Databases

Biological data accumulates at a rapid rate (See Figure 1). According to GOLD database https://gold.jgi.doe.gov/, as of 2019, there are $> 16,000$ genomes sequenced, mostly bacterial. $> 5x10^6$ unique sequences are available. Experiment is still lagging behind... We have all this data, and as data scientists, there are many things we can do with it:

1. Compare them to find what is common and different among organisms (Comparative Genomics)

2. Find out how and which genes encode for which proteins

3. Identify changes that lead to disease

4. Associate structural and functional information with new gene sequences

What We Expect From a Biological Databases

Biological databases are collections of scientific data containing information about biological sequences, structures, function and systems. The data is gathered from scientific experiments, published literature and computational analysis. A useful database should, at the least, provide the following:

- Sequence and/or functional and/or structural information, as well as related bibliography and documentation.

- It should be well structured and indexed to allow efficient storage and search.

- Well cross-referenced (with other databases)

- It should be periodically updated and maintained

- Provide tools for data analysis and visualization, or at least formatted in a compatible way with known tools

Database Access: Most databases are accessible to the public free of charge. Everyone can read the data but very few users can make changes to it. Users can submit information to the database, but direct modification of the database is reserved to the institutes maintaining the database.

The organization of a database may be hierarchical, where the data is classified and organized at multiple levels. SCOP and CATH are examples of such databases. There are many databases available, and it is not always easy to keep track and ensure consistency of the data presented in them. Most databases limit themselves to a certain subject. Here are some of the major databases:

Nucleotide Sequence Databases

The International Nucleotide Sequence Database Collaboration (INSDC) http://www.insdc.org/ is a collaborative effort operated by the following organizations:

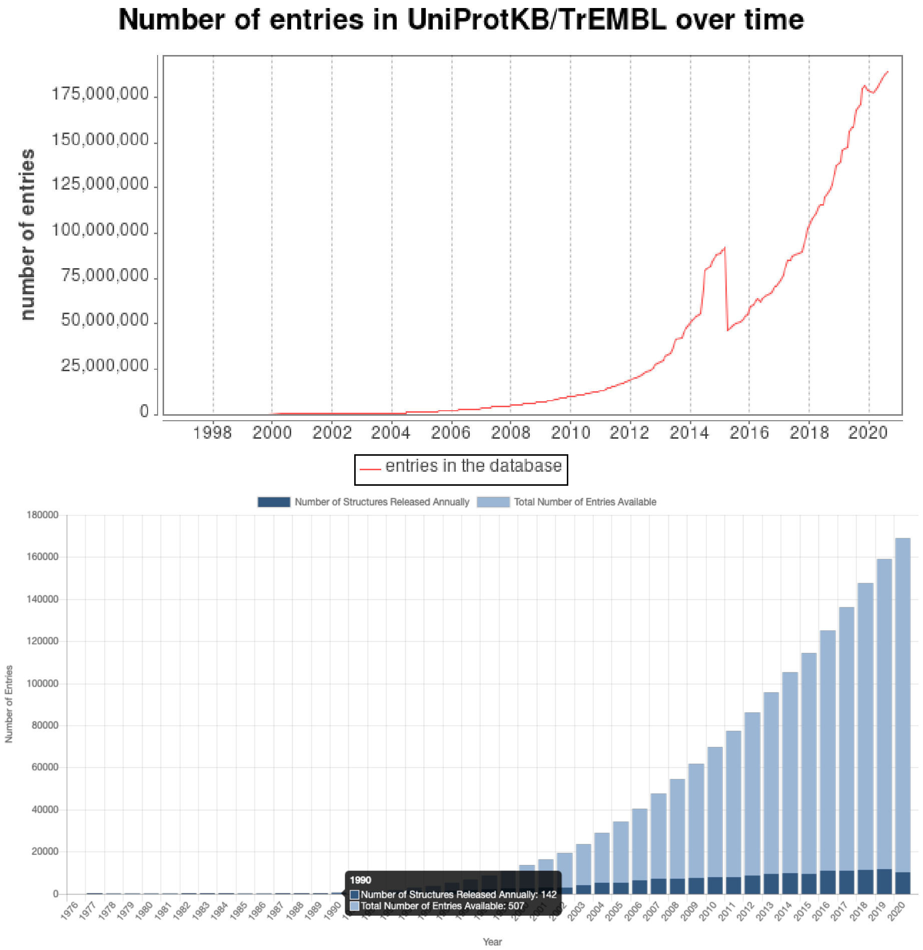


Figure 1: Growth of the Uniprot (top) and PDB (bottom).

- NCBI (National Center for Biotechnology Information): http://ncbi.nih.gov, which maintains the GenBank database, an annotated collection of all publicly available DNA sequences.
- EMBL-EBI (European Molecular Biology Laboratory, European Bioinformatics Institute): https://www.ebi.ac.uk/, which operates ENA, the European Nucleotide Archive.
- DDBJ (DNA Data Bank of Japan): http://www.ddbj.nig.ac.jp/

INSDC covers the spectrum of nucleotide sequences, from data raw reads, though alignments and assemblies to functional annotation, enriched with contextual information relating to samples and experimental configurations. Each one of the three individual databases provides tools for data submission, retrieval and analysis.

