PDB 101 Course Notes

[https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction]

0. Introduction

PDB archive == repository of **atomic coordinates** (+ other information) describing proteins

- *structures in the archive are determined using a balanced mixture of experimental observation and knowledge-based modeling => we should confirm that here is experimental evidence that supports the structure*

To determine the structure *(location of each atom relative to each other in the molecule)*:

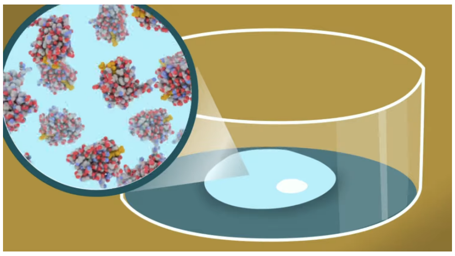
- scientist has some experimental data about the structure of the molecule - it is not sufficient to build an atomic model => need for additional knowledge *(e.g. the sequence of amino acids in the protein + we know the preferred geometry of atoms in a typical protein - the bond lengths and bond angles)*

1. X-ray Crystallography

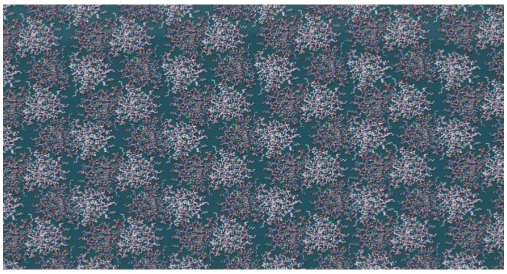
- experimental data: X-ray diffraction pattern

Steps:

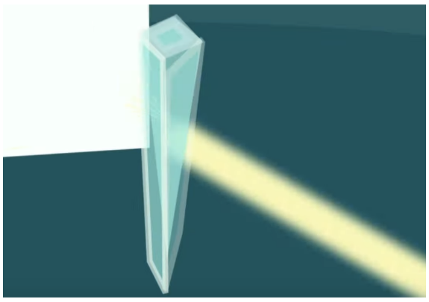
1. Purified and concentrated proteins form crystals.



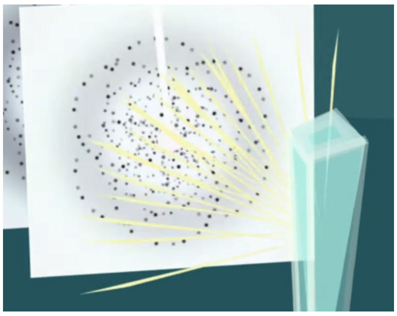
2. Within the crystal, many copies of the protein are arranged in symmetrical arrays.



3. X-ray beams strike the crystal.



4. The X-ray scatters into a spot pattern.



5. X-ray diffraction patterns are then analyzed to determine the positions of atoms in the protein.

- PDB contains 2 types of crystal structures:
  - 1. coordinate files (atomic positions for the final structure model),
  - 2. structure factors (the intensity and phase of the X-ray spots in the diffraction pattern) - *an image of the electron density map can be created (e.g. Astex viewer)*

- X-ray crystallography can provide very detailed atomic information, showing every atom in a protein or nucleic acid along with atomic details of ligands, inhibitors, ions, and other molecules that are incorporated into the crystal
- crystallization is difficult + limits the types of proteins that may be studied by it: excellent for determining structures of rigid proteins *(form nice, ordered crystals)*, but flexible proteins are more difficult (crystallography relies on having many molecules aligned in exactly the same orientation, like a repeated pattern in wallpaper *and flexible portions of protein will often be invisible in crystallographic electron density maps, since their electron density will be smeared over a large space*)
- accuracy of the atomic structure depends on the quality of the crystals (accuracy measures: **resolution** of crystallographic structure - *measures the amount of detail that may be seen in the experimental data*, and **R-value** - *measures how well the atomic model is supported by the experimental data found in the structure factor file*)

2. X-ray Free Electron Lasers (XFEL)
    - new method thanks to new technology termed serial femtosecond crystallography

Steps:
1. XFEL is used to create pulses of radiation that are extremely short (lasting only femtoseconds) and extremely bright.
2. Stream of tiny crystals (nanometers-micrometers) is passed through the beam.
3. Each X-ray pulse produces a diffraction pattern from a crystal *(often burning it up in the process)*.
4. Data set is compiled from as many as tens of thousands of these individual diffraction patterns.

