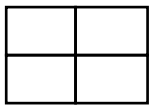
Silvia Onesti
silvia.onesti@elettra.eu

Elettra Sincrotrone Trieste

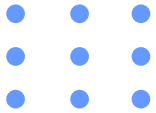
Macromolecular crystallography - Part 2

Small molecule crystal

small unit cell

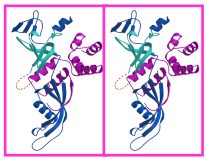


large reciprocal lattice

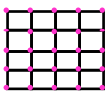


a few, very strong reflections; diffract to high resolution

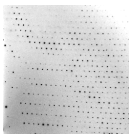
Protein crystal



large unit cell



small reciprocal lattice



many, very weak reflections; diffract to low resolution

How to solve macromolecular structures:

MIR (multiple isomorphous replacement)

Older method (Cambridge, 60') – relies on binding "heavy" atoms to the crystal and compare he diffraction pattern to the native. Trial and error search for good heavy atoms, it may take longer to get it right

MR (molecular replacement)

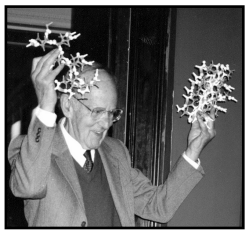
Older method (Cambridge, 70'-80') – relies on the expected similarity between the protein and another whose structure is known. Cannot solve de novo structures. Requires high homology (30% sequence identity?)

MAD (multiwavelength anomalous dispersion)

Relies on the absorption of specific wavelengths due to electronic transitions within the atom core. Similar to MIR but generally far quicker and more accurate. Requires high specification synchrotron radiation.

Some of the players...

Max Perutz (1914-2002), the inventor of MIR



In 1953, Perutz showed that the diffracted X-rays from protein crystals could be phased by comparing the patterns from crystals of the protein with and without heavy atoms attached.
In 1959, he employed this method to determine the molecular structure of hemoglobin. This work resulted in his sharing with John Kendrew the 1962 Nobel Prize for Chemistry.

Error treatment in MIR

The real breakthrough in using MIR for the determination of protein structures came when people learned how to deal with errors.

Blow D.M. & Crick F.H.C. (1959) "The treatment of errors in the isomorphous replacement method". *Acta Crystallogr.* **12**, 794-802

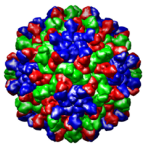
David Blow (1931-2004)

Francis Crick (1916-2004)

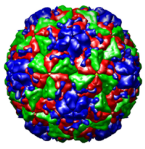
Error treatment is very complex – but it was an essential step in solving protein structures.

Molecular Replacement

The Molecular replacement method was mostly developed by Michael Rossmann. He used the structure of Tomato Bushy Stunt Virus, a plant virus, to determine the crystal structure of the Human Rhinovirus 14 (the common cold virus) in the early 80's.



Tomato Bushy Stunt Virus



Human Rhinovirus 14

However the theoretical basis were developed much earlier:
Rossman, M. G. and Blow, D. M. (1962). *ActaCryst*. 15:24-31.

Multiwavelength anomalous dispersion (MAD)

Close to the *absorption edge*, small changes in λ give rise to significant changes in the phase of the scattered wave (anomalous dispersion).

Using MAD, we can collect multiple "derivatives" from the same crystal just by changing the X-ray wavelength very slightly. Need tunable X-rays sources (synchrotrons!). Intensity changes (i.e. signals) are smaller than for MIR, but all the data are collected from the same crystal, so there is no non-isomorphism. MAD phases are more accurate that MIR phases.

We need anomalous scatterers!

At wavelengths of around 1 Å there are no electronic transitions for the "light" atoms (C, N, H, O, S, P) of biological molecules. Protein molecules may contain metals such as Zn, Fe and Cu which have accessible absorption edges.

The most common method is to replace Met residues by SeMet, as Se has an absorption edge at 0.98 Å

Phase problem solved? Calculating an electron density map

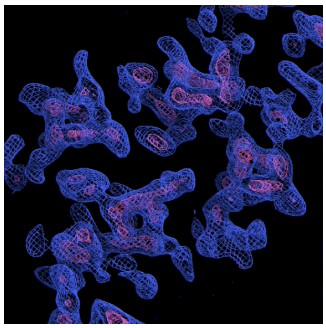
Once we have decent experimental estimates of the phases (from MIR, MR or MAD) we can obtain an electron density map.

