Acta Cryst. (1987). A43, 6–13

Isomorphous Replacement: Effects of Errors on the Phase Probability Distribution

By Thomas C. Terwilliger* and David Eisenberg

Molecular Biology Institute and Department of Chemistry and Biochemistry, University of California, Los Angeles, California 90024, USA

(*Received* 7 *December* 1985; *accepted* 16 *April* 1986)

Abstract

A simple framework for the analysis of the effects of errors in the method of isomorphous replacement is developed. This framework is used to derive phase probability distributions similar to those given by Blow & Crick [*Acta Cryst.* (1959). **12**, 794–802]. The present analysis suggests an improved means of calculating the mean-square 'lack-of-closure' residuals and shows that they differ by a factor of two for centric and acentric reflections. It is also shown that the lack-of-closure residuals have a straightforward interpretation and that they may be used to estimate the error in the heavy-atom model and the degree of isomorphism between native and derivative structures if anomalous differences have been measured.

Introduction

In the method of isomorphous replacement, information from several sources is generally combined by multiplying the various probability distributions for each phase. Consequently the accuracy of these probability distributions is of some importance and a detailed understanding of the effects of errors on these distributions would be valuable.

An expression often used to calculate a phase probability distribution $P(\varphi)$ for a reflection is (Blow & Crick, 1959)

$$P(\varphi)\propto \exp\left[-(F^o_{PH}-F^c_{PH})^2/2E^2\right] \qquad (1a)$$

where φ is the native phase and F^o_{PH} is the observed derivative structure-factor amplitude. F^c_{PH} is a calculated derivative structure-factor amplitude, given by

$$F^c_{PH}=|F^o_P\exp(i\varphi)+\mathbf{f}_\mathrm{H}|. \qquad (1b)$$

Here F^o_P is the observed native structure-factor amplitude and $\mathbf{f}_\mathrm{H}$ is an estimate of the structure factor of the heavy atoms present in the derivative structure but not in the native structure, calculated from a model.

Although it is clear that the choice of E^2 in $(1a)$ is crucial, the proper value of E^2 is not obvious. In practice (Ten Eyck & Arnone, 1976), the value of E^2 generally used is the mean-square value of the 'lack-of-closure' residual, $\langle(F^o_{PH}-F^c_{PH})^2\rangle$, where F^c_{PH} is the derivative structure-factor amplitude calculated using $(1b)$ at the 'best' native phase.

In this paper, we develop a simple framework for the analysis of the various errors in the isomorphous replacement method. We show that phase probability distributions similar to those given by Blow & Crick (1959) may be derived, beginning with very basic assumptions. This derivation shows that current methods of estimating E^2 are justified except that the value of E^2 for centric reflections is twice that for acentric reflections. We also suggest a somewhat improved method of estimating E^2. Finally, we show how the lack-of-closure residuals may be used to estimate the lack of isomorphism between native and derivative structures as well as the error in the heavy-atom model if anomalous differences have been measured.

Errors in calculated and measured derivative structure-factor amplitudes

Even if the native phase φ is known, the calculated derivative structure-factor amplitude F^c_{PH} for a particular reflection calculated using $(1b)$ is not generally equal to the measured derivative structure-factor amplitude F^o_{PH}. Aside from anomalous-scattering effects, there are three principal reasons for this: errors in measurement and scaling, errors in the heavy-atom model, and lack of isomorphism between native and derivative structures. The purpose of this section is to analyze these sources of error. Once we have done this, we will be able to write F^o_{PH} and F^c_{PH} in terms of variables which have known probability distributions. This will allow us to write an expression for the native-phase probability distribution.

Errors in measurement

The most obvious errors in $(1a)$ are the errors in measuring and scaling of native and derivative structure-factor amplitudes. We assume here that σ_P and σ_{PH}, the uncertainties in measurement of native and derivative structure-factor amplitudes, respectively, are known (*e.g.* from comparison of intensities of symmetry-related reflections). Assuming further that

[*] Present address: Department of Biochemistry and Molecular Biology, University of Chicago, 920 East 58th Street, Chicago, Illinois 60637, USA.

the errors in measurement have a Gaussian distribution, we may express the probability of measuring a value of the derivative structure-factor amplitude between F^o_{PH} and $F^o_{PH}+\mathrm{d}F^o_{PH}$, given that the true derivative structure-factor amplitude is F_{PH}:

$$P(F^o_{PH}|F_{PH})\,\mathrm{d}F^o_{PH} \propto \exp\left[-(F^o_{PH}-F_{PH})^2/2\sigma^2_{PH}\right]\mathrm{d}F^o_{PH}. \quad (2)$$

An analogous probability distribution may be written for the observed native structure-factor amplitude.

Errors in the heavy-atom model

A major uncertainty in F^c_{PH}, the calculated derivative structure-factor amplitude, is the error in the heavy-atom model leading to the estimate $\mathbf{f_H}$ of the heavy-atom structure factor. Let $\mathbf{F_H}$ denote the true heavy-atom structure factor, define the 'residual heavy-atom structure factor' $\boldsymbol{\eta}$ as the difference between the true and estimated values of the heavy-atom structure factor: $\boldsymbol{\eta}\equiv\mathbf{F_H}-\mathbf{f_H}$ (see Fig. 1 and Table 1). Then, if the native and derivative structures are perfectly isomorphous, the true native $\mathbf{F_P}$ and derivative $\mathbf{F_{PH}}$ structure factors are related by $\mathbf{F_{PH}}=\mathbf{F_P}+\mathbf{f_H}+\boldsymbol{\eta}$. We will assume that $\mathbf{f_H}$ and $\boldsymbol{\eta}$ are not correlated, which will generally be the case unless incorrect sites have been included in the heavy-atom model.

