

Biological Database

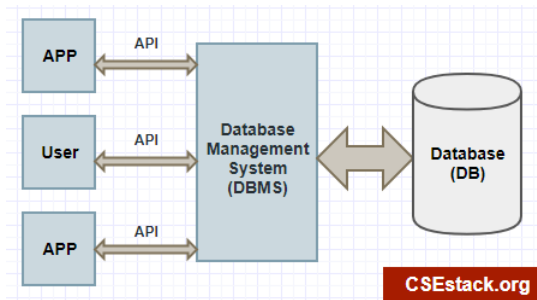
Luis Garreta

Electiva de Bioinformática
MAESTRÍA EN INFORMÁTICA BIOMÉDICA
Universidad del Bosque
Bogotá-Colombia

March 5, 2023

What is a database?

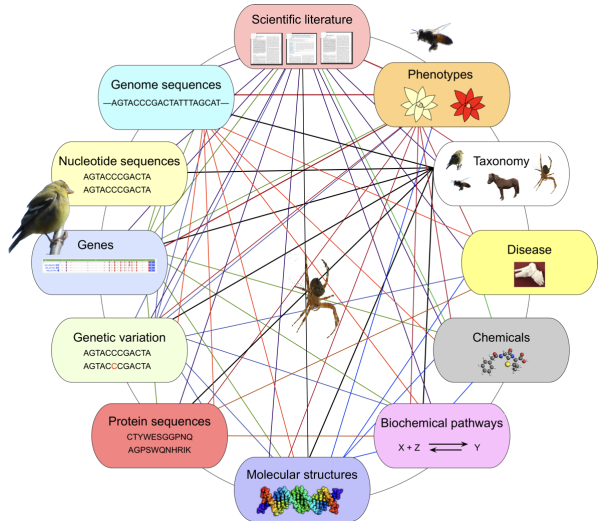
An organized collection of data or information that is electronically stored and accessible from a computer system



The organized nature of the database makes it easy to access, manage, periodically update, and rapidly search the required data/information from a suitable computer system

Biological databases

Datasets relevant to biological sciences such as **molecular biology** and **bioinformatics** are called biological databases

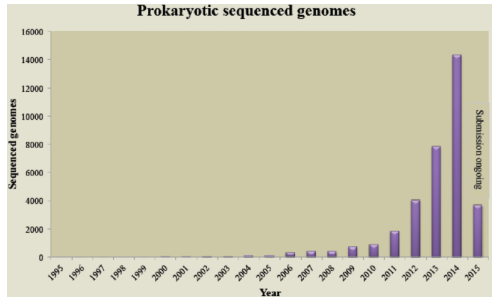


Why are Biological DBs important?

- ➊ Advances in molecular biology and technology
- ➋ Data analysis
- ➌ Data indexing and helps remove the data redundancy
- ➍ Central component of bioinformatics: data mining tools

1. Advances in molecular biology and sequencing technology

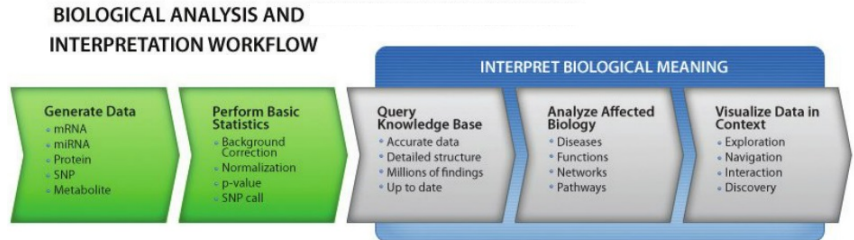
Due to rapidly advancing molecular biology, proteomics, and low-cost high-throughput genome sequencing technologies,



Huge amounts of biological information such as raw sequencing datasets, proteomes, etc. are being generated at a very rapid rate.

2. Biological analysis and drawing of meaningful conclusions

Biological analysis helps researchers transform statistically significant data into evidence-backed insights about affected biology by leveraging biological knowledge



