X. SOME NOTES ON X-RAY CRYSTALLOGRAPHY

No technique has provided more information or more detailed information on biochemical structure than x-ray crystallography. At the moment there are 100,000 structures on file with the Protein Data Bank (PDB), roughly 85% having been solved by crystallography. However, the quality of those structures varies significantly. As with any experimental result, crystallographic data can vary in quality, and as with any human practice, so can the analysis. The goal of these notes is to provide you with some simple tools to be used in evaluating the quality of the structures obtaining from the PDB.

Crystals

Growth

The first, and sometimes most frustrating, step in x-ray crystallography is in obtaining crystals of the molecule of interest. However difficult this is for small organic compounds, it's worse for proteins, which are large, irregularly shaped, and often conformationally labile. Protein crystal growth is typically performed by vapor diffusion. A drop (on the nano- to microliter scale) of concentrated solution of the protein (> 10 mg/mL) is mixed with a solution containing a precipitant (typically a salt, organic solvent, or soluble wax) and then sealed into a chamber containing a larger volume of that precipitant solution "in the well" (Figure X.1). Since the precipitant is at a higher concentration in the well, there will be vapor diffusion between the well and the drop until the concentration of the precipitant in the drop is as high as in the well. This slow approach to a high concentration will, in ideal circumstances, promote growth of the crystal – to dimensions as large as 1 mm (not so big).

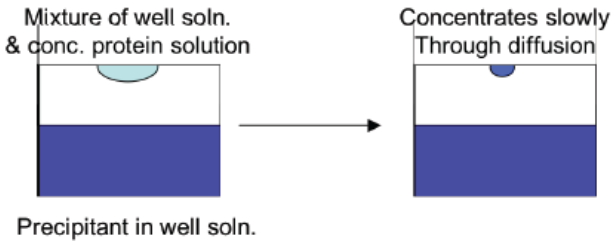

Figure X.1. Vapor diffusion method for protein crystallization As the mixture of protein and precipitant concentrate, crystals may form.

Note that the conditions required to create protein crystals (high salt or organic solvent, sometimes extremes of pH) are not necessarily similar to the physiological origins of the protein. In addition, a protein in the crystalline state is not quite the same thing as a protein in the soluble state. Nevertheless, it's worth noting that protein crystals are very different than salt crystals. Sometimes as much as 80% of the volume of the crystal is solvent, with the protein acting as a sort of ordered network through which the bulk solvent flows (see Figure X.2). Texturally, the protein crystal is more like a gel than a rock – easy to squish, and not prone to sharp cracks.

The Unit Cell

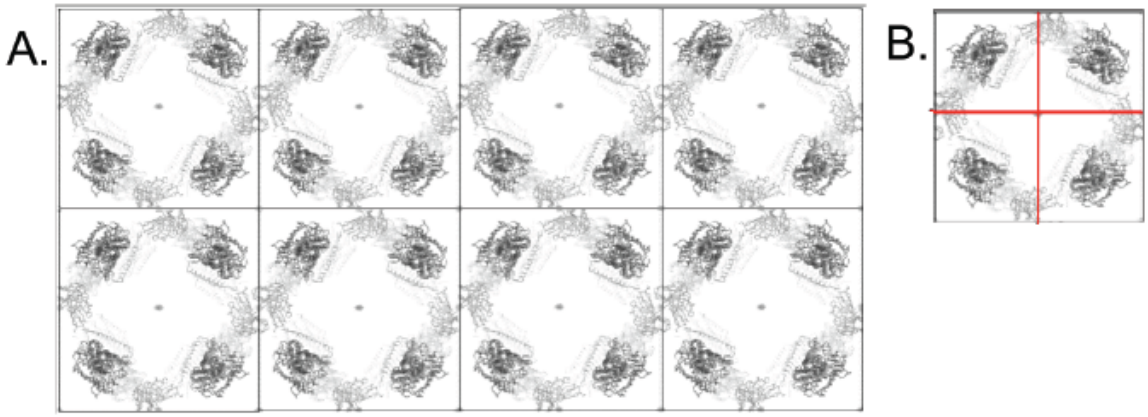

Figure X.2. (A) A cross-section of a protein crystal showing the tops of eight unit cells. Each is related to each other by simple translation along the x or y axes. (B) Each unit cell contains four identical objects related to each other by a four-fold rotation axis perpendicular to the page. The contents of one-quarter of the unit cell is, in this instance, the asymmetric unit.