At the end of the day, the only criterion for determining how good is your MIR/MR/MAD solution is whether the map is interpretable.

Can you build a polypeptide chain?

Electron density maps

After all this effort, we have a 3D map showing the shape of the protein:

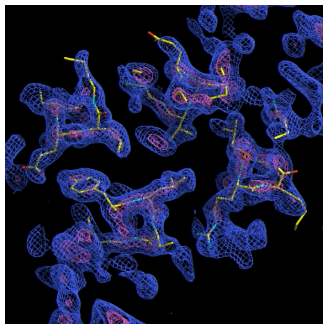


Electron density map displayed at two contour levels:

blue = 1 r.m.s (1σ);
magenta = 3 r.m.s (3σ)

Electron density maps

The task now is to try to fit an atomic model of the protein to the map…



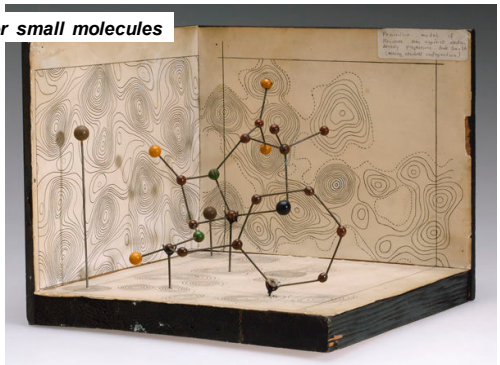
Electron density map displayed at two contour levels:

blue = 1 r.m.s (1σ);
magenta = 3 r.m.s (3σ)

Atomic model fitted to the map (in yellow)

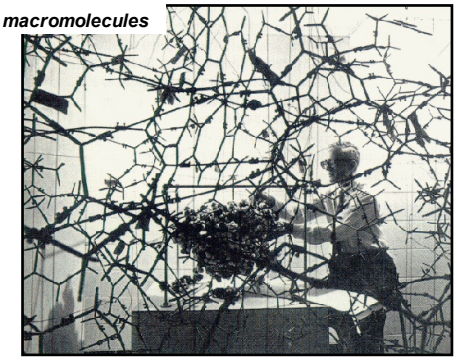
Model building in the early days…

for small molecules



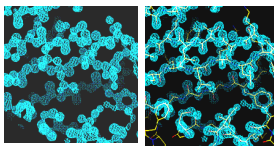
Model building in the early days…

for macromolecules



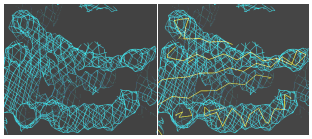
Maps and resolution

The task of model building is to interpret the electron density maps in light of chemical knowledge, basic stereochemistry, chemical sequence, etc…
The level of interpretation depends on the resolution of the map:



Here is a 1 Å map

At very high resolution, individual atoms can be seen and fitted in the electron density blobs: the problem therefore is reduced to 'join-the-dots'



Here is a 6 Å map

At very low resolution only large features can be seen - for example helices look like rods and β-sheets can barely be detected.

Refinement

Manual model building is not sufficient to build a completely accurate model but is required to get to the starting point for refinement.
Refinement is a process of optimisation of the atomic model to match the observed data and to conform to ideal stereochemistry

Successful structure determination usually requires several alternated rounds of model-building and refinement.

During refinement one calculate the expected diffraction pattern from the current model, compare it to the experimental diffraction data, and minimise the square of the differences.

Two problems:

- non linearity and presence of multiple minima
- low ratio observation parameters (especially at low res.); compensate by using stereochemical constraints.

The crystallographic R-factor

During the cycles of refinement, we calculate "*R* factors" to assess the progress.

