Acta Cryst. (1996). D52, 749–757

Correlated Phasing of Multiple Isomorphous Replacement Data

THOMAS C. TERWILLIGER[a]* AND JOEL BERENDZEN[b]

[a]*Structural Biology Group, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA, and* [b]*Biophysics Group, Mail Stop D454, Los Alamos National Laboratory, Los Alamos, NM 87545, USA*

(Received 14 August 1995; accepted 16 January 1996)

Abstract

Substantial highly correlated differences sometimes exist between a series of heavy-atom derivatives of a macromolecule and the native structure. Use of such a series of derivatives for phase determination by multiple isomorphous replacement (MIR) has been difficult because MIR analysis has treated errors as independent. A simple Bayesian approach has been used to derive probability distributions for the phase in the case where a group of MIR derivatives have correlated errors. The utility of the resulting 'correlated-phasing' method has been examined by applying it to both simulated and real MIR data sets that contain sizeable correlated errors and it has been found that it can dramatically improve MIR phase estimates in these cases. Correlated phasing is applicable to situations where derivatives exhibit substantial correlated changes in protein conformation or crystal packing or where correlated errors in heavy-atom models are large. Correlated phasing does not substantially increase the complexity of phase computation and is suitable for routine use.

1. Introduction

In the method of multiple isomorphous replacement (MIR), the phase problem of crystallography is solved using information from X-ray diffraction data on crystals of the 'native' macromolecule and on several 'derivative' crystals that differ from the native through binding of heavy atoms at a small number of sites in each asymmetric unit. An electron-density map that shows the locations of atoms in the native structure can then be obtained in four steps. First, heavy-atom locations are deduced by difference Patterson or direct methods. Next, a detailed model for the heavy-atom positions in each derivative is built and refined. The refined heavy-atom models are then used to obtain an estimate of the phase of each structure factor for the native crystals. Finally, the phases and measured amplitudes of structure factors for the native crystals are used in a Fourier synthesis to obtain an electron-density map. It is the third step, phasing, with which this paper is concerned. Over the past several decades MIR has proven spectacularly useful in phasing macromolecular data sets. Despite the recent successful applications of multi-wavelength anomalous diffraction phasing techniques (Karle, 1980; Hendrickson, 1991) and direct methods (*e.g.*, Miller *et al.*, 1993) to this problem, MIR remains the workhorse of new macromolecular structure determinations.

MIR is limited by the requirement that the derivatives be highly isomorphous and that the heavy-atom sites be well modeled. If the derivative structures differ substantially from the native or if many of the heavy atoms in a derivative cannot be located, then phasing may prove impossible. Defects in the heavy-atom models or any lack of isomorphism between native and derivative crystals will contribute to uncertainty in the resulting phase in much the same way as do errors in measurement (Terwilliger & Eisenberg, 1987). A serious lack of isomorphism that leads to differences between amplitudes of native and derivative structure factors of 40%, for example, makes the derivative almost worthless for MIR.

It is both common and disappointing to obtain non-isomorphous derivatives, and it would be very helpful if some way were available to use such derivatives in phasing. One scenario in which even poorly isomorphous derivatives could be useful in phase determination is when the derivatives all have the same non-isomorphism with respect to the native. In such a case, the differences among the derivatives, which would be due almost entirely to the different arrangements of heavy atoms, could yield substantial phase information, although in general practice it has been difficult this information except by ignoring the native structure altogether and simply defining one of the derivatives as the 'native'. Although the problem can be addressed in this way, such a procedure will be missing any phasing information that is present in the differences between the native and the derivative structures.

Lack of isomorphism is only one type of correlated error that could exist among derivatives. Others could arise from undetected sites of heavy-atom substitution that are present in each crystal but missing in the heavy-atom models, errors in data collection or scaling in common for all derivatives, or (since MIR phase calculations involve differences between each derivative and the native amplitude for each structure factor)

errors in the measurement of the native diffraction data. The correlated error as a result of measurement errors in the diffraction data from the native crystals was analyzed some time ago (Einstein, 1977) and improved methods of carrying out phase calculations to account for this effect were developed. A more general treatment, however, that can account for correlated lack of isomorphism, correlated scaling errors, and correlated errors in heavy-atom models has not been available up to now.

In this work, we use a Bayesian approach (Box & Tiao, 1973; Box, 1980) to obtain phase information in the presence of errors that are correlated among a set of MIR derivatives. The usefulness of a Bayesian approach to this problem is that it allows a detailed description of the possible sources of error to be used in calculating probability distributions for the native phase. In particular, this approach allows the explicit incorporation of information on the extent of correlation of the errors in each derivative. We previously used a Bayesian approach to derive expressions for phase-probability distributions for a single native-derivative pair (Terwilliger & Eisenberg, 1987). The phase-probability distribution we obtained was similar to the one proposed by Blow and Crick (Blow & Crick, 1959) and in general use at the time, but the derivation led to a more detailed interpretation of the lack-of-closure errors in terms of lack of isomorphism and errors in the heavy-atom model. At that time we found it necessary to assume that if more than one derivative was included in phase-probability calculations, errors were not correlated among the derivatives. This allowed the calculation of independent native phase-probability distributions based on each derivative, and a simple multiplication of these to yield the overall probability distribution for the native phase. In the present derivation, we take advantage of correlations among errors in a way that can substantially improve estimates of phases.

