Solving protein crystal structures

Hugo L. Monaco
Biocrystallography Laboratory,
Department of Biotechnology, University of Verona,
Strada Le Grazie 15, Verona 37134, Italy

1. The phase problem

The ultimate goal of an X-ray diffraction study of a protein crystal is to produce a model of the molecule, i.e a list of the coordinates of its atoms in some selected coordinate system. The model is built by fitting atoms to the electron density of the asymmetric unit of the crystal. We recall that the asymmetric unit is the smallest unit from which the crystal structure can be generated by making use of the symmetry operations of the space group of the crystal. The asymmetric unit can be one molecule, several molecules or a subunit of an oligomeric molecule. Knowing the electron density in the crystal, measured for example in electrons per $Å^3$, is equivalent to knowing the relative position of the atoms. As seen in previous lessons, the electron density is the Fourier transform of the structure factors. In the case of a crystal the relationship between these two quantities can be expressed in a summation as follows:

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

where $\rho(x,y,z)$ is the electron density, a function of the spatial coordinates x, y and z and h, k an l are three integers, V is a proportionality constant, the volume of the unit cell of the crystal and can be neglected since we are basically interested in relative values, F_{hkl} are the structure factors, in the case of a crystal, a function of the three indexes, h, k and l, three integers.
Structure factors are defined as the ratio of the amplitude of the radiation scattered by the sample to that scattered by a single electron at the origin. In the case of a crystal their value is different from 0 at defined positions in space which are characterized by the three integers, h, k and l. They are complex numbers and as such can be written as follows.

$$F_{hkl} = |F_{hkl}|\ e^{i\phi}$$

where $|F_{hkl}|$ is the amplitude and ϕ is the phase. Alternatively structure factors can be written as the sum of a real and an imaginary part

$$F_{hkl} = F_r + i\ F_i$$

where F_r is the real and F_i the imaginary part
Experimentally we can only measure the intensity of the radiation scattered by our sample which is the product of the structure factor times its complex conjugate.
.

$$I_{hkl} = F_{hkl} \cdot F_{hkl}^*$$

$$I_{hkl} = |F_{hkl}|\ e^{i\phi}\ |F_{hkl}|\ e^{-i\phi} = |F_{hkl}|^2$$

Thus the diffraction experiment yields only the structure factor amplitude and the phase term is not measured directly but has to be determined or estimated indirectly. This is one of the major obstacles in the road to a structure determination by X-ray diffraction methods and is known as "the phase problem". In the sections that follows we will discuss how the phase problem is solved in macromolecular crystallography.

2. The multiple isomorphous replacemente method.

This is historically the first successful method used to solve macromolecular structures and although other phasing methods are currently available it can still be said that the M.I.R. method is still central in macromolecular crystallography. The method uses a minimum of two heavy atom isomorphous derivatives of the protein crystal. A perfect isomorphous derivative is one in which the native protein and the derivative crystal belong to identical space groups and have identical unit cell parameters. The only difference between the two is intensity changes in the reflections. The derivative is prepared by reacting native protein crystals with a "heavy atom" which means an atom with a large atomic scattering factor i.e with a large number of electrons. We recall that protein crystals have mother liquor channels that can be used as a route for heavy atoms to get in contact with the protein molecules and hopefully react in selected points substituting in the structure disordered solvent molecules. This method used to prepare the derivative is called by "soaking" and is usually the method of choice. Attempts to react the protein with the heavy atom in solution followed by crystallization used to be made but are currently not common since they often result in either the derivatized protein not crystallizing or in the formation of non-isomorphous crystals.

If the heavy atom derivative has been successfully prepared, we can say that in it the total electron density is the sum of the electron density of the protein plus that of the heavy atom:

$$\rho_{PH} = \rho_P + \rho_H$$

If we Fourier transform this equation the result is that

$$\mathbf{F_{PH}} = \mathbf{F_P} + \mathbf{F_H}$$

Where $\mathbf{F_{PH}}$ is the structure factor of the derivative, $\mathbf{F_P}$ is the structure factor of the protein and $\mathbf{F_H}$ is the structure factor of the heavy atom in the unit cell of the protein, obviously a theoretical quantity since crystals of the space group and unit cell parameters of the protein containing only the heavy atom cannot be prepared. The experimental measurements that can be made are the intensities of native protein and derivative which, as we saw above, are proportional to the structure factor amplitudes of protein and derivative. Thus our problem is to derive the phases of $\mathbf{F_P}$ with the amplitudes $|\mathbf{F_P}|$ and $|\mathbf{F_{PH}}|$.

For any crystal we can calculate the structure factors if we know the positions of the atoms using the following equation:

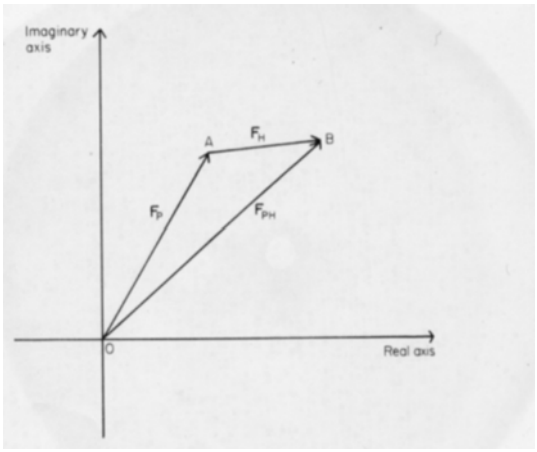
$$F_{hkl} = \sum_i f_i \, e^{2\pi i(hx_i+ky_i+lz_i)}$$

where h, k and l are the indexes identifying each structure factor, f_i is the atomic scattering factor of each atom in position (x_i y_i z_i) and the summation is done over all the atoms constituting the partial or total structure.

