In the Name of God, the Merciful, the Compassionate

Introduction to Bioinformatics
03: Biological Databases

Instructor: Hossein Zeinali

Amirkabir University of Technology

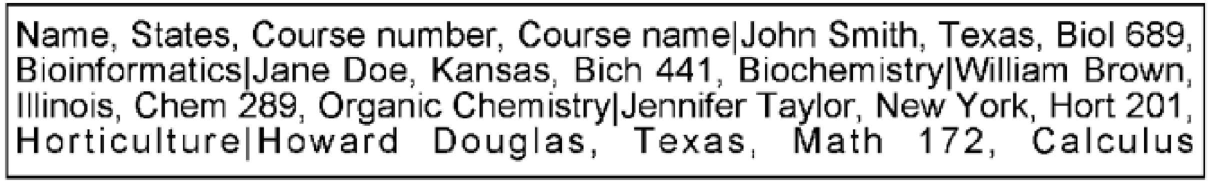
What is a Database?

- A ***database*** is a computerized archive used to store and organize data in such a way that information can be retrieved easily via a variety of search criteria.

Types of Databases

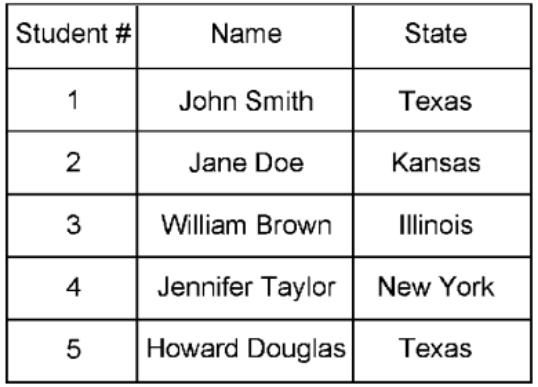
3 Major types of electronic databases:

1. **Flat files** - simple text files
   - no organization to facilitate retrieval



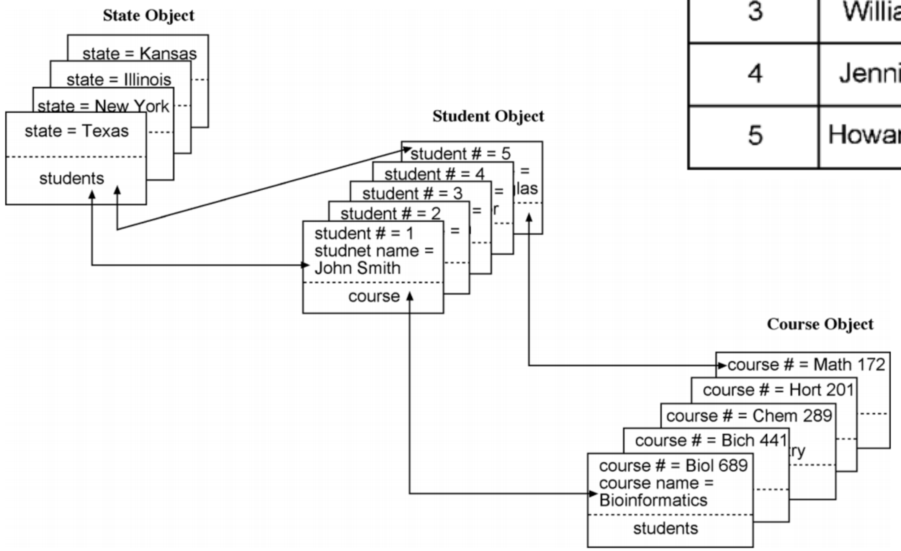
Name, States, Course number, Course name|John Smith, Texas, Biol 689, Bioinformatics|Jane Doe, Kansas, Bich 441, Biochemistry|William Brown, Illinois, Chem 289, Organic Chemistry|Jennifer Taylor, New York, Hort 201, Horticulture|Howard Douglas, Texas, Math 172, Calculus

2. **Relational** - data organized as tables ("relations")
   - shared features among tables allows rapid search