2. The correlated phasing model

We begin by developing a model for the derivative structure factors that includes the correlated and non-correlated sources of error, and we estimate the parameters of the error distributions from the data. We then integrate over the error distributions to obtain an expression for the probability distribution for the native phase. As in our previous treatment of the single isomorphous replacement case (Terwilliger & Eisenberg, 1987), the spirit of these calculations will be along the lines of the Blow–Crick formulation, and we shall be approximating many of the component probability distributions and complex sums to first order (that is, by normal distributions).

2.1. *Correlated and uncorrelated errors*

We describe the effects of X-ray diffraction from the arrangement of atoms in the asymmetric unit of the native protein crystals by a (complex) native structure factor, $\mathbf{F_P}$, for a particular reflection. For the jth derivative crystal, the corresponding derivative structure factor, $\mathbf{F_{PH_j}}$, is given by the native structure factor plus a contribution arising from the total changes due to the heavy atoms, which we write as,

$$\mathbf{F_{PH_j}} = \mathbf{F_P} + (\mathbf{F^c_{H_j}} + \mathbf{R} + \mathbf{S_j}). \qquad (1)$$

The first term in the change, $\mathbf{F^c_{H_j}}$, is the calculated structure factor of the heavy atoms based on the current model, which describes the heavy-atom positions, occupancies, and Debye–Waller factors. The second change term, $\mathbf{R}$ describes the error in the change that is correlated across all derivatives, so $\mathbf{R}$ is not indexed by j. The last change term, $\mathbf{S_j}$, represents the error in the change that is specific to the jth derivative.

The sum $\mathbf{R}+\mathbf{S_j}$ accounts for all errors due to inadequacies of the model, whether arising from non-isomorphism or errors in the heavy-atom model, but it does not include experimental errors in the measurement of $|\mathbf{F_P}|$. We account for the errors in measurement of the native amplitudes by writing the observed amplitude of the native structure factor F^o_P as the sum of the amplitude of $\mathbf{F_P}$ and a measurement error δ_P.

This leads to an expression for the (complex) structure factor for derivative j of

$$\mathbf{F_{PH_j}} = (F^o_P - \delta_P)\exp(i\varphi) + \mathbf{F^c_{H_j}} + \mathbf{R} + \mathbf{S_j}, \qquad (2)$$

where φ is the crystallographic phase of the native protein, the quantity that we are trying to determine in the phasing step. Note that (2) is an expression for the derivative structure factor itself, not our measurement of it. Although in principle it is possible to proceed farther from this expression without additional assumptions, calculations can be greatly simplified if we allow that the amplitude of the native structure factor, F_P, is measured accurately enough that $\delta_P \ll F^o_P$ and also that the total difference between native and derivative structure factors is small compared to F_P. These assumptions are used in the Blow–Crick treatment of phasing and experience has shown them to be generally quite good. We can then write F_{PH_j}, the magnitude of $\mathbf{F_{PH_j}}$, as approximately given by

$$F_{PH_j} \simeq |\mathbf{F^c_{PH_j}}| - \delta_P + R + S_j, \qquad (3)$$

where $\mathbf{F^c_{PH_j}} = F^o_P\exp(i\varphi) + \mathbf{F^c_{H_j}}$, $\delta_P = F_P - F^o_P$ and R and S_j refer to the components of $\mathbf{R}$ and $\mathbf{S_j}$ along the direction of $\mathbf{F^c_{PH_j}}$. Finally, noting that δ_P and R are the same for all derivatives and that S_j is unique to derivative j, rewriting $|\mathbf{F^c_{PH_j}}|$, the calculated amplitude of the derivative structure factor, as $F^c_{PH_j}(\varphi)$, and the observed derivative structure factor $F^o_{PH_j}$ as the sum of F_{PH_j} and a measurement error, δ_{PH_j}. we obtain

$$F^o_{PH_j} \simeq F^c_{PH_j}(\varphi) + \Delta + S_j + \delta_{PH_j}, \quad (4)$$

where $\Delta = -\delta_P + R$. The amplitude of the derivative structure factor therefore differs from that calculated based on the measured amplitude of the native structure factor and the heavy-atom model by a term correlated across all derivatives, Δ, and terms unique to the jth derivative, S_j and δ_{PH_j}. The utility of the present approach will largely depend on whether Δ is sizeable relative to the S_j and δ_{PH_j}.

2.2. *Probability distribution for the native phase*

To obtain a probability distribution for the native phase φ we begin by using Bayes' rule (Box & Tiao, 1973) to write an expression for the posterior probability distribution for φ given that we have made a measurement F^o_P of the native structure factor and measurements $F^o_{PH_1} \ldots F^o_{PH_n}$ of the n derivative structure factors,

$$p(\varphi|F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}) \propto p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi)p_o(\varphi). \quad (5)$$

The prior probability distribution for the native phase, $p_o(\varphi)$, is usually flat and uninformative because we do not know anything beforehand about the native phase φ. However, if there is information available from another experiment, such as a multi-wavelength anomalous diffraction experiment, this probability distribution should reflect this prior information.

