Advanced Physical Chemistry
Crystal Structure Determination by X-ray Diffraction
Virginia B. Pett
The College of Wooster

Structural information we get from X-ray diffraction experiment

Type and position of atoms in unit cell

x, y, z coordinates of atoms

bond lengths

intermolecular contact distances give indication of hydrogen bonds, ionic interactions, van der Waals forces, ring stacking

torsion angles

conformation of molecule

Thermal motion

possible reactivity

possible alternate conformations

Requirements for single-crystal X-ray diffraction experiment

- Single crystal
- X rays from X-ray generator or synchrotron radiation
- Detector
- Goniometer to move crystal and detector into position to satisfy Bragg's law
- Computer and software
- Crystallographer

Information we measure/deduce from the diffraction pattern

Diffraction angle theta (θ) ➞ (*hkl*) planes

Geometry of the diffraction pattern ➞ size and shape of unit cell

Density of crystal + unit cell info ➞ number of molecules in unit cell

Systematic absences of diffraction spots ➞ symmetry of lattice

Intensity of each diffraction spot + phase of diffracted beam ➞ ➞ type of atoms, x,y,z coordinates of atoms in unit cell. This last step is the difficult one!

The diffraction experiment in greater detail

Diffraction depends upon constructive and destructive interference of X rays scattered by the atoms in the crystal.

The detector records the direction and intensity of X rays diffracted by crystal: thousands or millions of diffraction data recorded on photographic film, fluorescent screen, or electronically. The data set is called the diffraction pattern.

Each diffraction spot is produced by the constructive interference of the X rays from a set of planes (*hkl*) in the crystal. The spacing of the diffraction pattern is inversely dependent upon the spacing of the planes. Braggs's Law: $n\lambda = 2d \sin\theta$

The intensity of a diffraction spot depends upon the position of atoms relative to each set of planes. Therefore, the intensities contain the information about the positions of the atoms in the unit cell.

The resolution of the structure depends upon the spacing of the planes from which diffraction measurements were made. A high-resolution structure is one where diffraction data have been observed from closely spaced planes. For small molecules, high resolution is less than 1.0 Å. For a protein, 1.5-2.0 Å is high resolution.

We want to calculate electron density at grid points in the unit cell; atoms are at points x,y,z of high electron density

Phase problem: we know only the amplitude of the diffracted X rays, not the phases. We have measured the intensities of the diffracted X rays, but we can calculate only the absolute values of the structure factors, since they are the square roots of the observed intensities.

"Solving the phase problem" = getting approximate phases = getting a trial model for the molecule

Heavy atom method = Patterson method. If we have a structure in which one or two atoms have many more electrons than the other atoms, the Patterson method is appropriate.

First, a Patterson map is calculated directly from the intensities of the spots. Since this is an $F(hkl)^2$ map, we don't need to know the phases. This Patterson map has peaks and valleys like a geological map. The map shows *vectors* between two atoms. The higher the peaks, the greater the electron density in the two atoms. The strongest vectors give us information about vectors between the one or two "heavy" atoms. We can deduce where the heavy atoms are from the Patterson map. Then we can calculate approximate structure factors, using only the x,y,z position of the one or two heavy atoms (eq 1). The structure factor summation is a truncated Fourier summation, since only the positions of a few heavy atom(s) are known.

The electron density map based on the calculated structure factors is likewise approximate (eq 2). Nevertheless, we can usually see additional atoms (or the whole structure) in the electron density map. When the new atoms are added to the structure factor series, the calculation is more accurate, and the electron density map is more accurate also. At this point we can often find hydrogen atoms, and then proceed to refinement.

$$F(hkl) = \sum_j f_j e^{2\pi i(hx+ky+lz)} \quad (1)$$

$$\rho(x,y,z) = \frac{1}{V}\sum_h \sum_k \sum_l |F(hkl)| \cos[2\pi(hx+ky+lz) - \phi(hkl)] \quad (2)$$

In eq 1, x,y,z are the atomic positions of only the heavy atoms initially. At a later stage, all the atoms are included in the structure factor summation. In eq 2, x,y,z are the arbitrary grid points in the unit cell where we want to calculate what the electron density is at that point. ϕ(hkl) is the *phase* of the diffraction from planes hkl. We know the approximate phase for each reflection from eq 1. The phase information is combined with the observed diffraction amplitude [absolute value of F(hkl)], and an approximate electron density map is calculated.

Multiple Isomorphous Replacement (MIR). This method of solving protein structures uses the *anomalous scattering* from the heavy atom to locate the heavy atom in the unit cell. The rest of the structure solution continues as explained above. Often more than one heavy atom derivative is required to phase the reflections. A common problem is that the native structure and the derivative structure aren't exactly the same (aren't isomorphous), making the phasing inaccurate.

Molecular replacement from similar structure or from modeled structure (HIV protease structures were done this way). Approximate phases can be calculated from the known structure, since we know the positions of the atoms (eq 1). The phase information is combined with the observed diffraction amplitudes for the unknown structure, and an approximate electron density map is calculated (eq 2).

Direct Methods=probability method. We know that electron density has to be non-negative. There are also probability relationships between the phases of the reflections that have related hkl indices. Using these two assumptions, Karle and Hauptmann developed a way to estimate phases for the intense reflections, and won the Nobel Prize for their achievement. An approximate electron density map is calculated from eq 2, giving the location of some of the atoms. Other atoms can be located in the electron density, and by this "bootstrap" method the structure can be completed.

Multiple Wavelength Anomalous Dispersion (MAD) uses the anomalous scattering from a heavy atom or atoms, observed at several wavelengths, to solve the phase problem algebraically. This is a powerful method for solving protein structures when molecular replacement cannot be used (no similar protein structure exists). This method avoids the problems associated with the MIR method.

Refining the trial structure = getting better phases = getting a better model for the molecule. Refinement requires the following information:

Trial model requires approximate x,y,z coordinates of at least some of the atoms, from one of the methods described above. The trial model allows calculation of approximate structure factors.

Calculated electron density map, using observed diffraction intensities with approximate phases from trial model.

Least squares calculation to match calculated (from model) and observed (from diffraction pattern) structure factors as closely as possible while restraining bond lengths and angles to typical values from previous crystallographic results

Atomic connectivity of the small molecule, or sequence of protein or nucleic acid is very helpful, but not strictly necessary.

For refining protein structures, standard geometry of α–helix or β–sheet secondary structures is used.

Sometimes the electron density of symmetry-related subunits is averaged to improve the clarity of the electron density map.

Methods of refinement:

Stacked plastic sheets and wire models (old)

Computer graphics display of electron density contours and real time manipulation of model

Computer program for least squares calculation followed by Fast Fourier Transform (FFT) computer program for calculating electron density maps

Molecular dynamics refinement on supercomputer using force field calculations restrained by the observed intensities

The *R factor* or residual error is a measure of how closely the trial structure reproduces the experimental information. A low R factor is desirable. Accurately determined small-molecule structures have R factors of 5 % or less; for protein structures the range is 15-20 %.

The end result of a structural determination by single crystal X-ray diffraction is the set of atoms in the structure with x,y,z coordinates and temperature factors. (Protein structures usually do not include hydrogen atom positions, since the resolution isn't high enough.) Usually these coordinates are deposited in the Cambridge Structural Database (CSD) or the Protein Data Bank (PDB), where they are available to investigators around the world.