Acta Cryst. (1993). D**49**, 522–529

Derivation by Statistical Methods of Phase Information from Multiple-Wavelength Anomalous Diffraction Data. Basic Questions, 'Best' Electron-Density Map, Implementation and Tests

BY M. CHIADMI

Laboratoire de Physique, Faculté des Sciences Pharmaceutiques et Biologiques, Université Paris V, 4 *Avenue de l'Observatoire*, 75270 *Paris CEDEX* 06, *France*

AND R. KAHN, E. DE LA FORTELLE AND R. FOURME

LURE (CNRS, CEA, MESR), Université Paris-Sud, Bâtiment 209*D*, 91405 *Orsay CEDEX, France*

(*Received* 1 *April* 1993; *accepted* 9 *July* 1993)

Abstract

A probability distribution of the structure factor is established from the analysis of the effects of errors involved in the multiple-wavelength anomalous diffraction (MAD) method. This probability distribution, derived from those of the intensities, is two-dimensional for acentric reflections and unidimensional for centric reflections. It permits, using the centroid of the distribution, the calculation of the modulus and the phase of the 'best' structure factor. The procedure for extracting the phase and its figure of merit is presented. Tests performed on simulated data show the contribution of this method with respect to other methods which use a distribution of only the phase as a function of the error of closure.

Introduction

In the field of biological crystallography, one of the most important consequences of the use of synchrotron radiation has been the practical implementation of the multiple-wavelength anomalous diffraction method (MAD), a phasing method which has now been used to solve a number of macromolecular structures (for reviews, see *e.g.* Fourme & Hendrickson, 1990; Hendrickson, 1991; Smith 1991).

MAD is a newcomer to the pallette of *de novo* phasing methods such as multiple isomorphous replacement (MIR) or single isomorphous replacement with anomalous scattering (SIRAS). In these methods, the phasing procedure can be divided into two basic steps. The first step is the determination of the structural parameters of a subset of atoms which have, with respect to the rest of atoms, particular properties such as a large atomic number and/or significant anomalous scattering of the X-ray wavelength(s) used in data collection. The second step is the use of these parameters in the estimation of Fourier coefficients suitable for the calculation of an electron-density map of the total structure. In practice, the distinction between the two steps is not so clear cut, especially because parameters of the subset and Fourier coefficients are often refined together. The MAD method relies on a subset of atoms with significant anomalous scattering. This subset of atoms, A, is called hereafter the partial structure. The pertinent parameters of the partial structure are atomic coordinates, occupancies, temperature factors and the anomalous components of the atomic scattering factors at each wavelength. The first step of the MAD phasing procedure is the determination of these parameters. For each **h**, the second step makes a pivotal use of a reference structure factor calculated from these parameters in order to estimate the phase (and possibly modulus) of the Fourier coefficient to be used in the calculation of the initial electron-density map. In most applications of MAD, the reference structure factor is the wavelength-dependent part of the structure factor of the partial structure, and the result is an estimation of the 'normal' part, ${}^0F_T$, of the structure factor of the total structure (the subscript T refers to the total structure, and the superscript 0 to a wavelength-independent quantity) (Fig. 1).

When compared to the MIR or SIRAS methods, MAD has several advantages: all measurements can be made on a single sample in which no chemical changes occur, thereby supressing problems and limitations related to the lack of isomorphism between the native crystal and the heavy-atom derivatives. In contrast, intensity differences in the MAD method are generally weak. This requires well designed experiments, a careful methodology to optimize both the strength of the signals and the accuracy in the measurement of small intensity differences, and a careful treatment of errors.

Let us assume that the collection, analysis and scaling of multiple-wavelength diffraction data have

been performed, providing for each reflection **h**, n intensity measurements $|{}^{\lambda}F_T(\pm\mathbf{h})|^2$ (the superscript λ is the wavelength of the measurement; $\pm\mathbf{h}$ refers to the anomalous pair **h**, $-\mathbf{h}$). The determination of the structural parameters of the partial structure may be performed using single-wavelength data, although not under optimal conditions, In effect, the Fourier map calculated with coefficients $[|{}^{\lambda}F_T(\mathbf{h}) - |{}^{\lambda}F_T(-\mathbf{h})|]^2$ (Rossmann, 1961) is a sharpened Patterson that represents the distribution of interatomic vectors for the anomalous scatterers, but with a degraded signal-to-noise ratio (Karle, 1980). An intrinsically more powerful method relies on that analysis of multiple-wavelength data by an algebraic procedure first proposed in principle by Karle (1980). Karle expressed all quantities in terms of two structure factors: ${}^0F_A$ which corresponds to the *normal* scattering of the anomalous scatterers and ${}^0F_P$ which corresponds to the normal scattering of the rest of the atoms. Hendrickson (1985) used the formulation in terms of ${}^0F_A$ and ${}^0F_T$ in his implementation of the algebraic method (program *MADLSQ*). With such notations as in Fig. 1, the equations for a single type of anomalous scatterers are

$$\begin{aligned} |{}^{\lambda}F(\pm\mathbf{h})|^2 = {} & |{}^0F_T|^2 + a(\lambda)\,|{}^0F_A|^2 \\ & + b(\lambda)\,|{}^0F_T|\,|{}^0F_A|\cos({}^0\varphi_T - {}^0\varphi_A) \\ & \pm c(\lambda)\,|{}^0F_T|\,|{}^0F_A|\sin({}^0\varphi_T - {}^0\varphi_A), \quad (1) \end{aligned}$$

with

$$a(\lambda) = ({}^{\lambda}f'^2 + {}^{\lambda}f''^2)/(f^0)^2,$$

$$b(\lambda) = 2\,{}^{\lambda}f'/f^0,$$

and

$$c(\lambda) = 2\,{}^{\lambda}f''/f^0.$$

The quantities derived by *MADLSQ* are $|{}^0F_A|$, $|{}^0F_T|$ and the phase difference ${}^0\varphi_T - {}^0\varphi_A$. From the set of $|{}^0F_A|^2$, the structure of anomalous scatterers can by solved by standard methods such as Patterson or direct methods, and then refined by least-squares techniques.

