Protein Data Bank

24-08-2024

1

Why is it important to understand a protein's structure?

- Information of the protein structure gives understanding of how a protein works.

2

US data center for the global Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB) archive of 3D structure data for large biological molecules (proteins, DNA, and RNA)

PDBe is a founding member of the Worldwide Protein Data Bank which collects, organises and disseminates data on biological macromolecular structures.

PDBj - maintains PDB/BMRB/EMDB archives of macromolecular structures – under the Joint Usage and Research activities of the Institute for Protein Research.

WORLDWIDE PDB PROTEIN DATA BANK

a single PDB archive of macromolecular structural data that is freely and publicly available to the global community.

Biological Magnetic Resonance Data bank (BMRB) collects, annotates, archives, and disseminates spectral and quantitative data derived from NMR spectroscopic investigations of biological macromolecules and metabolites.

The Electron Microscopy Data Bank - public repository for electron cryo-microscopy maps and tomograms of macromolecular complexes and subcellular structures.

3

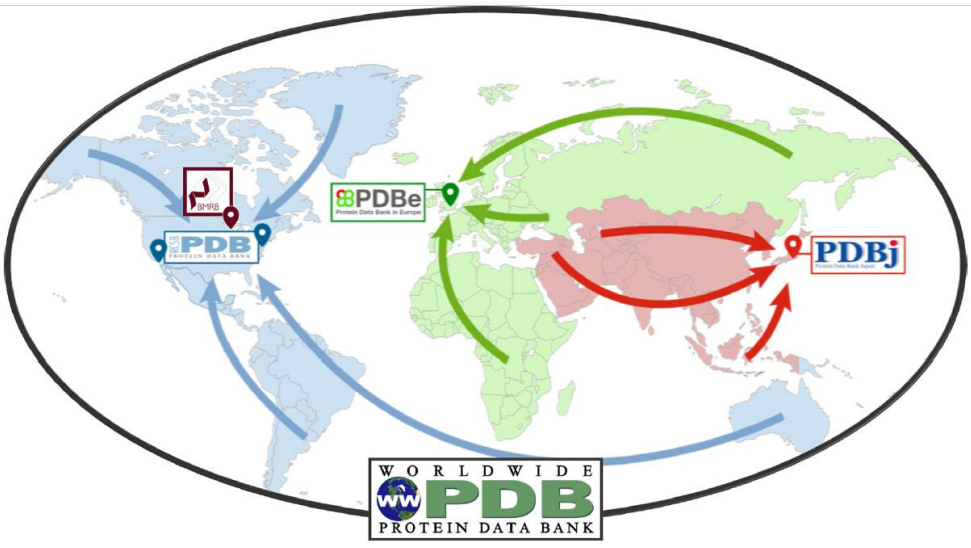


Figure 2. World map indicating the locations of wwPDB partner sites and color-coded to indicate the regions from which each accepts PDB depositions.

4

- The “PDB Archive” is a collection of flat files in three different formats:
  1. the legacy PDB format : PDB format is **the legacy file format (**.pdb, .ent, .brk.)
  2. the PDBx/mmCIF format : Macromolecule Crystallographic Information File
  3. the Protein Data Bank Markup Language (PDBML) format. - **provides a representation of PDB data in XML format** (Extensible Markup Language (XML) is a file format and markup language that can store, transmit, and reconstruct data in a format that's both human-readable and machine-readable).