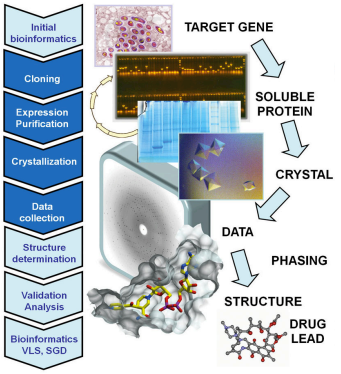
$$R_{factor} = \sum \frac{\left| |F_{obs}| - |F_{calc}| \right|}{|F_{obs}|}$$

R = "residual" = fractional difference between observed and calculated diffraction - a sort of "fractional error"

To monitor the refinement, we calculate *R* after each cycle; if things are going well, *R* reduces.

However, with complicated refinements like these, it is possible to "over-fit" the data – for cross-validation we take away 5% of the data (which we do not use in the refinement, to monitor the agreement. This is known as R_{free}

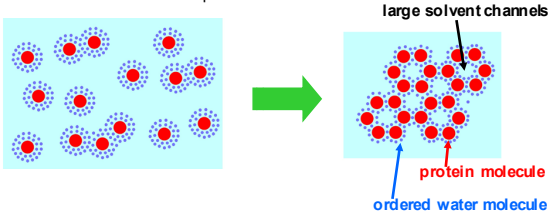
The crystallographic pipeline



- no protein, no party: a lot of your time will be spent in cloning, protein expression and purification
- no crystal, no party! crystallization is the real bottleneck!
- once you get crystals, the fun begins:
  - DATA COLLECTION & PROCESSING
  - SOLVING THE PHASE PROBLEM
  - MODEL BUILD AND REFINEMENT

Crystallisation

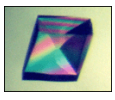
Protein crystallisation is the transfer of protein molecules from an aqueous solution to an ordered solid phase.



Protein crystallisation is controlled by:
- thermodynamic factors governing the solubility
- kinetic factors governing nucleation and growth

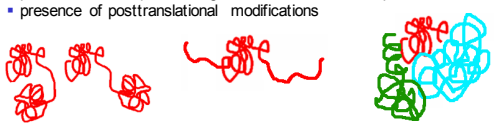
Easy/difficult to crystallise?

A lot of interesting proteins are difficult to crystallise.

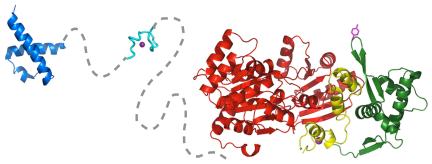


Best cases:
- single proteins
- rigid domain structure
- one dominant conformation

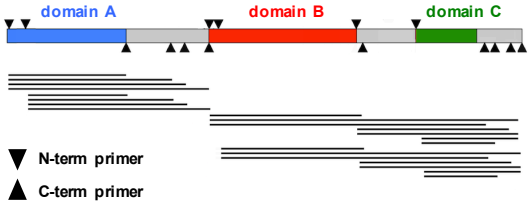
Worst cases:
- multi-domain proteins with flexible unstructured linkers
- proteins with flexible N- and/or C-termini
- proteins that are part of large macromolecular complexes
- presence of posttranslational modifications



We are more and more confronted with proteins that contain unstructured regions, exhibit polymorphism, have multi-domain architectures: these are also challenging target for crystallisation.



"Optimise" proteins for crystallography

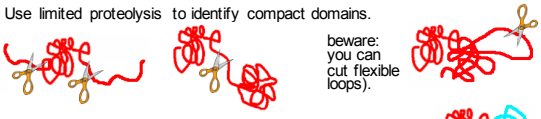


Clone each of the fragment in more than one vector, i.e. N-term His-tag, C-term His-tag, SUMO-tag, etc…

"Optimise" proteins for crystallography

Use bioinformatics (database searches, sequence alignments) to identify "core domains" that can be expressed in a soluble form.

Use limited proteolysis to identify compact domains.



beware: you can cut flexible loops).

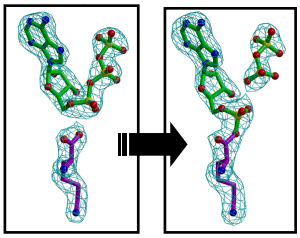
Co-express proteins that are part of the same complex.

Add ligands/inhibitors/cofactors/metals to stabilise one conformation.

