Deducing protein atomic structure from x-ray diffraction patterns of protein crystals

1. Review of where we are, and the remaining challenge—the phase problem

It seems then that we have almost achieved our long sought-after goal to predict the atomic structure of a molecule, based on the diffraction pattern of crystals composed of such molecules. From last lecture, we deduced ways of inferring the ‘Structure Factors’ from the diffraction pattern, where the said factors are equivalent to the magnitude of the Fourier coefficients for the Fourier series representation of the electron density within a unit cell. If we also knew the phases of each Fourier coefficient \((\varphi_{h,k,l})\), which seems like a minor detail, we could directly plug into the Electron Density Equation, and calculate directly the electron density

\[
\rho(x, y, z) = \frac{1}{V} \sum_{h,k,l=0}^{\infty} F(h,k,l) \cos \left[ 2\pi \left( \frac{hx}{a} + \frac{ky}{b} + \frac{lz}{c} \right) - \varphi_{h,k,l} \right]
\]

where \(x, y, z\) give the position within a unit cell, with corresponding dimensions \(a, b, c\); \(V\) is the volume of a unit cell in angstroms\(^3\); \(F(h,k,l)\) turns out to be proportional to the square root of the intensity at the potential diffraction pattern spot corresponding to a given \(h,k,l\) planar cosine. As an aside, for purposes of deductions later on in this lecture, another more convenient (but equivalent) way of writing the Electron Density Equation is (the complex part will zero out)

\[
\rho(\mathbf{x}) = \frac{1}{V} \sum_{h,k,l=0}^{\infty} F(\mathbf{h}) \exp \left[ 2\pi j (\mathbf{h} \cdot \mathbf{x}) \right]
\]

where \(\mathbf{x} = \begin{bmatrix} x/a \\ y/b \\ z/c \end{bmatrix}\); \(\mathbf{h} = [h \ k \ l]\); and \(F(\mathbf{h}) = |F(\mathbf{h})| \exp[-j\varphi_{h,k,l}]\)

Having calculated \(\rho(x, y, z)\), we could then thread a protein backbone through the electron density, to the best of our ability. Then we could check the accuracy of our molecular threading, by recalculating the structure factors according to the so-called Structure Factor Equation.
DEDUCING MOLECULAR STRUCTURE FROM DIFFRACTION PATTERNS

deduced at the end of Chapter 7. The result we achieved was the Structure Factor Equation below. $F(h, k, l)$ here is the complex structure factor.

$$F(h, k, l) = \sum_{m=1}^{n} f_m \exp \left[ j2\pi \left( h x_m / a + k y_m / b + l z_m / c \right) \right]$$

where $m$ indexes overall all atoms in the unit cell, $f_m$ is the scattering factor of atom $m$ (related to total electron mass and how dense or still the atom is), and $f_{h,k,l,m}$ is the atomic structure factor for atom $m$. This equation would predict the complex structure factors $F(h,k,l)$

$$= F(\mathbf{h}) = |F(\mathbf{h})| \exp \left[ -j\varphi_{h,k,l} \right]$$

based on the positions and net electron scattering mass of individual atoms (as we have threaded them) within the unit cell. We could then compare these calculated structure factors against those we measured experimentally from the diffraction patterns. In this way, we could iteratively improve our atomic model of the molecule. Job done!

Not quite. The fatal flaw in this ‘garden rosy path’ is that we have thus far lacked an experimental method to determine the phases $(\varphi_{h,k,l})$, and most of the needed information is actually embedded in these phases. Thus, our initial calculation of $\rho(x, y, z)$ would be essentially worthless.

The problem of determining the phases of the structure factors then becomes a critical and very difficult challenge. No direct means of measuring the phase are yet available, as they might be for electromagnetic waves at lower frequencies (as in radio or light waves).

2. Determining phase by multiple isomorphous replacement of heavy atoms

The approach here is to give ourselves some GPS landmarks in the diffraction pattern, so that we can unravel the phases. In particular, we implant heavy atoms in a few distinct places within the crystal, without disturbing the overall structure of the crystal. These heavy atoms diffract very strongly relative to other atoms normally found in proteins; thus, their signature in the overall diffraction pattern is obvious (Figure 1), despite the small number of heavy atom ‘beacons.’ For example, when soaked into crystals, the heavy atom Hg$^{2+}$ reacts specifically with surface-exposed cysteine residues.