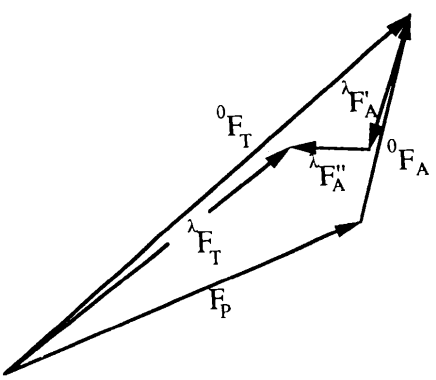

Fig. 1. Vector diagram of structure factors pertinent to the MAD method for one reflection **h** at a single wavelength (see notations in the text).

The second step makes use of the partial structure to obtain estimates of ${}^0F_T$. As for MIR and SIRAS methods, the main problem is the treatment of errors in the derived quantities. Methods which have been used to date for the treatment of errors fall into two broad categories. One has been developed in the context of the algebraic analysis (Hendrickson, 1985) and proceeds as follows. The phase ${}^0\varphi_A$ calculated from the refined model of the partial structure is used to derive ${}^0\varphi_T$ from the phase difference ${}^0\varphi_T - {}^0\varphi_A$. The phase and amplitude of the Fourier coefficient ${}^0F_T$ are thus determined. Two electron densitites based on weighted ${}^0F_T$ coefficients are traced with ${}^0\varphi_A$ calculated for the two enantiomorphs of the partial structure, and the correct map is distinguished by chemical reasonableness or by symmetry considerations. Weights are based on residuals in the least-squares fit of the phase equations to the observations.

The second type of treatment is based on probability methods. Most applications to date of a probability analysis to MAD were adaptations of the error treatment for MIR (Blow & Crick, 1959). In the description of Hendrickson & Lattman (1970), a Gaussian distribution is ascribed to errors of closure in F^2 values instead of F. A probability distribution function for the phase information is derived from the intensity measurements. Fourier coefficients used for the calculation of the electron-density map have centroid phases and figure-of-merit-weighted moduli. The correct enantiomorph of the partial structure is selected as in the algebraic method. This approach was used either with simulated MAD data (Phillips & Hodgson, 1980) or real data (Kahn *et al.*, 1985; Kahn, Fourme, Bosshard, Chiadmi *et al.*, 1986; Korszun, 1988; Guss *et al.*, 1988). Hendrickson *et al.* (1989) and Pähler, Smith & Hendrickson (1990) used it as an alternative way to derive weights for the ${}^0F_T$ Fourier synthesis in the algebraic analysis and the probability distribution was cast in the *A*, *B*, *C*, *D* representation of Hendrickson & Lattman (1970). These simple probability analyses have been quite useful in providing an evaluation of the reliability of phases. Furthermore, the *A*, *B*, *C*, *D* formulation facilitates the combination of MAD information with information from other sources, such as isomorphous replacement, solvent flattening, non-crystallographic symmetry and partial models. As for MIR methods, probability methods are required to exploit fully and rigorously all the available information. Thus, our goal is the formulation and implementation of a probability theory for MAD phasing which includes the refinement of parameters of the partial structure model. The first step in this procedure is reported in this article.

Theory

At the wavelength λ, the scattering factor of the anomalous species k is

$$^{\lambda}f_k = f_k^0 + {}^{\lambda}f'_k + i\,{}^{\lambda}f''_k,$$

and the contribution of the partial structure to the structure factor from the total structure is then

$$^{\lambda}F_A = {}^0F_A + {}^{\lambda}F'_A + i\,{}^{\lambda}F''_A.$$

The expression of the structure factor from the total structure, $^{\lambda}F_T$, can be written as

$$^{\lambda}F_T = {}^0F_T + {}^{\lambda}F'_A + i\,{}^{\lambda}F''_A.$$

The coefficients ^0F_T follow Friedel's law and their Fourier transform is a real map. Let us note $^{\lambda}\varphi_T$, $^0\varphi_T$, $^0\varphi_A$, $^{\lambda}\varphi'_A$ and $^{\lambda}\varphi''_A$ as the phases of $^{\lambda}F_T$, ^0F_T, ^0F_A, $^{\lambda}F'_A$ and $^{\lambda}F''_A$, respectively. Then

$$|^{\lambda}F_T|\exp(i\,{}^{\lambda}\varphi_T) = |^0F_T|\exp(i\,{}^{\lambda}\varphi_T) + |^{\lambda}F'_A|\exp(i\,{}^{\lambda}\varphi'_A) + i|^{\lambda}F''_A|\exp(i\,{}^{\lambda}\varphi''_A). \quad (2)$$

Let us define

$$X = |^0F_T|\cos{}^0\varphi_T, \quad (3a)$$

$$Y = |^0F_T|\sin{}^0\varphi_T, \quad (3b)$$

$$^{\lambda}x = |^{\lambda}F'_A|\cos{}^{\lambda}\varphi_A' - |^{\lambda}F''_A|\sin{}^{\lambda}\varphi''_A, \quad (3c)$$

and

$$^{\lambda}y = |^{\lambda}F'_A|\sin{}^{\lambda}\varphi'_A + |^{\lambda}F''_A|\cos{}^{\lambda}\varphi''_A. \quad (3d)$$

Squaring equation (2), and after some straightforward rearrangements, one gets the equation

$$^{\lambda}I(\mathbf{h}) = |^{\lambda}F_T(\mathbf{h})|^2 = (X + {}^{\lambda}x)^2 + (Y + {}^{\lambda}y)^2. \quad (4)$$

The intensity of the mate reflection $-\mathbf{h}$ is obtained by inverting the sign preceding the term in F'' in expressions (3). Each intensity measurement pertaining to the anomalous pair $\pm\mathbf{h}$ at the wavelength λ will provide such an equation and will contribute to the determination of the Fourier coefficient $^0F_T(\mathbf{h})$.