5

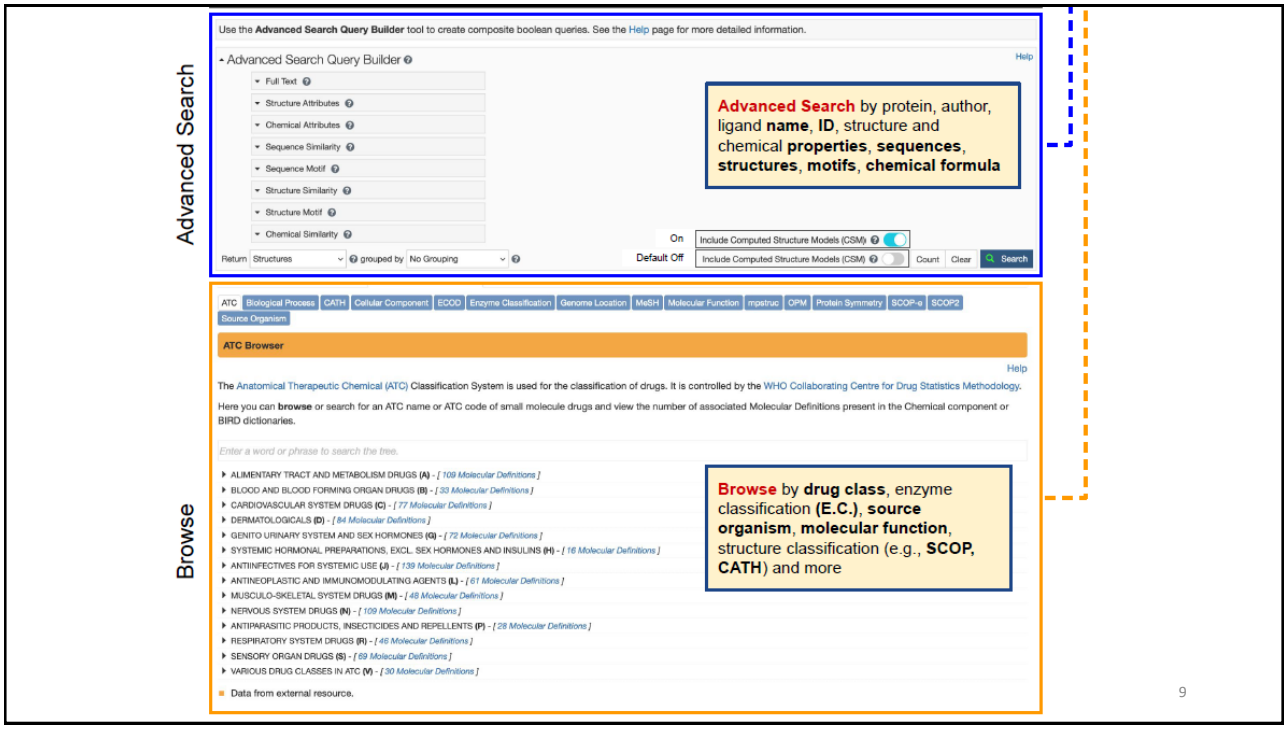
https://www.rcsb.org/

Importance of PDB

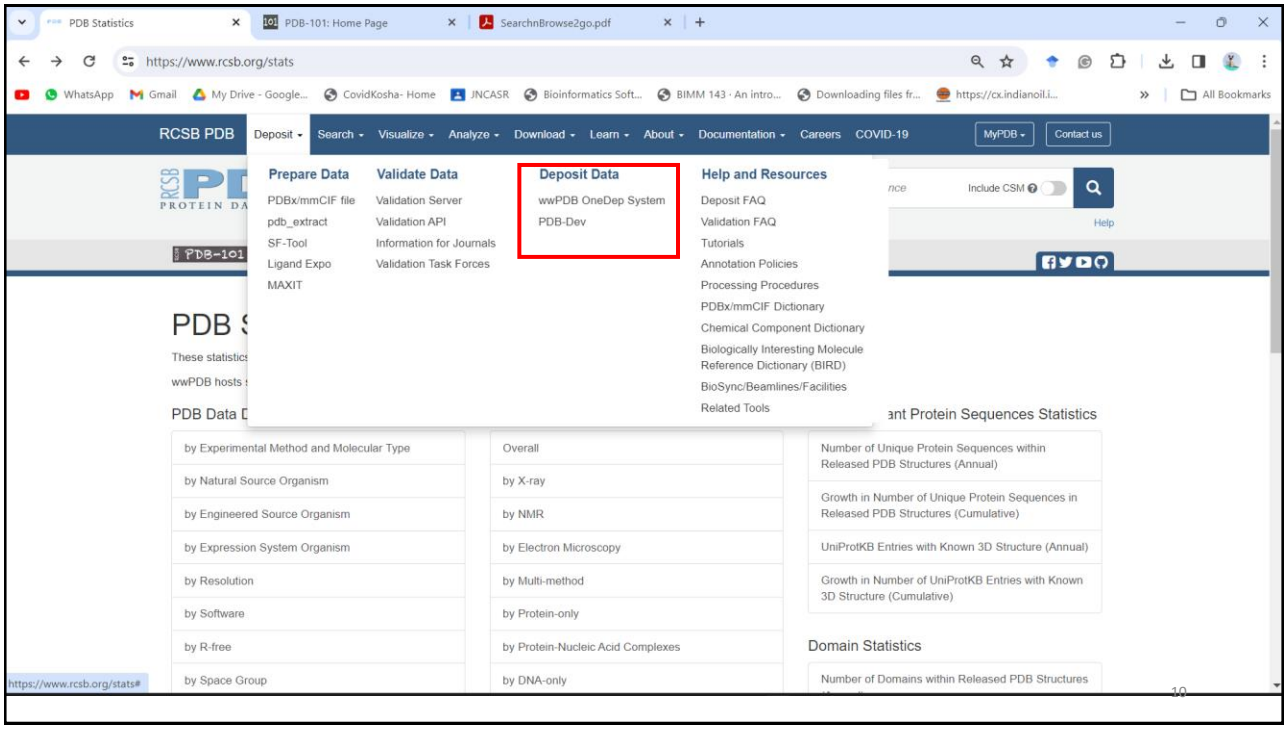
- It offers information on the 3D forms of proteins, nucleic acids, and complex components
- helps students to have the whole understanding, from protein synthesis to health and disease, of all aspects of biomedicine and agriculture.
- Visualization tools are also present through which one can visualize a structure in three-dimensional in the web browser itself.