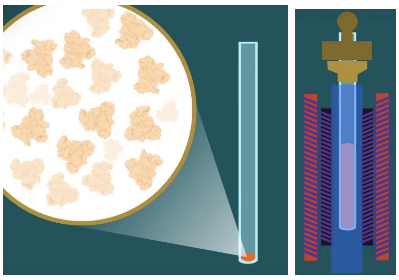
- very powerful method because it allows us to study molecular processes that occur over very short time scales *(e.g. absorption of light by biological chromophores)*

3. NMR Spectroscopy (Nuclear Magnetic Resonance)

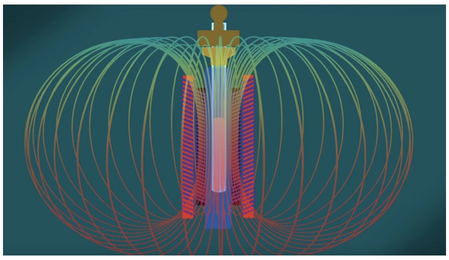
- experimental data: information on the local conformation and distance between atoms that are close to one another

Steps:

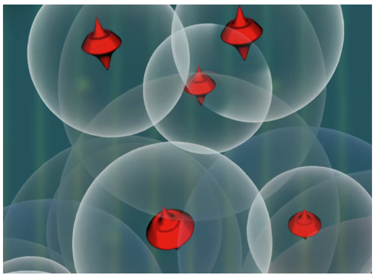
1. Purified protein is mixed with special solvent and inserted into NMR probe.



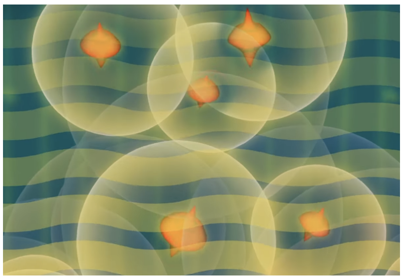
2. The sample is exposed to strong magnetic field.



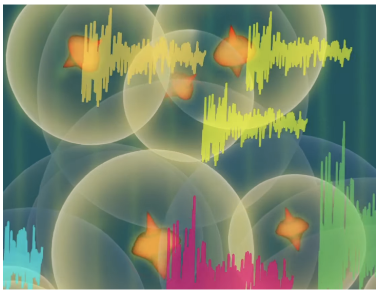
3. It makes nuclei of certain atoms spin (e.g. hydrogen).



4. When the sample is probed with radio waves, the nuclei become excited and resonate.

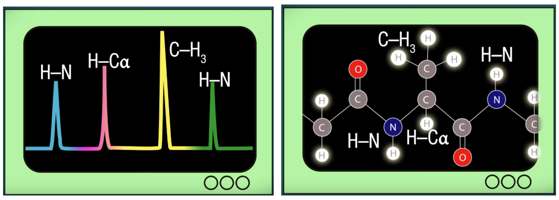


5. The frequencies are measured and recorded.



(the surrounding atoms determine how high/ low the frequencies are)

6. Using computational methods, the measurements are converted into a graph that represents the frequencies as peaks with specific locations for specific atom groups.



