Acta Cryst. (1997). D53, 571–579

Bayesian Correlated MAD Phasing

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. E-mail: terwilliger@lanl.gov*

(Received 22 October 1996; accepted 8 April 1997)

Abstract

A Bayesian treatment for phase calculation in the multiwavelength anomalous diffraction (MAD) technique is presented. This approach explicitly treats effects of errors correlated among measurements at different wavelengths and between Bijvoet pairs. The resulting method, which is called Bayesian correlated MAD phasing, gives proper statistical consideration to all data and does not give special treatment to data from a particular wavelength. Results obtained using Bayesian correlated MAD phasing and two other strategies on both a model test case and on data obtained in two actual MAD experiments are compared. Although all procedures performed well when the completeness of the data was high, it is shown that Bayesian correlated MAD phasing is more robust with respect to incompleteness of data than the other methods are. At 60% completeness the improvement over other methods for the examples given was nearly 50% in the correlation coefficients, and made a substantial difference in the interpretability of an electron-density map.

1. MAD data and its analysis

In the multiwavelength anomalous diffraction (MAD) method, crystallographic phases are estimated from the wavelength dependence of diffracted intensities when an X-ray beam from a tunable source is stepped over an absorption edge of some heavy atoms present in small numbers in the asymmetric unit of a crystal (Karle, 1980). MAD experiments measure amplitudes of Bijvoet pairs F^+ and F^- for a crystal at two or more wavelengths chosen so that the f' values for the heavy atoms vary as much as possible among wavelengths and so that the f'' values are as large as possible (Hendrickson, 1985). As in a multiple isomorphous replacement (MIR) experiment, the locations of heavy atoms within the asymmetric unit are generally obtained using a Patterson or a difference Patterson synthesis, and parameters describing these heavy atoms are refined. Estimates of heavy-atom structure factors calculated from the heavy-atom model at each wavelength, together with the observed F^+ and F^-, form the basis for MAD phase calculations.

A major strength of MAD lies in the perfect isomorphism of a given crystal at different wavelengths, in contrast to MIR in which the heavy-atom derivatives are frequently insufficiently isomorphous with the native structure to be usable. MAD has proven exceptionally useful for phasing macromolecular structures (Hendrickson, 1991) and in a number of laboratories has become the technique of first choice (*e.g.* Hendrickson *et al.*, 1989; Ramakrishnan, Finch, Graziano, Lee & Sweet, 1993; Leahy, Aukhil & Erickson, 1996; Peat *et al.*, 1996).

Although MAD phasing is conceptually somewhat like MIR, analysis of MAD data is less straightforward and a number of approaches to MAD phasing have been proposed and used. One widely used approach (Karle, 1980; Hendrickson, 1985) combines an estimate from the measured data of the phase difference $\Delta\varphi$ (between the overall structure factor for a reflection and the heavy-atom structure factor for that reflection) with an estimate of the heavy-atom structure factor calculated from the heavy-atom model to give a phase estimate for the overall structure factor. While this approach has the advantage of treating all data on an equal basis, it is not ideal in that the information embodied in the heavy-atom model is not directly used in estimating $\Delta\varphi$. Some related approaches make use of the heavy-atom model in a manner analogous to MIR phasing (Pahler, Smith & Hendrickson, 1990; Burling, Weis, Flaherty & Brünger, 1996).

Another approach treats the data measured at one wavelength as a 'native' data set and data at the other wavelengths as isomorphous derivatives just as in the MIR method (Ramakrishnan *et al.*, 1993; Ramakrishnan & Biou, 1996). Heavy-atom structure factors for comparison of 'native' and 'derivative' data sets are then calculated using the differences in heavy-atom scattering factors at the wavelengths used from the 'native' and derivative data sets. Although this approach is simple and fairly robust, it has the disadvantage that data at the 'native' wavelength is treated in a special way. If data are missing at that wavelength then no

phasing information is available for the missing reflections.

Yet another approach to MAD phasing is one in which all the measured MAD data are converted to a form identical to single isomorphous replacement with anomalous scattering (SIRAS) (Terwilliger, 1994*b*, 1996). No matter how many wavelengths data are collected at, the information obtained in a MAD experiment consists essentially of three independent numbers: an anomalous difference that (if small enough) scales between wavelengths according to the value of f'' at each wavelength, a dispersive difference that scales between pairs of wavelengths according to the differences in values of f', and a value of the structure factor corresponding to all atoms but the heavy atoms in the structure. These three values are readily calculated from the original MAD data and can be used to construct a pseudo-native data set and a pseudo-derivative with anomalous scattering data set that can be analyzed with conventional SIRAS approaches (*e.g.* Blundell & Johnson, 1976). This pseudo-SIRAS approach has the advantage that all the observed MAD data are treated equally, but the disadvantage that this approach depends on a synthetic data set derived from the observed MAD data, a scheme for which statistical evaluation of the most probable phase is complicated.