Protein Sequence Databases

UniprotKB/SwissProt, http://www.uniprot.org/ is a universal resource, resulting from a merger of several databases. It is a manually annotated, non-redundant protein sequence database. The UniProt databases are the UniProt Knowledge base (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc). UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR). The manually annotated and reviewed entries (the Swiss-Prot database), contain over 500,000 entries as of 2019. TrEMBL (Translated EMBL Nucleotide Sequence Data Library) entries are computationally analyzed and not reviewed. They are awaiting manual annotation. As of 2019 TrEMBL more than 146,000,000 entries.

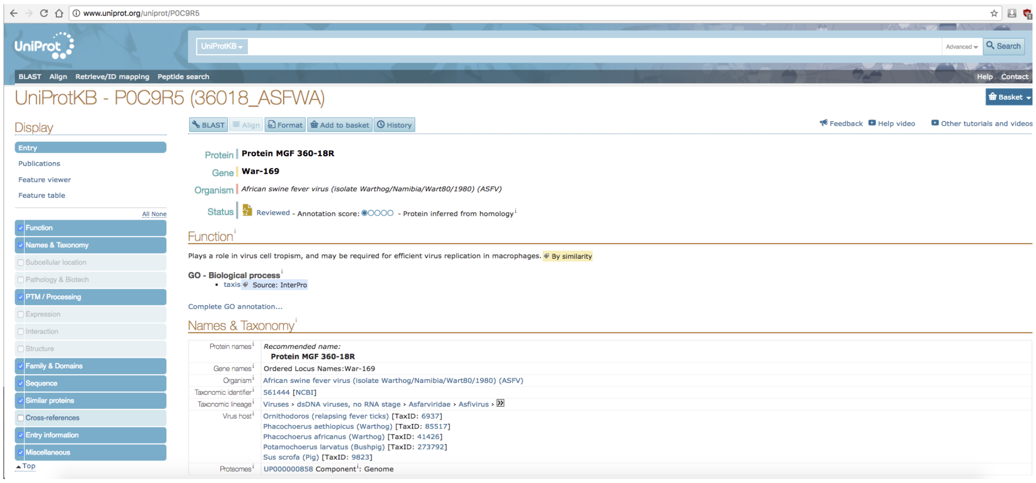
UniprotKB provides several analysis tools: BLAST, align, Retrieve/IDmapping and peptide search.

Figure 2 shows an example of a Uniprot entry. Figure 3 shows a Uniprot search example. The search terms are polymerase from human (H. Sapiens) and only reviewed entries. The boxes on the left of each entry can be checked and the selected sequence(s) can be search with BLAST, aligned and saved.

The Pfam database – http://pfam.xfam.org/ is a database of protein families based on conserved regions.

1 The Protein Databank (PDB)

Now that we have reviewed the basics of protein chemistry, let us turn our attention to the tools. The most important source of information about protein structure is the Protein Databank (PDB), maintained by the Research Collaboratory for Structural Bioinformatics (RCSB). In addition to being an entry point to the structural data itself, the PDB web site, http://www.rcsb.org/pdb, contains links to many tools database you can apply to individual protein structures as you search the database. Information from the database is made available through the Protein Structure Explorer interface. For each protein, you can view the molecular structure using 3D display tools such as JMol and the Java QuickPDB viewer. PDB files and file headers can be viewed as HTML and downloaded in a variety of formats. Links to the protein structure classification databases CATH, FSSP, and SCOP are provided, along with the tools CE (Combinatorial Extension) and VAST (Vector Alignment Search Tool), which search for structures based on structural alignment. Average geometric properties, including dihedral angles, bond angles, and bond lengths can be



UniProtKB - P0C9R5 (36018_ASFWA)

Protein | Protein MGF 360-18R

Gene | War-169

Organism | *African swine fever virus (isolate Warthog/Namibia/Wart80/1980) (ASFV)*

Status | Reviewed - Annotation score: ●○○○○ - Protein inferred from homology

Function

Plays a role in virus cell tropism, and may be required for efficient virus replication in macrophages. By similarity