7. This information is further refined and combined with additional NMR experiments to determine the 3D structure.

- technique is currently limited to small/ medium proteins *(large proteins present problems with overlapping peaks in the NMR spectra)*
- major advantage: it provides information on proteins in solution *(opposed to those locked in a crystal/ bound to a microscope grid)* - premier method for studying the atomic structures of flexible proteins
- *typical NMR structure will include an ensemble of protein structures, all of which are consistent with the observed list of experimental restraints - structures in ensemble will be very similar to each other in*

regions with strong restraints, and very different in less constrained portions of the chain - these areas with fewer restraints are the flexible parts of the molecule => do not give a strong signal

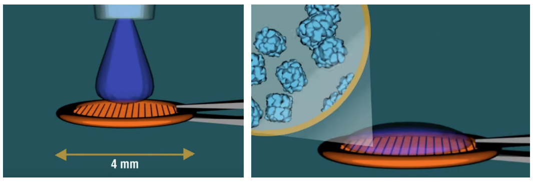
- PDB contains two types of structures:
    - 1. full ensemble from the structural determination (each structure designated as a separate model),
    - 2. minimized average structure,
    - *+ 3. list of restraints (hydrogen bonds, disulfide linkages, distances between hydrogen atoms that are close to one another, and restraints on the local conformation and stereochemistry of the chain)*

4. 3D Electron Microscopy (3DEM)

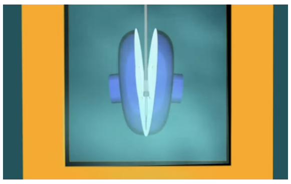
- experimental data: image of the overall shape of the molecule

Steps:

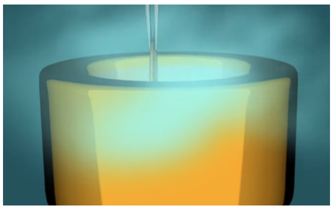
1. Tiny amount of purified protein is placed onto copper grid.



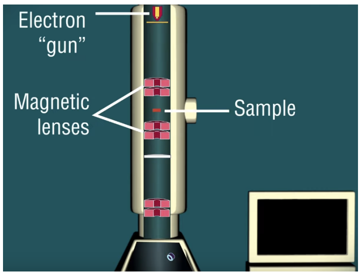
2. Special machine spreads the sample in single layer on the grid.



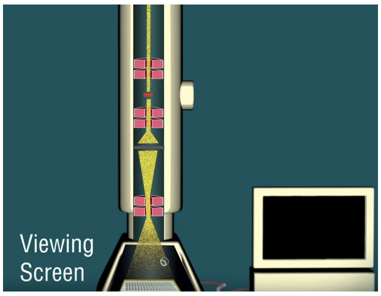
3. Sample if frozen in liquid ethane.



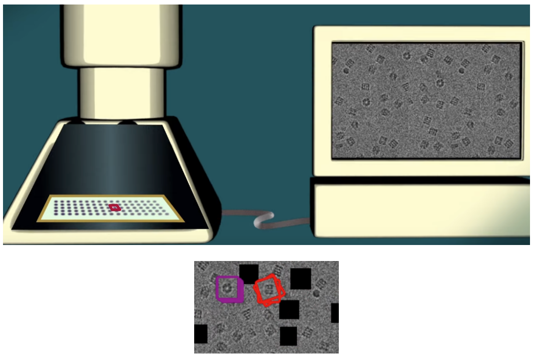
4. Sample is loaded into electron microscope.



5. Sample is exposed to a beam of accelerated electrons and images are captured.

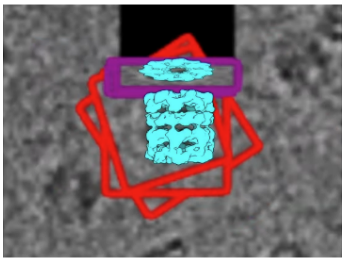


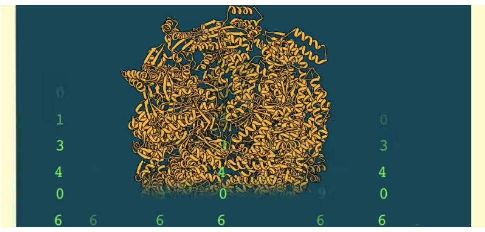
6. Usually, the proteins are in many different orientations. Images are grouped by orientation.



