Acta Cryst. (1994). D**50**, 11–16

MAD Phasing: Bayesian Estimates of F_A

BY THOMAS C. TERWILLIGER

*Mail Stop M*880, *Life Sciences Division, Los Alamos National Laboratory, Los Alamos, NM* 87545, *USA*

(*Received* 12 *January* 1993; *accepted* 3 *August* 1993)

Abstract

A Bayesian approach is applied to the calculation of Patterson functions and cross-Fourier maps in the analysis of multi-wavelength anomalous-diffraction (MAD) data. This procedure explicitly incorporates information available *a priori* on the likely magnitudes of partial structure factors (F_A) corresponding to the anomalously scattering atoms, uses weighted-average estimates of F_A, and incorporates estimates of errors in the data that are not represented in the instrumental uncertainties. The method is demonstrated by application to MAD data collected on selenomethionine-containing gene V protein.

Introduction

The use of multi-wavelength anomalous-diffraction (MAD) data has recently become a powerful tool for structure determination of biological macromolecules by X-ray diffraction (Karle, 1980; Hendrickson, 1991). In this technique, structure-factor amplitudes are measured at several wavelengths near to and distant from an absorption edge for an atom present at a small number of sites in the asymmetric unit of the crystal. The anomalous and dispersive components of the scattering factors are then used to estimate both the magnitudes of structure factors (F_A) corresponding to the anomalously scattering atoms, and the phase difference ($\Delta\varphi$) between these structure factors and those corresponding to all atoms in the structure (F_Z).

A key step in determining phases with the MAD technique is the determination of the positions of anomalous scatterers in the unit cell. In the widely used technique developed by Karle & Hendrickson (Karle, 1980, 1989; Hendrickson, 1985, 1991; Pähler, Smith & Hendrickson, 1990), the most probable values of F_A, obtained using a least-squares approach, are used to calculate a Patterson function from which the locations of the anomalous scatterers are identified. A model describing these atoms is then refined, and the model partial structure for the anomalously scattering atoms is used with the MAD data to obtain estimates of $\Delta\varphi$, and hence, of phases for the entire structure. While this method has been quite effective, the first step in the technique can yield estimates of F_A that are unrealistically high, particularly if the data contains large errors in measurement (Pähler *et al.*, 1990). These large F_A, which can be a substantial fraction of all the structure factors, must be identified and rejected before a Patterson synthesis is calculated (Yang, Hendrickson, Crouch & Satow, 1990). An alternate approach for identifying the positions of anomalously scattering atoms in the structure is to calculate an anomalous-difference Patterson function using anomalous differences at a single wavelength (Ramakrishnan, Finch, Graziano, Lee & Sweet, 1993). While this anomalous-difference Patterson and the related dispersive-difference Patterson using differences between structure-factor amplitudes at different wavelengths are very useful, it is not straightforward to combine the information from all the anomalous and dispersive differences into a single Patterson function in this method.

In the present work, a Bayesian approach is used to estimate values of F_A from MAD data. In this approach, information on the number and types of anomalously scattering atoms in the asymmetric unit is used to obtain probability distributions for the expected values of F_A. This information is used together with the MAD data to obtain estimates of F_A that are weighted toward values that are, *a priori*, more likely. Additionally, values of parameters of interest such as F_A are obtained using their weighted-average values rather than their most probable values. This procedure is analogous to phasing using figure-of-merit weighting and the 'best' phase in the method of multiple isomorphous replacement (MIR), rather than using the 'most probable' phase and unit weighting (Terwilliger & Eisenberg, 1987).

Methods

The principal experimental information available for each reflection in the MAD technique consists of observations of Bijvoet pairs of structure-factor amplitudes for this reflection (F^+ and F^-) measured at several X-ray wavelengths. The quantities to be determined in the present formulation are the magnitude of the structure factor corresponding to the anomalously scattering atoms (F_A), the magnitude of

ISSN 0907-4449

the structure factor corresponding to all other atoms in the unit cell (F_o), and the phase difference between these two structure factors (α). These quantities are chosen because they are all independent of each other. The quantities used by Hendrickson (Hendrickson, 1991) are closely related to these and can be readily calculated from them.