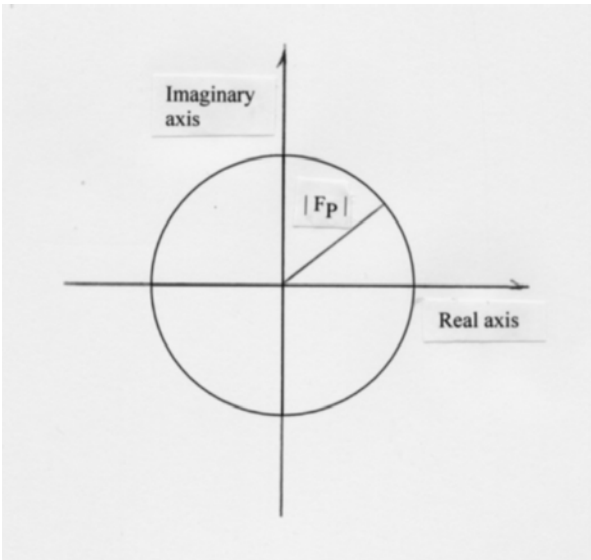
We will describe below how the position of the heavy atoms in the unit cell of the protein can be determined but for the time being let us assume that it has been done. If so we can the calculate $\mathbf{F_H}$, both the amplitude and the phase.

$$F_H = | F_H | \ e^{i\phi_H}$$

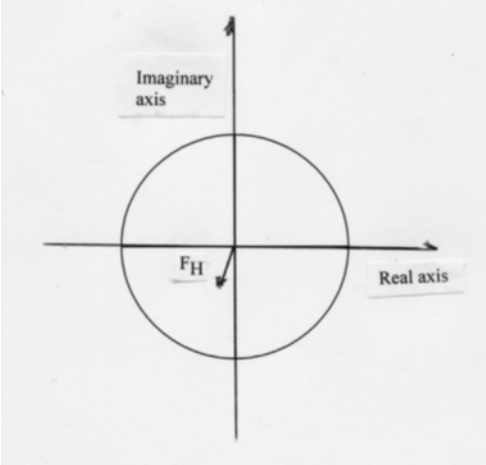
We will now make use of two types of diagrams, vector diagrams in which structure factors are represented as vectors and the Harker constructions to explain the relationships between structure factors and phases.
The three structure factors of the equation $F_{PH} = F_P + F_H$ are represented in a vector (or Argand) diagram below:



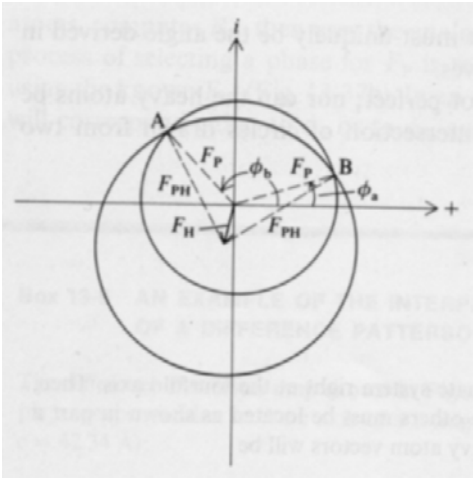
If we have measured the intensities of the native protein and one heavy atom derivative we know the amplitude of two structure factors, $| F_P |$ and $| F_{PH} |$, and if we have determined the substructure of one heavy atom derivative we know amplitude and phase of its contribution,
$F_H = | F_H | \ e^{i\phi_H}$
knowing the amplitude but not the phase of the native protein structure factor, F_P is equivalent to knowing that the amplitude lies on a circle whose radius is $| F_P |$, we know the length but not the position of the vector representation of F_P because there is an infinite number of possible phases



but the contribution of F_H is completely defined in amplitude and phase in the vector diagram.



and so if we now trace another circle with a radius equal to $|F_{PH}|$ but centred at the end of F_H and not at the origin there are only two points that satisfy the condition that $F_{PH} = F_P + F_H$, the intersections of the two circles, A and B and we have reduced the possible phases of F_P, the structure factors of the protein from an infinite number to two!



The next step is to choose between these two possible phases and in order to do so the simplest way is by preparing a second heavy atom derivative. Examination of the expression given above used to calculate F_H indicates clearly that the most important factor in preparing the second derivative is that the position of the atom should be different from that of the first. The atomic scattering factor has no influence on the phase which is the most important part to determine a very different F_H that will let us choose easily the correct phase for the protein.
The method is called multiple isomorphous replacement (M.I.R) because historically more than two derivatives were prepared to improve the quality of the protein phases. The reason is that there are many errors involved in the operation and more derivatives tended to imply better phases.
Many variations of this basic ideas are used nowadays to phase protein crystals, the major problem of this method was and still is lack of isomorphism between the native protein and the derivatives. In the following section we discuss the methods used to determine the position of the heavy atoms in the unit cell of the protein

3. Solving the heavy atom substructure

The classical method used to find the positions of the heavy atoms in the protein unit cell use the Patterson function calculated with the differences between the structure factor amplitudes of the derivative and the protein.
The Patterson function is the Fourier transform of the intensities of the reflections of the crystal and as such presents some analogies with the electron density function.