As long as there are several minor heavy-atom sites *not* included in the heavy-atom model (as will generally be the case), the probability distribution for $\boldsymbol{\eta}$ will be nearly a two-dimensional Gaussian distribution (Wilson, 1949). Defining $\boldsymbol{\eta}\equiv(\eta_x,\eta_y)$, we may therefore write that, for acentric reflections, the probability that η_x is between η_x and $\eta_x+\mathrm{d}\eta_x$, and that η_y is between η_y and $\eta_y+\mathrm{d}\eta_y$ is

$$P(\boldsymbol{\eta})\,\mathrm{d}^2\boldsymbol{\eta}\equiv P(\eta_x,\eta_y)\,\mathrm{d}\eta_x\,\mathrm{d}\eta_y \propto \exp\left[-(\eta_x^2+\eta_y^2)/H^2\right]\mathrm{d}\eta_x\,\mathrm{d}\eta_y \quad (3a)$$

where the mean-square value of the residual heavy-atom structure-factor amplitude $\langle|\boldsymbol{\eta}|^2\rangle$ is H^2. For centric reflections, which can have only one of two possible phases, the probability distribution for $\boldsymbol{\eta}$ is a one-dimensional Gaussian with the same mean-square amplitude H^2 (Wilson, 1949). Denoting the residual heavy-atom structure factor by the scalar η, we have

$$P(\eta)\,\mathrm{d}\eta \propto \exp-(\eta^2/2H^2)\,\mathrm{d}\eta. \quad (3b)$$

Table 1. *Summary of principal variables*

Variable	Symbol	Observed quantity (if any)
Native structure factor and phase	$\mathbf{F_P}\equiv F_P\exp(i\varphi)$	F^o_P
Derivative structure factors (elements of a Friedel pair)	$\mathbf{F^+_{PH}}, \mathbf{F^-_{PH}}$	$\bar{F}^o_{PH}$
Average derivative structure factor	$\mathbf{\bar{F}_{PH}}$	
Anomalous difference	Δ_{PH}	Δ^o_{PH}
Heavy-atom structure factor		
non-anomalous part	$\mathbf{F_H}$	
anomalous part	$\boldsymbol{\delta}$	
Estimated heavy-atom structure factor (non-anomalous part)	$\mathbf{f_H}$	
Residual heavy-atom structure factor (non-anomalous part)	$\boldsymbol{\eta}$	
Lack-of-isomorphism structure factor	$\boldsymbol{\mu}$	

Basic relationships

$$\mathbf{F_{PH}}\equiv\mathbf{F_P}+\boldsymbol{\mu}+\mathbf{F_H}$$
$$\mathbf{F_H}\equiv\mathbf{f_H}+\boldsymbol{\eta}$$
$$\mathbf{F^+_{PH}}\equiv\mathbf{\bar{F}_{PH}}+\boldsymbol{\delta}$$
$$\mathbf{F^{-*}_{PH}}\equiv\mathbf{\bar{F}_{PH}}-\boldsymbol{\delta}$$
$$\Delta_{PH}\equiv\tfrac{1}{2}(|\mathbf{F^+_{PH}}|-|\mathbf{F^-_{PH}}|)$$

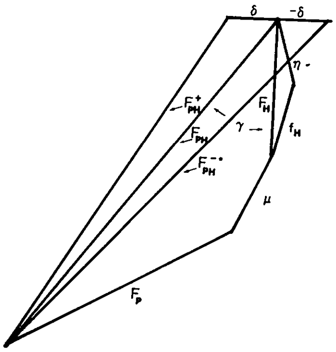

Fig. 1. The relationship of structure factors and other quantities discussed in the text. Definitions of all symbols are summarized in Table 1.

Notice that although the mean-square values H^2 are equal in the acentric and centric cases, the denominators of the exponentials in (3*a*) and (3*b*) differ by a factor of two: this is the difference between one-dimensional and two-dimensional Gaussian distributions with equal mean-square values.

Lack of isomorphism

Another major error in the calculated derivative structure-factor amplitude F^c_{PH} is the effect of non-isomorphism between native and derivative structures. We define the lack-of-isomorphism structure factor $\boldsymbol{\mu}$ as the difference between true native and derivative structure factors caused by all sources except heavy-atom substitution in the derivative structure.

$$\boldsymbol{\mu}\equiv\mathbf{F_{PH}}-(\mathbf{F_P}+\mathbf{F_H}), \quad (4)$$

where $\mathbf{F_{PH}}$, $\mathbf{F_P}$, and $\mathbf{F_H}$ are defined above.

The probability distribution for the lack-of-isomorphism structure factor $\boldsymbol{\mu}$ is, like that for the residual heavy-atom structure factor $\boldsymbol{\eta}$, very nearly a two-dimensional Gaussian for acentric reflections and a one-dimensional Gaussian for centric reflections, as is shown in the Appendix. We can therefore express the probability distribution for the lack-of-isomorphism structure factor for centric reflections as

$$P(\boldsymbol{\mu})\,\mathrm{d}\boldsymbol{\mu} \propto \exp(-\mu^2/2M^2)\,\mathrm{d}\mu \qquad (4a)$$

where $\langle\mu^2\rangle = M^2$. For acentric reflections a nearly identical treatment may be applied. Defining $\boldsymbol{\mu} = (\mu_x, \mu_y)$, we may write

$$P(\boldsymbol{\mu})\,\mathrm{d}^2\boldsymbol{\mu} \equiv P(\mu_x, \mu_y)\,\mathrm{d}\mu_x\,\mathrm{d}\mu_y$$
$$\propto \exp[-(\mu_x^2+\mu_y^2)/M^2]\,\mathrm{d}\mu_x\,\mathrm{d}\mu_y \qquad (4b)$$

where the mean value $\langle|\boldsymbol{\mu}|^2\rangle$ is again M^2.

