Introduction to single crystal X-ray analysis XIII. Phase determination in protein structure analysis

Akihito Yamano*

1. Introduction

The principle of single crystal X-ray structure analysis is the same for organic/inorganic materials and proteins. However, although the steps of structural analysis are the same, there are major differences in the method of executing each step between structural analysis of small molecules and proteins. One of the steps where there is a major difference is the phase determination method for solving the phase problem, regarded as the central problem of single crystal X-ray structure analysis.

The main method of phase determination in small molecule structure analysis is the direct method of inferring phase via statistical processing of diffraction intensity. In protein structure analysis, on the other hand, phase determination using the direct method is impractical, and thus phase is determined experimentally. In the direct method, a roughly assigned initial phase is improved by using a phase relation which takes the magnitude of diffraction intensity as a clue. Therefore, in protein crystals, which have a comparatively large lattice, and limited atomic species and deviation of the electron distribution in the crystal, the magnitude of diffraction intensity is small compared with a small molecule crystal, and thus it is difficult in principle to apply the direct method.

As an experimental phase determination method for protein structure analysis, the MIR (Multiple Isomorphous Replacement) method serves as the classical approach. The MIR method derives the phase of the target (native) protein by using the slight shifts in phase which occur when heavy atoms are incorporated. This paper explains the MIR method, and the MAD (Multiple Anomalous Dispersion) method which uses the wavelength dependence of anomalous dispersion.

2. Preparation of heavy atom derivative

2.1. Soaking

The first stage of experimental phase determination is preparation of heavy atom derivative. The typical method of preparing the heavy atom derivative is a technique called "soaking" where the native protein crystal is soaked in heavy atom solution. The soaking procedure itself is simple (Fig. 1(a)), but there are many parameters to be varied, such as the heavy atom type, concentration, and soaking time, therefore a considerable amount of labor is required. This is because there is a need to evaluate whether the heavy atoms have been incorporated or not by actually measuring data, and there are many cases where, to obtain one derivative, it is necessary to repeat steps from soaking to measurement/evaluation at least a few times to as many as a few tens of times.

As a modified method of preparing a heavy atom derivative, there is a technique using vaporized iodine[1]. In particular, the method using solid iodine is extremely easy. The process involves simply cutting a piece of iodine to the appropriate size under a microscope, affixing it to the side of a crystallization drop with grease or a similar substance, and then resealing (Fig. 1(b)). It is also possible to use an iodine solution or a mixed solution of iodine and potassium iodide.

The method of preparing a heavy atom derivative using vaporized iodine has the following advantages not available with other methods:

- The crystal yellows as incorporation proceeds, so it is easy to confirm progress

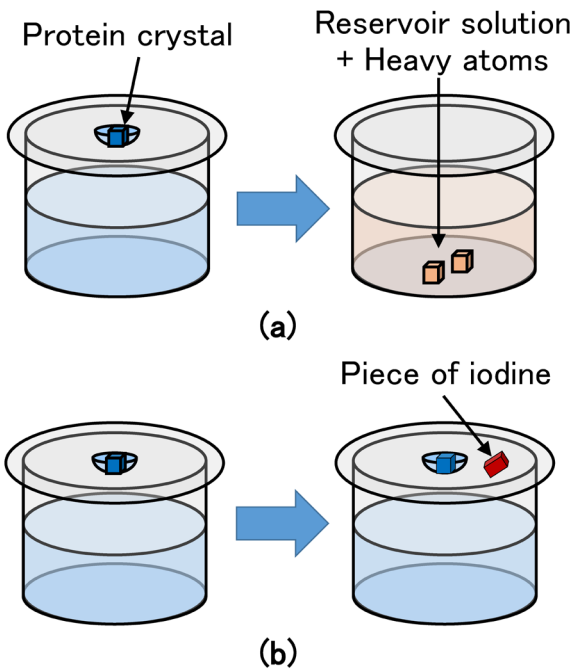


Fig. 1. Schematic diagram of procedure for preparing heavy atom derivative. (a) Method of preparing ordinary heavy atom derivative. (b) Method of preparing heavy atom derivative using vaporized iodine. A piece of iodine is attached to cover glass with grease etc., and sealed. Instead of the piece of iodine, it is also possible to use an iodine solution, or an iodine + KI solution.