$$P(x,y,z) = 1/V \sum_h \sum_k \sum_l I_{hkl} \; e^{-2\pi i(hx+ky+lz)}$$

since $I_{hkl} = F_{hkl} \cdot F_{hkl}^* = |F_{hkl}|^2$ and $F_{hkl} = \sum_i f_i \, e^{2\pi i(hx_i+ky_i+lz_i)}$

$$I_{hkl} = \sum f_i \; f_k \; e^{2\pi i [\, h(x_i - x_k) + k(y_i - y_k) + l(z_i - z_k) \,]}$$

Thus the Patterson function is like an electron density of pseudo-atoms with atomic scattering factors that are the product of the actual scattering factors of all the possible combinations of the atoms effectively present in the unit cell and whose peaks are found in positions that correspond to the differences between the values of the actual coordinates of the atoms.
In order to find the position of the heavy atom ideally one would like to have $| F_H |^2$ but that amplitude is not available because it would be the scattering amplitude of the heavy atom in the unit cell of the protein and so is estimated as the difference $| F_{PH} | - | F_P |$

and so to find the positions of the heavy atoms in the unit cell the difference Patterson function is calculated as follows

$$\Delta P = 1/V \sum_h \sum_k \sum_l [\, |F_{PH}| - |F_P| \,]^2 \; e^{-2\pi i(hx+ky+lz)}$$

After interpreting the Patterson map either manually or automatically by means of one of the very powerful existing computer programs one is in a position to calculate $F_H = | F_H | \; e^{i\phi_H}$
and then proceed as described above.
This method is by no means the only possibility to solve the substructure but it is historically the oldest, an alternative is the use of direct methods for phasing. Direct methods were originally developed for the determination of small molecule structures. They attempt to derive the structure factor phases directly from the amplitudes through the use of mathematical relationships. When applied to the problem of finding the position of the heavy atoms what is done is simulate a small molecule structure using as input the differences $|F_{PH}| - |F_P|$.
If more than one derivative is used, the positions of the second, third, etc. heavy atoms is found using the difference Fourier methods which will be described in the last section of these notes.

4. Using anomalous scattering.
In all the equations we have seen so far the atomic scattering factor is consider to be a real number which is true if the wavelength used for the diffraction experiment is not near a wavelength absorbed by the atom. Since

$F_{hkl} = \sum f_i \, e^{2\pi i(hx_i+ky_i+lz_i)}$ and $I_{hkl} = F_{hkl} \cdot F_{hkl}^*$

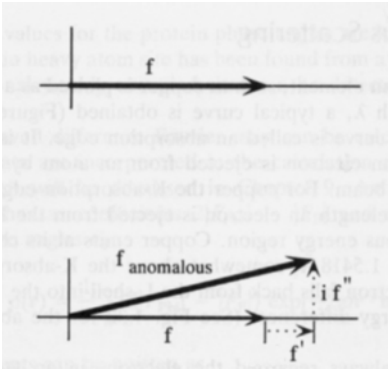
$$I_{h\,k\,l} = \sum f_i \, e^{2\pi i(hx_i+ky_i+lz_i)} \sum f_i \, e^{-2\pi i(hx_i+ky_i+lz_i)}$$

and for the reflection with the same indexes but having opposite sign

$$I_{\bar{h}\bar{k}\bar{l}} = \sum f_i \, e^{-2\pi i(hx_i+ky_i+lz_i)} \sum f_i \, e^{2\pi i(hx_i+ky_i+lz_i}$$

and thus $I_{h\,k\,l} = I_{\bar{h}\bar{k}\bar{l}}$

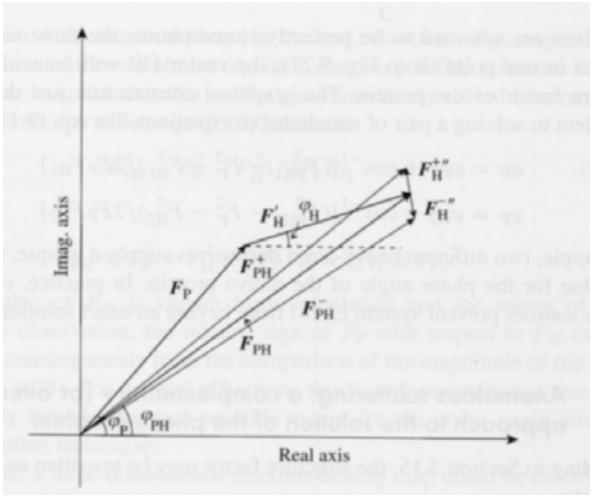
This relationship is known as Friedel's law and it applies for as long as all the atomic scattering factors can be considered real numbers. The two reflections with identical indexes having opposite signs are called Friedel pairs. When the X-rays wavelength approaches the absorption edge wavelength of an atom its atomic scattering factor cannot be considered only real anymore, the phenomenon is called anomalous scattering and Friedel's law is not valid anymore.
In a vector representation the atomic scattering factor has now a real and an imaginary part as indicated in the following figure:



