IC251 – Basics of Bioinformatics (4 Credits)

Bioinformatics

Broad applications

Medicine

- ✓ Molecular medicine
- ✓ Personalised medicine
- ✓ Gene therapy
- ✓ Drug development
- ✓ Antibiotic resistance

Agriculture

- ✓ Crop improvement
- ✓ Improve nutritional quality
- ✓ Development of drought resistant varieties
- ✓ Insect resistance

Phylogenetic

- ✓ Evolutionary studies

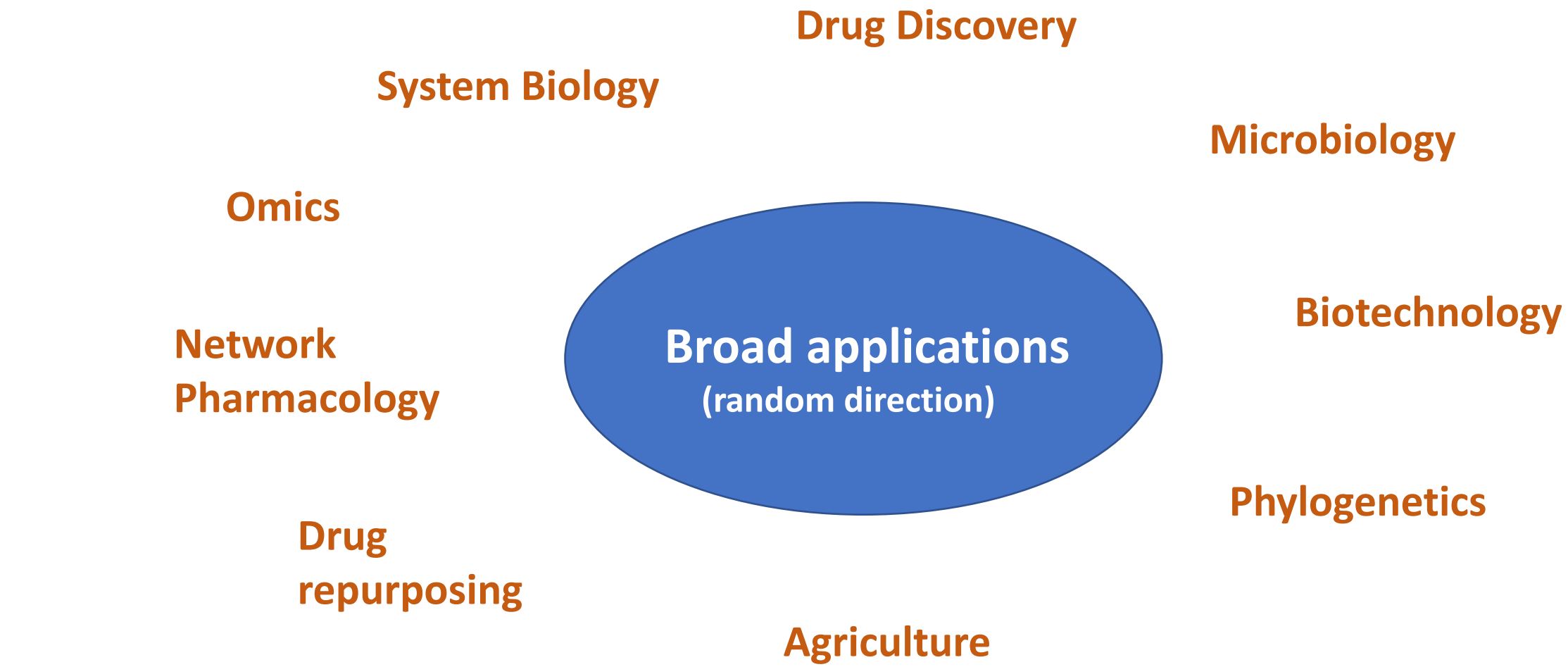
Microbial genome applications

- ✓ Waste cleanup
- ✓ Biotechnology
- ✓ Microbial genome sequencing

Source:

Essential Bioinformatics by Jin Xiong

Bioinformatics



Source:

Essential Bioinformatics by Jin Xiong

Bioinformatics

Applications
(another direction)

- Essential for basic genomic and molecular biology research.
- Major impact on many areas of biotechnology and biomedical sciences.

- ✓ Bioinformatics plays a vital role in the areas of structural genomics, functional genomics, and nutritional genomics.
- ✓ It covers emerging scientific research and the exploration of proteomes from the overall level of intracellular protein composition (protein profiles), protein structure, protein-protein interaction, and unique activity patterns (e.g. post-translational modifications).
- ✓ Bioinformatics is used for transcriptome analysis where mRNA expression levels can be determined.
- ✓ Bioinformatics is used to identify and structurally modify a natural product, to design a compound with the desired properties and to assess its therapeutic effects, theoretically.
- ✓ Cheminformatics analysis includes analyses such as similarity searching, clustering, QSAR modeling, virtual screening, etc.
- ✓ Bioinformatics plays an increasingly important role in almost all aspects of drug discovery and drug development.
- ✓ Bioinformatics tools are very effective in prediction, analysis and interpretation of clinical and preclinical findings.