7

8


Advanced Search
Use the Advanced Search Query Builder tool to create composite boolean queries. See the Help page for more detailed information.
Advanced Search Query Builder
Full Text
Structure Attributes
Chemical Attributes
Sequence Similarity
Sequence Motif
Structure Similarity
Structure Motif
Chemical Similarity
Return Structures
grouped by No Grouping
On
Include Computed Structure Models (CSM)
Default Off
Include Computed Structure Models (CSM)
Count
Clear
Search
Advanced Search by protein, author, ligand name, ID, structure and chemical properties, sequences, structures, motifs, chemical formula
Browse
ATC
Biological Process
CATH
Cellular Component
ECOD
Enzyme Classification
Genome Location
MeSH
Molecular Function
mpstruc
OPM
Protein Symmetry
SCOP-e
SCOP2
Source Organism
ATC Browser
Help
The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of drugs. It is controlled by the WHO Collaborating Centre for Drug Statistics Methodology.
Here you can browse or search for an ATC name or ATC code of small molecule drugs and view the number of associated Molecular Definitions present in the Chemical component or BIRD dictionaries.
Enter a word or phrase to search the tree.
ALIMENTARY TRACT AND METABOLISM DRUGS (A) - [109 Molecular Definitions]
BLOOD AND BLOOD FORMING ORGAN DRUGS (B) - [33 Molecular Definitions]
CARDIOVASCULAR SYSTEM DRUGS (C) - [77 Molecular Definitions]
DERMATOLOGICALS (D) - [84 Molecular Definitions]
GENITO URINARY SYSTEM AND SEX HORMONES (G) - [72 Molecular Definitions]
SYSTEMIC HORMONAL PREPARATIONS, EXCL. SEX HORMONES AND INSULINS (H) - [16 Molecular Definitions]
ANTIINFECTIVES FOR SYSTEMIC USE (J) - [139 Molecular Definitions]
ANTINEOPLASTIC AND IMMUNOMODULATING AGENTS (L) - [61 Molecular Definitions]
MUSCULO-SKELETAL SYSTEM DRUGS (M) - [48 Molecular Definitions]
NERVOUS SYSTEM DRUGS (N) - [109 Molecular Definitions]
ANTIPARASITIC PRODUCTS, INSECTICIDES AND REPELLENTS (P) - [28 Molecular Definitions]
RESPIRATORY SYSTEM DRUGS (R) - [46 Molecular Definitions]
SENSORY ORGAN DRUGS (S) - [69 Molecular Definitions]
VARIOUS DRUG CLASSES IN ATC (V) - [30 Molecular Definitions]
Data from external resource.
Browse by drug class, enzyme classification (E.C.), source organism, molecular function, structure classification (e.g., SCOP, CATH) and more

