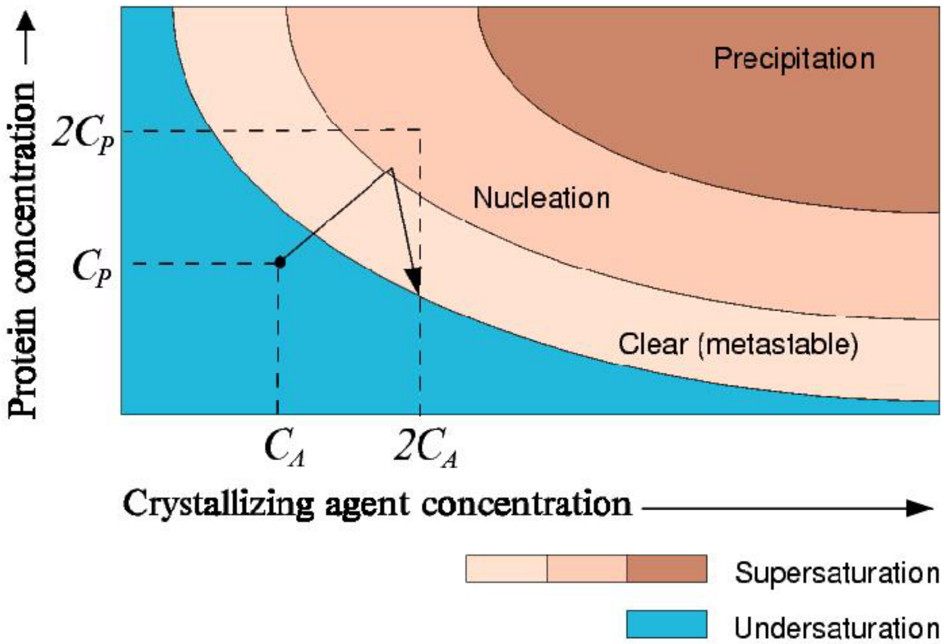
BC530 Class notes on X-ray Crystallography

Ethan A Merritt
web material: `http://skuld.bmsc.washington.edu/~merritt/bc530/`

September 28, 2017

Growing Crystals

It should be self-evident that in order to do crystallography one needs a crystal. So the first problem a protein crystallographer faces is to persuade a protein to become crystalline. Sometimes this is an insoluble problem (*pun intended*). Often the eventual success depends on cleverness in the molecular biology lab, where you can trim off recalcitrant bits of the peptide chain or introduce selective point mutations to enhance solubility, stability, or the propensity to form a crystal.

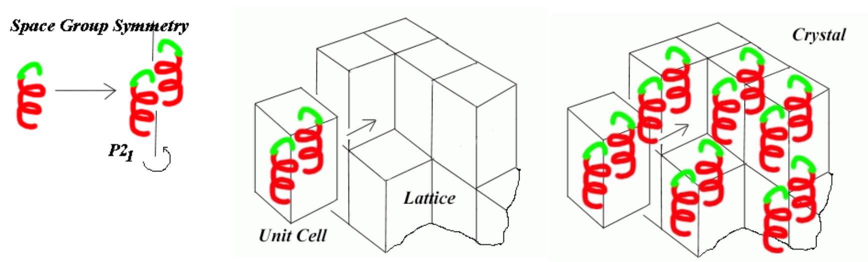


In principle the process of growing a crystal from a solution relies on a solubility phase diagram such as the one shown here. The details of the phase diagram, including the shape and position of the colored areas above, will be different for every protein. If the concentration of the protein is too dilute, it will stay in solution and not form a crystal (lower left). If the concentration of the protein is too great, it will come out of solution, possibly as a crystal but more likely as a poorly-ordered precipitate (upper right). Somewhere in between these two extremes is a region, perhaps a very small region, of the phase diagram in which there is an equilibrium between protein molecules in solution and protein molecules in a crystal lattice. The precise size, shape, and location of this region is strongly influenced by the presence of other molecular species in the crystallization solution. Usually the search for crystallization conditions consists of a search for suitable molecules ('crystallizing agents' in the diagram) whose concentration can be varied in order to shepherd the protein along a path through the phase diagram that takes successively through regions of solution->nucleation->crystal growth (the growth stage is labeled 'metastable' in the diagram). The path shown is typical for a vapor equilibration protocol, 'hanging drop' or 'sitting drop'. The initial condition in the drop is in the blue region; as water leaves the drop the concentration of all components increases until nucleation begins to occur; in the final segment of the path as shown, concentration of the crystallizing agent continues to increase as water vapor leaves, but the protein concentration in solution goes down as the individual protein molecules leave solution to join the growing crystal.

Crystallographic and Non-crystallographic Symmetry

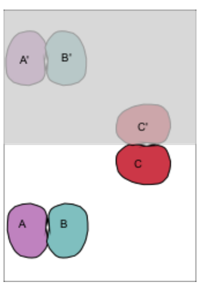
A crystal is regular 2-dimensional or 3-dimensional array of identical *Unit Cells*. This array is called the *Crystal Lattice*. The crystal lattice may, or may not, exhibit symmetry elements such as 2-fold, 3-fold, 4-fold, or 6-fold axes of rotation. This is called the *Crystallographic Symmetry*. There are exactly 230 possible combinations of rotational and translational symmetry that can exist in a 3-dimensional lattice. These are called *Space Groups*. Sometimes the crystallographic symmetry is instead called *Space Group Symmetry*. The simplest case is when no symmetry is present in the lattice beyond the pure translation that distinguishes one unit cell from its nearest neighbors. This is space group $P1$.