Crystals are ordered assemblies of matter. That order is reflected in their composition. The least reducible unit of the crystal as the **unit cell** – that block of matter that can be used to recreate the entire crystal by translation along the x, y and z axes Figure X.2A). It is somewhat similar to a collection of shoe boxes stacked vertically and horizontally to fill a room. In an average protein crystal there are often 10^{17} individual unit cells. The unit cell must be a parallelopiped (a prism having three pairs of parallel faces, each of which is a parallelogram, whose dimensions are defined by the cell axis lengths **a, b** and **c,** and whose faces are set off by angles of α, β and γ (the angle α denotes the angle between the b and c axes, β is between a and c, and γ is between a and b.) The most symmetric parallelopiped is a cube, in which all cell axes are the same length (a = b = c) and all cell angles are 90°. Such unit cells are called "cubic." The least symmetric units cells are "triclinic" having cell edges of differing lengths and differing cell angles, none being 90°. The differing classes of unit cells are listed in Table X.1.

The goal of the crystallographic experiment is to determine the position of every atom (or every non-hydrogen atom generally) inside the unit cell. Small unit cells may contain one protein chain and have unit cell edges of 20 Å. Large unit cells may contain several dozen protein chains and have cell edges above 400 Å. Clearly, the crystallographic problem grows with the size of the unit cell, since that (in part) determines how many atoms will need to be located.

However, it isn't always necessary to build up the entire unit cell, atom by atom. Often, unit cells contain internal symmetry that allows for a simplified problem. While the crystal is built by stacking up unit cells in three dimensions, the unit cell can also be built up from simpler sections, using symmetry properties internal to the unit cell. The classes of unit cells listed in Table X.1 list the minimal symmetry elements that are available within them. Triclinic crystals are the least promising, since there is no internal symmetry at all. However, an orthorhombic crystal, with three different cell edges but all angles of 90°, has considerable internal symmetry. Think of a typical shoebox, and

you have a picture of the orthorhombic crystal. Looking at any of the faces straight on, you can rotate the unit cell by 180°, flipping it over, and get back the same shape. Those 180° turns occur along "two-fold" axes that relate two halves of the box (or unit cell) to each other. That external appearance means that the two halves of the box contain identical arrangements of atoms. Because all three faces have two-fold symmetry, the unit cell is comprised of four symmetry related sections – the **asymmetric units** (see Figure X.1B).

Since the contents of the asymmetric unit are reproduced in other parts of the unit cell by simple, known symmetry operations, it isn't necessary to solve the position of every atom in the unit cell individually. Instead, one can solve the position of every atom in the asymmetric unit, and then use symmetry to reconstruct the full unit cell. In the cases of orthorhombic and higher symmetry unit cells, this is a real advantage and can dramatically reduce the time required for structure solution.

Table X.1. Classes of units cells frequently observed in biochemical crystals.

Class	**Rules for edges and angles**	**Minimal Internal Symmetry**	**Asymmetric Units per Unit Cell**
Triclinic	$a \neq b \neq c$ $\alpha \neq \beta \neq \gamma \neq 90^\circ$	none	one
Monoclinic	$a \neq b \neq c$ $\alpha \neq \gamma \neq 90^\circ$ $\beta = 90^\circ$	1 two-fold	two
Orthorhombic	$a \neq b \neq c$ $\alpha = b = \gamma = 90^\circ$	3 two-folds	four
Tetragonal	$a = b \neq c$ $\alpha = b = \gamma = 90^\circ$	1 four-fold & maybe 2 two-folds	four or eight
Trigonal	$a = b \neq c$ $\alpha = b = 90^\circ$ $\gamma = 120^\circ$	1 three-fold & maybe 1 two-fold	three or six
Hexagonal	$a = b \neq c$ $\alpha = b = 90^\circ$ $\gamma = 120^\circ$	1 six-fold & maybe 2 two-folds	six or twelve
Cubic	$a = b = c$ $\alpha = b = \gamma = 90^\circ$	symmetry all over the place	twelve to forty eight

X-Rays and Diffraction

X-ray diffraction is a misnomer – it's really x-ray scattering. X-rays scatter in three dimensions from electrons. If you shine an x-ray beam on any random chunk of matter, you'll see a fog of scattered x-rays on a detector set up on the opposite side of the sample. On the other hand, if you shine x-rays on a crystalline sample, the order of the sample leads to patterns of constructive and destructive interference and one obtains a "diffraction pattern" of discrete spots kind of like light reflected off of a disco ball. While the mathematical reasons for the appearance of the diffraction pattern are too involved for this presentation, there is a simplified model that is predictive of the pattern. Each spot is treated as a **reflection** that emanates from x-rays that are deflected off of parallel planes of atoms within the crystal. This is the Bragg model for x-ray diffraction.