We do not know the distribution on the right hand side of (5), but using (3) to calculate the $F^o_{PH_j}$ we can obtain the related probability distribution $p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi, F_P, R, S_1 \ldots S_n)$ assuming that the errors in measurement are normally distributed,

$$p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi, F_P, R, S_1 \ldots S_n) \propto \mathcal{N}(F^o_P - F_P, \sigma^2_P) \prod_j \mathcal{N}(F^o_{PH_j} - F_{PH_j}, \sigma^2_{PH_j}), \quad (6)$$

where $\mathcal{N}(x, \sigma^2) = 1/\sigma(2\pi)^{1/2} \exp(-x^2/2\sigma^2)$ represents a normal distribution with variance σ^2, and σ_P and σ_{PH_j} are the uncertainties in measurement of the native and jth derivative structure factors. (6) states that if we knew the values of F_P, φ, R, and the S_j, then the probability that we would measure a value $F^o_{PH_j}$ is normally distributed about F_{PH_j} and F^o_P would be normally distributed about F_P. If we obtain information about distributions for F_P, R and the S_j, we can obtain an estimate of $p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi)$ by integrating (6) over the 'nuisance' variables F_P, R and S_j in a process known as 'marginalization' (Box, 1980). Assuming that R and S_j are independent of the native phase φ, we can write,

$$p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi) \propto \int p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi, F_P, R, S_1 \ldots S_n) \times p_o(F_P)\mathrm{d}F_P\, p_o(R)\mathrm{d}R \prod_j p_o(S_j)\mathrm{d}S_j, \quad (7)$$

where $p_o(F_P)$, $p_o(R)$ and $p_o(S_j)$ are estimates of the prior probability distributions for F_P, R and S_j and the integrations are over all possible values of these variables. We will assume that the native structure-factor amplitude F_P is measured with sufficient accuracy that the prior probability distribution $p_o(F_p)$ does not contribute a substantial amount of additional information and may be ignored.

2.3. *Prior probability distributions*

We now make estimates of the prior distributions $p_o(R)$ and $p_o(S_1) \ldots p_o(S_n)$. R is the component along the direction of the native structure factor of the correlated portions of the errors from lack of isomorphism, modeling, and other sources. We assume that $F^c_{PH_j}$, δ_P, R, and S_j are independent of each other in the sense that the value of any of their products averaged over many reflections would be zero. This assumption implies that the probability distributions that govern the magnitudes of δ_P, R and the S_j will be independent of each other. So long as the previous assumption about δ_P, R and S_j being small relative to F_P holds, this should not be a problem. However, as extensively discussed by Read in a related context, the part of the errors present because of lack of isomorphism are not truly independent from the native structure factor (Read, 1986). Moreover, while the assumption of independence is reasonable if the errors in the heavy-atom model are due to heavy-atom sites not included in the model at all, it will be a poor assumption if the occupancies of heavy-atom sites are overestimated. In the latter case, the component of R due to the heavy-atom model error will be negatively correlated with $F^c_{PH_j}$.

We have argued before (Terwilliger & Eisenberg, 1987) that as long as the structure factor **R** is due to scattering or changes in scattering at a number of locations in the unit cell of the derivative crystals, its prior probability distribution can be quite reasonably described by Wilson statistics (Wilson, 1949). In this case the component R along the direction of the native structure factor will have a normal prior probability distribution with a variance dependent on the resolution of the reflection. We can write that

$$p_o(R) = \mathcal{N}(R, \alpha E^2), \quad (8)$$

where α is equal to the expected intensity factor (Stewart & Karle, 1976) for centric reflections and half this value for acentric reflections (Terwilliger & Eisenberg, 1987), and E^2 is a measure of the total correlated error.

A very similar analysis may be applied to the variable S_j, representing errors unique to the derivative j. Assuming again a normal distribution of errors and that the mean-square amplitudes of these errors for this derivative are given by αA_j^2, this leads to the prior probability distribution for S_j of,

$$p_o(S_j) = \mathcal{N}(S_j, \alpha A_j^2). \tag{9}$$

We can estimate the correlated error E^2 and the uncorrelated errors A_j^2 for each derivative using a method similar to the one we previously developed for estimation of errors for single isomorphous replacement (Terwilliger & Eisenberg, 1987). From (4), if we knew the native phase, φ, we could use the part of $F^o_{PH_j} - F^c_{PH_j}(\varphi)$ that is correlated between any two derivatives j and k to estimate the mean-square value of the correlated error Δ in a range of resolution, because the mean-square value of S_jS_k is zero if $j \neq k$. Referring to (4), this leads to an estimate of Δ^2 given by

$$\Delta^2 \simeq \langle [F^o_{PH_j} - F^c_{PH_j}(\varphi)][F^o_{PH_k} - F^c_{PH_k}(\varphi)] \rangle, \tag{10}$$

where centric and acentric reflections are treated separately, α is as defined above, and the averages are taken over reflections in a range of resolution. Further, noting that $\Delta = R - \delta$ and that $\langle R^2 \rangle = \alpha E^2$ and $\langle \delta^2 \rangle = \sigma_P^2$, we can write that $\langle \Delta^2 \rangle = \alpha E^2 + \sigma_P^2$. We do not know the value of φ in (10), so our best estimate of $\langle \Delta^2 \rangle$ for each reflection is obtained by averaging over all values of φ, weighted by the probability of obtaining φ, to be developed below. This yields,

$$E^2 \simeq \langle 1/\alpha \int [F^o_{PH_j} - F^c_{PH_j}(\varphi)][F^o_{PH_k} - F^c_{PH_k}(\varphi)] \times p(\varphi) \mathrm{d}\varphi \rangle - \langle \sigma_P^2/\alpha \rangle. \tag{11}$$

