Acta Cryst. (1971). B**27**, 1474

Some aids for breaking the phase ambiguity in the single isomorphous replacement method. By WAYNE A. HENDRICKSON,* *Laboratory for the Structure of Matter, Naval Research Laboratory, Washington, D.C.* 20390, *U.S.A.*

(*Received* 2 *November* 1970)

Methods are discussed for using the unambiguous, but often imprecise, phase information from a partial structure or the tangent formula in choosing between the usually rather precise phase alternatives left by a single isomorphous replacement. Phase probability distributions are used in making the phase selections and in weighting the results for Fourier synthesis. The procedures described have been used in the structure analysis of lamprey hemoglobin.

The phase probability distribution arising from a single isomorphous replacement (SIR) is, in general, bimodal with the two peaks disposed symmetrically about the centroid phase angle. The ambiguity of phase in such a distribution can be reduced by multiplicatively combining it with the distributions from other isomorphous replacements, anomalous scattering or other types of phase information. However, certain auxiliary phase information may be more advantageously employed merely to choose between the maxima of the SIR distribution. This alternative offers a more conservative approach for cases where the reliability of the auxiliary information, relative to the physically based SIR information, is in question. Phase determining relationships, particularly the tangent formula, have been suggested as means for making the phase choice (Coulter, 1965; Karle, 1966) and tests on model problems and on a protein have been reported (Coulter, 1965; Weinzierl, Eisenberg & Dickerson, 1969). Phase information from a partial structure might also be used for this purpose.

In the course of using the tangent formula for SIR phase selection with lamprey hemoglobin, some aids for breaking the SIR phase ambiguity have been developed. It is the purpose of this note to describe a simplified form of the SIR distribution and its use in the phase selection process, to suggest a measure of the intrinsic reliability of an SIR phase alternative, and to indicate a probabilistic means for assessing the certainty with which a choice is made.

Probability distributions for phase information from isomorphous replacement, as well as from several other sources, can be expressed in the representation,

$$P(\alpha)=N\exp\,(A\cos\alpha+B\sin\alpha+C\cos 2\alpha+D\sin 2\alpha)\,. \quad (1)$$

This form can be achieved by using a reformulated error model for isomorphous replacement data [Hendrickson & Lattman, 1970 (HL)] or alternatively by means described by Hendrickson (1971). A complete description of the derivation and advantages of (1) is given by HL. In the case where only SIR phase information is included, equation (1) can be further simplified. This can be done by shifting the origin of the distribution to the centroid phase angle, α_c. In the SIR case the centroid phase equals either φ, the phase of the heavy-atom group or $\varphi+\pi$ depending as $F_P^2+f^2$ is either less or greater than F_H^2. Here F_P and F_H respectively denote the structure factor amplitudes of the protein and heavy-atom derivative, and the heavy-atom contribution to the structure factor is given by $f\exp\,(i\varphi)$. Evaluation of the effect of this shift of origin on the A, B, C and D coefficients given in equations (6) of HL produces the SIR distribution,

$$P(\alpha')=N\exp\,(S\cos\alpha'-T\cos 2\alpha') \quad (2)$$

where $S=|\sqrt{A^2+B^2}|$, $T=|\sqrt{C^2+D^2}|$ and $\alpha'=\alpha-\alpha_c$. This result can also be demonstrated by using constraining properties of SIR distributions, namely, that $P(\alpha_c+\theta)=P(\alpha_c-\theta)$ for any θ and that $P(\alpha_c+\theta)\geq P(\alpha_c+\pi-\theta)$ and $P'(\alpha_c+\theta)\geq P'(\alpha_c+\pi-\theta)$ for θ between 0 and $\pi/2$, to evaluate the parameters of the transformation of equation (19) of HL in terms of the centroid phase angle. A useful consequence of this analysis is the result

$$\tan\alpha_c=B/A\,. \quad (3)$$

Once the SIR distribution has been cast in the form of (2), it is easy to determine the positions of its extrema. By using the condition that $P'(\alpha)=0$ at an extremum and inspecting the curvature, $P''(\alpha)$, at the extreme points to decide between maxima and minima, the extrema can be simply described in terms of S and T (Table 1). The case where $S\geq 4T$ results when errors preclude the equality, $\mathbf{F}_P+\mathbf{f}=\mathbf{F}_H$. Inequalities (5) of Blow & Rossmann (1961) pertain to this situation and are equivalent to those listed in Table 1. The equation giving the half-angle separating maxima in the case where $S<4T$,

$$\cos\delta=S/4T\,, \quad (4)$$

corresponds exactly to that based on geometric arguments (Kartha, 1961; Blow & Rossmann, 1961) as can be seen by expansion of S and T using equations (6) of HL.