Although all the approaches to MAD phasing described above work quite well, none makes full use of the heavy-atom model and at the same time takes into account the correlation of errors. All methods in common use thus far treat errors among measurements at different wavelengths as if they were independent, even though in practice errors frequently are significantly correlated due to errors in the heavy-atom model. Bayesian approaches (Box & Tiao, 1973; Box, 1980) to this problem are particularly attractive because through them information about possible sources of error can be explicitly described and used to improve phase estimates. Recently, Bayesian maximum likelihood refinement of heavy-atom parameters has proven very useful for macromolecular structure determination (Otwinowski, 1991), and Bayesian approaches to analysis of MIR data with correlated errors have been developed (Terwilliger & Berendzen, 1996). A very general Bayesian framework on which analysis of MAD data can be hung has been presented (Chiadmi, Kahn, de la Fortelle & Fourme, 1993). This has recently been implanted in the program *SHARP* (de la Fortelle & Bricogne, 1997) and applied to MAD structure determination (Boissy *et al.*, 1996).

In the present work, we employ a Bayesian approach to the specific problem of calculating from MAD data while taking into account correlated errors between measurements at different wavelengths and between Bijvoet pairs at the same wavelength. The resulting analysis approach, which we call 'Bayesian correlated MAD phasing', has advantages over the approaches mentioned above because it treats all the data on an equal basis, because it takes the heavy-atom model and its correlated errors into full consideration, and because it uses the measured data directly (*e.g.*, without conversion to pseudo-SIRAS data).

Table 1. *Selenium scattering factors*

System	Wavelength (Å)	f'	f''
Model	0.9798	−9.8	2.9
	0.9794	−8.6	4.9
	0.9000	−1.6	3.3
IF3-C	0.9802	−9.5	3.2
	0.9795	−7.4	5.8
	0.9150	−1.7	3.4
GVP	0.9798	−9.6	2.2
	0.9794	−7.7	5.8
	0.9000	−1.6	3.4

2. Bayesian estimation of MAD phases

The approach we will use here is closely related to our approach to MIR phasing in the presence of correlated non-isomorphism (Terwilliger & Berendzen, 1996). The key element in both is the recognition that the analyses of measured structure factors (whether related as for different derivatives, Bijvoet pairs, or for different wavelengths) are not independent. In MIR, this effect can arise from non-isomorphism that is shared among derivatives; errors in estimating derivative structure factors based on native structure factors and the heavy-atom model will then be correlated among derivatives. In severe cases of correlated error, proper treatment of this correlated error can dramatically improve the accuracy of phase calculations. Similarly, in a MAD experiment, errors in the heavy-atom model will be shared between all Bijvoet pairs and among all wavelengths. Such correlated errors can arise not only through failure to identify and model all heavy-atom binding sites, but also through the more common problem of inaccuracy in estimating heavy-atom coordinates and thermal parameters. Although we do not treat them explicitly here, other contributions to correlated error can include the effects of radiation damage in experiments where different wavelengths are collected in separate sweeps, and inaccuracies in scaling. In MAD phasing methods that involve intermediate estimates of structure factors without heavy-atom contributions (*e.g.* Pahler *et al.*, 1990; Burling *et al.*, 1996), errors in calculated synthetic structure factors will lead to correlated errors at all wavelengths, too.

We wish to obtain, for a particular reflection in a MAD experiment, a probability distribution for the complex structure factor $\mathbf{F}_k$ corresponding to all the atoms in the structure of interest at a given wavelength

λ_k in the presence of errors that may be correlated among wavelengths or between Bijvoet pairs. λ_k need not be a wavelength at which a measurement took place; it is merely a bookeeping device, a reference wavelength with particular values of f, f', and f'' for the anomalously scattering atoms to which all measurements and calculations will be referred. A useful reference 'wavelength' is one for which f' and f'' are defined to be equal to zero, that it, a 'wavelength' at which only normal scattering and no anomalous scattering occurs. In this case the Fourier synthesis obtained using the structure factors $\mathbf{F_k}$ will be an electron-density map corresponding to all the atoms in the structure. These $\mathbf{F_k}$ are equivalent to the ${}^0\mathbf{F}_T(\mathbf{h})$ in the MAD analysis approach described by Hendrickson (Hendrickson, 1991). We begin by describing the F^+ and F^- structure factors at each wavelength in terms of a structure-factor amplitude and phase at λ_k (F_k and θ), the heavy-atom structure factor calculated from a model, and correlated and non-correlated sources of error. We will then integrate over possible values of the error terms to obtain expressions for probability distributions governing F_k and θ. As in our previous treatment of MIR data, the calculations will be along the lines of the Blow–Crick formulation, and we will approximate many of the component probability distributions and complex sums to first order (*e.g.*, by normal distributions). It is important to note that only the error terms are approximated in this fashion. This means that the analysis is equally applicable to cases in which anomalous scattering effects are large or small. It only becomes inapplicable in cases where the errors in estimates of scattering from the heavy-atom model are large relative to the total scattering factors, a case that is rare in MAD phasing.

2.1. *Correlated and uncorrelated errors*

X-ray diffraction from the non-anomalously scattering atoms in a protein crystal can be described for a particular reflection by a complex structure factor $\mathbf{F}_N$ For a MAD experiment, we will have to consider a set of structure factors for each reflection, since measurements are typically made for both F^+ and F^- at each of several wavelengths. We can write the complex structure factor corresponding to F^+ at the measured wavelength λ_j, denoted by $\mathbf{F_j^+}$, as

$$\mathbf{F_j^+} = \mathbf{F}_N + [\mathbf{F_{H_j}^{+c}} + \mathbf{R}^+(\lambda_j) + \mathbf{S_j^+}]. \tag{1}$$