GO - Biological process

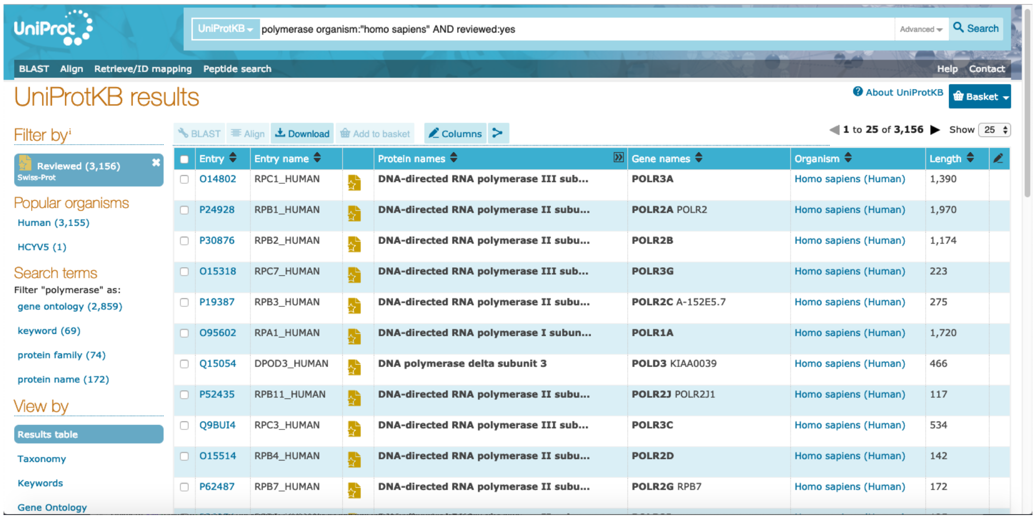
- taxis Source: InterPro

Complete GO annotation...

Names & Taxonomy

Protein names	*Recommended name:* **Protein MGF 360-18R**
Gene names	Ordered Locus Names:War-169
Organism	African swine fever virus (isolate Warthog/Namibia/Wart80/1980) (ASFV)
Taxonomic identifier	561444 [NCBI]
Taxonomic lineage	Viruses › dsDNA viruses, no RNA stage › Asfarviridae › Asfivirus ›
Virus host	Ornithodoros (relapsing fever ticks) [TaxID: 6937] Phacochoerus aethiopicus (Warthog) [TaxID: 85517] Phacochoerus africanus (Warthog) [TaxID: 41426] Potamochoerus larvatus (Bushpig) [TaxID: 273792] Sus scrofa (Pig) [TaxID: 9823]
Proteomes	UP000000858 Component: Genome

Figure 2: An example of a UniprotKB entry



polymerase organism:"homo sapiens" AND reviewed:yes

UniProtKB results

Filter by

Reviewed (3,156)
Swiss-Prot

Popular organisms

Human (3,155)

HCYV5 (1)

Search terms

Filter "polymerase" as:

gene ontology (2,859)

keyword (69)

protein family (74)

protein name (172)

View by

Results table

Taxonomy

Keywords

Gene Ontology

1 to 25 of 3,156

□	Entry	Entry name		Protein names	Gene names	Organism	Length
□	O14802	RPC1_HUMAN		DNA-directed RNA polymerase III sub...	POLR3A	Homo sapiens (Human)	1,390
□	P24928	RPB1_HUMAN		DNA-directed RNA polymerase II subu...	POLR2A POLR2	Homo sapiens (Human)	1,970
□	P30876	RPB2_HUMAN		DNA-directed RNA polymerase II subu...	POLR2B	Homo sapiens (Human)	1,174
□	O15318	RPC7_HUMAN		DNA-directed RNA polymerase III sub...	POLR3G	Homo sapiens (Human)	223
□	P19387	RPB3_HUMAN		DNA-directed RNA polymerase II subu...	POLR2C A-152E5.7	Homo sapiens (Human)	275
□	O95602	RPA1_HUMAN		DNA-directed RNA polymerase I subun...	POLR1A	Homo sapiens (Human)	1,720
□	Q15054	DPOD3_HUMAN		DNA polymerase delta subunit 3	POLD3 KIAA0039	Homo sapiens (Human)	466
□	P52435	RPB11_HUMAN		DNA-directed RNA polymerase II subu...	POLR2J POLR2J1	Homo sapiens (Human)	117
□	Q9BUI4	RPC3_HUMAN		DNA-directed RNA polymerase III sub...	POLR3C	Homo sapiens (Human)	534
□	O15514	RPB4_HUMAN		DNA-directed RNA polymerase II subu...	POLR2D	Homo sapiens (Human)	142
□	P62487	RPB7_HUMAN		DNA-directed RNA polymerase II subu...	POLR2G RPB7	Homo sapiens (Human)	172

Figure 3: An example of a UniprotKB search. The search terms are polymerase from human (H. Sapiens) and only reviewed entries.

displayed in tabular format with extremes and deviations noted. Sequences can be viewed and labeled according to secondary structure, and sequence information downloaded in FASTA format. You can go directly to the page for a particular protein of interest by entering that protein's four-letter PDB code in the Explore box on the PDB's main page. The PDB can also be searched using two different search tools, SearchLite and SearchFields. SearchLite is a simple search tool that allows you to enter one or more search terms separated by boolean operators into a single search field. SearchFields is a tool for advanced searches that provides a customizable search form that allows you to use separate keywords to search each PDB header field.

SearchFields supports options for searching a dozen of the most important fields in the PDB header, as well as crystallographic information. SearchFields also allows the database to be searched using FASTA for sequence comparison, as well as secondary structure features or short sequence features. From the individual protein page generated by the Structure Explorer, the PDB provides a menu of links through which to connect to other tools. These features are still evolving rapidly. Table 9-2 provides a brief overview of the PDB protein page. We also encourage you to explore the PDB site regularly if you are interested in tools for protein structure analysis.