The scheme leading to Fourier coefficients suitable for the calculation of an electron-density map from MAD measurements is as follows. For the n measurements pertaining to $\pm\mathbf{h}$, it is assumed that estimations of the parameters in $^{\lambda}x$ and $^{\lambda}y$ are available. Then the 'best' values of X and Y in equation (4) (X_B and Y_B), which are optimal with respect to some specified criterion, are determined taking into account all intensity measurements, their error distributions and error distributions for $^{\lambda}x$ and $^{\lambda}y$. The 'best' coefficient, $F_B(\mathbf{h})$, to be used in the Fourier synthesis is

$$F_B(\mathbf{h}) = [^0F_T(\mathbf{h})]_B = (X_B + iY_B). \quad (5)$$

If there is only one atomic species with significant anomalous scattering, the equations are obtained from (4) after rotating the reference axis by $^0\varphi_A$, giving

$$^{\lambda}I(\pm\mathbf{h}) = |^{\lambda}F_T(\pm\mathbf{h})|^2 = (X' + {}^{\lambda}x')^2 + (Y' \pm {}^{\lambda}y')^2, \quad (6)$$

where

$$^{\lambda}x' = |^0F_A|\,{}^{\lambda}f'/f^0,$$

$$^{\lambda}y' = |^0F_A|\,{}^{\lambda}f''/f^0,$$

$$X' = |^0F_T|\cos(^0\varphi_T - {}^0\varphi_A),$$

and

$$Y' = |^0F_T|\sin(^0\varphi_T - {}^0\varphi_A).$$

This set of equations is equivalent to the set of equations (1). They were derived independently on the basis of Karle (1980), and their first applications were reported by Fourme, Kahn & Chiadmi (1985).

Derivation of $F_B(\mathbf{h})$

Let us define the random variable $F(\mathbf{h}) = X(\mathbf{h}) + iY(\mathbf{h})$. To each possible set $\{F(\mathbf{h})\}$, where $F(\mathbf{h})$ can take any value in the complex plane, we associate a probability distribution $p[\{F(\mathbf{h})\}]$ which reflects the uncertainties on the measurements and on the parameters of the partial structure. This distribution is more precisely defined in the next section; as for all other distributions in the text, its normalization is implicit.

The criterion chosen in this article for determining $F_B(\mathbf{h})$ from this distribution is taken from Blow & Crick (1959). It consists of minimizing the r.m.s. (root mean square) of the discrepancy between the Fourier transform $\rho_B(\mathbf{r})$ calculated from a unique set of Fourier coefficients $\{F_B(\mathbf{h})\}$ and the continuum of the transforms $\rho[\mathbf{r},\{F(\mathbf{h})\}]$. Each of these transforms derives from the set of coefficients $\{F(\mathbf{h})\}$ and is weighted by the probability distribution $p[\{F(\mathbf{h})\}]$ associated with this set.

Let us define $\Delta\rho[\mathbf{r},\{F(\mathbf{h})\}]$ as $\rho_B(\mathbf{r}) - \rho[\mathbf{r},\{F(\mathbf{h})\}]$. According to Blow & Crick, its variance $\langle\Delta\rho^2\rangle$ over a unit cell, for a given set, is

$$\langle\Delta\rho^2\rangle_{\text{cell}} = 1/V^2\sum_{\mathbf{h}}|F_B(\mathbf{h}) - F(\mathbf{h})|^2.$$

The weighted mean over all sets is

$$\langle\Delta\rho^2\rangle = \int p[\{F(\mathbf{h})\}]\langle\Delta\rho^2\rangle_{\text{cell}}\,\mathrm{d}\xi,$$

where $\mathrm{d}\xi$ is the differential element of the $2m$-dimensional space of $\{F(\mathbf{h})\}$, and m is the number of reflections.

In the approximation that all $F(\mathbf{h})$ are distributed independently, *i.e.*

$$p[\{F(\mathbf{h})\}] = \prod_{\mathbf{h}} p[F(\mathbf{h})],$$

then

$$\langle\Delta\rho^2\rangle = 1/V^2\sum_{\mathbf{h}}\{\int|F_B(\mathbf{h}) - F(\mathbf{h})|^2 p[F(\mathbf{h})]\,\mathrm{d}X(\mathbf{h})\,\mathrm{d}Y(\mathbf{h})\},$$

or, with a simplified notation

$$\langle \Delta\rho^2 \rangle = 1/V^2 \sum_{\mathbf{h}} \langle |F_B(\mathbf{h}) - F(\mathbf{h})|^2 \rangle.$$

The minimum of this quantity is obtained when every term of the sum is minimal. As each term can be written as

$$\langle |F_B(\mathbf{h}) - F(\mathbf{h})|^2 \rangle = X_B^2 + Y_B^2 - 2(X_B\langle X\rangle + Y_B\langle Y\rangle) + \langle X^2 + Y^2 \rangle,$$

its minimum is obtained when X_B and Y_B satisfy the relations

$$X_B = \langle X \rangle$$

and

$$Y_B = \langle Y \rangle.$$

Then

$$F_B(\mathbf{h}) = \langle F(\mathbf{h}) \rangle = \int\int F(\mathbf{h}) p[F(\mathbf{h})] \mathrm{d}X \mathrm{d}Y.$$