The second term on the right, $\mathbf{F_{H_j}^{+c}}$, is the calculated structure factor of the heavy atoms at λ_j, which includes the effects of including anomalous scattering. This calculation is based on the current model, which describes the heavy-atom positions, occupancies, and Debye–Waller factors. The model is presumed to contain some errors, which we describe by the third and fourth terms, $\mathbf{R}^+(\lambda_j) + \mathbf{S_j^+}$. $\mathbf{R}^+(\lambda_j)$ corresponds to errors in the heavy-atom model itself and as such is assumed to be correlated across all wavelengths and to have a simple wavelength dependence determined by the nature of the anomalously scattering atoms in the structure. The last term, $\mathbf{S_j^+}$, represents any error in $\mathbf{F_{H_j}^{+c}}$ that is specific to wavelength λ_j. The sum $\mathbf{R}^+(\lambda_j) + \mathbf{S_j^+}$ accounts for all errors arising from inadequacies of the model, whether arising from errors in the heavy-atom model or other sources. A similar expression can be written for $\mathbf{F_j^-}$. Note that we have not yet included experimental errors. Application of (1) again allows us to express $\mathbf{F_j^+}$ in terms of the complex structure factor at the reference wavelength λ_k as

$$\begin{aligned}\mathbf{F_j^+} = \mathbf{F_k^+} &+ (\mathbf{F_{H_j}^{+c}} - \mathbf{F_{H_k}^{+c}}) \\ &+ [\mathbf{R}^+(\lambda_j) - \mathbf{R}^+(\lambda_k)] + (\mathbf{S_j^+} - \mathbf{S_k^+}).\end{aligned} \tag{2}$$

Defining $\mathbf{F_k^+} + \mathbf{F_{H_j}^{+c}} - \mathbf{F_{H_k}^{+c}}$ as $\mathbf{F_j^{+c}}$, putting in the explicit wavelength dependence of the $\mathbf{R}^+$'s (the heavy-atom model error terms), which are proportional to $f + f'(\lambda_j) + if''(\lambda_j)$, and defining the reference wavelength λ_k so that both $f'(\lambda_k)$ and $f''(\lambda_k)$ are equal to zero, we obtain

$$\mathbf{F_j^+} = \mathbf{F_j^{+c}} + f_j'\mathbf{R} + if_j''\mathbf{R} + (\mathbf{S_j^+} - \mathbf{S_k^+}), \tag{3}$$

where $\mathbf{R}$ is the normalized error in the non-anomalous part of the heavy-atom scattering factor. An expression for $\mathbf{F_j^-}$ can be written that differs by the substitution of $-i$ for the i in (3). At this point, calculations can be considerably simplified if we allow that the error terms are small compared with $\mathbf{F_j^{+c}}$. This assumption was used in our previous treatment of correlated MIR phasing and will generally be quite good. It will only be a poor assumption if the errors in the model describing the anomalously scattering atoms are exceptionally large and at the same time the scattering from these atoms is exceptionally strong. In the usual case where the error terms are small, the contributions to F_j^+, the amplitude of $\mathbf{F_j^+}$, will be dominated by the components parallel to $\mathbf{F_j^{+c}}$. Introducing new notation we can write F_j^+ as approximately given by

$$F_j^+ \simeq |\mathbf{F_j^{+c}}| + f_j'R' + f_j''R'' + S_j^+, \tag{4}$$

where R' and R'' refer to the components of $\mathbf{R}$ parallel and perpendicular to $\mathbf{F_j^{+c}}$, respectively, and where S_j^+ refers to the component of $\mathbf{S_j^+} - \mathbf{S_k^+}$ parallel to $\mathbf{F_j^{+c}}$. Rewriting $|\mathbf{F_j^{+c}}|$, the calculated amplitude for this reflection at wavelength λ_j, as F_j^{+c}, and expressing the observed wavelength structure factor F_j^{+o} as the sum of F_j^+ and a measurement error, ε_j^+, we obtain

$$F_j^{+o} \simeq F_j^{+c} + f_j'R' + f_j''R'' + S_j^+ + \varepsilon_j^+. \tag{5}$$

The amplitude of the measured structure factor for the F^+ observation at λ_j therefore differs from that calculated based from the structure factor at the

arbitrary wavelength λ_k by two terms with correlations across all wavelengths, $f_j'R'$ and $f_j''R''$, and two terms unique to the jth F^+ observation, S_j^+ and ε_j^+. Generalizing (5) for F^+ and F^- measurements we can write that

$$F_j^o \simeq F_j^c + f_j'R' \pm f_j''R'' + S_j + \varepsilon_j, \qquad (6)$$

where the positive branch is for F^+ and the negative branch is for F^-.

2.2. *Probability distribution for F_k and θ*

To obtain a probability distribution for the structure-factor amplitude and phase at wavelength λ_k (F_k and θ), we begin by using Bayes' rule (Box & Tiao, 1973) to write an expression for the posterior probability distribution for F_k and θ given that we have made measurements $F_1^o \ldots F_n^o$ of the n structure factors are various wavelengths for this reflections,

$$p(F_k, \theta | F_1^o \ldots F_n^o) \propto p(F_1^o \ldots F_n^o | F_k, \theta) p_0(F_k, \theta), \qquad (7)$$

where the measurements $F_1^0 \ldots F_n^o$ can each be either an F^+ or and F^- observation. The prior probability distribution for the structure factor at wavelength λ_k, $p_0(F_k, \theta)$, is usually flat and uninformative because we do not know anything about this structure factor in advance. However, if there is information available from another experiment this probability distribution can reflect this prior information.