Avoid/encourage post-translational modifications (mutagenesis of target residues, mimic modifications by mutagenesis, change expression system/cell lines, so that the process does/does not occur).

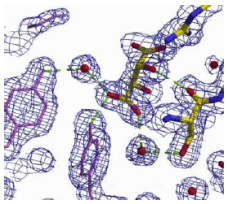
Does crystallisation affects the structure?

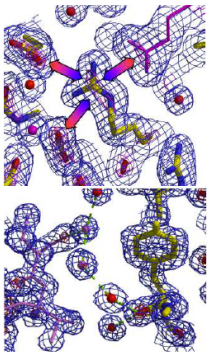
- The forces that hold molecules in a lattice are very much weaker than those that hold protein structures together, so gross conformational changes triggered by crystallisation are unlikely (but they do occur). However the crystal lattice may favour a conformation that is not dominant in solution.
- Proteins crystallised in different crystal forms are often identical or almost identical.
- Some enzymes retain their biological activity within the crystal (a strong indication of a native-like structure).



Holding protein crystals together

Very few & weak interactions hold the fragile protein crystals together:
- H-bonds
- VdW interactions
- salt bridges
- often mediated by H_2O





How to read a PX paper

Read Materials and Methods (even if on Supplementary materials)!

How many molecules in the asymmetric unit? if there are multiple copies, check for differences that may reveal flexibility or alternative configurations.

If you are interested in a substrate/inhibitor ideally there should be a picture showing the electron density map around it.

Is the level of detail in the interpretation compatible with the resolution?

Check Data Collection statistics

how good was the data? at which resolution? how complete? is the "claimed" resolution realistic?

Check Refinement statistics

how well does the model agree with the data?

TABLE 1

How to use a PDB file

"All crystallographic models are *not* equal. … The brightly coloured views of a protein model, *which are in fact more akin to cartoons than to molecules*, endow the model with a concreteness that exceeds the intentions of the thoughtful crystallographer. It is impossible for the crystallographer, with vivid recall of the massive labour that produced the model, to forget its shortcomings. It is all too easy for users of the model to be unaware of them. It is also all too easy for the user to be unaware that, through temperature factors, occupancies, undetected parts of the protein, and unexplained density, crystallography reveals more than a single molecular model shows."

Crystallography Made Crystal Clear,
by Gale Rhodes
University of Southern Maine,
Portland, USA

How to use a PDB file

How were the coordinates generated?
- crystallography (roughly 90% of the PDB)
- NMR (roughly 10%)
- modelling (roughly 1-2%)

What do I want to get out of it?
- the topology of the molecule? which aa is on a particular surface?
- the hydrogen bonding pattern?
- the network of water molecules?
- the detailed interactions of a substrate/inhibitor?

Resolution? which is the level of details I can trust?

Which is the "biological unit"?

Temperature (B) factors

Thermal motion ↦ atoms oscillate around an average position
Time-scale of a diffraction experiment >>> period of thermal oscillation

➡ we see a time-averaged electron-density distribution

Electron density gets spread out over a larger area and is weaker.

In MX a high temperature factor suggests both thermal motion and static disorder.

If portions of a chain have very high mobility or disorder, they produce low electron density, making it impossible to assign positions to atoms in such portions.

We model this by assigning to each atom a "temperature factor" – the higher, the more mobile/disordered. In many graphics programs there is a temperature colour scheme that assigns warm colours (red-orange) to high temperature factors, and cool colours (blue-green) to low ones.

Occupancies

The PDB has an "occupancy" estimate for each atom – from 0 to 1.00.

Even at medium-high resolution (1.8-1.5) it is not possible to refine occupancies and particularly to decouple them from temperature effects.

Cases in which one may refine occupancies:

- at high resolution sometimes people model multiple side-chain conformations and the occupancies of the two configurations are refined
- if we know a residue is a Lysine, but we see only density for Alanine, we may build a lysine but flag the atoms that we don't see as zero occ.
- to get an estimate for the occupancy of a substrate/inhibitor, even at medium resolution, one can fix the B factors to a level compatible with the average of the surrounding atoms and do a cycle of Occ refinement

Disordered residues

Different strategies - example a disordered lysine, with no electron density beyond Cβ:

- build an alanine and call the residue alanine

  The PDB validation system may complain

- model a lysine (based on ideal torsion angles and chemical constraints) and set all the occupancies beyond Cβ to zero.

