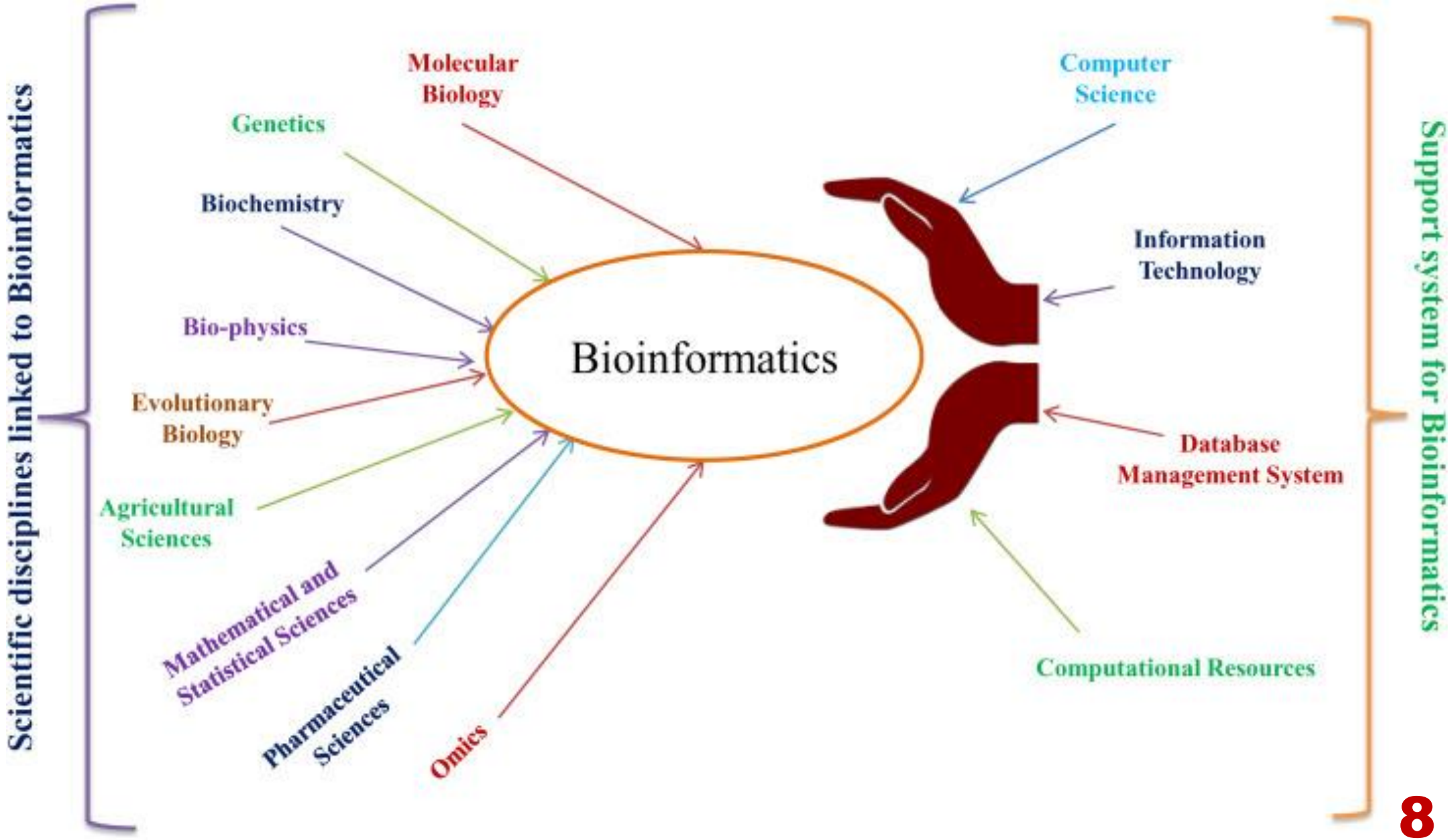
Bioinformatics Vs Computational Biology

- **Computational biology**
  - development of algorithms and statistical models to analyze biological data