We have now obtained the probability distributions for the principal sources of error in (1*a*): the errors in measurement of native and derivative structure-factor amplitudes, the errors in the heavy-atom model and the lack of isomorphism between native and derivative structures. The errors from these sources are essentially uncorrelated with each other and with the native and derivative structure factors so that we may treat them all as independent variables in calculating a probability distribution for the native phase φ. Note that given $\mathbf{F_P}$, $\mathbf{f_H}$, $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ we can write an expression for the derivative structure factor $\mathbf{F_{PH}}$ (Fig. 1):

$$\mathbf{F_{PH}} = \mathbf{F_P} + \boldsymbol{\mu} + \mathbf{f_H} + \boldsymbol{\eta}. \qquad (5)$$

Calculation of the native-phase probability distribution

A straightforward application of statistical methods may now be used to calculate the native-phase probability distribution. The central element in this calculation is the relationship known as Bayes' rule (Hamilton, 1964), which we illustrate now. Suppose we have some knowledge of the probability distribution for a variable (from a previous experiment), so that before making a new measurement we may write (*a priori*) that the probability it is between x and $x+\mathrm{d}x$ is $P_0(x)\,\mathrm{d}x$. Now we make a new measurement of this variable, call its value y. Suppose that we know enough about our procedure to say that if the parent value of y is x, then the probability that we would measure y between y and $y+\mathrm{d}y$ is $P(y|x)\,\mathrm{d}y$. Bayes' rule states that we may calculate a new probability distribution for x based on the *a priori* information $P_0(x)$, the measurement y, and the probability that we would have measured y if x were its parent value:

$$P(x) = P_0(x)P(y|x) \Big/ \int_{\text{all } x} P_0(x)P(y|x)\,\mathrm{d}x \qquad (6)$$

or, more simply,

$$P(x) \propto P_0(x)P(y/x). \qquad (7)$$

In the present case, we are interested only in the probability distribution for φ, the native phase. In order to obtain this, however, we must first calculate the joint probability distribution for φ, F_P, $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$. Then we will integrate this joint probability distribution over all values of F_P, $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$, in order to obtain the probability distribution for φ. The set of variables φ, F_P, $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ corresponds to the single variable x in the preceding example.

We have *a priori* information about some of the fundamental variables in this calculation. Consider first an acentric reflection. If there is phase probability information from other sources, denote this *a priori* distribution by $P_0(\varphi)$. We assume that these other sources are independent so that the *a priori* phase probability distribution does not depend on any of the variables of current interest. An *a priori* probability distribution for the native structure-factor amplitude is also known (Wilson, 1949),

$$P_0(F_P) \propto F_P \exp(-F_P^2/\Sigma) \qquad (8)$$

where Σ is the mean-square native structure-factor amplitude at the appropriate resolution. In (3*a*) and (4*a*) we have *a priori* probability distributions for $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$. Before making any measurements of F_P^o and F_{PH}^o, then, we may write that the joint probability distribution for φ, F_P, $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$ is

$$P_0(\varphi, F_P, \boldsymbol{\eta}, \boldsymbol{\mu}) \propto P_0(\varphi)P_0(F_P)P(\boldsymbol{\eta})P(\boldsymbol{\mu}). \qquad (9)$$

Now, if we apply Bayes' rule after making measurements F_P^o and F_{PH}^o of the native and derivative structure-factor amplitudes, we may use (2) and the analogous equation for F_P^o to express P:

$$P(\varphi, F_P, \boldsymbol{\eta}, \boldsymbol{\mu}) \propto P_0(\varphi, F_P, \boldsymbol{\eta}, \boldsymbol{\mu})$$
$$\times P(F_P^o|F_P)P(F_{PH}^o|F_{PH}) \qquad (10)$$

where $F_{PH} \equiv |\mathbf{F_{PH}}|$ is calculated using (5). Finally, to obtain the probability distribution for φ alone, we may integrate (10) over all possible values of F_P, $\boldsymbol{\eta}$, $\boldsymbol{\mu}$:

$$P(\varphi) \propto \int P(\varphi, F_P, \boldsymbol{\eta}, \boldsymbol{\mu})\,\mathrm{d}F_P\,\mathrm{d}^2\boldsymbol{\eta}\,\mathrm{d}^2\boldsymbol{\mu} \qquad (11)$$

or, using (2), (3), (4*b*) and (8),

$$P(\varphi) \propto P_0(\varphi) \int \exp[-F_P^2/\Sigma - |\boldsymbol{\eta}|^2/H^2$$
$$-|\boldsymbol{\mu}|^2/M^2 - (F_P - F_P^o)^2/2\sigma_P^2$$
$$-(F_{PH} - F_{PH}^o)^2/2\sigma_{PH}^2]F_P\,\mathrm{d}F_P\,\mathrm{d}^2\boldsymbol{\eta}\,\mathrm{d}^2\boldsymbol{\mu} \qquad (12)$$

where F_{PH} is calculated from (5). Equation (12) is an exact expression which includes all the information discussed here relating to the probability distribution for the native phase φ. If we assume further that σ_P, σ_{PH}, $|\boldsymbol{\eta}|$ and $|\boldsymbol{\mu}|$ as well as $|\mathbf{f_H}|$ are small relative to F_P^o and F_{PH}^o, (12) may be integrated using first-order approximations to yield an approximate phase probability distribution for acentric reflections:

$$P(\varphi) \propto \exp\{-(F_{PH}^o - F_{PH}^c)^2$$
$$\times [H^2 + M^2 + 2\sigma_P^2 + 2\sigma_{PH}^2]^{-1}\}P_0(\varphi), \qquad (13)$$

where F^c_{PH} is evaluated using (1*b*). Except for the *a priori* probability distribution for the native phase $P_0(\varphi)$, this is equivalent to the formula given by Blow & Crick (1959), as shown in (1*a*), where E^2 is given by

$$E^2 = E^2_{\rm ACENT} = H^2/2 + M^2/2 + \sigma^2_P + \sigma^2_{PH}. \qquad (14)$$

Our procedure for calculating the phase probability distribution, which begins from very basic assumptions, has yielded the same result as this much earlier work except that the contributions of various sources to E^2 are now separated into defined components.