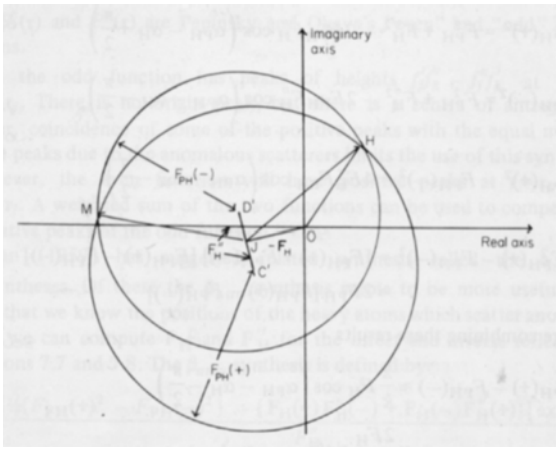
and

$$f_a = f + f' + i\,f''$$

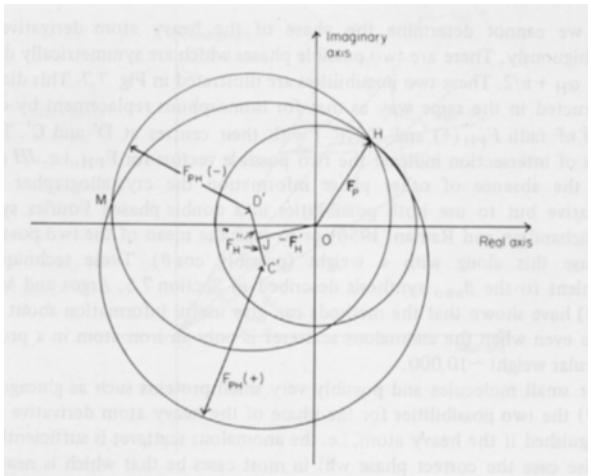
Anomalous scattering is not important for the atoms that are normally found in proteins with the exception of sulphur which has a signal that by appropriately choosing the wavelength of the X-rays can be used for phasing. The phenomenon becomes important for atoms with a large number of electrons like those used for heavy atom phasing and measurement of the anomalous signal becomes thus an important complement of the isomorphous replacement method.
If a heavy atom derivative has a non negligible anomalous signal the two reflections with indexes with opposite signs, that are called Bijvoet pairs instead of Friedel pairs, will not have equal intensities and there will be a measurable difference in the structure factor amplitudes of the two members of the pair. A vector representation of the two structure factors of the derivative, indicating with a plus and a minus the signs of the indexes of the Bijvoet pair and the two contributions of the anomalous scattterer is shown in the following figure:



There will be thus two different positions for the F_H vector representation corresponding to the two members of the Bijvoet pair and there will be two circles of slightly different radii centred on those two vectors corresponding to the two amplitudes of $\mathbf{F_{PH}}$ (+) and $\mathbf{F_{PH}}$ (-).as shown in the following Harker diagram:



If this diagram is superimposed with the circle representing the structure facture amplitude of the native protein, $|F_P|$



It becomes clear that measurement of the Bijvoet pairs of a heavy atom derivative leads to unambiguous phasing of the protein structure factors. One might say that this is equivalent to having two heavy atom derivatives, but in fact it is better because in this second case the two sets of

structure factor measurements are provided by the same crystal and therefore lack of isomorphism between the two derivatives is no longer a problem

5. The multiple wavelength anomalous dispersion (MAD) technique.

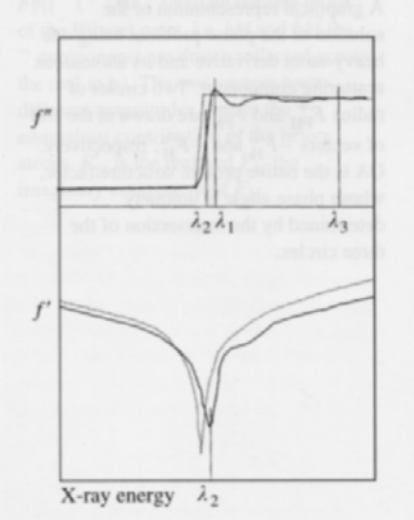
In the presence of anomalous scattering the magnitude of the real and imaginary parts of the atomic scattering factor are strongly dependent on the wavelength of the X-rays used for the diffraction experiment

Thus in the equation

$$f_a = f + f' + i\, f''$$

f is independent of the wavelength and falls off with the scattering angle θ whereas the anomalous part of the scattering factor is strongly wavelength dependent but is virtually independent of the angle θ. Perhaps more important is the fact that the wavelength dependence of f' and f'', the real and imaginary parts of the anomalous contribution is different.

As the following figure shows



the minimum in the values of f' is positioned at a slightly lower energy with respect to the maximum of f''. A MAD experiment exploits this behaviour to extract phase information.

A data set measured at the wavelength λ_1 corresponding to the peak of f''' of the anomalous scatterer will show the largest differences in the Bijvoet pairs because the imaginary component of the anomalous contribution is a maximum. A data set measured at the minimum of the real component of the anomalous scattering contribution will present very large differences in the structure facture amplitude when compared to the values obtained measuring the same intensities at the wavelength λ_3, distant from the maximum in f''. Using data collected at these three different wavelengths one can set up the equations that will yield the phase information. It is a obvious requisite of the method the possibility to carefully select the wavelength for the different measurements which can only be done if the data are collected at a synchrotron.