- **Bioinformatics**
  - collection and storage of biological information
  - derives knowledge from computer analysis of biological data


Practical Activities
Wet Lab Use
(water, Chemicals)
In Silico
(Computer)
Dry Lab
Computer Offline
Web Lab
(Web based – On Line)

Applications of Bioinformatics



History of Bioinformatics

1950-1970: Phase 1 – Protein Sequencing and Protein database

- **Pre-History**
  - Protein sequencing – Edman degradation
  - Molecular Evolution (Functionally related proteins are similar) - Zuckerkandl and Pauling

- **Margarett Dayhoff and Co-Workers** (Mother of Bioinformatics)
  - Produced the first biological database – “Atlas of protein Sequence and Structures”.
  - Developed COMPROTEIN – a protein assembly computer program
  - Introduced 3-letter and 1-letter code for amino acid
  - Introduced PAM - a substitution matrix for determining distance between proteins
- CLUSTAL – A multiple sequence alignment program for protein sequences was developed.

1970-1980: Phase 2 – DNA Sequencing

- 1975: Sanger-Coulson - Plus-minus method
- 1977: Maxam-Gilbert – Chemical degradation method
- 1977: Sanger's sequencing – Chain termination method

1970-1980: Phase 2 – Genome Sequencing

- 1976: First genome sequenced – Bacteriophage MS2 (RNA virus) by Fiers et al. (1976)
- 1977 – Second genome sequenced – Bacteriophage Φ174 (DNA virus) by Sanger et al. (1977).
- 1995 – Third genome (First cellular genome) – *Haemophilus influenza* by Venter et al. (1995)

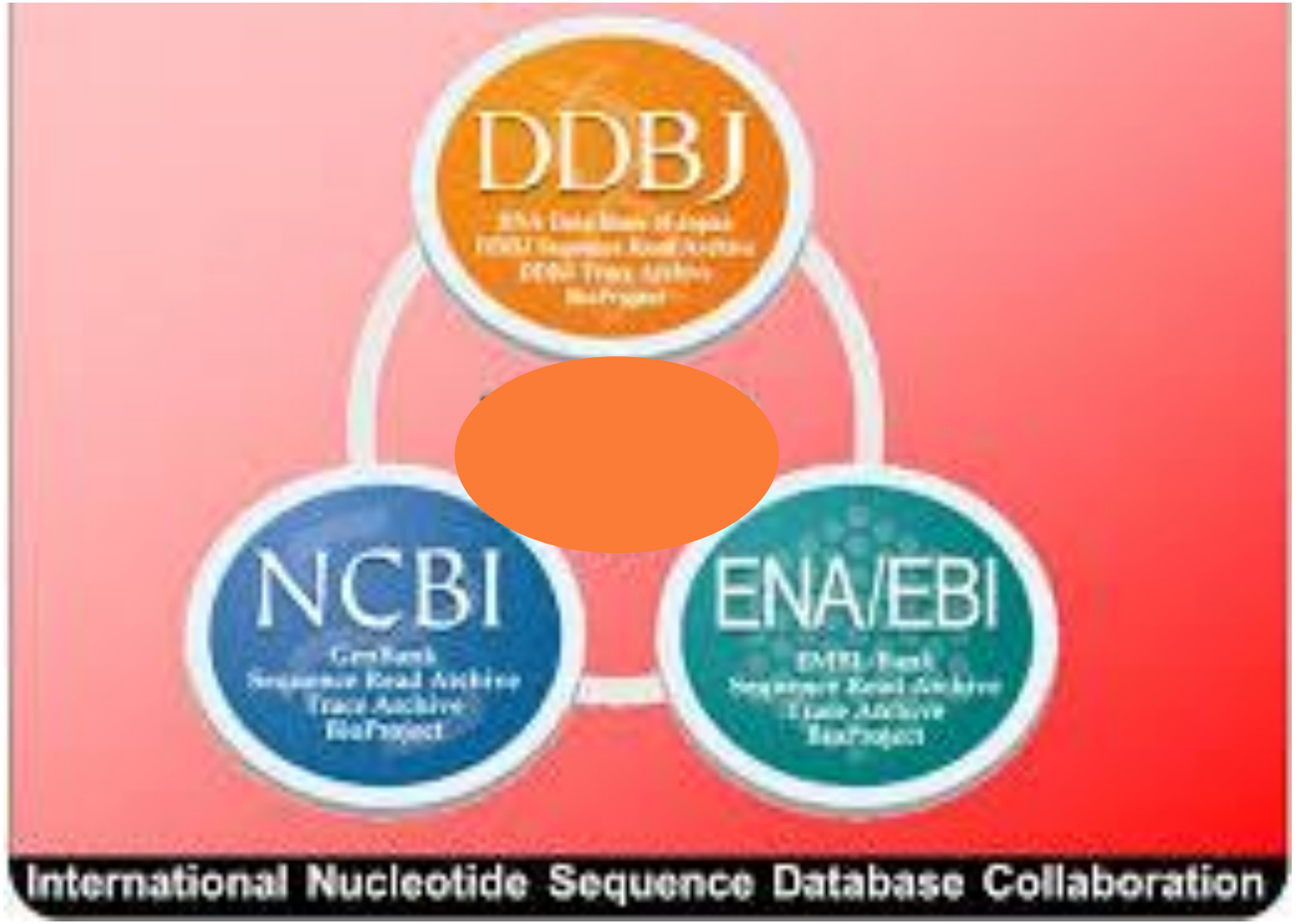
1980 -1990: Phase 3 – Advances in Biology and Computer Sciences