We cannot directly calculate the probability distribution $p(F_1^o \ldots F_n^o | F_k, \theta)$ on the right hand side of (7), but using (4) and (5) we can write the related probability distribution $p(F_1^o \ldots F_n^o | F_k, \theta, R', R'', S_1 \ldots S_n)$ assuming that the ε_j are normally distributed,

$$p(F_1^o \ldots F_n^o | F_k, \theta, R', R'', S_1 \ldots S_n) \propto \prod_{j=1,n} \mathcal{N}(F_j^o - F_j, \sigma_j^2), \qquad (8)$$

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 the σ_j are the uncertainties in measurement of F_j^o. (8) states that if we knew the values of F_k, θ, R', R'', and the S_j, then the probability that we would measure a value F_j^o is normally distributed about F_j, calculated from (4) using the F^+ or F^- term as appropriate.

If we obtain information about distributions for R', R'', and the S_j, we can obtain an estimate of $p(F_1^o \ldots F_n^o | F_k, \theta)$ by integrating (8) over the 'nuisance' variables R', R'' and S_j in the process known as 'marginalization' (Box, 1980). Assuming that R', R'' and S_j are independent of F_k and θ, we can write,

$$p(F_1^o \ldots F_n^o | F_k, \theta) \propto \int p(F_1^o \ldots F_n^o | F_k, \theta, R', R'', S_1 \ldots S_n) \times p_0(R')\, \mathrm{d}R' p_0(R'') \prod_{j=1,n} p_0(S_j)\, \mathrm{d}S_j, \qquad (9)$$

where $p_0(R')$, $p_0(R'')$ and $p_0(S_j)$ are estimates of the prior probability distributions for R', R'' and S_j.

2.3. *Prior probability distributions for the errors*

We now make estimates of the prior distributions $p_0(R')$, $p_0(R'')$, and $p_0(S_1) \ldots p_0(S_n)$. We assume as in our treatment of correlated MIR phasing that the probability distributions that govern the magnitudes of R', R'' and the S_j are independent of each other so that the value of any of their products averaged over many reflections would be zero. So long as the previous assumptions that R', R'' and S_j are small relative to F_k hold, this is reasonable. However, while the assumption of independence is good if the errors in the heavy-atom model arise from inaccurate positioning of heavy-atom sites or from underestimates of occupancies of heavy-atom sites, it will be a poorer assumption if the occupancies of heavy-atom sites are overestimated. In this case, the components of R' and R'' from the heavy-atom model error will be negatively correlated with F_j^c.

So long as error in the heavy-atom structure factor **R** arises from scattering or changes in scattering at a number of locations in the unit cell of the crystals, its prior probability distribution can be quite reasonably described by Wilson statistics (Wilson, 1949). In this case the components R' and R'' along the direction of the native structure factor will have normal prior probability distributions with variances. dependent on the resolution of the reflection. Since R' and R'' are orthogonal projections of the same structure factor **R**, their prior probability distributions are identical and given by,

$$p_0(R') = p_0(R'') = \mathcal{N}(|\mathbf{R}|, \alpha E^2), \qquad (10)$$

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 normalized error in the heavy-atom model.

We can estimate the normalized error in the heavy-atom model E^2 using a method similar to the one we previously developed for estimation of errors for single isomorphous replacement and for correlated MIR phasing (Terwilliger & Eisenberg, 1987). From (5) we can write two correlations, each of which is expected to yield a reasonable estimate of E^2 for an appropriate range of resolution if the values of F_k and θ were known exactly. Changing notation slightly to refer to anomalous differences and average amplitudes of structure factors at each wavelength, these are,

$$\langle (\Delta_i^o - \Delta_i^c)(\Delta_j^o - \Delta_j^c)/4\alpha f_i'' f_j'' \rangle \simeq E^2, \qquad (11)$$

$$\langle (\overline{F_i^o} - \overline{F_i^c})(\overline{F_j^o} - \overline{F_j^c})/\alpha f_i' f_j' \rangle \simeq E^2, \qquad (12)$$

where Δ_i^o is the anomalous difference $(F_i^{+o} - F_i^{-o})$, $\overline{F_i^o}$ is the average structure factor $(F^{+o} + F^{-o})/2$, and the angled brackets represent averages over all pairs of measured wavelengths.

We do not know the value of F_k and θ in (11) and (12), so our best estimate of E^2 from each reflection is given by the weighted average of the terms in (11) and (12), integrated over all values of F_k and θ, and weighted by the phase probability to be developed below.

The quantities S_j, representing errors unique to a particular F^+ or F^- measurement at wavelength λ_j, can be analyzed in a similar fashion. Assuming a normal distribution of errors and that the mean-square amplitudes of these errors for this wavelength are given by αA_j^2, this leads to the prior probability distribution for S_j of

$$p_0(S_j) = \mathcal{N}(S_j, \alpha A_j^2). \tag{13}$$

In a MAD experiment, the principal sources of error should ordinarily be instrumental uncertainties in measurement and errors in the heavy-atom model. If these are the only errors, then the values of A_j^2 will all be zero and the S_j can be neglected. In some cases, there may be errors in measurement not reflected in the estimated variances. σ_j^2 and in those cases the appropriate values of A_j^2 may be non-zero. We will assume here that the A_j^2 in a particular shell of resolution all have the same value, A^2, and that this value is usually zero. A method to estimate its value if it is non-zero will be described below.

