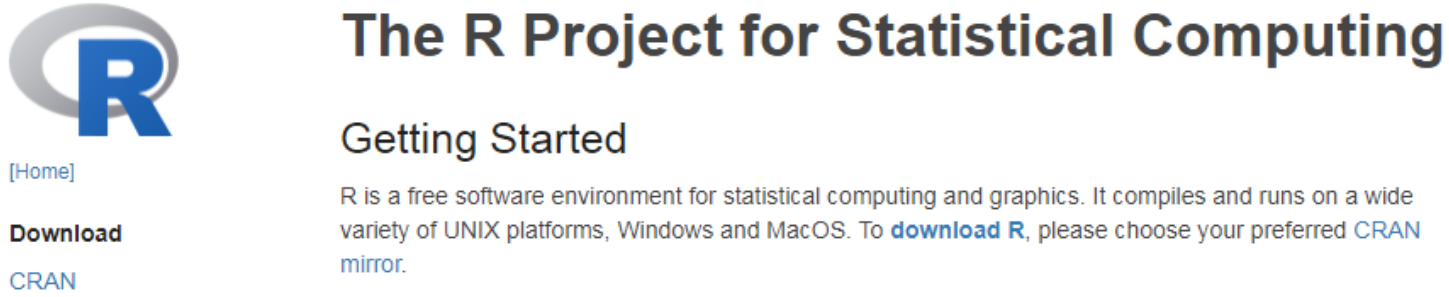
Applied Bioinformatics

Autumn 2022

Wenjun Shen
wjshen@stu.edu.cn

Course description

- This course is designed to introduce the most basic and important concepts, methods and tools used in Bioinformatics. Topics covered in the course include principles and methods used for sequence alignment, phylogenetic tree construction, bulk and single cell transcriptomic data analysis. The scripting language R, which is gaining widespread usage for bioinformatics and computational biology will be used.
- Upon completion of the course, students should be more comfortable working with the vast amounts of biomedical and genomic data and be able to use the bioinformatics tools to solve the problems on their own research.



Course outline

- Introduction to bioinformatics and biological databases
- Sequence alignment
- Phylogenetic trees
- Microarray data analysis
- Bulk RNA-seq data analysis
- Single-cell RNA-seq data analysis
- Introduction to R programming

Grading

- Homework – 20%
- Midterm exam – 25%
- Final exam – 55%

Course materials

- Moodle
- Baidu Netdisk

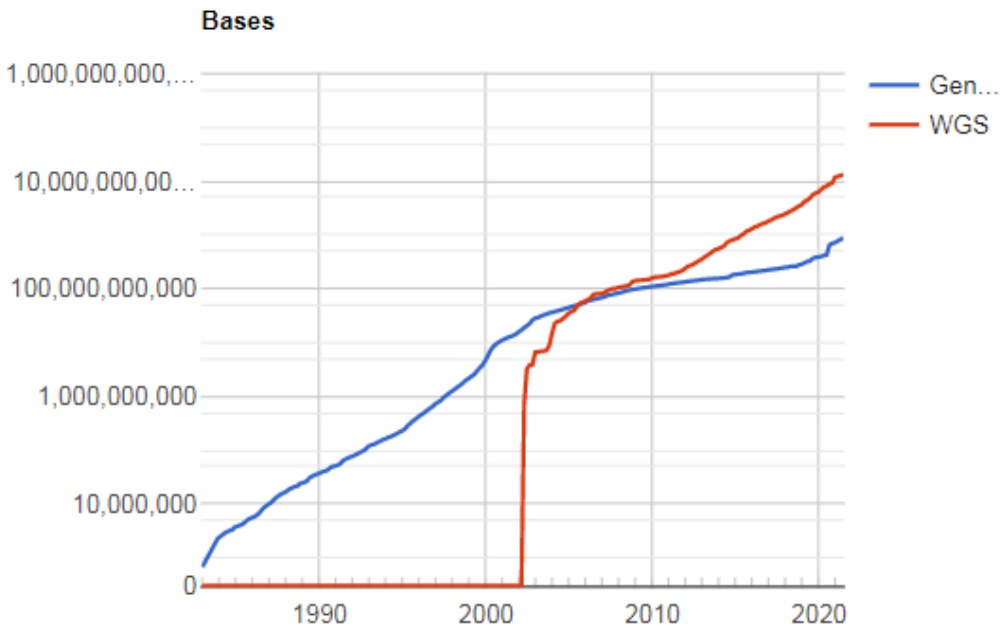
Textbooks and Other Learning Resources

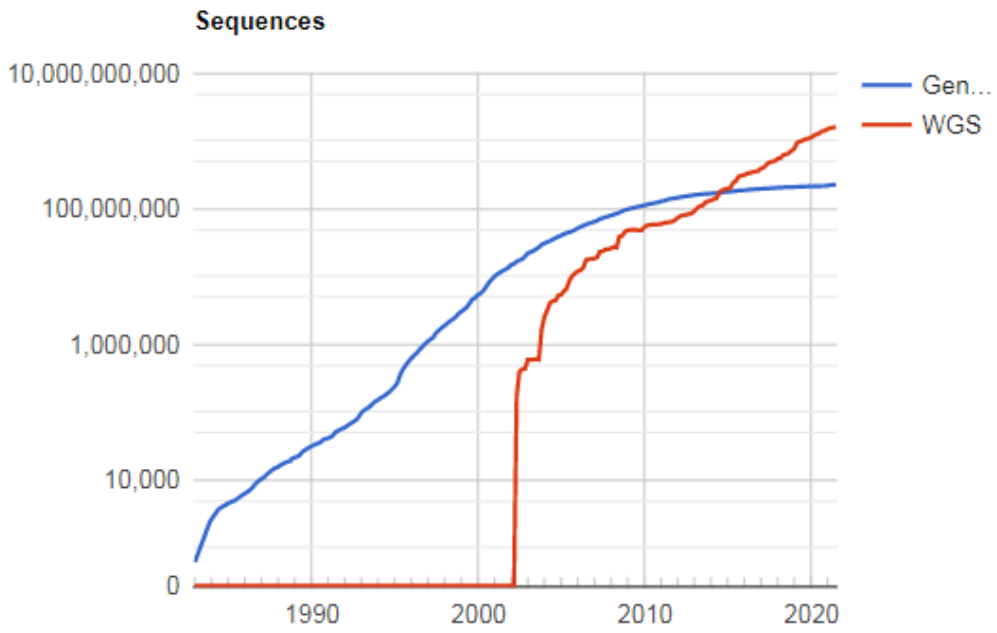
- **Textbooks:**
- (1) Essential Bioinformatics (1st edition), by Jin Xiong.
- **Supplemental textbooks:**
- [1] Bioinformatics with R Cookbook (1st edition), by Paurush Praveen Sinha.
- **Online R resources:**
- [1] Quick-R: quick online reference for data input, basic statistics and plots
- [2] Thomas Girke's R & Bioconductor manuals

Introduction to bioinformatics and biological databases

Why do we need bioinformatics

GenBank and WGS Statistics





		GenBank		WGS	
Release	Date	Bases	Sequences	Bases	Sequences
244	Jun 2021	~866 billion	~227 milliion	~13 trillion	~1 billiion

What's bioinformatics

- Bioinformatics is the discipline of quantitative analysis of information relating to **biological macromolecules** (e.g. DNA, RNA, protein) with the aid of computers.
- Development and implementation of computer programs that enable efficient access to, use and management of, various types of biological information.

Bioinformatics vs. computational biology

- Bioinformatics is limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products.
- Computational biology encompasses all biological areas that involve computation, but do not necessarily involve biological macromolecules.
  - mathematical modeling of ecosystems
  - application of the game theory in behavioral studies
  - phylogenetic construction using fossil

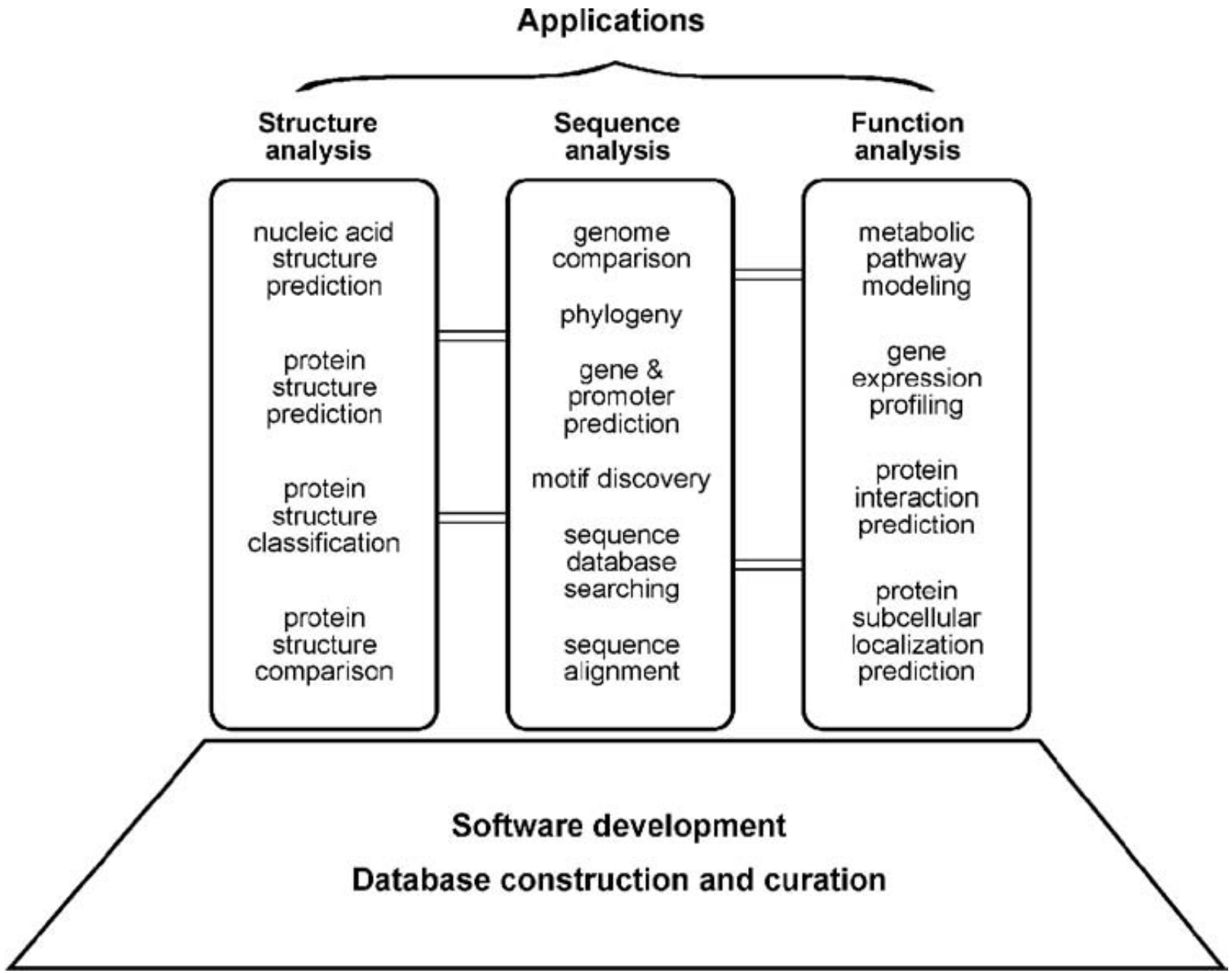
Bioinformatics goal

The primary goal of bioinformatics is to increase the understanding of biological processes using primarily computational methods including: pattern recognition, data mining, machine learning algorithms, and visualization.

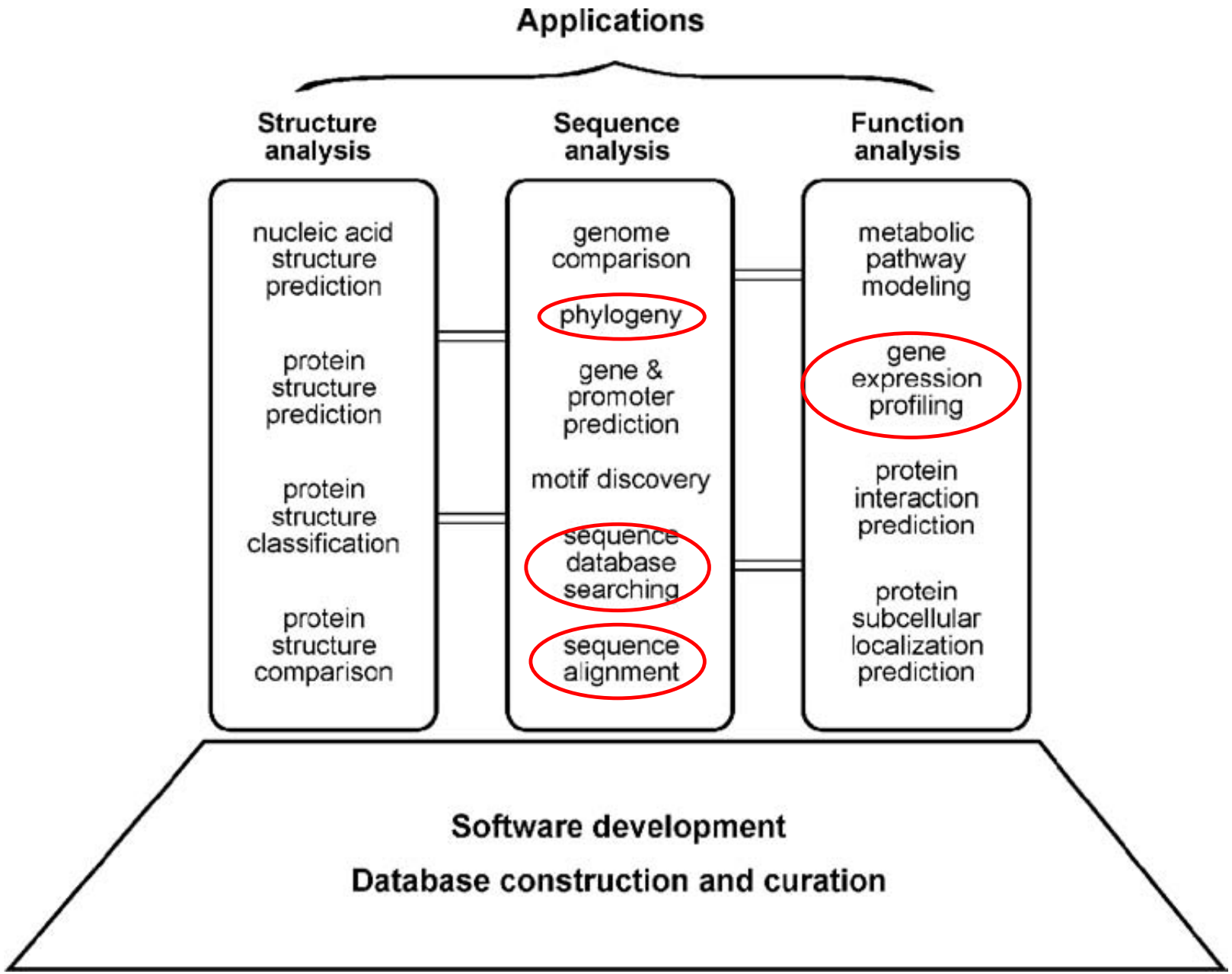
Scope

Bioinformatics consists of two subfields:

1. the **development** of computational tools and databases
   - writing software for sequence, structural, and functional analysis
   - the construction and curating of biological databases
2. the **application** of these tools and databases in generating biological knowledge to better understand living systems
   - these tools are used in three areas of genomic and molecular biological research: molecular sequence analysis, molecular structural analysis, and molecular functional analysis



The three aspects of bioinformatics analysis are not isolated but often interact to produce integrated results.



The three aspects of bioinformatics analysis are not isolated but often interact to produce integrated results.