7. Images grouped by orientation are computationally combined to reconstruct the 3D shape of molecule.

8. Atoms are fit into the map to derive the 3D structure of the protein.





- 3DEM used to determine 3D structures of large macromolecular assemblies
- *imaging of many thousands of different single particles preserved in a thin layer of non-crystalline ice (cryo-EM) - these views show the molecule in myriad different orientations, a computational approach akin to that used for computerized axial tomography or CAT scans in medicine will yield a 3D mass density map*
- *cryo-electron tomography provides structural information at slightly lower resolution (i.e., protein domains and secondary structural elements)*

The practice of combining multiple experimental approaches is often referred to as Integrative or Hybrid Methods (I/HM) - proven to be very useful for multimolecular structures (complexes of ribosomes, tRNA and protein factors, and muscle actomyosin structures).

PDB contains structures for:

- ribosomes,
- oncogenes,
- drug targets,
- and even whole viruses.

- *we can often find multiple structures for a given molecule, or partial structures, or structures that have been modified or inactivated from their native form*

File formats: PDB, mmCIF, XML

- usually consists of header (info about the protein) + sequence of atoms and their coordinates
- *in a typical entry, there is a diverse mixture of biological molecules, small molecules, ions, and water - the names and chain IDs can be used to help sort these out*

In structures determined from crystallography, atoms are annotated with temperature factors that describe their vibration and occupancies that show if they are seen in several conformations.
NMR structures often include several different models of the molecule.

1. Beginner’s Guide to PDB Structures and the PDBx/mmCIF Format

PDBx/mmCIF format includes data items relevant to macromolecular crystallographic experiments

- overcomes limitations of the legacy PDB file format and supports data representing large structures, complex chemistry, and new and hybrid experimental methods
- PDB file format is not modified or extended to support new content (will become outdated)
- *supported by visualization applications: Jmol, Chimera, and OpenRasMol and structure determination systems: CCP4, Phenix*

Syntax & Format:

```
_data_item_category_name.data_item_attribute_name

_category.attribute        value      // key-value data category
_category.attribute_beta   90.00

// tabular data category (multiple values for each token):
_loop
_atom_site.id
_atom_site.label_sth
_atom_site.sth
_atom_site.string_example
1 SOME_LABEL 6.913   ‘vic, veci, carkou’
2 QUAK       8.888   ‘quak’
3 NECONECO   1.000   ‘A.B.C., Neco’

# hash at line beginning would indicate a comma/ separate categories

=> category is a tabular data structure where data items are the rows and the stored information are the
columns:
----------------------------------------------------------------
|__________________________atom_site___________________________|
| .id          | 1                   | 2      | 3              |
| .label_sth   | SOME_LABEL          | QUAK   | NECONECO       |
| .sth         | 6.913               | 8.888  | 1.000          |
| .string_name | ‘vic, veci, carkou’ | ‘quak’ | ‘A.B.C., Neco’ |
----------------------------------------------------------------
```

- if there are multiple columns within a data item/ group of data items in the same category, the category is preceded by a `loop_ token`

Parent-child relationships:

- created when data item occurs in multiple categories *(most commonly occurs for labels and identifiers which are reused throughout the dictionary)*

Chemical component dictionary - descriptions of all of the monomers and ligands in PDB structures (`CHEM_COMP_DICTIONARY` category group)

2. Dealing with Coordinates

- primary information stored in PDB: coordinate files (list of atoms in each structure and their 3D location in space + summary about the structure, sequence, and experiment)
- files are available in formats: PDBx/mmCIF, PDB, XML

Atomic-level Data

- PDB entry contains atomic coordinates for a collection of proteins, small molecules, ions and water
- each atom is identified by a sequential number, specific atom name, the name and number of the residue it belongs to, a one-letter code to specify the chain, its x, y, and z coordinates, and an occupancy and temperature factor (stored in the `_atom_site category`)

`ATOM` record - used to identify proteins or nucleic acid atoms

`HETATM` record - to identify atoms in small molecules

- most molecular graphics programs enable to color identified portions of the molecule selectively - pick out all of the carbon atoms and color them/ one particular amino acid and highlight it (by default, many molecular graphics programs do not display the water molecules)