The approach descibed here for analysis of MAD data, as well as the nomenclature, is similar to that used previously for calculation of phase probability distributions in the multiple isomorphous replacement (MIR) method (Terwilliger & Eisenberg, 1987). In essence, Bayes's rule (Hamilton, 1964) is used to estimate the relative probability, $P(F_A, F_o, \alpha)$, that each possible set of values of F_A, F_o and α are correct. Then the best estimate of any quantity such as α is the weighted average, over all values of F_A, F_o and α, of this quantity.

To carry out this calculation, it is necessary to first obtain two probability distributions. The first is an *a priori* probability distribution for F_A and the second is a probability distribution for observed data given a set of values of F_A, F_o and α. These two probability distributions can then be combined to yield the probability that any particular combination of values of F_A, F_o and α is correct.

A priori probability distribution for F_A

As described above, the least-squares method for analyzing MAD data often leads to overestimates of F_A. These estimates can be much larger, in fact, than the possible range of values of F_A given the types and numbers of anomalously scattering atoms in the unit cell. In the analysis used here, the values of F_A are restricted to a range that is reasonable, given information on the anomalously scattering atoms, by calculating an *a priori* probability distribution for the magnitude of the structure factor as a result of anomalously scattering atoms. This will be possible if information is available concerning the number and type of anomalously scattering atoms, an approximate value of the thermal factors associated with these atoms is known, and the measured structure-factor amplitudes have been put on an absolute scale.

If there are many anomalously scattering atoms in the unit cell, the *a priori* probability distribution for F_A is given for acentric reflections by,

$$P_o(F_A) \propto F_A \exp -(F_A^2/\Sigma^2). \quad (1)$$

Here Σ^2 is the mean square value of F_A within an appropriate range of resolution, and is obtained directly from the number and type of anomalously scattering atoms and their thermal factors (Wilson, 1949). In practice, the thermal factors for anomalously scattering atoms are not known exactly and the number of such atoms is small. The contribution of (1) to the overall probability distribution will not depend strongly on the precise value of Σ^2, however. A reasonable estimate of Σ^2 may therefore be obtained by assuming that the thermal factors for anomalously scattering atoms are the same as the average for all other atoms in the unit cell, and by neglecting the relatively small effects of the limited number of atoms on the probability distribution.

In this analysis, the *a priori* probability distributions for F_o and α are assumed to be constants. That is, all values of these parameters are assumed to be equally likely before measurements of them are made. In the case of F_o, an *a priori* probability distribution is unneccessary as it is generally fairly well defined from the MAD experiment. In the case of the phase angle, α, all possible values are assumed to be equally likely as no information on the arrangement of atoms in the asymmetric unit of the unit cell is ordinarily known at the start of the MAD experiment.

Probability distribution for experimental data at a given X-ray wavelength, given parent values for F_A, F_o and α

The experimentally observable data in MAD analyses are the structure-factor amplitudes (F^+ and F^-), measured at several X-ray wavelengths. For the purpose of this analysis, it is useful to convert these measurements to estimates of the average structure-factor amplitude, $\overline{F}$, and an anomalous difference, $\Delta_{\rm ANO}$. This form is advantageous because errors in measurement of the Bijvoet pairs are often highly correlated, while errors in the average structure-factor amplitude and in the anomalous difference are much more likely to be independent. Additionally, this formulation allows separate estimation of systematic errors in the two quantities, as has been performed in analyses of MIR data (*e.g.* Matthews, 1966; Terwilliger & Eisenberg, 1987). It can be written that, for measurements at a particular X-ray wavelength and for a particular reflection

$$\overline{F} = \tfrac{1}{2}(F^+ + F^-) \quad (2)$$

$$\Delta_{\rm ANO} = (F^+ - F^-). \quad (3)$$

These quantities then have associated experimental uncertainties of $\sigma_{\overline{F}}$ and $\sigma_{\rm ANO}$, respectively. If one of the members of a Bijvoet pair is missing, the other can still be used to estimate $\overline{F}$, though this estimate will have an additional uncertainty of half the r.m.s value of the anomalous differences in the corresponding resolution range. It is assumed that $\overline{F}$ and $\Delta_{\rm ANO}$ may have additional errors associated with them, and denote these additional errors $E_{\overline{F}}$, and $E_{\rm ANO}$, respectively. The estimation of these additional errors is discussed in a later section.