1.1 The PDB file format

Every PDB entry contains a header with information about the molecule. This information includes:

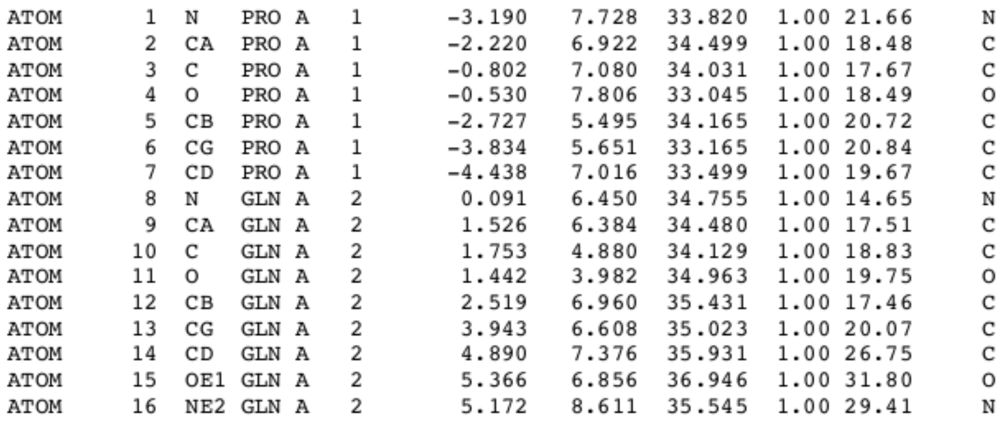
- The name of the protein and what species it came from.
- Details about how the structure was determined – X-ray crystallography, NMR etc., the resolution and other experimental and chemical details.
- A literature reference.
- The amino acid sequence and secondary structure information.
- Disulphide bridges.

The header is followed by the structural information itself. The most essential information for modeling a protein structure is the relative position of each atom, given as (x, y, z) Cartesian coordinates. Popular imaging methods such as X-Ray Crystallography, Nuclear Magnetic Resonance (NMR) and Cryogenic Electron Microscopy (Cryo-EM) are used to experimentally obtain relative atom positions from protein crystals and solutions. This is precisely the information provided by Protein Databank (PDB) format coordinate files. PDB format consists of lines of information in a text file. Each line of information in the file is called a record. A PDB file generally contains several different types of records, arranged in a specific order to describe a structure. At the top of the file is an optional header which contains information about the structure: The names of molecules, primary and secondary structure information, sequence database references, where appropriate, and ligand and biological assembly information, details about data collection and structure solution, and bibliographic citations. The header is followed by records that describe, line by line, the atomic coordinates of a protein molecule.

The atomic coordinates in a PDB file are shown in Figure 5. The meaning of each record is given in Figure 6