The advantage of this method is that usually all the data are measured from a single crystal which avoids the problems due to lack of isomorphism, a caveat is that the wavelengths have to be very carefully chosen, which is very simple in the case of λ_3 but more complicated in the case of λ_1 and λ_2. The method relies on the presence in the crystal of an anomalous scatterer that can give a measurable signal, and has become very popular to phase proteins that are recombinantly expressed in microorganisms. With the advent of the techniques of modern Molecular Biology it has, in fact,

become rather simple to substitute the amino acid methionine with selenomethionine, and thus replace a sulphur with a selenium at every position containing a methionine in the sequence. The presence of a rather limited number of selenium atoms in the unit cell is, in general sufficient, to generate an observable signal.

6. The molecular replacement method.

Is an alternative method of phasing which is based on positioning the model of a different, but rather similar molecule, or a fragment in the unit cell of the crystal whose structure is being determined. This operation is possible because more than 70,000 protein models are available in the protein data bank and chances are that the fold that our unknown belongs to is already present in the data set. This method of phasing is an obvious choice when trying to determine the three dimensional structure of a protein of a different species (and therefore quite similar) from one already solved. It is also very useful when dealing with protein complexes in which one or more partners have been studied and their X-ray structures solved.

If X is the set of coordinates of the model to be used for phasing and X' are the coordinates in the new crystal form the transformation one is trying to do is simply described by the following equation:

$$\backslash X' = [\, C \,] X + t$$

Where [C] is the matrix that rotates the coordinates so that the model is now oriented as the molecule present in the crystal and t is the translation that places the rotated model in the correct positions in the unit cell. The variables in this problem are thus six, three angles that rotate the model in the proper orientation and three translation vectors required to place it in the positions of the new unit cell.

The basic idea behind the methods used to find these six variables is quite simply stated. From the model one has calculated structure factors that can be used to calculate intensities and from them a Patterson map can be produced. Another Patterson map can be calculated from the experimental X-ray diffraction intensities of the other crystal form. The first of these two maps will depend on the orientation of the model and of its position in the new unit cell. If the model is oriented and positioned in the unit cell as the real molecules, the two Patterson maps should be closely correlated.

In a Patterson map there are contributions due to all the vectors between atoms pairs present in the same or in different molecules. Those corresponding to atoms of the same molecule are shorter and are called self-Patterson vectors, while those that relate atoms present in different molecules are called cross-Patterson vectors. The first group of vectors depends only on the molecule and therefore is expected to be very similar for model and molecule in the crystal to be solved while the second group depends on the packing of the molecules in the two unit cells.

The problem of finding the three variables required to position the model in the new unit cell is divided into two parts, in the first a rotation function is calculated in order to correctly orient the model, in the second part a translation function is calculated to superimpose the model to the molecules.

A rotation function can be defined as follows:

$$R(C) = \int_{v} P_{cryst}(u)\, P_{mod}(Cu)\, du$$

where C is the matrix that rotates the coordinates of the model, and $P_{cryst}(u)$ is the Patterson function of the crystal.

This function should have a maximum when the two Pattersons overlap and that should happen when the model has been properly rotated.
Once the orientation of the model has been defined, the second step is to correctly place it in the new unit cell. When the model is translated in the new unit cell, symmetry related molecules move accordingly and, when all the models are in their correct position, the Patterson they generate will superimpose with those of the experimental Patterson function. So in principle, finding the translation vectors is rather similar to rotating appropriately the model but the vectors used are different, in the case of the rotation it is the self Patterson vectors while for the translation it is the cross-Patterson vectors that have to be used. This principle has been implemented in the definition of several translation functions which are described in detail in the references given at the end of this chapter. Here we will briefly consider one of them. The Patterson function due to the cross vectors of molecules 1 & 2 can be calculated as

$$P_{12}(u) = \int_v \rho_1(x)\, \rho_2(x+u)\, dx$$

if molecule 1 is translated in the unit cell, molecule 2 will also move to satisfy the symmetry operations of the space group and the function P_{12} will change its value. If P_{obs} is the experimental Patterson function, i.e the function calculated with the experimental intensities, then a translation function T (t) can be defined as follows:

$$T(t) = \int_v P_{obs}(u)\, P_{12}(u, t)\, du$$

this function will have a maximum when the two Pattersons superimpose.
The first step after a solution of the rotation and translation problem has been found is to check the position of the model in the new unit cell, the different molecules should obviously not clash and should be at reasonable distances from one another. If a convincing solution has been found then one can proceed with rigid body refinement (see below), first with the entire molecule and, in a second stage, with the elements of secondary structure. Eventually the phases calculated with the model have to be used along with the observed structure factor amplitudes to calculate electron density maps to be used for fitting the correct side chains and then proceed with the refinement of the structure.

7. Interpretation of the electron-density maps and model building

If a phase for the structure factors of the protein is available, then an electron density map of the crystal asymmetric unit can be calculated using the equation