Bragg's Law – A Simple Approach to Understanding X-Ray Diffraction

Diffraction is a common laboratory phenomenon. UV/Vis spectrophotometers use diffraction gratings to select particular wavelengths of light for illumination of a sample. Diffraction gratings are closely spaced mirrored surfaces that reflect light of varying wavelengths at varying angles (think of holding a compact disk up to a light at an angle). In most x-ray diffraction experiments a single wavelength is selected (typically between about 0.7 to 2.0 Å). Thus "monochromatic" x-rays will only successfully reflect at a given angle if the "mirrors" are spaced at an appropriate distance. Based on the wavelengths of x-rays, these distances turn out to be appropriate for the dimensions of the unit cell in the crystal.

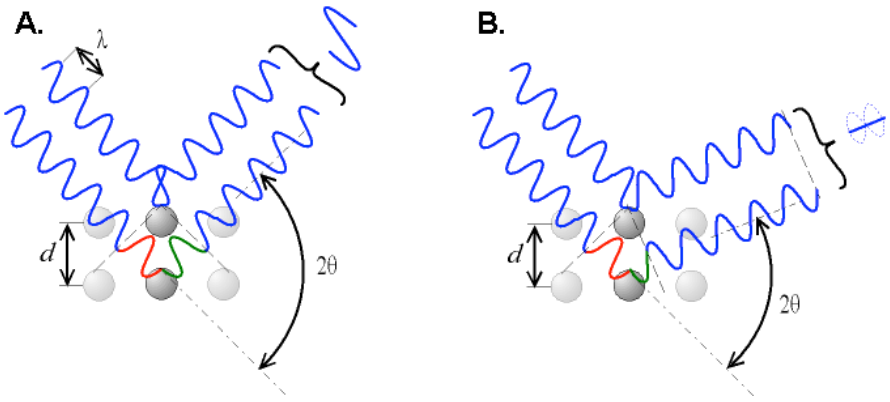

Figure X.3. (A) Constructive and (B) destructive interference from planes of atoms separated by a distance "d". (Stolen from our friends at Wikipedia.)

Imagine a simple crystal with one unit cell dimension of **d** and atoms at the corners of the unit cell. Atoms will lie along planes that are separated by that distance **d**, and you can imagine x-rays reflecting from those planes (Figure X.3). If x-rays are directed at that set of planes with an incident angle of θ, they will reflect off at an angle of θ as well. However, not every angle will lead to the observation of reflected x-rays. Because of the possibility of destructive interference between x-rays reflecting off the parallel planes in differing phases, it's important that the difference in path length between each reflected x-ray be an integral number of wavelengths. Some simple trigonometry can be used to show that the following relationship can be used to predict the successful angle of diffraction:

$h\lambda = 2d\sin\theta$

In this equation, **h** is an integer and **d** is the spacing between planes**.** This is Bragg's Law (Figure X.3).

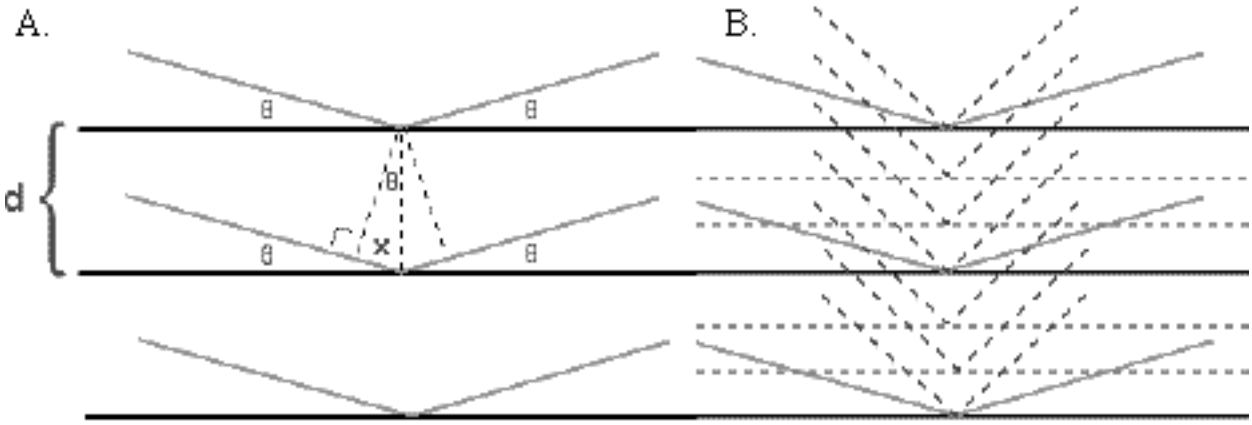

Figure X.4. (A) Bragg diffraction. X-rays diffract from parallel planes at incoming and exiting angles of θ. The path difference, 2x, must be an integral number of wavelengths. (B) Higher resolution diffraction, the dashed blue lines are diffracting from planes separated by d/3 at a sharper angle.

