PDB
The protein data bank

Thomas Hamelryck
thamelry@binf.ku.dk
BIO/DIKU, University of Copenhagen
November 2019

What is the PDB?

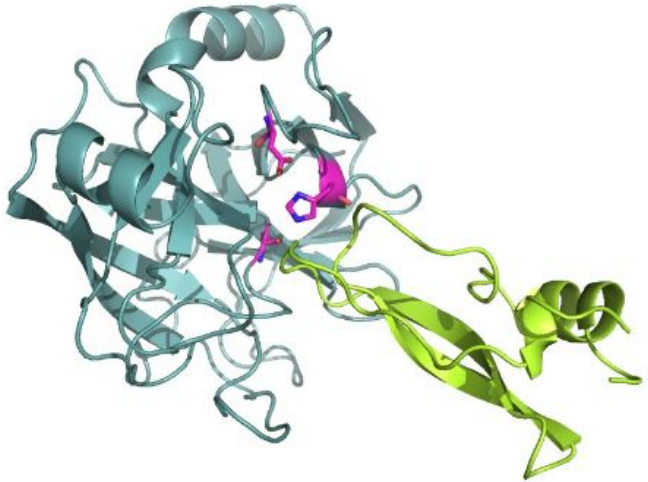
- www.rcsb.org
- Large repository of protein structures since 1971
  - 4 letter identifiers, eg. 1PTC or 1FAT
- 157.935 structures (Nov. 2019) and growing quickly
  - Highly redundant
  - 45.612 distinct protein sequences at 50% identity
- Mainly X-ray crystallography and NMR (12%)
- Advanced querying
- Freely available
- World-wide collaboration
  - Most scientific journals demand PDB submission

Exercise: Lysozyme

- Lysozyme – work in groups of 2
  a. Find all structures of Lysozyme. How many structures?
  b. What is the highest resolution structure?
  c. You obtained a lot of structures. Let's use “Advanced search”.
  d. Narrow down using “Macromolecular name”
  e. Narrow down knowing that lysozyme is approximately 130 amino acids in length (say, between 110-140).
  f. Narrow down using X-ray resolution (say, maximum 2Å).
  g. Remove structures with 95% similarity.
  h. How many structures are left? What do you conclude given the numbers?

PDB files

- Each structure is the outcome of an experiment and is given a unique 4 letter ID code (the PDB identifier).
    - Download the PDB file for 2PTC



PDB header

- Title, references, authors, experimental method and details, ligands, missing atoms
- Resolution, R-factor
- Secondary structure
  - Helices, sheets, disulfide bridges

```
HELIX    1  H1 SER E  164  ILE E  176  1SNGL ALPHA TURN,REST IRREG.       13
HELIX    2  H2 LYS E  230  VAL E  235  5CONTIGUOUS WITH H3                 6
HELIX    3  H3 SER E  236  ASN E  245  1CONTIGUOUS WITH H2                10
HELIX    4  H4 SER I   47  GLY I   56  1                                 10
SHEET    1  S1 2 ALA I  16  ALA I  25  0
SHEET    2  S1 2 GLY I  28  GLY I  36 -1
SSBOND   1 CYS E   22    CYS E  157                          1555   1555  2.03
SSBOND   2 CYS E   42    CYS E   58                          1555   1555  2.03
```

PDB header

- Symmetry transformations
  - REMARK 300, 350
  - Rotations and translations
- Asymmetric unit
  - Describes the smallest structure to which symmetry operations can be applied to generate the crystal repeating unit.
- Biological assembly
  - The quaternary structure.
  - The assembly that is believed to be the functional form of the molecule.

Exercise 2: Biological assembly

- For proteins 2HHB, 1HHO and 1HV4, determine how the asymmetric unit and biological assembly relate to each other.
  - Use the visualization box on the left.
  - Try the 3D view.