An analogous procedure may be used to obtain an expression for the phase probability distribution for centric reflections:

$$\begin{aligned}P(\varphi) \propto P_0(\varphi) \textstyle\int \exp\,[-F^2_P/2\Sigma - \eta^2/2H^2 \\ - \mu^2/2M^2 - (F_P - F^o_P)^2/2\sigma^2_P \\ - (F_{PH} - F^o_{PH})^2/2\sigma^2_{PH}]\,{\rm d}F_P\,{\rm d}\eta\,{\rm d}\mu \quad (15)\end{aligned}$$

which may be integrated, using the same approximations as above, to yield

$$\begin{aligned}P(\varphi) \propto \exp\{-(F^o_{PH} - F^c_{PH})^2 \\ \times [2H^2 + 2M^2 + 2\sigma^2_P + 2\sigma^2_{PH}]^{-1}\}P_0(\varphi). \quad (16)\end{aligned}$$

This expression is identical to that for acentric reflections except that E^2 for centric reflections is

$$E^2_{\rm CENT} = H^2 + M^2 + \sigma^2_P + \sigma^2_{PH}. \qquad (17)$$

That is, aside from errors in measurement, the centric E^2 is twice the acentric E^2. This difference is entirely due to the difference between the expressions for one-dimensional and two-dimensional Gaussian distributions with equal mean-square values [*e.g.* (3*a*) and (3*b*)].

Calculation of the phase probability distribution in the presence of anomalously scattering atoms in the derivative structure

In order to calculate the phase probability distribution in the presence of anomalously scattering heavy atoms in the derivative structure, we need expressions for the average derivative structure factor $\bar{\mathbf{F}}_{\mathbf{PH}}$ and the anomalous difference Δ_{PH} (North, 1965). Since anomalous differences are rarely measured with an accuracy much better than the typical values of the anomalous differences, relatively crude approximations will suffice here. If we assume that the heavy-atom structure-factor amplitude $|\mathbf{F_H}|$ is small relative to the native structure-factor amplitude $|\mathbf{F_P}|$, the average derivative structure-factor amplitude $\bar{F}_{PH}$ and the anomalous difference Δ_{PH} are given approximately by (see the Appendix and North, 1965):

$$\bar{F}_{PH} \equiv \tfrac{1}{2}(|\mathbf{F}^+_{\mathbf{PH}}| + |\mathbf{F}^-_{\mathbf{PH}}|) \doteq F_{PH}, \qquad (18)$$

$$\Delta_{PH} \equiv \tfrac{1}{2}(|\mathbf{F}^+_{\mathbf{PH}}| - |\mathbf{F}^-_{\mathbf{PH}}|) \doteq -1/\kappa F_H \sin\gamma \qquad (19)$$

where γ is the angle between $\mathbf{F_H}$ and $\mathbf{F^+_{PH}}$, $\mathbf{F_{PH}}$ is from (5), and κ is the ratio of real to anomalous scattering for the heavy atoms. Notice that the anomalous difference is relatively insensitive to lack of isomorphism.

An expression analogous to (12) for the phase probability distribution when anomalous differences have been measured is:

$$\begin{aligned}P(\varphi) \propto P_0(\varphi) \textstyle\int \exp\,[-F^2_P/\Sigma - |\boldsymbol{\eta}|^2/H^2 - |\boldsymbol{\mu}|^2/M^2 \\ -(F_P - F^o_P)^2/2\sigma^2_P - (\bar{F}_{PH} - \bar{F}^o_{PH})^2/2\sigma^2_{PH} \\ - (\Delta_{PH} - \Delta^o_{PH})^2/2\sigma^2_{\rm ANO}]F_P\,{\rm d}F_P\,{\rm d}^2\boldsymbol{\eta}\,{\rm d}^2\boldsymbol{\mu} \\ (20)\end{aligned}$$

where $\bar{F}_{PH}$ and Δ_{PH} are calculated using (18) and (19), $\sigma_{\rm ANO}$ is the uncertainty in the anomalous difference, and the integral is over all possible values of F_P, $\boldsymbol{\eta}$ and $\boldsymbol{\mu}$. With the same approximations which led to (12), (20) may be integrated to yield an approximate phase probability distribution when anomalous differences have been measured:

$$\begin{aligned}P(\varphi) \propto \exp\{-(\bar{F}^o_{PH} - \bar{F}^c_{PH})^2 \\ \times [H^2 + M^2 + 2\sigma^2_P + 2\sigma^2_{PH}]^{-1} \\ - (\Delta^o_{PH} - \Delta^c_{PH})^2[H^2/\kappa^2 + 2\sigma^2_{\rm ANO}]^{-1}\}P_0(\varphi) \\ (21)\end{aligned}$$

where $\bar{F}^c_{PH}$ and Δ^c_{PH} are calculated using (18) and (19).

The first term in the exponent of (21) is identical to the exponent in (13). The second term is due to the effects of anomalous scattering and is identical to that given by North (1965) except that $E^2_{\rm ANO}$ is given by the expression

$$E^2_{\rm ANO} = (H^2/2\kappa^2) + \sigma^2_{\rm ANO}. \qquad (22)$$

Notice that $E^2_{\rm ANO}$ does not depend on the lack of isomorphism (M^2).