9

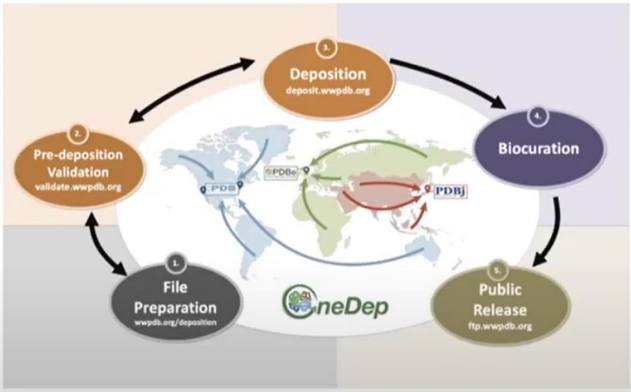

PDB Statistics
PDB-101: Home Page
SearchnBrowse2go.pdf
https://www.rcsb.org/stats
RCSB PDB
Deposit
Search
Visualize
Analyze
Download
Learn
About
Documentation
Careers
COVID-19
MyPDB
Contact us
Prepare Data
PDBx/mmCIF file
pdb_extract
SF-Tool
Ligand Expo
MAXIT
Validate Data
Validation Server
Validation API
Information for Journals
Validation Task Forces
Deposit Data
wwPDB OneDep System
PDB-Dev
Help and Resources
Deposit FAQ
Validation FAQ
Tutorials
Annotation Policies
Processing Procedures
PDBx/mmCIF Dictionary
Chemical Component Dictionary
Biologically Interesting Molecule Reference Dictionary (BIRD)
BioSync/Beamlines/Facilities
Related Tools
PDB-101
PDB Data
by Experimental Method and Molecular Type
by Natural Source Organism
by Engineered Source Organism
by Expression System Organism
by Resolution
by Software
by R-free
by Space Group
Overall
by X-ray
by NMR
by Electron Microscopy
by Multi-method
by Protein-only
by Protein-Nucleic Acid Complexes
by DNA-only
Protein Sequences Statistics
Number of Unique Protein Sequences within Released PDB Structures (Annual)
Growth in Number of Unique Protein Sequences in Released PDB Structures (Cumulative)
UniProtKB Entries with Known 3D Structure (Annual)
Growth in Number of UniProtKB Entries with Known 3D Structure (Cumulative)
Domain Statistics
Number of Domains within Released PDB Structures

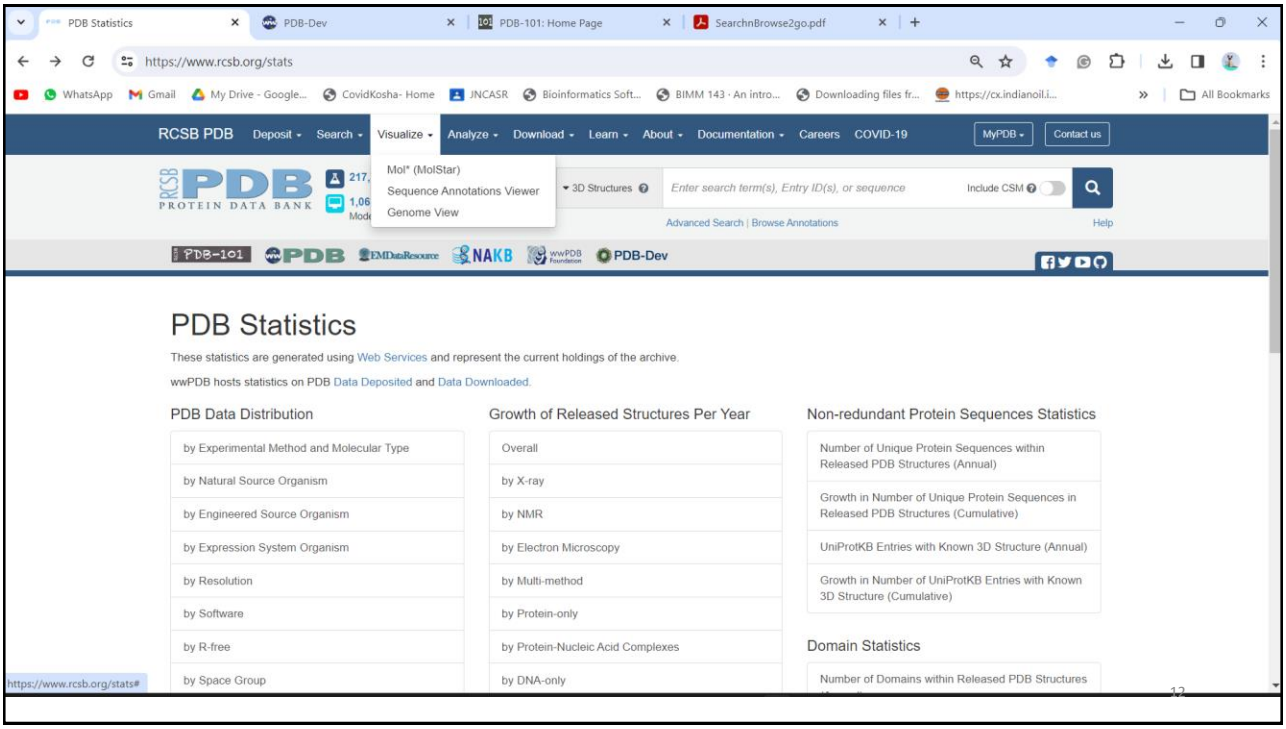
10

OneDep: a wwPDB unified deposition, biocuration, and validation tool

- PDBx/mmCIF framework
- Supports crystallographic, 3DEM, and NMR methods
- Geographical workload distribution
  - Maximize customer service
- Follow standard biocuration procedure/guidelines
  - Published online
- Provide validation report at four stages