The optimal structure factor F_B, which can be used as such in the calculation of an electron-density map, is then given by

$$|F_B| = (X_B^2 + Y_B^2)^{1/2}$$

and

$$\varphi_B = \mathrm{Arg}\,(X_B + iY_B).$$

Determination of the probability distribution $p[F(\mathbf{h})]$

Conditional probability distributions

On the basis of (4), the probability distribution of $F(\mathbf{h})$ is directly deduced from the probability distribution, $p(\mathbf{I})$, of the intensity, $\mathbf{I} = {}^{\lambda}I(\mathbf{h})$, by taking into account the parameters ${}^{\lambda}F'_A$, ${}^{\lambda}F''_A$ and the scale factors for each data set. These quantities can be expressed from the global parameters, noted as G_i, which are ${}^{\lambda}f'_k$, ${}^{\lambda}f''_k$, occupancies, coordinates and temperature factors of the anomalous scatterers and the scale factors ${}^{\lambda}K$.

The expression of the probability distribution of the structure factor also depends on the type of reflection.

(i) *Acentric phases.* To get the probability distribution of $F = (X + iY)$, we use the change of variables

$$X + {}^{\lambda}x = (I)^{1/2} \cos\psi$$

and

$$Y + {}^{\lambda}y = (I)^{1/2} \sin\psi,$$

where $I = {}^{\lambda}I(\mathbf{h})$ and ψ is the phase of ${}^{\lambda}F(\mathbf{h})$. The variables I and ψ being independent,

$$p(I,\psi) = p(I)p(\psi).$$

Assuming that we have no prior information on $p(\psi)$, it is chosen uniform. Thus,

$$p(\psi) = 1/2\pi,$$

and

$$p(I,\psi) = (1/2\pi)p(I).$$

The elementary conditional probability distribution of the structure factor is

$$p(X,Y|\ldots G_i\ldots) = 2p(I,\psi) = (1/\pi)p(I).$$

Hence, this probability is proportional to $p(I)$.

(ii) *Centric phases.* Since the phases are constrained to a value φ_R modulo π, $F(\mathbf{h})$ is a unidimensional variable, so it is convenient to multiply (2) by $\exp(-i\varphi_R)$. We define the parameters

$$X_c = |{}^{0}F_T| \cos({}^{0}\varphi_T - \varphi_R),$$

$${}^{\lambda}x_c = |{}^{\lambda}F'_A| \cos({}^{\lambda}\varphi'_A - \varphi_R),$$

and

$${}^{\lambda}y_c = |{}^{\lambda}F''_A| \cos({}^{\lambda}\varphi''_A - \varphi_R),$$

all of these cosine values being ± 1. Thus, the intensity has the form

$${}^{\lambda}I = (X_c + {}^{\lambda}x_c)^2 + {}^{\lambda}y_c{}^2,$$

and the probability distribution of $F = X_c \exp(i\varphi_R)$ is

$$p(F|\ldots G_i\ldots) = p(X_c|\ldots G_i\ldots) = 2|X_c + {}^{\lambda}x_c| p({}^{\lambda}I).$$

Combination of distributions

At this stage, a set of conditional probability distributions of the structure factor, deriving from those of the intensities, has been obtained. In essentially all published studies using probability methods for the derivation of phases from MAD measurements, the following approximations have been made: (i) independence of the information associated with the individual observations ${}^{\lambda}I(\pm\mathbf{h})$; (ii) no systematic errors on these observations; (iii) no errors on the global parameters G_i. Then, with p_j referring to the elementary probability distribution associated with the measurement j ($j = 1\ldots n$) pertaining to $\pm\mathbf{h}$:

for a reflection with a centric phase

$$p(X_c) = \prod_j p_j(X_c|\ldots, G_i, \ldots) \propto \prod_j |X_c + {}^{\lambda}x_c| p_j({}^{\lambda}I);$$

for a reflection with an acentric phase

$$p(X,Y) = \prod_j p_j(X,Y|\ldots, G_i, \ldots) \propto \prod_j p_j({}^{\lambda}I).$$

In fact, one has to take into account the uncertainties on the global parameters G_i. These uncertainties are expressed by the probability distribution noted $p(\ldots, G_i, \ldots)$ which can be estimated, for instance, from the variance–covariance matrix obtained

during the refinement procedure of the parameters G_i.

Then, for an acentric reflection, the probability distribution of the structure factor is (Einstein, 1977)

$$p(X,Y)=\int\ldots\int p(\ldots,G_i,\ldots)\prod_j p_j(X,Y|\ldots,G_i,\ldots)\ldots \mathrm{d}G_i\ldots.$$

For a centric reflection, the probability distribution of the structure factor is

$$p(X_c)=\int\ldots\int p(\ldots,G_i,\ldots)\prod_j p_j(X_c|\ldots,G_i,\ldots)\ldots \mathrm{d}G_i\ldots.$$

Figure of merit

In the conventional error treatment in the MIR method (Blow & Crick, 1959), the amplitude of each Fourier coefficient is weighted by a figure of merit. In our case, the figure of merit does not appear explicitly in the expression of F_B, but it can be defined as follows.