All other space groups are more complicated. The figure below shows another simple, and very common, space group called $P2_1$.It has one additional symmetry operation, consisting of a rotation by 180ř and a simultaneous translation vertically. (This is called a 2-fold screw axis, but you don't need to worry about the naming conventions). As shown in the figure, if we start with a single unique copy of the molecule this symmetry operation yields a second, identical, copy elsewhere in the unit cell. In solving the crystal structure, we only need to model the one unique copy. To be more precise: in space group $P2_1$each unit cell is subdivided into two identical volumes related by the 2-fold screw axis symmetry operation. Since they are identical, we only need to determine the contents of one of these volumes. This is called the *Asymmetric Unit* of the unit cell. In more complicated space groups there are more space group symmetry operations, and the asymmetric unit is a smaller fraction of the unit cell.



The asymmetric unit itself may contain multiple copies of the crystallized molecule(s). This is called *Non-crystallographic Symmetry*. In this case the crystallographer must choose whether to model these multiple copies independently, or create only a single model and replicate it however many times is required. Another way of looking at this is to say that the crystallographer must decide whether these multiple copies are identical to each other or whether they are significantly different. If there are significant differences, then often the next question is whether the differences are biologically relevant or merely an artifact of being packed into slightly different environments in the unit cell. Generally if there is only low resolution X-ray diffraction data then it is safest to treat the copies as being identical, since this is the simplest model and requires the least amount of data to refine. If high resolution data is available then it is usually possible to treat the copies independently from the start, and then compare the results afterward to ask if there truly are any significant differences.

Figure 1: **Mixture of crystallographic and non-crystallographic symmetry**. A and B are related by noncrystallographic symmetry; C and C' are related by crystallographic symmetry.



A final complication is that proteins often come in the form of dimers, trimers, or higher-order molecular assemblies. Thus the biologically active form of the protein may already exhibit internal symmetry. For example, a dimer made up of two copies of the same subunit usually, although not always, contains an internal 2-fold axis. When this dimer is crystallized, the two component subunits may or may not end up in the same asymmetric unit. If they do, then the two subunits are related by non-crystallographic symmetry; as just described, the crystallographer can choose to model them independently and compare them afterward. But it can also happen that this dimer will crystallize in a space group containing a 2-fold axis as part of the space group symmetry. If the dimer positions itself in the unit cell such that the internal 2-fold axis of the dimer is the same as the 2-fold axis of the crystal, then the asymmetric unit will by definition only contain one half of the dimer (one subunit). Figure 1 contains an example of both cases. The contents of the top half of the cell (grey background) are related to contents of the bottom half by a crystallographic 2-fold axis. Dimer AB is made up of two monomers which are crystallographically independent. The two halves are not constrained to be identical (e.g. A might have a ligand bound while B does not) Dimer CC' is made up of two crystallographically identical halves. They are necessarily identical to each other.

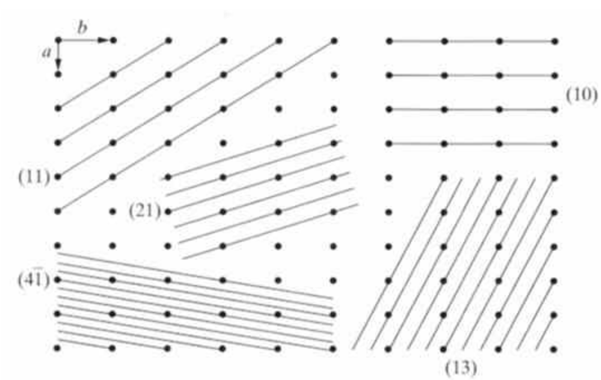
Bragg's Law and the resolution of the experiment

When a crystal is illuminated by a beam of X-rays, the X-rays interact with every atom in the crystal (actually with the electrons that belong to that atom). Since the nature of a crystal is that it is a regular repeating array of objects, whatever interaction the X-rays have with a particular atom is repeated identically for each copy of that atom. Because of this regular repeat, incoming X-ray that are diffracted by the atoms in one unit cell of the crystal interfere with X-rays that are diffracted by the equivalent atoms in the next unit cell over, and the one after that, and so on. This interference can either be constructive or destructive. We see a net outgoing ("reflected") X-ray beam only at a well-defined set of crystal orientations and illumination angles that result in constructive interference. This set of conditions is described by Bragg's Law.

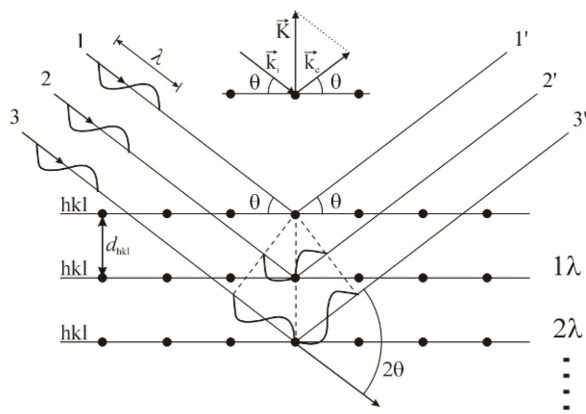
$$\text{Bragg's Law} \qquad n\lambda = 2dsin(\theta)$$

Here λ is the wavelength of the X-rays, d is the distance between the planes of atoms defined by a particular set of Bragg indices hkl, and θ is the incidence angle of the X-ray beam striking the planes. The spacing d is called the **resolution** of that reflection. Reflections with larger values of h,k, and l define planes with closer spacing (smaller d).

Beware: relections with small d are called "high resolution"; reflections with large d are called "low resolution". This is confusing.