Given values of the parameters F_A, F_o and α, then the values of the average structure-factor amplitude, $\bar{F}$, and of the anomalous difference, Δ_{ANO}, can be immediately calculated. Along with estimates of the total errors in measurement, and the assumption of a Gaussian distribution of errors, a probability distribution of $\bar{F}$ and Δ_{ANO} can be written as,

$$P(\bar{F}, \Delta_{ANO}|F_A, F_o, \alpha) \propto \exp -\frac{1}{2}\left(\frac{\varepsilon^2_{\bar{F}}}{E^2_{\bar{F}} + \sigma^2_{\bar{F}}} + \frac{\varepsilon^2_{ANO}}{E^2_{ANO} + \sigma^2_{ANO}}\right), \quad (4)$$

where the differences between observed values of $\bar{F}$ and Δ_{ANO} and those calculated from the values of F_a F_o and α, are $\varepsilon_{\bar{F}}$ and ε_{ANO}, respectively.

Probability distribution for F_A, F_o and α, given observed values of F and Δ_{ANO} at several X-ray wavelengths

Applying Bayes' rule (Hamilton, 1964), it can now be written that, after making measurements of $\bar{F}$ and Δ_{ANO} at several X-ray wavelengths, the probability distribution for F_A, F_o and α is the product of the *a priori* probability distribution for F_A and the probability distributions at each wavelength for $\bar{F}$ and Δ_{ANO} given parent values of F_A, F_o and α:

$$P(F_A, F_o, \alpha|\{\bar{F}, \Delta_{ANO}\}_{\lambda_1 \ldots \lambda_N}) \propto P_o(F_A) \prod_{\lambda_1 \ldots \lambda_N} P(\bar{F}, \Delta_{ANO}|F_A, F_o, \alpha), \quad (5)$$

where the product is over all X-ray wavelengths used ($\lambda = \lambda_1 \ldots \lambda_N$). (5) gives an estimate of the relative likelihood that a particular set of values of F_A, F_o and α is correct.

Estimation of F_A, F_A^2, F_o and α

Using the probabiliity distribution given in (5), the value of any quantity of interest, x, that depends on F_A, F_o and α can be estimated by averaging its value over all possible values of F_A, F_o and α, weighting by the likelihood that this set is correct:

$$\langle x \rangle = \frac{\int x\, P(F_A, F_o, \alpha|\{\bar{F}, \Delta_{ANO}\}_{\lambda_1 \ldots \lambda_N})\, dF_A\, dF_o\, d\alpha}{\int P(F_A, F_o, \alpha|\{\bar{F}, \Delta_{ANO}\}_{\lambda_1 \ldots \lambda_N})\, dF_A\, dF_o\, d\alpha}, \quad (6)$$

where the integration is over all possible values of the three variables.

As this integration over three variables can be time consuming, an additional simplification is made in implementing (6). It is assumed that F_o is quite sharply defined by the experimental data, so that integration over this variable is not as important as for F_A and α, which are not as precisely defined. Instead, for each set of values of F_A and α, the value of F_o that is most probable is found and 'integration' over this variable is carried out only at this point of maximum probability.

Estimation of errors in measurement not included in instrumental uncertainties

In many cases, the accuracy of X-ray diffraction data is limited by errors such as those caused by inaccuracies in data collection and in scaling or absorption corrections that are difficult to estimate. It is assumed that these additional errors do not vary strongly from reflection to reflection within a range of resolution in a data set. These errors are estimated in the fashion developed for estimation of comparable errors in the MIR method (Terwilliger & Eisenberg, 1987). In each case, the weighted-average value of the squared differences between observed and calculated values of $\bar{F}$ and Δ_{ANO} is used as an estimate of the total mean-square error in these variables. For anomalous differences, for example, an estimate of the total error in a particular measurement is $\langle \varepsilon^2_{ANO} \rangle$, where the average is taken using (6). Then it is possible to estimate the mean value of errors not included in the instrumental uncertainties from the mean difference between $\langle \varepsilon^2_{ANO} \rangle$ and σ^2_{ANO}:

$$E^2_{ANO} = (1/N) \Sigma \{\langle \varepsilon^2_{ANO} \rangle - \sigma^2_{ANO}\} \quad (7)$$

where the summation is over reflections in a range of resolution. A similar relationship is used to estimate the errors in measurement of $\bar{F}$.