2.4. *The Bayesian correlated MAD phasing equation*

We are now in a position to calculate the probability distribution for the desired structure factor at wavelength λ_k, $\mathbf{F_k}$. Substituting (8), (10) and (13) into (9), and using (4) to replace F_j, we can write that

$$p(F_1^o \ldots F_n^o | F_k, \theta) \propto \int \mathcal{N}(R', \alpha E^2)\, dR' \mathcal{N}(R'', \alpha E^2)\, dR'' \prod_{j=1,n} \mathcal{N}[F_j^c - (F_j^o + f'R' \pm f''R'' + S_j), \sigma_j^2] \mathcal{N}(S_j, \alpha A_j^2)\, dS_j \tag{14}$$

where the + or − signs are taken for the F^+ and F^- observations, respectively. The integrations over the S_j can each be carried out independently, leading to

$$p(F_1^o \ldots F_n^o | F_k, \theta) \propto \int \mathcal{N}(R', \alpha E^2)\, dR' \prod_{j=1,n} \mathcal{N}(R'', \alpha E^2)\, dR'' \mathcal{N}[F_j^o - (F_j^c + f_j'R' \pm f_j''R''), \sigma_j^2 + \alpha A_j^2]. \tag{15}$$

Finally, noting that R' and R'' are each present in each term of the product in (15), carrying out the integrations over R' and R'', and substituting of the result into (7), yields the Bayesian correlated MAD phasing equation,

$$p(F_k, \theta) \propto p_0(F_k, \theta) \exp(-\chi_B^2/2), \tag{16}$$

where

$$\chi_B^2 = \chi^2 - \frac{C_1^2 D_2 + C_2^2 D_1 - 2C_1 C_2 D_{12}}{D_1 D_2 - D_{12}^2}. \tag{17}$$

The first term in (17), χ^2, is the familiar statistic from the case that there are no correlated errors R' and R'', namely

$$\chi^2 = \sum_{j=1,n} (F_j^o - F_j^c)^2 / \sigma_{B_j}^2, \tag{18}$$

where

$$\sigma_{B_j}^2 = \alpha A_j^2 + \sigma_j^2. \tag{19}$$

The second term in (17) is a correction due to the correlation of errors, whose terms are,

$$C_1 = \sum_{j=1,n} f_j'(F_j^o - F_j^c)/\sigma_{B_j}^2, \tag{20}$$

$$C_2 = \sum_{j=1,n} \pm f_j''(F_j^o - F_j^c)/\sigma_{B_j}^2, \tag{21}$$

$$D_1 = 1/\alpha E^2 + \sum_{j=1,n} f_j'^2/\sigma_{B_j}^2, \tag{22}$$

$$D_2 = 1/\alpha E^2 + \sum_{j=1,n} f_j''^2/\sigma_{B_j}^2, \tag{23}$$

$$D_{12} = \sum_{j=1,n} \pm f_j' f_j''/\sigma_{B_j}^2, \tag{24}$$

and where the + and − choices again correspond to F^+ and F^- measurements, respectively. If there is no correlated heavy-atom error then $E^2 = 0$ and the correction term is zero as well. The Bayesian correlated MAD phasing equations (16)–(24) are similar to the correlated MIR phasing equation (Terwilliger & Berendzen, 1996), except that for MAD phasing there are effectively two correlations to consider, correlations between Bijvoet pairs and correlations among measurements at different wavelengths, instead of a single principal correlation for MIR phasing, the correlated non-isomorphism.

As mentioned above, in most circumstances the values of the A_j^2 will be zero for MAD phasing and $\sigma_{B_j}^2$ becomes simply the experimental uncertainty σ_j^2. In some cases, however, there may be errors in measurement not reflected in the variances σ_j^2. In this case, an overall estimate of A^2 to be applied to all reflections in a range of resolution may be estimated by identifying the value of A^2 that leads to an overall χ^2 value equal to the number of measurements. This is similar to the procedure of renormalizing variances by finding a scale factor that leads to a reduced χ^2 value of approximately unity (Bevington, 1969).

Note that (16)–(24) yield a probability distribution for the vector $\mathbf{F}_k$. This means that a two-dimensional integration over F_k and θ must be carried out to obtain all the available information on this structure factor. The centroid or 'best' electron-density map is then one in which all values of $\mathbf{F}_k$ are considered and are weighted by their relative probabilities. Similarly, the 'best' estimate of F_k is the one in which this quantity is