(a) Different sets of Bragg planes defined by the crystal lattice. Each plane can be described by 2 indices (hk) for a 2-dimensional crystal as in the figure, 3 indices (hkl) for a 3-dimenensional crystal.

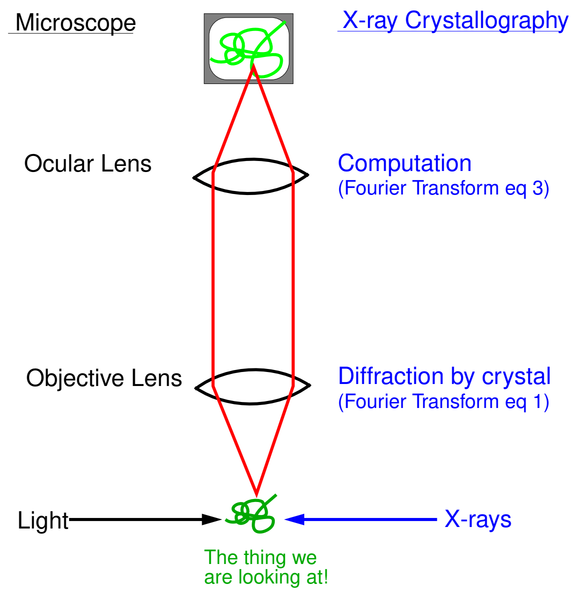


(b) Bragg's Law for a particular set of Bragg planes

The resolution of an X-ray data set is determined by the highest resolution reflections in contains. Ideally the data set is **complete**, which means that all possible reflections with that resolution have been measured. The total number of reflections in a complete data set is proportional to $\frac{1}{d^3}$. This means that high resolution data sets contain a lot more information. Incomplete data sets are sometimes described as having an **effective resolution** calculated based on the number of reflection measured rather than on the resolution of the highest resolution reflection measured.

The intensity of diffracted X-rays drops at higher resolution both because the intrinsic scattering power of the atoms decreases at larger values of θ and because the internal ordering of the crystal is not perfect. Those rare protein crystal with near-perfect internal ordering may allow measurement of data to $1\AA$ resolution or better. More typically the diffracted intensities fall off to noise level somewhere in the range $2\AA$ resolution to $10\AA$ resolution. $2-3\AA$ resolution is "good".

X-ray crystallography is analogous to light microscopy



Fourier transform as "objective lens"

The overall scattering of light from an object can be described by a Fourier summation arising from the piecewise contribution of scattering from all the individual "bits" of the object. In the case of X-ray diffraction, these bits are electrons. If we knew the precise location of every electron, we could describe the overall scattering as a grand sum over all of them. For the purposes of X-ray crystallography we will make two modifications to this description.

1. Because it is a crystal, instead of a single object there are many, evenly spaced copies of the object. The overall scattering becomes discretized into individual reflections (Bragg's Law). The regular 3-dimensional array of Bragg reflections is by convention indexed by the subscripts h, k and l.

2. Although nature 'knows' where each electron is, in practice we will approximate this by subdividing the unit cell into little boxes, and describing how many electrons lie in each box. These boxes are by convention indexed by x, y and z.

Disregarding a few constants and the imperfection of our crystal lattice, we can write our Fourier transform as

$$F_{hkl} = \sum_x \sum_y \sum_z \rho(x,y,z)\, e^{2\pi i(hx+ky+lz)} \tag{1}$$

Fourier transform as "ocular lens"

In X-ray crystallography, a numerical computation substitutes for the function of the ocular lens in an optical microscope. This Fourier transformation creates an image of the unit cell contents (an electron density map) starting from our observed data (the thousands upon thousands of Bragg reflections). The basic formula used in this procedure is

$$\rho(x,y,z) = \frac{1}{V} \sum_h \sum_k \sum_l F_{hkl}\, e^{-2\pi i(hx+ky+lz)} \tag{2}$$

$\rho(x,y,z)$ is again the electron density inside the little box of the unit cell with coordinates x, y, z. All the little boxes taken together constitute a model for how the electron density in the whole unit cell is distributed.

Smears instead of spheres - Since the electrons are associated with individual atoms, this means we should ideally find 6 electrons worth of density at the location of each carbon atom in the cell, 8 electrons worth of density at the location of each oxygen atom, and so on. Except in the case of very high-resolution data, however, each little box generally contains contributions from multiple nearby atoms. So instead of seeing a nice little sphere for each atom you see a smear.

The Phase Problem

The central problem in crystallography is the fact that the term F_{hkl} in Eq (1) and Eq (2) is a complex number. It has both an amplitude $|F_{hkl}|$ and a phase α_{hkl}. So we could rewrite the Fourier transform in Eq. (2) as

$$\rho(x,y,z) = \frac{1}{V}\sum_h\sum_k\sum_l |F_{hkl}|\, e^{i\alpha_{hkl}}\, e^{-2\pi i(hx+ky+lz)} \tag{3}$$

The amplitude is easy to get; it is the square root of the X-ray intensity we measured for the spot corresponding to that Bragg reflection hitting the X-ray detector. The sticking point is that we have no measured value for the phase α_{hkl}. Crystallographers have a wide array of techniques to deal with the unfortunate fact that the phases are not measurable. None of these are foolproof, and all of them are complicated, so we only summarize briefly the most common ones.