- DNA cloning
- PCR
- Software development
- Establishment of **NCBI** (in US), **EMBL** (in UK) and **DDBJ** (in Japan)

Note:

- In view of genomic data accumulation, bioinformatics emerged as a tool for genomic analysis
- NCBI + EMBL + DDBJ = INSDC (International Nucleotide Sequence Database Collaboration)

INSDC – International Nucleotide Sequence Database Collaboration



1990-2005: Phase 4 – Era of genomics

- 1990 - Commencement of Human Genome Project (HGP)
- 1990 - 2000 – Genome sequencing of model organisms
- 2001 - 1st draft of human genome released
- 2003 - Final draft of human genome released

2005 and After: Phase 5 - Postgenomic Era

HGP Spinoff Projects and NGS

- **HGP Spinoffs**
  - HapMap Project – Haplotype map to describe human sequence variation.
  - 1000 Genomes – Sequencing of not less than 1000 genomes in the world.
  - ENCODE (the Encyclopedia of DNA elements) – Functional elements in human genome.
  - Cancer Genome Project – Expression profile of normal, precancer, and cancer cells.
- **Next Generation Sequencing (NGS)**
- **Third Generation Sequencing**
- **Fourth Generation Sequencing**
- **Other Omics – Transcriptomics, proteomics, metabolomics, metagenomics etc.**
- **Multi-Omics Studies**

Emerging and Future Perspective: Phase 6

Artificial Intelligence (AI) and Related Areas (Machine Learning, Neural Networks etc.

- Protein folding problem
- Solving of protein folding problem – Alphafold 2
- Novel Protein designing

Databases

What is a database?

A repository or collection of related information

Data Mining
(Knowledge discovery)

Finding new patterns and relationship in large amount of data.

Data Retrieval
(Data acquisition)

Accessing and extracting specific biological information from various **databases** and other data storage systems (cloud storage, file systems etc.).

Data Retrieval Methods

- ❑ Downloading
- ❑ Copying and Pasting

Key Considerations in Bioinformatics Data Storage

Data Volume

Bioinformatics datasets can be massive, requiring large storage capacity and efficient data management strategies.