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

note that the number of structure factors that have to be measured and phased can be large, in general several thousand, and their number is related to the resolution of the electron density map.
In a standard plate or frame (electronic plate) the reflections corresponding to the low resolution information are those close to the centre and we can imagine in the plate circles of different radii that correspond to different resolutions that can be easily calculated using Bragg's law. The larger the radius the higher the resolution. And so in general we speak of 5.0, 3.5 o 2.0 Å resolution maps. Electron density maps calculated at different resolutions contain different amounts of information and there is a minimum threshold below which tracing a new chain becomes very difficult if not impossible. This limit is usually placed at around 3.0 – 3.5 Å. If the phases are reasonable and the

resolution is high enough tracing the chain is, in general not difficult. Having information on the amino acid sequence of the protein is very important although not essential if the resolution is say higher that 2.0 Å. Some amino acids can be very easily distinguishable whereas others are not, for example the electron density of the side chain of a threonine and a valine look pretty much the same and in any attempt to produce a "crystallographic sequence" the positions where those amino acids are located will be ambiguous. In general though sequence information tends to be readily available because many proteins are cloned and recombinantly expressed ad if the clone is available so is the sequence. If the resolution of the maps is very high tracing the chain can be a very easy task that can be done automatically by the available model building software.
In every case the model which is fit to the density is a virtual model produced by the existing software, that after three decades of evolution, has become very flexible and sophisticated. There are several valid options of freely available model building software and most laboratories have more that one choice so that the operator can select the program he/she likes best. The relevant references are given at the end of the chapter. The task of model building consists of fitting a model to the calculated density, the first can be manipulated while the second cannot. All the graphic programs have dictionaries with the low energy conformers of the side chains of all the amino acids that the operator can choose to build the model. The time taken by the entire process is very strongly dependent on the quality of the maps and the experience of the person carrying out the task.

8. Structure refinement

If a model of the macromolecule has been built then we have the coordinates of all the atoms of the protein present in the asymmetric unit of the crystal and therefore we can calculate any structure factor using a modified version of the equation we have seen before,

$$F_{hkl} = \sum_i f_i \, e^{2\pi i(hx_i+ky_i+lz_i)}$$

$$F_{hkl} = K \sum_i f_i \, e^{-(B_i \, Sen^2\,\theta)/\lambda^2} \, e^{2\pi i(hx_i+ky_i+lz_i)}$$

Where K is a proportionality constant and the atomic scattering factor f_i has been corrected for the thermal motions that are normally present in every crystal, B_i is called the thermal parameter of atom i, θ is ½ the scattering angle of the reflection and λ is the wavelength of the X-rays.
The thermal parameter is adjusted together with the coordinates and it is related to the amplitude of the atomic vibrations about the equilibrium position according to the equation

$$B = 8 \pi \langle U \rangle^2$$

Where $\langle U \rangle^2$ is the mean square amplitude of the atomic vibrations.
If even a single atom is moved in the unit cell, all calculated structure factors vary. The process of refinement consists in moving slightly the atoms in the model so as to minimize the difference between the calculated and the observed structure factor amplitude.
The quantity that is calculated to follow the progress of refinement is called the R factor and it is defined as follows:

$$R = \frac{\sum \; | F_{obs}(h,k,l) - K F_{calc}(h,k,l)|}{\sum \; | F_{obs}(h,k,l) |}$$

where F_{obs} is the observed structure factor amplitude, F_{calc} is the calculated structure factor amplitude, K is a scale factor and the summations are performed over all the indexes h, k, and l.
The problem of finding the set of parameters of the model that give the best fit to the experimental data is normally approached in Crystallography using the least squares method and minimizing the quantity

$$S = \sum_{w(h,k,l)} [F_{obs}(h,k,l) - F_{calc}(h,k,l)]^2$$

where w (hkl) is used to weight the differences using as a criterion for example the precision of the measurement, the summation is extended over all the reflections that have been measured and each F_{calc} depends on all the parameter of the model
For an accurate definition of the parameters the method requires that the system be largely over-determined, i.e. the ratio of observations to parameters (coordinates and B factors) has to be of the order of 10 or higher. Whereas this ratio is, in general, easy to reach when dealing with small molecules it is very seldom attained when dealing with macromolecules.
Thus a fundamental question to be solved in protein crystal refinement is that normally the number of observations available, the structure factor amplitudes measured, is not adequate for an efficient use of the least squares method of crystallographic refinement. In dealing with this problem, there are two ways to improve the ratio of observables to parameters i.e. either increase the observables or reduce the number of parameters to be determined. In the second case what is done is to move large portions of the molecule, even the entire molecule and not individual atoms. In this way the number of coordinates is substantially reduced. The method is called constrained or rigid body refinement and is widely used in protein Crystallography. For example when a structure is solved by molecular replacement the first step is normally to perform rigid body refinement of the entire molecule to make sure that the position of the molecule in the unit cell is optimized. Rigid body refinement of elements of secondary structure and side chains have also been found to be very useful.
The alternative to reducing the parameters is to increase the observations that is to include observations from other sources different from the structure factor amplitudes, typically geometric restraints based on the known distances and valence angles of the amino acids, that are very well determined and are not expected to vary beyond certain values in the new model being refined.
Thus to the equation with the summation of the differences between F_{obs} and F_{calc} the following summation can be added:

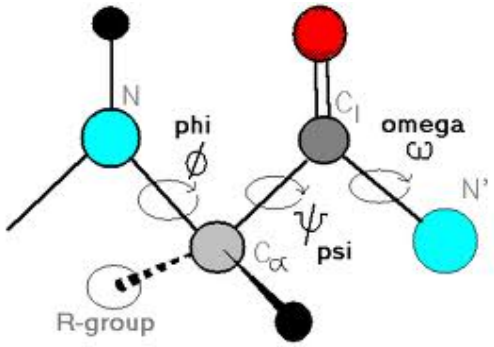
$$S' = \sum w_j [d_{j\ ideal} - d_{j\ calc}]^2$$

where $d_{j\ ideal}$ is an ideal value for a specific distance d_j and $d_{j\ calc}$ is the value calculated from the model and the weighting w_j factor is included to take care of the standard deviation of the distribution of the distances d_j.
Similar equations can be written for other geometric restraints, torsion angles, deviations of the peptide bond from planarity, the volume of chiral groups, etc.
When equations such as the example given before are used in the least squares refinement the refinement of the structure is said to be restrained.
Macromolecular refinement can also be done using molecular dynamics and in this case what is restrained is the energy of the system and the simulation usually spans a very short period of time. The simulation starts with a set of coordinates and produces a family of conformations with the constraints imposed by the X-ray data which reduce this number to those that correspond to lower values for the R factor.

