Phase Problem in X-ray Crystallography, and Its Solution

Secondary article

Kevin Cowtan, *University of York, UK*

X-ray crystallography can provide detailed information about the structure of biological molecules if the 'phase problem' can be solved for the molecule under study. The phase problem arises because it is only possible to measure the amplitude of diffraction spots: information on the phase of the diffracted radiation is missing. Techniques are available to reconstruct this information.

Article Contents

- The Phase Problem in X-ray Crystallography
- Patterson Methods
- Direct Methods
- Multiple Isomorphous Replacement
- Multiwavelength Anomalous Dispersion (MAD)
- Molecular Replacement
- Phase Improvement
- Summary

The Phase Problem in X-ray Crystallography

X-ray diffraction provides one of the most important tools for examining the three-dimensional (3D) structure of biological macromolecules. The physics of diffraction requires that in order to resolve features of atomic structure it is necessary to employ radiation with a wavelength of the order of atomic spacing or smaller. X-rays have suitable wavelength, and interact with the electrons of the atoms. However, the interaction between X-rays and electrons is weak, and such energetic radiation causes ionization of the constituent atoms of the molecule, damaging the molecule under study. Therefore it is necessary to examine a vast number of molecules simultaneously: This is achieved by using a crystal containing many copies of a molecule in a regular lattice.

When a crystal is exposed to X-rays, the radiation is scattered to form a diffraction pattern. The X-rays are scattered from every point in the crystal, with a strength in proportion to the concentration of electrons at that point. All the X-rays scattered along any particular direction interfere with each other, giving rise to detailed features in the diffraction pattern that depend on the arrangement of atoms in the crystal. Analysis of the diffraction pattern may therefore allow the arrangement of atoms to be deduced.

The intensity of the radiation scattered in any particular direction from the crystal depends on whether X-rays scattered along that direction interfere constructively or destructively. This in turn depends on the position and spacing of electron density features (in particular atoms) within the crystal. Examples of constructive interference, leading to strong scattering along a particular direction, and of destructive interference, leading to weak scattering, are shown in **Figure 1**.

The crystal, by its nature, contains a regular lattice of identical molecules, and thus every feature of the electron density will be repeated at regular intervals. The basic repeating unit from which the crystal is constructed is called the unit cell. It is convenient to define crystal axes a, b and c defining the unit cell in three dimensions. Scattering along directions reflecting the lattice repeat will be reinforced by every repeat of the unit cell, and will be strong; scattering along all other directions will be weak. As a result, the full diffraction pattern of the crystal is a pattern of spots, forming a three-dimensional lattice with

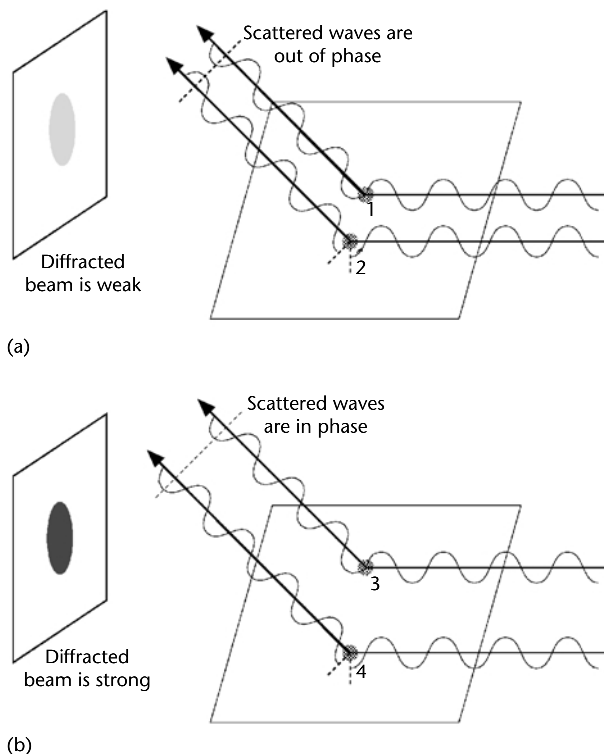

Figure 1 The diffraction pattern from a crystal depends on the arrangement of atoms in the crystal. In (a) the scattered waves from atoms 1 and 2 have opposite phases and cancel out, so the scattered beam is weak along that direction. In (b) the scattered waves from atoms 3 and 4 combine to give a strong scattered beam.