Components of biological database

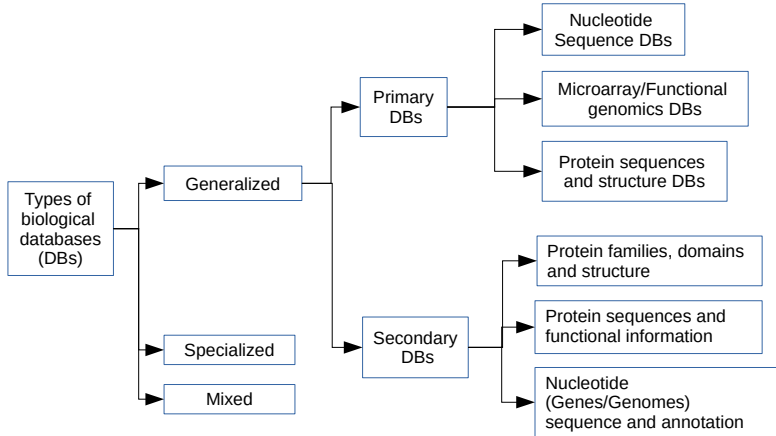
- ① **Entity** - An entity refers to the thing we want to store in a database. Eg. DNA sequences, Genes, Bibliographic references, etc.
- ② **Fields** - The properties of an entity are called fields. Eg. Gene name, gene sequence, mutation (if any), etc.
- ③ **Records** - A record typically refers to a combination of all the fields for a given entity. For eg. Record for gene BRCA1 in GenBank
- ④ **Identifier** - The unique name which identifies a record. In the case of a simple database, a single file contains multiple records. Among these

Example of a tabular database

Entity - Movies		Fields			
		ID	Title	Year	Director
Identifier	movie 1	Section 375	2019	Ajay Bahl	
	movie 2	Har Kisse Ke Hisse: Kaamyaab	2020	Hardik Mehta	Record
	movie 3	A Wednesday	2008	Neeraj Pandey	
	movie 4	Pink	2016	Aniruddha Roy Chowdhury	
	movie 5	Parched	2016	Leena Yadav	

Record

Types of biological databases



Primary databases

- Archival databases
- Contain experimentally derived datasets such as:
 - nucleotide sequences
 - protein sequences
 - structural information of macromolecules.
- This basic information can be accompanied by:
 - functional annotation,
 - bibliographies, and
 - links to other databases.
- Data is directly submitted by researchers.
- Once submitted, the data is assigned an **accession number**
 - It is permanent and becomes a part of the scientific record

Primary databases:

Nucleotide sequence databases

- The European Nucleotide Archive (ENA - EMBL),
- The National Center for Biotechnology Information GenBank (NCBI GenBank),
- The DNA Data Bank of Japan (DDBJ),



Primary databases:

Microarray/Functional genomics database

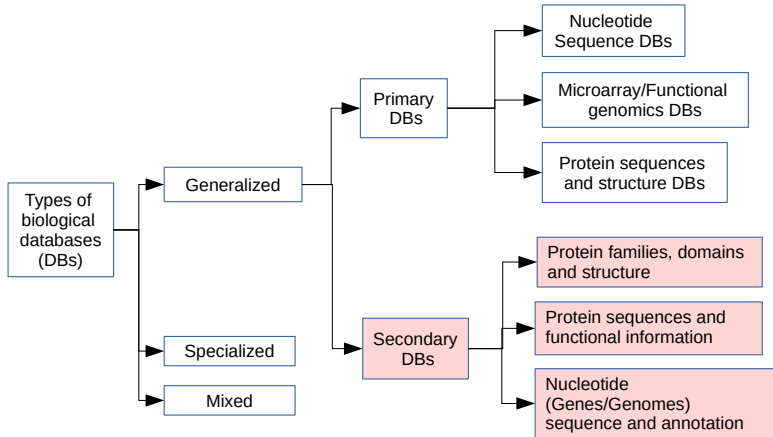
- Gene Expression Omnibus (GEO)
- Array Express Archives etc.
-

Primary databases:

Protein sequences and structure databases

- Protein sequences:
 - Swiss-Prot
 - Protein Information Resource (PIR)
- Protein structures:
 - Protein Databank (PDB).

Secondary databases



Secondary databases

Store the information derived from the analysis of primary datasets.

- Contain **highly curated information** derived from analysis of primary resources and literature, using:
 - Complex computational analysis
 - Manual analysis
- These databases often store information about:
 - conserved domain structure/sequences,
 - signal sequences, and
 - active site residues

Secondary databases:

Protein families, domains and structure databases

- InterPro
- PROSITE
- SCOP
- CATHand
- NCBI Conserved Domain Database (CDD)

Secondary databases:

Protein sequences and functional information databases

- UniProt Knowledgebase (UniProtKB)
- ...

Secondary databases:

Nucleotide (Genes/Genomes) sequence and annotation databases

- Ensembl,
- NCBI UniGene,
- The European Bioinformatics Institute (EBI) Genomes (EBI Genomes),
- ...

Specialized databases

These databases cater to the needs of specific research interests.

- Ribosomal Project Database (RDP),
- HIV sequence database,
- The Saccharomyces Genome Database (SGD),
- Mouse Genome Database (MGD), and
- Antibiotic Resistance Genes Database (ARDB),
- etc.