9. Model validation.

In addition to the conventional R factor defined above, a much more reliable indicator of the quality of the model is the Rfree, defined as the R factor but applied to a subset of reflections that are not used in the refinement of the model. The subset is a limited portion of the data, usually between 5 and 10% of the total reflections, chosen at random before the refinement begins and used only to calculate the Rfree. Its value is higher than the conventional R factor calculated with the data used for refinement and it is a much better indicator of the absence (or presence) of any major errors in the model. In addition the deviations from ideality of the bond lengths, torsion angles, planarity, etc are important indicators of the model quality and there is an overall consensus of the values that can be accepted for all these parameters.

Another very useful indicator of model quality is the Ramachandran plot, which is a plot of the ψ vs Φ angles, the polypeptide chain backbone angles for every residue in the protein model. The two angles are defined in the following figure



and every amino acid in the sequence gives one point in the plot. Steric hindrance does not allow every possible combination of the ψ Φ values and some of them are more favourable than others. Thus the space in the Ramachandran plot can be divided into different areas: most favoured combinations, additionally allowed, generously allowed and forbidden regions in the plot. Prolines and glycines can occupy special areas in the plot, the first because the possible angles are more strictly restricted and the second because, they are less bulkier than the average amino acid and therefore can have ψ Φ combinations that are not allowed fro other amino acids. Outliers and residues with a very unusual combination should be very carefully checked and eventually corrected.

Another control that should be made after refinement is to check the values of the B factors of the different atoms of the model. The atoms with high B values are those that on the average move more about their equilibrium position, and we expect them to be those of the amino acids at the N and C terminus of the chain and those of the side chains exposed to the solvent. We also expect those portions of the chain that we know are more flexible to have higher B values. If this is not the case we should have an explanation, maybe the presence of intermolecular contacts in the crystal that reduce the mobility. Also unusually high B values may be an indication of disorder in the crystals or incorrectly built portions of the model.

At the end of the refinement it is a standard practice to include solvent molecules in the model; the result is that the R factor is invariably lowered although sometimes the same is not true for the Rfree. The number of solvent molecules in the model should not be exaggerated and their position should be reasonable in the sense that their distance from the protein should be such that a particular contact with the protein atoms should explain the presence of the solvent molecule there.

The validation process uses software that is freely accessible and there are servers that will do it automatically, a step that is also done at the time that coordinates and structure factors are deposited in the Protein Data Bank.

10. The difference Fourier synthesis and the study of protein function

Once the structure of a protein has been solved and refined, it becomes possible to undertake functional studies based on the existence of the new model. In particular studies of the association of proteins with small molecules, substrates, products or transition state analogues in the case of enzymes or ligands if dealing with transport proteins become feasible and relatively easy to carry out. Such studies are very informative and give much useful information on the function of the protein. In particular if the crystals of the unliganded and liganded protein are isomorphous the difference Fourier technique can be applied. The co-crystals, i.e. those containing the ligand can be prepared by soaking the apoprotein in mother liquor containing the small molecule of interest or alternatively by co-crystallizing the protein in the presence of the small molecule.

The addition of the small molecule to the crystals of the macromolecule changes the electron density so that

$$\rho_{PL} = \rho_P + \rho_L$$

where ρ_{PL} is the electron density of the macromolecule-small molecule complex, ρ_P is the electron density of the protein and the electron density of the ligand.

Similarly to what we saw for isomorphous heavy atom derivatives if we we Fourier transform this equation the result is that

$$\mathbf{F_{PL}} = \mathbf{F_P} + \mathbf{F_L}$$

what we would like to have is the electron density of the ligand in the unit cell of the macromolecule so that a model of the small molecule can be built and the interactions with the macromolecule defined.

The fundamental assumption in these studies is that F_L is relatively small so that, the phase of the native protein does not change very much in the complex and the phase of F_{PL} can be approximated by that of F_P and in the difference Fourier equation

$$\Delta\rho = 1/V \sum_h \sum_k \sum_l [\,|F_{PL}| - |F_P|\,]\, e^{i\phi_p}\, e^{2\pi i(hx+ky+lz)}$$

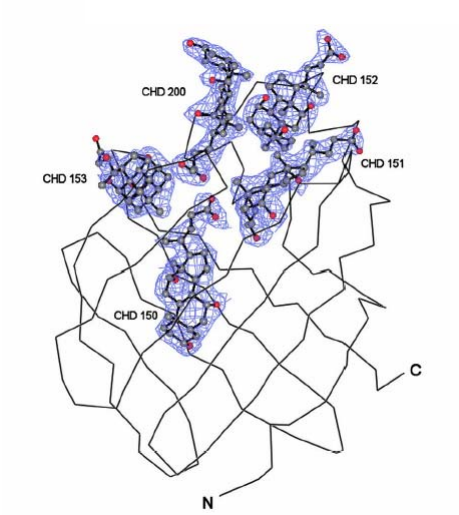
ϕp, the phase of the apoprotein can be used to calculate the difference electron density corresponding to the presence of the ligand in the co-crystals.