Another way to think of **h** is as a divisor of **d**. Bragg's law will hold for x-rays diffracting off electrons that fall on imaginary planes that divide up the unit cell into **n** slices (Figure X.4).[1] The newly written equation states that $\lambda = 2(d/h)\sin\theta$. **"d/h"** is the **resolution** of diffraction, usually measured in units of Å. By increasing **h** we divide up the unit cell into ever smaller sections and sample the electron density within the unit cell ever more finely. This is like looking at an elephant through a fence of wood slats, where there are narrow gaps between the slats. If the slats are 10 ft wide (low resolution), you'll get some rough sense of how big the elephant is. At 1 ft. wide (medium resolution), the slats are coming often enough to get a rough shape. If the slats are about 2 inches wide (high resolution), you really can pretty much see everything you need to see of the elephant. So it is with crystals. The most important factor used to describe the quality of a crystallographic model is its resolution. The lower the number (the current macromolecular record is 0.6 Å) the more detail you see. Typically, we describe resolution in qualitative terms as well as numerical terms (and these are somewhat arbitrary): **Low resolution** is data taken at 3 Å and above, **moderate resolution** goes from 2 – 3 Å, **high resolution** is between 1-2 Å and **atomic level resolution** is below 1 Å. Instead of an elephant, we're looking at the distribution of electron density in the crystal. At any point lower than 3 Å resolution its possible to make a good estimate of the shape of the molecule, but with increasing resolution comes increasing confidence.

[1] In all honesty, this a pretty sketchy way of thinking of diffraction, but it works fine as qualitative explanation. There are plenty of great books that will give a quantitative view that is much more comprehensive but more difficult to comprehend.

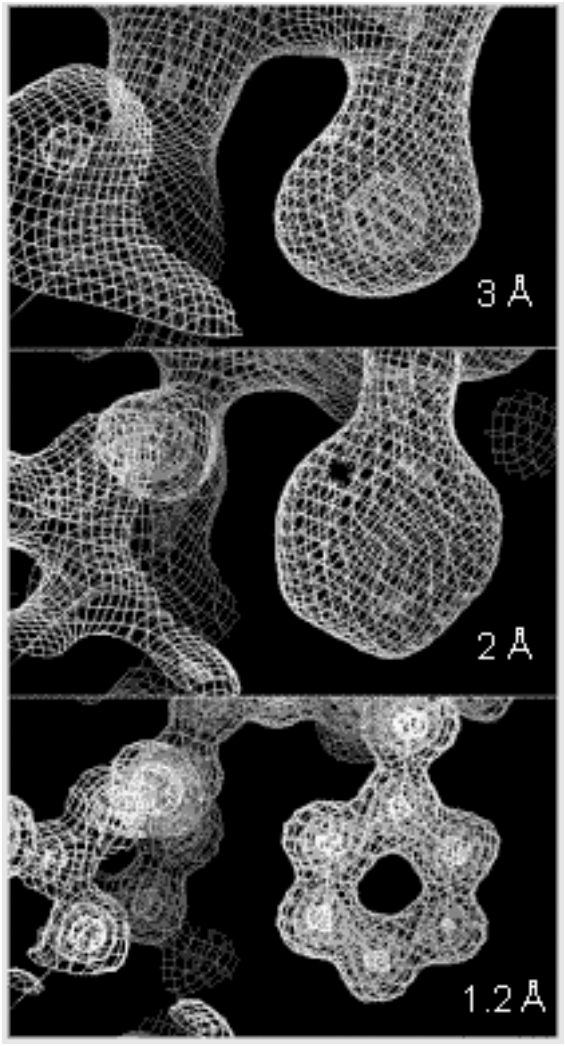

Figure X.3 Shows the electron density around the same portion of a molecule solved to three differing levels of resolution. This image was taken from Bernhard Rupp's page at http://www-structure.llnl.gov/Xray/101index.html. Note that the electron density becomes more precise and more specific at higher resolution (lower d/h spacing).

Difficulties

Ideally, all crystal structures should be solved to the highest possible resolution. The best resolution achieved for a protein structure, to date, is about 0.6 Å At that level of sharpness, one can see individual atoms as balls of electron density, and even hydrogen atoms are visible. Sadly, it is often the case that we don't get diffraction even nearly that far. The biggest problem is order with in the crystals. Since diffraction requires the ordered arrangement of 10^{18} unit cells in a crystal and ordered molecules within each unit cell, high resolution diffraction requires an extremely well ordered crystal. Protein crystals are often not well-ordered. Where small molecule crystals are often solid samples of a given compound, protein crystals are really more of a gel, where protein molecules stack loosely and are surrounded by as much as 80% solvent. At the very least surface residues have more disordered side chains than do core residues, which are usually tightly packed. But there may be

variability in how the loosely packed proteins contact each other, and, at an even larger scale, the crystal is itself a "mosaic" collection of many smaller crystals, which may be more or less ordered with respect to itself. A crystal that has sufficient order to diffract to 1 Å is a rare and beautiful thing. More commonly, one finds oneself battling to get diffraction below 2 Å, and sometimes you're grateful for 3 Å resolution. Large molecules and assemblies of molecules, like the ribosome and viruses present classic problems in obtaining high resolution data.