(i) Let us first consider the situation (as in the conventional treatment of errors in the MIR method) where the probability distributions of the modulus and of the phase of every Fourier coefficient $F=|F|\exp i\varphi$ are independent, *i.e.*

$$p(F)=p(|F|,\varphi)=p(|F|)p(\varphi).$$

Then

$$F_B=\langle|F|\exp i\varphi\rangle=\langle|F|\rangle\langle\exp i\varphi\rangle=m\langle|F|\rangle.$$

$\langle\Delta\rho^2\rangle$ can be expressed as a function of the figure of merit, $|m|$, as

$$\langle|F_B-F|^2\rangle=\langle|m\langle|F|\rangle-|F|\exp i\varphi|^2\rangle=\langle|F|^2\rangle-|m|^2\langle|F|\rangle^2,$$

and, since in the MIR method the modulus is a measured quantity, $|F_P|$, one obtains the usual expression

$$\langle\Delta\rho^2\rangle=1/V^2\sum_{\mathbf{h}}|F_p(\mathbf{h})|^2[1-|m(\mathbf{h})|^2]. \qquad (7)$$

(ii) If the probability distribution is a function of both the modulus and the phase, then

$$\langle|F_B-F|^2\rangle=\langle|\langle F\rangle-F|^2\rangle=\langle|F|^2\rangle-|\langle F\rangle|^2,$$

and thus

$$\langle\Delta\rho^2\rangle=1/V^2\sum_{\mathbf{h}}\langle|F|^2\rangle\left(1-\frac{|\langle F\rangle|^2}{\langle|F|^2\rangle}\right).$$

By analogy with equation (7), the figure of merit, $|m|$, is

$$|m|=\frac{|\langle F\rangle|}{\langle|F|^2\rangle^{1/2}}=\frac{|F_B|}{\langle|F|^2\rangle^{1/2}}. \qquad (8)$$

Probability distribution of the intensity

The formalism developed above uses any valuable expression of the probability distribution of the intensity. The use of a Gaussian distribution is justified by the following considerations. The uncertainty in an intensity measurement must take into account the counting statistics (Poisson distribution) and other errors as a result of intensity fluctuations of the X-ray beam, absorption, extinction, radiation damage *etc.* The estimation of this uncertainty can be made from the measurements of equivalent reflections, which gives a more realistic evaluation than taking into account only the counting statistics. In this case, French & Wilson (1978) suggest that the resulting intensity follows a quasi-normal distribution. The probability distribution of the measurement $\mathcal{I}$ of the intensity I is then

$$p(\mathcal{I}|I)\propto\exp-[(\mathcal{I}-I)^2/(2\sigma_I^2)],$$

where I is given by equation (4), and σ_I is the standard deviation of errors associated with the distribution $\delta=\mathcal{I}-I$.

The intensity I can only be known by means of a probability distribution, which is, according to Bayes rule

$$p(I|\mathcal{I})\propto p(\mathcal{I}|I)p_0(I),$$

where the function p_0 represents the information known *a priori* on the quantity I. In the narrow range of I in which $p(\mathcal{I}|I)$ is significantly different from 0, $p_0(I)$ can be considered uniform. The distribution $p(I|\mathcal{I})$ is the only information available about I and has been already quoted as $p(I)$. Then

$$p(I)\propto\exp[-(\mathcal{I}-I)^2/(2\sigma_I^2)].$$

The value of the standard deviation σ_I is estimated from the standard deviation $\sigma_{\mathcal{I}}$ of the intensity measurements.

Representation of the phase information with *A, B, C, D* coefficients

In order to combine the information from the MAD source with the information from other sources, the probability distribution of the structure factor is expressed in practical terms as a phase distribution, $f(\varphi)$, associated with a single modulus value.

The best phase value is calculated from $f(\varphi)$, which is determined from $p(|F|,\varphi)$, and is expressed as

$$\varphi_B=\mathrm{Arg}\,[\int f(\varphi)\exp(i\varphi)\mathrm{d}\varphi].$$

This equation could be written as

$$|m|\exp(i\varphi_B)=\int f(\varphi)\exp(i\varphi)\mathrm{d}\varphi,$$

which corresponds to the usual definition of the figure of merit.

In the case where the probabilities of the modulus and of the phase are decoupled, the function, $f(\varphi)$, is simply the margin distribution $p(\varphi)$ of $p(|F|,\varphi)$ (*i.e.* $\int|F|p(|F|,\varphi)\mathrm{d}|F|$).

We derive below the expression of $f(\varphi)$ in the general case where the phase and the modulus are coupled. As

$$F_B = |F_B| \exp(i\varphi_B) = |m|\langle|F|^2\rangle^{1/2} \exp(i\varphi_B)$$
$$= \int\int |F| \exp(i\varphi) p(|F|,\varphi) |F| \mathrm{d}|F| \mathrm{d}\varphi,$$

then

$$f(\varphi) = (1/\langle|F|^2\rangle^{1/2}) \int |F|^2 p(|F|,\varphi) \mathrm{d}|F|.$$

$f(\varphi)$ can be expressed using A, B, C, D coefficients (Hendrickson & Lattman, 1970) as

$$f(\varphi) \propto \exp[A\cos(\varphi) + B\sin(\varphi) + C\cos(2\varphi) + D\sin(2\varphi)],$$

where

$$A = \int \ln[f(\varphi)] \cos(\varphi) \mathrm{d}\varphi,$$
$$B = \int \ln[f(\varphi)] \sin(\varphi) \mathrm{d}\varphi,$$
$$C = \int \ln[f(\varphi)] \cos(2\varphi) \mathrm{d}\varphi$$