Because each pair of derivatives gives one estimate of E^2 and because the extent of correlation may vary between pairs, we choose to use the minimum value of (11) for any pair of derivatives as our estimate of E^2. A similar argument leads to the relation

$$\langle [F^o_{PH_j} - F^c_{PH_j}(\varphi)]^2 \rangle \simeq \langle \Delta^2 + S_j^2 + \delta^2_{PH_j} \rangle \simeq \alpha E^2 + \alpha A_j^2 + \langle \sigma^2_{PH_j} \rangle + \langle \sigma_P^2 \rangle, \tag{12}$$

and an estimate for A_j^2 of

$$A_j^2 = \langle 1/\alpha \int [F^o_{PH_j} - F^c_{PH_j}(\varphi)]^2 p_o(\varphi) \mathrm{d}\varphi \rangle - E^2 - \langle \sigma^2_{PH_j}/\alpha \rangle - \langle \sigma_P^2/\alpha \rangle. \tag{13}$$

2.4. *The correlated phasing equation*

Substituting (6), (8) and (9) into (7), and using (3) and the relation $\delta = F^o_P - F_P$ to replace F_{PH_j} with $F^c_{PH_j} + (F^o_P - F_P) + S_j$, we obtain

$$\begin{aligned} p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi) &\propto \int \mathcal{N}(F^o_P - F_P, \sigma_P^2)\mathrm{d}F_P \int \mathcal{N}(R, \alpha E^2)\mathrm{d}R \\ &\times \prod_j \int \mathcal{N}(F^o_{PH_j} - [F^c_{PH_j} + R + (F^o_P - F_P) + S_j], \sigma^2_{PH_j}) \\ &\times \mathcal{N}(S_j, \alpha A_j^2)\mathrm{d}S_j. \end{aligned} \tag{14}$$

The integrations over the S_j can be carried out independently, leading to

$$\begin{aligned} p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi) &\propto \int \mathcal{N}(F^o_P - F_P, \sigma_P^2)\mathrm{d}F_P \int \mathcal{N}(R, \alpha E^2)\mathrm{d}R \\ &\times \prod_j \mathcal{N}\{F^o_{PH_j} - [F^c_{PH_j} + R + (F^o_P - F_P)], \sigma^2_{PH_j} + \alpha A_j^2\}. \end{aligned} \tag{15}$$

Substituting $\Delta = R + (F^o_P - F_P)$ and noting that in the integration over R, F_P is fixed so that $\mathrm{d}\Delta = \mathrm{d}R$, this can be rewritten as

$$\begin{aligned} p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi) &\propto \int \mathcal{N}(F^o_P - F_P, \sigma_P^2)\mathrm{d}F_P \\ &\times \int \mathcal{N}(\Delta - [F^o_P - F_P], \alpha E^2)\mathrm{d}\Delta \\ &\times \prod_j \mathcal{N}(F^o_{PH_j} - [F^c_{PH_j} + \Delta], \sigma^2_{PH_j}) + \alpha A_j^2). \end{aligned} \tag{16}$$

Reversing the order of integration and integrating over F_P leads to,

$$\begin{aligned} p(F^o_P, F^o_{PH_1} \ldots F^o_{PH_n}|\varphi) &\propto \int \mathcal{N}(\Delta, \alpha E^2 + \sigma_P^2)\mathrm{d}\Delta \\ &\times \prod_j \mathcal{N}(F^o_{PH_j} - [F^c_{PH_j} + \Delta], \sigma^2_{PH_j} + \alpha A_j^2). \end{aligned} \tag{17}$$

Finally, integration over Δ and substitution of the result into (5), yields the correlated phasing equation,

$$\begin{aligned} p(\varphi) \propto p_o(\varphi) \exp -\tfrac{1}{2} \Bigg(&\sum_j \{[F^o_{PH_j} - F^c_{PH_j}(\varphi)]^2/(\sigma^2_{PH_j} + A_j^2)\} \\ &- \left\{ \sum_j [F^o_{PH_j} - F^c_{PH_j}(\varphi)]/(\sigma^2_{PH_j} + A_j^2) \right\}^2 / [1/(E^2 + \sigma_P^2)] \\ &+ \sum_j 1/(\sigma^2_{PH_j} + A_j^2) \Bigg). \end{aligned} \tag{18}$$

The first term in the exponent corresponds to Blow–Crick phasing, that is, phasing based on independent derivatives. The second term accounts for the correlation between errors in the derivatives. Note that if $E^2 + \sigma_P^2 = 0$ – if there is no correlated error – then the second term will equal zero.

The correlated-phasing method described here can be thought of in much the same way as difference refinement, a method first used by Fermi, Perutz, Dickinson & Chien (1982) and recently examined in detail by us (Terwilliger & Berendzen, 1995), if the 'model error' terms of difference refinement are

replaced by the 'correlated errors' of correlated phasing. In Blow–Crick phasing, each derivative is used independently in phasing. In correlated phasing, the difference between the observed amplitude of a structure factor for one derivative and the corresponding calculated amplitude is used as an estimate of the correlated error for that reflection. This estimate of the correlated error is then subtracted from the measured amplitude of a structure factor for other derivatives that share correlated errors. The 'corrected' amplitudes for these other derivatives then can be used with the native amplitude to form a more accurate phasing estimate than could be obtained with independent phasing. Of course, the correlated phasing formulation does all this at once, not in sequential subtraction steps.