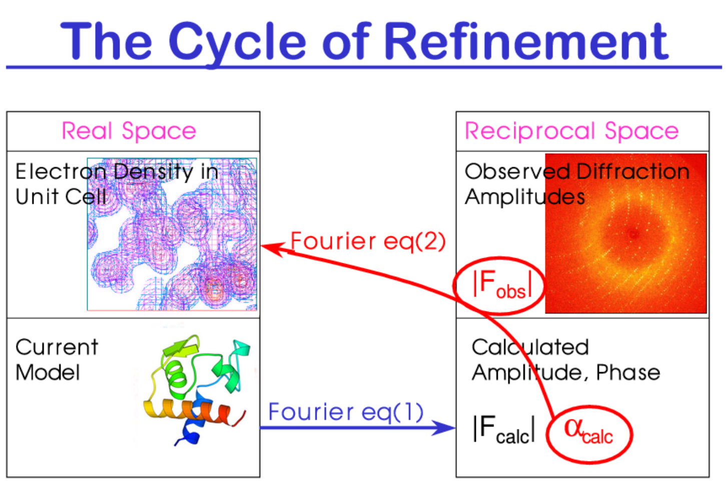
Molecular Replacement (MR)

The simplest technique to understand is called molecular replacement. It applies only to the case where we already have a model for the protein structure, usually based on the structure of a homologous protein that was solved previously. Fortunately, as more and more structures are solved and deposited in the PDB (Protein Data Bank) it becomes more likely that a suitable starting point for molecular replacement can be found. The basic idea of MR is that we can evaluate trial placements of the supposedly homologous structure into the current crystal unit cell, each time calculating both the amplitudes and phases that would hypothetically be measured from the hypothetical crystal. If both the structural homology and the selected orientation are sufficiently good, the calculated phases from the hypothetical crystal may be good enough to use as approximate phases for the actual crystal.

In principle the starting model could come from something other than a previous crystal structure, e.g. an NMR determination of structure in solution or an *ab initio* model of protein folding. In practice this almost never works. In fact, successful use of a calculated model to do molecular replacement is a good benchmark for progress in the state of the art in fold prediction.

The important point is that once we have an initial approximation for the phases, we can use Eq (3) to generate a first-pass map of electron density throughout the unit cell. Of course if the initial approximation of the phases is bad, then we get a bad map from applying Eq (3) even if our measured data set is quite good. We just have to hope that the map is good enough to proceed with the task of building and refining a model of the unit cell contents.

As the model of the unit cell contents improves, we can use it to generate better approximate phases α_{calc}. These in turn give a better map, which helps us to improve the model, which gives better phases, and so it goes. A schematic of this iterative cycle is shown below.



Direct Methods

"Direct Methods" refer to the inference of phase estimates directly from the measured data, without resort to a prior structural model. These methods depend on probabilistic distributions of phase correlation between small subsets of Bragg reflections. When the unit cell contains a small number of atoms, these joint probability distributions are sharp enough to allow inference of the phase of a newly-considered reflection by examining the previously-assigned phases for other reflections. In this way a small number of known, or assumed, starting phases can be bootstrapped to assign phases to the rest of the data. In practice iterative phase inference is carried out many times, each time starting from a set of randomly assigned phases for a small subset of the data. An external criterion, e.g. atomicity of the corresponding electron density, is used to decide when one or more of the trials has succeeded.

The simplest example of such a joint phase distribution is Sayre's Law, also known as the "triplet" relationship. It states that if there are 3 strong Bragg reflections F_{-H}, F_K, and F_{H-K} then the sum of their phases is approximately zero.

$$\varphi_{-H} + \varphi_K + \varphi_{H-K} \approx 0$$

A key extension of this relationship is the "tangent formula" introduced by Karle & Hauptman (1956). In this equation the E_K are normalized structure factors, basically a rescaled version of the observed amplitudes F_K.

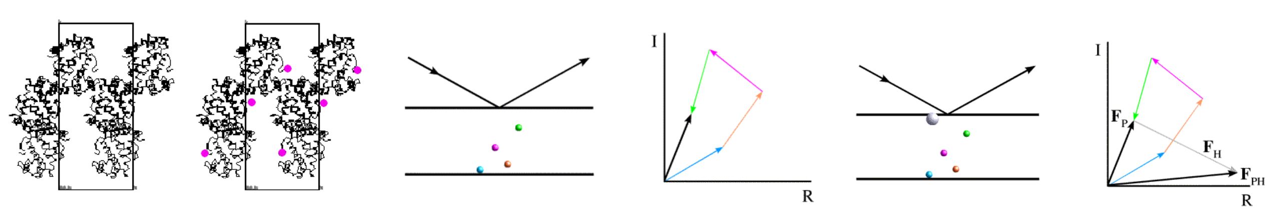
$$tan(\varphi_H) \simeq \frac{\sum_K E_K E_{H-K} sin(\varphi_K + \varphi_{H-K})}{\sum_K E_K E_{H-K} cos(\varphi_K + \varphi_{H-K})}$$

Direct methods have become the mainstay of small molecule crystallography. Sadly for crystallographers working on larger structures, the underlying probability distributions become flatter as the number of atoms in the asymmetric unit increases. Not only do the inferred phases become less accurate, the number of trial starting phase assignments that would have to be considered becomes computationally intractable. So direct methods are not generally useful for determining macromolecular structures from a single data set. Nevertheless, they can be crucial in overcoming an initial hurdle presented by the techniques summarized below (SIR, MIR, SAD, that of locating a small number of strongly scattering atoms mixed in with the hugely larger number of weakly scattering H/C/N/O atoms.

Isomorphous Replacement (SIR or MIR)

This method depends on measuring matched data from the original *native* crystal, and a single (SIR) or multiple (MIR) *derivative* crystals. Each derivative crystal differs from the native crystal by the presence of additional metal atoms, usually introduced by soaking the native crystal in buffer containing a suitable metal salt, e.g. $PtCl_4$. Unfortunately, soaking often destroys or distorts the original crystal. Even if the crystal survives the lattice may become non-isomorphous, making it unusable for phasing.