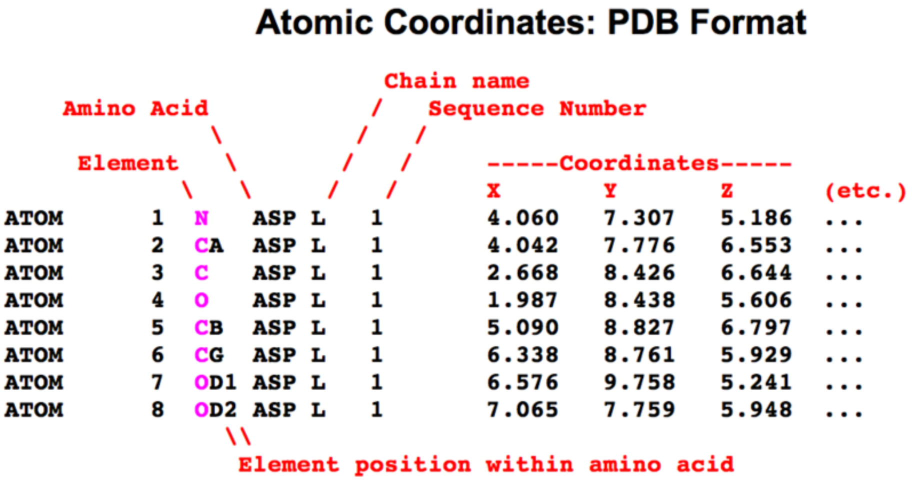
```
HEADER    CHROMOSOMAL PROTEIN                     02-JAN-87   1UBQ
TITLE     STRUCTURE OF UBIQUITIN REFINED AT 1.8 ANGSTROMS RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: UBIQUITIN;
COMPND   3 CHAIN: A;
COMPND   4 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGANISM_TAXID: 9606
KEYWDS    CHROMOSOMAL PROTEIN
EXPDTA    X-RAY DIFFRACTION
AUTHOR    S.VIJAY-KUMAR,C.E.BUGG,W.J.COOK
REVDAT   5   09-MAR-11 1UBQ    1       REMARK
REVDAT   4   24-FEB-09 1UBQ    1       VERSN
REVDAT   3   01-APR-03 1UBQ    1       JRNL
REVDAT   2   16-JUL-87 1UBQ    1       JRNL   REMARK
REVDAT   1   16-APR-87 1UBQ    0
JRNL        AUTH   S.VIJAY-KUMAR,C.E.BUGG,W.J.COOK
JRNL        TITL   STRUCTURE OF UBIQUITIN REFINED AT 1.8 A RESOLUTION.
JRNL        REF    J.MOL.BIOL.                   V. 194   531 1987
JRNL        REFN                   ISSN 0022-2836
JRNL        PMID   3041007
JRNL        DOI    10.1016/0022-2836(87)90679-6
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   S.VIJAY-KUMAR,C.E.BUGG,K.D.WILKINSON,R.D.VIERSTRA,
REMARK   1  AUTH 2 P.M.HATFIELD,W.J.COOK
REMARK   1  TITL   COMPARISON OF THE THREE-DIMENSIONAL STRUCTURES OF HUMAN,
REMARK   1  TITL 2 YEAST, AND OAT UBIQUITIN
REMARK   1  REF    J.BIOL.CHEM.                  V. 262  6396 1987
REMARK   1  REFN                   ISSN 0021-9258
REMARK   1 REFERENCE 2
REMARK   1  AUTH   S.VIJAY-KUMAR,C.E.BUGG,K.D.WILKINSON,W.J.COOK
REMARK   1  TITL   THREE-DIMENSIONAL STRUCTURE OF UBIQUITIN AT 2.8 ANGSTROMS
REMARK   1  TITL 2 RESOLUTION
REMARK   1  REF    PROC.NATL.ACAD.SCI.USA        V.  82  3582 1985
REMARK   1  REFN                   ISSN 0027-8424
REMARK   1 REFERENCE 3
REMARK   1  AUTH   W.J.COOK,F.L.SUDDATH,C.E.BUGG,G.GOLDSTEIN
REMARK   1  TITL   CRYSTALLIZATION AND PRELIMINARY X-RAY INVESTIGATION OF
REMARK   1  TITL 2 UBIQUITIN, A NON-HISTONE CHROMOSOMAL PROTEIN
REMARK   1  REF    J.MOL.BIOL.                   V. 130   353 1979
REMARK   1  REFN                   ISSN 0022-2836
REMARK   1 REFERENCE 4
REMARK   1  AUTH   D.H.SCHLESINGER,G.GOLDSTEIN
REMARK   1  TITL   MOLECULAR CONSERVATION OF 74 AMINO ACID SEQUENCE OF
REMARK   1  TITL 2 UBIQUITIN BETWEEN CATTLE AND MAN
REMARK   1  REF    NATURE                        V. 255   423 1975
REMARK   1  REFN                   ISSN 0028-0836
REMARK   2
REMARK   2 RESOLUTION.    1.80 ANGSTROMS.
```

Figure 4: The beginning of the header of Ubiquitin (PDB:1UBQ)



```
ATOM      1  N   PRO A   1      -3.190   7.728  33.820  1.00 21.66       N
ATOM      2  CA  PRO A   1      -2.220   6.922  34.499  1.00 18.48       C
ATOM      3  C   PRO A   1      -0.802   7.080  34.031  1.00 17.67       C
ATOM      4  O   PRO A   1      -0.530   7.806  33.045  1.00 18.49       O
ATOM      5  CB  PRO A   1      -2.727   5.495  34.165  1.00 20.72       C
ATOM      6  CG  PRO A   1      -3.834   5.651  33.165  1.00 20.84       C
ATOM      7  CD  PRO A   1      -4.438   7.016  33.499  1.00 19.67       C
ATOM      8  N   GLN A   2       0.091   6.450  34.755  1.00 14.65       N
ATOM      9  CA  GLN A   2       1.526   6.384  34.480  1.00 17.51       C
ATOM     10  C   GLN A   2       1.753   4.880  34.129  1.00 18.83       C
ATOM     11  O   GLN A   2       1.442   3.982  34.963  1.00 19.75       O
ATOM     12  CB  GLN A   2       2.519   6.960  35.431  1.00 17.46       C
ATOM     13  CG  GLN A   2       3.943   6.608  35.023  1.00 20.07       C
ATOM     14  CD  GLN A   2       4.890   7.376  35.931  1.00 26.75       C
ATOM     15  OE1 GLN A   2       5.366   6.856  36.946  1.00 31.80       O
ATOM     16  NE2 GLN A   2       5.172   8.611  35.545  1.00 29.41       N
```

Figure 5: An example of atomic coordinates in the PDB



```
                 Atomic Coordinates: PDB Format

                                   Chain name
    Amino Acid                    /   Sequence Number
              \                  /   /
     Element   \                /   /     -----Coordinates-----
            \   \              /   /      X         Y        Z        (etc.)
ATOM      1  N   ASP L   1      4.060    7.307    5.186    ...
ATOM      2  CA  ASP L   1      4.042    7.776    6.553    ...
ATOM      3  C   ASP L   1      2.668    8.426    6.644    ...
ATOM      4  O   ASP L   1      1.987    8.438    5.606    ...
ATOM      5  CB  ASP L   1      5.090    8.827    6.797    ...
ATOM      6  CG  ASP L   1      6.338    8.761    5.929    ...
ATOM      7  OD1 ASP L   1      6.576    9.758    5.241    ...
ATOM      8  OD2 ASP L   1      7.065    7.759    5.948    ...
              \\
               Element position within amino acid
```