Coordinate data

- Protein atoms

	Atom number	Name	Res type	C	Res number	Coordinates x	y	z	Occupancy	B-factor	Element
ATOM	1	N	ILE	E	16	18.871	65.715	12.731	1.00	21.86	N
ATOM	2	CA	ILE	E	16	19.782	64.969	13.587	1.00	21.86	C
ATOM	3	C	ILE	E	16	21.173	64.987	12.945	1.00	21.86	C
ATOM	4	O	ILE	E	16	21.316	64.450	11.815	1.00	21.86	O
ATOM	5	CB	ILE	E	16	19.336	63.476	13.649	1.00	21.86	C
ATOM	6	CG1	ILE	E	16	17.903	63.230	14.154	1.00	18.04	C
ATOM	7	CG2	ILE	E	16	20.336	62.527	14.373	1.00	18.04	C
ATOM	8	CD1	ILE	E	16	17.785	63.415	15.666	1.00	18.04	C
ATOM	9	N	VAL	E	17	22.160	65.538	13.640	1.00	21.82	N
ATOM	10	CA	VAL	E	17	23.595	65.525	13.234	1.00	21.82	C

Coordinate data

- HETATMS (ligands, covalently bonded groups)
- Waters

```
HETATM 2086 CA    CA E 462       6.326  59.439   2.555  0.50 14.17           CA

HETATM 2087  O   HOH E 401      11.822  64.932  15.453  1.00 15.66           O
HETATM 2088  O   HOH E 402      21.349  65.697  21.825  1.00 17.85           O
HETATM 2089  O   HOH E 403      19.871  75.111  18.175  1.00 17.40           O
```

Exercise - coordinate data

- Chain identifiers for multiple chains
  - How many chains in 2PTC?
  - What are they?
- Occupancy is less than 1.0 when multiple conformations are present
  - What is the occupancy of the calcium ion and why?
- Residue numbers of chain E start at 16
  - Why are the first 15 missing?
- How many HETATMS are there and what do they belong to?

PDB Footer

- Connectivity data
  - Within non-standard (HET) residues.
  - From non-standard (HET) residues to standard residues.
  - Disulphide bridges.
- What do the last 2 connect lines below describe?

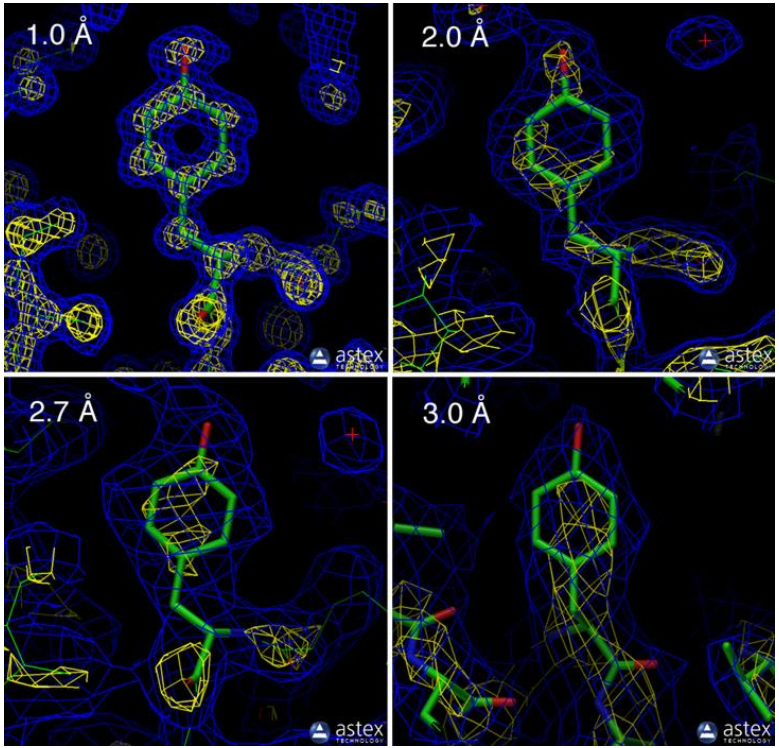
```
CONECT 2038 1872
CONECT 2070 1673
CONECT 2086  385  397  421  461
CONECT 2086 2131 2169
```

Structure quality

- Resolution
- Refinement (R-factor)
- B-factors (temperature factors)
- Model geometry

Structure quality - resolution

- Resolution
  - Measures the amount of information collected in a crystallography experiment.
  - The higher the resolution (lower Å), the higher the accuracy of the atomic positions.