Of ultimate concern is the use of the phase information in a Fourier synthesis. Kartha (1961) and Blow & Rossmann (1961) have proposed that in the absence of other phase information the structure factor amplitudes should be entered with SIR centroid phases and weighted by $\cos\delta$. Where errors are taken into account (Blow & Crick, 1959), these weights are readily generalized to the figures-of-merit,

$$m=\frac{\int_0^\pi \cos\alpha'\,P(\alpha')\,\mathrm{d}\alpha'}{\int_0^\pi P(\alpha')\,\mathrm{d}\alpha'}\,. \quad (5)$$

Auxiliary phase information can be combined directly with the SIR information by using the product of the respective phase probability distributions in the phase determination. However, experience indicates that in the case of rather imprecise phase information, such as from the tangent formula applied to limited data, it may be more efficacious to use this information merely to choose between the alternative most probable SIR phases, $\alpha=\alpha_c\pm\delta$ (Weinzierl, Eisenberg & Dickerson, 1969). The selected phase can then be used in a Fourier synthesis and weighted according to its reliability. Whereas the weight given by (5) measures both the ambiguity in the SIR distribution and the precision of the separate maxima, having broken the ambiguity one desires a weight which reflects just the precision of the chosen phase alternative. A useful measure of this precision can be achieved by using an intrinsic figure-of-merit,

* A Resident Postdoctoral Research Associate of the National Research Council.

Table 1. *Extrema of* $P(\alpha')$

Condition	Physical situation $\alpha_c=\varphi$	$\alpha_c=\varphi+\pi$	Real extrema $\alpha'=0$	$\alpha'=\pi$	Other
$S\geq 4T$	$F_P+f\leq F_H$	$\lvert F_P-f\rvert\geq F_H$	max	min	
$S<4T$	$F_P+f>F_H$	$\lvert F_P-f\rvert<F_H$	min	min	max at $\alpha'=\pm\delta$, $\delta=\cos^{-1}(S/4T)$

$$m'=\frac{\int_\delta^\pi \cos(\alpha'-\delta)\,P(\alpha')\,d\alpha'}{\int_\delta^\pi P(\alpha')\,d\alpha'}\,, \tag{6}$$

derived from a distribution symmetrized about one of the SIR phase alternatives as shown in Fig. 1. The symmetrized distribution is, of course, only an approximation to the phase probability intrinsic to an SIR alternative. The observable region, δ to π, of the alternative at δ is contaminated by tails from the distribution centered at $-\delta$ and is truncated from the complete hemi-distribution which would go on to $\pi+\delta$. The effects of these factors on m' are opposite, contaminating tails tend to increase m' and truncation tends to decrease it. In any event these effects are scarcely deleterious since intrinsic figures-of-merit derived from 2Å lamprey hemoglobin data proved entirely reasonable. If desired, a circular normal distribution, $\exp[k\cos(\alpha'-\delta)]$, corresponding to the symmetrized distribution can be constructed from the intrinsic figure-of-merit by inverting the relationship, $m'=I_1(k)/I_0(k)$, where I_n is a Bessel function of imaginary argument. Although the effort hardly seems justified, such a circular distribution might be used to improve the symmetrized distribution by extending it to the unobserved region and removing tail contributions. However, in this connection it is well to note that the SIR distribution cannot be reconstructed from a sum of normal distributions centered at $\pm\delta$ since such a sum necessarily shifts the maxima.