History of bioinformatics

- Protein sequence and structure wave
- Gene expression wave
- DNA sequencing wave

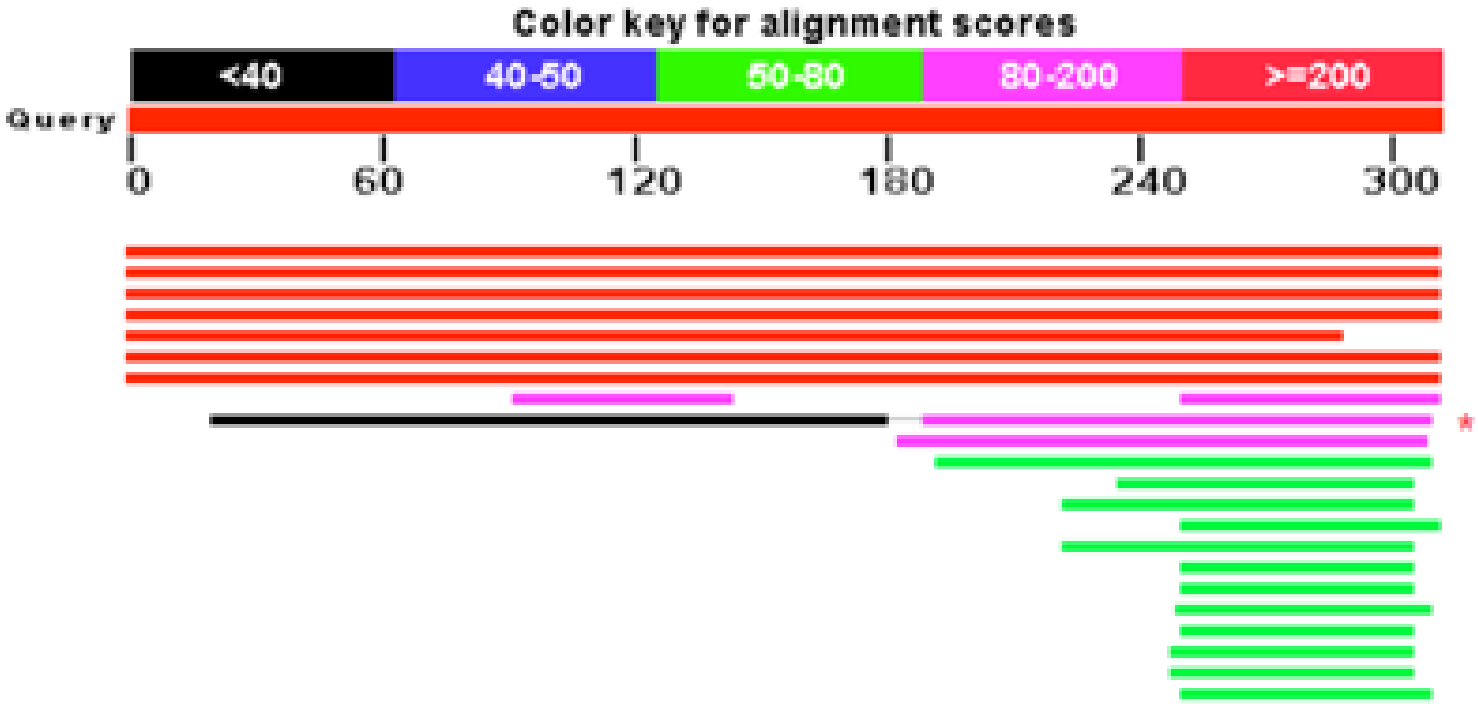
Protein sequence and structure wave

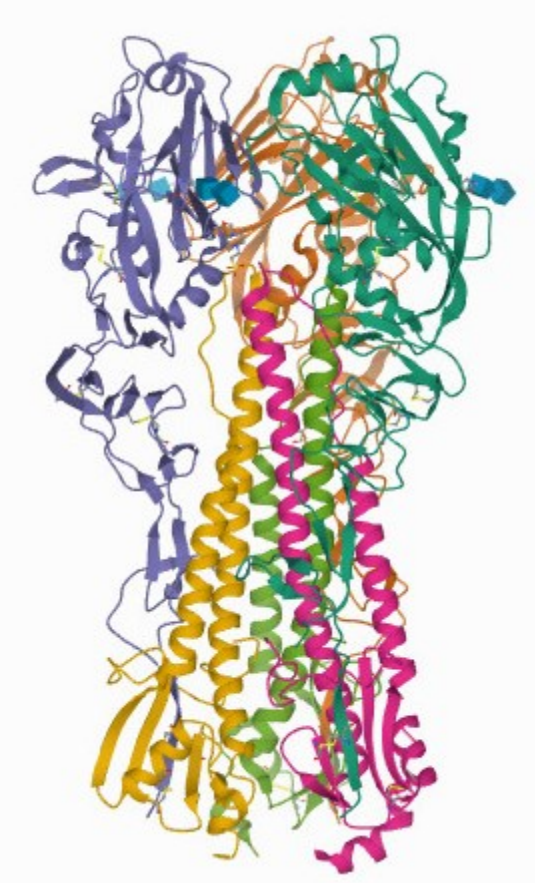
- 1955 - Frederick Sanger and coworkers invented a method to determine the protein sequence of bovine insulin
- 1970 – Needleman-Wunsch algorithm was developed to determine the similarity of two sequences

```
HBA_HUMAN          1 MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLS     50
                     ||||..||:|:||||||:|.|..||||||||||.||||||||||||||:|
HBA_MOUSE          1 MVLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVS     50

HBA_HUMAN         51 HGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFK    100
                     ||||||||||||||||||.:|..|:||:|.||||||||||||||||||||
HBA_MOUSE         51 HGSAQVKGHGKKVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFK    100

HBA_HUMAN        101 LLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR    142
                     ||||||||||||:|.||:||||||||||||||||||||||||||
HBA_MOUSE        101 LLSHCLLVTLASHHPADFTPAVHASLDKFLASVSTVLTSKYR    142
```

Protein sequence and structure wave

- 1973 – Protein Data Bank project was started
- 1990 – BLAST – fast pairwise alignment algorithm

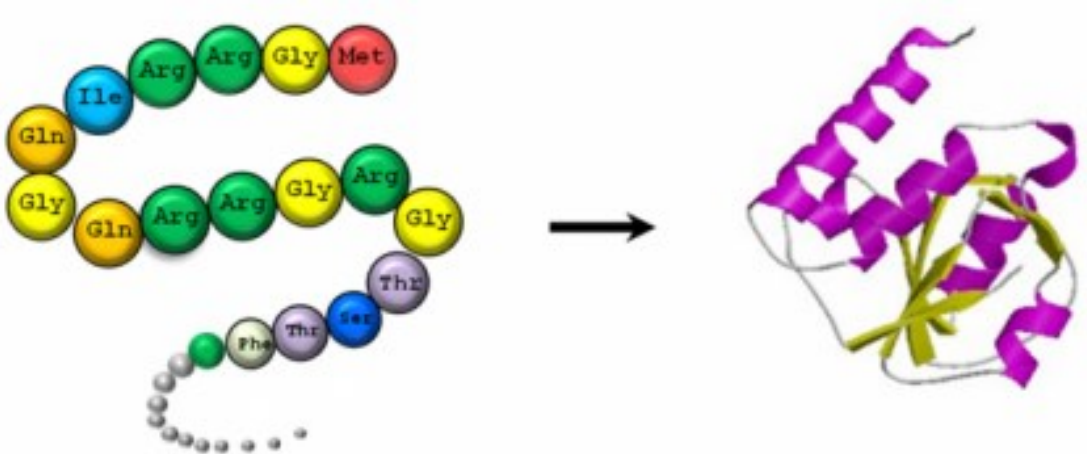


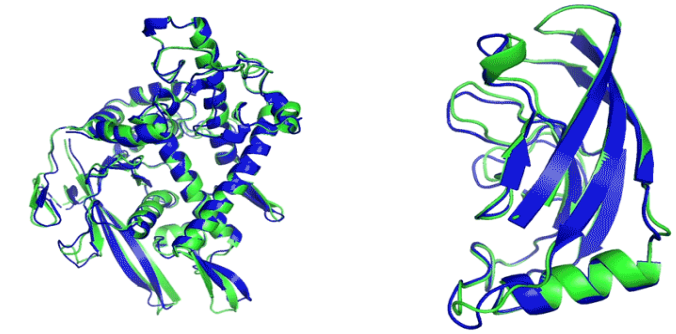


H5N1 influenza virus hemagglutinin

Protein sequence and structure wave

- 1994 – Critical Assessment of Structure Prediction (CASP) competition
  - ✓Predict protein structure based on sequence
  - ✓Use the experimentally solved protein structure as the gold standard to evaluate the computational predicted structure is correct or not
- 2018 – DeepMind's AlphaFold – can determine highly-accurate structures in a matter of days





T1037 / 6vr4
90.7 GDT
(RNA polymerase domain)

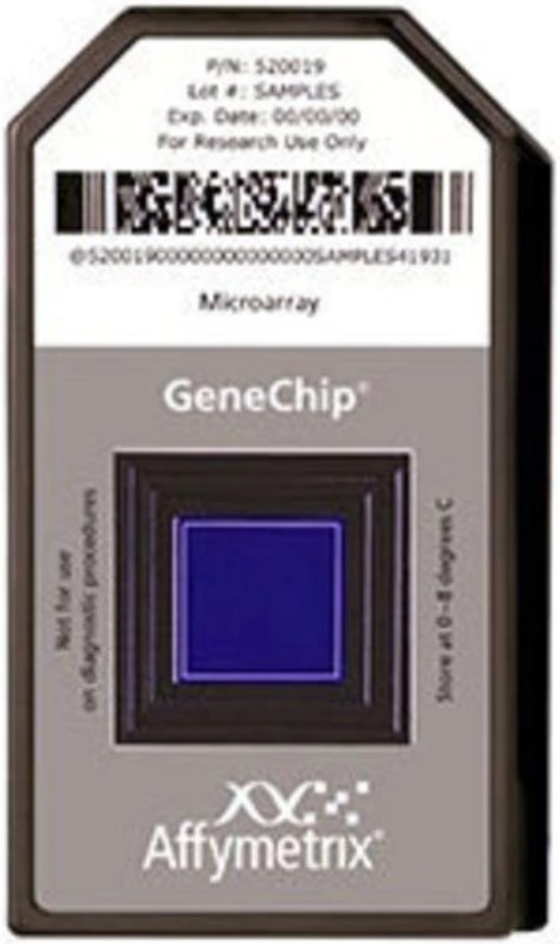
T1049 / 6y4f
93.3 GDT
(adhesin tip)

Experimental result
Computational prediction

DeepMind's AlphaFold

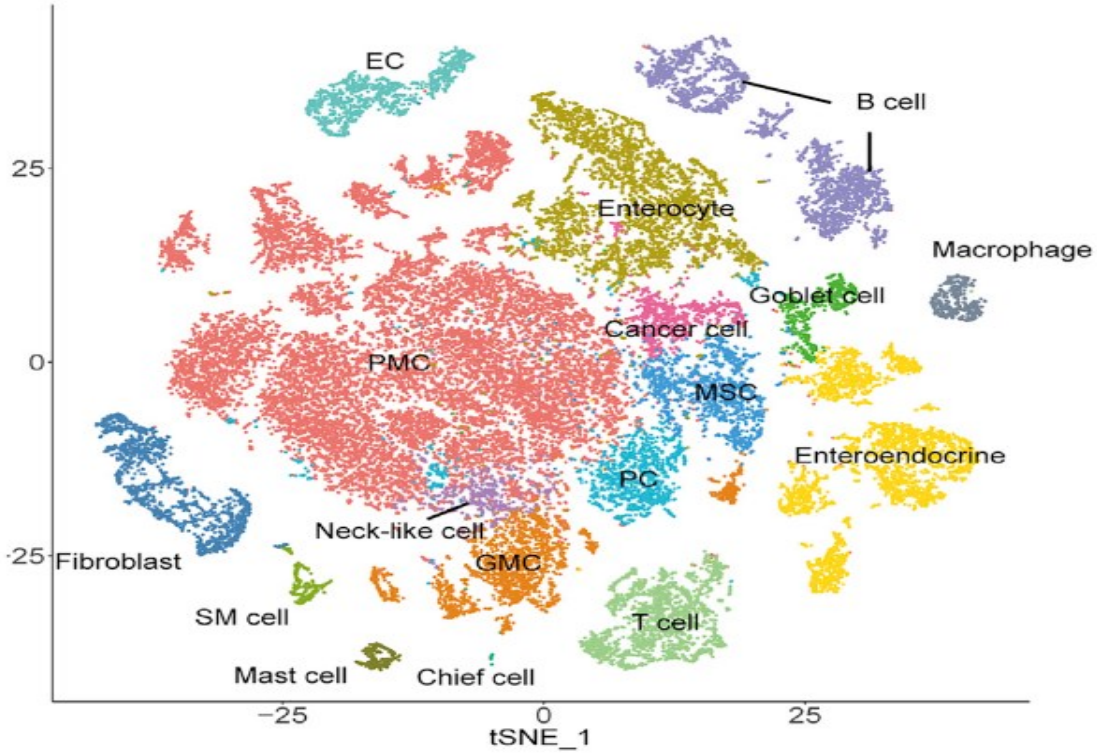
Gene expression wave

- 1977 – Northern blot was invented to measure the expression of a single or a few genes in a cell condition
- 1995 – cDNA microarrays were developed at Stanford to measure hundreds or thousands of genes at a time
  - ✓Created in the lab
  - ✓Sometimes have very significant artifacts
- Late 1990s – Affymetrix microarrays measuring ~6 million probes
  - ✓Commercial microarray platform
  - ✓Results are much more reproducible

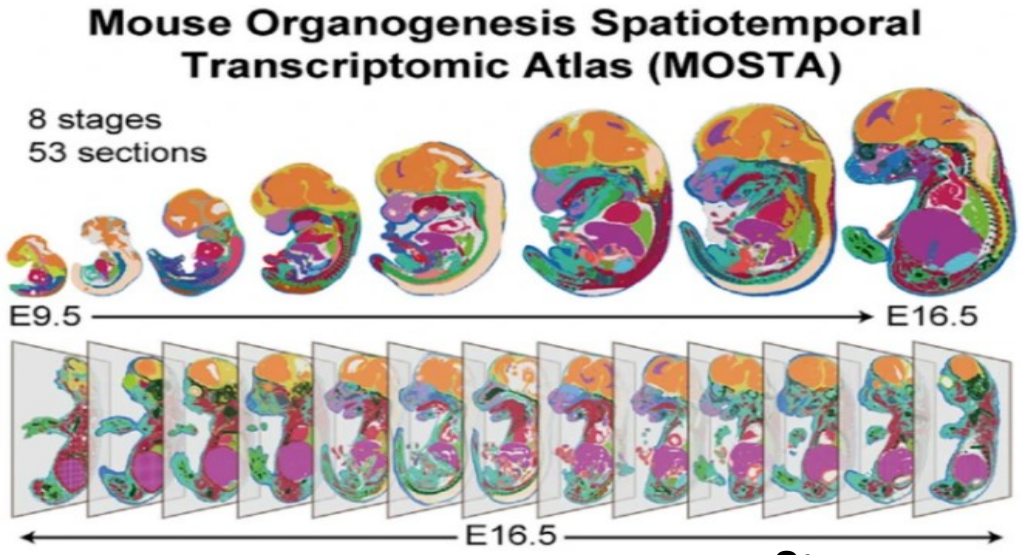


Gene expression wave

- Issues with microarrays:
  - needing to know the sequence *a priori*
  - cross-hybridization artifacts
  - poor quantification of lowly and highly expressed genes
- Mid 2000s – RNA-seq - sequencing based methods - bulk tissue
- 2009 – Single-cell RNA-seq - cellular expression within a bulk tissue
- 2016 – Spatial transcriptomics – get transcriptomic data and know the positional context of those cells in a tissue



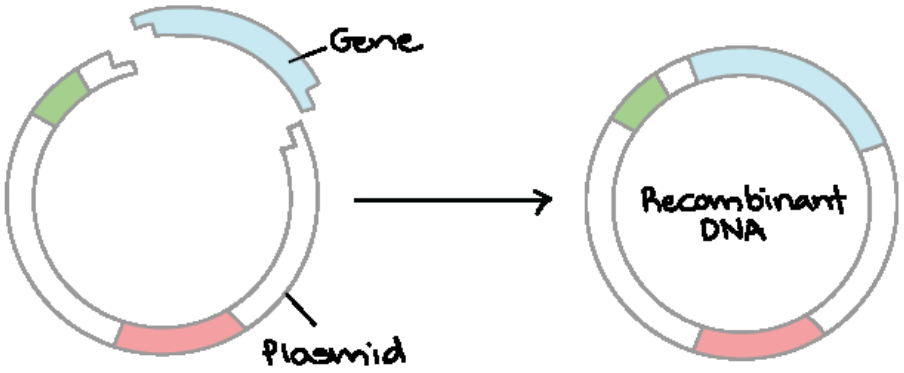
Zhang et al., Cell Reports 2019

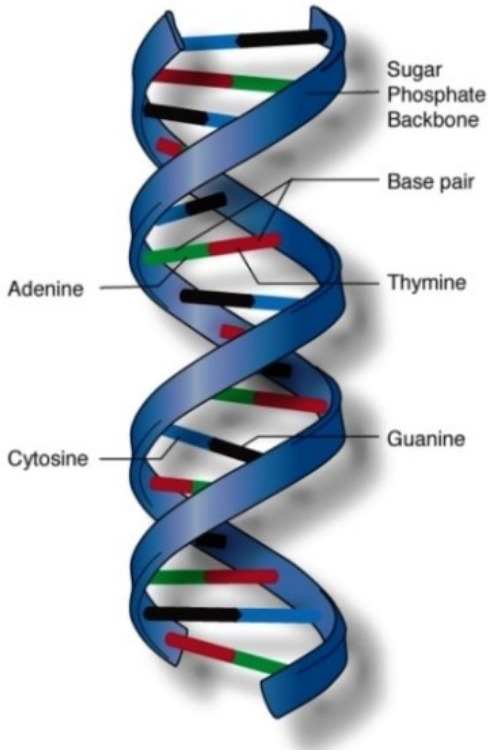


Stereo-seq

DNA sequencing wave

- 1953 – DNA structure – Watson and Crick discovered the DNA structure to be a double helix
- 1972 – Recombinant DNA – a piece of DNA that were created by combining at least two fragments from multiple sources