Student #	Name	State
1	John Smith	Texas
2	Jane Doe	Kansas
3	William Brown	Illinois
4	Jennifer Taylor	New York
5	Howard Douglas	Texas

3. **Object-oriented** - data organized as "objects"
   - objects associated hierarchically



Biological Databases

- Currently in all 3 types
  - – MANY flat files despite the obvious drawbacks of them

- What are goals of biological databases?
  - – Information retrieval
  - – Knowledge discovery

UniProt Universal Protein Resource
prosite
Fasta Similarity and Homology Searching
WU-Blast2 Basic Local Alignment Search Tool
IntEnz
SeqVISTA
Pfam
PRINTS Protein Fingerprints Database
InterPro
CAST
GENOMES
ClustalW Multiple Sequence Alignment Tool
emboss
ProDom
SMART
Alternative Splicing Database
Integr8
CSA Catalytic Site Atlas
PDBj Protein Data Bank Japan
TIGR
tigr fams
PANTHER Classification System
caBIG
SeqHound
PIRSF
Gene3D
BioLayout
NCBI
LION
Superfamily
SCOP
DALI
GeneQuiz
PubMed.gov U.S. National Library of Medicine National Institutes of Health
caBIO
MOUNT SINAI HOSPITAL
CATH Protein Structure Classification
MSD
e! Ensembl
Parasite Blast
DDBJ DNA Data Bank of Japan
wellcome trust sanger institute
SWISS-MODEL
ModBase
GENOME REVIEWS
IUPHAR
KEGG
PathPort The Pathogen Portal Web Project