* Application Laboratories, Rigaku Corporation.

- Soaking can be stopped and resumed at any time
- Only tyrosine side chains exposed to the solvent are iodized
- Isomorphism is high
- When determining phase, it is possible to use the high electron density and large anomalous dispersion of iodine

The likely reason for the good isomorphism is that the heavy atom solution is not added directly to the crystallization drop, and thus there is little change in the environment around the crystal, such as osmotic pressure, and only the ortho positions of the tyrosine residue exposed to the solvent are iodized, and as a result there is little effect on the direct and indirect intermolecular contact attributable to the change in structure of the protein molecule itself. Since the location where iodine is present is tyrosine, this also serves as clue for model construction. The drawback of the method using vaporized iodine is that it cannot be used in cases where PEG is the precipitant, or in cases where a tyrosyl residue exposed to the solvent is not present. Vaporized iodine absorbed into a crystallization drop containing PEG ends up crystallizing in the drop.

In phase determination using the MAD method, selenomethionine is normally incorporated using genetic engineering techniques, and thus heavy atom soaking is unnecessary. Although this also depends on the quality of data collection, it has been reported that phase can be determined if about one atom of selenium is present for every 150 amino acid residues[(2)].

2.2. Determination of possibility of incorporating heavy atoms via measurement of X-ray fluorescence

In Section 2.1, it was noted, as one of the difficulties of preparing a heavy atom derivative, that the possibility of incorporation must be evaluated by actually measuring data. One way of avoiding this is the method of measuring X-ray fluorescence (XRF). An attachment for measurement of X-ray fluorescence for single crystals (Fig. 2(a)) can be mounted onto an existing single crystal X-ray diffractometer, and XRF can be measured while carrying out data collection. To eliminate the possibility of XRF signal contamination from the heavy atom solution part in the case of a crystal mount, it is necessary to perform back-soaking, in which the heavy atom derivative crystal is soaked beforehand in a solution that does not contain heavy atoms. Derivatives prepared using the vaporized iodine above are also advantageous regarding this point. This is because the incorporated iodine has covalent bonds to a specific site in the tyrosyl residue side chain, and thus it is nearly impossible that heavy atoms are removed during back-soaking, a frequent problem with derivatives obtained using the ordinary soaking method.

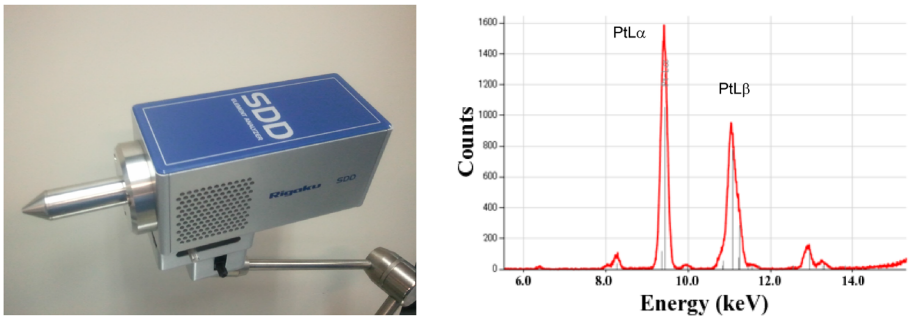
If XRF is obtained corresponding to the incorporated atomic species, then heavy atoms exist in the crystal (Fig. 2(b)). In heavy atom derivatives that contribute to phase determination, it is not enough for heavy atoms to simply be present; they must be bonded regularly to specific sites across the entire crystal. The occupancy is also crucial. However, if no XRF signal is present, this means heavy atoms are not present in the crystal, and the crystal can be rejected without data collection. Using the XRF attachment enables significant labor-saving.