and

$$D = \int \ln[f(\varphi)] \sin(2\varphi) \mathrm{d}\varphi.$$

Implementation

Our first experience with the use of probability methods for the analysis of MAD data occurred with the crystal structure study of a 10 kDa protein, the *Opsanus tau* parvalbumin (Kahn *et al.*, 1985; Kahn, Fourme, Bosshard, Chiadmi *et al.*, 1986). Tb-labelled parvalbumin crystals were prepared. Bijvoet pairs were measured at three wavelengths close to the Tb L_{III} absorption edge, using synchrotron radiation and the spherical drift proportional chamber of the MARK I diffractometer (Kahn *et al.*, 1982). Inter-scaling between the six data sets was performed, and a quasi-absolute scaling was applied. The Tb partial structure was solved by anomalous Patterson techniques (Rossmann, 1961) and refined with the program *ANOLSQ* (Hendrickson & Teeter, 1981). The phasing technique was derived from MIR methods, as described by Phillips & Hodgson (1980), and implemented in the program *MWASD* written by one of us (RF). Reflections **h** recorded at the wavelength λ_1 at which ${}^{\lambda}f''$ is maximum were used as a pseudo-native data set. The total phase probability $P(\varphi)$ was taken as the product of the distributions formed from all the possible pairings of the pseudo-native set with the other sets. As we had six data sets (hkl and $h\bar{k}l$ at three wavelengths), it was possible to calculate five 'lack-of-closure' values $\varepsilon(\varphi)$ for each reflection of the pseudo-native set. In order to get a real electron-density map, the complex number ${}^{\lambda_1}F_T$ was corrected for the imaginary wavelength-dependent part, ${}^{\lambda_1}F''_A$, of the partial structure. Finally, the electron-density map was calculated with corrected centroid phases and figure-of-merit-weighted amplitudes. This probability method had several drawbacks: (i) it overestimates the role of the pseudo-native set; (ii) Fourier coefficients have to be corrected in order to give a real electron density (the current version of *MWASD* gives ${}^{0}F_T$ coefficients); (iii) only the phase of the Fourier coefficient is treated as a random variable. In order to overcome these problems, the formalism of equations of type (6) was developed (Fourme *et al.*, 1985) after the seminal ideas of Karle (1980) and used for the treatment of centric reflections in the parvalbumin study, as mentioned by Kahn *et al.* (1985). The next step, in 1986, was a program called *MADBEST*, designed for a single type of anomalous scatterer and able to treat acentric as well as centric phases with the formalism of (6). *MADBEST* assumes that the partial structure has been solved and refined, and that ${}^{\lambda}f'$ and ${}^{\lambda}f''$ values of sufficient accuracy have been determined, which provides values for ${}^{\lambda}x$ and ${}^{\lambda}y$ in each equation of type (6). Intensity data sets must be submitted to the usual scaling procedures and put preferably on a quasi-absolute scale. Under the assumptions and simplifications given in *Combinations of distributions*, *MADBEST* calculates the total bidimensional phase probability distributions and Fourier coefficients F_B with figures of merit. As an example, results of MAD phasing with *MADBEST* for one reflection **h** are given in Figs. 2(*a*) to 2(*c*) for a hypothetical example with error-free values, except for Gaussian errors on intensity measurements.

In order to test bidimensional integration, three sets of calculated structure factors (each of them including Bijvoet pairs) were prepared, using the atomic model of the refined structure of the *Opsanus tau* parvalbumin (space group $P2_12_12$) refined at 2 Å resolution (Chiadmi, 1991); the model included a single Tb^{3+} ion with the anomalous-scattering-factor values reported in the paper of Kahn *et al.* (1985) (respectively $f' = -18, -14, -24$; $f'' = 19, 11, 10$). Six 'ideal' intensity data sets were obtained by squaring these structure amplitudes, and errors with a Gaussian distribution of adjustable variance were added to produce intensity data sets with random errors. *MADBEST* was applied to these simulated data, using error-free ${}^{\lambda}x$ and ${}^{\lambda}y$ values. The mean-squared differences of the phase $\langle\Delta\varphi^2\rangle$, where $\Delta\varphi = \varphi_B - \varphi_{\text{exact}}$, and density $\langle\Delta\rho^2\rangle$ were calculated for various simulated intensity data sets with increasing errors. The same data were processed with *MADBEST* and *MADABCD* (Pähler *et al.*, 1990). The resulting errors in phase and in electron density, shown in Table 1, are lower with *MADBEST*. The improvement of $\langle\Delta\varphi^2\rangle$ and $\langle\Delta\rho^2\rangle$ is marginal for very accurate data, but becomes better for data with greater errors (Table 1). Phase distributions obtained

for a particular reflection, using *MADBEST* and *MADABCD*, respectively, are shown in Fig. 3.

Discussion and perspectives

The enzyme cutinase (molecular weight 22 kDa) was selected as a test with real data. Crystals of this enzyme (space group $P2_1$) diffract to very high resolution. The three-dimensional structure had been solved and refined at 1.6 Å resolution (Martinez, De Geus, Lauwereys, Matthyssens & Cambillau, 1992). We used a single-site (occupancy $\simeq 0.6$) mercury derivative of a point mutant, introducing a cysteine residue. Data at 1.7 Å resolution were measured on a single crystal of this derivative at three wavelengths close to the Hg L_{III} absorption edge, using the MARK II diffractometer at LURE (Kahn, Fourme, Bosshard & Saintagne, 1986). Values for the Hg anomalous-scattering factors at the three wave-