Figure 6: An example of atomic coordinates in the PDB

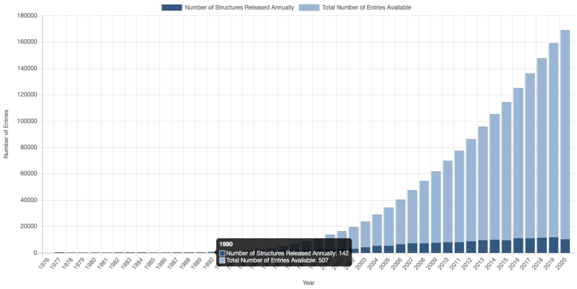


Figure 7: Number of structures in the PDB over the years, as of 2020

Table 1: Breakdown of structures on the PDB as of 2017 (from the RCSB website)

Experimental Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	113840	1906	5818	4	121568
NMR	10571	1229	246	8	12054
ELECTRON MICROSCOPY	1328	30	473	0	1831
HYBRID	105	3	2	1	111
other	200	4	6	13	223
Total	126044	3172	6545	26	135787

Size of the PDB over the years: The PDB was established in 1971. The number of deposited structures began to increase dramatically due to improved experimental techniques. In October 1998, the management of the PDB became the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB). As of 2017, there are over 135,000 structures in the PDB, with thousands of new structures being added every year.

The breakdown by experimental method shows that the majority of structures were obtained by X-ray crystallography. Out of the rest, most of the structures were obtained using NMR. Other methods are still not as popular.

PDB-101: In recent years the RCSB has provided extensive research and educational tools through the PDB-101 portal (`https://pdb101.rcsb.org/`). Most prominently, the ”Molecule of the Month” series provides a curated introduction to the structures available in the PDB. It presents a short description of selected molecules from the PDB. Each article includes an introduction to the structure and function of the molecule, a discussion of the relevance of the molecule to human health and welfare, and suggestions for how visitors might view these structures and access further details.

1.2 The wwPDB

In recent years, the major database for macromolecular structures is the worldwide PDB (wwPDB) at `http://www.wwpdb.org/`. It is a joint effort of the RCSB, the Protein Data Bank Europe (at the European Bioinformatics Institute, EBI), the Protein Databank Japan (based at Osaka University), and the Biological Magnetic Resonance Data Bank (BMRB).

2 Structural Classification of Proteins

Protein structure classification is important because it gives you an entry point into the world of protein structure that is independent of sequence similarity. Proteins are grouped not by functional families, but according to what kind of secondary structure (alpha helix, beta sheet, or both) they have. Within those larger classes, subclasses are defined based on how the secondary structures in the protein are arranged. The focus in protein classification is on finding proteins that have similar chemical architectures; it doesn't matter if their sequences are related. Over the years, we've learned from classification that there are far fewer unique protein folds than there are protein sequence families. Protein chemists often are interested in the information that can be extracted from broader structural classes of proteins, since analyzing that information can help them better understand how proteins fold.

here isn't really a consensus as to how to classify protein structures quantitatively. Instead, structures end up in qualitatively named classes such as ”greek key,” ”helix bundle,” and ”alpha-beta barrel.” These fold classes are useful in that they draw attention to prominent structural features and create a frame of reference for classifying structure. However, qualitative classifications don't lend themselves to automated analysis, and such protein classification databases still require the involvement of expert curators. If you're simply concerned with finding the close structural relatives of a published protein structure, there are a number of online classification databases in which existing structures have been annotated by a combination of automated analysis and input from protein structure experts. There are also automated tools for finding structural neighbors by structure alignment, though like any alignment method, these tools require you to understand the significance of comparison scores when analyzing results. If you are interested in doing your own analysis of a protein structure, there are several structure classification processes and tools that might help.

2.1 The SCOP Database

The Structural Classification of Proteins (SCOP) database is a largely manual classification of protein structural domains based on similarities of their structures and amino acid sequences [SCOP]. A motivation for this classification is to determine the evolutionary relationship between proteins. SCOP was created in 1994 in the Centre for Protein Engineering and the Laboratory of Molecular Biology.

The source of protein structures is the Protein Data Bank. The unit of classification of structure in SCOP is the protein *domain*. What the SCOP authors mean by ”domain” is suggested by their statement that small proteins and most medium sized ones have just one domain, and by the observation that human hemoglobin, which has an $\alpha 2 \beta 2$ structure, is assigned two SCOP domains, one for the α and one for the β subunit.

The shapes of domains are called ”folds” in SCOP. Domains belonging to the same fold have the same major secondary structures in the same arrangement with the same topological connections. 1195 folds are given in SCOP version 1.75. Short descriptions of each fold are given. For example, the ”globin-like” fold is described as core: 6 helices; folded leaf, partly opened. The fold to which a domain belongs is determined by inspection, rather than by software.