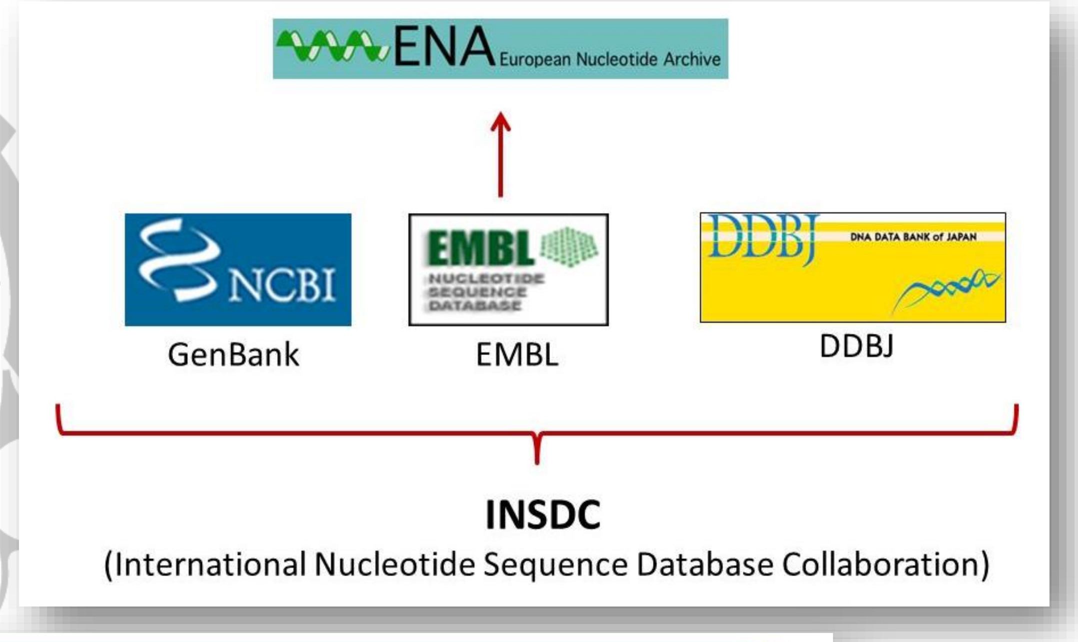
Types of Biological Databases

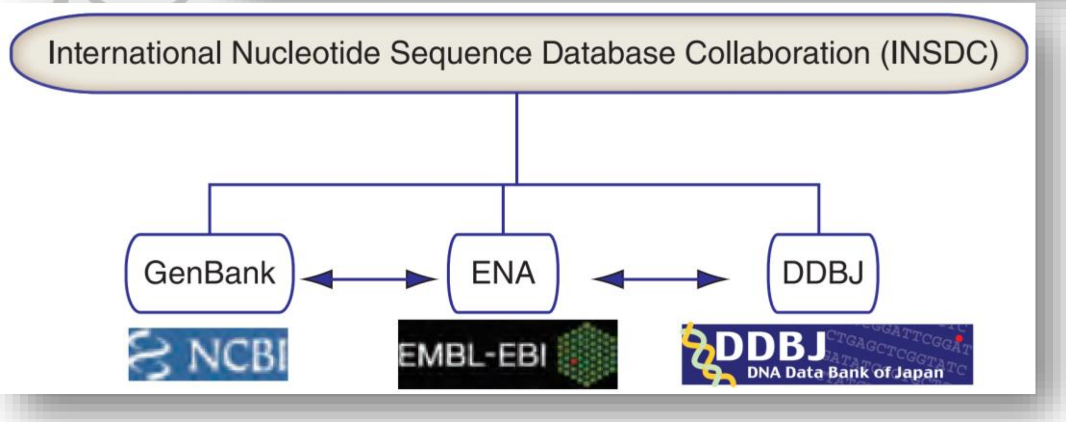
1. Primary
   - Simple archives of sequences, structures, images, etc.
   - Raw data, minimal annotations, not always well curated!
2. Secondary
   - Enhanced with more complete annotation of sequences, structures, images, etc.
   - Usually curated!
3. Specialized
   - Focused on a particular research interest or organism
   - Usually highly curated

Examples of Biological Databases

1- Primary

- **DNA sequences**
  - GenBank - US
  - European Molecular Biology Lab (EMBL)
  - DNA Data Bank of Japan (DDBJ)
- **Structures (Protein, DNA, RNA)**
  - PDB - Protein Data Bank (3D)
    - Flat file format
    - Archives atomic coordinates
  - NDB - Nucleic Acid Databank


ENA European Nucleotide Archive
NCBI
GenBank
EMBL NUCLEOTIDE SEQUENCE DATABASE
EMBL
DDBJ DNA DATA BANK of JAPAN
DDBJ
INSDC
(International Nucleotide Sequence Database Collaboration)


International Nucleotide Sequence Database Collaboration (INSDC)
GenBank
ENA
DDBJ
NCBI
EMBL-EBI
DDBJ DNA Data Bank of Japan

Examples of Biological Databases

2- Secondary

- **Protein sequences**
  - Swiss-Prot, TreEMBL, PIR
  - These recently combined into UniProt
  - Pfam, Blocks, DALI

3- Specialized

- **Species-specific (or "taxonomic" specific)**
  - HIV sequence, Flybase, WormBase, AceDB, PlantDB, GenBank EST
- **Molecule-specific, disease-specific**

Pitfalls of Biological Databases

- There are many errors in sequence databases
- Lack of documentation: quality or reliability of data
- Limited mechanisms for "data checking" or preventing propagation of errors (*esp. annotation errors!!*)
- Redundancy: NCBI has now created a *nonredundant* database, called RefSeq
- Inconsistency in annotation
- Incompatibility (format, terminology, data types, etc.)

Information Retrieval from Biological Databases

- **2 most popular retrieval systems:**
    - **ENTREZ -** developed and maintained by **NCBI**
    - **Sequence Retrieval Systems (SRS) -** maintained by **EBI**
- **Both:**
    - Provide access to multiple databases
    - Allow complex queries

References

- Mostly used:
  - – Essential bioinformatics, Chapter 2 (Introduction to Biological Databases)
- Second reference:
  - – Bioinformatics and functional genomics, Chapter 2 (Access to Sequence Data and Related Information)

Thanks for your attention