Because each metal atom is very electron-dense it contributes strongly to the Bragg diffraction, changing the relative amplitudes of the various Bragg reflections. The basic idea is that although we do not know the phases for either the native or derivative Bragg reflections, we can use the difference in measured amplitudes to figure out where the introduced metal atoms are. This is often done by direct methods applied to estimated amplitudes F_H corresponding to the scattering contribution of the introduced heavy atom[s] alone. This in turn allows calculation of approximate starting phases from which we can start the iterative improvement process. *[Note: The H in F_H here and in the figure below stands for "heavy atoms", in contrast to F_P where P stands for "protein atoms". This nomenclature is conventional, but it is confusing because in most other contexts F_H is used instead as a short form of F_{hkl}]*

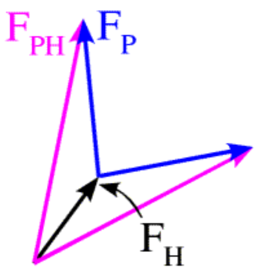


(a) Isomorpous replacement depends on adding metal atoms (usually by soaking) to some consistent set of positions in the original unit cell, without perturbing the original contents.

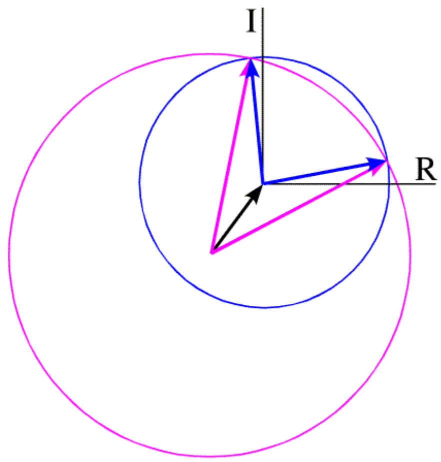
(b) The net X-ray scattering present in a Bragg reflection is the sum of the scattering contributed by each atom in the crystal's unit cell. The total amplitude $|F_{hkl}|$ and phase φ_{hkl} can be represented as a vector F_P lying in the Real/Imaginary plane. This vector is the sum of constituent vectors representing the scattering by each atom in the protein.

(c) If we add in a metal atom without perturbing the original atomic positions, there is a new net scattering vector F_{PH} that is the sum of the original protein scattering vector F_P and the added metal scattering vector F_H. Electron-rich metal atoms contribute very strongly to the total scattering; i.e. F_H can be large.

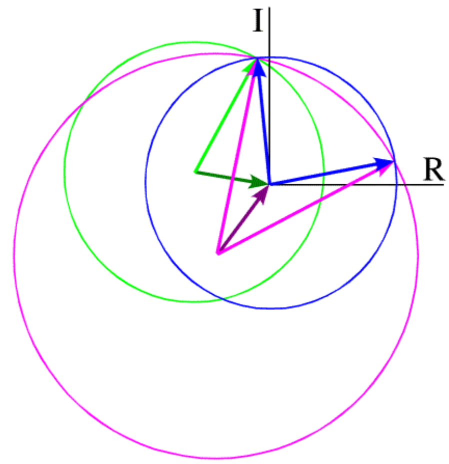
The figures below show geometrically how an estimate for the missing quantity φ_P , can be generated from the original amplitude measurement F_P from the native crystal, the amplitude measurements F_{PH1}, F_{PH2}, ... from one or more derivative crystals, and an initial model we construct for the substructure consisting of only the added metal atoms.



(d) Two possible ways that the measured amplitude F_P and the estimated amplitude F_H can sum to the amplitude F_{PH} measured from a derivative crystal.



(e) The SIR case: one native measurement (blue circle with radius F_P) and one derivative measurement (magenta circle with radius F_{PH}). This representation was introduced by David Harker, and is known as the "Harker construction". The small vector in center represents the correct amplitude F_H and phase for the metal atom[s] alone. The two possible values for the phase of the protein atoms correspond to the two intersections of the circles.

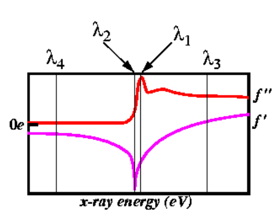


(f) The MIR case: additional data measured from a different derivative crystal (green circle with radius F_{PH2}) resolves the ambiguity in phase from SIR. Only one of the original intersections is also an intersection for the new data.

So we have reduced the problem of finding phases for the native protein with the simpler problem of finding phases for the N substructures consisting of only the metal atoms from each of N derivative crystals. Since the number of metal atoms is usually small, this task is no harder than solving the structure of a small molecule.

Anomalous Scattering (SAD or MAD)

This method is conceptually similar to SIR/MIR, with the major difference that the "native" and "derivative" data can be measured from the same crystal. Instead of modifying the Bragg diffraction by soaking in new metal atoms, we experimentally manipulate the strength of scattering by existing metal atoms. Due to a phenomenon called *anomalous scattering*, the scattering power of the metal changes as a function of the X-ray wavelength. The change is particularly dramatic at specific wavelengths that are characteristic of the particular metal (see Figure below). Thus changing the wavelength of the X-rays used in the experiment increases or decreases the contribution of the metals to the overall scattering. Since not all proteins contain metals, the most common procedure of SAD or MAD phasing involves substituting selenomethionine for the usual sulfur-containing methionine. The scattering power of Se varies sharply near an X-ray wavelength of 0.97Å, which is very convenient for data collection. The use of anomalous scattering has several advantages over isomorphous replacement. (1) The isomorphism is perfect (all data sets come from the same crystal). (2) In the presence of anomalous scattering, $F_{hkl} \neq F_{-h-k-l}$. This means that there are effectively twice as many observations in each data set that contribute phase information. This is presented in more detail on my web site `http://www.bmsc.washington.edu/scatter`.