3. Evaluation of correlated phasing using test data

We constructed model data to find out in what circumstances correlated phasing is useful. We examined how high the correlation among errors in the derivatives must be before correlated phasing has a substantial effect, and we examined the use of correlated phasing in cases where there were substantial errors in the measurement of the amplitude of the native structure factor. In each case, correlated phasing was compared with Blow–Crick (independent) phasing using the same heavy-atom parameters.

Model data were constructed based on a 'native' peptide structure with 51 atoms (in seven residues) in space group *P*222. Known model native structure factors were calculated from this structure, and 'measured' native structure factors were obtained from these by additional normally distributed random variable to simulate a measurement error of 5%, except as noted below. Derivative structure factors were constructed by adding three additional terms to the native structure factors. The first was a structure factor of a heavy-atom partial structure with one heavy-atom site in the asymmetric unit of each derivative. The second and third were terms representing non-isomorphism between native and derivatives that was either correlated or not correlated among derivatives. Each of the non-isomorphism terms were two-dimensional normal distributions for acentric reflections and one-dimensional normal distributions for centric reflections (Wilson, 1949). The correlation of errors among the derivatives was adjusted by varying the r.m.s. values of these non-isomorphism terms. Three derivative data sets were used in each case.

The model data sets were analyzed with the *HEAVY* package of programs using origin-removed difference Patterson refinement of heavy-atom parameters (Terwilliger & Eisenberg, 1983) and either Blow–Crick phasing or correlated phasing. The same set of heavy-atom parameters and scaling factors was used for each phasing method. The resulting phases were compared to the 'true' native phases used to generate the model data sets, and a map using these phases and their figures of merit were calculated and evaluated at the positions of atoms in the 'true' native structure.

Fig. 1 compares Blow–Crick and correlated-phasing applications for a series of model data sets where the r.m.s. total lack-of-isomorphism error for each derivative was fixed at 20% of the r.m.s. native amplitude, and where the correlated part of these errors was varied from 0 to 100% of the total. When the mean-square correlated errors were less than about 50% of the total mean-square error, correlated phasing yielded only slight improvement over Blow–Crick phasing in accuracy of native

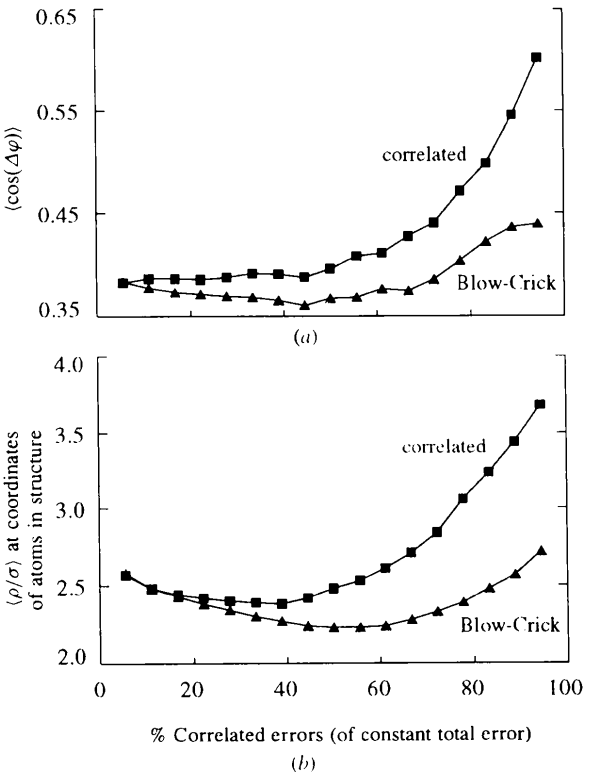


Fig. 1. Correlated and Blow–Crick phasing with constant total errors and varying correlated errors. Data sets consisting of a native and three derivatives were analyzed as described in the text. Native and derivative errors in measurement were 5%. The r.m.s. total lack-of-isomorphism error for the amplitudes of structure factors for each derivative was 20% of the r.m.s. amplitude of the native structure factor. The correlated lack of isomorphism was varied from 0 to 100% of the total lack-of-isomorphism error. The abscissa is the percentage of mean-square errors in amplitudes of structure factors that are correlated among the three derivatives. (*a*) Agreement between phases calculated with Blow–Crick (triangles) and correlated (squares) phasing. The value of the mean effective figure of merit of the map, $\langle\cos(\Delta\varphi)\rangle$, where Δ is the error in native phase, is shown. (*b*) Mean value of electron-density maps at positions of atoms in the native structure calculated using Blow–Crick (triangles) and correlated (squares) phasing. The electron-density values are normalized to the r.m.s. electron-density value of the maps averaged over the asymmetric unit.

phases and the quality of the resulting electron-density map. When the correlated errors were above 80% of the total, however, correlated phasing resulted in a marked improvement in phase accuracy and quality of the map. The mean effective figure of merit of the map was improved from 0.44 to 0.60 and the mean value of ρ/σ at coordinates of atoms in the structure increased from 2.7 to 3.7σ using correlated phasing when the correlated errors were 95% of the total, for example. In contrast, as the fraction of correlated error increases, the quality of phasing obtained using Blow-Crick phasing holds fairly constant. Correlated phasing takes advantage of the correlation of errors so that the effective figure of merit and mean ρ/σ of the map dramatically increase when the errors are highly correlated, even though the same total error is still present.

