Biological Databases: An Introduction

Dr Safdar Ali
Department of Biological Sciences
Aliah University, Kolkata
Date: 04/04/2020

Biological Database

- Data: Information
- Database: Collection of Information
- What is information in Biology?
  - Sequence
  - Structure
  - Interactions
  - Reactions
- What are Biological Databases?

Types of Database

Basis of Classification: Source of information

Primary	Secondary
Herein the source of information is research, hence also called archival.	Herein, the source is primary databases, and hence also known as curated.
Examples GenBank; DDBJ, PDB	Examples RefSeq; GeneCards; OMIM

!Students are advised to explore these databases to understand the differences!

Types of Database

Basis of Classification: Status of Maintainer

- NCBI, EMBL or Govt entity
- Academic group/Scientist
- Commercial entity

Types of Database

Basis of Classification: Type of Data

- **Sequence**
  - Nucleotide (GenBank; DDBJ)
  - Protein (SwissProt)
- **Structure** (PDB; MSD)
- **Pathway** (KEGG, BioSilico)
- **Microarray** (MIAME)
- **Disease** (OMIM)
- **Literature** (PubMed; Scopus)

Types of Database

Basis of Classification: Accessibility of Data

- Publicly available
- Available with copyright
- Accessible but not downloadable
- Available for a charge/fee
- Selectively available (Restricted)

Types of Database

Basis of Classification: Organism

- Human
- Drosophila
- Yeast

Nucleotide Databases

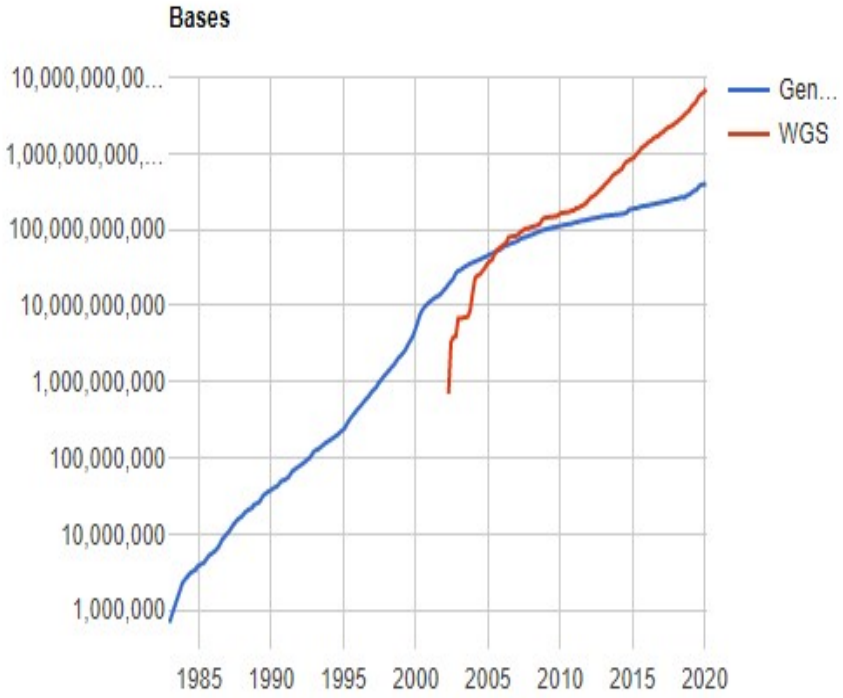
Three publicly accessible databases

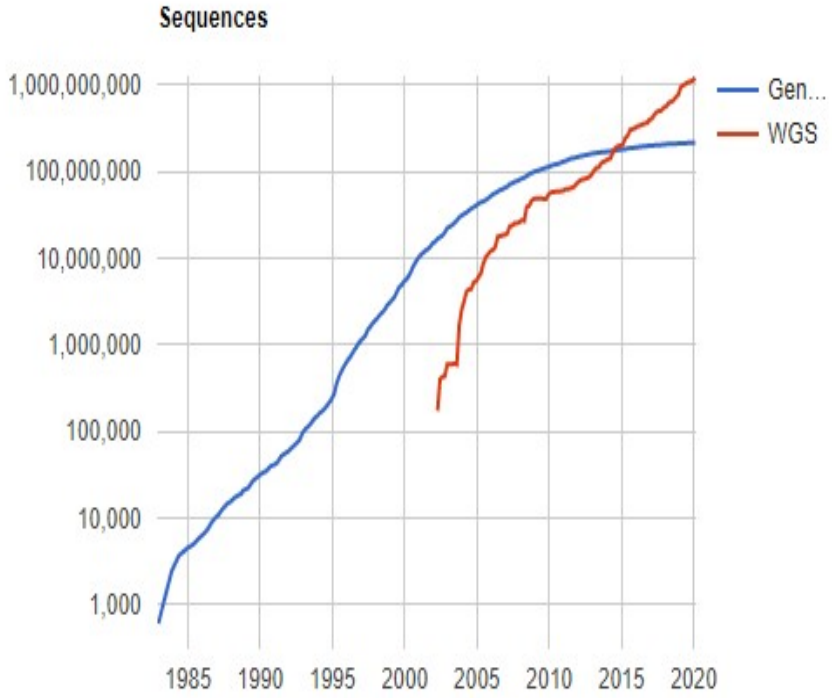
GenBank at the National Center for Biotechnology Information (NCBI) of the National Institutes of Health (NIH) in Bethesda (Benson et al., 2009)

DNA Database of Japan (DDBJ) at the National Institute ofGenetics in Mishima (Miyazaki et al., 2004)

European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database at the European Bioinformatics Institute (EBI) in Hinxton, England (Kulikova et al., 2007).

These three databases share their sequence data daily. They are coordinated by the **International Nucleotide Sequence Database Collaboration (INSDC).**





https://www.ncbi.nlm.nih.gov/genbank/statistics/

Accession Number

- Alphanumeric code specific for a particular data
- Each database has its own pattern for Accession no
- It helps in data classification and retrieval

GenBank Accession no Format

Nucleotide:	1 letter + 5 numerals
	2 letters + 6 numerals
	2 letters + 8 numerals
Protein:	3 letters + 5 numerals
	3 letters + 7 numerals
WGS:	4 letters + 2 numerals for WGS assembly version + 6 or more numerals
	6 letters + 2 numerals for WGS assembly version + 7 or more numerals
MGA:	5 letters + 7 numerals

https://www.ncbi.nlm.nih.gov/Sequin/acc.html

RefSeq Accession No

- The RefSeq projects are NCBI sequence annotation projects snd their Accession nos can be distinguished from GenBank accessions by an underbar in the third position.

Accession prefix	Molecule type	Comment
AC_	Genomic	Complete genomic molecule, usually alternate assembly
NC_	Genomic	Complete genomic molecule, usually reference assembly
NG_	Genomic	Incomplete genomic region
NT_	Genomic	Contig or scaffold, clone-based or WGS[a]

https://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.refseq_accession_numbers_and_mole/?report=objectonly

Thank You