Chains

- **biological molecules are hierarchical: atoms -> residues -> chains -> assemblies**
- coordinate files contain ways to organize and specify molecules at all levels
- in PDBx/mmCIF format, the looping nature of the records makes it easy to represent different chains and multiple molecules

Segment from entry `4hhb` showing the transition from chain A to chain B, where the chain is designated in the `_atom_site.label_asym_id` record and further identified in the `_atom_site.label_entity_id record`:

```
loop_
_atom_site.group_PDB
_atom_site.id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1    N  N    . VAL A 1 1   ?  6.204   16.869    4.854   1.00 49.05 ? 1  VAL A N   1
ATOM   2    C  CA   . VAL A 1 1   ?  6.913   17.759    4.607   1.00 43.14 ? 1  VAL A CA  1
<snip>
ATOM   1069 O  OXT  . ARG A 1 141 ? -9.474   13.682   -9.742   1.00 31.52 ? 1  ARG A OXT 1
ATOM   1070 N  N    . VAL B 2 1   ? 9.223   -20.614    1.365   1.00 46.08 ? 1  VAL B N   1
ATOM   1071 C  CA   . VAL B 2 1   ? 8.694   -20.026   -0.123   1.00 70.96 ? 1  VAL B CA  1
```

TER records are used to separate protein and nucleic acid chains:

- **indicates that the chains are not physically connected to each other**

```
ATOM   1067  NH1 ARG A 141         -10.147   7.455   -6.079  1.00 23.24   N
ATOM   1068  NH2 ARG A 141         -8.672    8.328   -4.506  1.00 33.34   N
ATOM   1069  OXT ARG A 141         -9.474   13.682   -9.742  1.00 31.52   O
TER    1070      ARG A 141
ATOM   1071  N   VAL B   1   9.223   -20.614          1.365  1.00 46.08   N
ATOM   1072  CA  VAL B   1   8.694   -20.026         -0.123  1.00 70.96   C
ATOM   1073  C   VAL B   1   9.668   -21.068         -1.645  1.00 69.74   C
ATOM   1074  O   VAL B   1   9.370   -22.612         -0.994  1.00 71.82   O
```

MODEL/ENDMDL keywords indicate multiple molecules in a single file:

- *MODEL keyword also used in biological assembly files to separate the many symmetrical copies of the molecule that are generated from the asymmetric unit*

Temperature Factors

If we were able to hold an atom rigidly fixed in one place, we could observe its distribution of electrons in an ideal situation. The image would be dense towards the center with the density falling off further from the nucleus. The experimental electron density distributions, however, usually have a wider distribution -> due to vibration of the atoms/ differences between the many different molecules in the crystal lattice. The observed electron density will include an average of all these small motions, yielding a slightly smeared image of the molecule.

- motions + resultant smearing of the electron density are incorporated into the atomic model by a B-value or temperature factor (amount of smearing is proportional to the magnitude of the B-value)

`_atom_site.B_iso_or_equiv`

- value < **10**: model of the atom is very sharp (the **atom is not moving much** => is in the same position in all of the molecules in the crystal),
- > **50**: atom is **moving so much that it can barely been seen** (often for atoms at the surface of proteins, where long side chains are free to wag in the surrounding water)

```
. . .
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
ATOM   1 N  N   . VAL A 1 1   ? 6.204   16.869  4.854   1.00 49.05 ? 1   VAL A N   1
ATOM   2 C  CA  . VAL A 1 1   ? 6.913   17.759  4.607   1.00 43.14 ? 1   VAL A CA  1
ATOM   3 C  C   . VAL A 1 1   ? 8.504   17.378  4.797   1.00 24.80 ? 1   VAL A C   1
```