Figure 1. Underlined reflections show differences upon heavy atom labelling (panel b). This exemplar is better in the original. The key is that one compares the order of intensity in adjacent reflection pairs, as underlined. The leftmost underlined pair is perhaps the best example.
Now, we can in principle obtain magnitudes of the complex structure factors for the protein itself \((F_p(\mathbf{h}, k, l) \) or \(F_p\)) and for the protein + heavy atom \((F_{HP}(\mathbf{h}, k, l) \) or \(F_{HP}\)) . From vector addition rules for complex numbers, we know that \(F_{HP} = F_p + F_H\), where \(F_H\) is the complex structure factor for the heavy atoms alone (Figure 2).

From the intensity of diffraction pattern reflections, we have the intensities \(I_{HP}\) and \(I_P\). As an approximation, we can calculate the magnitude of \(F_H\), as follows. It turns out this approximation is usually good enough for purposes of locating heavy atom positions via the Patterson function approach described later below.

\[
|F_H|^2 \sim \text{abs}(|F_{HP}|^2 - |F_P|^2) = \text{abs}(I_{HP} - I_P)
\]

This estimate is necessary for getting the position of the heavy atoms within the unit cell by the Patterson approach described below. Once we know the position of the heavy atoms within the unit cell, we could also determine the phase of this \(F_H\) term, according to the Structure Factor Equation for the heavy atoms alone, as stated below.

\[
F_H(\mathbf{h}, k, l) = \sum_{k=1}^{q} f_k \exp \left[ j2\pi \left( \frac{h x_k/a + k y_k/b + l z_k/c}{f_{\lambda h, k}} \right) \right]
\]

where there are \(q\) heavy atoms per unit cell, and \(k\) indexes over the heavy atoms in a unit cell. We will develop how the positions of the heavy atoms can be determined in section 3, but for the moment let’s assume we can do this.

Armed with the approximate magnitude of \(F_H\) and its phase, we can then determine the phase of structure factors for the protein alone, at least for those reflections that are appreciably perturbed by heavy metal addition. The approach is as follows. From Figure 2, we know that the following complex vector equality must hold: \(F_p = F_{HP} - F_H\) . We know the magnitude and phase of \(F_{HP}\) and we know the magnitudes of \(F_{PH}\) and \(F_P\) (but not their phases). Hence, on the complex plane, we can add \(-F_H\) to the origin, and draw a circle with radius \(|F_{HP}|\) (Figure 3a). This circle represents the constraint on \(F_P\), by the right hand of the equation \(F_p = F_{HP} - F_H\); \(F_P\) must reside somewhere on the circle sketched in Figure 3a. The other constraint on actual complex vector \(F_P\) is that it must reside on a circle, with origin at the graph origin, and with radius equal to the
magnitude of $F_P$. This second constraint is shown by the red circle in Figure 3b. Reproducing the circle from Figure 3a onto Figure 3b, superimposes the second constraint onto the deliberations in Figure 3b. Hence, the actual complex vector $F_P$ must satisfy both constraints and reside at the intersection of red and gray circles. We have thus determined (for many $h$, $k$, $l$ reflections) the phase of $F_P$, up to an ambiguity of two choices per reflection.

To get around this still critical ambiguity, we repeat the entire process with another heavy atom, which must label at different spots within the unit cell, different enough to produce a different $F_H$, but similar enough to effect some of the same reflections as did the first heavy metal. This is where the *multiple* in multiple isomorphous replacement comes in. This gives rise to determination of $F'_H$ for the second heavy atom. Repeating the process of determining $F_P$ with this second heavy atom yields the analysis in Figure 4a. The unique solution for $F_P$ is the one that satisfies both heavy atom constraints (circled in red). Thus, we have determined phases and magnitudes for $F_P$, for reflections appreciably affected by heavy atom substitution. From these, we can calculate a crude first estimate of $\rho(x, y, z)$, as given by the Electron Density Equation.

The catch with this method is that it is hard enough to get a good normal protein crystal. Multiple isomorphous replacement requires that we do it twice more, with different heavy atoms. In times of yore, this represented a significant challenge. In section 4, we will talk about a much improved method of phasing, called multiple anomalous diffraction (MAD). This method only requires that we make the protein crystal twice (with and without ‘anomalous’ heavy atoms), yet gives us much more accurate estimates of phase. Also, the insertion of these anomalous heavy atoms is much easier than for say Hg$^{2+}$. Because MAD also requires the capability of determining where heavy atoms (latent or actual) are situated within a unit cell, we will first develop how such fiduciary positions can be determined.
3. Determining the position of heavy atoms by the Patterson function