reciprocal directions and spacings from the real lattice. An individual diffraction image, obtained from a single crystal orientation, is a 2-D section through this pattern. This is referred to as the reciprocal lattice. The diffraction spots are labelled by three integer values called Miller indices referring to their position in the diffraction pattern. A simulated two dimensional crystal structure, and the corresponding diffraction pattern, are shown in **Figure 2**.

The diffraction spots are of different intensities, dependent on the spacings of features of the electron density within each individual unit cell. The density within the cell is related to the intensities of the diffraction spots by a mathematical relationship known as a Fourier transform (Bragg, 1915). As a result, if the contents of the unit cell are known, the diffraction pattern may be predicted. Conversely, if the scattering from the crystal is known, it is possible to reconstruct the contents of the unit cell.

It is only necessary to determine the electron density in a single unit cell rather than the whole crystal; or equivalently, through the Fourier relationship, it is only necessary to consider the diffraction spots, and not the space in between. Each spot may then be represented by a single wave representing the magnitude and relative phase (**Figure 3**) of the X-rays scattered along that direction. The mathematical representation of this wave is called a structure factor.

Unfortunately, it is only possible to measure the amplitude of the diffraction pattern spots by experimental means; the phase information is missing: this is the ‘phase problem’ of X-ray crystallography. Without phase information it is impossible to reconstruct the electron density in the unit cell.

The diffraction pattern has some other important features. If there is symmetry within the unit cell, the diffraction pattern will show a related symmetry, with the addition that in the absence of anomalous scattering (see later) Friedel opposites (reflections related by inversion through the origin, e.g. (1,3,2) and $(-1,-3,-2)$) are always equal in magnitude and opposite in phase. Reflections that are scattered through higher angles represent finer features in the electron density. The smallest feature that can be resolved in the electron density (if phases are available) depends on the Bragg spacing of the highest-angle reflections that can be measured, and is called the ‘resolution’ of the data set.

Patterson Methods

The structure of small molecules may be solved even in the absence of phase information. While the phases determine the positions of the peaks of electron density across the unit cell, and therefore the positions of the atoms, the presence of a strong diffraction spot alone give a strong indication that features must be present with the corresponding spacing. Thus, the structure factor magnitudes alone contain information about the spacing of atoms in the structure.