  Dangerous: not many people beyond crystallographers look at the Occupancy column - I had people measuring distances involving residues with zero occupancies…

- build an alanine and call the residue lysine

Atomic details

PX usually cannot resolve the positions of hydrogen atoms, except for the small number of structures at resolution below 1.2 Å. Some newer X-ray crystal diffraction PDB files contain hydrogen positions; these hydrogens were added by modelling.

PX cannot reliably distinguish nitrogen from oxygen from carbon. This means that the chemical identity of the terminal side-chain atoms is uncertain for Asn, Gln and Thr and is usually inferred from the protein environment of the side chain (*i.e.* the side chain orientation which forms the most hydrogen bonds or makes the best electrostatic interactions is selected and built by the crystallographer as the most plausible choice).

Sometimes there is also uncertainty about whether an atom that is not part of the protein is a bound water oxygen or a metal ion.

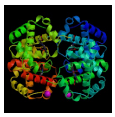
However, there are methods that can be used to identify metals (soak with and isoelectric heavier metal: Mg/Mn, Zn/Cd/Hg, Ca/Ba, Na/Rb; use anomalous data, check potential ligands/stereochemistry…)

Biological unit - 1

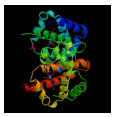
When crystallographic structures are deposited in the PDB, the primary coordinate file generally contains one asymmetric unit (A.U.)

An asymmetric unit may contain:
- one biological molecule
- a portion of a biological molecule
- multiple biological molecules

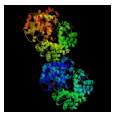
Haemoglobin ($\alpha_2\beta_2$), provides a good example of each of these cases:



Entry 2HHB contains **one** molecule ($\alpha_2\beta_2$) in the A.U



Entry 1HHO contains **half** a molecule ($\alpha\beta$) in the A.U. A crystallographic 2-fold axis generates the 4 chains of the haemoglobin molecule.



Entry 2HV4 contains **two** molecules (**2** $\alpha_2\beta_2$) in the A.U.

Biological unit - 2

The biological unit is the macromolecule that has been shown to be or is believed to be functional.

Depending on the asymmetric unit, space group symmetry operations consisting of either rotations or translations must be performed in order to obtain the complete biological unit. However, if the asymmetric unit contains multiple biological molecules, then one copy may be selected.

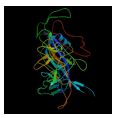
In the examples before:

- 2HHB: no operation necessary
- 1HHO: application of a crystallographic symmetry operation (a 180 rotation around a crystallographic two-fold axis) to produces the complete biological unit.
- 1HV4: PDB file contains **two** structurally similar, but not exactly identical copies of the biological unit. Need to select one.

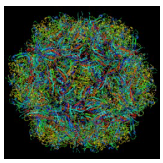
Biological unit - 3

Occasionally, a molecule may appear multimeric in the crystal, but this has not been proven through other studies to be biologically relevant.

In certain cases, most notably viral capsids, the coordinate file may contain only part of the asymmetric unit. Here, the complete asymmetric unit can be generated by applying NCS operators to the coordinates. This complete asymmetric unit in turn may either form the biological unit (coat protein) or, in some complicated cases, only part of the biological unit. In the latter cases crystallographic symmetry operators may have to be applied to form the full biological unit (viral capsid).



1P5Y: PDB file content (1 chain)



The viral capsid (biological unit) need 60 copies of the chain shown on the left!

Biological unit - what to do?

In PDB files, info about the biological unit is given in Remarks 300/350 (in older files other Remarks may be used).

Remark 300 provides a description of the biological unit in free text.

