Bioinformatics

Programs

What to know?

- **Must know at least the principles behind the programs.**
- **Don't just treat them as a black box.**
- **To understand the results, the user should have some idea of:**
  - **how they work**
  - **what assumptions they make**

Source: https://chagall.med.cornell.edu/BioinfoCourse/presentations2010/Lecture1_2010.pdf

Bioinformatics

Some common resources/programs

- NCBI
- EMBL
- Expasy
- BLAST – homology searching
- Clustal Omega – sequence alignment
- Modeller – protein structure prediction
- Swiss-Model – protein structure prediction
- PHYLIP – Phylogenetic analysis

Source: https://chagall.med.cornell.edu/BioinfoCourse/presentations2010/Lecture1_2010.pdf

National Center for Biotechnology Information (NCBI)

NCBI

What is NCBI?

- The National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM) at the U.S. National Institutes of Health, is a leader in the field of bioinformatics.
- It studies computational approaches to fundamental questions in biology and provides online delivery of biomedical information and bioinformatics tools.
- NCBI hosts approximately 40 online literature and molecular biology databases—including PubMed, PubMed Central, and GenBank—that serve millions of users around the world.
- The databases and resources are organized here into various concept areas: literature, genomes, variation, health, genes and gene expression, nucleotide, proteins, and small molecules and biological assays.
- Three additional categories encompass tools, infrastructure, and metadata.

https://www.ncbi.nlm.nih.gov/books/NBK143764/

The NCBI Handbook, 2nd edition

National Center for Biotechnology Information (NCBI)

NCBI

Brief History

- **The establishment of the National Center for Biotechnology Information (NCBI) in November of 1988 occurred primarily through the convergence of three independent but related actions. They were:**
  - **1984-86—Advocacy groups convened meetings on Capitol Hill to educate legislators and their staffs on the value of supporting genomic research.**
  - **1986—NLM's Long Range Plan was completed; it contained a recommendation that a new NLM Division be created to manage and process molecular biology information.**
  - **1987—The House Select Committee on Aging, Chaired by Senator Claude Pepper, introduced a Bill to establish the NCBI.**

More at: https://www.ncbi.nlm.nih.gov/

https://www.ncbi.nlm.nih.gov/books/NBK148949/

The NCBI Handbook, 2nd edition

Lecture – Biological databases

Biological databases

Types of data

- nucleotide and protein sequences
- protein structures
- genomes
- genetic expression
- bibliography
- metabolic pathway
- Human disease

Biological databases

Types of data

Nucleotide sequence databases

- GenBank (from NCBI)
- EMBL - European Nucleotide Archive
- DDBJ

Protein sequence databases

- Atlas
- PIR
- Swiss-Prot
- UniProt

Structure databases

- PDB
- NDB
- MMDB (from NCBI)
- SCOP
- CATH

Biological databases

Types of data

Nucleotide sequence databases

- GenBank (from NCBI)
- EMBL - European Nucleotide Archive
- DDBJ

Biological databases

Nucleotide sequence databases

GenBank

(from NCBI- National Center for Biotechnology Information)

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

Biological databases

Nucleotide sequence databases

GenBank (from NCBI)

- GenBank is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences.
- A GenBank release occurs every two months.

International Nucleotide Sequence Database Collaboration (INSDC)

- GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI.
- These three organizations exchange data on a daily basis.

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

Biological databases

Nucleotide sequence databases

GenBank (from NCBI)

- GenBank is a public collection of annotated sequences hosted by the NCBI.
- Among other kinds of sequences Genbank includes messenger RNAs, genomic DNAs and ribosomic RNA.

Some characteristics:

- It is a public repository, any one can send sequences to it.
- There are sequences of different qualities, anything submitted is stored.
- There could be multiple sequences for the same gene or for the same mRNA
- A sequence can have several versions that represent the modifications done by the authors.

https://bioinf.comav.upv.es/courses/biotech3/theory/databases.html

Biological databases

Nucleotide sequence databases

GenBank (from NCBI)

- **Due to the huge amount of sequences stored to ease the search the databases are split in different divisions.**

These divisions follow two criteria:

Species/Taxonomy

⇩

Among the taxonomical divisions we can find: primate, rodent, other mammalian, invertebrate an others.

Type of sequence

⇩

The other divisions are related to the kind of sequences like: EST, WGS, HTGS, and many others.