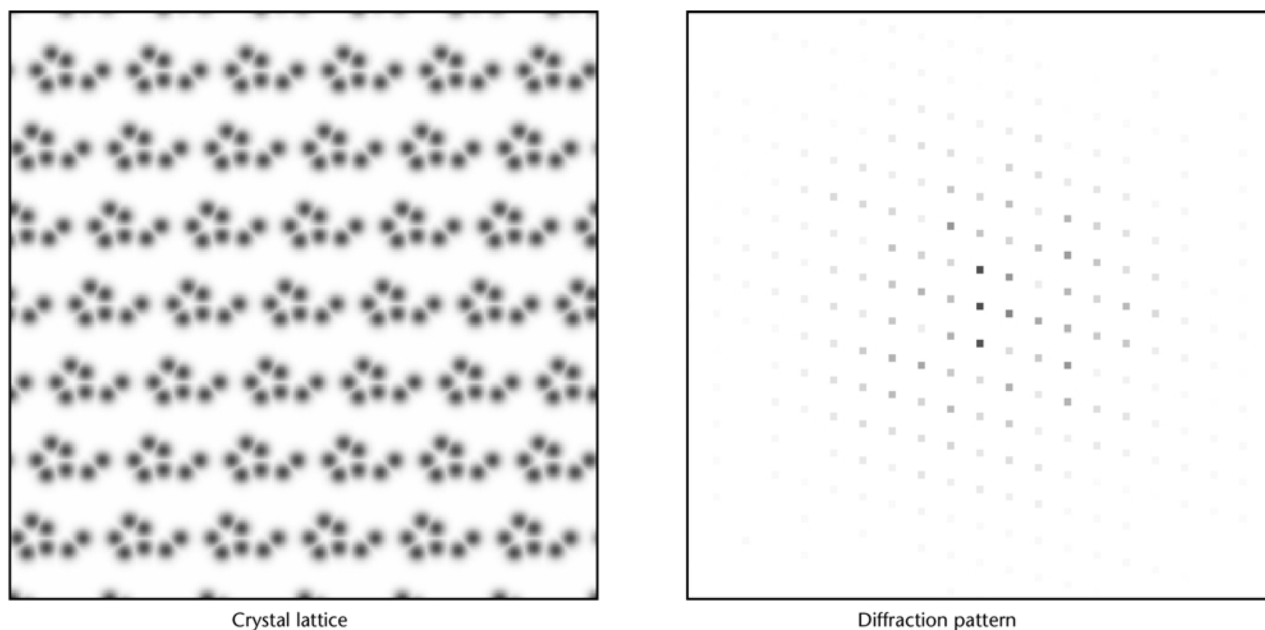

Figure 2 A simulation of a two-dimensional crystal and its diffraction pattern. Note that the lattice repeat in the crystal gives rise to a pattern of spots in the diffraction pattern, and that the lattice directions in the crystal are orthogonal to the lattice directions in the diffraction pattern.

 ENCYCLOPEDIA OF LIFE SCIENCES / © 2001 Macmillan Publishers Ltd, Nature Publishing Group / www.els.net

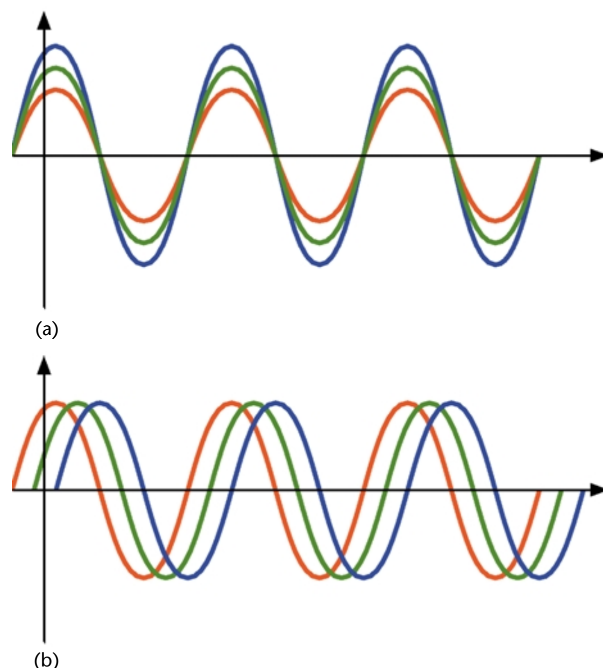

Figure 3 Each reflection in the diffraction pattern can be described as a wave with a certain magnitude and phase. The magnitude determines the size of the wave (a), and the phase determines where the peaks occur (b).

This information may be accessed by calculation of the Patterson function (Patterson, 1934). The Patterson function is obtained by calculating a map using the squared structure factor magnitudes, and all the phases set to zero. Instead of peaks at the atomic positions, the Patterson map shows peaks at every position that corresponds to an interatomic vector in the structure.