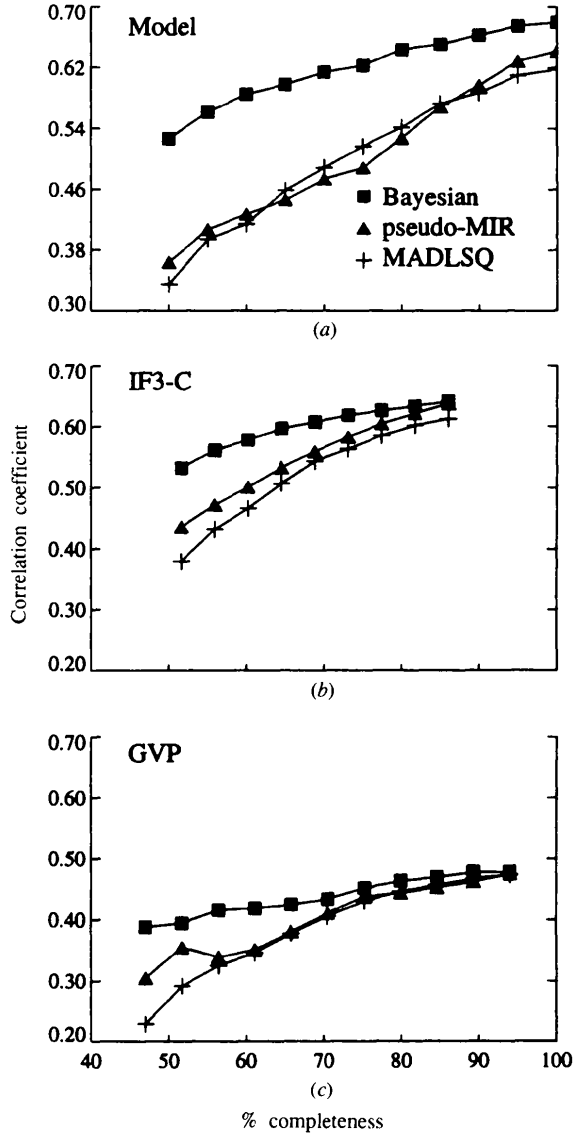


Fig. 1. Analysis of three-wavelength MAD data with varying completeness. The quality of the MAD phasing was evaluated in each case by calculating the correlation coefficient between the 'best' (centroid) electron-density map for each method and a reference electron-density map calculated either from known structure factors for the simulated data set or from the electron-density map calculated from the refined structures of the two proteins. (*a*) A model MAD data set containing 1730 reflections with random phases in space group *C*2 from 3 to 20 Å constructed previously was used as a test data set (Terwilliger, 1994*a*). Simulated MAD data from the model was calculated at three wavelengths as shown in Table 1. Random errors (mean of 5%) were added to the MAD data to simulate measurement errors. The heavy-atom model used throughout consisted of two of the three Se atoms. The heavy-atom parameters were refined using origin-removed difference Patterson refinement after conversion of the full MAD data set to pseudo-SIRAS data using *MADMRG* (Terwilliger, 1994*b*). (*b*) and (*c*) Analysis of three-wavelength MAD data from the C-terminal domain of initiation factor 3 (IF3-C) and gene V protein (GVP) with varying amounts of missing data. MAD data sets collected on selenomethionne-containing IF3-C and GVP were analyzed as in (*a*) by calculating the correlation coefficient between MAD electron-density maps and maps calculated from the refined coordinates of IF3-C (Biou *et al.*, 1995, PDB entry 1tig.ent) or GVP (Skinner *et al.*, 1994, PDB entry 1bgh.ent). The IF3-C MAD data had 86% of the possible data from 20 to 1.95 Å with $F > 0$ in space group *C*2, and the GVP data set had 94% of the possible data from 20 to 2.6 Å with $F > 0$ in space group *C*2. The heavy-atom models used for the IF3-C and GVP data sets contained two and one Se atoms in each asymmetric unit, respectively.

averaged over all F_k and θ, weighted in the same way. In practice, we find that the two-dimensional integration is very nearly approximated by a one-dimensional integration over θ, and choosing for each θ the most likely value of F_k. This is the method that will be used here.

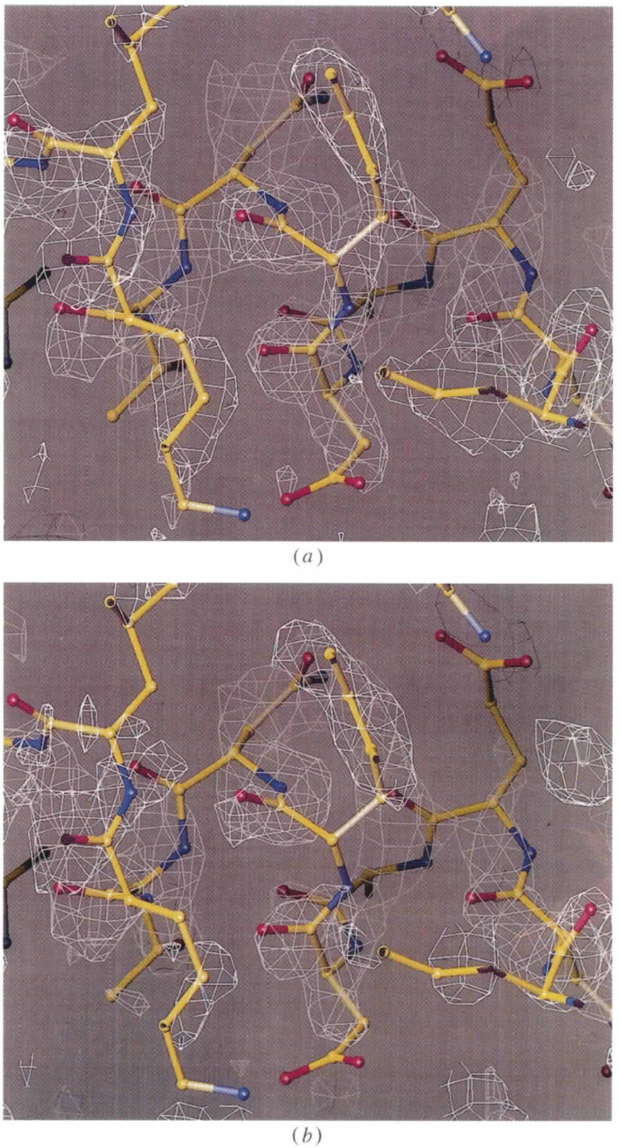
3. Evaluation of Bayesian correlated MAD phasing