Results and discussion

Test of Bayesian analysis using model MAD data

We first tested the approach described here by using it to analyze model sets of data containing variable amounts of error. The exact model data was based on a set of measured structure-factor amplitudes, random native phases, and Se atoms as the anomalously scattering atoms (see legend to Table 1). A group of 15 data sets differing from the exact one by 'random' errors with r.m.s (root-mean-square) values of 1–15% were then constructed. Statistics on some of these data sets are listed in Table 1. These 15 data sets were analyzed with the models described here, as implemented in the program *FABEST*, and with the procedure of Hendrickson (1985), using the program *MADLSQ*, in order to obtain estimates of F_A for use in Patterson syntheses to simulate the first steps in a MAD structure determination. In both cases, exact values of f' and f'' for selenium and of the scale factors among data sets were used in the analysis, so that it was not necessary to determine or refine them.

Using each method, some reflections could not be successfully analyzed for values of F_A. In the Baye-

Table 1. *Characteristics of model MAD data*

An exact model data set was constructed using the 1763 measured structure-factor amplitudes from 10 to 3 Å resolution from the *C*2 crystal form of gene V protein, which has 682 non-H atoms in the asymmetric unit (Skinner *et al.*, 1993). These were placed on an approximate absolute scale and had an overall thermal factor of $B = 32$ Å^2. Arbitrary phases were assigned to each reflection. The two Se atoms were placed at positions (0.141, 0.344, 0.219) and (0.484, 0.500, 0.094) and were each given thermal factors of 20.0 Å^2. Structure factors for the protein plus Se atoms were calculated for X-ray wavelengths of $\lambda_a = 0.9000$, $\lambda_h = 0.9794$ and $\lambda_c = 0.9797$ Å to make up an exact model data set. Values of the scattering factors used for the Se atoms were, $f'(\lambda_a) = -1.622$, $f'(\lambda_h) = -8.639$, $f'(\lambda_c) = -9.851$, $f''(\lambda_a) = 3.284$, $f''(\lambda_h) = 4.879$, $f''(\lambda_c) = 2.858$. The 15 data sets analyzed in Fig. 1 were derived from the exact data set by adding random errors of 1–15% as described in the text. Statistics for the exact data set and data sets with 2, 4, 6, 8 and 10% errors are listed. The mean value of structure factors at λ_a were $\langle \bar{F}_{\lambda_a} \rangle = 254$ in all cases. In the table, the normalized mean value of the error in $\bar{F}_{\lambda_a}$ is $\langle \sigma(\bar{F}_{\lambda_a}) \rangle / \langle \bar{F}_{\lambda_a} \rangle$, the normalized mean absolute value of the anomalous difference at λ_h is $\langle \Delta_{ANO,\lambda_h} \rangle / \langle \bar{F}_{\lambda_a} \rangle$, and the normalized mean absolute value of the dispersive difference between structure factors at λ_c and λ_a is $\langle |\bar{F}_{\lambda_c} - \bar{F}_{\lambda_a}| \rangle / \langle \bar{F}_{\lambda_a} \rangle$. Note that the errors in $\bar{F}_{\lambda_a}$ are smaller than the errors in the structure factors as each $\bar{F}_{\lambda_a}$ is the average of Bijvoet pair. For these model data sets, the statistics do not vary substantially with resolution and an average is shown.