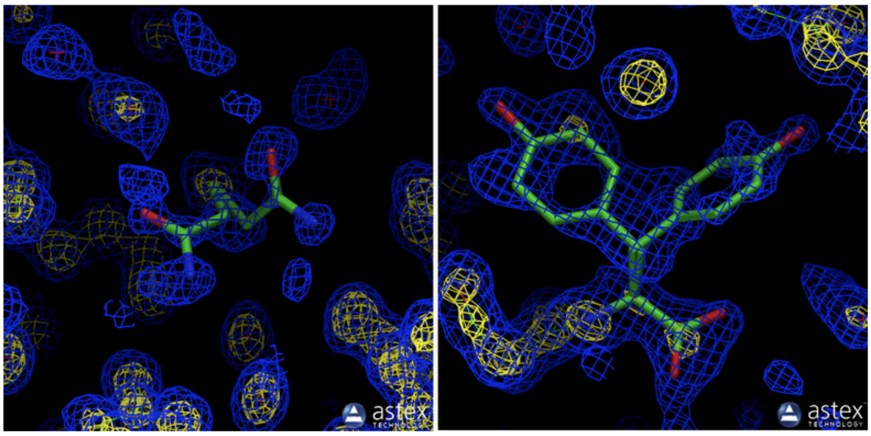
- **temperature factors == a measure of confidence in the location of atom**

Occupancy and Multiple Conformations

- macromolecular crystals are composed of many individual molecules packed into a symmetrical arrangement
- in some crystals: slight differences between each of these molecules *(a sidechain on the surface may wag back and forth between several conformations/ substrate may bind in two orientations in an active site/ a metal ion may be bound to only a few of the molecules)*

- => observe **occupancy**: for most atoms it has value **1** (== **the atom is found in all of the molecules in the same place in the crystal**), if a metal ion binds to only half of the molecules in the crystal –> occupancy of 0.5 *(we will see a weak image of the ion in the electron density map)*
- occupancies are also commonly used to identify side chains or ligands that are observed in multiple conformations -> it indicate the fraction of molecules that have each of the conformation - 2+ atom records are included for each atom, with occupancies like 0.5 and 0.5, or 0.4 and 0.6, or other fractional occupancies that sum to a total of 1

Alternate conformations in Myoglobin:



(alternate conformations: `_atom_site.label_alt_id`, occupancy: `_atom_site.occupancy`)

```
. . .
_atom_site.label_alt_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.label_entity_id
_atom_site.label_seq_id
_atom_site.pdbx_PDB_ins_code
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.pdbx_formal_charge
_atom_site.auth_seq_id
_atom_site.auth_comp_id
_atom_site.auth_asym_id
_atom_site.auth_atom_id
_atom_site.pdbx_PDB_model_num
<snip>
ATOM     65  C   GLN A   8   7.602  12.149  22.631  1.00   8.08    C
ATOM     66  O   GLN A   8   8.769  12.399  22.918  1.00   8.39    O
ATOM     67  CB AGLN A   8   5.987  11.822  24.520  0.57  13.03    C
ATOM     68  CB BGLN A   8   5.948  11.968  24.580  0.43   9.68    C
ATOM     69  CG AGLN A   8   7.030  11.303  25.506  0.57  16.30    C
ATOM     70  CG BGLN A   8   6.967  12.094  25.688  0.43  12.07    C
```

3. Biological Assemblies

- Biological Assembly and Asymmetric Unit are the same for many PDB entries - some are different (mostly those solved by X-ray crystallography)

The primary coordinate file of a crystal structure typically contains just one crystal asymmetric unit and may or may not be the same as the biological assembly. This introduction describes the terms asymmetric unit and biological assembly, lists where information about these can be found in various files formats (PDB and mmCIF), and explains how biological assembly files in the PDB archive are derived. Since the PDBML format is derived from the mmCIF format file, a separate discussion of this format is not included here.

4. Missing Coordinates and Biological Assemblies

o

5. Primary Sequences and the PDB Format

o

6. Hierarchical Structure of Proteins

o

7. Exploring Carbohydrates in the PDB Archive

o

8. Small Molecule Ligands

o

9. Molecular Graphics Programs

o

10. Computed Structure Models

o

12. Resolution

o

13. R-value and R-free

o

14. Structure Factors and Electron Density

o

```
text sample
text sample
```

```
text sample
```