DNA sequencing wave

- 1977 – Sanger sequencing - DNA sequencing
- 1985 – Polymerase chain reaction (PCR) – amplify a piece of DNA millions to billions of copies, with enough copies you can use sanger sequencing to figure up specific sequence of some gene
- 1988 – National Center for Biotechnology Information (NCBI)
  - houses a series of databases relevant to biotechnology and biomedicine
  - an important resource for bioinformatics tools and services

DNA sequencing wave

- 1990 – Basic Local Alignment Search Tool (BLAST) - programs compare nucleotide or protein sequences to sequence databases
- 1990 – 2003 Human Genome Project

Main goals:

✓identify all the approximate 20,000-25,000 genes in human DNA

✓determine the sequences of the 3 billion chemical base pairs that make up human DNA

✓store this information in databases

✓improve tools for data analysis

DNA sequencing wave

- 2003 – International HapMap project
  - find genetic variants affecting health, disease and responses to drugs and environmental factors
- 2003 – ENCODE project
  - aims to identify all functional elements in the human genome
- 2006-2014 The Cancer Genome Atlas (TCGA) project
  - a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types

DNA sequencing wave

- 2008-2015
  - 1000 Genomes Project
  - establish by far the most detailed catalogue of human genetic variation

A global reference for human genetic variation

The 1000 Genomes Project Consortium

Nature **526**, 68–74 (2015) | Cite this article

Abstract

The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

Introduction to Biological Databases

What is a database

- **Structured** collection of information
- Consists of basic units called records or entries
- Each record consists of fields, which hold **pre-defined** data related to the record
  - For example, a protein database would have protein entries as records and protein properties as fields (e.g., name of protein, length, amino-acid sequence)
- Data retrieval is the main purpose of all databases.
- Knowledge discovery - the identification of connections between pieces of information that were not known when the information was first entered.
  - For example, sequence databases can perform extra computational tasks to identify sequence homology or conserved motifs.

Types of biological databases

- Primary databases
- Secondary databases
- Specialized databases

Types of biological databases

1. Primary databases

- Contain original biological data by experimentalists
- Archives of raw sequence or structural data
- Content controlled by the submitter
- Examples: GenBank, EMBL, DDBJ, SRA, SNP, GEO, PDB

2. Secondary databases

- Contain computationally processed or manually curated information
- Based on original information from primary databases
- Content controlled by third party (e.g., NCBI)
- Examples: SWISS-Prot, PIR, Refseq, UniGene, SCOP

Types of biological databases

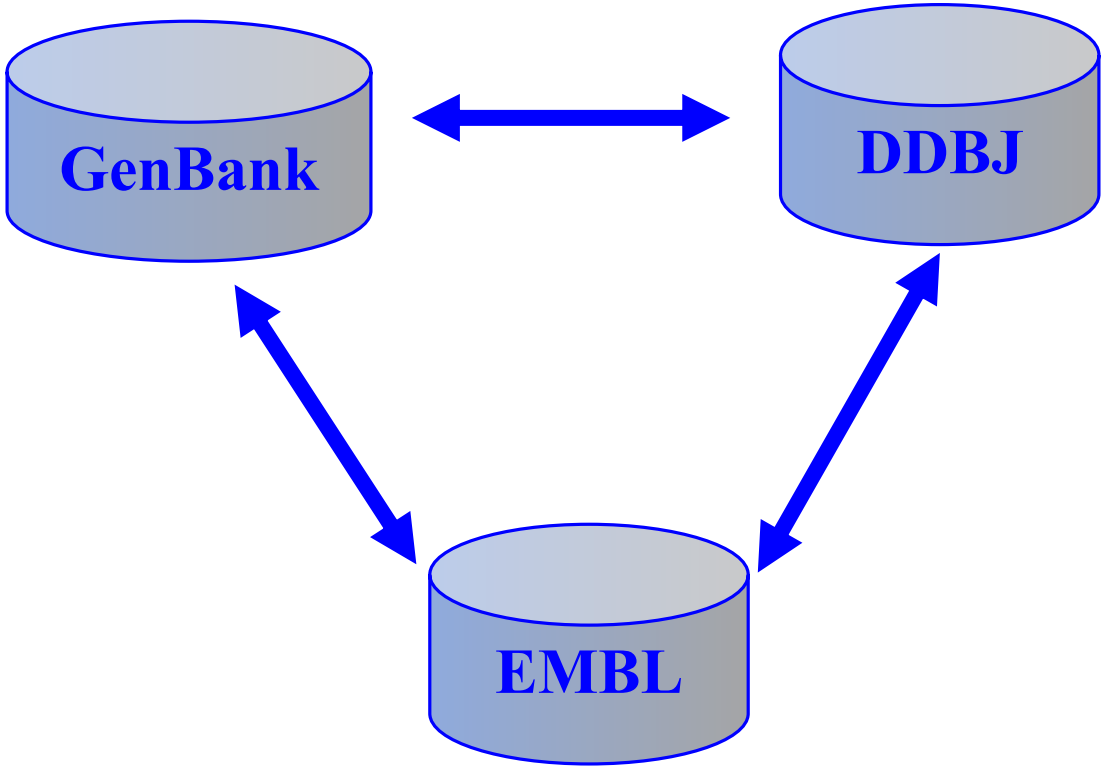
3. Specialized databases

- Cater to a particular research interest
- Specialize in a particular organism or a particular type of data
- For example, Flybase (drosophila), HIV sequence database, and Ribosomal Database

Nucleic acid sequence databases

1. GenBank - NCBI
2. EMBL - the European Molecular Biology Laboratory database
3. DDBJ - the DNA Data Bank of Japan

- All freely available on the internet
- Most of the data in the databases are contributed directly by authors with a minimal level of annotation.



- These three public databases closely collaborate and exchange new data daily.
- Each of the individual databases has a slightly different kind of format to represent the data

GenBank - 1982

- The genetic sequence database at the National Center for Biotechnology information (NCBI)
- The most complete collection of annotated nucleic acid sequence data for almost every organism
- Includes genomic DNA, mRNA, cDNA, ESTs, high throughput raw sequence data, and sequence polymorphisms
- Two ways to search for sequences in GenBank

a. text-based keywords search

b. molecular sequences to search by sequence similarity (BLAST)

ncbi.nlm.nih.gov/genbank/

NCBI Resources How To Sign in to NCBI

GenBank Nucleotide Search

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Other

COVID-19 Information

Public health information (CDC) | Research information (NIH)

SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

GenBank Overview

What is GenBank?

GenBank ® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). 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.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Access to GenBank

There are 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.

GenBank Resources

GenBank Home

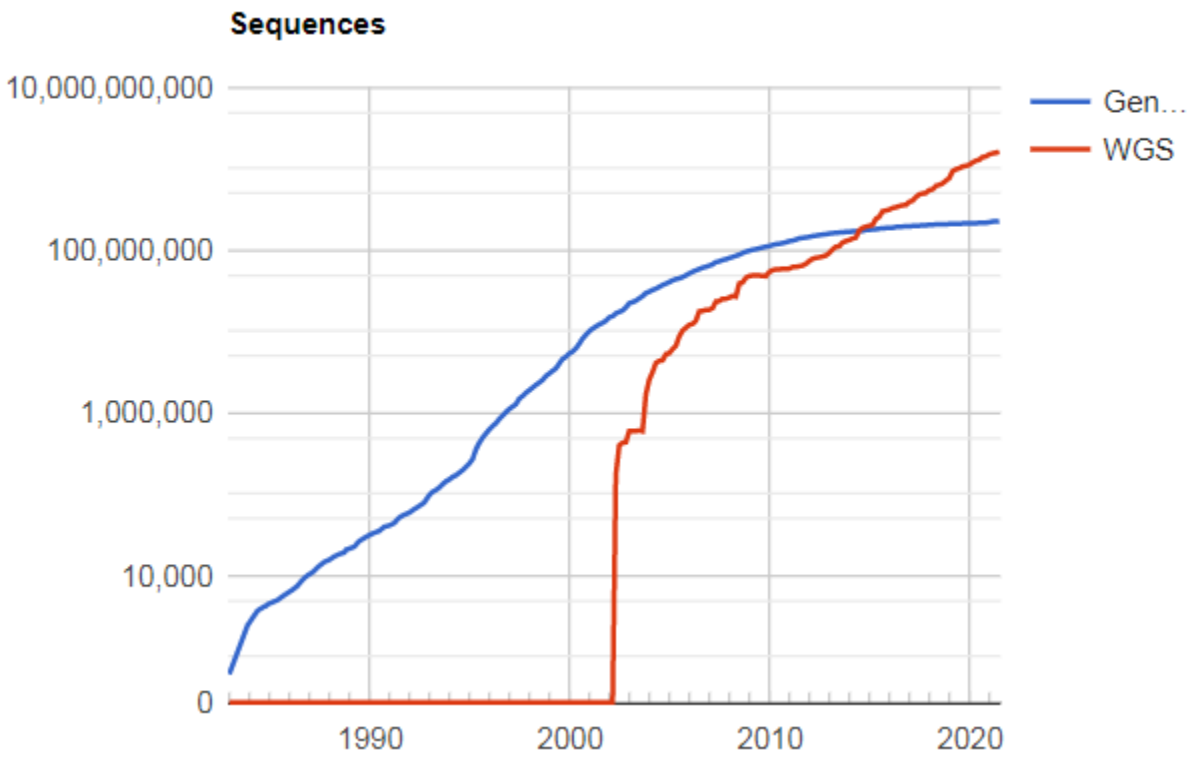
Submission Types

Submission Tools

Search GenBank

Update GenBank Records

Growth of GenBank



		GenBank	
Release	Date	Bases	Sequences
3	Dec-82	680338	606
74	Dec-92	120242234	97,084
133	Dec-02	2.851E+10	22,318,883
193	Dec-12	1.484E+11	161,140,325
244	Jun-21	8.66E+11	227,888,889

GenBank sequence format

Sequence files are produced as **flat files (GBFF):**

1. Header

Describes the origin of the sequence, identification of the organism, and unique identifiers associated with the record.

2. Features

Includes annotation information about the gene and gene product, as well as regions of biological significance reported in the sequence, with location and qualifiers.

3. Sequence

DNA sequences

Header

Accession number: a unique database identifier

Sequence length

Molecule type

GenBank divisions BCT for bacterial sequences

```
LOCUS       AB000100                2992 bp    DNA     linear   BCT 15-MAY-2009
DEFINITION  Synechococcus elongatus PCC 7942 genes for intrinsic membrane
            protein, malK-like protein, cyanase, complete cds.
ACCESSION   AB000100
VERSION     AB000100.1  GI:2330514
KEYWORDS    .
SOURCE      Synechococcus elongatus PCC 7942
  ORGANISM  Synechococcus elongatus PCC 7942
            Bacteria; Cyanobacteria; Oscillatoriophycideae; Chroococcales;
            Synechococcus.
REFERENCE   1
  AUTHORS   Harano,Y., Suzuki,I., Maeda,S., Kaneko,T., Tabata,S. and Omata,T.
  TITLE     Identification and nitrogen regulation of the cyanase gene from the
            cyanobacteria Synechocystis sp. strain PCC 6803 and Synechococcus
            sp. strain PCC 7942
  JOURNAL   J. Bacteriol. 179 (18), 5744-5750 (1997)
   PUBMED   9294430
REFERENCE   2  (bases 1 to 2992)
  AUTHORS   Omata,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-DEC-1996) Contact:Tatsuo Omata School of Agricultural
            Sciences, Nagoya University, Department of Applied Biological
            Sciences; Chikusa, Nagoya, Aichi 464-01, Japan
COMMENT     On Aug 16, 1997 this sequence version replaced gi:1943948.
```

Header

Sequence name

the name and taxonomy of the source organism

whether the sequence is complete or partial

```
LOCUS       AB000100                 2992 bp    DNA     linear   BCT 15-MAY-2009
DEFINITION  Synechococcus elongatus PCC 7942 genes for intrinsic membrane
            protein, malK-like protein, cyanase, complete cds.
ACCESSION   AB000100
VERSION     AB000100.1  GI:2330514
KEYWORDS    .
SOURCE      Synechococcus elongatus PCC 7942
  ORGANISM  Synechococcus elongatus PCC 7942
            Bacteria; Cyanobacteria; Oscillatoriophycideae; Chroococcales;
            Synechococcus.
REFERENCE   1
  AUTHORS   Harano,Y., Suzuki,I., Maeda,S., Kaneko,T., Tabata,S. and Omata,T.
  TITLE     Identification and nitrogen regulation of the cyanase gene from the
            cyanobacteria Synechocystis sp. strain PCC 6803 and Synechococcus
            sp. strain PCC 7942
  JOURNAL   J. Bacteriol. 179 (18), 5744-5750 (1997)
   PUBMED   9294430
REFERENCE   2  (bases 1 to 2992)
  AUTHORS   Omata,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-DEC-1996) Contact:Tatsuo Omata School of Agricultural
            Sciences, Nagoya University, Department of Applied Biological
            Sciences; Chikusa, Nagoya, Aichi 464-01, Japan
COMMENT     On Aug 16, 1997 this sequence version replaced gi:1943948.
```

Header

Version number

Gene index

```
LOCUS       AB000100                 2992 bp    DNA     linear   BCT 15-MAY-2009
DEFINITION  Synechococcus elongatus PCC 7942 genes for intrinsic membrane
            protein, malK-like protein, cyanase, complete cds.
ACCESSION   AB000100
VERSION     AB000100.1  GI:2330514
KEYWORDS    .
SOURCE      Synechococcus elongatus PCC 7942
  ORGANISM  Synechococcus elongatus PCC 7942
            Bacteria; Cyanobacteria; Oscillatoriophycideae; Chroococcales;
            Synechococcus.
REFERENCE   1
  AUTHORS   Harano,Y., Suzuki,I., Maeda,S., Kaneko,T., Tabata,S. and Omata,T.
  TITLE     Identification and nitrogen regulation of the cyanase gene from the
            cyanobacteria Synechocystis sp. strain PCC 6803 and Synechococcus
            sp. strain PCC 7942
  JOURNAL   J. Bacteriol. 179 (18), 5744-5750 (1997)
   PUBMED   9294430
REFERENCE   2  (bases 1 to 2992)
  AUTHORS   Omata,T.
  TITLE     Direct Submission
  JOURNAL   Submitted (26-DEC-1996) Contact:Tatsuo Omata School of Agricultural
            Sciences, Nagoya University, Department of Applied Biological
            Sciences; Chikusa, Nagoya, Aichi 464-01, Japan
COMMENT     On Aug 16, 1997 this sequence version replaced gi:1943948.
```

Accession number

- Unique identifiers which permanently identify sequences in the database
- If the sequence annotation is revised at a later date, the accession number remains the same, but the version number is incremented as is the gi number.
- Assigned and communicated to authors within two working days of the receipt of submission
- This is the number that should be cited in publications.