3. Principle of phase determination

3.1. MIR method

As the name indicates, in the MIR (Multiple Isomorphous Replacement) method a crystal containing no heavy atoms (known as a native crystal) is prepared, together with derivative crystals in which the native crystal has been soaked in solution containing heavy atoms so that the heavy atoms are incorporated, and the phase of the native crystal is derived from the shift in phase between the two.

Figure 3 shows the principle of phase determination using the MIR method. Structure factors can be regarded as waves, and therefore it is convenient to consider these on the complex plane. Information from the native crystal is indicated in blue. The diffraction intensity of the native crystal can be measured through data measurement, but phase cannot. This situation corresponds to the fact that the length of the arrow for the structure factor, in itself a vector, is known, but its endpoint position is unknown. Here is where the heavy



(a) (b)

Fig. 2. (a) XRF attachment for single crystal X-ray diffractometer. (b) Results of measurement with (a) after soaking native crystal for 10 min in 10 mM K_2PtCl_4 solution.

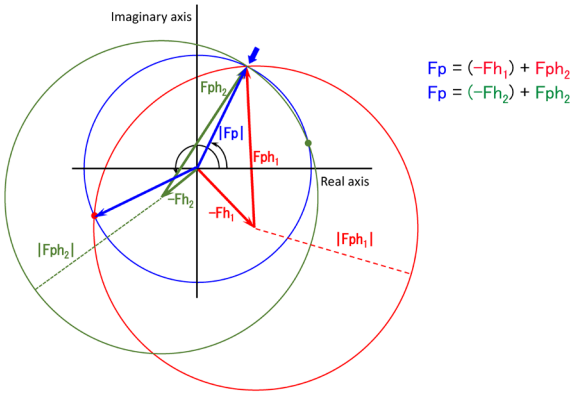


Fig. 3. Diagram showing principle of the MIR method (Argand diagram). The ideal phase relationship for a specific diffraction point is shown. F_p: Structure factor of native protein, F_{ph1}: Structure factor of 1st heavy atom derivative, F_{ph2}: Structure factor of 2nd heavy atom derivative, F_{h1}: Structure factor of only heavy atoms of 1st heavy atom derivative, F_{h2}: Structure factor of only heavy atoms of 2nd heavy atom derivative. The values $|F_p|$, $|F_{ph1}|$, and $|F_{ph2}|$ are, respectively, the magnitudes of the structure factors of the native protein, 1st heavy atom derivative, and 2nd heavy atom derivative, and these can be measured through diffraction intensity measurement.

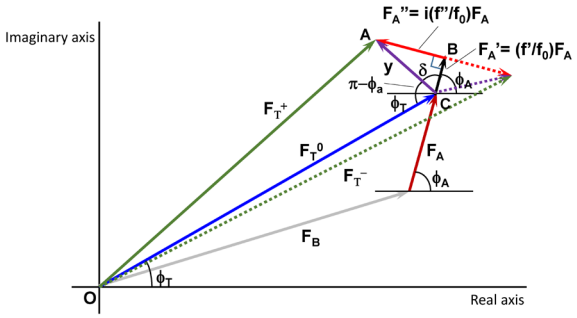


Fig. 4. Diagram showing principle of the MAD method (Argand diagram). The ideal phase relationship for a specific diffraction point is shown. F_B: Structure factor of all atoms without anomalous dispersion, F_A: Structure factor of parts that do not contribute to anomalous dispersion of all atoms with anomalous dispersion, F'_A: Among structure factors of parts that contribute to anomalous dispersion of all atoms with anomalous dispersion, structure factors of parts with the same phase as parts that do not contribute to anomalous dispersion, F''_A: Among structure factors of parts that contribute to anomalous dispersion of all atoms with anomalous dispersion, structure factors of parts having a 90 degree phase difference with parts that do not contribute to anomalous dispersion.

atom derivative comes into play.