Estimation of $E^2_{\rm ACENT}$, $E^2_{\rm CENT}$ and $E^2_{\rm ANO}$

Equations (13), (16) and (19) require estimates of $E^2_{\rm ACENT}$, $E^2_{\rm CENT}$ and $E^2_{\rm ANO}$, which are not available directly from measurements. To develop a method for estimating them we first calculate the expected value of the centric, acentric and anomalous mean-square lack-of-closure residuals, evaluated in the usual fashion (Blow & Crick, 1959; Ten Eyck & Arnone, 1976), assuming that the native phase φ is known precisely. Using the same first-order approximations as above, it may be shown, using (3*b*), (4*a*) and (5), that the mean-square value of the centric lack-of-closure residual is

$$\langle(F^o_{PH} - F^c_{PH})^2\rangle \doteq H^2 + M^2 + \langle\sigma^2_P\rangle + \langle\sigma^2_{PH}\rangle \\ \text{(centric).} \quad (23)$$

If we assume once again that the native phase is known, the expected values of the acentric and anomalous lack-of-closure residuals are given, using (3*a*), (4*b*), (5), (*A*16), (18) and (19), by

$$\langle (F^o_{PH} - F^c_{PH})^2\rangle \doteq H^2/2 + M^2/2 + \langle \sigma^2_P\rangle + \langle \sigma^2_{PH}\rangle \quad \text{(acentric)} \qquad (24)$$

and

$$\langle (\Delta^o_{\rm PH} - \Delta^c_{PH})^2\rangle \doteq H^2/(2\kappa^2) + \langle \sigma^2_{\rm ANO}\rangle \qquad (25)$$

where the constant κ is defined above.

Comparing (23)-(25) with (14), (17) and (22), it is clear that the appropriate values of E^2 for centric, acentric and anomalous differences in the formulations of Blow & Crick (1959) and North (1965) are the usual mean-square lack-of-closure residuals calculated using the correct native phases. It must be noted that centric and acentric reflections are to be treated separately, however.

The standard methods of estimating the lack-of-closure errors (Blow & Crick, 1959; Ten Eyck & Arnone, 1976) rely on the values of the lack-of-closure residuals evaluated at the 'best' or 'most probable' native phases. If the native phases are not known with certainty, then this procedure will yield inaccurate values of the lack-of-closure errors (Blow & Matthews, 1973). Since a probability distribution for the phase is available [(21), for example], improved estimates of the lack-of-closure errors may be obtained simply by averaging the standard lack-of-closure residuals over all native phases, weighting by the probability that each phase is correct. We therefore define the 'averaged' lack-of-closure residuals $\bar{E}^2_{\rm ACENT}$, $\bar{E}^2_{\rm CENT}$ and $\bar{E}^2_{\rm ANO}$ as follows:

$$\bar{E}^2_{\rm CENT} \equiv \left\langle \sum_{j=1}^{2} P(\varphi_j)[F^o_{PH} - F^c_{PH}(\varphi_j)]^2 \right\rangle, \qquad (26)$$

$$\bar{E}^2_{\rm ACENT} \equiv \langle \textstyle\int P(\varphi)[\bar{F}^o_{PH} - \bar{F}^c_{PH}(\varphi)]^2\, {\rm d}\varphi\rangle, \qquad (27)$$

$$\bar{E}^2_{\rm ANO} \equiv \langle \textstyle\int P(\varphi)[\Delta^o_{PH} - \Delta^c_{PH}(\varphi)]^2\, {\rm d}\varphi\rangle, \qquad (28)$$

where the normalized phase probability distribution $P(\varphi)$ includes all information available and the angle brackets indicate an average over all appropriate reflections in a given range of resolution.

Estimation of the mean-square residual heavy-atom and lack-of-isomorphism structure factors

Based on the preceding section, we can expect that the averaged lack-of-closure residuals defined by (26)-(28) will be reasonable estimates of $E^2_{\rm ACENT}$, $E^2_{\rm CENT}$ and $E^2_{\rm ANO}$. Also, if we recall (14), (17) and (22), it is evident that the averaged lack-of-closure residuals may be used to estimate the mean-square residual heavy-atom structure factor (H^2) and the mean-square lack-of-isomorphism structure factor (M^2) at a given resolution. From (14), (17), (26) and (27) we have

$$H^2 + M^2 \doteq \bar{E}^2_{\rm CENT} - \langle \sigma^2_P\rangle - \langle \sigma^2_{PH}\rangle, \qquad (29)$$

$$H^2 + M^2 \doteq 2(\bar{E}^2_{\rm ACENT} - \langle \sigma^2_P\rangle - \langle \sigma^2_{PH}\rangle) \qquad (30)$$

where the angle brackets indicate an average over all centric or acentric reflections in the appropriate range of resolution. The sum of H^2 and M^2 may therefore be obtained either from the averaged centric lack-of-closure residual or the averaged acentric lack-of-closure residual.

The average anomalous lack-of-closure residual $\bar{E}^2_{\rm ANO}$ depends on H^2 but not on M^2, so that if anomalous differences have been measured, separate estimates of H^2 and M^2 may be obtained. Using (22) and (28), we may write

$$H^2 \doteq \langle \kappa^2\rangle(\bar{E}^2_{\rm ANO} - \langle \sigma^2_{\rm ANO}\rangle), \qquad (31)$$

where the values of κ are the same as those used in (21). Since the value of κ is typically greater than unity, it is clear that estimates of H^2 obtained from (31) will generally be fairly coarse. Notice that the estimate of H^2 does require an accurate estimate of the uncertainties in measurement of the anomalous differences $\langle \sigma^2_{\rm ANO}\rangle$.

Discussion

We have shown here that, with reasonable approximations, the phase probability distributions given by Blow & Crick (1959) and by North (1965) and Matthews (1966) can be derived from basic premises. Blow & Crick (1959), citing empirical evidence, assumed that the errors in observed and calculated derivative structure-factor amplitudes were distributed in a Gaussian fashion. Approximations similar to those we have used led them to (1*a*). On the basis of a detailed analysis of errors, we have found that the assumption of Gaussian distributions is very reasonable, justifying (1*a*). More importantly, the present analysis has yielded a straightforward interpretation of the E^2 for centric, acentric and anomalous differences [(14), (17) and (22)] and an improved method of estimating them from the averaged lack-of-closure residuals [(26)-(28)].