Measurement

As a practical matter, good data depends on good data collection. The crystal is mounted in an intense x-ray beam of a single wavelength, and data collection may last from 1-48 hours depending on the source of the x-rays, the detector, and the quality of the crystal(s). Frequently the crystal is flash frozen at liquid nitrogen temperatures (100 K) to permit extended data collection or stability in the particularly intense beams available at synchrotron beam lines. The most common detectors currently are CCD devices (like in digital cameras, but huge numbers of megapixels) and imaging plates (like in phosphorimagers). A typical experiment requires many images be taken, so that the crystal is viewed across a range of angles. The more symmetry in the crystal, the fewer degrees of rotation are required. In the best case, one only needs to rotate about 22.5°, but in the worst case 180° of rotation are required. Figure X.5 shows a typical diffraction image, collected at a synchrotron, of a moderately decent crystal.

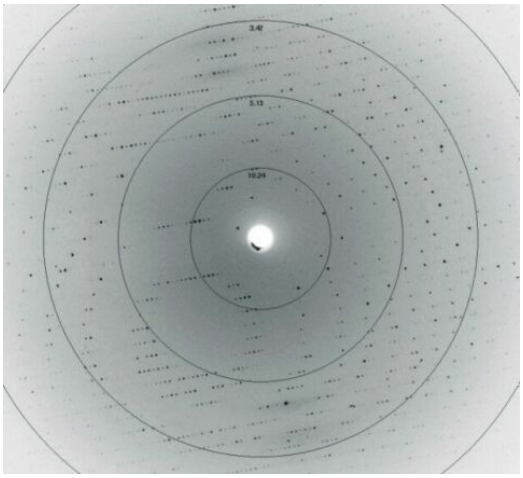

Figure X.5 Diffraction image from a protein crystal taken from http://userpage.chemie.fu-berlin.de/~psf/ifv_psfx.htm. The concentric rings note the resolution at that angle of diffraction. The inner most ring reads 10.2 Å while the outermost ring is at about 2 Å. Note that the spots are weaker at the edges, demonstrating the difficulty of obtaining good quality, high resolution data.

Diffraction and the Fourier Transform – Calculating Electron Density

Consider a 1-dimensional crystal, in which atoms are placed at an even spacing along a line – essentially at the ends of 1-D unit cells (Figure X.6). When x-rays diffract from this crystal, they will do so at angle prescribed by Bragg's Law. The diffraction pattern registered at a detector will have spots registered at increasing distance from the origin – that is, the point on the detector that would

be struck by the original, direct x-ray beam. Each reflection is numbered ("indexed") according to its distance from the origin using the variable "h".