If the position of the heavy atom is known, the structure factor of the heavy atom part can be calculated. The structure factor is expressed as a vector, and therefore it can be written in like Fh_1 in Fig. 3. However, for parts other than heavy atoms in the heavy atom derivative, the only thing known is the diffraction intensity, and thus a circle is drawn, whose radius is the magnitude of the structure factor, centered at the endpoint of the structure factor of the heavy atom (red circle in Fig. 3). At the intersection points of the red circle and blue circle, deriving from the derivative and native crystal, structure factor vector operations are satisfied only for the heavy atom derivative, native crystal, and heavy atom. However, there are two intersection points, and two possibilities for the phase angle.

Thus a second heavy atom derivative is prepared, and the same process is repeated. For the second derivative too, the position of the heavy atom is determined, and the structure factor of the heavy atom part is calculated (Fh_2 in Fig. 3). If a circle is drawn, centered at the endpoint of Fh_2 and with the magnitude of the heavy atom derivative as its radius, then in this case too there are two intersection points, but only one intersection point does not conflict with the two intersection points obtained with the first derivative (thick arrow in Fig. 3). This results in determination of a single possibility for the two phase angles, and thus the phase of the protein crystal is determined.

3.2. MAD method

The MAD method tends to be regarded as a newer technique than the MIR method, but the theory itself was developed at the same time as the MIR method[3), 4)]. The MAD method appeared later because measurement must be done at different wavelengths, and thus its practical application required the advent of synchrotron radiation beamlines for protein structure analysis. Whereas the MIR method uses the fact that the structure factor changes depending on differences in the atomic species, and differences in the bonding site and occupancy, the MAD method uses the fact that the structure factor changes because the anomalous dispersion differs for different wavelengths.

Figure 4 shows the phase relationship for a single diffraction point when there is a source of anomalous dispersion. The principle of phase determination in the MAD method[5)] will now be derived from this diagram. Assuming that all of the heavy atoms which serve as sources of anomalous dispersion are the same atom, and letting F_H be the structure factor for heavy atoms as a whole, F_A be the part that does not contribute to anomalous dispersion, F'_A be the wavelength dependent part that has the same phase as F_A and contributes to anomalous dispersion, and F''_A be the wavelength dependent part with a 90 degree phase difference from F'_A that contributes to anomalous dispersion, then:

$$F_H = F_A + F'_A + F''_A = F_A + \frac{f'}{f_o}F_A + \frac{if''}{f_o}F_A$$

$$\therefore \quad |F'_A| = \frac{f'}{f_0}|F_A|, \ |F''_A| = \frac{f''}{f_0}|F_A| \qquad (1)$$

Because,

$$
\begin{aligned}
F_H &= \sum_j f_0 \exp\{2\pi i (h_j \cdot r_j)\} + \sum_j f' \exp 2\pi i \{2\pi i (h_j \cdot r_j)\} \\
&\quad + \sum_j if'' \exp 2\pi i \{2\pi i (h_j \cdot r_j)\} \\
&= F_A + \frac{f'}{f_0} \sum_j f_0 \exp\{2\pi i (h_j \cdot r_j)\} \\
&\quad + i \frac{f''}{f_0} \sum_j f_0 \exp\{2\pi i (h_j \cdot r_j)\} \\
&= F_A + \frac{f'}{f_0} F_A + i \frac{f''}{f_0} F_A
\end{aligned}
$$

The △ABC is a right triangle, and thus if the Pythagorean theorem is applied, and Equation (1) is substituted in,

$$y^2 = |F_A'|^2 + |F_A''|^2 = \left(\frac{f'^2 + f''^2}{f_o^2} \right) |F_A|^2 \quad (2)$$