Remark 350 presents all transformations, both crystallographic and non-crystallographic, to be applied to the coordinates in the entry.

```
REMARK 350 GENERATING THE BIOMOLECULE
REMARK 350 COORDINATES FOR A COMPLETE MULTIMER REPRESENTING THE KNOWN
REMARK 350 BIOLOGICALLY SIGNIFICANT OLIGOMERIZATION STATE OF THE
REMARK 350 MOLECULE CAN BE GENERATED BY APPLYING BIOMT TRANSFORMATIONS
REMARK 350 GIVEN BELOW. BOTH NON-CRYSTALLOGRAPHIC AND
REMARK 350 CRYSTALLOGRAPHIC OPERATIONS ARE GIVEN.
REMARK 350
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C, D
REMARK 350 BIOMT1 1 1.000000 0.000000 0.000000 0.00000
REMARK 350 BIOMT2 1 0.000000 1.000000 0.000000 0.00000
REMARK 350 BIOMT3 1 0.000000 0.000000 1.000000 0.00000
REMARK 350 BIOMOLECULE: 2
REMARK 350 APPLY THE FOLLOWING TO CHAINS: E, F, G, H
REMARK 350 BIOMT1 2 1.000000 0.000000 0.000000 0.00000
REMARK 350 BIOMT2 2 0.000000 1.000000 0.000000 0.00000
REMARK 350 BIOMT3 2 0.000000 0.000000 1.000000 0.00000
```

Here the PDB contains 2 copies (2 biomolecules), each made up of 4 chains.

Biomol1: A+B+C+D
Biomol2: E+F+G+H

Biological unit - another example

A complex example - a viral capsid -- Entry 1P5Y

The coordinate file contains 1/60 of the asymmetric unit. The asymmetric unit, which in this case is the same as the biological molecule, is created by applying the 60 matrices given in remark 350 to chain A.

```
REMARK 350 GENERATING THE BIOMOLECULE
REMARK 350 COORDINATES FOR A COMPLETE MULTIMER REPRESENTING THE KNOWN
REMARK 350 BIOLOGICALLY SIGNIFICANT OLIGOMERIZATION STATE OF THE
REMARK 350 MOLECULE CAN BE GENERATED BY APPLYING BIOMT TRANSFORMATIONS
REMARK 350 GIVEN BELOW. BOTH NON-CRYSTALLOGRAPHIC AND
REMARK 350 CRYSTALLOGRAPHIC OPERATIONS ARE GIVEN
REMARK 350 REMARK 350
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A
REMARK 350 BIOMT1 1 1.000000 0.000000 0.000000 0.00000
REMARK 350 BIOMT2 1 0.000000 1.000000 0.000000 0.00000
REMARK 350 BIOMT3 1 0.000000 0.000000 1.000000 0.00000
REMARK 350 BIOMT1 2 0.313670 0.838710 0.445170 17.37754
REMARK 350 BIOMT2 2 -0.779790 0.495040 -0.383210 117.73824
REMARK 350 BIOMT3 2 -0.541780 -0.226940 0.809300 73.47149
REMARK 350 BIOMT1 3 0.848650 0.432000 0.305240 -18.99369
REMARK 350 BIOMT2 3 0.097760 0.439030 -0.893140 88.98281
REMARK 350 BIOMT3 3 -0.519850 0.787800 0.330350 124.14395
.....
REMARK 350 BIOMT1 60 0.162270 -0.123390 0.979000 -26.80613
REMARK 350 BIOMT2 60 -0.123390 -0.986900 -0.103830 23.40252
REMARK 350 BIOMT3 60 0.979000 -0.103930 -0.175370 34.77368
```

Matrix 2 is highlighted as an example -- the rotation in red and the translation in blue.

Skip matrices 4-59

Structures can be VERY useful in biology! However, remember that:

- Crystal lattice interactions may favour a conformation that is not dominant in solution (non a *very* common case, but a possibility).
- When interpreting a structure in terms of biochemistry, keep in mind that the structure may represent only a snapshot in a complex reaction pathway.
- When only the structure of an isolated protein is known, do not trust blindly models of complexes with other proteins or with nucleic acid.
- To FULLY UNDERSTAND the behaviour of a protein in the cell one needs to integrate data from cell biology, biochemistry, molecular biology, bioinformatics, various structural biology techniques, computational biology, biophysical chemistry, nanobiophysics, single molecule studies.

Elettra Structural Biology Laboratory – Facebook page

https://www.facebook.com/Elettra-Structural-Biology-Lab-2406267665542 84