3.1. *Application to a model data set*

We first examined the utility of Bayesian correlated MAD phasing by applying it to a set of model MAD data we had previously used to test analysis of MAD data as pseudo-SIRAS data (Terwilliger, 1994*a*,*b*). This data set contains 1730 reflections from 20 to 3 Å at three wavelengths. The amplitudes of structure factors are derived from a model of a polypeptide chain with 86 amino acids containing three Se atoms. The heavy-atom model used in analysis included only two of the three Se atoms. One set of heavy-atom parameters was used for all the phase calculations so that changes in model parameters would not bias comparison of the phasing formulations. Fig. 1(*a*) shows a comparison of the quality of electron-density maps obtained using Bayesian correlated MAD phasing, using pseudo-MIR phasing with λ_1 data as 'native' (Ramakrishnan *et al.*, 1993; Ramakrishnan & Biou, 1996), and using *MADLSQ* and the *MADSYS* suite of programs (Hendrickson, 1991). When all the data is included in the analysis, the three phasing methods give comparable results, with correlation coefficients between calculated and known electron-density maps ranging from 0.62 to 0.68. When fewer data are available for analysis, however, Bayesian correlated MAD phasing performs better than either the MIR-like or the *MADLSQ* procedures. At a completeness of 60%, for example, Bayesian correlated MAD phasing yields a correlation coefficient to the known map of 0.58, while the MIR-like procedure has a correlation coefficient of 0.43 and the *MADLSQ* approach yields a correlation coefficient of 0.41. There are most likely several reasons for the relative insensitivity of the correlated phasing procedure to missing data. One is that in cases where measurements of both F^+ and F^- at the 'native' wavelength are missing, the MIR-like procedure produces no phase information at all, and the *MADLSQ* procedure has a greatly reduced accuracy. A second reason is that only the Bayesian correlated MAD phasing method takes into account the very large correlated error that is present in this simulated data set and model, in which the structure factor from the third selenium site was not included.

3.2. *Application to two real cases*

We next compared the three approaches on two structure that have recently been solved using MAD phasing, the C-terminal domain of initiation factor 3 (IF3-C), which was solved using MIR-like MAD phasing (Biou, Shu & Ramakrishnan, 1995), and gene V protein (GVP) from bacteriophage $f1$, which was solved using the pseudo-SIRAS approach with *MADMRG* (Skinner *et al.*, 1994). For both data sets, heavy-atom parameters were again obtained just once and used for all the phase calculations on that data set. The quality of the phasing obtained was evaluated by calculation of the correlation coefficient of the resulting electron-density maps to those calculated from the refined models of IF3-C and GVP, respectively. Fig. 1(*b*) shows a comparison of correlated MAD phasing



(*a*)

(*b*)

Fig. 2. Portion of electron-density maps obtained for IF3-C using (*a*) Bayesian correlated MAD phasing and (*b*) *MADLSQ*. The electron-density maps are contoured at 1.5 times their r.m.s. values and are superimposed on the refined structure of IF3-C (Biou *et al.*, 1995).

with MIR-like phasing and *MADLSQ*-based phasing on IF3-C. This data set had about 86% of the data with $F > 0$, and when all the available data was included in the analysis all three methods give correlations in the range 0.61–0.64 to the F_{calc} map. As fewer data are available for analysis, however, Bayesian correlated MAD phasing once again proves more robust than the other two methods, yielding a correlation of 0.58 at a completeness of 60% compared with 0.50 for the MIR-like procedure and 0.47 for *MADLSQ*.

A similar result was obtained for GVP. When all available data is used to calculate phases, all three procedures work equally well, with correlation coefficients to the F_{calc} map ranging from 0.47 to 0.48. Bayesian correlated MAD phasing is again less sensitive than the other two approaches to the effects of missing data. When 61% of the data is used, for example, the Bayesian correlated MAD phasing procedure still has a correlation coefficient of 0.42 while the MIR-like and *MADLSQ* procedures each have a correlation of 0.35.

We next examined whether the insensitivity of the Bayesian correlated MAD procedure to missing data demonstrated in Fig. 1 would have a marked effect on one's ability to interpret an actual electron-density map. A region of the electron-density maps calculated for IF3-C using correlated phasing and using *MADLSQ*, based on 60% of the data are shown in Fig. 2, along with the coordinates of the refined model of IF3-C. In this calculation, the estimated figure of merit of the map phased with Bayesian correlated MAD phasing was 0.65, with 6002 or 6005 reflections phased, and the correlation of this map with a map calculated from the refined coordinates of IF3-C was 0.58. The estimated figure of merit of the *MADLSQ*-phased map was 0.85, with 3522 of 6005 reflections phased, and the correlation of this map to the F_{calc} map was 0.47. It is clear from Fig. 2 that in this case the higher correlation of final and Bayesian correlated MAD phasing maps corresponds to features that would facilitate map interpretation.