Young, J. *et al.*, Structure, 2017, doi:





Sequence Annotation Viewer

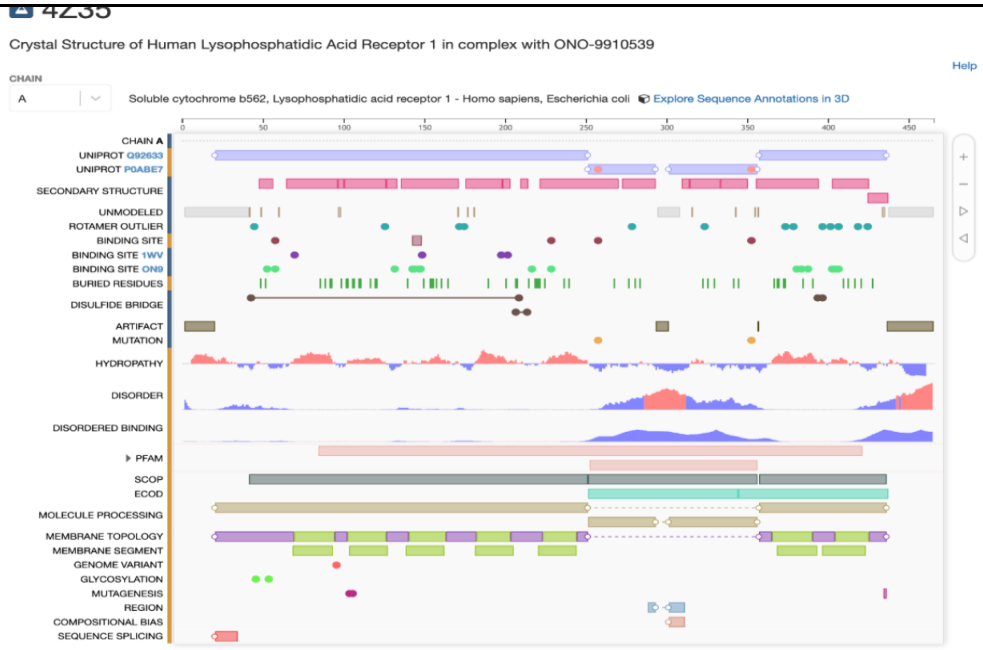


Figure 1: Sequence Annotations Viewer for PDB entry 4z35 showing features from the PDB entry (marked by blue vertical bar on the left) and features derived from UniProt, SCOP, PFam etc. (marked by orange vertical lines on the left).