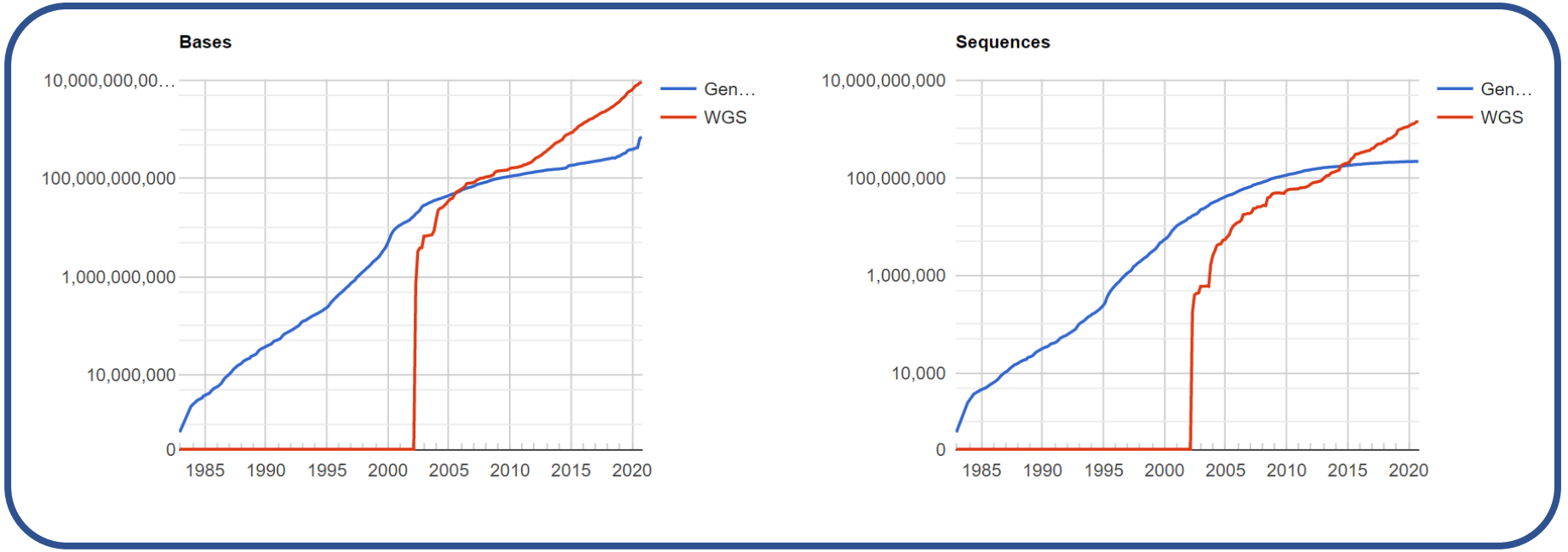
Note: If we are looking for reads coming from the Next Generation Sequencing Technologies they are stored in a special division called Sequence Read Archive (SRA).

https://bioinf.comav.upv.es/courses/biotech3/theory/databases.html

Bioinformatics

GenBank (from NCBI)

Growth of GenBank (Gen…) and whole genome sequencing (WGS)



- ✓ **GenBank, beginning with Release 3 in 1982.**
- ✓ **From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months.**

Source: https://www.ncbi.nlm.nih.gov/genbank/statistics/#:~:text=Notes%20on%20GenBank%20statistics&text=From%201982%20to%20the%20present,doubled%20approximately%20every%2018%20months.

Biological databases

Nucleotide sequence databases

GenBank (from NCBI)

GenBank Data Usage

- **The GenBank database is designed to provide and encourage access within the scientific community to the most up-to-date and comprehensive DNA sequence information.**
- **Therefore, NCBI places no restrictions on the use or distribution of the GenBank data.**
- **However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted.**
- **NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.**

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

Biological databases

Nucleotide sequence databases

GenBank (from NCBI)

Access to GenBank

Several ways to search and retrieve data from GenBank.

- **Search GenBank for sequence identifiers and annotations with Entrez Nucleotide.**
- **Search and align GenBank sequences to a query sequence using BLAST (Basic Local Alignment Search Tool). See BLAST info for more information about the numerous BLAST databases.**
- **Search, link, and download sequences programatically using NCBI e-utilities.**
- **The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1 and ftp://ftp.ncbi.nlm.nih.gov/genbank.**

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

Biological databases

Nucleotide sequence databases

EMBL – European Molecular Biology Laboratory

European Nucleotide Archive

https://www.ebi.ac.uk/ena/browser/home

Biological databases

Nucleotide sequence databases

EMBL -European Nucleotide Archive

- ✓ Maintained at the European Bioinformatics Institute (EBI)
- ✓ Contain data of DNA and RNA sequences

- The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation.