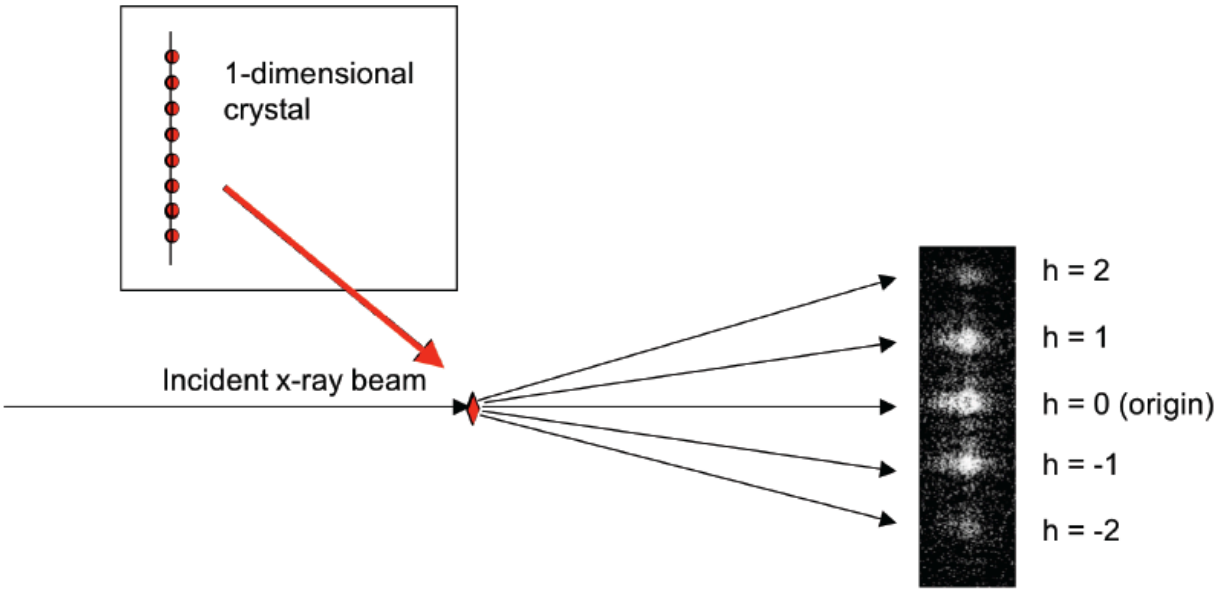

Figure X.6. Diffraction from a 1-D crystal yields a diffraction pattern in 1-D. The reflections are indexed by the value of h (related to the integer values in Bragg's Law).

The structure of the crystal and the diffraction pattern are related to each other by the Fourier transform. Each can be used to calculate the other. The diffraction pattern is constructed of the scattered x-rays. Each can be described using a **structure factor**, F_h, which has properties of amplitude ($|F_h|$), phase (α) and frequency (h) – like a wave. The structure factor for each reflection can be calculated directly from the distribution of electron density, $\rho(x)$, in the unit cell as shown below. Note that electron density in a crystal is a continuous function, so we integrate.[2]

$$F_h = \int_0^1 \rho(x) \cdot \exp[2\pi i h x] \cdot dx$$

Similarly, the electron density can be calculated from the structure factors as follows. Note that the diffraction pattern is a non-continuous function with indexed components, so we sum.

$$\rho(x) = \sum_h F_h \cdot \exp[-2\pi i h x]$$

[2] A curious person might wonder about the "exp[2πihx]" term. This is a computationally easier way of expressing the cosine function. Note that $e^{(i\phi)} = \cos(\phi) + i\sin(\phi)$. In typical Fourier transform calculations, only the real component of this expression is needed

X-ray crystallography proceeds by measuring the structure factors (the diffraction pattern) and using it to calculate electron density. If we know all the parameters of the diffracted x-rays, we can recalculate the positions of the atoms in a crystal, such as the slightly more complex 1-D crystal in Figure X.7. Note that the more reflections that you can measure, the more accurate the calculation of the positions of the atoms (though they can never be known with perfect precision).

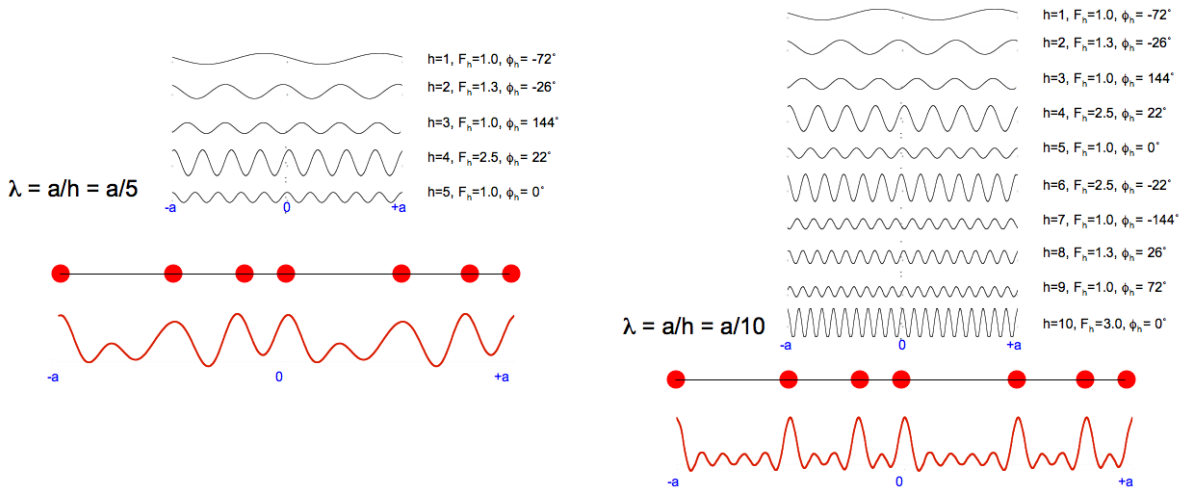

Figure X.7. Illustration of the Fourier transform. On the left we use the first five reflections (lowest resolution) to calculate the distribution of electron density in a 1-D crystal (two unit cells are shown, each has 3 atoms). On the right, we use 10 terms to obtain a higher resolution picture of the crystal. Note that the small ripples are not real and reflect an imperfection in the calculation that is inherent to the Fourier transform.