The Patterson function has been an effective tool for solving small molecules; however, its usefulness falls quickly as the number of atoms increases. For a structure of N atoms, the Patterson function will contain $N(N-1)$ interatomic vectors, many of them overlapped. This approach becomes unusable for structures of more than 20–50 atoms, unless there is a subset of atoms with high atomic number.

Direct Methods

For small and intermediate-sized molecules the atoms are normally well ordered, and as a result structure factors may be measured to very high diffraction angles. The high-angle diffraction spots give information about finely spaced features in the unit cell. In this case the missing phase information may be reconstructed directly from mathematical relationships between the structure factors. Since the phases come directly from the observed diffraction pattern, these methods are referred to as 'direct methods'.

The phase relationships on which direct methods rely depend on placing constraints on the electron density in the unit cell, for example that it is everywhere positive, or that it is clumped into distinct atomic peaks. If one structure factor is known in both magnitude and phase, then it can be inferred that atoms are more likely to be located in some regions of the cell and less likely to be in other regions. This places restrictions on the possible phases of other structure factors, which must reinforce likely areas in order to produce sharp peaks at atomic positions. The strongest relationship of this form is the three phase invariant: in this case, the constraints of positivity and atomicity imply that when three reflections whose Miller indices sum to zero are strong, the phases of those reflections must sum to a value near zero (Cochran, 1952).

A direct methods calculation might then proceed as follows. Phases are chosen for a few strong reflections, then phases for other reflections are generated using phase relationships among strong reflections. Once enough phases have been calculated, the electron density may be calculated and can be interpreted in terms of atomic positions.

Unfortunately, the phase relationships become weaker as the number of atoms in the structure increases; furthermore this approach only works when data can be collected to high resolution. The use of multisolution methods, by which a larger set of starting phases is chosen at random, and the calculation is repeated many times until a reasonable structure is obtained, has allowed direct methods to solve structures of up to 2000 atoms. However, most proteins cannot be solved by this approach because structure factors cannot be measured to atomic resolution.

Multiple Isomorphous Replacement

Larger structures such as proteins are typically less rigid than smaller molecules, and thus even in a crystal there will be significant disorder between unit cells. This limits the resolution of the diffraction pattern to typically 2 or 3 Å. Since these molecules may also contain many thousands of atoms, a different approach to phasing must be adopted. The most common approach to this problem has been multiple isomorphous replacement (MIR) (Green *et al.*, 1954).

Information about the unknown phases may be obtained by making a known change to the contents of the unit cell and measuring the effect on the diffraction pattern. In practice, this involves introducing a reactive group containing a heavy metal ion to the protein without disturbing the structure of the protein: that is, the two structures must be 'isomorphous'. Since the heavy atom scatters more strongly than the rest of the atoms in the

structure, it is usually possible to locate the heavy atoms alone by use of direct methods or Patterson methods. Once the location of the heavy atoms is known, the scattering from those atoms may be calculated both in magnitude and in phase.

Diffraction patterns are measured from both the native and the heavy-atom crystals, and the structure factors for the two crystals are compared. If a structure factor from the heavy-atom crystal is significantly stronger than the corresponding structure factor from the native crystal, then the scattering from the heavy atoms must be interfering constructively with the scattering from the rest of the crystal (**Figure 4a**). In this case the (unknown) phase of the native structure factor must be close to the (known) phase of the scattering from the heavy atom alone.

Similarly, if a structure factor from the heavy-atom crystal is significantly weaker than the corresponding structure factor from the native crystal, then the scattering from the heavy atoms must be interfering destructively and the phases must be offset by nearly 180° (**Figure 4b**). In most cases, however, the scattering will lie between these extremes. The magnitude of the combined scattering gives an indication of the size of the difference between the native and heavy-atom phases, but it does not tell us whether the native phase leads or lags the heavy-atom phase. To resolve this ambiguity and reduce noise introduced owing to lack of isomorphism between the crystals and to experimental error, it is usually necessary to use multiple crystals with different heavy atoms bonded to different sites on the protein.

