
Lecture 30 X-ray Crystallography

Microbiology 343

David Wishart, University of Alberta, Edmonton, AB

Protein Structure Determination by X-ray Crystallography

Classically, the determination of macromolecular structures through X-ray crystallography requires 5 steps:

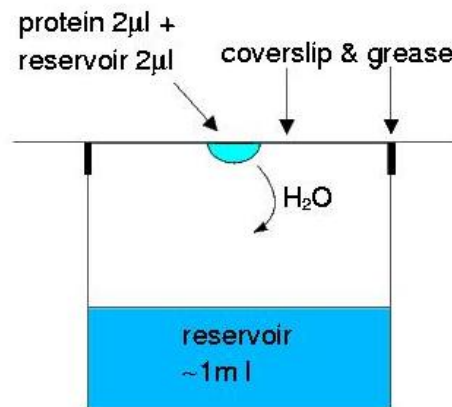
- 1) Crystallization
- 2) Data collection
- 3) Phase determination
- 4) Chain tracing
- 5) Refinement

Crystallization

The first prerequisite to solve a 3D structure of a protein or any other macromolecule is a well ordered crystal that diffracts X-rays strongly. Crystallization, especially with proteins, can be quite difficult to achieve and crystal growth can be painstakingly slow (taking months or even years before a crystal grows to more than 0.5 mm). Crystallization is key to all aspects of X-ray crystallography yet it is the least understood and least controllable component to the whole process. Crystallization is regarded by many as a "**black art**" and represents the weakest link in macromolecular structure determination.



A **pure** (>97%) and **homogeneous** protein sample is crucial to a successful crystallization attempt. In addition, crystallization is also critically dependent on pH (close to the pI), temperature (low temp is best), nature of the solvent (non-chaotropic), nature of the precipitant (PEG or ammonium sulfate) and the presence of added ions or ligands. Crystals are most often formed when protein molecules are precipitated from a **supersaturated solution**. The most frequently used technique for creating a supersaturated solution is the hanging drop method. This involves placing a drop (10 μ l of a 10 mg/mL buffered solution) of the concentrated protein solution onto a cover slip, inverting the cover slip and then placing it over a 96-well plate containing saturated salt or PEG solutions in each well. In doing so, the protein solution in the drop is brought very gradually to supersaturation by loss of water to the larger reservoir containing such precipitants as salt solutions or polyethylene glycol.

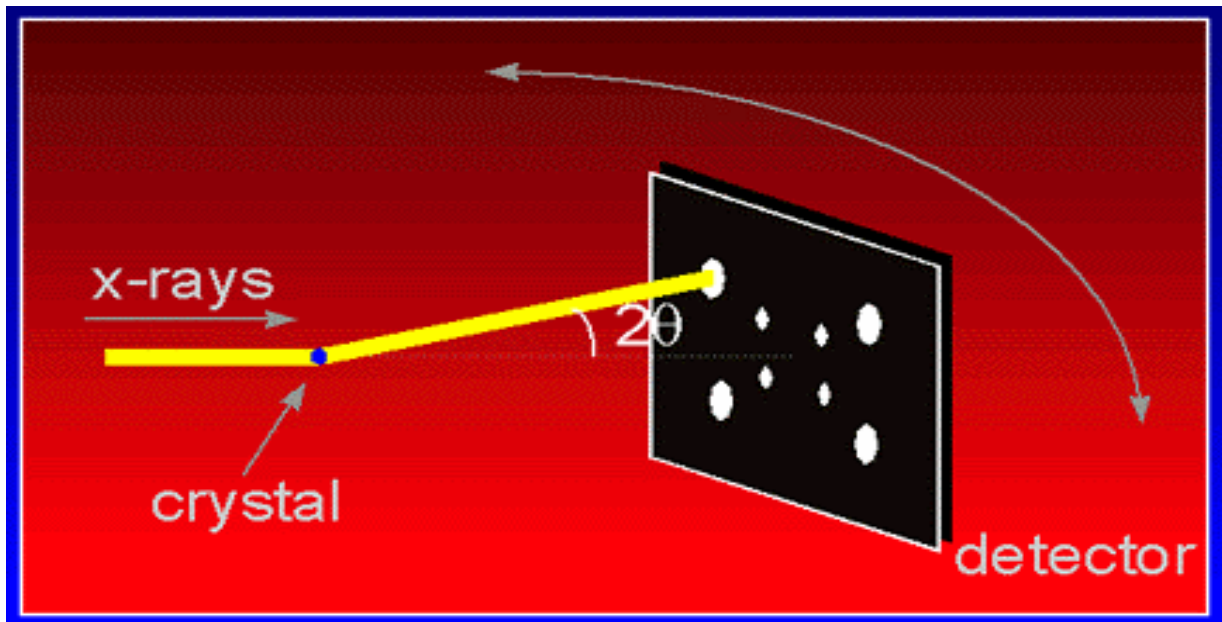


Crystals can vary tremendously in their size, water content (30 - 60 %) and overall geometry or shape. Furthermore, the conditions of crystallization can also affect the packing of individual molecules. As a result it is important to know and accurately characterize the crystal form that has been generated when the crystal growth is complete.