Important biological databases

The National Center for Biotechnology Information GenBank (NCBI GenBank)

- NCBI is a part of the National Library of Medicine (NLM) under the U.S. National Institute of Health (NIH).
- NCBI has been playing a leading role in bioinformatics by providing online access to:
 - biological datasets
 - biological information, and
 - computational resources/tools to the millions of researchers across the world.
- NCBI hosts approximately 40 online biological databases ,
 - GenBank is one among them.
 - Accessible online at <https://www.ncbi.nlm.nih.gov/>

GenBank at NCBI

NCBI

Resources

How To

NCBI

National Center for Biotechnology Information

All Databases

Search

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation


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

March 10 Webinar: Where to find data for your research organism! 01 Mar 2021

Do you work with data from organisms outside the traditional set of model

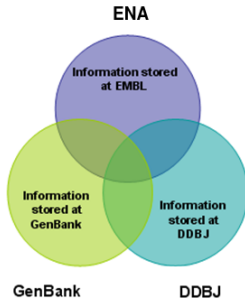
Important Update About How You Log Into your NCBI Accounts 26 Feb 2021

As mentioned in a previous blog post, we are transitioning to using third party login.

Biological Database

Luis Garreta

GenBank is part of the International Nucleotide Sequence Database Collaboration



- GenBank is a publically available collection of nucleotide sequences, their protein sequences along with annotations.
- These three organizations exchange data on daily basis.

Who can submit data to GenBank?

Being a primary database, the GenBank accepts:

- Sequence submission directly from **individual scientists** and **laboratories**.
- It also accepts batch submissions from **large-scale sequencing projects**.

Being a free public repository, **any researcher** can submit the sequences to GenBank without incurring any financial cost.

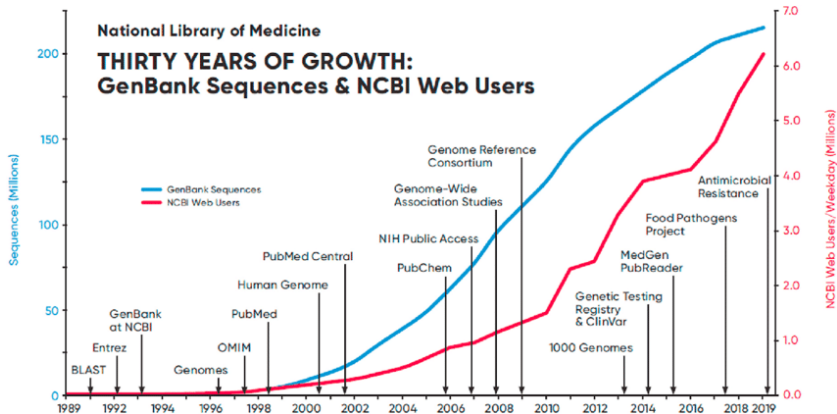
What Kind of Data Can be Submitted to GenBank?

Examples of submission types included in GenBank

- mRNA Sequences
- Prokaryotic Genes
- Eukaryotic Genes
- rRNA and/or ITS
- Viral Sequences
- Transposon or Insertion Sequences
- Microsatellite Sequences
- Pseudogenes
- Cloning Vectors
- Phylogenetic or Population Sets
- Non-coding RNAs

<https://www.ncbi.nlm.nih.gov/books/NBK566998/>

GenBank thirty years of growth

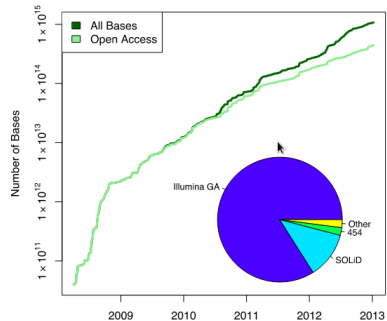


Sequence Read Archive (SRA) at NCBI

The SRA stores raw sequencing data from the next generation of sequencing platforms including:

- Roche 454 GS System®,
- Illumina Genome Analyzer®,
- Life Technologies AB SOLiD System® ,
- Helicos Biosciences Heliscope®,
- Complete Genomics®, and
- Pacific Biosciences SMRT®.

<https://www.ncbi.nlm.nih.gov/sra>



Some of the unique features of Genbank

- For the same gene or genome, multiple sequences of varying quality are available in GenBank.
- Essentially, anything is stored.
- A sequence can have several versions to represent the various modification done by the author,
- Unique identifier (Accession number) for each GenBank record
 - Accession number is permanent
 - Changes in the seq. version but not in the accession number.
 - For eg. `ACCESSION AF000001`, `VERSION AF000001.5`
- Each record is assigned to a specific division based on:
 - the Source taxonomy or
 - Sequencing strategy used to obtain the data.
 - There are 12 taxonomic divisions and 8 functional divisions