13

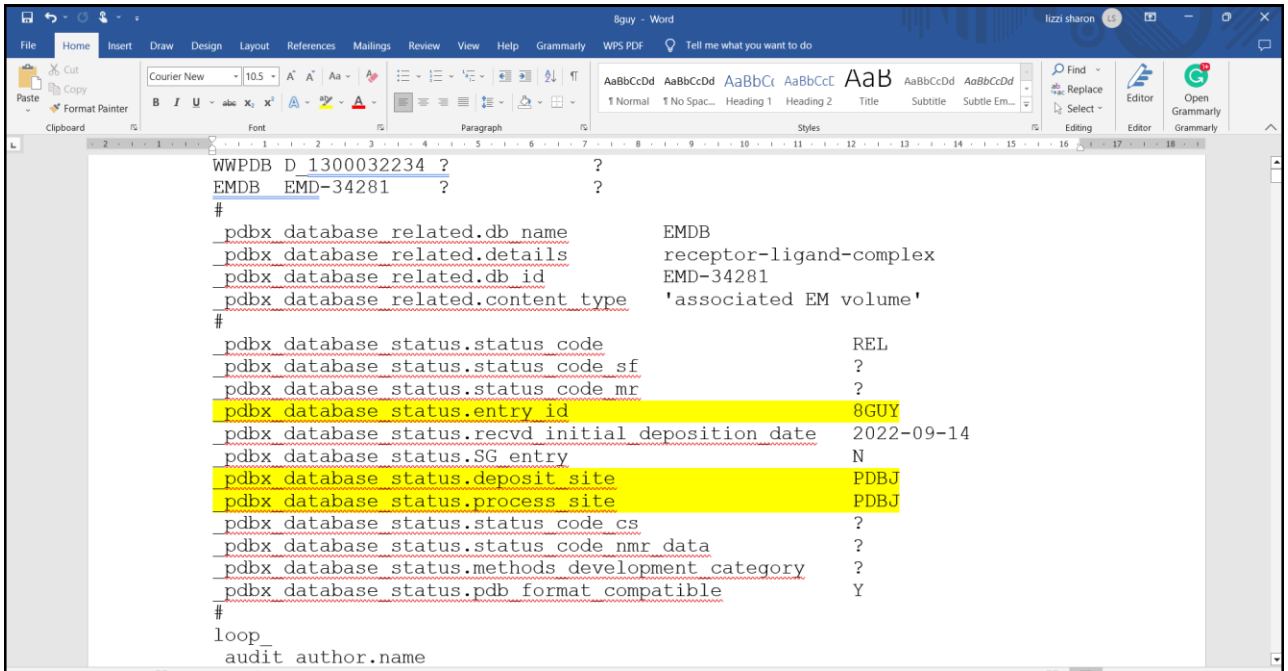
Structure of a PDB file

14

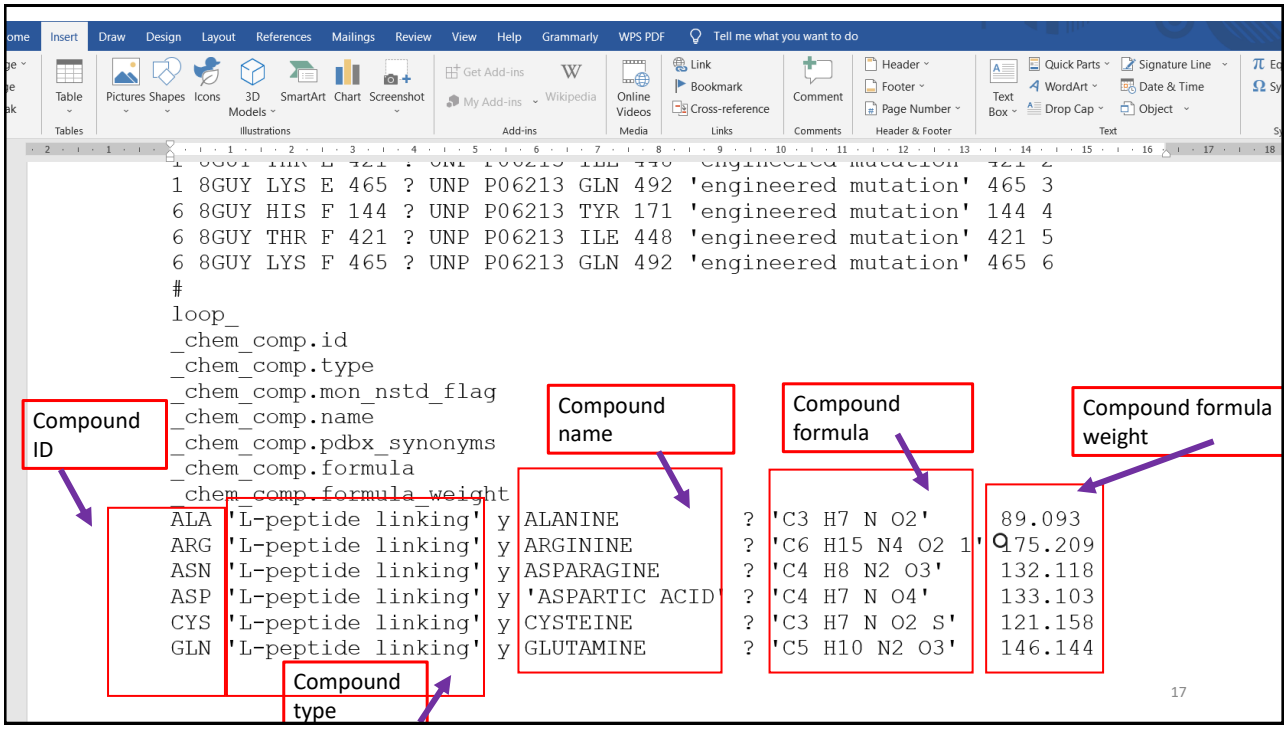
What does PDB file contain?