International Nucleotide Sequence Database Collaboration (INSDC)

- GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI.
- These three organizations exchange data on a daily basis.

https://www.ebi.ac.uk/ena/browser/home

More at: https://www.ebi.ac.uk/ena/browser/about

Biological databases

Nucleotide sequence databases

DDBJ – DNA Data Bank of Japan

Biological databases

Nucleotide sequence databases

DDBJ – DNA Data Bank of Japan

- **DDBJ Center collects nucleotide sequence data as a member of INSDC and provides freely available nucleotide sequence data and supercomputer system, to support research activities in life science.**

International Nucleotide Sequence Database Collaboration (INSDC)

- **GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI.**
- **These three organizations exchange data on a daily basis.**

https://www.ddbj.nig.ac.jp/aboutus-e.html#history

Biological databases

Nucleotide sequence databases

DDBJ – DNA Data Bank of Japan

- **The principal purpose of DDBJ operations is to improve the quality of INSD, as public domains.**
- **When researchers make their data open to the public through INSD and commonly shared in world wide, DDBJ Center make efforts to describe information on the data as rich as possible, according to the unified rules of INSD, by using DDBJ.**
- **Currently, DDBJ Center is in operation at Research Organization of Information and System National Institute of Genetics (NIG) in Mishima, Japan with endorsement of MEXT; Japanese Ministry of Education, Culture, Sports, Science and Technology.**

https://www.ddbj.nig.ac.jp/aboutus-e.html#history

Biological databases

Types of data

- Atlas
- PIR
- Swiss-Protein
- UniProt

Protein sequence databases

Biological databases

Protein sequence databases

PIR – Protein Information Resource

Biological databases

Protein sequence databases

PIR – Protein Information Resource

The Protein Information Resource (PIR) is an integrated public bioinformatics resource to support genomic, proteomic and systems biology research and scientific studies.

History

- PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers in the identification and interpretation of protein sequence information.
- Prior to that, the NBRF compiled the first comprehensive collection of macromolecular sequences in the "Atlas" of Protein Sequence and Structure, published from 1965-1978 under the editorship of Margaret O. Dayhoff.
- Dr. Dayhoff and her research group pioneered in the development of computer methods for the comparison of protein sequences, for the detection of distantly related sequences and duplications within sequences, and for the inference of evolutionary histories from alignments of protein sequences.

https://proteininformationresource.org/pirwww/about/

Biological databases

Protein sequence databases

PIR – Protein Information Resource

History 

- For over four decades, beginning with the Atlas of Protein Sequence and Structure, PIR has provided protein databases and analysis tools freely accessible to the scientific community including the Protein Sequence Database (PSD).
- In 2002 PIR, along with its international partners, EBI (European Bioinformatics Institute) and SIB (Swiss Institute of Bioinformatics), were awarded a grant from NIH to create UniProt, a single worldwide database of protein sequence and function, by unifying the PIR-PSD, Swiss-Prot, and TrEMBL databases.

https://proteininformationresource.org/pirwww/about/

https://proteininformationresource.org/

Biological databases

Protein sequence databases

UniProt – Universal Protein Resource

Biological databases

Protein sequence databases

UniProt

- UniProt is a collaboration between:
  - the European Bioinformatics Institute (EMBL-EBI),
  - the SIB Swiss Institute of Bioinformatics
  - and the Protein Information Resource (PIR).
- Across the three institutes more than 100 people are involved through different tasks such as database curation, software development and support.