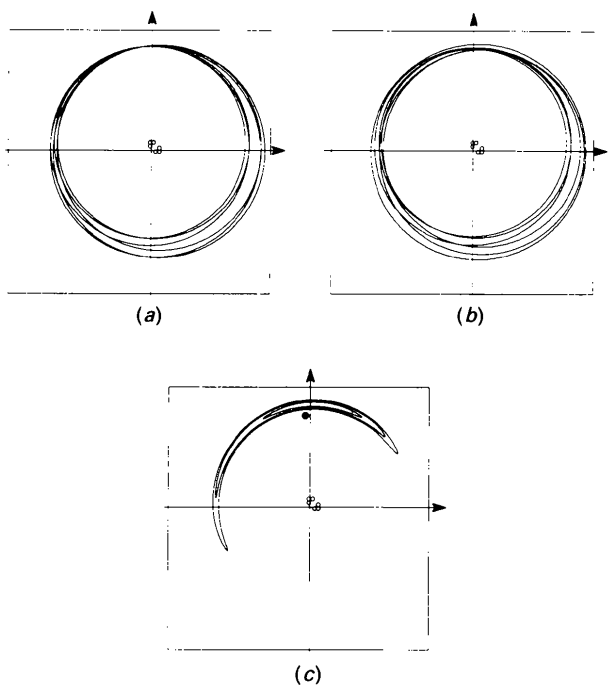

Fig. 2. Hypothetical example of MAD phasing of a reflection **h** from Bijvoet-pair measurements at three wavelengths with a single type of anomalous scatterer, (*a*) assuming the following error-free values: $|^0F_T| = 175$; $|^0F_A| = 40$; $^0\varphi_T = 105°$; $^0\varphi_A = 45°$; $|^{A_1}F'_A| = 5$; $|^{A_2}F'_A| = 10$; $|^{A_3}F'_A| = 15$; $|^{A_1}F''_A| = 7.5$; $|^{A_2}F''_A| = 12.5$; $|^{A_3}F''_A| = 7.5$, plots of six circles of radius $|^AF_T(\pm\mathbf{h})|$ with centres at $(-^Ax, -^Ay)$. (*b*) same as (*a*), but Gaussian-distributed errors were added to intensities with $\sigma[^AI_T(\pm\mathbf{h})]/^AI_T(\pm\mathbf{h}) = 0.075$. (*c*) Total probability distribution $p(X,Y)$ corresponding to (*b*). Isovalue curves are drawn at σ, 3σ and 5σ. The modulus and the phase are 198.0 and 94.0°, respectively, for the true F, 196.2 and 84.4° for the most probable F, 174.3 and 93.1° for F_B. The figure of merit is 0.90. The black dot shows the extremity of F_B.

Table 1. *Test of probability methods using 2 Å intensity data calculated from the atomic model of Opsanus tau parvalbumin with Gaussian-distributed errors added*

$|^0F_A|$, $^Af'$ and $^Af''$ values are error free. UPM and BPM: unidimensional (program *MADABCD*) and bidimensional (program *MADBEST*) probability methods, respectively. $\Delta\varphi$: r.m.s. difference between the phase estimated by a probability method and the phase calculated from the model. $\Delta\rho$: r.m.s. difference over the unit cell between the electron density calculated from the probability method and the electron density of the model. N_{ref}: number of unique acentric reflections **h** in the data set (each of these reflections has been generated twice, together with an equal number of reflections $-\mathbf{h}$).

$$R_{sym} = \sum_{\mathbf{h}}\sum_{i}|I_i(\mathbf{h}) - \langle I(\mathbf{h})\rangle| / \sum_{\mathbf{h}}\sum_{i} I_i(\mathbf{h}).$$

	$\Delta\varphi$ (°)	$\Delta\rho$ (e Å^{-3})
(*a*) $N_{ref} = 4910$; $R_{sym} = 0.028$		
BPM	17.3	0.062
UPM	17.4	0.063
(*b*) $N_{ref} = 4918$; $R_{sym} = 0.056$		
BPM	29.2	0.093
UPM	30.0	0.095
(*c*) $N_{ref} = 4924$; $R_{sym} = 0.084$		
BPM	38.0	0.112
UPM	39.4	0.115

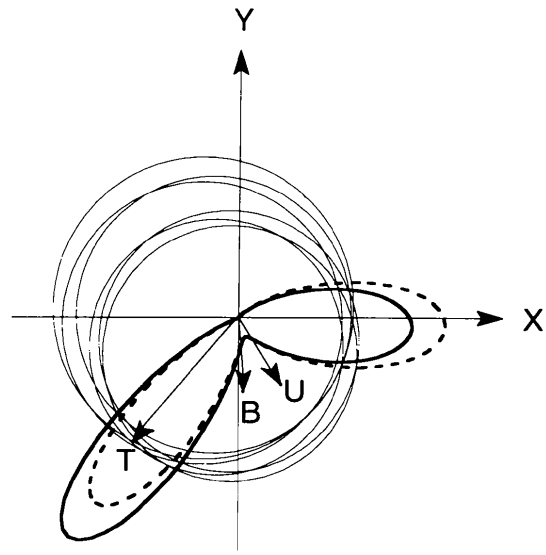

Fig. 3. Phase distributions for a particular Bragg reflection: $f(\varphi)$ (solid line) from *MADBEST* and $p(\varphi)$ (dashed curved line) from *MADABCD*. The two functions are normalized and their amplitude is proportional to the distance from the centre. $\Delta\rho$ is the contribution of the reflection to the total error of the electron density. Data are simulated from the *Opsanus tau* parvalbumin structure (see text) with $\sigma(I)/I = 0.15$ at three wavelengths (six circles). The true structure factor (T) is: $|F| = 142.0$, $\varphi = 230°$. The best structure factors are: (i) using *MADBEST* (*B*): $|F| = 64.6$, $\varphi = 274°$ ($|m| = 0.53$, $\Delta\rho = 1.3 \times 10^{-3}$ e Å^{-3}). (ii) *MADABCD* (*U*): $|F| = 69.9$, $\varphi = 303°$ ($|m| = 0.50$, $\Delta\rho = 1.7 \times 10^{-3}$ e Å^{-3}).