From the difference of reflection intensities of the \(HP\) and \(P\) contexts, we can approximately determine \(|F_H|^2 \sim \text{abs}(|F_{HP}|^2 - |F_P|^2) = \text{abs}(I_{HP} - I_P)\). Without knowledge of phases, we could calculate a ‘close cousin’ of the Electron Density Equation for the heavy atoms, called the (difference) Patterson Function, given by

\[
\Delta P(\mathbf{u}) = \frac{1}{V} \sum_{h,k,l=0}^\infty \left| F_H(\mathbf{h}) \right|^2 \exp \left[ 2\pi j \left( \mathbf{h} \cdot \mathbf{u} \right) \right]
\]

where \(\mathbf{u} = \begin{bmatrix} u/a \\ v/b \\ w/c \end{bmatrix}\), and \(\mathbf{h} = [h \ k \ l]\)

This is actually equivalent to the cross-correlation function of heavy-atom electron densities \(C_H\), which is a function of the vector between points. This can be seen, somewhat miraculously, as follows.

\[
C_H(\mathbf{u}) = \int_{\text{entire crystal volume}} \rho(\mathbf{x}) \cdot \rho(\mathbf{x} + \mathbf{u}) \, d\mathbf{x}
\]

\[
= \int_{\text{entire crystal volume}} \left( \frac{1}{V} \sum_{\mathbf{h}'} F_H(\mathbf{h}') \exp \left[ -2\pi j \left( \mathbf{h}' \cdot \mathbf{x} \right) \right] \right) \left( \frac{1}{V} \sum_{\mathbf{h}} F_H(\mathbf{h}) \exp \left[ -2\pi j \left( \mathbf{h} \cdot (\mathbf{x} + \mathbf{u}) \right) \right] \right) \, d\mathbf{x}
\]

\[
= \frac{1}{V^2} \sum_{\mathbf{h}} \sum_{\mathbf{h}'} F_H(\mathbf{h}') \cdot F_H(\mathbf{h}) \exp \left[ -2\pi j \left( \mathbf{h} \cdot \mathbf{u} \right) \right] \int_{\text{entire crystal volume}} \exp \left[ -2\pi j \left( \mathbf{h} + \mathbf{h}' \cdot \mathbf{x} \right) \right] \, d\mathbf{x}
\]

\[
= \frac{N}{V} \sum_{\mathbf{h}} F_H(\mathbf{h}) \cdot F_H(\mathbf{h}) \exp \left[ -2\pi j \left( \mathbf{h} \cdot \mathbf{u} \right) \right]
\]

\[
= \frac{N}{V} \sum_{\mathbf{h}} \overline{F_H(\mathbf{h})} \cdot F_H(\mathbf{h}) \exp \left[ -2\pi j \left( \mathbf{h} \cdot \mathbf{u} \right) \right]
\]

\[
= \frac{N}{V} \sum_{h,k,l=0}^\infty \left| F_H(\mathbf{h}) \right|^2 \exp \left[ 2\pi j \left( \mathbf{h} \cdot \mathbf{u} \right) \right]
\]

This is equivalent (up to a factor \(N\), the number of unit cells) to \(\Delta P(\mathbf{u})\). The last 2-3 steps requires Friedel’s law, given by

\[
F_H(h,k,l) = \sum_{k=1}^q f_k \exp \left[ j2\pi \left( h x_k / a + k y_k / b + l z_k / c \right) \right]
\]

\[
F_H(-h,-k,-l) = \sum_{k=1}^q f_k \exp \left[ -j2\pi \left( h x_k / a + k y_k / b + l z_k / c \right) \right] = \overline{F_H(h,k,l)}
\]

because \(f_k\) factors must be real.
With this insight, we can ask, what would the cross-correlation function of heavy-atom electron densities look like? It isn’t too bad for a few heavy atoms. Consider the three atoms within the unit cell shown in Figure 5a. All the cross correlation peaks from within this unit cell would show up as peaks in the Patterson function shown in Figure 5b.

Considering cross correlation peaks from heavy atoms in adjacent unit cells, we get the complete picture shown in Figure 5c. You can easily deduce that there will be $q \cdot (q-1)$ Patterson peaks for $q$ heavy atoms per unit cell.