Features

Sequence length

Scientific name of the organism

taxonomy identification number

```
FEATURES             Location/Qualifiers
     source          1..2992
                     /organism="Synechococcus elongatus PCC 7942"
                     /mol_type="genomic DNA"
                     /strain="PCC 7942"
                     /db_xref="taxon:1140"
                     /clone_lib="constructed in pBluescript II KS-"
     gene            121..912
                     /gene="cynB"
     CDS             121..912
                     /gene="cynB"
                     /codon_start=1
                     /transl_table=11
                     /product="intrinsic membrane protein"
                     /protein_id="BAA21794.1"
                     /db_xref="GI:2330515"
                     /translation="MVRTPVPLYLRWAVSILSVLAFLAIWQIAAASGFLGKTFPGSLR
                     TLQDLFGWLSDPFFDNGPNDLGIGWNLLISLRRVAIGYLLATVVAIPLGIAIGMSALA
                     SSIFSPFVQLLKPVSPLAWLPIGLFLFRDSELTGVFVILISSLWPTLINTAFGVANVN
                     PDFLKVSQSLGASRWRTILKVILPAALPSIIAGMRISMGIAWLVIVAAEMLLGTGIGY
                     FIWNEWNNLSLPNIFSAIIIIGIVGILLDQGFRFLENQFSYAGNR"
```

Features

Gene location

Gene name

```
FEATURES             Location/Qualifiers
     source          1..2992
                     /organism="Synechococcus elongatus PCC 7942"
                     /mol_type="genomic DNA"
                     /strain="PCC 7942"
                     /db_xref="taxon:1140"
                     /clone_lib="constructed in pBluescript II KS-"
     gene            121..912
                     /gene="cynB"
     CDS             121..912
                     /gene="cynB"
                     /codon_start=1
                     /transl_table=11
                     /product="intrinsic membrane protein"
                     /protein_id="BAA21794.1"
                     /db_xref="GI:2330515"
                     /translation="MVRTPVPLYLRWAVSILSVLAFLAIWQIAAASGFLGKTFPGSLR
                     TLQDLFGWLSDPFFDNGPNDLGIGWNLLISLRRVAIGYLLATVVAIPLGIAIGMSALA
                     SSIFSPFVQLLKPVSPLAWLPIGLFLFRDSELTGVFVILISSLWPTLINTAFGVANVN
                     PDFLKVSQSLGASRWRTILKVILPAALPSIIAGMRISMGIAWLVIVAAEMLLGTGIGY
                     FIWNEWNNLSLPNIFSAIIIIGIVGILLDQGFRFLENQFSYAGNR"
```

Features

Coding sequence

Translated protein accession number

Translated protein sequences

```
FEATURES             Location/Qualifiers
     source          1..2992
                     /organism="Synechococcus elongatus PCC 7942"
                     /mol_type="genomic DNA"
                     /strain="PCC 7942"
                     /db_xref="taxon:1140"
                     /clone_lib="constructed in pBluescript II KS-"
     gene            121..912
                     /gene="cynB"
     CDS             121..912
                     /gene="cynB"
                     /codon_start=1
                     /transl_table=11
                     /product="intrinsic membrane protein"
                     /protein_id="BAA21794.1"
                     /db_xref="GI:2330515"
                     /translation="MVRTPVPLYLRWAVSILSVLAFLAIWQIAAASGFLGKTFPGSLR
                     TLQDLFGWLSDPFFDNGPNDLGIGWNLLISLRRVAIGYLLATVVAIPLGIAIGMSALA
                     SSIFSPFVQLLKPVSPLAWLPIGLFLFRDSELTGVFVILISSLWPTLINTAFGVANVN
                     PDFLKVSQSLGASRWRTILKVILPAALPSIIAGMRISMGIAWLVIVAAEMLLGTGIGY
                     FIWNEWNNLSLPNIFSAIIIIGIVGILLDQGFRFLENQFSYAGNR"
```

Sequence

```
ORIGIN
        1 ctgcagccgc cgactgaaat ctatcgggaa gaaaagctcg cttacgacac ctttaacccg
       61 caggatccag tcgcttacct cgcatctcaa aagcagaaat acggggagata aacacaactt
      121 atggtgagaa ctcctgtacc gctttaccta cgttgggcgg tctccatcct cagcgtgctt
      181 gcgttcctag ccatttggca aattgcggca gcttcaggat ttttaggcaa aacttttcct
      241 ggctccctgc gcactttgca ggatttgttt ggatggcttt cagatccctt ctttgataac
      301 ggccccaatg acttagggat tggctggaac ttactgatta gtttgcgtcg cgttgcgatc
      361 ggctacctgc tggcaacagt tgttgcaatt cctttgggga ttgcaatcgg tatgtcggcg
      421 ctagcttcca gtatttttc gccctttgtg caactcctga agccagtttc acctttggcc
      481 tggttgccga ttggtctctt cttattccga gattcggaat tgacgggtgt ttttgtcatc
      541 ctgatttcga gtctgtggcc aacgttgatc aacacagcgt ttggggtggc gaatgtcaat
      601 cctgactttt tgaaggtttc gcaatctttg ggagctagtc gttggcgcac gattctgaag
      ...
     2581 gcaggatcgt aacttggctt tttggacctg caccgcttgg acggatgaag gagccatgcg
     2641 tcggttcatg agagcggatg cccacgggca ggccatgacg aaattgatgg attggtgcag
     2701 cgaagcctca gtcgtccatt ggcagcagga tcagccagac ttgcccgact ggcaggaagc
     2761 tcaccgccgc atgatcgcgg agggggcgccc ctccaaagtg aaccatcctt cggctgccca
     2821 ccaagcattt caggtcgatc cgccgcgccg cgcctagctc agtgactgcg gtcgcgctgt
     2881 cttgcatcat tgcttcgctc taccagcccg gatcgctggc acagtccacg gtgatctcac
     2941 ccgaggcggc atcgggaatc gcagtgatac agccgcagac tggctcgcca tc
//
```

ebi.ac.uk

EMBL-EBI | Services | Research | Training | About us | EMBL-EBI

EMBL-EBI

The home for big data in biology

We help scientists exploit complex information to make discoveries that benefit humankind.

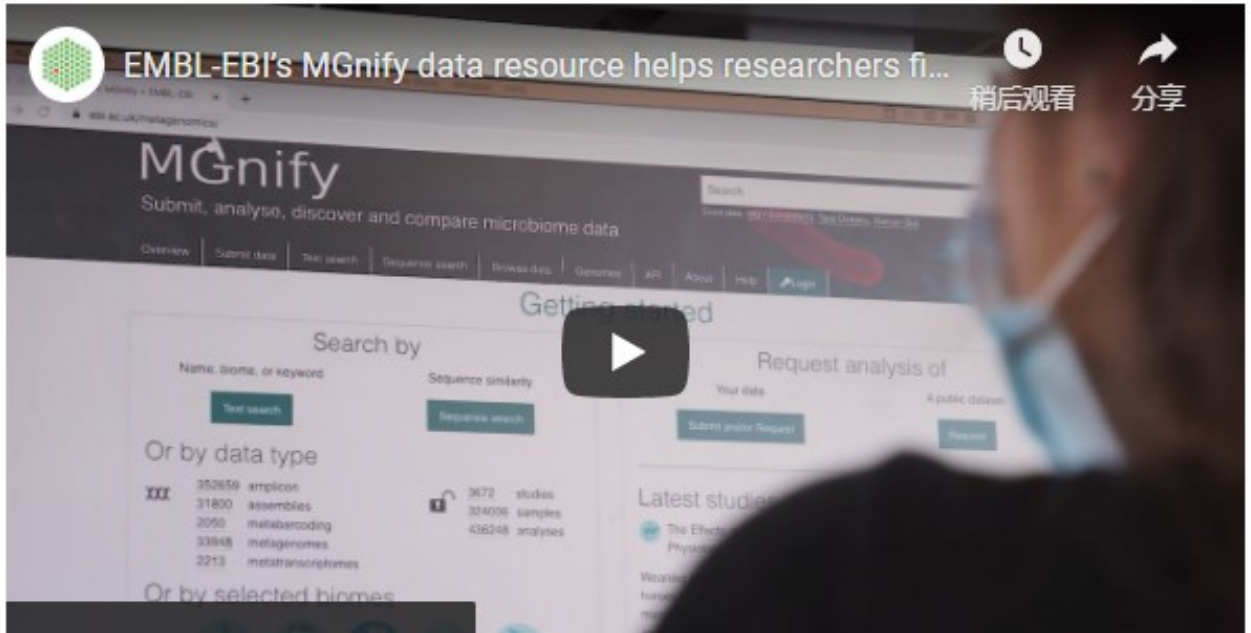
Find tools and resources or deposit data.

Explore dozens of biological data resources with our Search service.

Find a gene, protein or chemical | All | Search

Example searches: blast keratin bfl1 | Build query

Featured topic



Latest news

02 Aug 2021

Using ultrafast technology for protein sequencing

22 Jul 2021

DeepMind and EMBL release the most complete database of predicted 3D structures of human proteins

EMBL sequence format

Sequence files is produced as **flat files:**

1. Header

Describes the origin of the sequence, identification of the organism, and unique identifiers associated with the record.

2. Features

Includes annotation information about the gene and gene product, as well as regions of biological significance reported in the sequence, with location and qualifiers.

3. Sequence

DNA sequences

Header

Accession number: a unique database identifier

Version number

DNA topology 'circular' or 'linear'

Molecule type

Taxonomic division

```
ID   DQ286969; SV 1; linear; genomic DNA; STD; HUM; 1098 BP.
XX
AC   DQ286969;
XX
DT   05-DEC-2005 (Rel. 86, Created)
DT   08-DEC-2005 (Rel. 86, Last updated, Version 4)
XX
DE   Homo sapiens APOE (APOE) gene, promoter region and 5' UTR.
XX
KW   .
XX
OS   Homo sapiens (human)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae;
OC   Homo.
XX
RN   [1]
RP   1-1098
RX   DOI; 10.1016/j.molbrainres.2005.02.001.
RX   PUBMED; 15893602.
RA   Du Y., Chen X., Wei X., Bales K.R., Berg D.T., Paul S.M., Farlow M.R.,
RA   Maloney B., Ge Y.W., Lahiri D.K.;
RT   "NF-(kappa)B mediates amyloid beta peptide-stimulated activity of the human
RT   apolipoprotein E gene promoter in human astroglial cells";
RL   Brain Res. Mol. Brain Res. 136(1-2):177-188(2005).
XX
RN   [2]
RP   1-1098
RA   Du Y., Chen X., Wei X., Bales K.R., Berg D.T., Paul S.M., Farlow M.R.,
RA   Maloney B., Ge Y.-W., Lahiri D.K.;
RT   ;
RL   Submitted (10-NOV-2005) to the INSDC.
RL   Psychiatric Research, Indiana University School of Medicine, 791 N. Union
RL   Drive, Indianapolis, IN 46202, USA
XX
DR   MD5; e807dca94e11182865811f67dc4b365f.
```

Features

```
FH   Key             Location/Qualifiers
FH
FT   source          1..1098
FT                   /organism="Homo sapiens"
FT                   /chromosome="19"
FT                   /map="19q13.2"
FT                   /mol_type="genomic DNA"
FT                   /db_xref="taxon:9606"
FT   gene            <1..>1098
FT                   /gene="APOE"
FT   misc_binding    199..208
FT                   /gene="APOE"
FT                   /bound_moiety="NF-kB"
FT   misc_feature    532..541
FT                   /gene="APOE"
FT                   /note="nonfunctional NF-kB binding site"
FT   regulatory      1022..1027
FT                   /gene="APOE"
FT                   /regulatory_class="TATA_box"
FT   mRNA            1055..>1098
FT                   /gene="APOE"
FT                   /product="APOE"
FT   5'UTR           1055..>1098
FT                   /gene="APOE"
```

Sequence

the numbers of A, G, C, and T in the sequence

```
SQ   Sequence 1098 BP; 218 A; 350 C; 289 G; 241 T; 0 other;
     ggaacttgat gctcagagag gacaagtcat ttgcccaagg tcacacagct ggcaactggc      60
     agagccagga ttcacgccct ggcaatttga ctccagaatc ctaaccttaa cccagaagca     120
     cggcttcaag cccctggaaa ccacaatacc tgtggcagcc agggggaggt gctggaatct     180
     catttcacat gtggggaggg ggctcccctg tgctcaaggt cacaaccaaa gaggaagctg     240
     tgattaaaac ccaggtccca tttgcaaagc ctcgactttt agcaggtgca tcatactgtt     300
     cccacccctc ccatcccact tctgtccagc cgcctagccc cactttcttt tttttctttt     360
     tttgagacag tctccctctt gctgaggctg gagtgcagtg gcgagatctc ggctcactgt     420
     aacctccgcc tcccgggttc aagcgattct cctgcctcag cctcccaagt agctaggatt     480
     acaggcgccc gccaccacgc ctggctaact tttgtatttt tagtagagat ggggtttcac     540
     catgttggcc aggctggtct caaactcctg accttaagtg attcgcccac tgtggcctcc     600
     caaagtgctg ggattacagg cgtgagctac cgcccccagc ccctcccatc ccacttctgt     660
     ccagcccccct agccctactt tctttctggg atccaggagt ccagatcccc agcccccctct   720
     ccagattaca ttcatccagg cacaggaaag gacagggtca ggaaaggagg actctgggcg     780
     gcagcctcca cattcccctt ccacgcttgg cccccagaat ggaggagggt gtctggatta     840
     ctgggcgagg tgtcctccct tcctggggac tgtgggggt ggtcaaaaga cctctatgcc      900
     ccacctcctt cctccctctg ccctgctgtg cctggggcag ggggagaaca gcccacctcg     960
     tgactggggg ctggcccagc ccgccctatc cctgggggag ggggcgggac agggggagcc    1020
     ctataattgg acaagtctgg gatccttgag tcctactcag ccccagcgga ggtgaaggac    1080
     gtccttcccc aggagccg                                                  1098
//
```

EMBL identifier	GenBank identifier
ID	LOCUS
DE	DEFINITION
AC	ACCESSION
SV	VERSION
KW	KEYWORDS
OS	SOURCE
OC	ORGANISM
DT	
RN	REFERENCE
RA	AUTHORS
RT	TITLE
RL	JOURNAL
RX	MEDLINE
RC	REMARK
RP	
CC	COMMENT
DR	
FH	FEATURES
FT	
SQ	BASE CONTENT
空格	ORIGIN
//	//

Alternative sequence format

FASTA

- A single definition line that begins with a right angle bracket (>)
- A plain sequence in standard one-letter symbols starts in the second line

```
>AY539659.1 Homo sapiens CD45 (PTPRC) gene, exon 4 and partial cds
GATTGACTACAGCAAAGATGCCCAGTGTTCCACTTTCAAGTGACCCCTTACCTACTCACACCACTGCATT
CTCACCCGCAAGCACCTTTGAAAGAGAAAATGACTTCTCAGAGACCACAACTTCTCTTAGTCCAGACAAT
ACTTCCACCCAAGTATCCCCGGACTCTTTGGATAATGCTAGTGCTTTTAATACCACAG
```

- It is readable by many bioinformatics analysis programs.
- The drawback of this format is that most annotation information is lost.

FASTA file

```
>AAT64830 A/Akita/4/1993 1993// HA H3N2 Human
MKTIIALSYILCLVFAQKLPGNDNSTATLCLGHHAVPNGTLVKTITNDQIEVTNATELVQSSSTGRICDS
PHRILDGKNCTLIDALLGDPHCDGFQNKEWDLFVERSKAYSNCYPYDVPDYASLRSLVASSGTLEFINED
FNWTGVAQDGGSYACKRGSVNSFFSRLNWLHKLEYKYPALNVTMPNNGKFDKLYIWGVHHPSTDSDQTSL
YVRASGRVTVSTKRSQQTVIPNIGSRPWVRGQPSRISIYWTIVKPGDILLINSTGNLIAPRGYFKIRNGK
SSIMRSDAPIGNCSSECITPNGSIPNDKPFQNVNRITYGACPRYVKQNTLKLATGMRNVPEKQTRGIFGA
IAGFIENGWEGMV
>AAT64720 A/Amsterdam/1609/1977 1977// HA H3N2 Human
MKTIIALSYIFCLVFAQDLPGNDNSTATLCLGHHAVPNGTLVKTITNDQIEVTNATELVQSSSTGKICDN
PHRILDGINCTLIDALLGDPHCDGFQNEKWDLFVERSKAFSNCYPYDVPDYASLRSLVASSGTLEFINEG
FNWTGVTQNGGSSACKRGPDNGFFSRLNWLYKSGSTYPVQNVTMPNNDNSDKLYIWGVHHPSTDKEQTDL
YVQASGKVTVSTKRSQQTVIPNVGSRPWVRGLSSRVSIYWTIVKPGDILVINSNGNLIAPRGYFKMRTGK
SSIMRSDAPIGTCSSECITPNGSIPNDKPFQNVNKITYGACPKYIKQNTLKLATGMRNVPEKQTRGIFGA
IAGFIENGWEGMI
>AAT64790 A/Amsterdam/4112/1992 1992// HA H3N2 Human
MKTIIALSYILCLVFAQKLPGNDNSTATLCLGHHAVPNGTLVKTITNDQIEVTNATELVQNSSTGRICDS
PHRILDGKNCTLIDALLGDPHCDDFQNKEWDLFVERSKAYSNCYPYDVPDYASLRSLVASSGTLEFINED
FNWTGVAQSGESYACKRGSVKSFFSRLNWLHESEYKYPALNVTMPNNGKFDKLYIWGVHHPSTDREQTSL
YVRASGRVTVSTKRSQQTVIPNIGSRPWVRGLSSRISIYWTIVKPGDILLINSTGNLIAPRGYFKIRTGK
SSIMRSDAPIGTCSSECITPNGSIPNDKPFQNVNRITYGACPRYVKQNTLKLATGMRNVPEKQTRGIFGA
IAGFIENGWEGMV
```

Protein databases

- Uniprot
- Swiss-Prot
- TrEMBL

protein sequence databases

- PDB

protein structure database

```
>pdb|5IBL|E Chain E, Hemagglutinin
GLFGAIAGFIEGGWTGMVDGWYGYHHQNEQGSGYAADLKSTQNAIDEITNKVNSVIEKMNTQFTAVGKEF
NHLEKRIENLNKKVDDGFLDIWTYNAELLVLLENERTLDYHDSNVKNLYEKVRSQLKNNAKEIGNGCFEF
YHKCDNTCMESVKNGTYDYPKYSEEAKLNREEIDGV
```

Universal Protein Resource (UniProt)

1. UniProtKB – Protein knowledgebase
   a. Swiss-Prot – manually annotated and reviewed
   b. TrEMBL – automatically annotated and is not manually reviewed
2. UniParc – protein Archive contains most of the publicly available protein sequences in the world, stores each unique sequence only once, and contains only protein sequences
3. UniRef – non-redundant reference clusters, provide clustered sets of sequences from the UniProtKB and selected UniParc records

   For example, UniRef100 combines identical sequences and sub-fragments with 11 or more residues from any organism into a single UniRef entry.
4. Proteomes – set of proteins thought to be expressed by an organism

   Provides proteomes for species with completely sequenced genomes

uniprot.org

UniProt

UniProtKB

BLAST Align Retrieve/ID mapping Peptide search SPARQL

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of

UniProtKB

UniProt Knowledgebase

Swiss-Prot (565,254)

Manually annotated and reviewed.