lengths were $f' = -18, -14, -16$ and $f'' = 7, 10, 4$. The Hg site could be readily located; atomic coordinates, occupancy and an isotropic temperature factor were refined. Taking into account only measurement errors, the two electron densities built with phases given by *MADBEST* and *MADABCD* are of comparable quality; although the molecular envelope and some structural details are very clear, pieces of secondary structure are difficult to interpret. After a detailed analysis of results, taking into account the sequence of batches used for the three-wavelength data collection, it became clear that the crucial point to improve the quality of MAD phases was refinement of the parameters of the partial structure, especially scattering factors $^{\lambda}f'$ and $^{\lambda}f''$. Because of the monochromator of the experimental setup, we had difficulty taking into account the fine structure of the absorption edge and in maintaining and reproducing precisely defined values of the wavelength (see Weis, Kahn, Fourme, Drickamer & Hendrickson, 1991, for a similar case). A least-squares method based on the idea that structure-factor estimates for acentric reflections are implicit functions of the parameters to be refined (Bricogne, 1982, 1984) was first used. This method was effective, but some problems with instabilities still remained (De La Fortelle, Martinez, Kahn & Fourme, 1992). These difficulties are similar to those found in the MIR method with respect to the refinement of occupancies from acentric reflections. An improved method for refining the global parameters was then developed. As shown by Bricogne (1991*a*,*b*), the probability distribution for structure-factor moduli is best described by a Rice function instead of a Gaussian. The Rice function provides the analytical basis for a maximum-likelihood refinement of the global parameters of the partial structure which overcomes well known difficulties in the problem of bias-free refinement of heavy-atom parameters. The integrated value of a Rice-based likelihood distribution over the complex plane is an optimal estimator of the relevance of the partial structure (Bricogne, 1991*a*). A computer program, implementing this new treatment and using several features developed in this article – especially the analysis in the complex plane of the phase-probability function – has been written, in collaboration with G. Bricogne (to be published). It is designed to accommodate both MIR and MAD data or any mixture of the two. Analysis of the cutinase data will be a severe test, as this is a typical case of MAD analysis with systematic – albeit well identified – errors and low anomalous and dispersive ratios.

References

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

BRICOGNE, G. (1982). *Computational Crystallography*, edited by D. SAYRE, pp. 223–230. New York: Oxford Univ. Press.

BRICOGNE, G. (1984). *Methods and Applications in Crystallographic Computing*, edited by S. R. HALL & T. ASHIDA, pp. 141–151. Oxford: Clarendon Press.

BRICOGNE, G. (1991*a*). *Isomorphous Replacement and Anomalous Scattering, Proceedings of the CCP4 Study Weekend, 25–26 January* 1991, pp. 60–68. Warrington, England: SERC Daresbury Laboratory.

BRICOGNE, G. (1991*b*). *Crystallographic Computing*, edited by D. MORAS, A. D. PODJARNY & J. C. THIERRY, pp. 257–297. Oxford: Clarendon Press.

CHIADMI, M. (1991). Doctoral Thesis, Univ. Paris-Sud, Orsay, France.

DE LA FORTELLE, E., MARTINEZ, C., KAHN, R. & FOURME, R. (1992). Abstr. 4th Int. Conf. Biophys. Synchrotron Radiat., Tsukuba, Japan, p. 201.

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

FOURME, R. & HENDRICKSON, W. A. (1990). *Synchrotron Radiation and Biophysics*, edited by S. S. HASNAIN, pp. 156–175. Chichester: Ellis Horwood.

FOURME, R., KAHN, R. & CHIADMI, M. (1985). In *LURE Activity Report* 1983 1985, p. 74. LURE, Orsay, France.

FRENCH, S. & WILSON, K. (1978). *Acta Cryst.* A**34**, 517–525.

GUSS, J. M., MERRITT, E. A., PHIZACKERLEY, R. P., HEDMAN, B., MURATA, M., HODGSON, K. O. & FREEMAN, H. C. (1988). *Science*, **241**, 806–811.

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

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

HENDRICKSON, W. A. & LATTMAN, E. E. (1970). *Acta Cryst.* B**26**, 136–143.

HENDRICKSON, W. A., PÄHLER, A., SMITH, J. L., SATOW, Y., MERRITT, E. A. & PHIZACKERLEY, R. P. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 2190–2194.

HENDRICKSON, W. A. & TEETER, M. M. (1981). *Nature (London)*, **290**, 107–113.

KAHN, R., FOURME, R., BOSSHARD, R., CAUDRON, B., SANTIARD, J.-C. & CHARPAK, G. (1982). *Nucl. Instrum. Methods*, **201**, 203–208.

KAHN, R., FOURME, R., BOSSHARD, R., CHIADMI, M., RISLER, J.-L., BRUNIE, S., WERY, J.-P., DIDEBERG, O. & JANIN, J. (1986). *Structural Biological Applications of X-ray Absorption, Scattering and Diffraction*, edited by B. CHANCE & H. D. BARTUNIK, pp. 297–308. London: Academic Press.

KAHN, R., FOURME, R., BOSSHARD, R., CHIADMI, M., RISLER, J.-L., DIDEBERG, O. & WERY, J.-P. (1985). *FEBS Lett.* **178**(1), 133–137.

KAHN, R., FOURME, R., BOSSHARD, R. & SAINTAGNE, V. (1986). *Nucl. Instrum. Methods*, A**246**, 596–603.

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

KORSZUN, Z. R. (1988). *J. Mol. Biol.* **196**, 413–419.

MARTINEZ, C., DE GEUS, P., LAUWEREYS, M., MATTHYSSENS, G. & CAMBILLAU, C. (1992). *Nature (London)*, **356**, 615–618.

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

PHILLIPS, J. & HODGSON, K. O. (1980). *Acta Cryst.* A**36**, 856–864.

ROSSMANN, M. G. (1961). *Acta Cryst.* **14**, 383–388.

SMITH, J. L. (1991). *Curr. Opin. Struct. Biol.* **1**, 1002–1011.

WEIS, W. I., KAHN, R., FOURME, R., DRICKAMER, K. & HENDRICKSON, W. A. (1991). *Science*, **254**, 1608–1615.