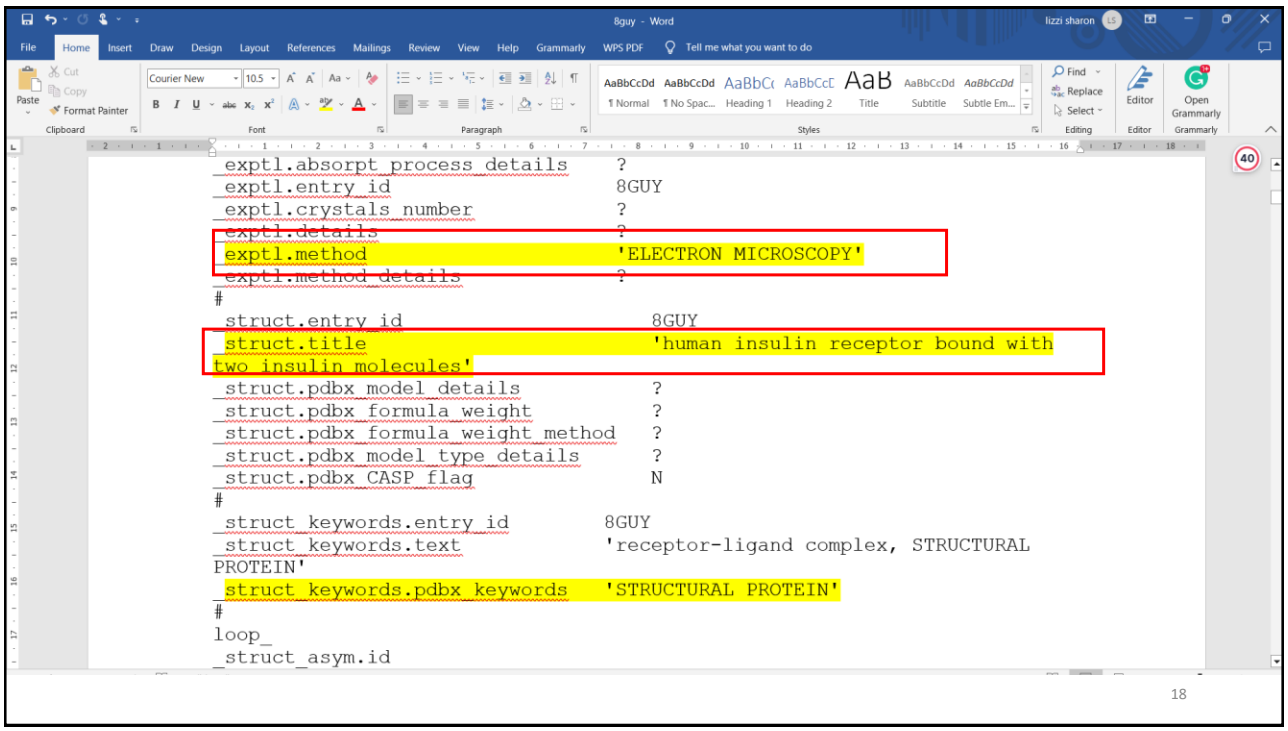
- The archives contain atomic coordinates, bibliographic citations, primary and secondary structure information, crystallographic structure factors and NMR experimental data.

15



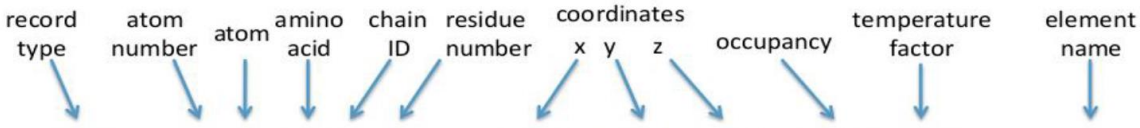
16


1 8GUY LYS E 465 ? UNP P06213 GLN 492 'engineered mutation' 465 3
6 8GUY HIS F 144 ? UNP P06213 TYR 171 'engineered mutation' 144 4
6 8GUY THR F 421 ? UNP P06213 ILE 448 'engineered mutation' 421 5
6 8GUY LYS F 465 ? UNP P06213 GLN 492 'engineered mutation' 465 6
#
loop_
_chem_comp.id
_chem_comp.type
_chem_comp.mon_nstd_flag
_chem_comp.name
_chem_comp.pdbx_synonyms
_chem_comp.formula
_chem_comp.formula_weight
ALA 'L-peptide linking' y ALANINE ? 'C3 H7 N O2' 89.093
ARG 'L-peptide linking' y ARGININE ? 'C6 H15 N4 O2 1' 175.209
ASN 'L-peptide linking' y ASPARAGINE ? 'C4 H8 N2 O3' 132.118
ASP 'L-peptide linking' y 'ASPARTIC ACID' ? 'C4 H7 N O4' 133.103
CYS 'L-peptide linking' y CYSTEINE ? 'C3 H7 N O2 S' 121.158
GLN 'L-peptide linking' y GLUTAMINE ? 'C5 H10 N2 O3' 146.144
Compound ID
Compound name
Compound formula
Compound formula weight
Compound type
17