Data Collection

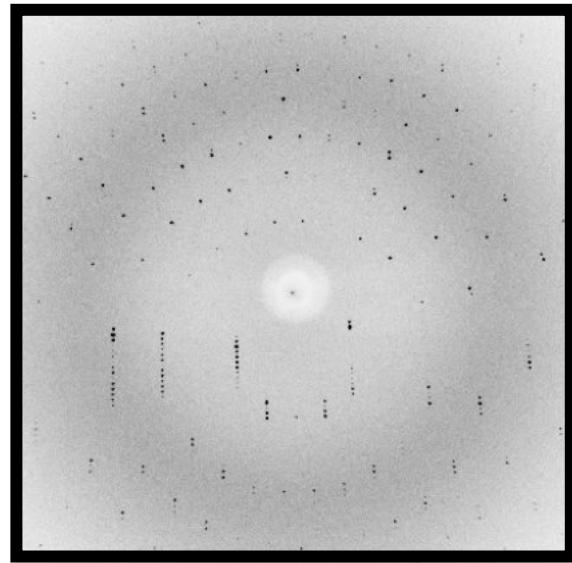
Once a crystal of sufficiently large size and stability has been generated, it is usually mounted in a thin glass tube and placed in front of a monochromatic (single wavelength) X-ray source. Most X-ray sources are **rotating anode** generators which create their X-rays by accelerating electrons into a copper (or other metal) plate which, in turn, ejects X-rays of a defined wavelength (1.54 Angstroms). In some cases, particularly if the crystals are small, the X-ray generator of choice is the **synchrotron**. This high-energy

particle accelerating device allows one to generate monochromatic X-ray beams that are 10 times more intense than the conventional rotating anode generators.



As the X-rays pass through the crystals (which are slowly rotated) they are deflected (diffracted) and then detected using either X-ray film or an electronic area detector (digital film). A typical X-ray diffraction experiment will involve the collection and detection of **tens of thousands** of diffraction spots. All of which have to be measured, entered and processed by computers.

Because X-rays are high energy forms of electromagnetic radiation and because their wavelength is roughly equal to the size of the objects they are interacting with, one gets a scattering or diffraction pattern instead of a "real" picture of the object under analysis. Diffraction patterns can also be generated by visible light when it interacts with images or slits that are comparable to the wavelength of light. In this regard it is possible to simulate what kind of diffraction patterns would be generated by using visible light as a pseudo X-ray source and small cut-out models to simulate the atoms in the molecule. As can be seen below, diffraction images bear no resemblance to the real image. In fact they are inverse Fourier transforms of the real image.



The appearance of intense bright spots and dark areas in between arises through constructive and destructive interference from the scattered X-rays. The interference patterns (their position and intensity) are determined from Bragg's law

$$1) \quad n\lambda = 2d\sin\theta$$

When there are many scattering centres, Bragg's law is more appropriately expressed in terms of a **structure factor** $F(S)$ which is defined as the ratio of the X-rays scattered by any real sample to that scattered by a single scattering centre at the origin. The square of the structure factor (actually when multiplied by its complex conjugate) $F(S) \times F(S)^* = |F(S)|^2$ is equal to the **intensity** $I(S)$ of the observed diffraction spot. The mathematical definition of the structure factor is:

$$2) \quad F(S) = \int_V \rho(r) e^{-2\pi i S \cdot r} dV$$

where $\rho(r)$ is the electron density distribution of the sample (or protein). The inverse **Fourier Transform** of the above equation, gives us the electron density distribution (i.e. the atomic arrangement of the molecule)

$$3) \quad \rho(r) = \frac{1}{V} \int dS e^{-2\pi i S \cdot r} F(S)$$

The above equations tell us that each diffraction spot is the result of interference of all X-rays with the same diffraction angle emerging from all atoms. For a typical protein