The hierarchical levels of SCOP are as follows.

1. Class: Types of folds, e.g., beta sheets.

2. Fold: The different shapes of domains within a class.

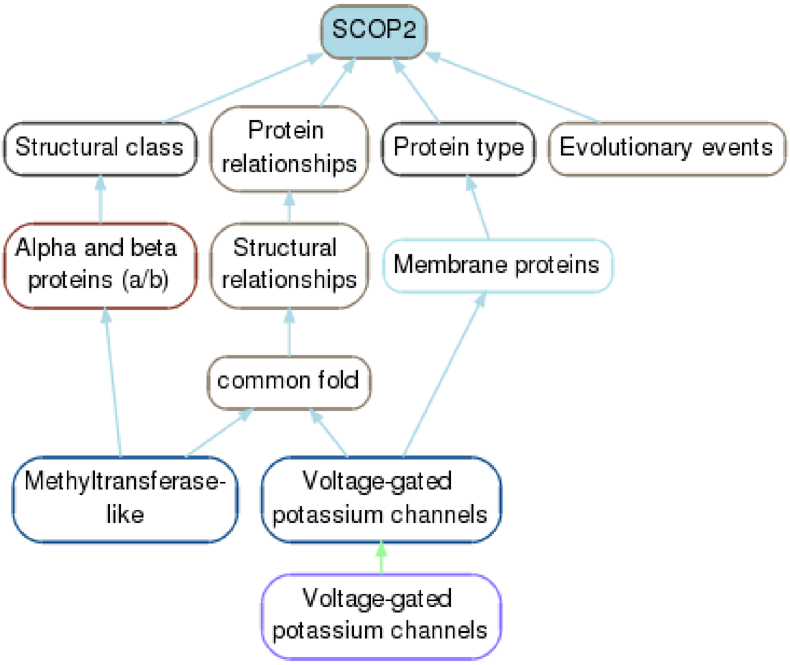
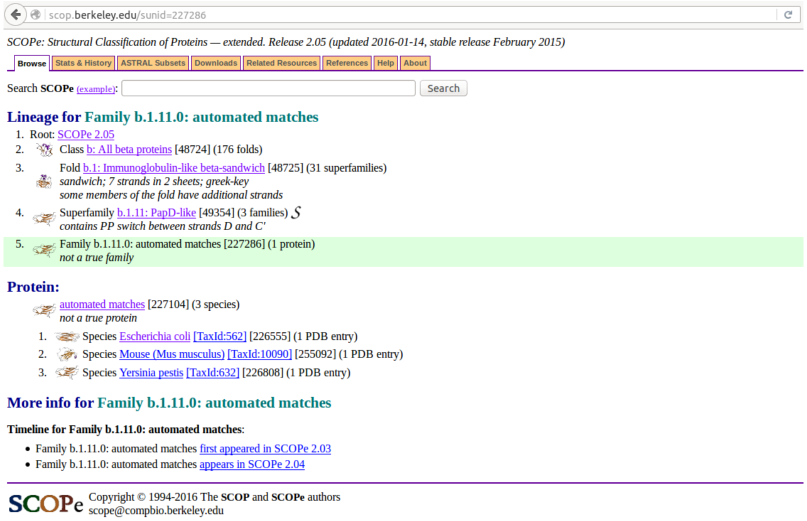


Figure 8: A SCOP2 graph structure example.

3. Superfamily: The domains in a fold are grouped into superfamilies, which have at least a distant common ancestor.
4. Family: The domains in a superfamily are grouped into families, which have a more recent common ancestor.
5. Protein domain: The domains in families are grouped into protein domains, which are essentially the same protein.
6. Species: The domains in ”protein domains” are grouped according to species.
7. Domain: part of a protein. For simple proteins, it can be the entire protein.

Work on SCOP concluded in June 2009 with the release of SCOP 1.75. SCOP is still available online but is no longer maintained or updated. The prototype of a new Structural Classification of Proteins 2 (SCOP2) database has been made publicly available. SCOP2 defines a new approach to the classification of proteins that is essentially different from SCOP, but retains its best features [SCOP2]. Rather than using a simple tree hierarchy, the classification of proteins is described in terms of a directed acyclic graph in which each node defines a relationship of particular type and is exemplified by a region of protein structure and sequence. Importantly, there can be more than one parental node for a child node that allows multiple routes to a particular relationship. Figure 8 shows an example of a graph structure.