Thus, to determine the structure of a molecule, you need two pieces of information from each indexed "reflection" that is measured during the diffraction pattern: the amplitude and the phase. The amplitude is related to the intensity of the x-ray striking the detector (the darker the spot, the higher the amplitude), but the phase cannot be measured. (The phase corresponds to the point in the wave's cycle at which the x-ray strikes the detector. Does it strike at a peak, a valley or somewhere in between?) Unfortunately, there is no way to measure that directly during data collection. As a result, some fairly tricky and subtle methods have been introduced to allow crystallographers a guess (sometimes a good guess) at the phases.

Obtaining the Phases

Phase determination is the greatest hurdle in most crystallographic experiments. The methods used require some complex geometric arguments to explain, so we'll leave those alone Instead, it should be sufficient to note that all of the methods listed in Table X.2 provide experimental phase information. Sometimes the initial determinations are quite good, such as when the protein has previously been crystallized under very similar conditions, and sometimes they are quite poor. The challenge is to overcome the latter situation and to obtain a reasonable structure despite limitations of the starting phases.

Table X.2 Some ways in which initial phases can be estimated.

Method	**Description**
Multiple Isomorphous Replacement (**MIR**)	Requires a derivative of the native protein with a heavy atom (like mercury or iodine) which scatters x-rays more strongly than other atoms in the molecule. By finding the positions of those atoms, one can make some good approximations of the phases for all reflections.
Multiple Anomalous Dispersion (**MAD**)	Requires substitution with an atom that scatters x-rays asymmetrically at a particular wavelength. Selenium, incorporated as selenomethionine, is very popular these days. The method requires tunable x-rays, typically obtained at synchrotrons.
Molecular Replacement (**MR**)	Requires that you have a model of a similar protein that can act as a first guess for the structure of your protein. You can make an initial estimate of the phases based on phases predicted for the model.
Molecular Difference	This is the easiest way to guess at phases. If the protein has already been crystallized under similar conditions and it crystallizes similarly (same unit cell, etc.) then chances are the phases for the new structure are similar to those of the old structure.

Structure Determination

Data Quality

Before attempting to solve the structure of a macromolecule by crystallography, one should have confidence in the data being used. Table X.3 describes several measures that are routinely used to report on the quality of data being used in an experiment.

Table X.3. Descriptors of data quality employed in crystallography.

Parameter	Description
Resolution	This typically describes the extent of the data from low to high resolution that is being used. The higher the resolution, the better, but that is often outside the control of the experimenter. *Usually a high resolution range will be given in parentheses. These are the weakest data and must be especially defended by the experimenter.*
$I/\sigma(I)$	Intensity vs. the standard deviation in intensity, or more simply, signal to noise. The higher this value the more reliable the experimental data. *Typically the signal to noise ratio is given for the high resolution shell. It should normally be about 2.0 or better. Lower values indicate that the experimenter is using very weak data to make their resolution look better than it is.*
Completeness	Within any resolution range, there are a defined number of data points available. Ideally, 100% of them will be measured, but anything over 90% is acceptable. *In the high range, it is common to see lower completeness. Recognize that incomplete data collection has a negative impact on structure solution, but it is better to include all the data you have, so incomplete high resolution data is acceptable.*
Redundancy	How many times was each reflection (data point) measured? As in any experiment, the more measurements the better. A redundancy of 3 is typically considered OK. Higher is better. Sometimes the experimenter will report total and unique reflections. The ratio of those two gives the redundancy.
R_{sym}	Or sometimes, R_{merge}. This measures the residual (R) between multiple measurements of the same reflection. It is the fractional or percent disagreement on average throughout the data set. It should be 10% or less, *though in the high resolution shell it may be up to 50%.*

The Model

The desired outcome of any crystallographic experiment is a model for the molecule of interest. The most obvious attribute of that model is the positions of the atoms within the molecules – the x, y and z coordinates. Crystallography also requires assignment of a fourth parameter for each atom – the so-called "B factor". Ideally, this is a measure of the thermal motion experienced by an atom in

the molecule, hence its alternate names, the "temperature factor" or "thermal parameter."[3] In fact, it describes all of the different things that contribute to varying position for an atom – rotation around bonds, disorder from unit cell to unit cell, and honest thermal motion. Typically, atoms on the surface experience a greater degree of positional disorder and will have larger B factors. Most models include B-factors that range from about 10-200. Assuming that the atom moves within a spherical region about a central point (not always a good assumption), then the radius of that motion can be calculated as:

$$\text{radius of motion} = \sqrt{B/8\pi^2}$$

For an atom with a B factor of 25 (pretty typical), the radius of motion is about 0.6 Å. This is obviously an important issue to consider when thinking about where atoms in a molecule reside. Big B-factors mean that they move around a lot within the crystal.