Source:
Essential Bioinformatics by Jin Xiong

https://microbenotes.com/bioinformatics-introduction-and-applications/

Bioinformatics

Applications

- Bioinformatics has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences.
- It has applications, for example, in knowledge-based drug design, forensic DNA analysis, and agricultural biotechnology.
- Computational studies of protein–drug interactions provide a rational basis for the rapid identification of novel leads for synthetic drugs.
- Knowledge of the three-dimensional structures of proteins allows molecules to be designed that are capable of binding to the receptor site of a target protein with great affinity and specificity.
- This informatics-based approach significantly reduces the time and cost necessary to develop drugs with higher potency, fewer side effects, and less toxicity than using the traditional trial-and-error approach.

Source:

Essential Bioinformatics by Jin Xiong

Bioinformatics

Applications

- It is worth mentioning that genomics and bioinformatics are now poised to revolutionize our healthcare system by developing personalized and customized medicine.
- The high speed genomic sequencing coupled with sophisticated informatics technology will allow a doctor in a clinic to quickly sequence a patient's genome and easily detect potential harmful mutations and to engage in early diagnosis and effective treatment of diseases.
- Bioinformatics tools are being used in agriculture. Plant genome databases and gene expression profile analyses have played an important role in the development of new crop varieties that have higher productivity and more resistance to disease.

Source:

Essential Bioinformatics by Jin Xiong

Bioinformatics

Types of data available

Huge data available 

- DNA/RNA sequence
- single-nucleotide polymorphisms (SNPs)
- protein sequence
- protein structure
- protein function
- organism-specific databases
- genomes
- gene expression
- biomolecular interactions
- molecular pathways
- scientific literature
- disease information

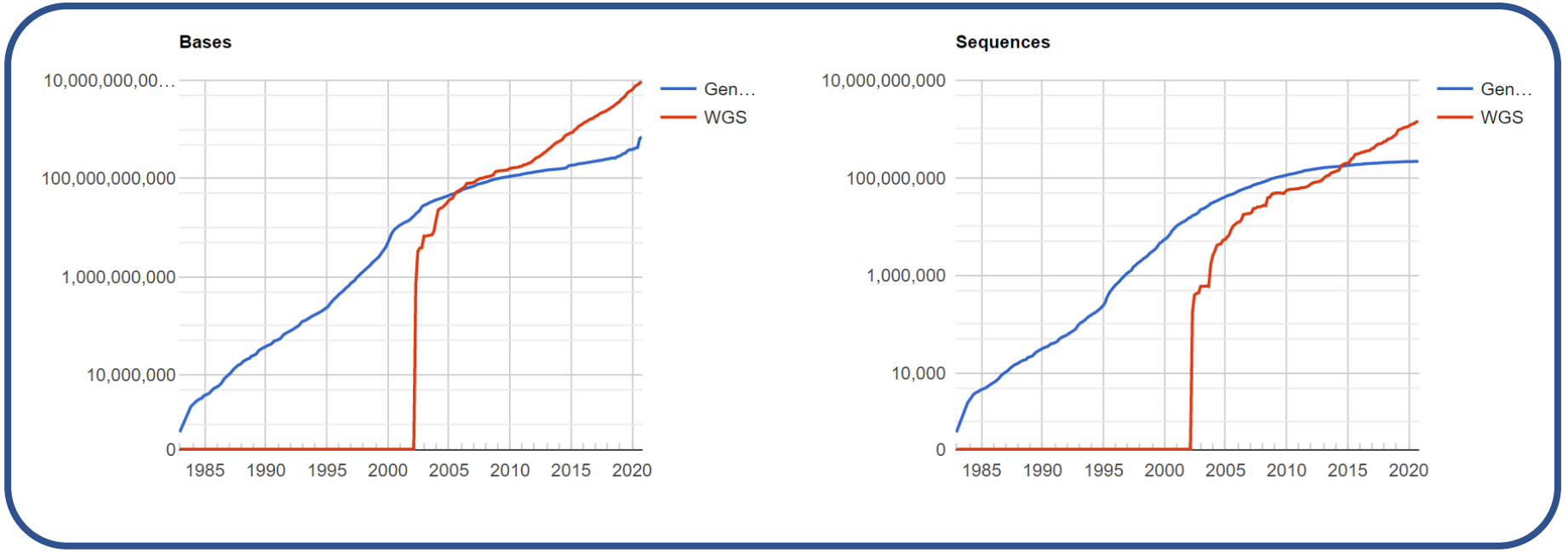
Source:

Essential Bioinformatics by Jin Xiong https://chagall.med.cornell.edu/BioinfoCourse/presentations2010/Lecture1_2010.pdf