Records with information extracted from literature and curator-evaluated computational analysis.

TrEMBL (219,174,961)

Automatically annotated and not reviewed.

Records that await full manual annotation.

UniRef

Sequence clusters

UniParc

Sequence archive

Proteomes

Proteome sets

Supporting data

Literature citations

Cross-ref. databases

Taxonomy

Diseases

XXX

Subcellular locations

Keywords

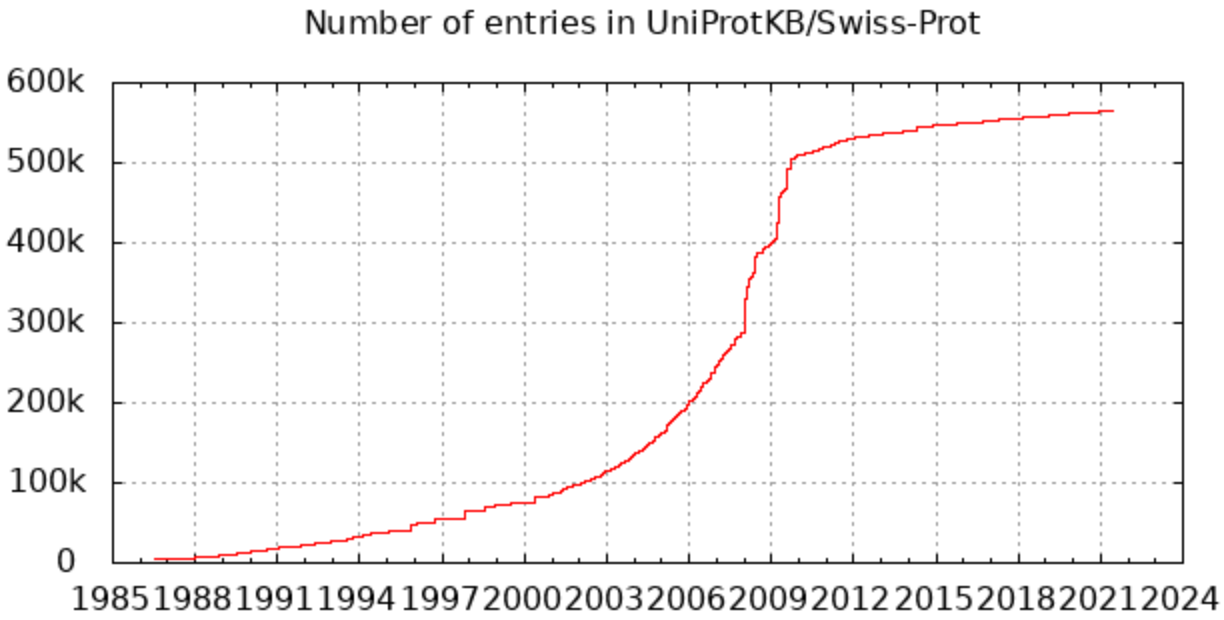
Swiss-Port - 1986

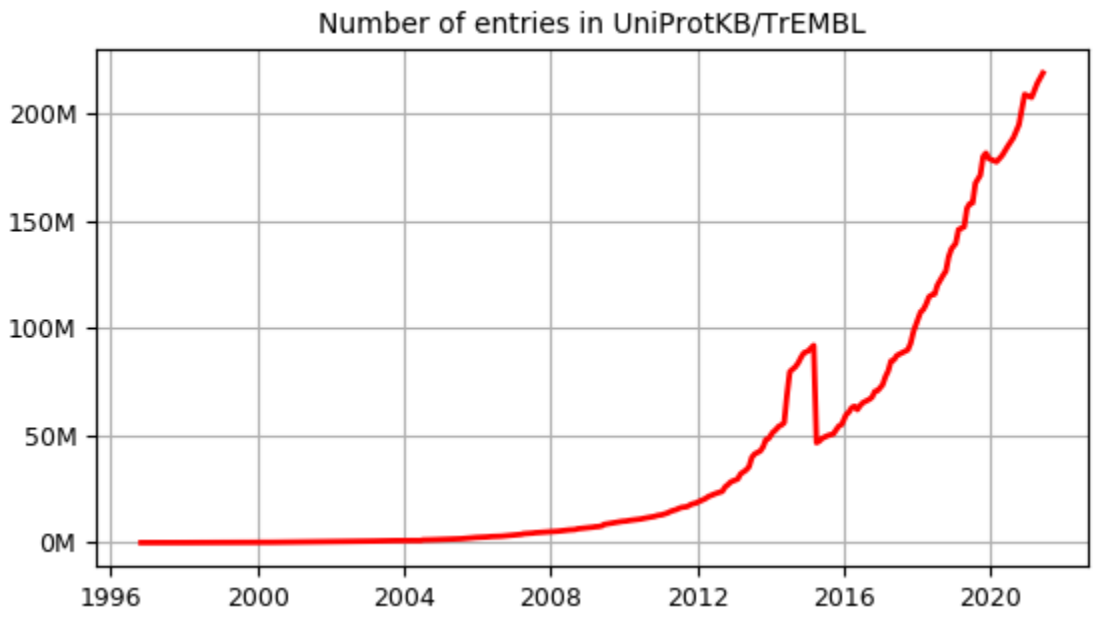
- A curated protein sequence database
- Annotation extracted from literature and curator-evaluated computational analysis
- Provide protein sequences with **a high level of annotation** (e.g., the description of protein function, structure domains and post translational modifications, etc.).
- Has a very low level of redundancy

TrEMBL

- Contains all translations of EMBL nucleotide sequence entries, which is not yet integrated in Swiss-Port
- Automatically computer-annotated and not reviewed
- Move to Swiss-Port after full manual annotation
- Currently, Swiss-Port have ~0.57 and TrEMBL have ~227 milliom sequences.

Growth of protein sequences





Protein data bank (PDB)

- The main primary database for 3D structures of biological **macromolecules** (proteins and nucleic acids)
- Archives **atomic coordinates** of macromolecules determined by x-ray crystallography, NMR spectroscopy or electron microscopy.
- Currently managed by the Research Collaboratory for Structural Bioinformatics (RCSB)
- Provide a variety of tools and resources for studying the structures of biological macromolecules and their relationship with other sequences, its function and diseases caused if any .

rcsb.org

RCSB PDB | Deposit | Search | Visualize | Analyze | Download | Learn | More | Documentation | Careers | MyPDB

RCSB PDB PROTEIN DATA BANK

181780 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Enter search terms or PDB ID(s).

Advanced Search | Browse Annotations

Help

PDB-101 | Worldwide PDB | EMDataResource | NDB Nucleic Acid Database | Worldwide Protein Data Bank Foundation

Celebrating 50 YEARS OF Protein Data Bank

Developers: Join the RCSB PDB Team | Explore Open Positions

Welcome

Deposit

Search

Visualize

Analyze

Download

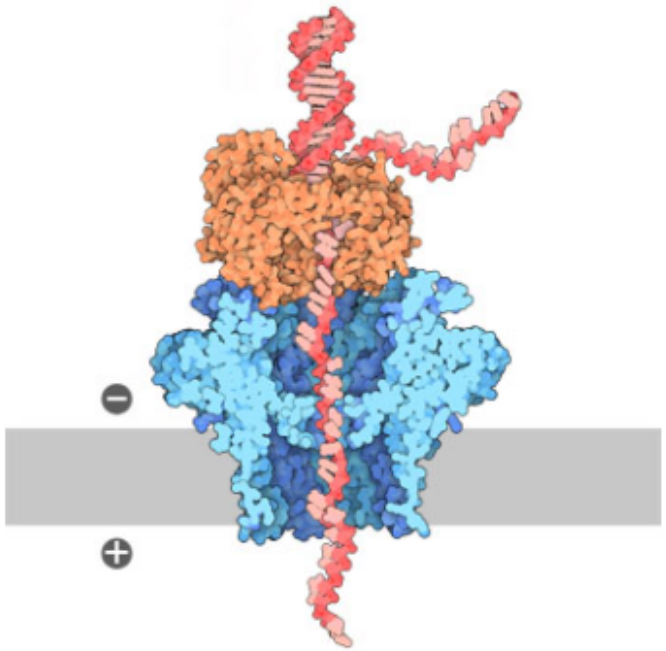
Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

September **Molecule of the Month**



DNA-Sequencing Nanopores

Contact Us

RCSB PDB PROTEIN DATA BANK

170966 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

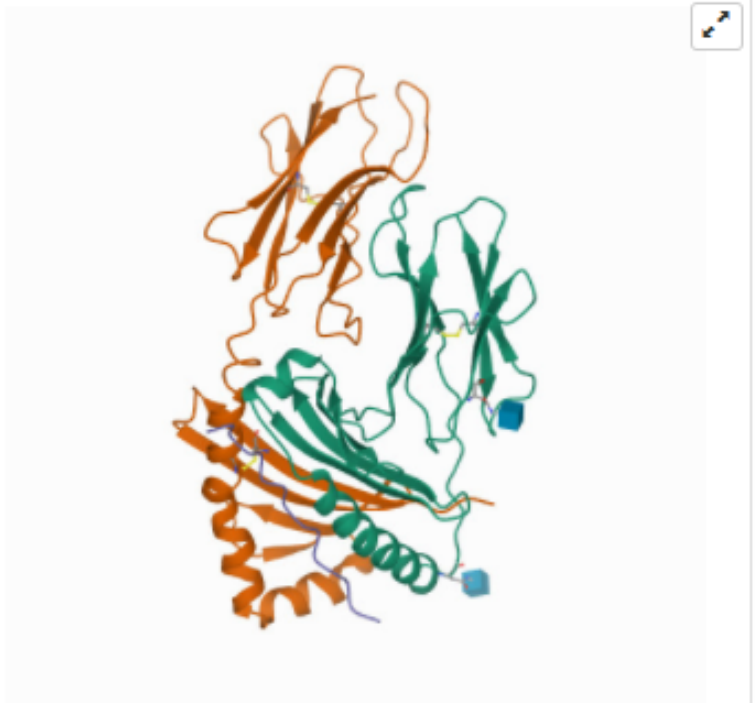
Enter search term(s)

Advanced Search | Browse Annotations

PDB-101 | WORLDWIDE PDB | EMDataResource | NDB NUCLEIC ACID DATABASE | Worldwide Protein Data Bank Foundation

Structure Summary | 3D View | Annotations | Experiment | Sequence | Genome

Biological Assembly 1



3D View: Structure | Electron Density | Ligand Interaction

Global Symmetry: Asymmetric - C1

PDBid

1A6A

Display Files | Download Files

THE STRUCTURE OF AN INTERMEDIATE IN CLASS II MHC MATURATION: CLIP BOUND TO HLA-DR3

DOI: 10.2210/pdb1A6A/pdb

Classification: COMPLEX (TRANSMEMBRANE/GLYCOPROTEIN)
Organism(s): Homo sapiens
Mutation(s): No

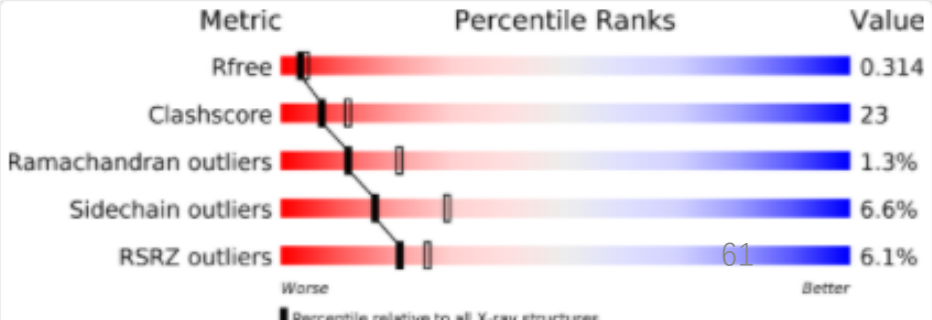
Deposited: 1998-02-22 **Released:** 1998-05-27
Deposition Author(s): Ghosh, P., Amaya, M., Mellins, E., Wiley, D.C.

Experimental Data Snapshot

Method: X-RAY DIFFRACTION
Resolution: 2.75 Å
R-Value Free: 0.325
R-Value Work: 0.246
R-Value Observed: 0.246