Equations (1*a*), (13), (16) and (21) are all approximate phase probability distributions. Given our assumptions, however, (12), (15) and (20) are exact expressions. Nearly any desired degree of accuracy in the phase probability distributions can be obtained by a suitably precise evaluation of these expressions. Green (1979) has analyzed expressions similar to (12) with few approximations, and found that, in cases where lack of isomorphism was not severe, the resulting phase probability distribution and that obtained using (13) were quite similar. We have also carried out numerical integrations of (12) and find that the resultant phase probability distribution is given very precisely by (13) when $|\mathbf{f}_{\rm H}|$, $|\boldsymbol{\eta}|$ and $|\boldsymbol{\mu}|$ are all less than 20% of F^o_P.

An error far more serious than the approximate evaluation of (12) [or of (15) or (20)] is that caused by using inaccurate values of the lack-of-closure errors. A factor-of-two change in the lack-of-closure error in ($1a$) is sufficient to convert a probability distribution with a 'figure of merit' of 0·45 into one with a figure of merit of 0·07, an error far greater than that typically caused by using ($1a$) instead of (12). This is why we have focused attention on the evaluation of these lack-of-closure errors and why we suggest the use of the averaged lack-of-closure residuals [(26)–(28)] for this purpose.

In order to determine the effects of this new procedure, we have calculated lack-of-closure residuals using data from a model native structure and from model derivative structures which included heavy-atom substitution and which were not completely isomorphous with the native structure. Table 2 compares the estimates of the mean-square lack-of-closure residuals obtained from (26)–(28) with their expected values from (14), (17) and (22) and with the conventional lack-of-closure residuals for a case with one derivative and a mean figure of merit of 0·47. In this test, lack-of-closure residuals calculated using the 'best' or 'most probable' phases were seriously underestimated, even for centric reflections. The averaged lack-of-closure residuals, however, were reasonable estimates of the true lack-of-closure errors, even in this case with one derivative. Also notice that, as predicted by (23) and (24), the mean-square centric lack-of-closure residual is essentially twice the mean-square acentric lack-of-closure residual.

The mean-square residual heavy-atom structure-factor amplitude (H^2) and the mean-square lack-of-isomorphism structure-factor amplitude (M^2) may be estimated from the averaged lack-of-closure residuals using (29)–(31), if the anomalous differences have been measured and uncertainties in these measurements are known. In the example in Table 2, anomalous differences were calculated, and in Table 3 are presented the estimates of H^2 and M^2 calculated from (29)–(31) with the data of Table 2. As noted earlier, the separate estimation of H^2 and M^2 depends on an accurate estimate of the *uncertainty* in measurement of the anomalous differences. These estimates were available in the present example, but are not available in all cases. When H^2 and M^2 can be accurately estimated, they may be used to determine when the derivative is 'solved', and all remaining differences between native and derivative structures are due to errors in measurement and lack of isomorphism.

As noted by Einstein (1977), our assumption that all *a priori* sources of phase information are independent of the present calculation is not always strictly correct. One set of native structure factors is often used in conjunction with each of several derivative sets of structure-factor amplitudes. Consequently, the phase probability distributions for the various derivative–native combinations are not independent.

Table 2. *Estimation of lack-of-closure errors using one derivative*

Native and derivative structure-factor amplitudes were calculated from a model for 407 acentric and 56 centric reflections from 2·0 to 2·1 Å resolution. The models were based on the partially refined structure of melittin form II crystals (space group $C222_1$; Terwilliger & Eisenberg, 1982). In order to introduce non-isomorphism between the derivative structure and the native structure, the coordinates of six protein atoms were moved a r.m.s. distance of 0·8 Å in the derivative structure. The actual mean-square 'lack-of-isomorphism' structure-factor amplitude (μ^2) in Table 3 was calculated using equation ($A3$) for centric reflections and the analogous expression for acentric reflections. The heavy-atom contribution to the derivative structure factor was calculated using five arbitrarily placed heavy-atom sites of equal occupancy. The magnitudes of the native and derivative structure factors were then calculated, and, after addition of a small random 'observational error', these amplitudes were used as the 'observed' native and derivative structure-factor amplitudes. The r.m.s. native structure-factor amplitude is 140, the r.m.s. 'observational' uncertainty in the native structure-factor amplitudes is 1·5, and the r.m.s. uncertainty in the derivative structure-factor amplitudes and anomalous differences is also 1·5.

In calculating phase probability distributions and lack-of-closure residuals, three of the five heavy-atom sites were included with the correct occupancy in the calculation of the 'estimates' of the heavy-atom structure factors ($\mathbf{f_H}$). The 'actual' mean-square residual heavy-atom structure-factor amplitude in Table 3 is the r.m.s. heavy-atom structure-factor amplitude due to the two sites not included in the model. The expected values of the residuals were calculated using the data in Table 3. The lack-of-closure residuals averaged over all phases were determined using equations (26)–(28). Those estimated at the 'true', 'best' or 'most probable' phases were calculated using the same equations, setting $P(\varphi)$ equal to zero except at the 'true', 'best' or 'most probable' phase.

For each of the three methods of estimating lack-of-closure residuals, phase probability distributions were evaluated using equations (16) and (21) which themselves require estimates of the lack-of-closure errors. Therefore, a procedure designed to simulate a real situation was followed. For each method, the average value of $(F^o_{PH}-F^o_P)^2$ over all centric reflections was used as a starting value of the centric lack-of-closure error, and the average value of $(\Delta^o_{PH})^2$ over all acentric reflections was used as a starting value of the anomalous lack-of-closure error. The acentric lack-of-closure error was always taken to be half the centric lack-of-closure error. Phase probability distributions were calculated and new estimates of the lack-of-closure errors were obtained from the corresponding lack-of-closure residuals. The new estimate of the centric lack-of-closure error was the weighted average of the centric lack-of-closure residual and twice the acentric lack-of-closure residual. This cycle was repeated once more and the residuals listed were found.

In calculating the lack-of-closure residuals, only reflections which satisfied the relation

$$F^o_{PH}>4(\sigma^2_P+\sigma^2_{PH}+H^2+M^2)$$

were included. This criterion resulted in the rejection of 21% of the centric and 12% of the acentric reflections. The mean figure of merit for the test was 0·47.