The extra electron density present in the crystals of the complex is then interpreted in terms of the ligand associated with the macromolecule and, since a model of the apoprotein in the same crystals is available. The side chains of the amino acids interacting with the ligand can be identified.

In some cases the presence of the ligand produces very drastic conformational changes in the macromolecule or affect the intermolecular contacts which can lead to the crystals cracking or eventually dissolving. If that is the case the only way out is to attempt to prepare the complex in solution and then crystallize it. Then chances are that, if crystals can be grown, they will not be isomorphous and if so the new structure will have to be solved from scratch but most likely the phase problem will be solvable using molecular replacement.

This rather simple technique has produced an enormous amount of biochemical information, sometimes very difficult if not impossible to obtain with other methods. An example is the presence of additional hydrophobic-binding sites present on the surface of a small transport protein molecule

which were suspected to exist (in addition to the canonical sites that were better characterized) but were only proven to be there when the techniques described here were applied to co-crystals of the protein.



References:

Textbooks

[1] Rupp, B. (2010**) Biomolecular Crystallography. Principles, Practice and Application to Structural Biology**. Garland Science New York.

[2] Zanotti, G. (2011**) Protein Crystallography in Fundamentals of Crystallography, Edited by Carmelo Giacovazzo**, Third Edition Oxford University Press New York.

[3] Blow D. (2002) **Outline of Crystallography for Biologists**. Oxford University Press New York.

[4] Blundell, T.L. & Johnson, L.N. (1976) **Protein Crystallography** Academic Press London

[5] Drenth, J (2002) **Principles of Protein X-ray Crystallography**. Second Edition Springer Verlag.

Papers

[6] Collaborative Computational Project Number 4. 1994. Acta Cryst. D50:760-767.

[7] Weeks CM, Adams PD, Berendzen J, Brunger AT, Dodson EJ, Grosse-Kunstleve RW, Schneider TR, Sheldrick GM, Terwilliger TC, Turkenburg MG, Usón I. (2001) Automatic solution of heavy-atom substructures Methods Enzymol. 2003;374:37-83.

[8] Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M. and Paciorek, W. (2003) Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. Acta Crystallogr. Sect. D Biol. Crystallogr. 59, 2023–2030

[9] Sheldrick, G. M. (2008) A short history of SHELX. Acta Crystallogr. Sect. A Found. Crystallogr. 64, 112–122

[10] Adams, P. D., Afonine, P. V., Bunk´oczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L. W., Kapral, G. J., Grosse-Kunstleve, R. W. et al. (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr. Sect. D Biol. Crystallogr. 66, 213–221

[11] Hendrickson WA. (1991) Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. Science. 254(5028):51-8.

[12] M. Cianci, M, Helliwell, JR, Helliwell, M, Kaucic, V, Logar, NZ, Mali, G and Tusar, NN. (2005) Anomalous scattering in structural chemistry and biology. Crystallography Reviews 11 (4): 245-335

[13] Navaza J. 1994. AMoRe: An automated package for molecular replacement. Acta Crystallogr. A5:157–163.

[14] Vagin A, Teplyakov A. 2000. An approach to multi-copy search in molecular replacement. Acta Crystallogr D Biol Crystallogr. 56: 1622–1624.

[15] McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. (2007) Phaser crystallographic software. J Appl Crystallogr. 40 (Pt 4):658-674.

[16] McRee, D. E. (1999). XtalView/Xfit—a versatile program for manipulating atomic coordinates and electron density. J. Struct. Biol. 125, 156–165.

[17] Potterton E, McNicholas S, Krissinel E, Cowtan K, Noble M. 2002. The CCP4 molecular graphics project. Acta Crystallogr D Biol Crystallogr. 58:1955–1957.

[18] Emsley, P., Lohkamp, B., Scott, W. G. and Cowtan, K. (2010) Features and development of Coot. Acta Crystallogr. Sect. D Biol. Crystallogr. 66, 486–501

[19] Brünger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. (1998) Crystallography & NMR system: A new software suite for macromolecular structure determination Acta Crystallogr D Biol Crystallogr. 54(Pt 5):905-21.

[20] Murshudov GN, Vagin AA, Dodson EJ. 1997. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr. 53:240–255.

[21] Adams PD, Afonine PV, Grosse-Kunstleve RW, Read RJ, Richardson JS, Richardson DC, Terwilliger TC. (2009) Recent developments in phasing and structure refinement for macromolecular crystallography. Curr Opin Struct Biol. 19(5):566-72.
[22] Laskowski RA, MacArthur MW, Moss DS, Thornton JM. 1993. PROCHECK: A program to check the stereochemical quality of protein structures. J Appl Crystallogr. 26:283–291.
[23] Kleywegt GJ. (2000) Validation of protein crystal structures Acta Crystallogr D Biol Crystallogr. 56(Pt 3):249-65.
[24] Read RJ, Adams PD, Arendall WB 3rd, Brunger AT, Emsley P, Joosten RP, Kleywegt GJ, Krissinel EB, Lütteke T, Otwinowski Z, Perrakis A, Richardson JS, Sheffler WH, Smith JL, Tickle IJ, Vriend G, Zwart PH. (2011) A new generation of crystallographic validation tools for the protein data bank Structure. 19(10):1395-412.