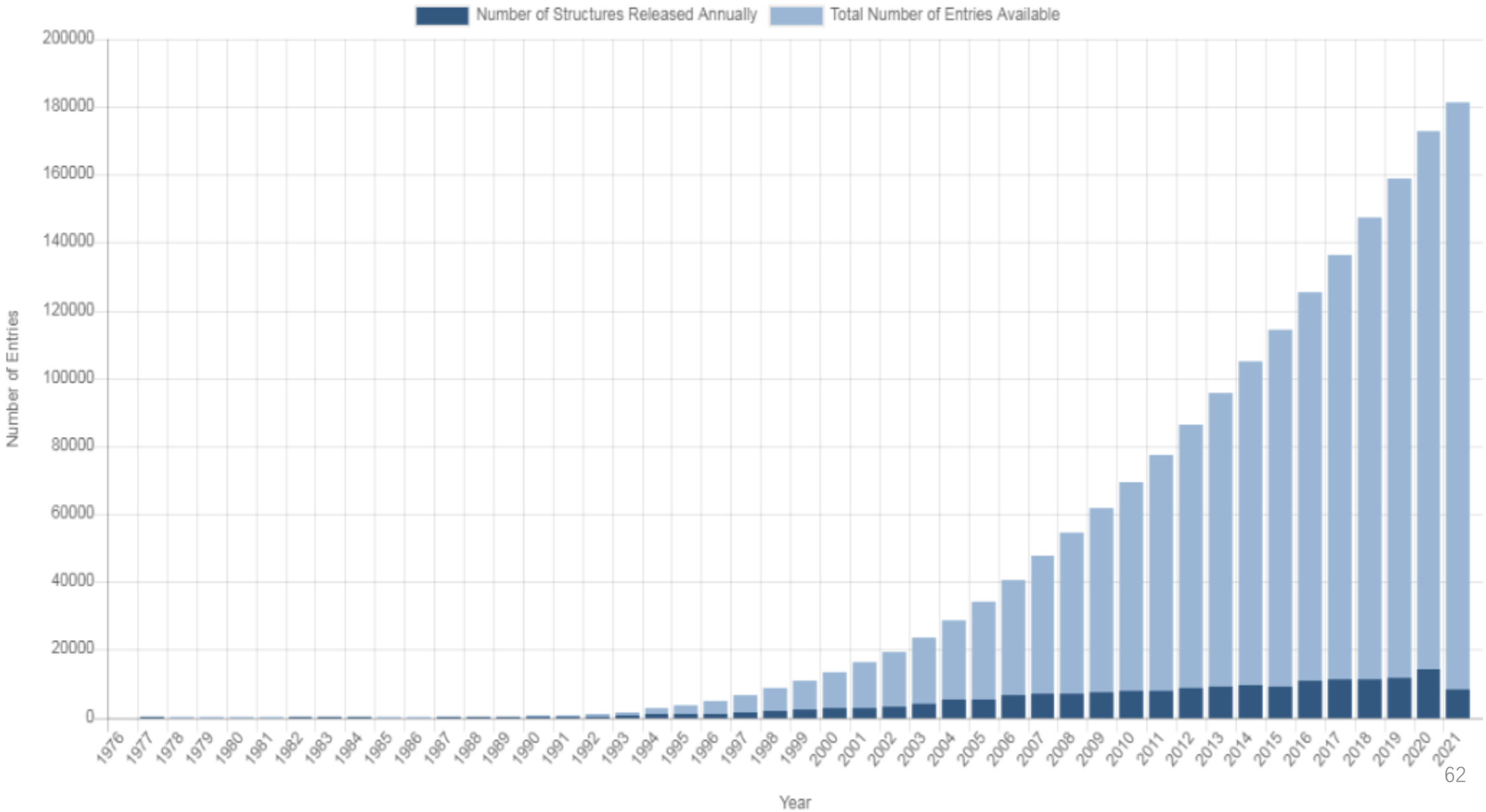
wwPDB Validation

3D Report | Full Report



PDB Statistics: Overall Growth of Released Structures Per Year

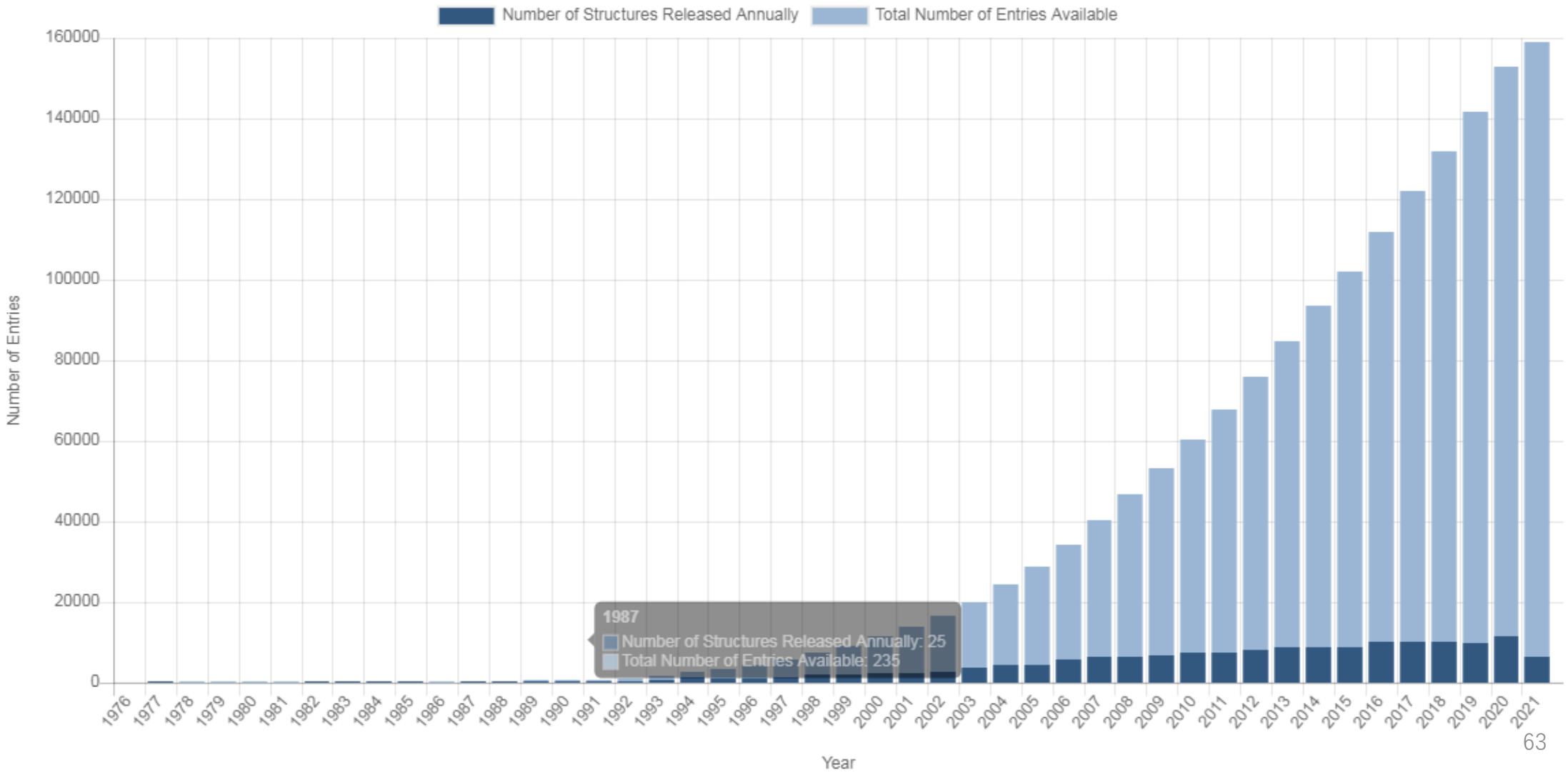
Other Statistics



Other Statistics

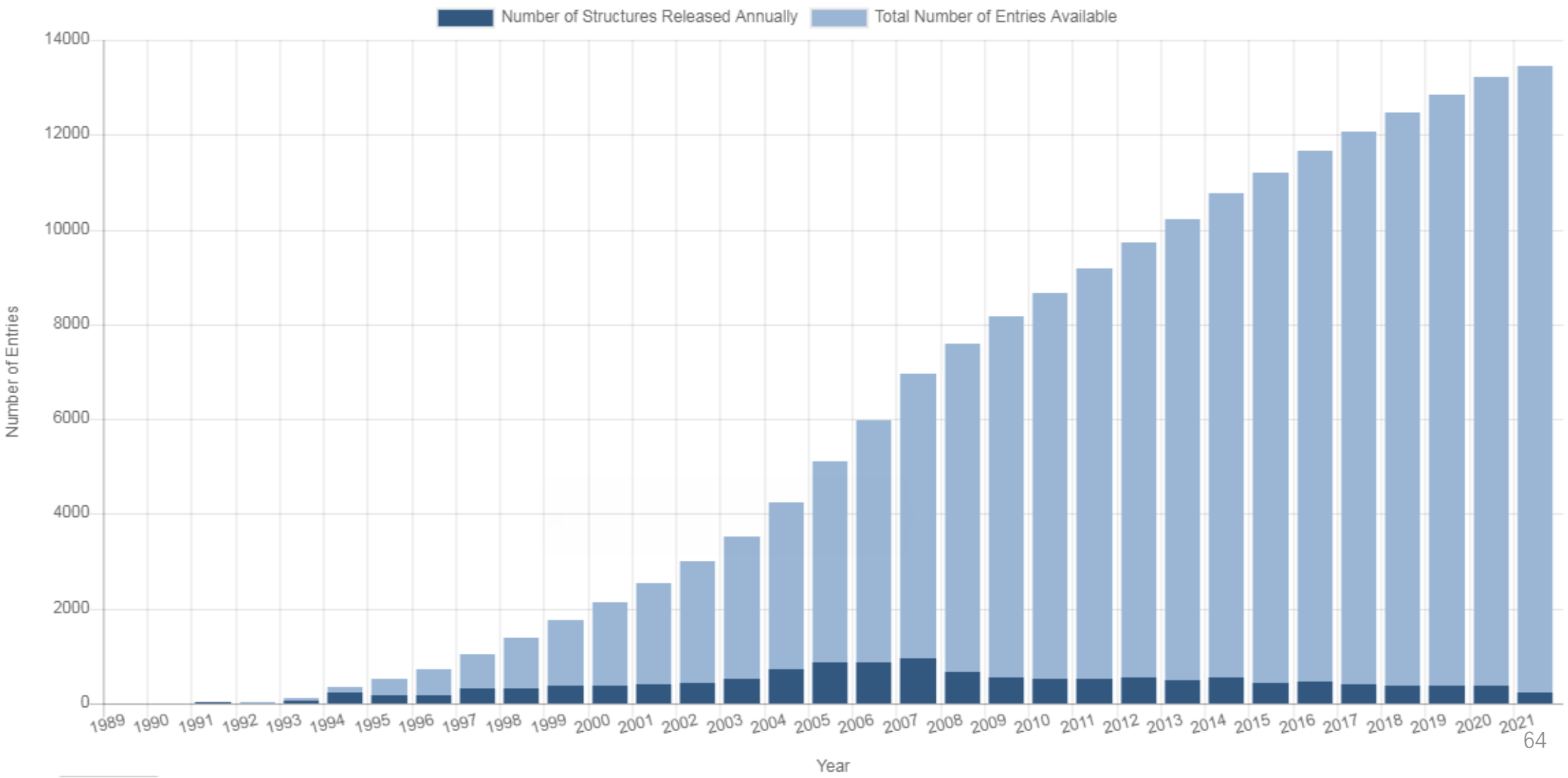
PDB Statistics: Growth of Structures from X-ray Crystallography Experiments Released per Year

Experimental methods such as X-ray crystallography, NMR spectroscopy, and 3D electron microscopy are used to determine the location of each atom relative to each other in the molecule.



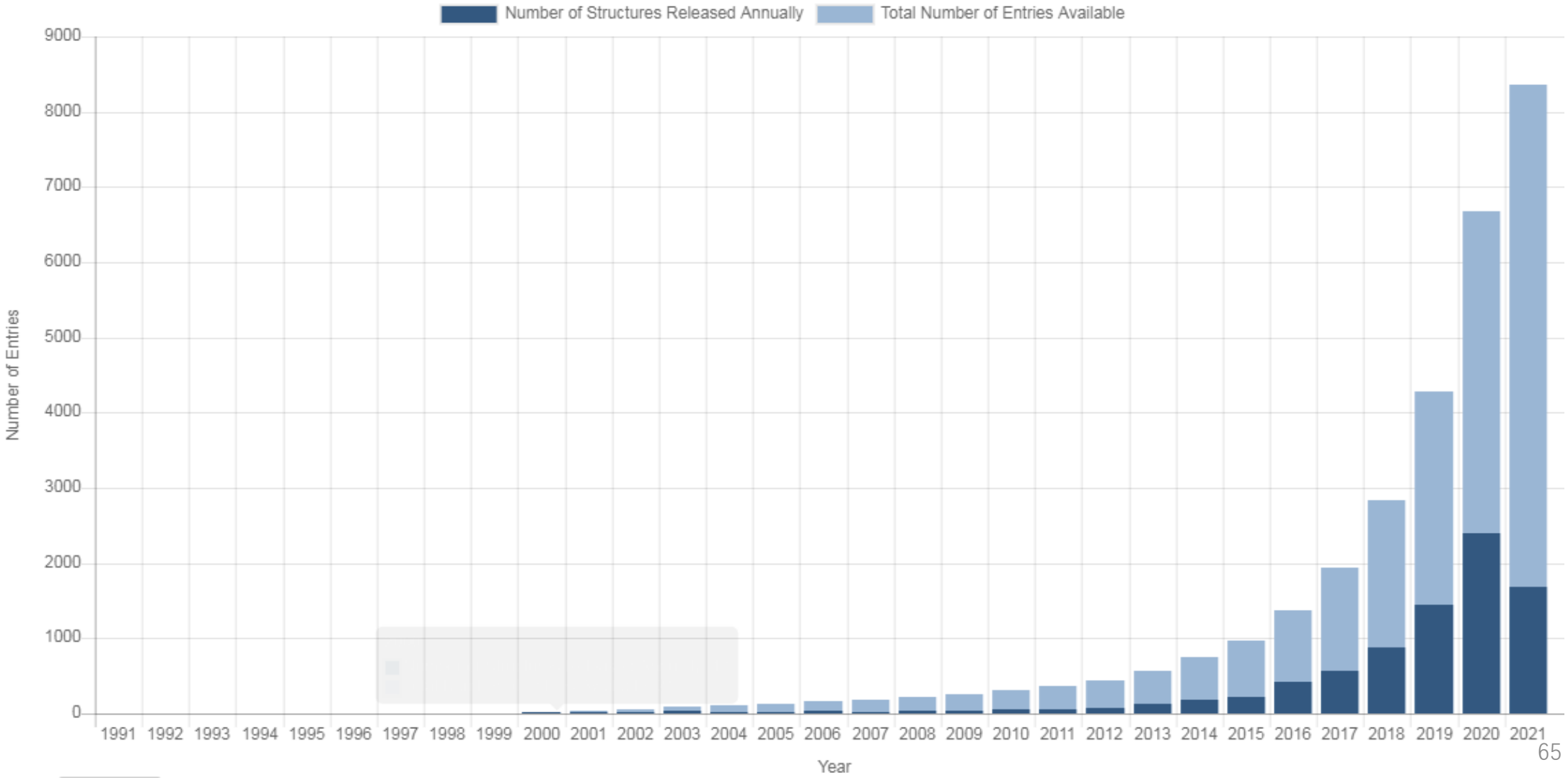
PDB Statistics: Growth of Structures from NMR Experiments Released per Year

Experimental methods such as X-ray crystallography, NMR spectroscopy, and 3D electron microscopy are used to determine the location of each atom relative to each other in the molecule.



PDB Statistics: Growth of Structures from 3DEM Experiments Released per Year

Experimental methods such as X-ray crystallography, NMR spectroscopy, and 3D electron microscopy are used to determine the location of each atom relative to each other in the molecule.



Header

Atomic coordinate

structure annotation

amino acid field

cofactor filed

```
HEADER    LYASE (CARBON-CARBON)                    03-JUL-95   1DNP
TITLE     STRUCTURE OF DEOXYRIBODIPYRIMIDINE PHOTOLYASE
... ...
SOURCE   2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI
KEYWDS    DNA REPAIR, ELECTRON TRANSFER, EXCITATION ENERGY TRANSFER,
KEYWDS   2 LYASE, CARBON-CARBON
... ...
ATOM     21  ND1 HIS A   3      55.365  27.866  62.971  1.00 11.07          N
ATOM     22  CD2 HIS A   3      57.200  28.354  61.894  1.00 13.12          C
ATOM     23  CE1 HIS A   3      56.124  26.783  62.981  1.00 13.03          C
ATOM     24  NE2 HIS A   3      57.243  27.052  62.334  1.00  8.19          N
ATOM     25  N   LEU A   4      55.580  32.694  59.656  1.00 12.61          N
ATOM     26  CA  LEU A   4      54.799  33.803  59.113  1.00 11.56          C
ATOM     27  C   LEU A   4      53.552  33.269  58.374  1.00  7.76          C
ATOM     28  O   LEU A   4      53.650  32.363  57.532  1.00  6.99          O
ATOM     29  CB  LEU A   4      55.656  34.683  58.174  1.00  9.03          C
ATOM     30  CG  LEU A   4      54.946  35.887  57.518  1.00  2.00          C
ATOM     31  CD1 LEU A   4      54.623  36.920  58.550  1.00  6.21          C
... ...
HETATM 7641 AN7  FAD B 472      27.855  78.556  29.073  1.00  4.55          N
HETATM 7642 AC5  FAD B 472      28.524  78.026  27.955  1.00  2.00          C
HETATM 7643 AC6  FAD B 472      29.848  77.609  27.724  1.00  3.40          C
HETATM 7644 AN6  FAD B 472      30.787  77.757  28.664  1.00  6.22          N
```

atom number

atom name

residue name

polypeptide chain identifier

residue number

x, y, z coordinates

occupancy

temperature factor

atom type

PDB format

1. Header
   - provides an overview of the protein and the quality of the structure
   - **PDBid** - consisting of four characters of either letters A to Z or digits 0 to 9 such as 1LYZ and 4RCR
   - name of the molecule
   - source organism
   - Bibliographic reference
   - methods of structure determination
   - resolution
   - crystallographic parameters
   - protein sequence
   - secondary structure information