Refinement and the R factor

How do we know that our model is improving as we work on it? Eq (1) gives predicted values of $|F_{hkl}|$ in addition to the phases. We can compare these predicted values, F_{calc}, to the measured values, F_{obs}. The usual way of making this comparison is to calculate the crystallographic residual R

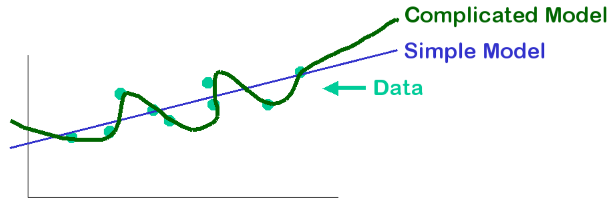
$$R = \frac{\sum ||F_{obs}| - |F_{calc}||}{\sum |F_{obs}|} \tag{4}$$

The better our model agrees with reality, the better $|F_{calc}|$ agrees with $|F_{obs}|$, and hence the smaller the residual R becomes. An approximate rule of thumb says that for a good, well-refined model, R reaches a value about 1/10 of the resolution of the data. So a good 2.0 Angstrom structure may be expected to have $R \leq 0.20$, while a good 1.5 Angstrom structure may be expected to have $R \leq 0.15$.

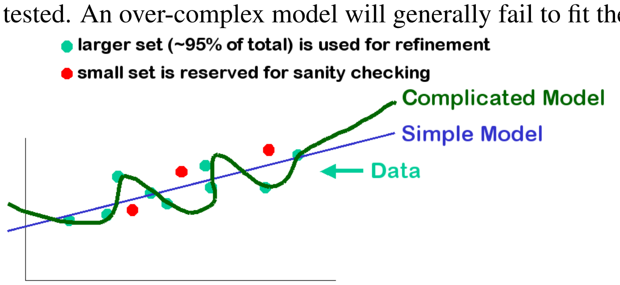
Refinement and R_{free}

A high value of R is always bad, but a low value of R does not guarantee that a structure model is good. R can be artificially lowered through the addition of unjustified parameters to the model. In any refinement (not just crystallography!) the total number of parameters being refined should remain substantially less than the number of observations.[1]

How not to fit a model: The green line passes through all the data points, but is an unjustifiably complex model. The blue line is a very simple linear model that is still a good fit to the data, though the fit is not perfect.



How to do it right: "Cross validation" (also called a "jack-knife test") is a sanity check on whether a model is too complicated to be justified. It is performed by reserved a small fraction of the total data (~5%) and not using it to construct or refine the model. After refinement, the ability of the model to explain this reserved set of observations is tested. An over-complex model will generally fail to fit the reserved data as well simpler, but valid, model.



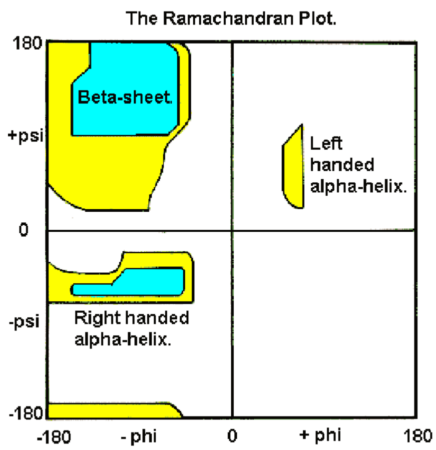
Crystallographers have occasionally gone astray by adding too many water molecules to the model (4 new parameters for each water molecule), by treating symmetry-equivalent protein models as being independent (doubling the number of parameters used), or refining B factors for every atom when the data resolution is too low to support it.

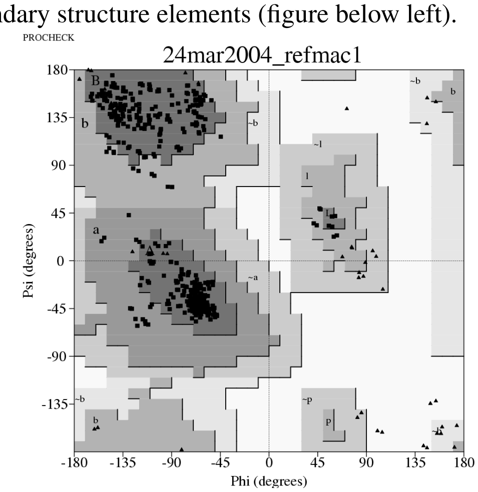
To avoid this, it is best to do cross validation by omitting a small subset of the data from the refinement, and using it only to calculate R_{free}. This calculation is is exactly the same as the one for $R4$ except that it is a sum over the reserved observations rather than the observations used for refinement. Any 'improvement' of the model that does not lower R_{free} as well as R should be looked upon with great suspicion. A good structure report will give the final R_{free} in addition to the final R. There is no precise rule for what the final R_{free} should be, but $0.05 < (R_{free} - R) < 0.10$ is pretty typical for proteins.

[1]Properly speaking, the requirement is that the ratio $\frac{\#parameters}{\#observations + \#restraints} < 1$. If this ratio ever gets larger than 1, the refinement is nonsense - think about fitting a line (2 parameters) through a single point (1 observation).

Reality checks - is the model OK?