Bioinformatics

Types of data available

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



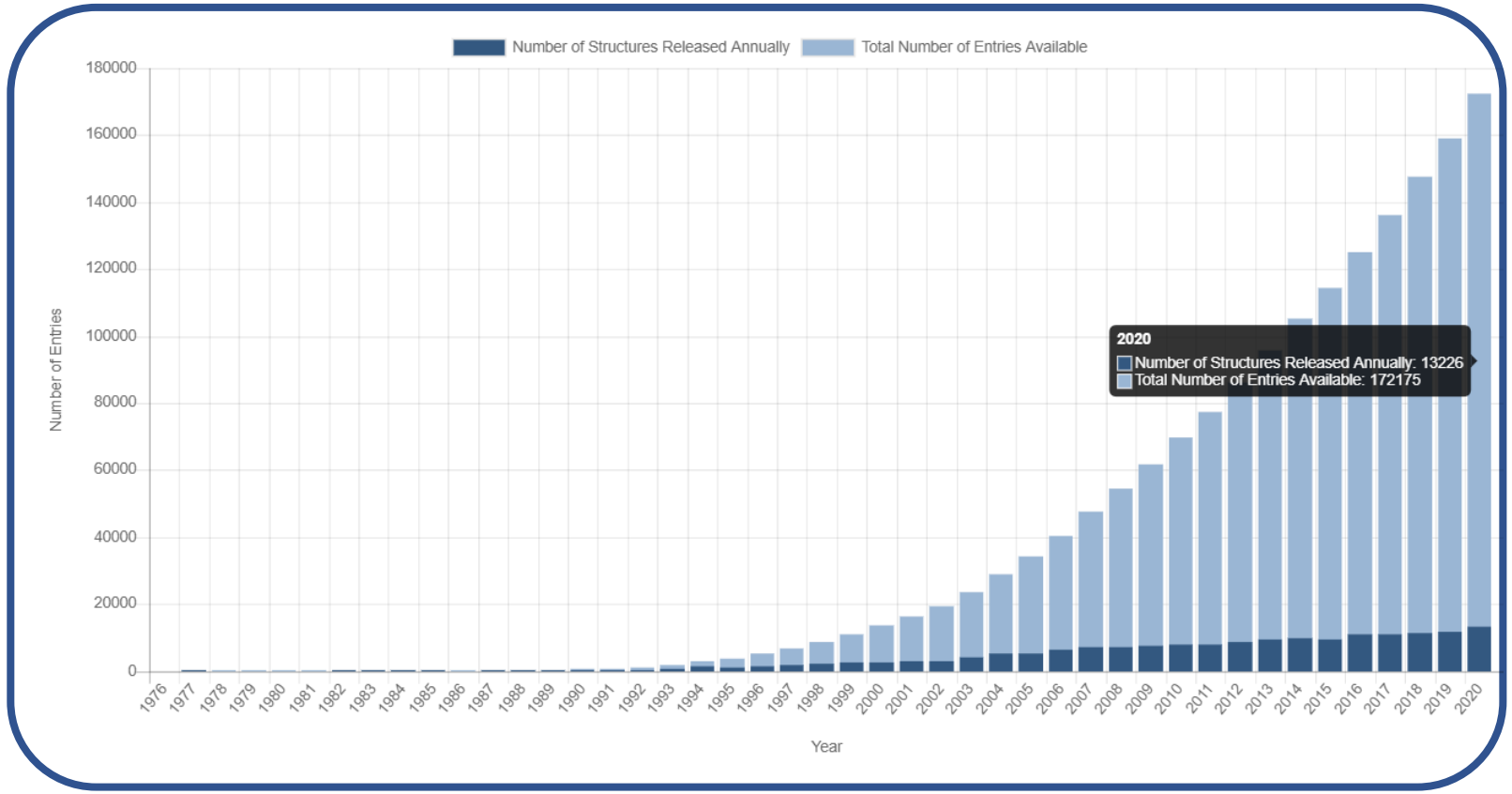
- ✓ **GenBank, beginning with Release 3 in 1982.**
- ✓ **From 1982 to the present, the number of bases in GenBank has doubled approximately every 18 months.**

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.

Bioinformatics

Types of data available

Growth of
Protein Data Bank (PDB)



Source: https://www.rcsb.org/stats

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.

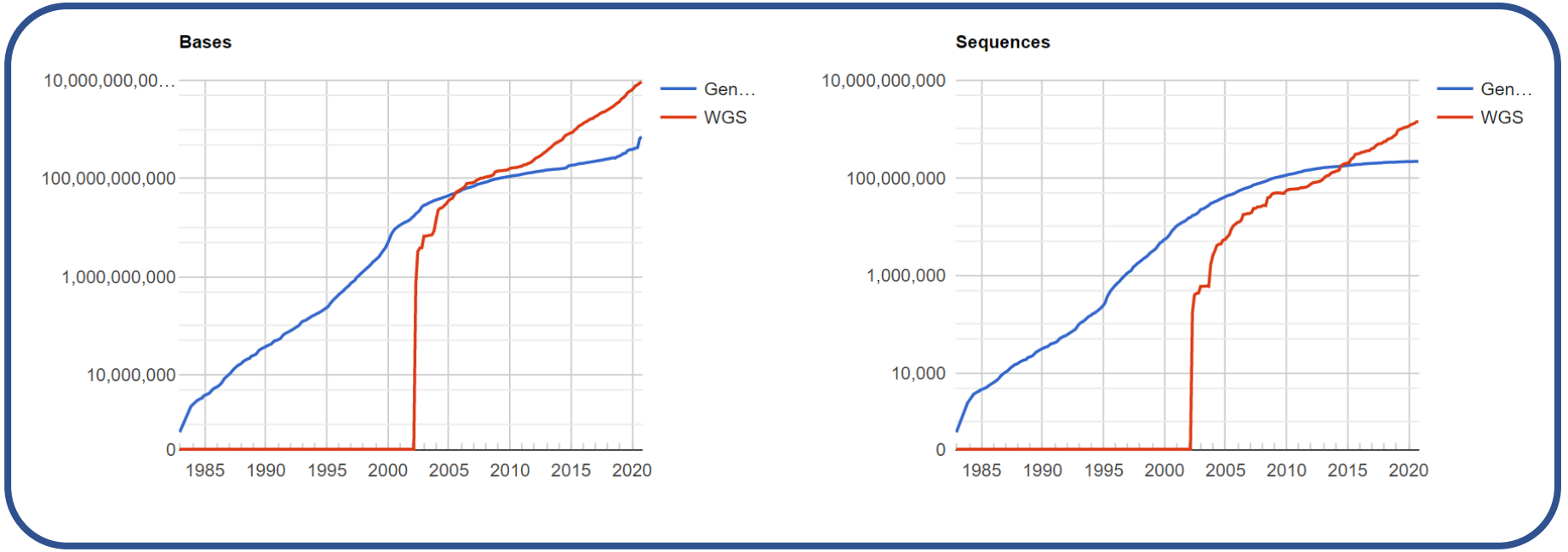
Note: If we are looking for reads coming from the Next Generation Sequencing Technologies they are stored in a special division called Sequence Read Archive (SRA).