	R.m.s. error in structure factors					
	0%	2%	4%	6%	8%	10%
$\langle \sigma(\bar{F}_{\lambda_a}) \rangle / \langle \bar{F}_{\lambda_a} \rangle$	0.000	0.014	0.028	0.042	0.057	0.071
$\langle \Delta_{ANO,\lambda_h} \rangle / \langle \bar{F}_{\lambda_a} \rangle$	0.053	0.060	0.074	0.091	0.109	0.129
$\langle \bar{F}_{\lambda_c} - \bar{F}_{\lambda_a} \rangle / \langle \bar{F}_{\lambda_a} \rangle$	0.058	0.060	0.068	0.078	0.090	0.102

sian method, reflections for which the maximum probability encountered in the integration in (6) was less than 0.001 were rejected. In the method of Hendrickson (1985), all reflections with estimated values of F_A greater than 300 were rejected, where the true mean value of F_A overall was 78. Both methods were able to analyze essentially all the data when errors were small. When errors were about 6%, however, the least-squares method (*MADLSQ*) was unable to obtain F_A for 15% of the reflections, and when errors were 10% this method obtained unrealistically large F_A values for 30% of the reflections. The Bayesian method was able to obtain reasonable F_A values for over 99.9% of all reflections over the entire range of errors tested.

The accuracy of estimates of F_A using the two methods are illustrated in Fig. 1, in which the lowest values of Patterson syntheses at expected positions of self and cross vectors are shown as a function of error in the data. These peak heights are shown as ratios to the r.m.s. of the origin-removed Patterson functions. For data with very small errors (1%), both methods yielded Patterson functions that were nearly identical to a Patterson synthesis based on the model data with no errors. When the error in the data was greater than about 4%, however, the peak heights in the Patterson function (relative to the r.m.s. values of the origin-removed maps) obtained after analysis with *MADLSQ* decreased much more rapidly than those obtained with *FABEST*. This decrease is probably because of our rejection of the estimates of F_A that are unrealistically large, many of which are also likely to represent values of F_A that are actually large as well. Simple truncation of the large values of F_A is not useful in this case, however. For each test data set with errors from 1 to 9%, when large values of F_A are truncated to a value of 300, the value of the Patterson function at positions of self and cross vectors is, after normalization to the r.m.s. value in the map as in the previous cases, lower than its value when large A are rejected.

In the analysis of MAD data developed by Hendrickson (1991), the value of $\Delta\varphi$, the phase difference between F_A and F_Z, estimated by *MADLSQ* is not ordinarily used directly, but is rather redetermined in a later step in which the F_A values are calculated from a model. In some cases, however, it might be useful to have a good estimate of $\Delta\varphi$ (or the related phase difference used in this work, α) before a model for the anomalously scattering atoms is obtained and refined. If some phase information were available from another source, for example, the values of $\Delta\varphi$ or α and F_A could be used to calculate a Fourier map showing the locations of the anomalously scattering atoms. In most cases additional phase information would consist of estimates of the phase of F_o, the structure factor from

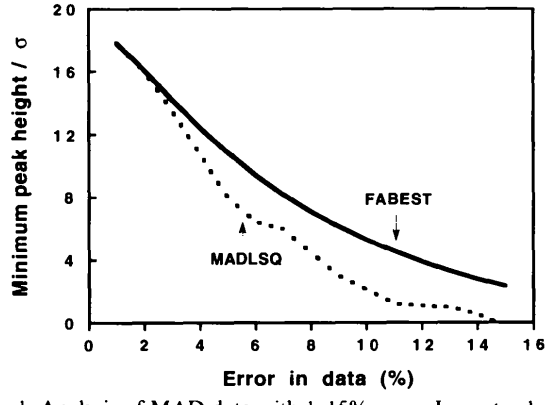

Fig. 1. Analysis of MAD data with 1–15% error. Lowest value of Patterson function at positions corresponding to self vectors and cross vectors based on values of F_A obtained by *FABEST* (——) or *MADLSQ* (- - -). Values are normalized to the r.m.s. value for the Patterson function, excluding the origin. Each data set was analyzed with *FABEST* (the procedure described here) and by *MADLSQ* (Hendrickson, 1985), and the two approaches were used to obtain estimates of the structure-factor amplitude for Se atoms (F_A), of the structure-factor amplitude for protein atoms plus the Se atoms (F_Z), and of the phase difference between these two structure factors ($\Delta\varphi$). The values of F_A^2 estimated using each method were used in Patterson syntheses, and the minimum value of the Patterson function at the two unique positions on the Harker section and the two positions corresponding to cross vectors between the two sites was determined in each case.