Also, from the △ABC and Equation (1),

$$y \cos\delta = |F_A'| = \frac{f'}{f_0} |F_A|, \quad y \sin\delta = |F_A''| = \frac{f''}{f_0} |F_A| \quad (3)$$

Here, if the law of cosines is applied to △OAB,

$$|F_T^+|^2 = |F_T^o|^2 + y^2 - 2|F_T^o| y \cos\{(\pi - \phi_a) + \phi_T\}$$

Since $(\pi - \phi_a) + \delta + \phi_A = \pi$, i.e., $\phi_a = \phi_A + \delta$, and $\cos(\pi - \alpha) = -\cos\alpha$ and $\cos(\alpha \mp \beta) = \cos\alpha \cos\beta \pm \sin\alpha \sin\beta$,

$$
\begin{aligned}
&= |F_T^o|^2 + y^2 + 2|F_T^o| y \cos\{(\phi_T - \phi_A) - \delta\} \\
&= |F_T^o|^2 + y^2 + 2|F_T^o| y \cos\delta \cos(\phi_T - \phi_A) \\
&\quad + 2|F_T^o| y \sin\delta \sin(\phi_T - \phi_A)
\end{aligned}
$$

Substituting in Equation (2) and Equation (3),

$$
\begin{aligned}
|F_T^+|^2 &= |F_T^o|^2 + ((f'^2 + f''^2)/(f_o^2)) |F_A|^2 \\
&\quad + 2(f'/f_0) |F_T^o| |F_A| \cos(\phi_T - \phi_A) \\
&\quad + 2(f''/f_0) |F_T^o| |F_A| \sin(\phi_T - \phi_A)
\end{aligned}
$$

If F_T^- is calculated in the same way, and it is assumed that $a(\lambda) = (f'^2 + f''^2)/(f_0^2))$, $b(\lambda) = 2(f'/f_0)$, and $c(\lambda) = 2(f''/f_0)$,

$$
\begin{aligned}
|F_T^+|^2 &= |F_T^o|^2 + a(\lambda) |F_A|^2 + b(\lambda) |F_T^o| |F_A| \cos(\phi_T - \phi_A) \\
&\quad \pm c(\lambda) |F_T^o| |F_A| \sin(\phi_T - \phi_A) \quad (4)
\end{aligned}
$$

In Equation (4), it is possible to find $a(\lambda)$, $b(\lambda)$, and $c(\lambda)$ for a specific wavelength from the absorption spectrum, and thus the number of unknowns dependent on the wavelength is three: F_T^o, F_A, and $(\phi_T - \phi_A)$. Therefore, if Friedel pairs can be measured for two or more different wavelengths, then four or more equations can be obtained for the three unknowns, and thus it is possible to determine F_T^o, F_A, and $(\phi_T - \phi_A)$. The largest change in diffraction intensity occurs near the absorption edge, and thus measurement is ordinarily done at three wavelengths: immediately before (Edge) and immediately after (Peak) the absorption edge (Edge), and distant (Remote) from the absorption edge.

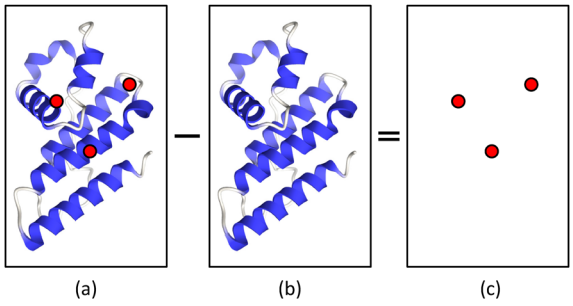


Fig. 5. (a) Structure for heavy atom derivative, (b) structure for native crystal, (c) structure for heavy atoms only.