Structure quality - refinement

- Refinement (R-factor)
  - Measures how well the structure fits the experimental X-ray data.
  - The lower the R-factor, the better the fit.
    - Free R-factor: uses cross validation
  - Well-refined protein structures have R-factors < 20%.

$$R = \frac{\Sigma||F_{obs}| - |F_{calc}||}{\Sigma|F_{obs}|}$$

R values:

0.6: Very bad

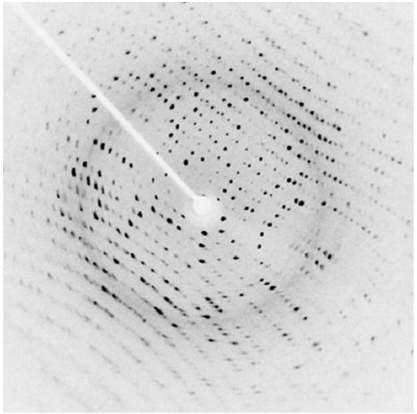
0.5: Bad

0.4: Recoverable

0.2: Good for Protein

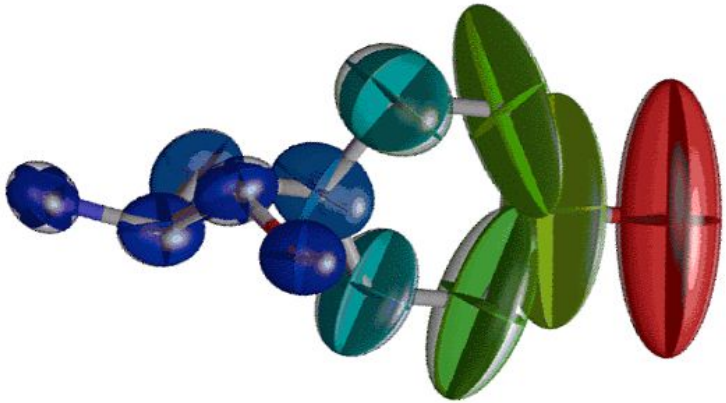
0.05: Good for small organic models

0.0: Perfect



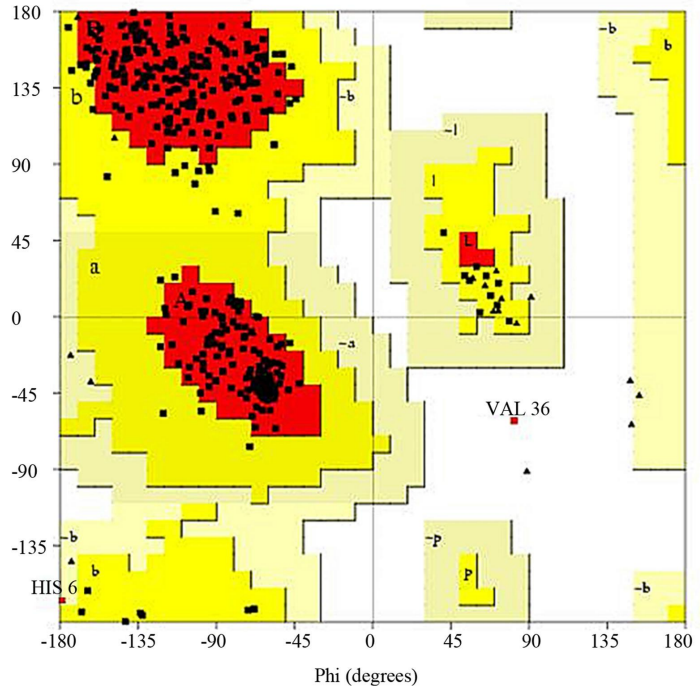
Structure quality - B factors

- B-factors or temperature factors
  - High B-factors (> 80) might indicate highly disordered parts of the structure or even error in the model.



Structure quality - Model geometry

- Model geometry
  - Bond distances
  - Bond angles
  - Dihedral angles
    - Ramachandran plot.



Exercise: structure quality

- Find 2DN1 and 5VMM
  - What are these structures?
- Assess the quality of each structure
  - Resolution
  - R-factor
  - B-factors
  - Ramachandran plot
- Are both high quality structures?

PDB & Biopython

- Download PDB file 1FAT with Biopython's Bio.pdb.
- Code:

```
from Bio.PDB import *

pdbl = PDBList()

pdbl.retrieve_pdb_file('1FAT')
```

Assignment

- Pick (any) 3 consecutive letters from your first or last name and look up the corresponding PDB file.
    - For example, I could pick 1THO, 1HOM, 1MAS or 1RYC etc.
- Briefly describe the structure (biology, ligands, experimental method etc.)
- Discuss structure quality.
- Discuss content of biological and asymmetric unit. Are they the same?
- Make two pictures of the content of the PDB file, with suitable captions.
- Submit as a PDF file.