non-anomalously scattering atoms in the unit cell, as described in the next section. To compare the utility of the Bayesian approach with that of the least-squares approach, however, it was most straightforward to evaluate the accuracy of determination of $\Delta\varphi$, the phase difference between F_A and F_Z, for each method. For reflections where the estimate of F_A made by *MADLSQ* was reasonable (less than 300 in our test data set), the average phase error for the two methods was very similar. For example, when the error in the data was 6%, the mean phase error in $\Delta\varphi$ estimated by *MADLSQ* for this subset of the data was 35.1° and that for *FABEST* was 35.4°. For the reflections where the F_A estimate made by *MADLSQ* was unrealistically large, however, the phase estimates were also quite inaccurate for this method, with a mean phase error of 73.5° when the error in the data was 6%. The Bayesian approach, on the other hand, yielded estimates of $\Delta\varphi$ equally accurate for this set of reflections as for the others, with a phase error of 34.0° for the same data. This means that the Bayesian approach could be quite useful in calculation of a Fourier map based on estimates of α and F_A.

Application of Bayesian analysis to Patterson syntheses of MAD data collected on gene V protein

The structure of the gene V protein of bacteriophage f1 has recently been determined using the MAD analysis methods described in this work and in the accompanying paper (Skinner *et al.*, 1993; Terwilliger, 1994). The diffraction data available in this case consisted of native data collected with Cu $K\alpha$ radiation and MAD data at three wavelengths collected on the 'wild-type' gene V protein containing two selenomethionine residues and on a mutant containing three selenomethionine residues. The MAD data on the wild-type selenomethionine-containing protein were of excellent quality and 94% complete to 2.6 Å, but that for the mutant were quite weak and only 65% complete to 2.5 Å (Skinner *et al.*, 1993). All three crystal forms were isomorphous and were in the space group *C*2.

In this structure determination, we used the Bayesian approach implemented in an earlier version of the program *FABEST* to estimate both F_A and α values and to calculate Patterson syntheses and cross-Fourier syntheses for the two selenomethionine-containing structures. Using the current version of *FABEST*, values of F_A and α could be obtained for 2692 reflections from 10 to 2.6 Å for the wild-type selenomethionine-containing protein, representing 100% of the measured reflections. The 'wild-type' protein contains two selenomethionine residues and a value of two Se atoms was assumed for the *a priori* probability distribution for F_A. The Patterson synthesis using the F_A data from the wild-type selenomethionine-containing protein yielded a clean map corresponding to a single Se atom in the asymmetric unit. The self-vector peak for this site had a height of 15.3 times the r.m.s. of the origin-removed map, and the next highest peak in the map other than the origin was just 7.8 times the r.m.s. value. It was assumed that one of the selenomethionine residues was disordered (this was later found to be residue 1 of the protein) and values of F_A and α were recalculated using an estimate of one Se atom in the asymmetric unit. This resulted in only a slight change in the Patterson synthesis, with the self-vector peak for the single site having a value of 15.7 times the r.m.s. in the origin-removed map.

This Patterson synthesis based on Bayesian estimates of F_A can be compared to anomalous- and dispersive-difference Patterson functions calculated directly from the MAD data. An anomalous-difference Patterson was obtained based on the anomalous differences measured at the wavelength of maximum f'' for selenium (λ_b) and a dispersive difference Patterson was obtained from the differences between measurements at the wavelength of minimum f' (λ_c) and a wavelength far from the absorption edge (λ_a). Both difference Patterson functions yielded the same single-site solution as the Patterson based on Bayesian estimates of F_A, but neither was as clear. The anomalous-difference Patterson had a self-vector peak 12.8 times the r.m.s. of the origin-removed map and the dispersive-difference Patterson had a self-vector peak 11.2 times the r.m.s. of the map.