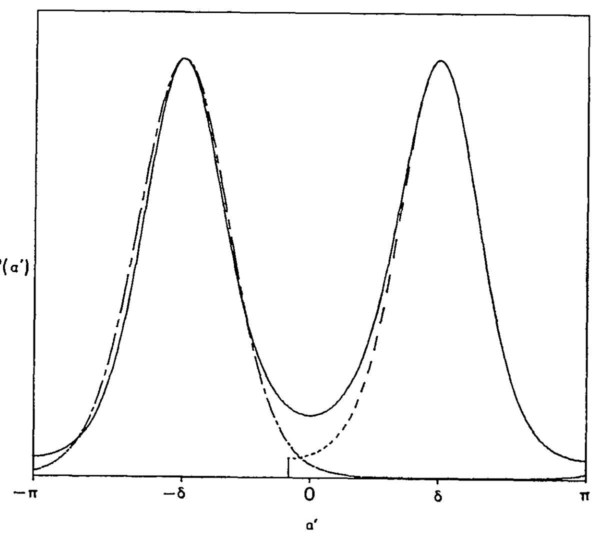

Fig. 1. Phase probability distributions used in weighting phase alternatives. An SIR distribution with $\delta=83{\cdot}2°$ and $m=0{\cdot}14$ is represented by ——, the symmetrized distribution corresponding to one alternative is represented by — — — and has $m'=0{\cdot}86$, and a circular normal distribution with $k=3{\cdot}9$ derived from the latter is centered at the other alternative and represented by — - — -. The Figure is designed to illustrate the symmetrizing operation and is not intended to portray a representative distribution. In many cases peaks are not so cleanly separated.

Choosing between SIR phase alternatives is not always a sure process. Sometimes the phase information used in selecting may be so unreliable as to make a choice unwarranted. Thus a means for judging the certainty of a choice is needed. When the selection is made with tangent formula or partial structural information, the associated probability function, $P_s(\alpha)=N_s\exp[K_s\cos(\alpha-\alpha_s)]$ where K_s is the coefficient of reliability and α_s is the phase of maximal probability (*cf.* equations (11) and (13) of HL), provides this means. The SIR phase angle alternative to which α_s is nearer is defined as α_1 and the more distant alternative is defined as α_2. Then the ratio of the probability, P_1, that α_1 be the alternative of choice to the probability, P_2, that α_2 be the choice is

$$R=\frac{P_1}{P_2}=\frac{P_s(\alpha_1)}{P_s(\alpha_2)}=\exp K_s[\cos(\alpha_1-\alpha_s)-\cos(\alpha_2-\alpha_s)]\,. \tag{7a}$$

Since the choice is restricted to either α_1 or α_2, $P_1+P_2=1$ and the probability that the phase be α_1 rather than α_2 is

$$P_1=\frac{R}{1+R}\,. \tag{7b}$$

Phases for cases where P_1 exceeds some cut-off, say 0·8, are accepted as α_1 and weighted by m', equation (6). Whereas when P_1 is less than the cut-off or $\delta=0(S\geq 4T)$, α_c is taken as the phase and weighted by m, equation (5). The choice of value for the P_1 cut-off must be somewhat arbitrary, but tests relating phase selections to other phases, such as from multiple isomorphous replacement, can aid the decision. An incorrect phase choice matters less when δ is small and one might consider making the cut-off a function of δ. Using the tangent formula, the above procedure has been successfully employed with lamprey hemoglobin in determining phases for data in the shell between 2·5 and 2·0 Å resolution.

Discussions with Drs Jerome Karle, Eaton Lattman and Alvin Siger contributed to this work and are greatly appreciated.

References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794.

Blow, D. M. & Rossmann, M. G. (1961). *Acta Cryst.* **14**, 1195.

Coulter, C. L. (1965). *J. Mol. Biol.* **12**, 292.

Hendrickson, W. A. (1971). *Acta Cryst.* B**27**, 1472.

Hendrickson, W. A. & Lattman, E. E. (1970). *Acta Cryst.* B**26**, 136.

Karle, J. (1966). *Acta Cryst.* **21**, 273.

Kartha, G. (1961). *Acta Cryst.* **14**, 680.

Weinzierl, J. E., Eisenberg, D. & Dickerson, R. E. (1969). *Acta Cryst.* B**25**, 380.