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-Prot
- 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.

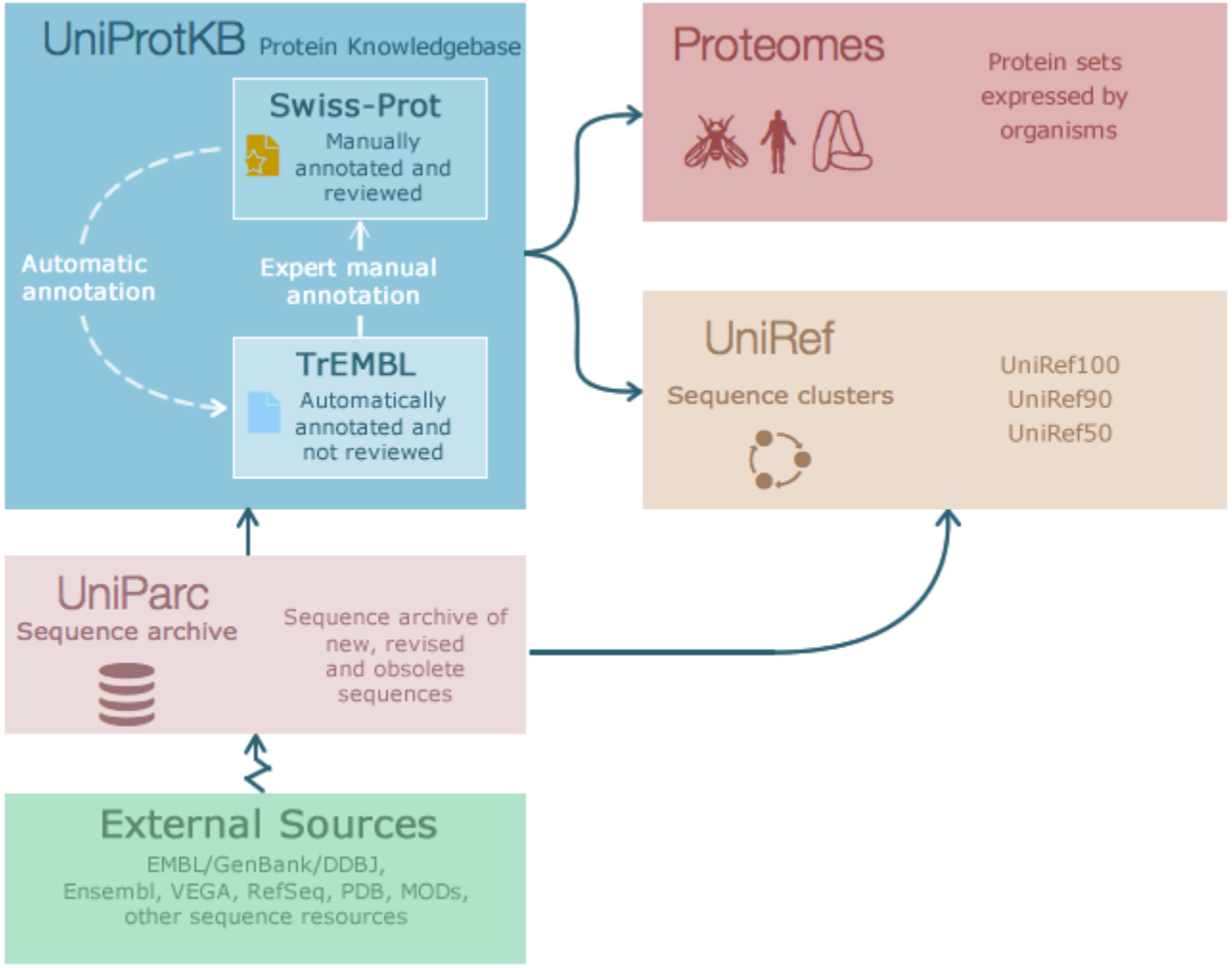
https://www.uniprot.org/

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

Biological databases

Protein sequence databases

Swiss-Prot

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

Biological databases

Protein sequence databases

Swiss-Prot

Swiss-Prot is an annotated protein sequence database, which was created at the Department of Medical Biochemistry of the University of Geneva and has been a collaborative effort of the Department and the EMBL, since 1987.