Fig. 2 further examines when correlated phasing might be useful. It shows that if all derivatives in a data set have completely correlated lack-of-isomorphism errors, both correlated and Blow-Crick phasing methods yield progressively poorer phasing estimates as this error is increased, but the worsening of phasing is far less using corelated phasing. As in Fig. 1, correlated phasing yields the most improvement over Blow-Crick phasing when the correlated errors are large.

Correlated phasing can also yield substantial improvements over Blow-Crick phasing in cases where the errors in the native amplitudes of structure factors are very large (Einstein, 1977). Fig. 3 illustrates a case with three derivative data sets where the native amplitude is measured with errors varying from 2 to 16%. Because the native amplitude is used with all three derivatives in phase calculation, errors in measurement of the native amplitude are correlated. Increases in measurement errors decrease the phasing quality using either method, but the decrease is far smaller using correlated phasing than with Blow-Crick phasing. Although it would be out of the ordinary to attempt to calculate phases using a native data set with a measurement error of 16%, many of the weaker data in an ordinary data set will have measurement errors of this size. Consequently these weak

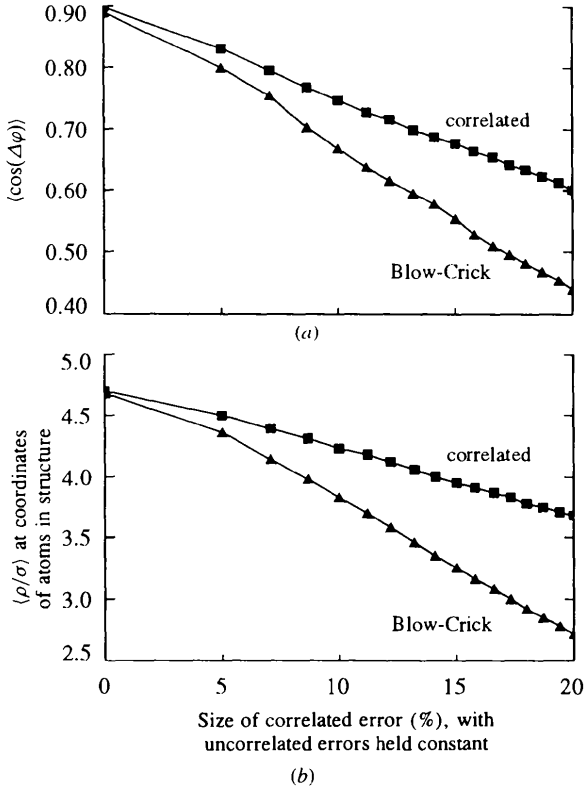


Fig. 2. Blow-Crick and correlated phasing with varying correlated errors. Data sets similar to those described in the legend to Fig. 1 were constructed, except that the lack-of-isomorphism error was entirely correlated among all derivatives, and the r.m.s. value of this error was varied from 0 to 20% of the r.m.s. native amplitude. Panels (*a*) and (*b*) are as in Fig. 1.

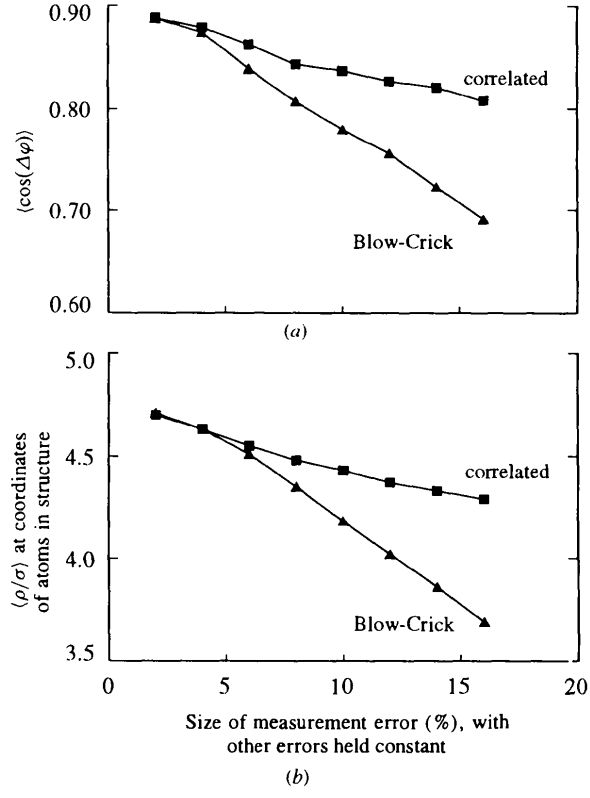


Fig. 3. Blow-Crick and correlated phasing with varying errors in measurement of amplitudes of native structure factors. Data sets were constructed as in Figs. 1 and 2, except that the error in measurement of the native amplitude was varied from 0 to 16%. The measurement errors for each derivative amplitude was 5%, and each derivative had an uncorrelated 2% lack-of-isomorphism error.

reflections could be analyzed more accurately using correlated phasing than using Blow–Crick phasing.