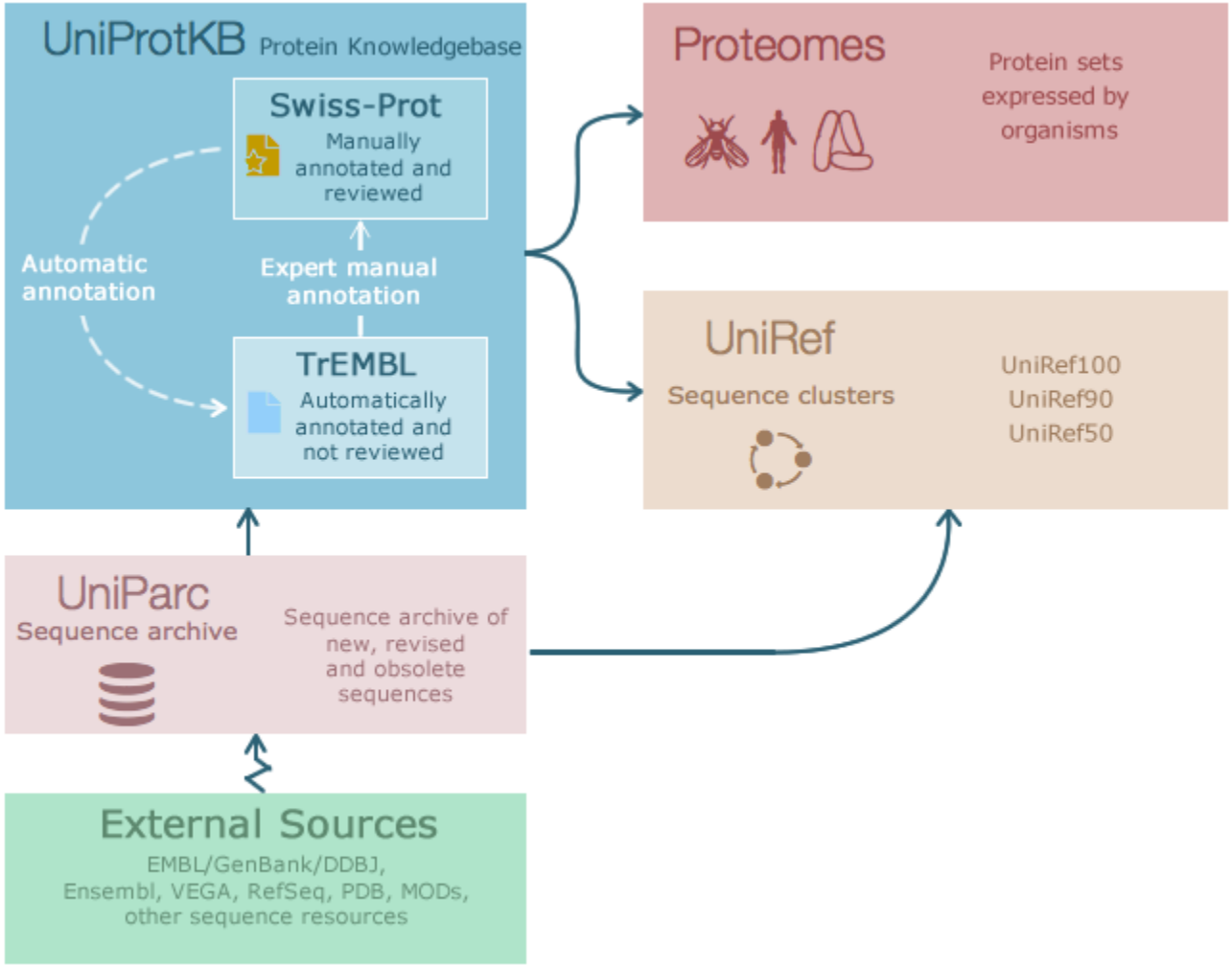
https://www.uniprot.org/

Biological databases

Protein sequence databases

UniProt

- The Universal Protein Resource (UniProt) is a comprehensive resource for protein sequence and annotation data.
- The UniProt databases are:
  - the UniProt Knowledgebase (UniProtKB),
  - the UniProt Reference Clusters (UniRef),
  - and the UniProt Archive (UniParc).
- The UniProt consortium and host institutions EMBL-EBI, SIB and PIR are committed to the long-term preservation of the UniProt databases.



Soure: https://www.uniprot.org/help/about

https://www.uniprot.org/

Biological databases

Protein sequence databases

UniProt

- EMBL-EBI and SIB together used to produce Swiss-Prot and TrEMBL,
- while PIR produced the Protein Sequence Database (PIR-PSD).
- These two data sets coexisted with different protein sequence coverage and annotation priorities.
- TrEMBL (Translated EMBL Nucleotide Sequence Data Library) was originally created because sequence data was being generated at a pace that exceeded Swiss-Prot's ability to keep up.
- Meanwhile, PIR maintained the PIR-PSD and related databases, including iProClass, a database of protein sequences and curated families.
- In 2002, the three institutes decided to pool their resources and expertise and formed the UniProt consortium.

https://www.uniprot.org/

Biological databases

Protein sequence databases

- **UniProt aims to store sequence and functional information for the proteins.**

UniProt

UniProt includes information divided in two sections:

Swiss-Prot

Swiss-Prot is reviewed manually by humans that add information by reviewing the literature.

TrEMBL

TrEMBL is automatically annotated

Swiss-Prot has information of a higher quality, but it has less sequences than TrEMBL.

https://www.uniprot.org/ https://bioinf.comav.upv.es/courses/biotech3/theory/databases.html

Biological databases

Protein sequence databases

UniProt – Uniref

Uniref

- **UniProt also hosts Uniref.**
- **This database aims to store one representative sequence for each protein without taking into account the species of origin.**
- **It clusters all the similar proteins and picks one for every cluster as a representative.**
- **There are clusters created at 100%, 90% and 50% identities.**

https://www.uniprot.org/help/uniref

https://bioinf.comav.upv.es/courses/biotech3/theory/databases.html