The mutant protein had three selenomethionine residues, and in the calculation of F_A and α, it was assumed that all three were present. A total of 2068 reflections from 10 to 2.5 Å were successfully analyzed by *FABEST*, corresponding to 99.7% of those measured. As these data were both weak and very incomplete, it was not surprising that the anomalous-difference Patterson, the dispersive-difference Patterson, and the Patterson calculated with Bayesian estimates of F_A were all very noisy. Of the three Patterson syntheses, only the anomalous-difference Patterson was readily interpretable, where a clear two-site solution to the Patterson function was obtained using the automatic search program *HASSP* (Terwilliger & Eisenberg, 1987) in which one site was identical to that found for the 'wild-type' protein. The self-vector peaks corresponding to the two sites were 6.6 and 3.6 times the r.m.s. of the origin-removed map, respectively, and the cross-vector peak was 3.7 times the r.m.s. of the origin-removed map. The dispersive difference Patterson had peaks corresponding to the same pair of sites, but only as 3.7 and 3.2 times the r.m.s. of the map. Similarly, the Patterson calculated from Bayesian

estimates of F_A had peaks at 3.9 and 3.9 times the r.m.s of the map and would not have been readily interpretable. The Bayesian approach was therefore not useful in this case with very weak and incomplete data.

Application of Bayesian approach to cross-Fourier analyses of MAD data collected on gene V protein

As the data from 'wild-type' and mutant selenomethionine-containing gene V proteins constituted two largely independent data sets, each could be used in a cross-Fourier analaysis to verify the locations of Se atoms in the other. To avoid biasing the outcome of this analysis, only one site, the site not in the wild-type structure, was used in the modeling for the mutant structure. Parameters describing one Se atom in the asymmetric unit were refined for the 'wild-type' structure, for example, as described in the accompanying paper (Terwilliger, 1994), and phases for the structure factors (F_o) corresponding to all atoms in the structure except for the anomalously scattering atoms were estimated. Then the phase difference, α, for the mutant structure was calculated with *FABEST* and was subtracted from the phase estimate for F_o to yield an estimate of the phase of F_A for the mutant structure, and a Fourier synthesis was carried out.

Using native phases calculated using the 'wild-type' MAD data, the cross-Fourier map for the mutant protein showed the site present in the 'wild type' and a second site identical to that found in the Patterson analyses. The height for this second site was 10.3 times the r.m.s. of the map. Similarly, using native phases calculated using the mutant MAD data and including only this second site in the model, the cross-Fourier map for the 'wild-type' protein showed the expected site with a height of 9.1 times the r.m.s. of the map. These cross-Fourier analyses were very important in the gene V protein structure determination because they showed that the Patterson solutions we had obtained were correct and because they verified the relative positions of the two sites in the two proteins (Skinner *et al.*, 1993).

Concluding remarks

We find that a Bayesian analysis of MAD data is very useful for obtaining estimates of both F_A and α. In many cases, the data obtained in MAD analyses of macromolecular structures is likely to contain errors of at least a few percent when both instrumental uncertainties and any scaling, absorption and other systematic errors are included. In these cases, a Bayesian approach including information on the likely values of F_A is helpful because it allows estimation of F_A values for almost measured reflections and limits them to values that are reasonable. The F_A values estimated in this way are useful in Patterson functions, and the α values can be helpful in calculating cross-Fourier maps in cases where phase information from more than one MAD or other experiment is available.

This work was supported by generous grants from the NIH and from the Laboratory Directed Research and Development Program of Los Alamos National Laboratory.

References

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

HENDRICKSON, W. A. (1985). *Trans. Am. Crystallogr. Assoc.* **21**, 11–21.

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

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

KARLE, J. (1989). *Acta Cryst.* A**45**, 303–307.

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

PÄHLER, A., SMITH, J. L. & HENDRICKSON, W. A. (1990). *Acta Cryst.* A**46**, 537–540.

RAMAKRISHNAN, C., 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., GUAN, Y., BELLAMY, H., SWEET, R. M., GRAY, C. W., KONINGS, R. N. H., WANG, A. H.-J. & TERWILLIGER, T. C. (1993). In preparation.

TERWILLIGER, T. C. (1994). *Acta Cryst.* D**50**, 17–23.

TERWILLIGER, T. C. & EISENBERG, D. (1987). *Acta Cryst.* A**43**, 6–13.

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

YANG, W., HENDRICKSON, W. A., CROUCH, R. J. & SATOW, Y. (1990). *Science*, **249**, 1398–1405.