PDB format

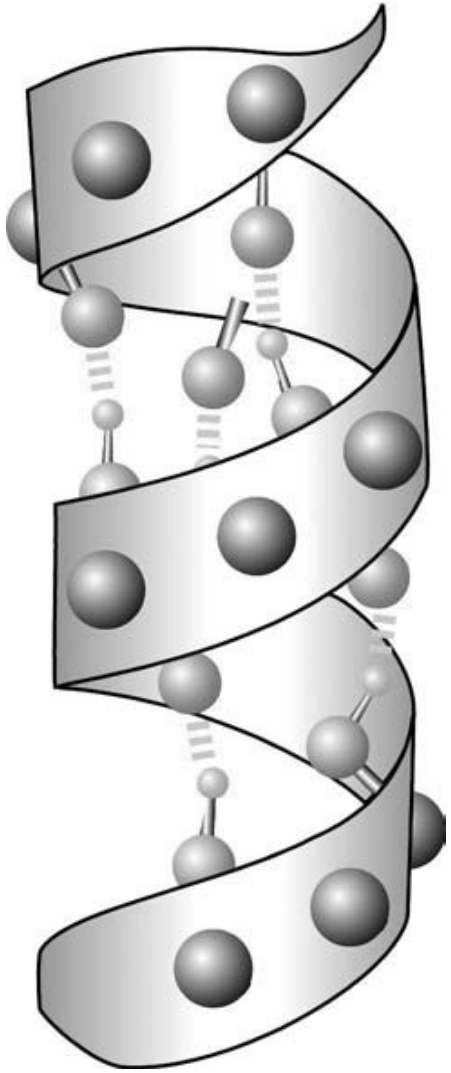
2. Atomic coordinate
   - atom number
   - atom name
   - residue name
   - residue number
   - polypeptide chain identifier
   - x, y, and z Cartesian coordinates

Protein secondary structure

- **Protein secondary structures**
  - stable **local** conformations of a polypeptide chain
  - critically important in maintaining a protein three-dimensional structure
  - chief elements of secondary structures are α-helices and β-sheets

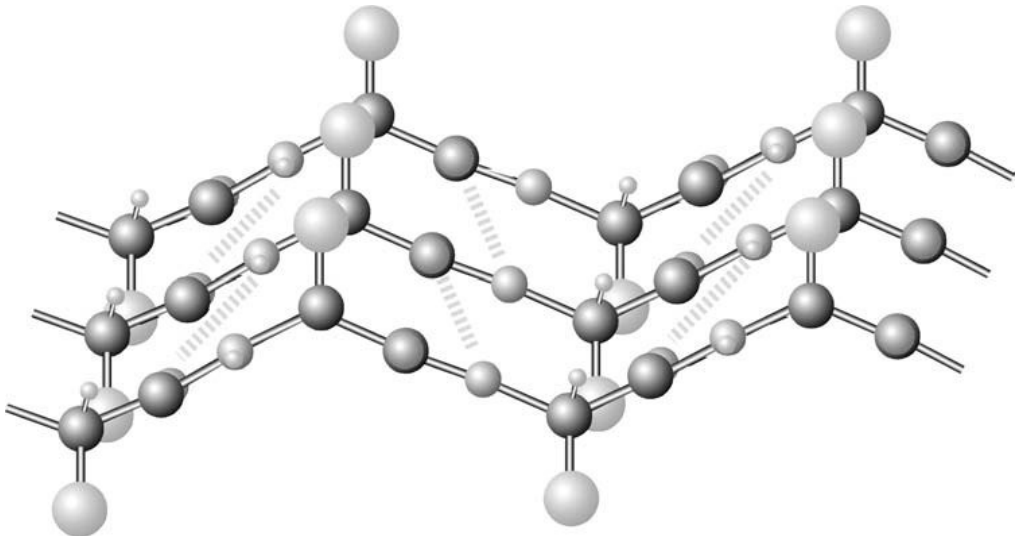
Secondary structural elements

- α-helices
  - a spiral-like structure with 3.6 amino acid residues per helical turn
  - the structure is stabilized by hydrogen bonds between residues i and i+4
  - nearly all known α-helices are right handed, exhibiting a rightward spiral form



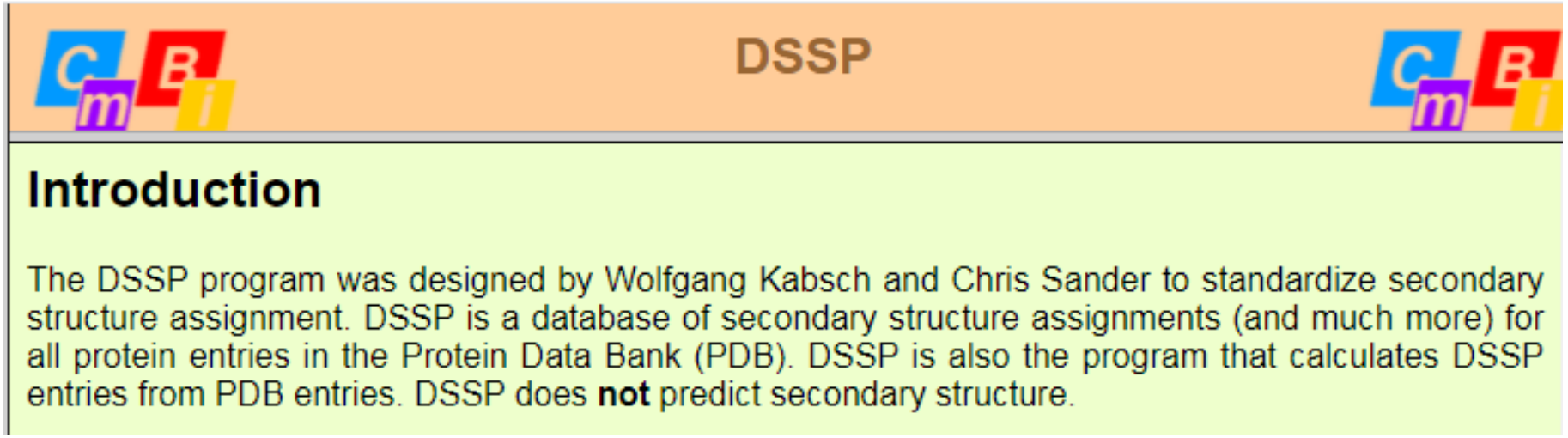
Secondary structural elements

- β-sheets
  - consists of two or more β-strands having an extended zigzag conformation
  - each region involved in forming the β-sheet is a β-strand
  - the structure is stabilized by hydrogen bonding between residues of adjacent strands



DSSP

- A database of secondary structure assignments
- Definition of secondary structure of proteins given a set of 3D coordinates
- Source of protein structure: PDB
- Software is available from http://www.embl-heidelberg.de/dssp/



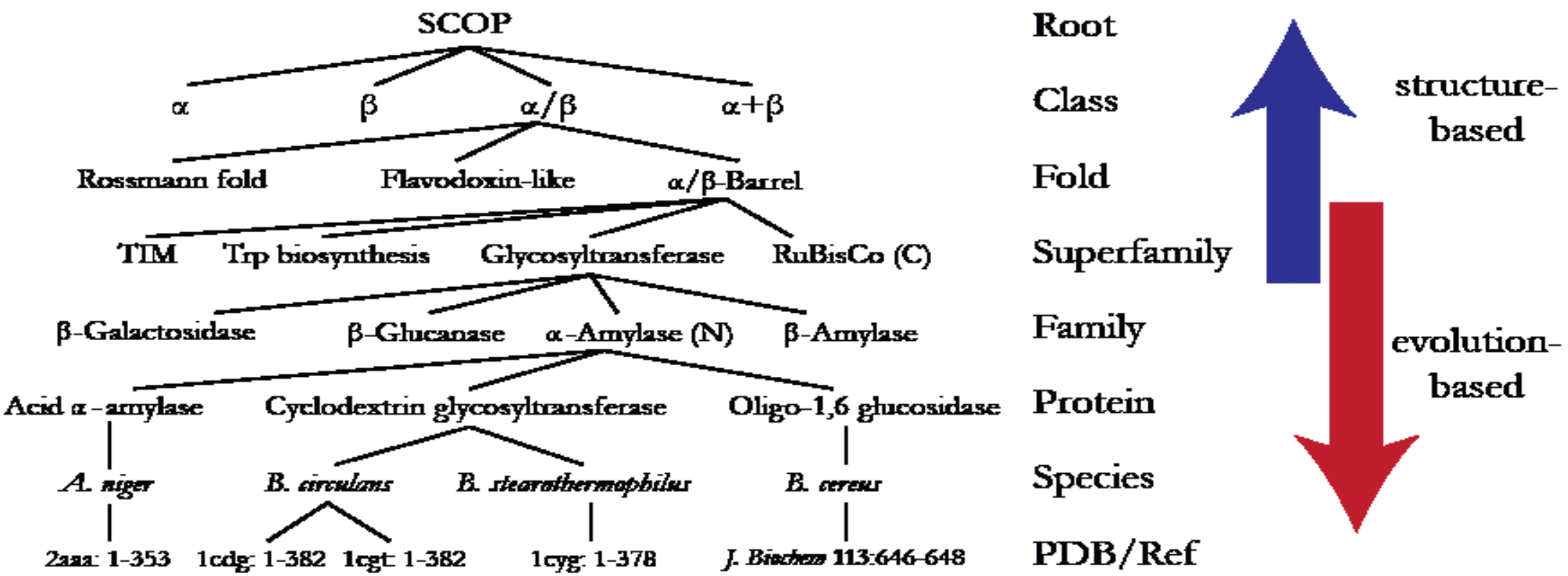
Protein structure classification

- One of the applications of protein structure comparison is structural classification.
- The reason to develop a protein structure classification system is to establish hierarchical relationships among protein structures and to provide a comprehensive and evolutionary view of known structures.
- Once a hierarchical classification system is established, a newly obtained protein structure can find its place in a proper category. As a result, its functions can be better understood based on association with other proteins.

Structural classification of proteins (SCOP)

- Created in 1994
- Source of protein structure: PDB
- Describe structural and evolutionary relationship between proteins of known structures
- It is constructed almost entirely based on manual comparison of structures by human experts.
  - Provides a relatively convincing structural classification system
  - Manual curation makes the classification more subjective, as the exact boundaries between levels and groups are sometimes arbitrary.

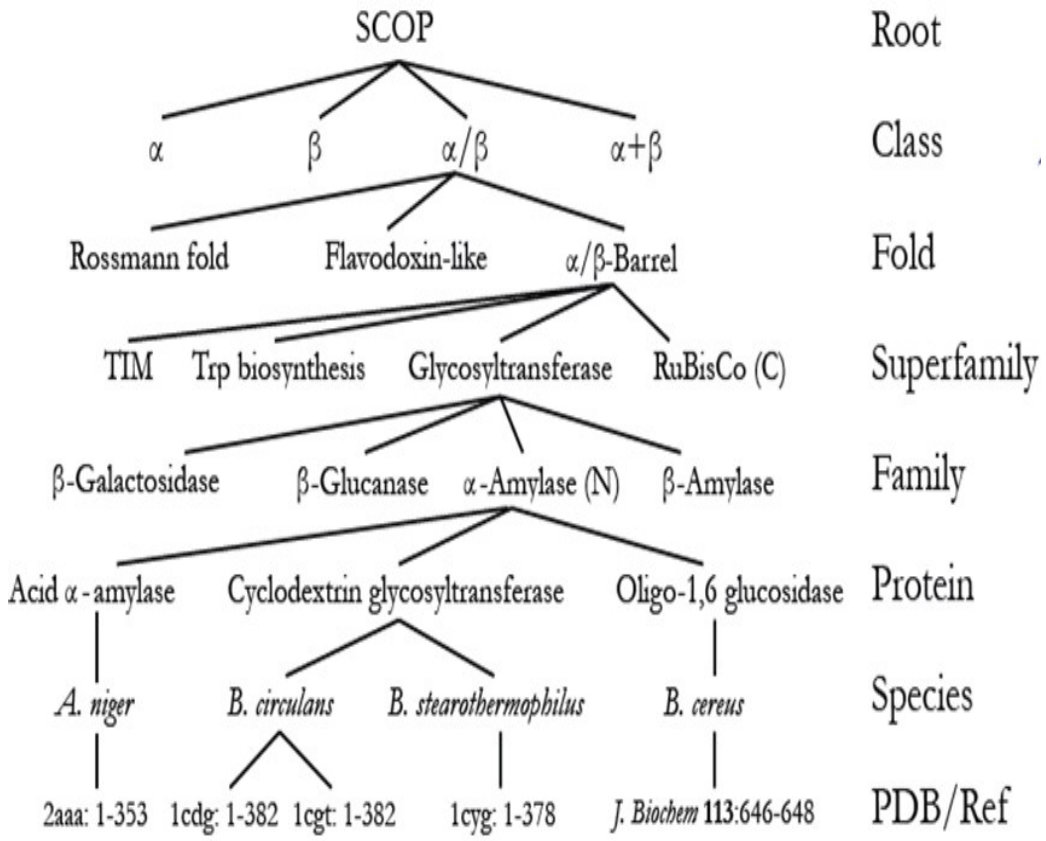
SCOP hierarchy



- The proteins are grouped into hierarchies of classes, folds, superfamilies, and families.
- From superfamily to upper levels, the hierarchy classifies groups of proteins by structure compositions.
- From family to lower levels, the hierarchy classifies groups of proteins by evolutionary relationships.

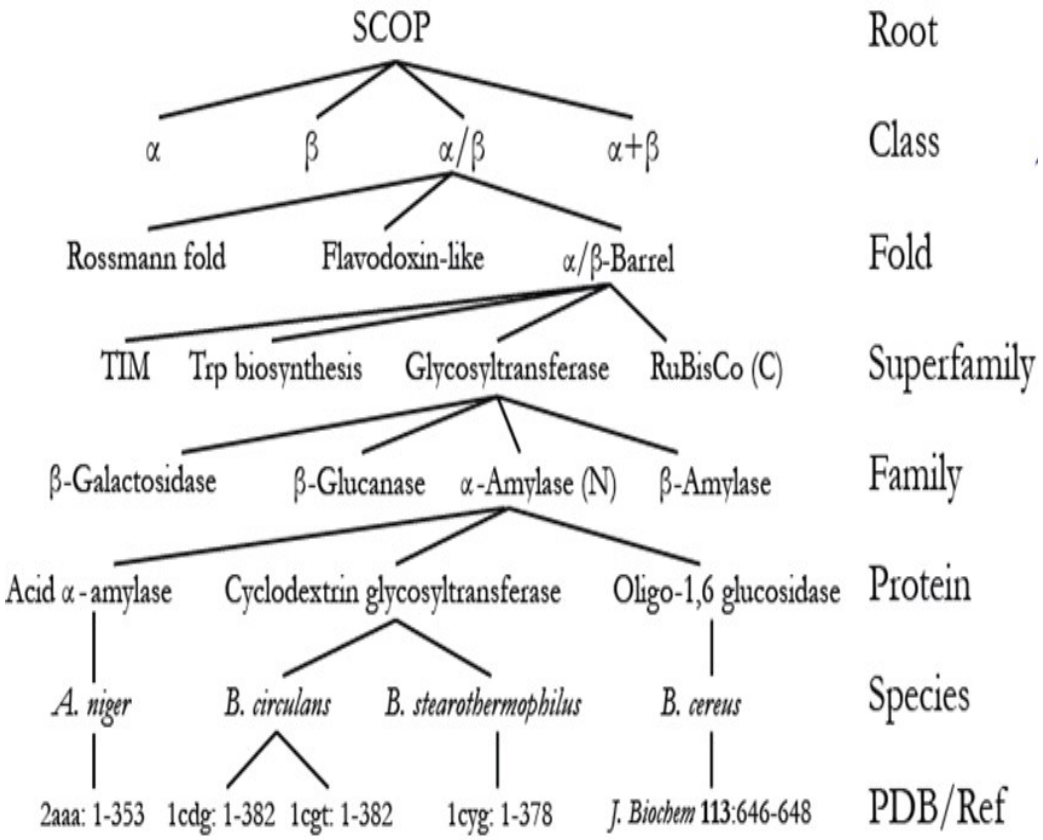
Hierarchical classification scheme

- Families - consist of proteins having high sequence identity (>30%)
  - clearly share close evolutionary relationships
  - normally have the same functionality
  - extremely similar protein structures
- Superfamilies - consist of families with similar structures, but weak sequence similarity
  - share a common ancestral origin, although the relationships between families are considered distant.