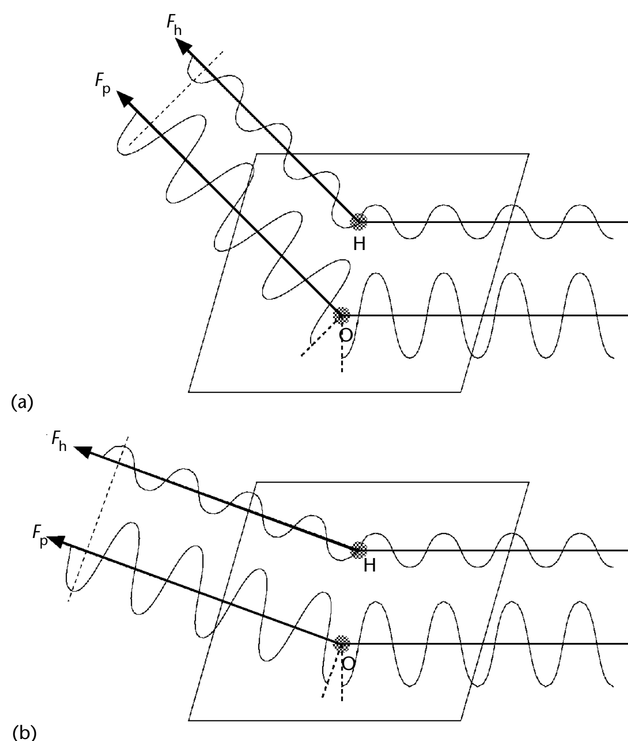

Figure 4 MIR allows estimation of phases when heavy atoms are introduced in known positions. F_p represents the total scattering from the native crystal, shifted to the origin of the crystal coordinates. F_h represents the additional scattering from a heavy atom located at H. If the addition of the heavy atom leads to stronger scattering along some direction, the phase of F_p must match that of F_h (a). If the addition of the heavy atom leads to weaker scattering, the phase of F_p must be opposite to that of F_h (b).

Multiwavelength Anomalous Dispersion (MAD)

Multiwavelength anomalous dispersion has become a popular alternative to multiple isomorphous replacement experiments following the introduction of tunable X-ray beam-lines at synchrotrons (Hendrickson and Ogata, 1997). The scattering from an atom is usually largely independent of wavelength; however, each atomic type has a few 'absorption edges' around which the scattering varies rapidly (in amplitude and phase) with wavelength. By varying the wavelength around the absorption edge for an atomic type, the contribution from those atoms to the total scattering can be varied. If the positions of the anomalously scattering atoms are known (for example, in the case of a few heavy atoms in the structure), phase information may be recovered by the same method as for MIR data.

Anomalous scattering has a further important property: near an absorption edge the scattering from an atom is shifted in phase. Normally the Friedel opposite of a reflection has the same magnitude with the negative of the phase; however, when there is anomalous scattering both phases are shifted in the same direction. Once the scattering from the rest of the nonanomalous atoms is included, the magnitudes of the reflection and its Friedel opposite are no longer equal.

Measurement of the difference between Friedel-opposite reflections provides additional phasing information; thus in theory it is possible to obtain an unambiguous phase estimate for a structure factor using only a single crystal, by measuring the scattering at the absorption edge and at a wavelength distant from the absorption edge. In practice, several wavelengths are used to provide better phase estimates.

This approach has several advantages over MIR. Since only a single crystal is required, the problems of growing multiple crystals with bound reactive compounds is avoided, as is the problem of nonisomorphism between crystals. However, the experiment requires very careful measurement of small differences in the diffraction pattern, and therefore must be performed to very high precision. Furthermore, locating all the atoms of a specific type in a structure requires that the structure contain a small

 ENCYCLOPEDIA OF LIFE SCIENCES / © 2001 Macmillan Publishers Ltd, Nature Publishing Group / www.els.net

number of heavier atoms that can be located by Patterson or direct methods.