4. Evaluation of correlated phasing using Trp–RS data

Doublié, Xiang, Gilmore, Bricogne & Carter (1994) recently described a very difficult determination of the structure of tryptophanyl tRNA synthetase (Trp-RS) from *Bacillus stearothermophilus*. This structure determination was difficult in large part because the three derivatives used for a key stage in phasing were exceptionally non-isomorphous to the native. Data from a selenomethionine-containing derivative that was isomorphous to the native had been collected and was potentially useful for phasing, but the positions of the selenium atoms could not be identified with the MIR phases obtained from the three non-isomorphous derivatives. The deficiency of the MIR phases obtained with these derivatives was overcome only by applying a phase permutation and likelihood scoring procedure along with maximum-entropy solvent flattening after MIR phasing had been carried out. Although the three derivatives used were not isomorphous to the native, they were relatively isomorphous to each other. The R factors comparing each of the three non-isomorphous derivatives to the native were from 41 to 42%, while those between the derivatives ranged from 16 to 29%. This indicated that the lack-of-isomorphism errors for the three non-isomorphous derivatives were highly correlated and suggested that correlated phasing might improve the accuracy of the phases from this experiment.

The MIR phasing using the three non-isomorphous derivatives was originally carried out using maximum-likelihood heavy-atom refinement procedures implemented in the program *MLPHARE* (Otwinowski, 1991), and including anomalous differences for all three derivatives (Doublié *et al.*, 1994). The resulting phases were used to calculate a difference Fourier synthesis for the positions of selenium atoms in the selenomethionine derivative using coefficients of $(F_{Se} - F_{nat})$, and subsequently these phases were used as the input for phase permutation and maximum-entropy solvent flattening. To compare correlated and Blow–Crick phasing methods directly using this as a test case, heavy-atom parameters for the three derivatives were re-refined by origin-removed difference Patterson refinement (Terwilliger & Eisenberg, 1983), and the newly refined parameters were used for either correlated or Blow–Crick phasing of difference Fourier syntheses based again on $(F_{Se} - F_{nat})$. Because the structure of TrpRS is now solved, the positions of the Se atoms in the selenomethionine derivative are known and the difference electron density at these positions could be used as a measure of the quality of the phasing. The mean value of ρ/σ (electron density at positions of Se atoms, normalized to the r.m.s. value of the map) using Blow–Crick phasing was 4.3. Using correlated phasing, the mean ρ/σ was increased to 5.3 and the difference electron density at each of the ten selenium positions was improved. In the original structure determination using maximum-likelihood heavy-atom refinement, the mean value of ρ/σ was only 3.2 (Doublié, *et al.*, 1994), probably due to the difficulties of using phase refinement of heavy-atom parameters in the presence of extreme non-isomorphism.

A second indication of the quality of MIR phases is the presence or absence of 'ghost' peaks at the locations of heavy-atom sites in these difference Fourier syntheses. If the phases are of high quality, these ghost peaks should be small or not present, while they may be very substantial if the phasing is poor. The mean value of ρ/σ at these ghost heavy-atom sites was reduced from 8.9 to just 3.8 by using correlated phasing. The effects of correlated phasing of the Trp-RS structure can be seen in another way in Fig. 4, which shows a portion of these difference Fourier syntheses in a region that contains four selenium locations and four of the six heavy-atom sites. Fig. 4(*a*) illustrates the difference Fourier synthesis obtained with Blow–Crick phasing. It would be difficult to identify the selenium sites even knowing that the very large peaks are simply 'ghost' peaks at the heavy-atom sites. In contrast, using correlated phasing (Fig. 4*b*), the ghost peaks are almost eliminated and the locations of the four Se atoms are clear.

Overall, Figs. 1–4 demonstrate that in cases where errors are highly correlated among derivatives in an MIR experiment, correlated phasing can result in a dramatic improvement in the quality of phases obtained. This improvement is possible because of the phasing information contained in the differences among the derivative amplitudes of structure factors that is used in correlated phasing but not in Blow–Crick phasing.

5. Conclusions

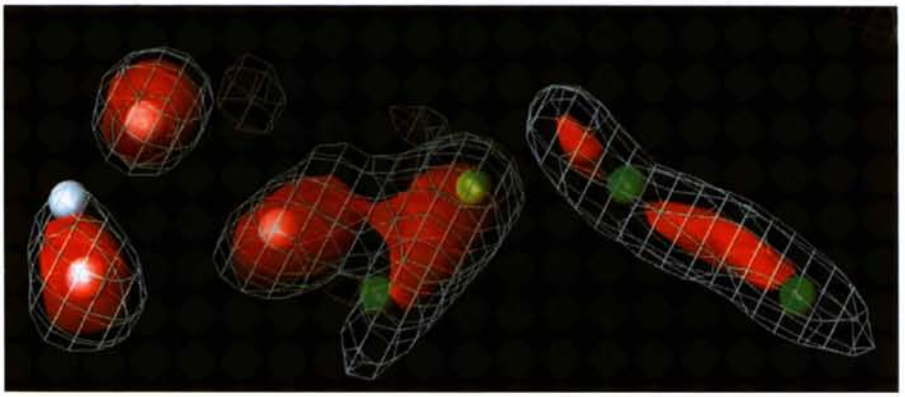
Our correlated phasing strategy is based on the fact that the errors for the various derivatives in a multiple-isomorphous replacement experiment are sometimes highly correlated. When errors are correlated in this way, the errors in the differences among derivatives can be substantially smaller than the errors for any one derivative. Correlated phasing is a way of using the phase information contained in the differences among derivatives as well as the usual information based on differences between each derivative and the native structure factors.

Correlated phasing will be an important tool for the analysis of multiple isomorphous replacement X-ray diffraction data where a substantial correlation of errors exists among the derivative data sets.