Model Building

Once a good data set is in place and phases are available, the electron density map is calculated. The protein sequence for a crystallized molecule is typically available, so the bonded structure of the protein is known – only the conformation is absent. Model building proceeds by placing residues, in correct sequence, into the electron density as appears appropriate. At high resolution, this can be straightforward and is often automated. At low resolution, it can be quite difficult to distinguish which residue goes into which ill-formed blob of electron density. Methionines and large aromatic residues are usually somewhat unique looking and typically provide toe-holds for model building. It is important to note that the software typically employed for model building restricts the experiment in only allowing residues with appropriate bond lengths and angles. The dihedrals, which define conformation, are allowed to vary – though in some instances there are restrictions on those as well to avoid steric conflicts that might arise during the construction of the model.

Refinement

The first attempt to build a model often yields an incomplete structure with poor fit to the electron density. Automated methods exist to modify the conformation of the model so that it better fits the electron density. This is the process known as **refinement**. The quality of the model is decided on two counts: (1) its agreement with the intensity data taken from the x-ray experiment and (2) its stereochemical reasonableness – that is, does it look like a real molecule, with normal bond lengths, angles, etc.

In judging the agreement with the data, the R-factor (often abbreviated $\mathbf{R}_{\mathbf{cryst}}$ or $\mathbf{R}_{\mathbf{work}}$) is measured. This value compares the measured amplitudes of the diffracted x-rays (given the abbreviation $\mathbf{F}_{\mathbf{obs}}$) and compares them to amplitudes that are predicted from the current model ($\mathbf{F}_{\mathbf{calc}}$; see equation X.1).

[3] Another name is the Debye-Waller factor, which is more commonly used in other spectroscopic techniques.

$$R_{cryst} = \frac{\sum \left| |F_{obs}| - |F_{calc}| \right|}{\sum |F_{obs}|} \qquad \text{(Eq. X.1)}$$

The better the agreement between F_{obs} and F_{calc}, the lower the R-factor. It is rare, even with modern refinement techniques, to see R-factors below 0.15 or 15%. Oftentimes you will see a separate measure of quality reported, called $\mathbf{R_{free}}$. R_{free} is calculated in an identical fashion to R_{cryst} except that a small subset of the total data (5-10%) is set aside and not used in refinement. If refinement is doing a good job, the R_{free} should sink along with R_{cryst}. However, it is possible to play games to try to optimize R_{cryst} without really doing a better job of modeling the protein. R_{free} tends to reveal that sort of tawdry behavior and makes a negative example of your cheating ways.[4]

The stereochemical quality of the model is easier to evaluate. Are the backbone dihedrals found in "allowed" regions of Ramachandran space? Are the deviations between the bond distances in your model and "ideal" bond distances small (< 0.02 Å) and so are the bond angles (< 2°). If so, then your protein looks like what we expect a protein to look like, and all is good. It should be recognized, however, that there is almost always a tension between fit of the crystallographic data and stereochemical parameters. At lower resolution, one frequently accepts the dictates of stereochemistry, because the electron density just can't be used to any precision. On the other hand, at atomic level resolution, it is possible to allow stereochemical concerns go. The data is of such good quality that it should be able to give a good looking model with no coaching from some table of what molecules "should" look like.

A Cautionary Note

So you have a model prepared from crystallographic data. The data was high resolution, the R-factor is low (as is R_{free}) and the model agrees with stereochemical conventions. It still may be bogus. Crystallography depends upon taking a soluble (or membrane-bound) protein and placing it into a dense, arrayed state using a variety of precipitants, including salts, organic solvents, and polymers. The pH may be far from physiologically relevant, and low temperature data collection may freeze out a portion of the structure that is normally quite flexible. Crystal contacts may block active sites and distort the surface of the molecule. How do you know if what you're seeing is real? Simple – the model is a guide for experiments. It is not an end unto itself, but rather a tool to be used in describing and predicting the behavior of the protein. Frequently, crystallized protein retain activity even in the crystal and can be seen to adopt structures wholly consistent with biochemical experiments. In fact, that is the rule rather than the exception. Nevertheless, always use a model with some skepticism. These structures are the product of human endeavor and are subject to all the errors that come with it. But even so, they provide some of the most compelling evidence for the means by which proteins achieve their unusually fine-tuned properties.

[4] A great description of these games and their failings has been published by Alwyn Jones and Gerard Kleywegt (1995) *Structure* **3,** 535.