- Swiss-Prot (created in 1986) is a high quality manually annotated and non-redundant protein sequence database, which brings together experimental results, computed features and scientific conclusions.
- UniProtKB/Swiss-Prot is now the reviewed section of the UniProt Knowledgebase.
- The TrEMBL section of UniProtKB was introduced in 1996 in response to the increased dataflow resulting from genome projects.
- It was already recognized at that time that the traditional time- and labour-intensive manual curation process which is the hallmark of Swiss-Prot could not be broadened to encompass all available protein sequences.
- UniProtKB/TrEMBL contains high quality computationally analyzed records that are enriched with automatic annotation and classification.
- These UniProtKB/TrEMBL unreviewed entries are kept separated from the UniProtKB/Swiss-Prot manually reviewed entries so that the high quality data of the latter is not diluted in any way.
- Automatic processing of the data enables the records to be made available to the public quickly.

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

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102476/#:~:text=SWISS%2DPROT%20(1)%20is,(EMBL)%2C%20since%201987.

Lecture – Structure databases

Biological databases

Types of data

Structure databases

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

Biological databases

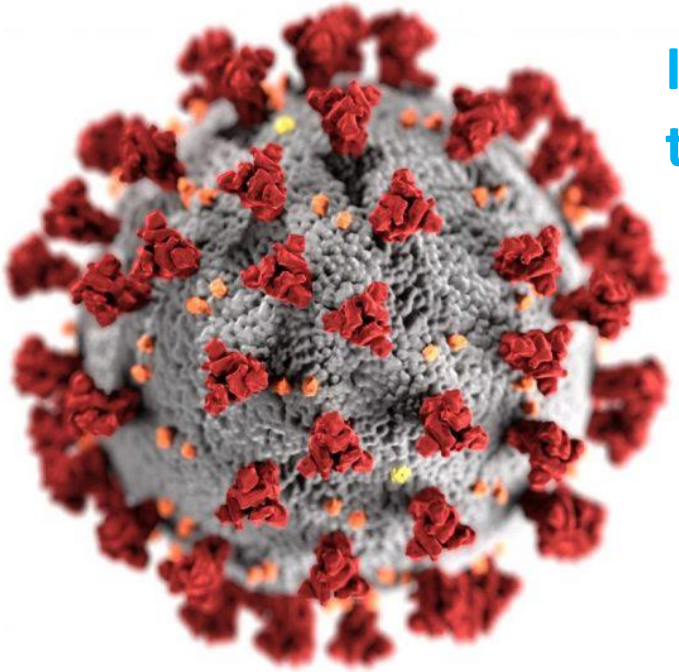
Structure databases

Protein Data Bank (PDB)

Biological databases

Structure databases

Protein Data Bank (PDB)



Identify the image?

Source:
https://www.clinicalomics.com/topics/patient-care/coronavirus/new-antibody-test-for-covid-19-targets-unique-region-of-spike-protein/

Spike protein

PDB code: 6XR8

https://www.rcsb.org/

Biological databases

Structure databases

Protein Data Bank (PDB)

What is it?

PDB is the single worldwide repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids.

These are the molecules of life that are found in all organisms including bacteria, yeast, plants, flies, other animals, and humans.

Understanding the shape of a molecule deduce a structure's role in human health and disease, and in drug development.

The structures in the archive range from tiny proteins and bits of DNA to complex molecular machines like the ribosome.

The PDB archive is available at no cost to users.

The PDB archive is updated weekly.

https://www.rcsb.org/

Biological databases

Structure databases

Protein Data Bank (PDB)

PDB – History

The PDB was established in 1971 at Brookhaven National Laboratory under the leadership of Walter Hamilton and originally contained 7 structures.

After Hamilton's untimely death, Tom Koetzle began to lead the PDB in 1973, and then Joel Sussman in 1994.

Led by Helen M. Berman, the Research Collaboratory for Structural Bioinformatics (RCSB) became responsible for the management of PDB in 1998.

In 2003, the wwPDB was formed to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community.

wwPDB consists of organizations that act as deposition, data processing and distribution centers for PDB data.

Stephen K. Burley became Director in 2014.

https://www.rcsb.org/

Biological databases

Structure databases

Protein Data Bank (PDB)

What is it?

The RCSB PDB supports a website where visitors can perform simple and complex queries on the data, analyze, and visualize the results.

The RCSB PDB has an international community of users, including biologists (in fields such as structural biology, biochemistry, genetics, pharmacology); other scientists (in fields such as bioinformatics, software developers for data analysis and visualization); students and educators (all levels); media writers, illustrators, textbook authors; and the general public.

The website (rcsb.org) is accessed by >1 million unique visitors per year.

RCSB PDB services have broad impact across research and education.

https://www.rcsb.org/

Biological databases

Structure databases

Protein Data Bank (PDB)

What is it?

wwPDB

The RCSB PDB is a member of the wwPDB, a collaborative effort with PDBe (UK), PDBj (Japan), and Biological Magnetic Resonance Data Bank (BMRB, USA) to ensure the PDB archive is global and uniform.

As the wwPDB archive keeper, the RCSB PDB updates the PDB archive at ftp://ftp.wwpdb.org weekly.

The structures included in each release are highlighted on the RCSB PDB home page and clearly defined on the FTP site.

These sites are maintained 24 hours a day, seven days a week.

A failover system automatically redirects internet traffic to a mirror site, if needed.