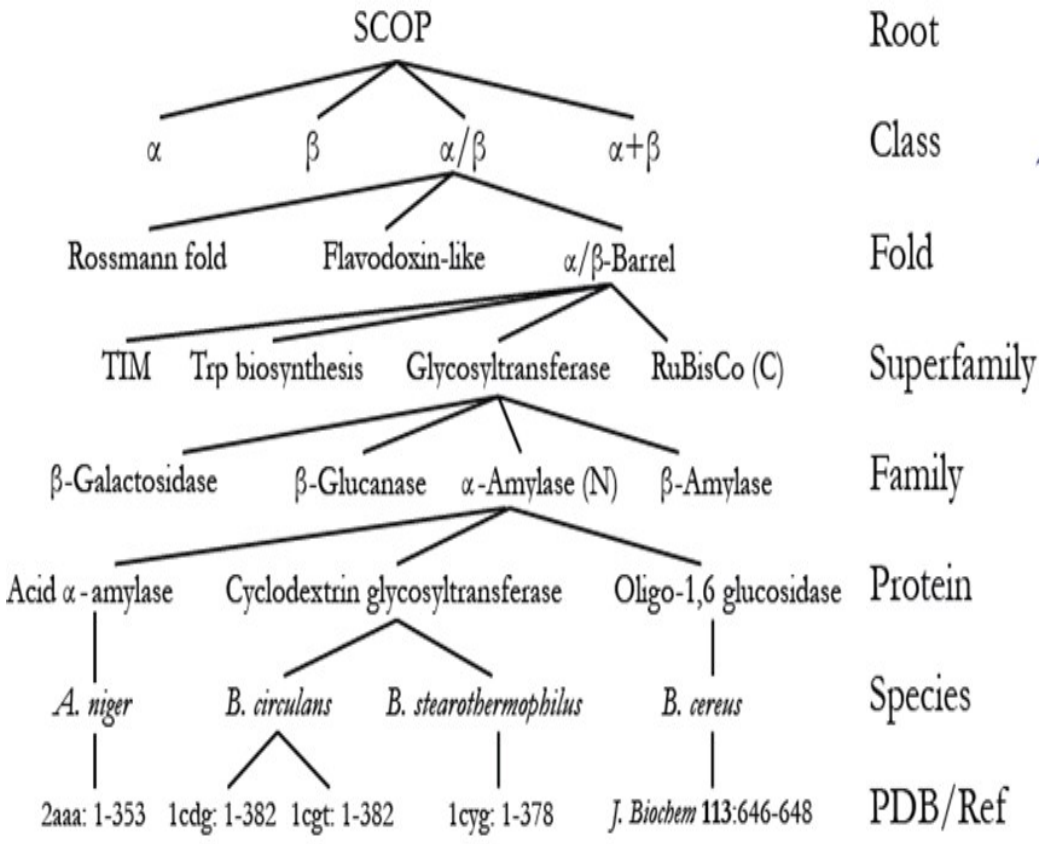
Hierarchical classification scheme

- Folds - consist of superfamilies with a common core structure
  - members within the same fold have similar overall secondary structures with similar orientation
  - may not have evolutionary relationships



Hierarchical classification scheme

- Classes – consist of folds with similar core structures
  - at the highest level of the hierarchy
  - distinguishes groups of proteins by secondary structure compositions such as all α, all β, α and β, and so on
  - Folds within the same class are essentially randomly related in evolution



Gene expression database

- Repositories for gene expression data
  - Microarry, RNAseq, single-cell RNA-seq
- Measure levels of mRNA under certain condition
- Gene Expression Omnibus (GEO) – NCBI
- ArrayExpress - EBI

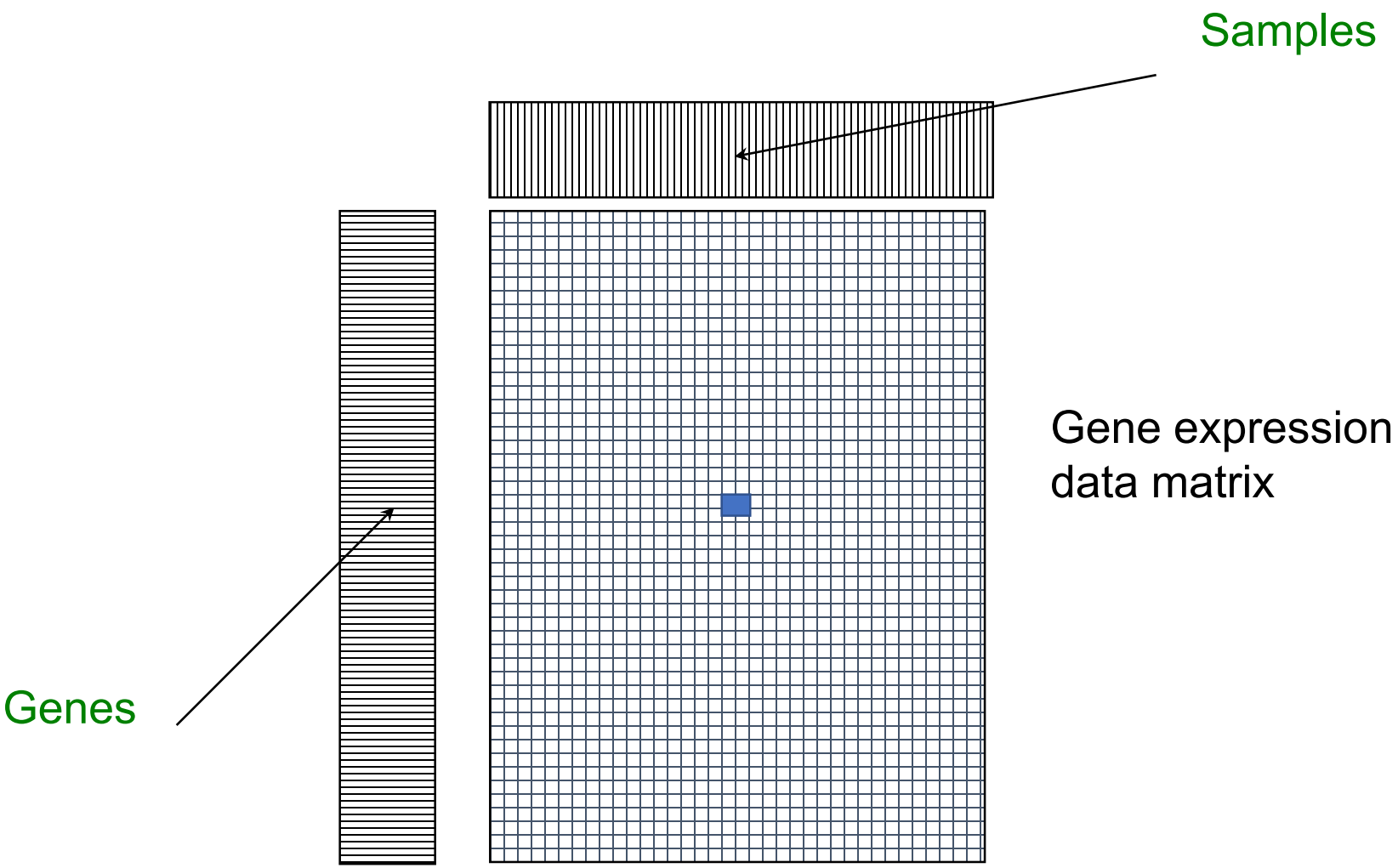
Where does the data come from

- Transparency/reproducibility of publications
  - Journals require research data to be released in published form for use by others
  - Databases offer single resource and standardized access
- Data was generated for a specific purpose, but is not limited to that purpose
  - Can be reanalyzed in a different context
  - Can be combined with other datasets
  - Can be used as independent validation

What is in a gene expression database

- Gene expression data in different forms:
  - Resolution:
    - Gene level
    - Transcript level
    - Exon level
  - Comprehensiveness:
    - Targeted arrays
    - Whole genome arrays
  - Different platforms (microarrays, RNAseq, scRNA-seq)
- Generally only gene expression, may have limited sample information

Gene expression matrix



The Cancer Genome Atlas (TCGA)

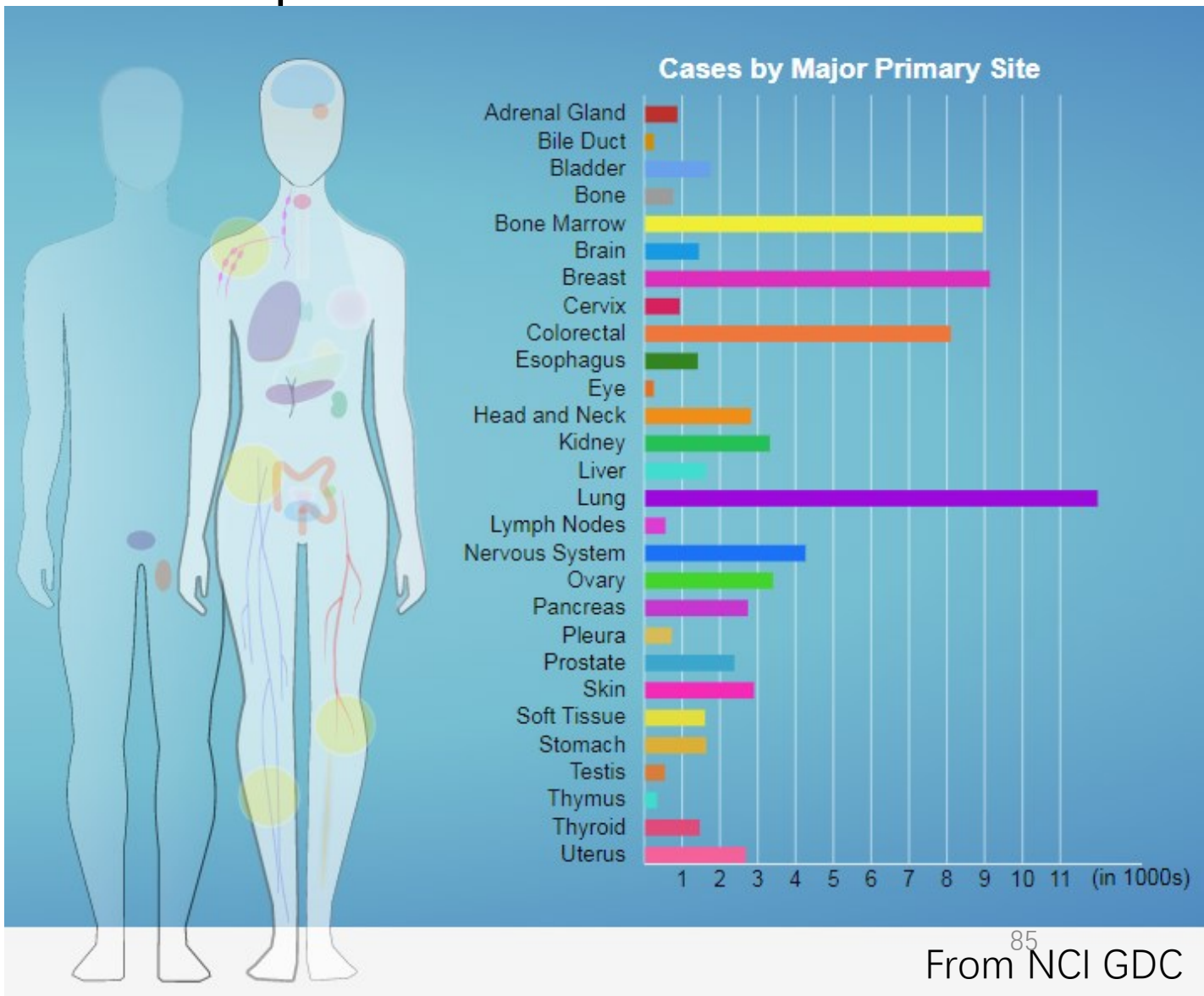
- A publicly accessible atlas of cancer related data from National Cancer Institute (NCI) and National Human Genome Research Institute.
- Phase I: initiated in 2006 to catalog genetic mutations causing cancer, using genome sequencing; focused on tumors having poor prognosis: GBM (glioblastoma multiforme), lung and ovarian cancer
- Phase II: transition in 2009, expanded to 20-25 different cancer types, involved complement genome sequencing with genomic characterization, including gene expression profiling, copy number variation, DNA methylation, miRNA profiles
- For a decade, TCGA sequenced and molecularly characterized over 11,000 patients and 33 cancer types

Selected cancers based on specific criteria

- Poor prognosis
- Overall public health impact
- Availability of samples meeting standards for patient consent
- Availability of samples meeting standards for quality and quantity that include:
  - Primary, untreated tumor with a source of matched normal tissue or blood sample
  - Frozen, sufficiently sized, resection samples
  - Samples composed of at least 80% tumor nuclei (threshold later lowered to 60% with improved sequencing technology and computational methods)
- With support from patients, patient advocacy groups, and doctors, many rare cancers were also included

Cancer measured at multiple scales

- mRNA & miRNA expression
- Copy number
- DNA Methylation
- Mutation (NGS)
- Pathology images
- Medical Images
- Treatment
- Survival Outcome
- ……



From NCI GDC

Retrieval system for biological databases

- Databases are required to provide efficient and user-friendly access to the data stored.
- Retrieval systems provide access to multiple databases for retrieval of integrated search results through cross-referencing links.
- Users do not have to visit multiple databases located in disparate places.

Entrez

- Entrez – developed and maintained by NCBI
- Entrez is a molecular biology database system that provides integrated access to over 20 databases:
  - protein sequence data from PIR-International, PRF, Swiss-Prot, and PDB
  - nucleotide sequence data from GenBank that includes information from EMBL and DDBJ
  - 3D structure data
  - Citations and abstracts, full papers from PubMed MEDLINE

  ……

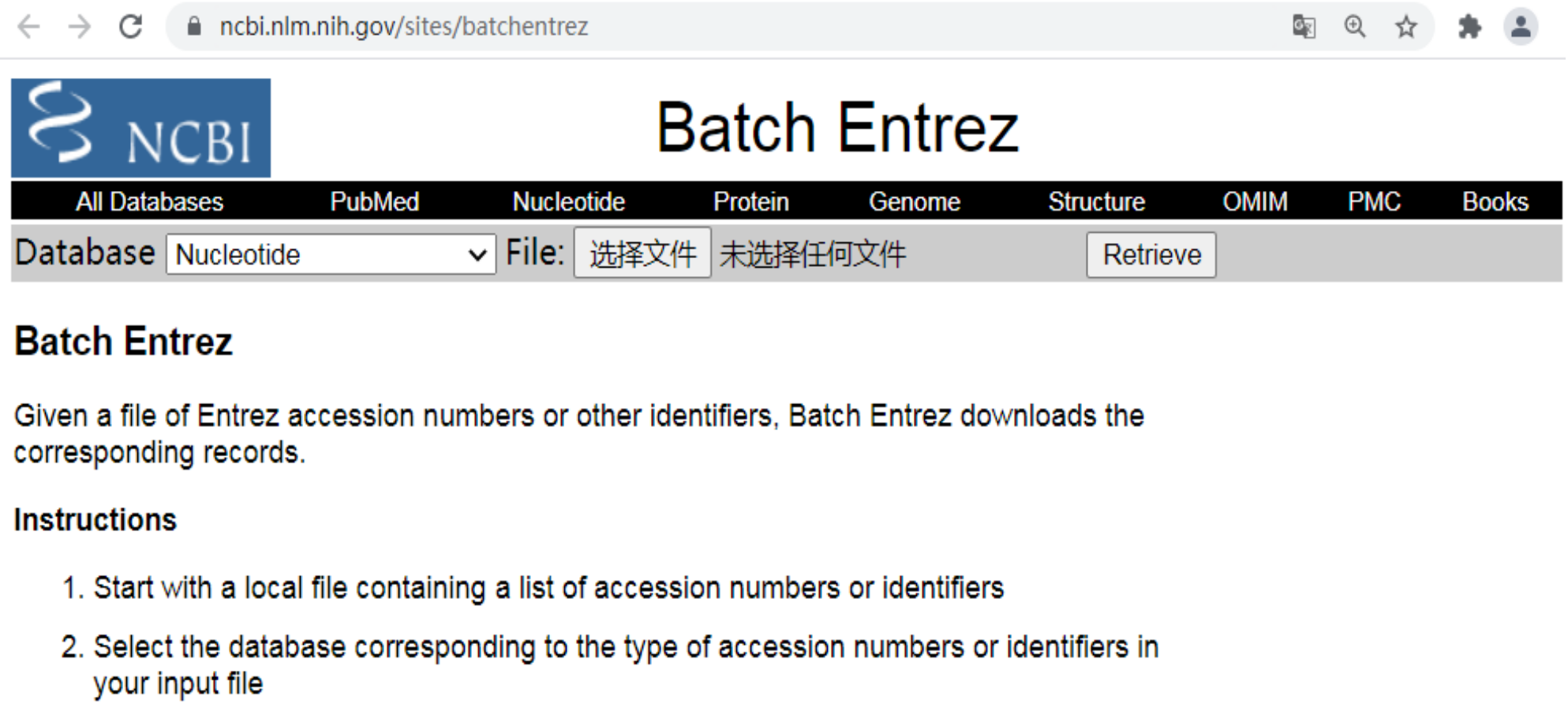
Entrez

- The key feature of Entrez is its ability to integrate information, which comes from cross-referencing between NCBI databases based on preexisting and logical relationships between individual entries.

  For example, in a nucleotide sequence page, one may find cross-referencing links to the translated protein sequence, genome mapping data, or to the related PubMed literature information, and to protein structures if available.

Batch Entrez

- Allow Batch downloads of large search results.



Review

Essential Bioinformatics:

- Chapter one
- Chapter two