crystal, each of the 20,000 diffraction spots represents scattered X-rays from each of around 1500 atoms. It is also worth noting that the degree of resolution which you can work with is strongly dependent on the number of spots and area (on the image plate) that the diffraction spots cover. The more diffraction spots that you see or record, the more information about the image is recoverable. Therefore, the amount of diffraction data measured directly affects the **resolution** of the structure or image that you generate by inverse Fourier transformation. If your crystals are said to diffract poorly, it usually means that only an inner core of diffraction spots are visible. If it was a perfect world, information about the phase, intensity, position and wavelength would all be measurable and known. Unfortunately, only the **intensity and wavelength** of the diffracted X-rays can be known. No information about the **phase** is detectable in a real X-ray data set. However, we need to know all three properties for all the diffracted data to determine the position of the scattering atoms in the macromolecule. How do we find the phases? This is called the phase problem in crystallography and it is the second most difficult aspect (next to the crystallization problem) of this technique to overcome.

Phase Determination

In small molecule crystallography the phase problem was solved quite some time ago using so-called direct methods. These methods, which are computationally very intensive have been applied to very small proteins (really peptides) with some success as well. However, to solve the phase problem for very large proteins, it has always been necessary to use a technique developed by Max Perutz and Andrew Kendrew called **Multiple Isomorphous Replacement (MIR)**. This involves the introduction of a heavy metal atom (like Mercury or Uranium) into the crystal that acts as a significant and new scattering center. These heavy atoms must be limited in number (2 or 3), they must bind to well defined places, and they should not change the structure of the protein or the crystal (i.e. they must be isomorphous). Typically many heavy metal complexes are formed between free **thiols on a protein**.

By comparing the intensity differences between the underivatized or unsubstituted molecule with the MIR molecule you can deduce the positions of the heavy atoms in the crystal unit cell. This, in effect, reduces the problem to solving the X-ray structure of a 2 or 3 atom molecule. These difference spectra are called **Patterson maps** and the vectors that can be drawn to the diffraction spots in these Patterson maps actually define the spatial arrangement of the heavy atoms. From this positional information we can back-calculate the **phases** and amplitudes of the heavy atoms. In addition, we know the

amplitude (intensity) information from the protein alone and we know the amplitude (intensity) of the protein with the heavy metals. So we have data for one set of phases and three amplitudes. This information is usually sufficient to give an estimate of the phase of the protein. Unfortunately the problem is **underdetermined** (one equation and two unknowns) so in order to distinguish between the two possible phase angles a second heavy metal MIR must be prepared. This makes it possible to distinguish which solution was correct and one is now able to calculate the phases for the protein of interest.

To summarize, the phase problem for macromolecules is solved as follows:

- 1) Diffuse a heavy metal into the protein crystal (make sure it is isomorphous)
- 2) Collect data from this derivative
- 3) Calculate the difference between the MIR derivative and the native crystal diffraction data (i.e. determine a Patterson map)
- 4) Determine the location of the heavy atoms in the unit cell
- 5) Determine the phases of the heavy atoms
- 6) Use these phases to estimate the protein phases
- 7) Repeat steps 1 to 6 a second or third time to completely determine the protein phases.

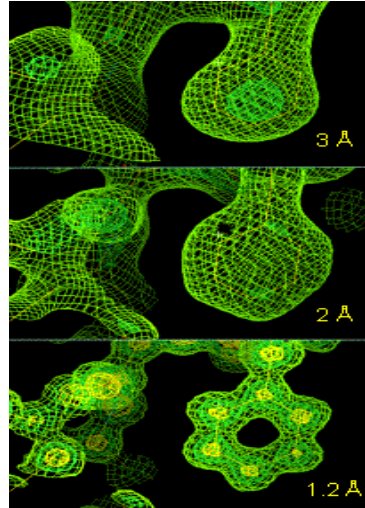
Chain Tracing

Once the phases to the protein have been determined, it is possible to use the following equation (and a fast computer) to calculate the electron density of the protein molecule:

$$4) \quad \rho(r) = 1/V \int dS e^{-2\pi i S \cdot r} F(S)$$

This electron density map then has to be fitted with the polypeptide chain of the protein of interest. This is a very **subjective** component to the project and can take many days and weeks if you are working with low quality data or the unit cell contains many polypeptide chains. Most chain tracing is done using computer graphics packages such as Xtal View, FRODO, BioSym and others which assist crystallographers with quickly and interactively placing, rotating and adjusting amino acid side chains. Prior to computer graphics, most of this work had to be done with large **tinker-toy** models and special optical devices. The quality of the electron density map depends strongly on the

resolution of the data. Real examples of what electron density maps look like at different levels of resolution are shown below:



As is evident with low resolution data, one is often left guessing. In this regard, building the initial model of a protein (often working with low resolution data) is usually a **trial and error** process. Maps showing continuous electron density from N to C terminus are extremely rare. More usually one produces a number of matches between the electron density and discontinuous regions of the sequence that one may be more certain about (particularly around Phe, Trp and Tyr residues). Eventually one is able to fit all the pieces together and this initial model is further refined.