Single numbers like R can at best be a global measure of reliability. They are not sensitive to small errors, and do not tell you if the model is better in one place than in another. Common sense, prior experience, and outside knowledge of general chemistry should be brought to bear also. We now have thousands of examples of well-refined protein structures, and this should lead us to be wary of a new structure that shows many odd bond lengths, torsion angles, unsatisfied hydrogen-bond donors/acceptors, hydrophilic residues in a hydrophobic environment, and so on. One of the oldest and best-known check of this sort is to plot the backbone torsion angles, φ and ψ, for each residue. This type of plot was first done by G N Ramachandran, who found that steric hindrance limited most residues to certain regions within the plot corresponding to regular secondary structure elements (figure below left).

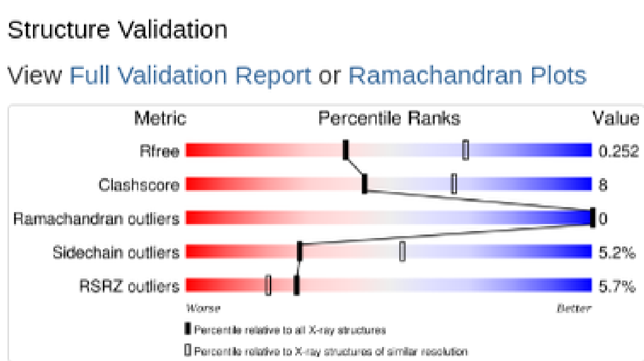




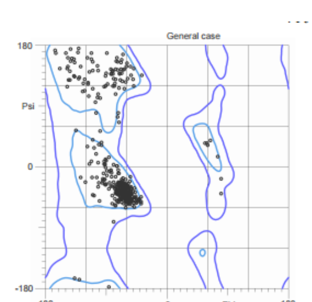
Right figure: The Ramachandran plot for an individual well-refined structure. Each black dot represents the ϕ, ψ values for one residue. Triangles represent glycine residues, which are less constrained because they have no sidechain. We expect to find that almost all of the residues from a well-refined structure lie in the core regions of the Ramachandran plot. This percentage is often reported in the results section of a journal article, and it may be the only clue you get as to the reliability of the structure other than the R factor.

Validation of PDB structures

The Protein Data Bank (PDB) is a repository of experimentally determined macromolecular structures. Essentially all previously reported protein crystal structures can be retrieved from the PDB. It now holds structures from NMR, electron microscopy, and a few other techniques. Because not all reported structures are of equal quality or reliability, the PDB has an extensive set validation checks that are made when a structure is deposited. The validation results are shown to the depositor immediately, are in theory provided to referees reviewing any papers that report the initial structure, and are available on the PDB web site to anyone who cares. The "front page" for each crystal structure in the PDB now shows how it scores on the validation tests compared to all other structures in the PDB. This comparison is presented in a graph like the one below.



(g) Summary of PDB validation checks, with links to full validation report



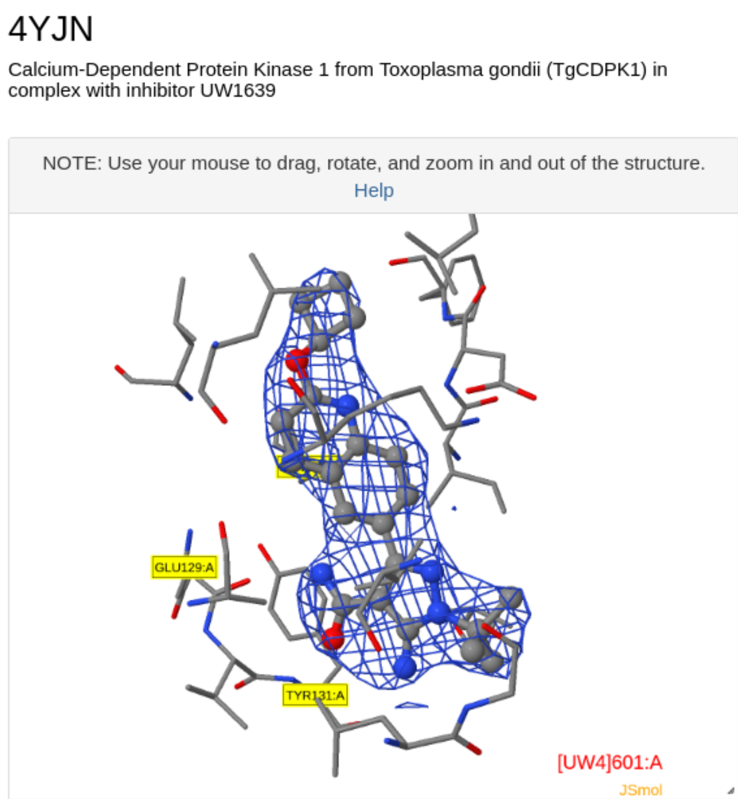
(h) Ramachandran provided as part of PDB validation

Validation of ligands

A primary reason for determining many crystal structures is to see how a small molecule binds to a protein a complex of proteins. Unfortunately this is often the least reliable part of a structure. This is because the tools for describing, modeling, refining, and validating the small molecules are not easy to use. Futhermore it is unfortunately true that the person who created the model for the ligand/inhibitor/drug/whatever is biased by prior expectations of what they will see in the crystal structure. The PDB validation tests now include checks for ligand bond lengths, bond angles, and agreement with the electron density map calculated from the deposited structure. Current estimates are that a third of all the ligands in PDB structures fail the validation checks, where "failures" range from minor deviation from expected bond angles to major disagreement with the electron density.

Mol	Type	Chain	Res	Atoms	RSCC	RSR	LLDF	B-factors($Å^2$)	Q<0.9
2	UW4	A	601	28/28	0.96	0.17	0.30	35,38,56,57	0

(i) Ligand validation table from full validation report for PDB entry 4YJN. The RSCC (realspace correlation coefficient) is one measure of agreement with the electron density (RSCC=1 is perfect correlation). LLDF is an estimate for the reliability of ligand atoms positions with compared to the reliability of nearby protein atoms. LLDF=1 means equally reliable. LLDF<1 means the ligand atoms are more reliably placed that the protein atoms. LLDF >> 1 means the ligand atom placement is unreliable.