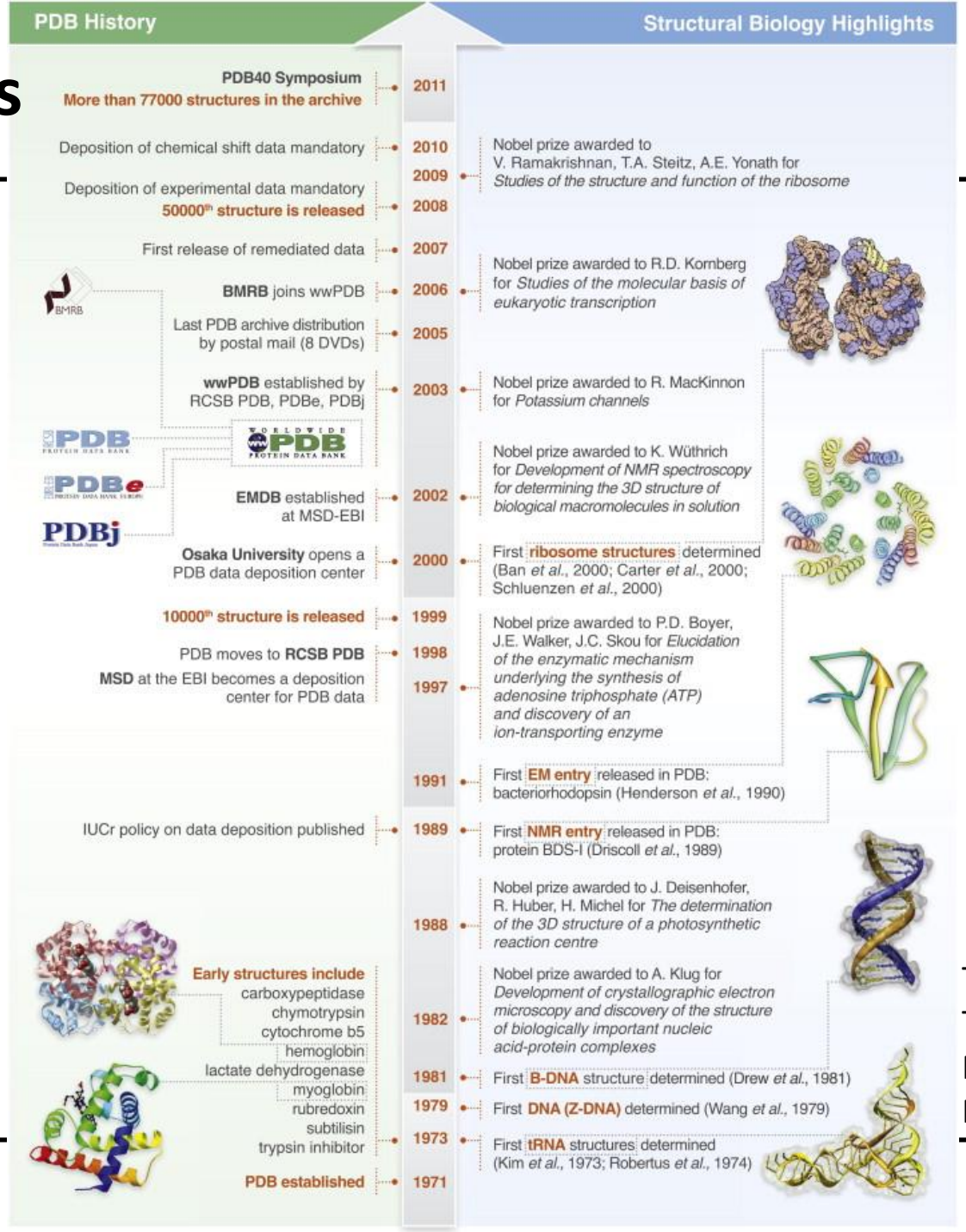
https://www.rcsb.org/

Biological databases

Structure databases

Protein Data Bank (PDB)

RCSB PDB History overview



Taken from Berman, 2012, The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future.

https://www.rcsb.org/

Biological databases

Structure databases

Protein Data Bank (PDB)

Experience PDB

COVID-19/SARS-CoV-2 Resources

PDB ID: 4 character alphanumeric code

https://www.rcsb.org/news?year=2020&article=5e74d55d2d410731e9944f52&feature=true