4. Derivation of heavy atom position

To calculate the phase of the native protein using the MIR method, it is necessary to know the structure factor of the heavy atom. Also, the phase obtained with the MAD method is $\phi_T - \phi_A$, and thus ϕ_A must be subtracted in order to calculate ϕ_T. That is, whether the MIR method or MAD method is used, it is necessary to determine heavy atom parameters such as heavy atom position using some method. This corresponds to determining the substructure of the heavy atoms only.

To determine the positions of heavy atoms with the MIR method, the expression $|Fph_1| - |Fp|$ is used, in which the diffraction intensity of the native crystal is subtracted from the diffraction intensity of the heavy atom derivative crystal. This is because, if the isomorphism of the heavy atom derivative is sufficiently high, then $|Fph_1| - |Fp|$ only includes the contribution from the heavy atom part, and therefore the derived structure corresponds to the structure of the heavy atoms only. This is easy to understand if it is considered in actual space. For both the structure of the native protein and the structure of the heavy atom derivative, it is difficult to directly calculate the phase because the protein part is included. However, if the native protein structure is subtracted from the structure of the heavy atom derivative, the result is the structure of the heavy atoms only, and this becomes extremely simple (Fig. 5). If this is the case, it should be possible to somehow determine the structure.

Until around the 1990, determination of the heavy atom positions was done through hand calculation, by calculating the Patterson function, a superposition of vectors between heavy atoms. This worked well as long as the applicable structure was simple, but as the applicable protein molecules grow large in size, the number of sites where heavy atoms are bonded also increases, and thus the problem can no longer be handled with hand calculation. At present, the typical approach is to use the direct method, or software where the Patterson function is analyzed automatically, e.g., using SOLVE, SHELXD, SHARP, or CRANK[6].

5. Conclusion

With the MIR method, it is possible to determine phase if diffraction data can be measured for a native crystal and two types of different heavy atom derivative crystals. However, the preparation of a single heavy atom derivative that contributes to phase determination requires a search for conditions and measurement/evaluation multiple times, and thus is the stage involving the most labor.

This situation has not changed even today, when there has been significant progress in measurement equipment. However, at present, if one heavy atom derivative can be obtained, it is also possible to determine phase using the SIR (Single Isomorphous Replacement) method, the SIRAS method which combines the SIR method with anomalous dispersion, and the SAD (Single-wavelength Anomalous Dispersion) method. In particular, phase determination is possible with a native crystal only if the SAD method can be applied while using the S-S bonds in proteins or using anomalous dispersion of sulfur contained in side chains of methionine.

With the MAD method, multiple data collections at different wavelengths are required, but this method has the advantage that only one heavy atom derivative crystal is needed. Therefore, the MAD method is effective in cases where the MIR method cannot be used because acquiring a native crystal is difficult either in principle or technically, where the protein contains metals to begin with, and other similar situations. Also, the MAD method is not affected by the quality of the isomorphism because it does not use a native crystal, and thus the phase quality depends on the data quality alone. Generally speaking, measurement using synchrotron radiation and its high count rate enables measurement of data with high reliability, and thus phase determined with the MAD method has high quality, and tends to facilitate the model construction which is the next step of structural analysis.

References

(1) H. Miyatake, T. Hasegawa and A. Yamano: *Acta Cryst.*, **D62** (2006), 280−289.
(2) W. Hendrickson, J. Horton and D. LeMaster: *EMBO J.*, **9** (1990), 1665−1672.
(3) J. Karle: *Int. J. Quant. Chem. Symp.*, **7** (1980), 357−367.
(4) W. Hendrickson: *J. Synchrotron Rad.*, **6** (1999). 845−851.
(5) For example, J. Drenth: *Principles of Protein X-ray Crystallography*, Springer-Verlag, New York, (1994), 207−212.
(6) S. R. Ness, R. A. G. de Graaff, J. P. Abrahams and N. S. Pannu: *Structure*, **12** (2004), 1753−1761.