The relative arrangement of heavy atoms must include one at the origin, and $(q-1)$ others at some $(q-1)$ of the Patterson peaks. A computer program can exhaustively try all the possibilities, and see which one gives a predicted Patterson function in agreement with that of the experimentally determined function. In this way, we can specify the relative position of the heavy atoms within the unit cell. To obtain the actual position of heavy atoms within the unit cell, we move around the relative assembly of heavy atoms within the unit cell, and plug into equation * (page 3) to match the predicted $|F_H|^2$ to the empirical estimate of $|F_H|^2$ ( $\sim abs(I_{HP}-I_P)$ ) obtained earlier on page 3. Once we have this match, we can use equation * to solve for the phases of $F_H$, so that we can solve for the phases of $F_P$ at reflections appreciably affected by heavy metal insertion (Figs. 3-4).

**Figure 5.** Generating a Patterson map from three heavy atoms as in panel a. Panel b shows peaks from correlations within a unit cell. Panel c shows all peaks, from correlations within and without a unit cell.
5. Determining phase by multiple anomalous diffraction (MAD)

Here, the most popular and powerful approach is to use a GPS landmark carried by selenium, which is part of selenomethiones, incorporated straight into the recombinant protein by appropriate manipulation of bacterial media when expressing proteins. At certain wavelengths ($\lambda_1$) of x-ray illumination, selenions behave just like other heavy atoms. This would give rise to the HP reflection pair $F_{\text{HP}}(+,\lambda_1)$ and $F_{\text{HP}}(−,\lambda_1)$ (Figure 6). Recall that a pair of reflections must occur as a property of diffraction, by analogy to our 1 dimensional case. $F_H$ and its Friedel pair (the conjugate of $F_H$), are as shown. The latter can be determined by the Patterson function analysis, and the Structure Factor Equation for heavy atoms.

At a second wavelength $\lambda_2$ (possible with synchrotron radiation to adjust wavelength) selenium doesn’t just scatter radiation, but puts a resonance phase advance on reflections. These can be calculated from standard tables for selenium and knowledge of $\lambda_2$. The net effect is to extend $F_H$ along its non-anomalous ($\lambda_1$) direction by $\Delta F_r(+)$, and then twist it orthogonally (this is the phase advance effect) by $\Delta F_i(+)$. This gives rise to $F_{\text{HP}}(+,\lambda_2)$ as shown. The net effect for the Friedel pair of $F_H$ is to extend the $\lambda_1$ conjugate of $F_H$ along its initial direction by $\Delta F_i(−)$, which is the conjugate of $\Delta F_i(+)$. It is the simple conjugate because there is nothing anomalous about the $\Delta F_i$ effect. However, the phase advance part, also adds orthogonally by $\Delta F_i(−)$, in the same sense as was effectuated for $F_H$. The resulting Friedel pair for $F_{\text{HP}}(+,\lambda_2)$ is $F_{\text{HP}}(−,\lambda_2)$, as shown.

Figure 6. Vector complex analysis of the effects of selenium, during normal and anomalous diffraction modes.
From Figure 6, we can deduce that
\[ F_{HP}(+, \lambda_1) = F_{HP}(+,-\lambda_2) - \Delta F_R(+) - \Delta F_i(+) \]
This gives rise to the two possible values of \( F_{HP}(+, \lambda_1) \) shown in Figure 7.

![Figure 7. Consideration of normal scattering mode of selenium yields estimate of \( F_{HP} \), with a two-fold ambiguity.]

To get a second constraint, consider how \( F_{HP}(+, \lambda_1) \) is related to \( F_{HP}(-, \lambda_2) \). Considering Figure 6, we have that
\[ F_{HP}(+, \lambda_1) = F_{HP}(-, \lambda_2) - \Delta F_R(-) - \Delta F_i(-) \]
\[ F_{HP}(+, \lambda_1) = F_{HP}(-, \lambda_2) - \Delta F_R(+) + \Delta F_i(+) \]
Deploying this in Figure 8, gives another pair of solutions for \( F_{HP}(+, \lambda_1) \), only one of which is compatible with the analysis in Figure 7. Thus, we have determined \( F_{HP}(+, \lambda_1) \). With knowledge of \( F_H(\lambda_1) \), we can calculate \( F_p(\lambda_1) \) including phase. In reality, several wavelengths are used to get a really good solution for \( F_{HP}(+, \lambda_1) \). This is powerful.

![Figure 8. Consideration of anomalous scattering mode of selenium yields unique determination of \( F_{HP} \), without ambiguity.]

6. Refinement by iteration
Real-space error removal on first pass includes: removal of negative electron density, low-pass filtering of electron density for divets, flattening the solvent portions, scaling overall amplitude so that the density of protein to solvent areas is appropriate for these two phases. At some point, a molecular model can be inserted.

**Figure 9.** Flow diagram of refinement strategy.