(j) Several applets are available from the PDB page for each structure to inspect the electron density map in the region of a ligand.

Key things to remember about X-ray diffraction

1. X-ray scattering is only sensitive to the position of electrons. The difference between an oxygen atom (8 electrons) and a nitrogen atom (7 electrons) is small. On the other hand there is a dramatic difference in visibility between Li^+ and K^+ despite their similar chemical properties. The local electron density at a particular position in the unit cell is denoted $\rho(x,y,z)$.

2. *Bragg's Law* tells us that X-rays scattered by an ideal crystal will go off in a discrete set of directions. These are called *Bragg reflections*, and are denoted F_{hkl}.

3. The *resolution* of an X-ray structure is determined by how many of the possible Bragg reflections are measured. That means you know the resolution even before you solve the structure! The achievable resolution is usually limited by the fact that the Bragg reflections get weaker as the scattering angle gets larger; eventually they cannot be distinguished from the background noise.

4. One Fourier transform describes the scattering of X-rays by the crystal to produce a set of Bragg reflections. This is Eq (1).

5. A second, closely related, Fourier transform describes how to reconstruct a model of the unit cell contents from the Bragg intensities (which we have measured) and associated phases (which we unfortunately cannot measure). This is Eq (3).

6. Every electron in the cell contributes to each F_{hkl}. Conversely, every Bragg reflection contains information about the local electron density ρ at each position (x,y,z) in the unit cell.

7. F_{hkl} is a complex number. It has both an amplitude and a phase. Unfortunately, we can only measure the amplitude.

8. Since we cannot measure the phases directly, we must start with an approximation for their values. Eq (3) then gives an initial electron density map, often quite a bad map, into which we must build a model for the molecules in the unit cell. As we refine the model, the calculated phases improve, which yields successively better maps.

9. Crystallography usually uses a single wavelength of X-rays, but for some experiments it is advantageous to use X-rays with a range of wavelengths. Bragg's Law still predicts a discrete diffraction pattern, but each wavelength of X-rays will produce a corresponding set of Bragg reflections.

Key things to remember about protein crystals

1. A crystal is regular 2-dimensional or 3-dimensional array of identical *Unit Cells*. This array is called the *Crystal Lattice*.

2. The overall symmetry of the unit cell is called the *Space Group*.

3. Depending on the space group, the unit cell consists of one or more copies of the *Asymmetric Unit*. Only one of these is unique; the others are generated by the space group symmetry operations.

4. The core of a crystallographic structure determination is to locate and identify the atoms in one asymmetric unit.

5. There can be more than one copy of the protein in a single asymmetric unit. In this case crystallography gives multiple views of the protein structure that may be compared with each other.

6. Conversely if the protein is a dimer (or trimer or higher-order assembly), then it can happen that only half (or a third, etc) of the protein is in the asymmetric unit.

7. In a typical protein crystal roughly half of the unit cell is occupied by the protein. The rest of it is filled by the aqueous buffer from which crystals were grown. Thus the protein's local environment in the crystal is not so very different from the environment in a living cell.

   (a) *Note:* This is very different from a small-molecule crystal, where typically the molecule that has been crystallized packs against itself so snugly that there is no space remaining for non-crystalline water or other solvent. This is the major reason that different techniques are needed for determining and interpreting small-molecule and macromolecule crystal structures.

Quality checklist

1. For a good model, the predicted values of Bragg reflections $|F_{calc}|$ agree well with the observed values $|F_{obs}|$. This gives a low value of the crystallographic residual R (Eq 4).

2. A high value of R is always bad, but a low value of R does not guarantee that a structure model is good. A reality check can be made if a small subset of the data is omitted from the refinement, and used only to calculate R_{free}. Typically a good protein structure will end up having $0.05 < (R_{free} - R) < 0.10$.

3. Rough rule-of-thumb: if there are more water molecules in the model than protein (or nucleic acid) residues then you should be suspicious. At very high resolution (≤ 1.5Å) this rule may be relaxed, as there are many more observations and at this resolution it is much easier to identify water molecules.

4. Missing data introduces noise into the electron density map, as the corresponding Bragg reflections are necessarily left out of the Fourier summation. Rough rule-of-thumb: The overall completeness of the data measured should be greater than 90%. Low resolution data and weak reflections should be included in refinement, because even weak data is more informative than not having a measurement at all.

5. The model should not be too complicated for the number of observations. This is true for all experimental science, not just crystallography. The number of Bragg reflections (observations) increases as the cube of the resolution. A high resolution structure determination (better than, say, 2.5 resolution) can support a more complex model and more detailed conclusions than a low resolution experiment. Example of a model that may be "too complicated": a low resolution structure of a dodecamer in which each subunit is modeled separately is 12 times as complicated as a model in which all subunits are treated as having identical geometry.

6. The protein structure should be believable! For instance, it is very unusual for more than a small handful of residues to lie outside the core regions of a Ramachandran (φ/ψ) plot. The structure should also make biological sense. It should explain, or at least not contradict, previous experimental work such as the effect of mutations or the specificity for particular substrates, inhibitors, or binding partners.

7. If there are small molecule ligands present, they should have believable stereochemistry and ligand atoms should match the experimental electron density maps just as the protein atoms should match that same map. The LLDF score in a PDB validation report is an attempt to quantify this. LLDF is expected to be $\approx$1 when the ligand and the protein fit the density equally well.