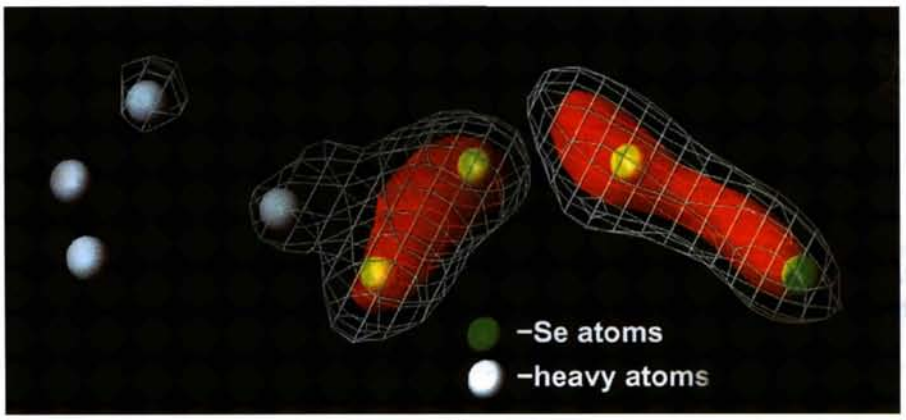
There are a number of situations where such a correlation of errors might arise. One is when correlated non-isomorphism exists within a group of derivatives. In such a case, each of a group of derivatives of a macromolecule changes conformation in the same way when the derivatives are formed. This change could be movement of a domain of the protein or movement of subunits relative to each other. It could also simply be a rotation of the entire protein relative to the crystal lattice, or a change in the dimensions or angles of the crystal lattice. As long as whatever changes occur are quite similar in each derivative, correlated phasing is likely to substantially improve the quality of phasing. Another situation where correlated errors can arise is if there are large errors in the native data, or if there are large and correlated errors in scaling of derivative data sets to the native. Similarly, correlated errors can be present if a group of low-occupancy heavy-atom sites are present in more than one derivative and are not included in the heavy-atom models for these derivatives, or if data for a particular derivative are measured more than once and the duplicate derivatives are included in phasing.

In all of these cases, the key factor determining the improvement in phasing that can be expected using correlated phasing is the extent of correlation of errors. If the errors are highly correlated, correlated phasing will improve the quality of phasing, but if errors are not substantially correlated, it will have little effect because there is little additional information contained in the differences among derivatives. It should be noted that if some derivatives in a MIR experiment have highly correlated errors and others do not, then the derivatives with correlated errors can be grouped together for correlated phasing, and the phase probabilities obtained from this correlated phasing group can be combined by simple multiplication with the essentially independent probabilities obtained from the uncorrelated derivatives.

The authors would like to thank C. Carter for generously providing Trp-RS diffraction and heavy-atom data as a test case for correlated phasing. The authors are also grateful for support from the National Institutes of Health, from the International Human Frontiers Organization, and from the Laboratory Directed Research and Development



(*a*)



(*b*)

Fig. 4. ($F_{Se} - F_{nat}$) difference Fourier syntheses for Trp–RS calculated using (*a*) Blow–Crick (1959) or (*b*) correlated phasing. The region of the difference Fourier maps surrounding the Se atoms in selenomethionine residues 92, 314, 318 and 322 is shown. Contours are at 3σ (grey net contours) and 4.5σ (solid red contours). Positions of Se atoms are indicated by yellow spheres if within the highest contour region or green spheres otherwise. Positions of heavy-atom sites used in phasing are indicated by white spheres.

program of Los Alamos National Laboratory. Correlated phasing has been implemented in version 4 of the package *HEAVY*, available from TT to whom enquiries may be directed at terwilliger@lanl.gov.

References

Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12,** 794–802.

Box, G. E. P. (1980). *J. R. Statist. Soc. A,* **H143,** 383–430.

Box, G. E. P. & Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis,* New York: John Wiley.

Doublié, S., Xiang, S., Gilmore, C. J., Bricogne, G. & Carter, C. W. Jr (1994). *Acta Cryst.* A**50,** 164–182.

Einstein, R. J. (1977). *Acta Cryst.* A**33,** 75–85.

Fermi, G., Perutz, M. F., Dickinson, L. C. & Chien, J. C. W. (1982). *J. Mol. Biol.* **155,** 495–505.

Hendrickson, W. A. (1991). *Science,* **254,** 51–58.

Karle, J. (1980). *Int. J. Quantum Chem.* **7,** 357–367.

Miller, R., DeTitta, G. T., Jones, R., Langs, D. A., Weeks, C. M. & Hauptman, H. A. (1993). *Science,* **259,** 1430–1433.

Otwinowski, Z. (1991). In *Isomorphous Replacement and Anomalous Scattering: Proceedings of the CCP4 Study Weekend 25–26 January 1991*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie. Warrington: Daresbury Laboratory.

Read, R. J. (1986). *Acta Cryst.* A**42,** 140–149.

Stewart, J. M. & Karle, J. (1976). *Acta Cryst.* A**32,** 1005–1007.

Terwilliger, T. C. & Berendzen, J. (1995) *Acta Cryst.* D**51,** 609–618.

Terwilliger, T. C. & Eisenberg, D. S. (1983). *Acta Cryst.* A**39,** 813–817.

Terwilliger, T. C. & Eisenberg, D. S. (1987). *Acta Cryst.* A**43,** 6–13.

Wilson, A. J. C. (1949). *Acta Cryst.* **2,** 318–321.