Refinement

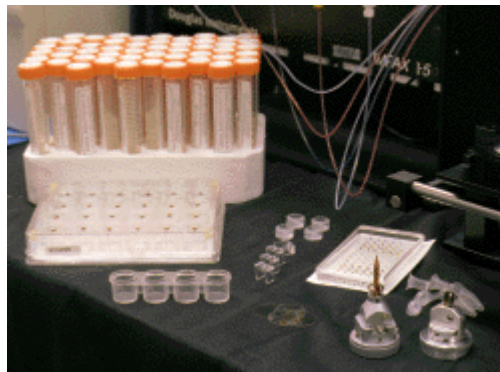
All initial 3D models of proteins contain errors. Most of these errors can be removed by crystallographic refinement. In this process the model is progressively changed to minimize the difference between the experimentally observed diffraction amplitudes and those calculated for a hypothetical crystal containing the model instead of the real molecule. The difference is called the **R factor** (where R stands for residual disagreement) and it is expressed as follows:

5)
$$R = \frac{\sum |F_{\text{obs}}| - |F_{\text{calc}}|}{\sum |F_{\text{obs}}|}$$

R factors can vary between 0.00 (which is exact agreement) and 0.59 which is total disagreement. In general, an R factor between **0.15 and 0.20** indicates a well determined or well-refined crystal structure for proteins. R factors of less than 0.05 are not uncommon for small molecules. Many refinement packages also include **energy minimization** and molecular dynamics that help sort out problems of steric clashes and atomic overlap that may not be obvious to the crystallographer when they initially build their model. These **indirect refinement** methods are a useful adjunct to direct refinement methods as often one method can help the other out of a local minimum.

HT X-Ray Methods

In the area of X-ray crystallography, protein crystallization continues to be the most problematic and time consuming aspect of the whole process. Fortunately, with the advent of robotic crystallization systems it is now possible to prepare and sample



thousands of crystallization conditions in a short time with minimal human effort. This high throughput testing is leading to more frequent crystallization successes in much shorter time periods.

Similarly the use of dynamic light scattering or ultracentrifugation as a rapid screen to determine the mono-dispersity of protein samples is leading to significant improvements in crystallization successes (Ferre-D'Amere et al. 1994). Recent studies indicate that monodisperse protein samples typically have a 75% chance of yielding suitable X-ray crystals, while polydisperse samples have only a 10% chance. This simple pre-screen could save significant time and effort in the candidate protein selection process.

In the area of X-ray technology, several significant advances are helping crystallographers collect, phase and analyze their data in remarkably short periods of time. For instance, the use of high intensity undulating X-ray synchrotron beams has had a major impact on the rate at which crystallographers can collect data and the crystals they can collect on.



Under appropriate cryogenic conditions (-170°) one can greatly limit radiation damage and collect data on remarkably small (<0.1 mm) or unstable crystals in minutes or even seconds. The use of charged couple devices (CCD) as detectors instead of X-ray film also allows for rapid and accurate digitization of diffraction data. With the advent of tunable X-ray sources from synchrotrons, the use of multi-wavelength anomalous dispersion (MAD) phasing with selenomethionyl labeled proteins (1 S-Met per 40 residues) now permits phase determination (minutes) of proteins without the use of isomorphous derivatives. Even the use of direct methods now means certain small proteins can be solved without the need for additional data sets or phasing data. With rapid phasing and structure solving programs such as SOLVE (<http://www.solve.lanl.gov/>) and the implementation of automated chain tracing systems such as molecular scene analysis and threading (Leherte et al., 1994), it is not unreasonable to have a protein structure solved within 24 hours of first mounting the crystal on the beam line.

These and other advances in X-ray crystallography have led to a number of modest attempts at structural proteomics or high throughput X-ray structure generation.

Based on a several published and unpublished results it appears that the typically success rate is about 10%. That is, only one in 10 proteins that is successfully expressed will have its X-ray structure solved within a year. Evidently the most significant bottleneck to structural proteomics via X-ray crystallography is not associated with data collection or analysis, rather it is with crystallization. In other words, the mechanics of X-ray structure determination (data collection, phasing, chain tracing and refinement) are very close to being fully automated. The only problem is that crystallization is not.