_exptl.absorpt_process_details ?
_exptl.entry_id 8GUY
_exptl.crystals_number ?
_exptl.details ?
_exptl.method 'ELECTRON MICROSCOPY'
_exptl.method_details ?
#
_struct.entry_id 8GUY
_struct.title 'human insulin receptor bound with two insulin molecules'
_struct.pdbx_model_details ?
_struct.pdbx_formula_weight ?
_struct.pdbx_formula_weight_method ?
_struct.pdbx_model_type_details ?
_struct.pdbx_CASP_flag N
#
_struct_keywords.entry_id 8GUY
_struct_keywords.text 'receptor-ligand complex, STRUCTURAL PROTEIN'
_struct_keywords.pdbx_keywords 'STRUCTURAL PROTEIN'
#
loop_
_struct_asym.id
18

PDB file format

A **textfile** that includes **atomic coordinates**, observed sidechain rotamers, secondary structure assignments, atomic connectivity, ...




```
ATOM      1  N   MET D   1      14.322  20.430  -2.337  1.00 17.78           N
ATOM      2  CA  MET D   1      14.423  20.285  -0.855  1.00 18.66           C
ATOM      3  C   MET D   1      15.153  21.479  -0.242  1.00 18.46           C
ATOM      4  O   MET D   1      15.811  22.241  -0.941  1.00 18.84           O
ATOM      5  CB  MET D   1      15.068  18.970  -0.457  1.00 20.20           C
ATOM      6  CG  MET D   1      16.569  18.895  -0.674  1.00 20.60           C
ATOM      7  SD  MET D   1      17.240  17.319  -0.103  1.00 22.81           S
ATOM      8  CE  MET D   1      16.378  16.194  -1.196  1.00 13.23           C
ATOM      9  N   LEU D   2      14.983  21.653   1.071  1.00 18.40           N
ATOM     10  CA  LEU D   2      15.568  22.825   1.718  1.00 19.14           C
ATOM     11  C   LEU D   2      17.093  22.722   1.765  1.00 18.53           C
ATOM     12  O   LEU D   2      17.655  21.647   1.945  1.00 19.07           O
ATOM     13  CB  LEU D   2      15.025  23.078   3.121  1.00 21.35           C
ATOM     14  CG  LEU D   2      15.438  24.404   3.773  1.00 22.45           C
ATOM     15  CD1 LEU D   2      14.856  25.606   3.049  1.00 23.53           C
ATOM     16  CD2 LEU D   2      15.042  24.430   5.244  1.00 23.83           C
```

19

"SEQRES records" - list of the primary sequence of the polymeric molecules present in the entry.

```
SEQRES   1 B   19   DT  DG  DG  DA  DG  DA  DT  DG  DA  DC  DG  DT  DC
SEQRES   2 B   19   DA  DT  DC  DT  DC  DC
SEQRES   1 A   63  MET ILE VAL PRO GLU SER SER ASP PRO ALA ALA LEU LYS
SEQRES   2 A   63  ARG ALA ARG ASN THR GLU ALA ALA ARG ARG SER ARG ALA
SEQRES   3 A   63  ARG LYS LEU GLN ARG MET LYS GLN LEU GLU ASP LYS VAL
SEQRES   4 A   63  GLU GLU LEU LEU SER LYS ASN TYR HIS LEU GLU ASN GLU
SEQRES   5 A   63  VAL ALA ARG LEU LYS LYS LEU VAL GLY GLU ARG
```

20

Several additional records are included in the PDB format to define modifications as they appear in the ATOM records:

Record Name	Describes
MODRES	Modifications to standard residues
HET	Nonstandard residues (as well as ligands, ions and water)
HETNAM	Full chemical name of the residue
HETSYM	Synonyms for the residue
FORMUL	Chemical formula of the residue

```
MODRES 1CAG HYP A    2   PRO   4-HYDROXYPROLINE
HET    HYP  A   2        8
HETNAM     HYP 4-HYDROXYPROLINE
HETSYN     HYP HYDROXYPROLINE
FORMUL   1  HYP    30(C5 H9 N O3)
```

21

Reference article

- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235-42. doi: 10.1093/nar/28.1.235. PMID: 10592235; PMCID: PMC102472.

22

Exercises

1. Download a tabular file (CSV) of **Ferritin protein**:
   1. PubMed ID
   2. Structure Keywords
   3. Structure title
   4. Resolution
   5. Ligand
   6. Experimental method

   Publication details:
   7. Title
   8. Publication year
   9. Journal Name
   10. DOI

   Oligosaccharide data:
   11. Molecular name
   12. Source organism
   13. Entry ID

23

Submit a print-out of the program and the table on Tuesday 27th August (10Marks – assignment)

Write a program to present table with the hits got with the key word Streptokinase C

SNo	Entry ID	organism	Title of the article	Journal name	Structure title	experimental method	resolution

24