4. Assessment of the approach

Analysis of a MAD data set that is accurately measured and highly complete can be readily accomplished with any of several existing methods (*e.g.*, Hendrickson, 1985, 1991; Pahler *et al.*, 1990; Burling *et al.*, 1996; Ramakrishnan *et al.*, 1993; Ramakrishnan & Biou, 1996; Terwilliger, 1994*b*, 1996). Our results and those of more extensive comparative studies (Ramakrishnan & Biou, 1996) indicate there are only small differences in the quality of the electron-density maps obtained with any of a variety of methods if the data are complete.

When a measured MAD data set is significantly incomplete, however, different approaches display a wide variation in the quality of phasing obtained. Our tests indicate that the phasing procedure described here, which treats all Bijvoet pairs and measurements of a reflection at different wavelengths on an equal footing and which takes into account the correlations of errors among these measurements, is much more robust to the effects of missing data than are others in common use. Fig. 1(*b*) shows that Bayesian correlated MAD phasing operating on a data set that is 70% complete, for example, can produce an electron-density map of a quality for which *MADLSQ* (Hendrickson, 1985) would have required a 86% completeness in order to equal. We conclude from these tests that Bayesian correlated MAD phasing will be a particularly useful tool in the analysis of MAD data sets that are missing a substantial fraction of the data. We also expect that correlated MAD phasing will be useful in cases where there are very small anomalous and dispersive differences because this approach takes the correlated errors that are present in these cases into consideration.

The estimates of the structure-factor amplitudes F_k obtained from this analysis can be expected to be improved over those measured at any particular wavelength because they include information from all the measured data. For example, if data are missing at wavelength λ_1 for a particular reflection, then a solvent-flattening analysis using the λ_1 data would be missing this reflection, while F_k data would still be available for this reflection if any data were measured at any other wavelength. For this reason, not only Fourier syntheses, but also any other analyses of the MAD data carried out after Bayesian correlated MAD phasing ordinarily should optimally be carried out using the F_k estimates, not using the raw measured data at one wavelength.

The authors wish to thank V. Ramakrishnan for use of the IF3-C MAD data. The authors are also grateful for support from the National Institutes of Health and from the Laboratory Directed Research and Development program of Los Alamos National Laboratory. Bayesian MAD phasing has been implemented in version 4 of the package *HEAVY*, available by contacting TT at terwilliger@lanl.gov.

References

Bevington, P. R. (1969). *Data Reduction and Error Analysis for the Physical Sciences*. New York: McGraw-Hill.

Biou, V., Shu, F. & Ramakrishnan, V. (1995). *EMBO J.* **16**, 4056–4064.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*. London: Academic Press.

Boissy, G., de La Fortelle, E., Kahn, R., Huet, J.-C., Bricogne, G., Pernollet, J.-C. & Brunie, S. (1996). *Structure*, **4**, 1429–1439.

Box, G. E. P. (1980). *J. R. Stat. Soc. A*, **143**, 383–430.

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

Burling, F. T., Weis, W. I., Flaherty, K. M. & Brünger, A. T. (1996). *Science*, **271**, 72–77.
Chiadmi, M., Kahn, R., de la Fortelle, E. & Fourme, R. (1993). *Acta Cryst.* D**49**, 522–529.
de la Fortelle, E. & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.
Hendrickson, W. A. (1985). *Trans. Am. Crystallogr. Assoc.* **21**, 11–21.
Hendrickson, W. A. (1991). *Science*, **254**, 51–58.
Hendrickson, W. A., Pahler, A., Smith, J. L., Satow, Y., Merritt, E. A. & Phizackerley, R. P. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 2190–2194.
Karle, J. (1980). *Int. J. Quantum Chem.* **7**, 357–367.
Leahy, D. J., Aukhil, I. & Erickson, H. P. (1996). *Cell*, **84**, 155–164.
Otwinowski, Z. (1991). *Isomorphous replacement and anomalous scattering: Proceedings of the CCP4 study weekend, 25–26 January 1991*, edited by P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.
Pahler, A., Smith, J. L. & Hendrickson, W. A. (1990). *Acta Cryst.* A**46**, 537–540.
Peat, T. S., Frank E. G., McDonald, J. P., Levine, A. S., Woodgate, R. & Hendrickson, W. A. (1996). *Nature (London)*, **380**, 727–730.
Ramakrishnan, V. & Biou, V. (1996). *Methods Enzymol.* **276**, 538–557.
Ramakrishnan, V., Finch, J. T., Graziano, V., Lee, P. L. & Sweet, R. M. (1993). *Nature (London)*, **362**, 219–223.
Skinner, M. M., Zhang, H., Leschnitzer, D. H., Bellamy, H., Sweet, R. M., Gray, C. M., Konings, R. N. H., Wang, A. H.-J. & Terwilliger, T. C. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 2071–2075.
Stewart, J. M. & Karle, J. (1976). *Acta Cryst.* A**32**, 1005–1007.
Terwilliger, T. C. (1994*a*). *Acta Cryst.* D**50**, 11–16.
Terwilliger, T. C. (1994*b*). *Acta Cryst.* D**50**, 17–23.
Terwilliger, T. C. (1996). *Methods Enzymol.* **276**, 530–537.
Terwilliger, T. C. & Berendzen, J. (1996). *Acta Cryst.* D**52**, 749–757.
Terwilliger, T. C. & Eisenberg, D. S. (1987). *Acta Cryst.* A**43**, 6–13.
Wilson, (1949). *Acta Cryst.* **2**, 318–321.