QUERY: PDB ID(s) **IN** (6XLU, 6XM0, 6XM3, 6XM4, 7CAH, 7JN2, 6ZME, 6ZRT, 6ZRU, 6WC1, 7JIR, 7JIT, 7JIV, 7JIW, 6ZWV, 6XEZ, 6XM5, 6ZDG, 6ZOW, 6ZP5, 6ZP7, 6XQB, 6ZSL, 7C7P, 7JFQ, 6ZBP, 6ZHD, 6ZOK, 6ZLW, 6ZM7, 6ZN5, 6ZON, 6ZP4, 6XEY, 6XR8, 6XRA, 6XS6, 6ZP0, 6ZP1, 6ZP2, 6ZOX, 6ZOY, 6ZOZ, 6ZOJ, 6XQS, 6XQT, 6XQU, 6XKL, 6XOA, 6XC2, 6XC3, 6XC4, 6XC7, 6XHM, 6XKM, 6XKF, 6XKH, 6XMK, 6ZFO, 6XCM, 6XCN, 6XE1, 6Z97, 6ZDH, 6ZGE, 6ZGG, 6ZGH, 6ZGI, 7C2L, 6XHU, 6XIP, 6ZCO, 6XFN, 6XG2, 7C8U, 6XG3, 6ZCT, 6ZCZ, 6ZER, 7C8W, 7C8V, 7CAN, 6XDG, 6X2G, 6XB0, 6XB1, 6XB2, 6XA4, 6XAA, 6XA9, 6XCH, 6XBG, 6XBH, 6XBI, 6XDC, 6XDH, 6Z2E, 6Z4U, 6M5I, 7BQ7, 7C8R, 7C8T, 6X6P, 7BYR, 5RHB, 5RHC, 5RHD, 5RHE, 5RHF, 6X4I, 6YZ5, 6YZ7, 6Z2M, 6Z43, 6M1V, 7BWJ, 7BZF, 7C2K, 6WPS, 6WPT, 6X29, 6X2A, 6X2B, 6X2C, 6WZO, 6WZQ, 6X1B, 7BW4, 7C2I, 7C2J, 7C01, 6WZU, 5RGT, 5RGU, 5RGV, 5RGW, 5RGX, 5RGY, 5RGZ, 5RH0, 5RH1, 5RH2, 5RH3, 5RH4, 5RH5, 5RH6, 5RH7, 5RH8, 5RH9, 5RHA, 6Y2G, 6Y2F, 6Y2E, 6W02, 6W01, 6Y84, 6W41, 6W4H, 6VSB, 6W4B, 6W61, 6W63, 6W75, 6VW1, 6W6Y, 6VXS, 6VWW, 6VYO, 6VYB, 6VXX, 6YB7, 5R84, 5R83, 5R7Y, 5R80, 5R82, 5R81, 5R7Z, 5REA, 5REC, 5REB, 5REE, 5RED, 5REG, 5REF, 5RE9, 5RE8, 5RE5, 5RE4, 5RE7, 5RE6, 5RFB, 5RFA, 5RFD, 5RFC, 5RFF, 5RFE, 5RFH, 5RFG, 5REY, 5REX, 5RF9, 5REZ, 5RF2, 5REP, 5RF1, 5RES, 5RF4, 5RER, 5RF3, 5REU, 5RF6, 5RET, 5RF5, 5REW, 5RF8, 5REV, 5RF7, 5REI, 5REH, 5REK, 5REJ, 5REM, 5REL, 5REO, 5RF0, 5REN, 5RFZ, 5RFY, 5RFR, 5RFQ, 5RFT, 5RFS, 5RFV, 5RFU, 5RFX, 5RFW, 5RFJ, 5RFI, 5RFL, 5RFK, 5RFN, 5RFM, 5RFP, 5RFO, 5RG0, 6M03, 6M17, 6M0J, 6M3M, 6LU7, 6LVN, 6LXT, 6LZG, 6W9C, 5R8T, 6M71, 6W9Q, 6YI3, 7BTF, 6WEN, 6WCF, 5RG1, 5RG2, 5RG3, 5RGG, 5RGH, 5RGI, 5RGJ, 5RGK, 5RGL, 5RGM, 5RGN, 5RGO, 5RGP, 5RGQ, 5RGR, 5RGS, 6M2N, 6M2Q, 6YLA, 6WIQ, 6WJI, 6WJT, 7BQY, 7BV2, 7BV1, 6LZE, 6M0K, 7BUY, 6W37, 6WEY, 6WKP, 6WKQ, 6WLC, 6YHU, 6YM0, 6YNQ, 6YOR, 6WKS, 6WNP, 6WOJ, 6WQF, 6WQ3, 6WQD, 6WRH, 6YT8, 6YWK, 6YWL, 6YWM, 6WRZ, 6WTC, 6WVN, 6YVA, 6YYT, 6YZ1, 7BRO, 7BRP, 7BRR, 7BZ5, 6WTJ, 6WTK, 6WTM, 6WTT, 6YVF, 6YZ6, 6WUU, 6WX4, 6WXC, 6WXD, 6YUN, 7C22)

Distinct conformational states of SARS-CoV-2 spike protein

https://www.rcsb.org/structure/6XRA

Biological databases

Structure databases

Nucleic Acids Database (NDB)

Biological databases

Structure databases

Nucleic Acids Database (NDB)

- A Portal for Three-dimensional Structural Information about Nucleic Acids.
- As of 9-Dec-2020 number of released structures: 11094

- The NDB contains information about experimentally-determined nucleic acids and complex assemblies.
- The goal of the NDB is to archive and distribute structural information about nucleic acids.
- The NDB was founded in 1992 by Helen M. Berman, Rutgers University, Wilma K. Olson, Rutgers University, and David Beveridge, Wesleyan University.
- The NDB Project is funded by the National Institutes of Health and has been funded by National Science Foundation and the Department of Energy in the past.

http://ndbserver.rutgers.edu/

Biological databases

Structure databases

Molecular Modeling Database (MMDB)

Biological databases

Structure databases

Molecular Modeling Database (MMDB)

Maintained by NCBI

- **Contains experimentally resolved structures of proteins, RNA, and DNA, derived from the Protein Data Bank (PDB),**
- **with value-added features such as explicit chemical graphs, computationally identified 3D domains (compact substructures) that are used to identify similar 3D structures, as well as links to literature, similar sequences, information about chemicals bound to the structures, and more.**
- **These connections make it possible, for example, to find 3D structures for homologs of a protein sequence of interest, then interactively view the sequence-structure relationships, active sites, bound chemicals, journal articles, and more.**

More at: https://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml

Lecture – Sequence file formats

Sequences

Sequence formats

Why so many formats?

- ✓ There are at least a couple of dozen sequence formats in existence at the moment. Some are much more common than others.
- ✓ Formats were designed so as to be able to hold the sequence data and other information about the sequence.
- ✓ Nearly every sequence analysis package written since programs were first used to read and write sequences has invented its own format.
- ✓ Nearly every collection of sequences that dares call itself a database has stored its data in its own format.