Native phase used to estimate lack-of-closure residuals	Mean-square lack-of-closure residual		
	Centric	Acentric	Anomalous
Expected values [equations (14), (17), (22)]	304	154	3·5
True φ	270	140	3·5
Averaged over φ	330	160	3·4
'Best' φ	210	80	1·4
'Most probable' φ	210	80	1·3

Table 3. *Estimation of H^2 and M^2 from the data in Table* 2

The r.m.s. residual heavy-atom structure-factor amplitude (H^2) was estimated from the data in Table 2 using equation (22). The value of κ was 5·8 in all cases. The r.m.s. lack-of-isomorphism structure-factor amplitude (M^2) was calculated from H^2 and equations (14) and (17).

	H^2	M^2
Actual value	90	210
Estimates from the single derivative	80	260

However, as long as the variances σ_P^2 of the observed native structure-factor amplitudes are small relative to the overall error $(\frac{1}{2}H^2+\frac{1}{2}M^2+\sigma_{PH}+\sigma_P^2)$, the errors in measurement of the native structure-factor amplitudes will not greatly affect the phase probability distribution.

Caution should be exercised when the various derivatives share heavy-atom sites. A substantial correlation between the phase probability distributions obtained for various derivatives may occur when there are *unknown* heavy-atom sites in common in the various derivatives. In this case, the residual heavy-atom structure factor $\boldsymbol{\eta}$ is not independent in the various derivatives and therefore the phase probability distributions based on the various derivatives are not independent. Of course, it is difficult to know when two derivatives share unknown heavy-atom sites, but it seems likely that, in cases where *known* heavy-atom sites are in common in two derivatives, unknown sites are also shared. A similar argument applies to the correlation between the lack-of-isomorphism structure factors $\boldsymbol{\mu}$ in various derivatives which share heavy-atom sites.

A Fortran program (*HEAVY*) which incorporates the results described here may be obtained from the Protein Data Bank, Brookhaven National Laboratory, Upton, Long Island, New York 11973, USA.

We thank Drs S. Sheriff, L. Weissman, W. Smith and R. Sweet for helpful discussions. This work was supported by USPHS grant GM-31299. TCT was the recipient of a National Science Foundation Graduate Fellowship.

APPENDIX

Probability distribution for the lack-of-isomorphism structure factor

Consider a centric reflection. In this case, the native structure factor F_P may be written as a scalar (Wilson, 1949).

$$F_P=2\sum_{j=1}^{N/2} f_j\cos(2\pi\mathbf{s}\,.\,\mathbf{x}_j) \qquad (A1)$$

where f_j and $\mathbf{x}_j$ are the form factors and coordinates of atom j, $\mathbf{s}\equiv h\mathbf{a}^*+k\mathbf{b}^*+l\mathbf{c}^*$ is the position of this reflection in reciprocal space, and there are N atoms in the unit cell. Then we may write an expression for the contribution F'_P of atoms present in the native structure to the derivative structure factor F_{PH}:

$$F'_P=2\sum_{j=1}^{N/2} f_j\cos[2\pi\mathbf{s}\,.\,(\mathbf{x}_j+\delta\mathbf{x}_j)] \qquad (A2)$$

where atom j has shifted position from $\mathbf{x}_j$ to $\mathbf{x}_j+\delta\mathbf{x}_j$ from native to derivative structures. An expression for the lack-of-isomorphism structure factor μ in this case is

$$\mu=2\sum_{j=1}^{N/2} f_j\{\cos[2\pi\mathbf{s}\,.\,(\mathbf{x}_j+\delta\mathbf{x}_j)]-\cos(2\pi\mathbf{s}\,.\,\mathbf{x}_j)\} \qquad (A3)$$

or, after rearrangement,

$$\mu=2\sum_{j=1}^{N/2} f_j\{\cos(2\pi\mathbf{s}\,.\,\mathbf{x}_j)[\cos(2\pi\mathbf{s}\,.\,\delta\mathbf{x}_j)-1] -\sin(2\pi\mathbf{s}\,.\,\mathbf{x}_j)\sin(2\pi\mathbf{s}\,.\,\delta\mathbf{x}_j)\}. \qquad (A4)$$

Now, using reasoning similar to that used by Wilson (1949), we note that $\cos(2\pi\mathbf{s}\,.\,\mathbf{x}_j)$ and $\sin(2\pi\mathbf{s}\,.\,\mathbf{x}_j)$ in ($A4$) vary in an essentially random fashion from -1 to 1. Therefore, μ is the sum of many uncorrelated random variables with mean values of zero. According to the central-limit theorem (Hamilton, 1964), μ will be distributed in a Gaussian fashion with an expected value of zero and a mean-square value equal to the sum of the mean-square values of the random variables. This can be shown to be

$$\langle\mu^2\rangle=4\sum_{j=1}^{N/2} f_j^2[1-\cos(2\pi\mathbf{s}\,.\,\delta\mathbf{x}_j)]. \qquad (A5)$$

If there are many terms in ($A5$) and the $\delta\mathbf{x}_j$ are randomly distributed, this expected mean-square value of μ will characterize all reflections at a given resolution (at other resolutions, the form factors f_i will be different). This leads to (4a) and a similar argument leads to (4b).

It may be shown that in cases of extreme non-isomorphism (for example, native and derivative structures unrelated) the lack-of-isomorphism structure factor μ will be strongly correlated with both the native and derivative structure factors. When the magnitude of μ is small relative to the native structure-factor amplitude, however, μ is correlated with neither. To show this for centric reflections (as before, a nearly identical argument applies to acentric reflections), define the correlation α between μ and F_P as the average value, over many reflections, of the product of μ and F_P, divided by the r.m.s. values of μ and F_P:

$$\alpha\equiv\langle\mu F_P\rangle/\langle\mu^2\rangle^{1/2}\langle F_P^2\rangle^{1/2}. \qquad (A6)$$

From ($A1$) and ($A3$), it can be shown that $\langle\mu F_P\rangle\doteq -\frac{1}{2}\langle\mu^2\rangle$, so that the correlation α is given by

$$\alpha\doteq-\tfrac{1}{2}\langle\mu^2\rangle^{1/2}/\langle F_P^2\rangle^{1/2}.$$

When $\langle\mu^2\rangle$ is small relative to $\langle F_P^2\rangle$, α is clearly small.