Data Complexity

Biological data comes in various formats and structures, requiring specialized storage solutions and data models.

Data Access Speed

Researchers need fast access to data for analysis and visualization, especially for large-scale studies.

Data Security

Protecting sensitive biological data from unauthorized access and cyber threats is crucial.

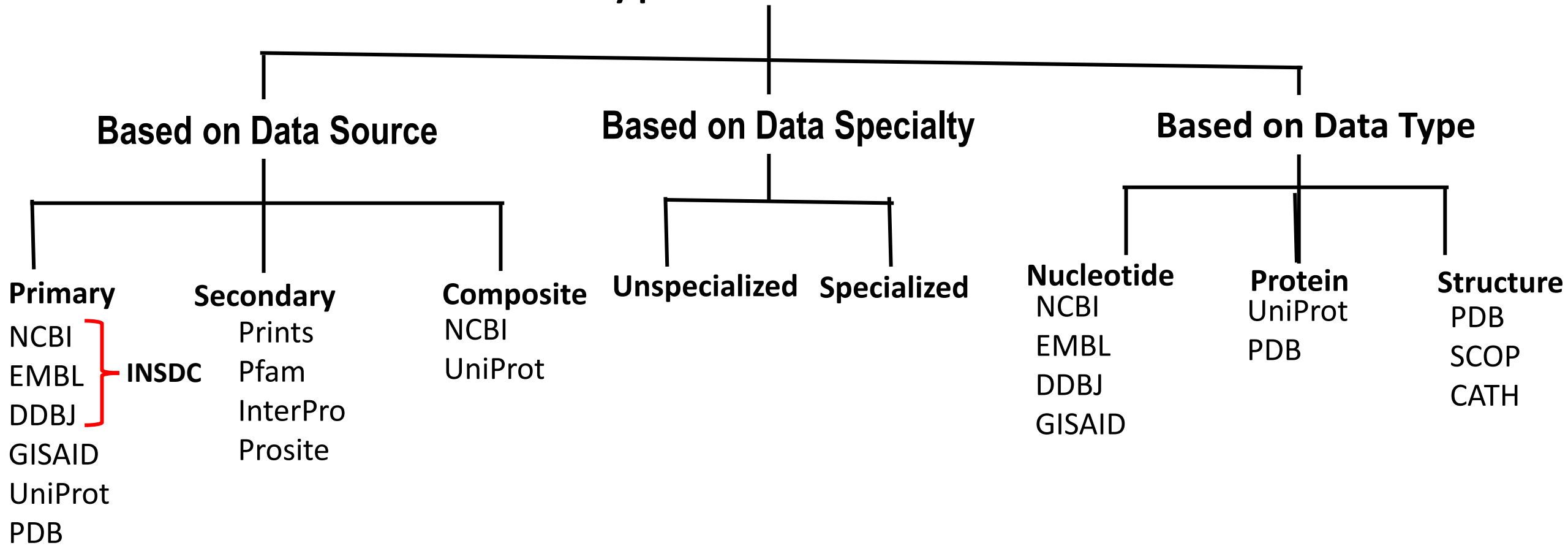
Integration with Analysis Pipelines

Storage systems should be compatible with common bioinformatics analysis tools and workflows.

Database Availability

- ❑ Freely available (e.g. GenBank, DDBJ, EMBL)
- ❑ Freely available but register and provide login details (e.g. GISAID)
- ❑ Not Free

Types of Database



Specialized Databases

- Literature (Bibliography) e.g. PubMed
- Sequence Databases e.g. NCBI GenBank, EMBL, DDBJ, UniProt
- Structure e.g. PDB
- Pathway e.g. KEGG
- Expression e.g. GEO
- MicroRNA
- Enzyme
- Small Molecule and Drugs e.g. PubChem, DrugBank
- Taxonomy

Thank You!