Source: http://emboss.sourceforge.net/docs/themes/SequenceFormats.html#seq

Sequences

Sequence formats

Text files

- ✓ There are different formats to store sequences in a text file. Text files should only include Plain text.
- ✓ Graphics or any other binary information are not allowed in text files.

1. Sequences in plain files

- ✓ We store the sequence in a text file by just writing the sequence. This files include only IUPAC characters.

Example – plain format

Microsoft WORD format is not a sequence format.

```
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTAC
GGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCTCCTGACTTTCCTCGCTTGGT
AGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCC
ATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAA
```

This kind of file is seldom used because it lacks any metadata to identify the sequence.

Note: A file in plain sequence format may only contain one sequence, while most other formats accept several sequences in one file.

Source: https://bioinf.comav.upv.es/courses/biotech3/theory/databases.html https://www.genomatix.de/online_help/help/sequence_formats.html

Sequences

Sequence formats

2. FASTA format
(most common format)

- ✓ **The FASTA file includes a name for the sequence and, optionally, some description.**
- ✓ **The sequence should be preceded by a line that starts with the symbol >. The name will be written after that symbol.**
- ✓ **If required, several sequences can be included in the same file.**

Example

```
>sequence1_name description
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCCCCTGGAGGGTAC
GGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGCCTCCTGACTTTCCTCGCTTGGT
AGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGGAAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCC
ATCCGCGCGCCGGGACAGAATGCCCTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAA
>sequence2_name description
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGA
```

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

Sequences

Sequence formats

3. **FASTQ format**

- ✓ A sequence file in FASTQ format can contain several sequences.
- ✓ FASTQ is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores.
- ✓ It is mainly used for storing the output of high-throughput sequencing instruments.

A FASTQ file usually uses four lines per sequence.

1. a '@' character, followed by a sequence identifier and an optional description.
2. the raw sequence letters.
3. a '+' character (separator), optionally followed by the same sequence identifier (and any description).
4. quality values for the sequence in Line 2.

Source: https://www.genomatix.de/online_help/help/sequence_formats.html https://sapac.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

Sequences

Sequence formats

3. **FASTQ format**

Example

```
@ML-P2-14:9:000H003HG:1:11102:17290:1073 1:N:0:TCCTGAGC+GCGATCTA
TTTGGTAACAGCATGAATTATTCTAGCCACTAAAACTCTATGAACATCTTGTGAAGGTTTCAGATAGAGCCTGAAGTACACAGAGAACAATTCTTAAAAAA
+
AAAAAEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE<AEEEEEEE
```

A FASTQ file usually uses four lines per sequence.

1. a '@' character, followed by a sequence identifier and an optional description.
2. the raw sequence letters.
3. a '+' character (separator), optionally followed by the same sequence identifier (and any description).
4. quality values for the sequence in Line 2 (e.g. Phred +33 encoded, using ASCII characters).

Source: https://www.genomatix.de/online_help/help/sequence_formats.html
https://sapac.support.illumina.com/bulletins/2016/04/fastq-files-explained.html

ASCII: American Standard Code for Information Interchange

Sequences

Sequence formats

4. EMBL format

- ✓ A sequence file in EMBL format can contain several sequences.
- ✓ One sequence entry starts with an identifier line ("ID"), followed by further annotation lines.
- ✓ The start of the sequence is marked by a line starting with "SQ" and the end of the sequence is marked by two slashes ("//").

Example

```
ID   AB000263 standard; RNA; PRI; 368 BP.
XX
AC   AB000263;
XX
DE   Homo sapiens mRNA for prepro cortistatin like peptide, complete cds.
XX
SQ   Sequence 368 BP;
     acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg        60
     ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg       120
     caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc       180
     aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag       240
     gcgcacccca ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga       300
     agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca       360
     gacctgaa                                                               368
//
```

Source: https://www.genomatix.de/online_help/help/sequence_formats.html

Sequences

Sequence formats

5. GenBank format

- ✓ A sequence file in GenBank format can contain several sequences.
- ✓ One sequence in GenBank format starts with a line containing the word LOCUS and a number of annotation lines.
- ✓ The start of the sequence is marked by a line containing "ORIGIN" and the end of the sequence is marked by two slashes ("//").

Example

```
LOCUS       AB000263                368 bp    mRNA    linear   PRI 05-FEB-1999
DEFINITION  Homo sapiens mRNA for prepro cortistatin like peptide, complete
            cds.
ACCESSION   AB000263
ORIGIN
        1 acaagatgcc attgtccccc ggcctcctgc tgctgctgct ctccggggcc acggccaccg
       61 ctgccctgcc cctggagggt ggccccaccg gccgagacag cgagcatatg caggaagcgg
      121 caggaataag gaaaagcagc ctcctgactt tcctcgcttg gtggtttgag tggacctccc
      181 aggccagtgc cgggcccctc ataggagagg aagctcggga ggtggccagg cggcaggaag
      241 gcgcacccca ccagcaatcc gcgcgccggg acagaatgcc ctgcaggaac ttcttctgga
      301 agaccttctc ctcctgcaaa taaaacctca cccatgaatg ctcacgcaag tttaattaca
      361 gacctgaa
//
```

Source: https://www.genomatix.de/online_help/help/sequence_formats.html