SCOPe: Structural Classification of Proteins — extended. Release 2.05 (updated 2016-01-14, stable release February 2015)

Search **SCOPe** (example):

Lineage for Family b.1.11.0: automated matches

1. Root: SCOPe 2.05
2. Class b: All beta proteins [48724] (176 folds)
3. Fold b.1: Immunoglobulin-like beta-sandwich [48725] (31 superfamilies)
   sandwich; 7 strands in 2 sheets; greek-key
   some members of the fold have additional strands
4. Superfamily b.1.11: PapD-like [49354] (3 families) *S*
   contains PP switch between strands D and C'
5. Family b.1.11.0: automated matches [227286] (1 protein)
   not a true family

Protein:

automated matches [227104] (3 species)
not a true protein

1. Species Escherichia coli [TaxId:562] [226555] (1 PDB entry)
2. Species Mouse (Mus musculus) [TaxId:10090] [255092] (1 PDB entry)
3. Species Yersinia pestis [TaxId:632] [226808] (1 PDB entry)

More info for Family b.1.11.0: automated matches

Timeline for Family b.1.11.0: automated matches:

- Family b.1.11.0: automated matches first appeared in SCOPe 2.03
- Family b.1.11.0: automated matches appears in SCOPe 2.04

SCOPe Copyright © 1994-2016 The **SCOP** and **SCOPe** authors
scope@compbio.berkeley.edu

Figure 9: The SCOPe database.

By 2009, the original SCOP database manually classified 38,000 PDB entries into a strictly hierarchical structure. With the accelerating pace of protein structure publications, the limited automation of classification could not keep up, leading to a non-comprehensive dataset. The Structural Classification of Proteins extended (SCOPe) database was released in 2012 with far greater automation of the same hierarchical system and is full backwards compatible with SCOP [SCOPe]. In 2014, manual curation was reintroduced into SCOPe to maintain accurate structure assignment. As of February 2015, SCOPe 2.05 classified 71,000 of the 110,000 total PDB entries. Figure 9 shows a screenshot of the SCOPe database.

2.2 The CATH Database

CATH is another database which classifies protein structures downloaded from the Protein Data Bank. It is a semi-automatic, hierarchical classification of protein domains initially published in 1997. The name CATH is an acronym of the four main levels in the classification. The four main levels of the CATH hierarchy are as in Table 2

Much of the work is done by automatic methods, however there are important manual elements to the classification. The very first step is to separate the proteins into domains. It is difficult to produce an unequivocal definition of a domain and this is one area in which CATH and SCOP differ. The domains are automatically sorted into classes and clustered on the basis of sequence similarities. These groups form the **H** levels of the classification. The topology level is formed by structural comparisons of the homologous groups. Finally, the **A**rchitecture level is assigned manually. **C**lass Level classification is done on the basis of 4 criteria:

1. Secondary structure content;

2. Secondary structure contacts;

Table 2: The four main levels of CATH classification

#	Level	Description
1	**C**lass	Overall secondary-structure content of the domain. (Equivalent to SCOP class)
2	**A**rchitecture	High structural similarity but no evidence of homology. (Equivalent to SCOP fold)
3	**T**opology	A large-scale grouping of topologies which share particular structural features
4	**H**omologous superfamily	Indicative of a demonstrable evolutionary relationship. (Equivalent to SCOP superfamily)

3. Secondary structure alternation score; and
4. Percentage of parallel strands.

CATH defines four classes: mostly-α, mostly-β, α and β, few secondary structures.

3 Molecular Visualization Tools

Visualizing Protein Structures Numerous tools are available for visualizing the structures stored in the PDB and other repositories. Most such tools allow a detailed examination of the molecule in a variety of rendering modes. For example, sometimes it may be useful to have a detailed image of the surface of the molecule as experienced by a molecule of water. For other purposes, a simple, cartoonish representation of the major structural features may be sufficient.

References

[SCOP] Loredana Lo Conte, Bart Ailey, Tim J. P. Hubbard, Steven E. Brenner, Alexey G. Murzin, and Cyrus Chothia. Scop: a structural classification of proteins database. *Nucleic Acids Research*, 28(1):257–259, 2000.

[SCOP2] Antonina Andreeva, Dave Howorth, Cyrus Chothia, Eugene Kulesha, and Alexey G. Murzin. Scop2 prototype: a new approach to protein structure mining. *Nucleic Acids Research*, 2013.

[SCOPe] Naomi K. Fox, Steven E. Brenner, and John-Marc Chandonia. Scope: Structural classification of proteins − extended, integrating scop and astral data and classification of new structures. *Nucleic Acids Research*, 42(D1):D304–D309, 2014. URL `http://nar.oxfordjournals.org/content/42/D1/D304.abstract`.

[JMol] Jmol: an open-source java viewer for chemical structures in 3d. URL `http://www.jmol.org/`.

[Pymol] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010.

[RasMol] R. A. Sayle and Milner E. J. White. RASMOL: biomolecular graphics for all. *Trends Biochem Sci*, 20(9), 1995.

[ProtExplorer] E. Martz. Protein explorer: easy yet powerful macromolecular visualization. *Trends Biochem Sci*, 27(2):107–109, 2002.

[Chimera] E.F. Pettersen, T.D. Goddard, C.C. Huang, G.S. Couch, D.M. Greenblatt, E.C. Meng, and T.E. Ferrin. Ucsf chimera–a visualization system for exploratory research and analysis. *J. Comput Chem.*, 25(13):1605–1612, 2004.

[VMD] W. Humphrey, A. Dalke, and K. Schulten. VMD - Visual Molecular Dynamics. *J. Molec. Graphics*, 14:33–38, 1996. URL `http://www.ks.uiuc.edu/Research/vmd/`.