MAD experiments are most commonly performed using protein in which all methionine residues have been modified to contain a selenium instead of a sulfur atom. Selenium is a moderately effective anomalous scatterer but has chemical properties very similar to sulfur.

Molecular Replacement

An alternative approach to the phase problem may be used when the molecule under study is similar to another molecule whose structure is already known. In this case the molecular replacement method (Rossmann, 1972) allows phases to be obtained from the known structure.

The molecular replacement calculation involves the solution of the rotation and translation functions: the known molecule is first rotated in three dimensions, and for each orientation structure factors are calculated from the model. The agreement between the calculated structure factors and the observed values from the diffraction experiment is used to identify the orientation of the known molecule that most closely matches that of the unknown molecule in the crystal.

Next, the oriented model is placed at every possible position in the unit cell, and again the agreement of the structure factors used to identify the correct translation. If the correct orientation and translation can be identified, then the model may finally be used to calculate phases for all the structure factors. Electron density maps may then be calculated using phases from the model structure and weighted magnitudes from the unknown structure. The resulting map may be examined to determine the unknown structure.

Phase Improvement

Once phase estimates (or phase probability distributions) are available from some experimental source, an initial electron density map may be calculated. At this point is is possible to apply chemical knowledge to improve the electron density map, and therefore the phases.

Since protein molecules are irregular in shape, they pack imperfectly in a crystal. The remaining voids are filled with disordered solvent. The electron density in these regions does not show identifiable features, so the electron density map may be improved by flattening the solvent (Wang, 1995). Conversely, sharpening the features inside the protein region adds further improvement.

Proteins frequently crystallize with more than one copy of the molecule in the unit cell, or asymmetric unit in the case of crystals with internal symmetry. Averaging the density between these copies also reduces the noise level in the map.

Once the map has been modified, it is used to calculate a new set of structure factors and phases. After combination with the observed magnitudes a new map may be calculated. The calculation may be repeated over several cycles.

Summary

X-ray crystallography provides a practical tool for imaging the atomic structure of biological molecules, provided the phases of the diffracted X-rays can be determined. For small molecules, and a few proteins that form particularly well-ordered crystals, Patterson and direct methods provide a mathematical solution to this problem. For larger biological molecules, MIR and MAD provide an effective experimental approach to phasing provided appropriate crystals can be grown. Once the structure of one molecule in a family is known, others may often be determined by molecular replacement using the known structure.

References

Bragg WH (1915) X-rays and crystal structure. *Philosophical Transactions of the Royal Society of London* **A215**: 253.

Cochran W (1952) *Acta Crystallographica* **5**: 65–67.

Green DW, Ingram VM and Perutz MF (1954) The structure determination of haemoglobin: sign determination by the isomorphous replacement method. *Proceedings of the Royal Society of London* **A225**: 287.

Hendrickson WA and Ogata CM (1997) Phase determination from multiwavelength anomalous diffraction measurement. *Methods in Enzymology* **276**: 494–523.

Patterson AL (1934) A Fourier series method for the determination of the components of interatomic distances in crystals. *Physical Review* **46**: 372.

Rossmann MG (1972) *The Molecular Replacement Method*. London: Gordon and Breach.

Wang BC (1985) Resolution of phase ambiguity in macromolecular crystallography. *Methods in Enzymology* **115**: 90–112.

Further Reading

Drenth J (1994) *Principles of Protein X-ray Crystallography*. New York: Springer-Verlag.

Blundell TL and Johnson LN (1976) *Protein Crystallography*. New York: Academic Press.

Glusker JP and Trueblood KN (1985) *Crystal Structure Analysis: A Primer*. New York: Oxford University Press.

Rhodes G (1999) *Crystallography Made Crystal Clear: A Guide for Users of Macromolecular Models*. San Diego: Academic Press.