Expressions for $\bar{F}_{PH}$ and Δ_{PH} in the presence of anomalously scattering heavy atoms

Suppose that all the heavy atoms in the derivative structure have the same ratio of 'real' to 'anomalous' scattering, given by $\kappa = (f + f')/f''$. Then $\mathbf{F_H}$, the non-anomalous part of the heavy-atom structure factor, is perpendicular to $\boldsymbol{\delta}$, the anomalous part of the heavy-atom structure factor:

$$\boldsymbol{\delta} = \exp(i\pi/2) F_{\mathbf{H}}/\kappa \qquad (A7)$$

so that the total heavy-atom structure factor is given by

$$\mathbf{F_H} + \boldsymbol{\delta} = \mathbf{F_H}[1 + \exp(i\pi/2)/\kappa]. \qquad (A8)$$

If there are several types of anomalously scattering heavy atoms present, ($A7$) may be generalized. Let

$$\mathbf{F_H} = \sum_j \mathbf{F}_{\mathbf{H},j} \qquad (A9)$$

and

$$\boldsymbol{\delta} = \sum_j \boldsymbol{\delta}_j = \exp(i\pi/2) \sum_j \mathbf{F}_{\mathbf{H},j}/\kappa_j \qquad (A10)$$

where $\mathbf{F}_{\mathbf{H},j}$, $\boldsymbol{\delta}_j$ and κ_j refer to all atoms of a particular type. Now define the angles α and ω by

$$\mathbf{F_H} \equiv |\mathbf{F_H}| \exp(i\alpha), \qquad (A11)$$

$$\boldsymbol{\delta} \equiv |\boldsymbol{\delta}| \exp(i\alpha + i\pi/2 + i\omega) \qquad (A12)$$

where ω is zero if all heavy atoms are identical. We may now define an effective value of κ which is related to the ratio of 'real' to 'anomalous' scattering for this particular reflection:

$$\kappa \equiv [(|\mathbf{F_H}|/|\boldsymbol{\delta}|) - \sin\omega]/\cos\omega \qquad (A13)$$

where κ is equal to each κ_j if all atoms are identical. Equations ($A9$)-($A12$) may now be combined to yield

$$\mathbf{F_H} + \boldsymbol{\delta} = \mathbf{F_H}[1 + \exp(i\pi/2)/\kappa]/[1 - |\boldsymbol{\delta}| \sin(\omega)/|\mathbf{F_H}|]. \qquad (A14)$$

If it is assumed that the ratio κ_j of real to anomalous scattering is large for all the heavy atoms present, then $|\boldsymbol{\delta}| \sin(\omega)/|\mathbf{F_H}|$ is nearly always small compared with unity. If this term is ignored, ($A14$) becomes

$$\mathbf{F_H} + \boldsymbol{\delta} \doteq \mathbf{F_H}[1 + \exp(i\pi/2)/\kappa]. \qquad (A15)$$

Equation ($A15$) cannot be used directly to calculate the total heavy-atom structure factor from its non-anomalous part because the factor κ ($A13$) depends on both $\boldsymbol{\delta}$ and $\mathbf{F_H}$ and may vary from reflection to reflection. In order to avoid this difficulty, we assume that the factor κ, based on $\boldsymbol{\delta}$ and $\mathbf{F_H}$, is the same as that which can be calculated from the heavy-atom model. Then

$$\boldsymbol{\delta} \doteq (\mathbf{f_H} + \boldsymbol{\eta}) \exp(i\pi/2)/\kappa \qquad (A16)$$

where κ is calculated from the heavy-atom model using expressions analogous to ($A11$)-($A13$).

Now we may obtain expressions for the total derivative structure factor including anomalous-scattering effects. Once again, let $\mathbf{F_{PH}}$ be the derivative structure factor due to the non-anomalous part of the scattering from the derivative structure:

$$\mathbf{F_{PH}} \equiv \mathbf{F_P} + \boldsymbol{\mu} + \mathbf{f_H} + \boldsymbol{\eta}. \qquad (A17)$$

Then the total structure factor for one element of this Friedel pair may be written as (see Fig. 1)

$$\mathbf{F}^{+}_{\mathbf{PH}} = \mathbf{F_{PH}} + \boldsymbol{\delta} \qquad (A18)$$

and the complex conjugate of the total structure factor for the other element of this Friedel pair may be written as

$$\mathbf{F}^{-*}_{\mathbf{PH}} = \mathbf{F_{PH}} - \boldsymbol{\delta}. \qquad (A19)$$

With first-order approximations, this leads directly to (18) and (19).

References

BLOW, D. M. & CRICK, F. H. C. (1959). *Acta Cryst.* **12**, 794-802.

BLOW, D. M. & MATTHEWS, B. W. (1973). *Acta Cryst.* **A29**, 56-62.

EINSTEIN, R. J. (1977). *Acta Cryst.* **A33**, 75-85.

GREEN, E. A. (1979). *Acta Cryst.* **A35**, 351-359.

HAMILTON, W. C. (1964). *Statistics in Physical Science.* New York: Ronald Press.

MATTHEWS, B. W. (1966). *Acta Cryst.* **20**, 82-86.

NORTH, A. C. T. (1965). *Acta Cryst.* **18**, 212-216.

TEN EYCK, L. F. & ARNONE, A. (1976). *J. Mol. Biol.* **100**, 3-11.

TERWILLIGER, T. C. & EISENBERG, D. (1982). *J. Biol. Chem.* **257**, 6010-6015, 6016-6022.

WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318-321.